DOCTOR OF PHILOSOPHY

Development of bioinfirmatic 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

1884

PRIFYSGOL
BANGOR
UNIVERSITY

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.

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*

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*

Chapter 3*

Chapter 4*

Chapter 5*

Chapter 6*

Chapter 7

Appendices

Acknowledgements

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.

Abbreviations

Abbreviation	Definition
AE	Axial element
ANOVA	Analysis of Variance
BLAST	Basic Local Alignment Search Tool
BTB	Blood-testis-barrier
CE	Central element
cDNA	Complementary DNA
CI	Confidence interval
CIBEX	Center for Information Biology Gene Expression Database
CNS	Central nervous system
CO	Crossover
CR	Chromatin remodelling co-repressor complex
CT antigen (or CTA)	Cancer testis antigen
CT gene	Cancer testis gene
CT-X (or X-CT) gene	Cancer testis gene encoded on the X chromosome
D loop	Displacement loop
DDD	Digital Differential Display
dHJ	Double Holliday junction
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
DSB	Double-strand break
DSBR	Double-strand break repair
EBI	European Bioinformatics Institute
EST	Expressed sequence tag
FC	Fold change
FDA	US Food and Drug Administration
G0 phase	Growth 0 phase
G1 phase	Growth 1 phase
GEO	Gene Expression Omnibus
GO	Gene ontology
H3K4me3	Lysine 4 of histone H3
HAT	Histone acetyltransferase

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

SC	Synaptonemal complex
SDSA	Synthesis-dependent strand annealing
SEREX	Serological analysis of cDNA expression libraries
SMD	Stanford Microarray Database
SNP	Single nucleotide polymorphism
ssDNA	Single-strand DNA
TA	Tumour antigen
TF	Transcription factor
tpm	Transcripts per million
TVF	Transverse filament
UTR	Untranslated region
VSN	Variance Stabilization and Normalization
WCE	Whole cell extract

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].

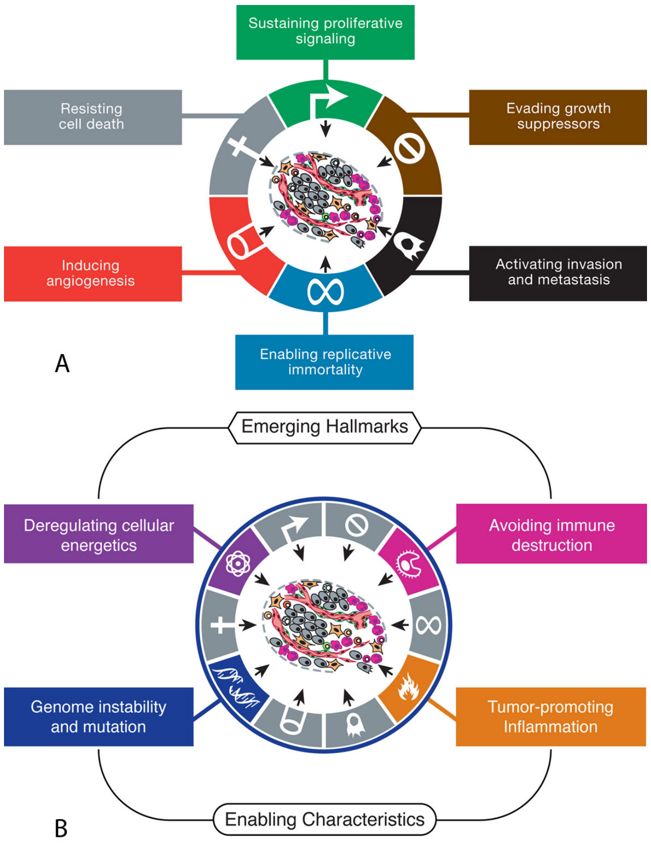

Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality
A
Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting Inflammation
Enabling Characteristics
B

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 identified 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

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 chemo-immunotherapies 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];

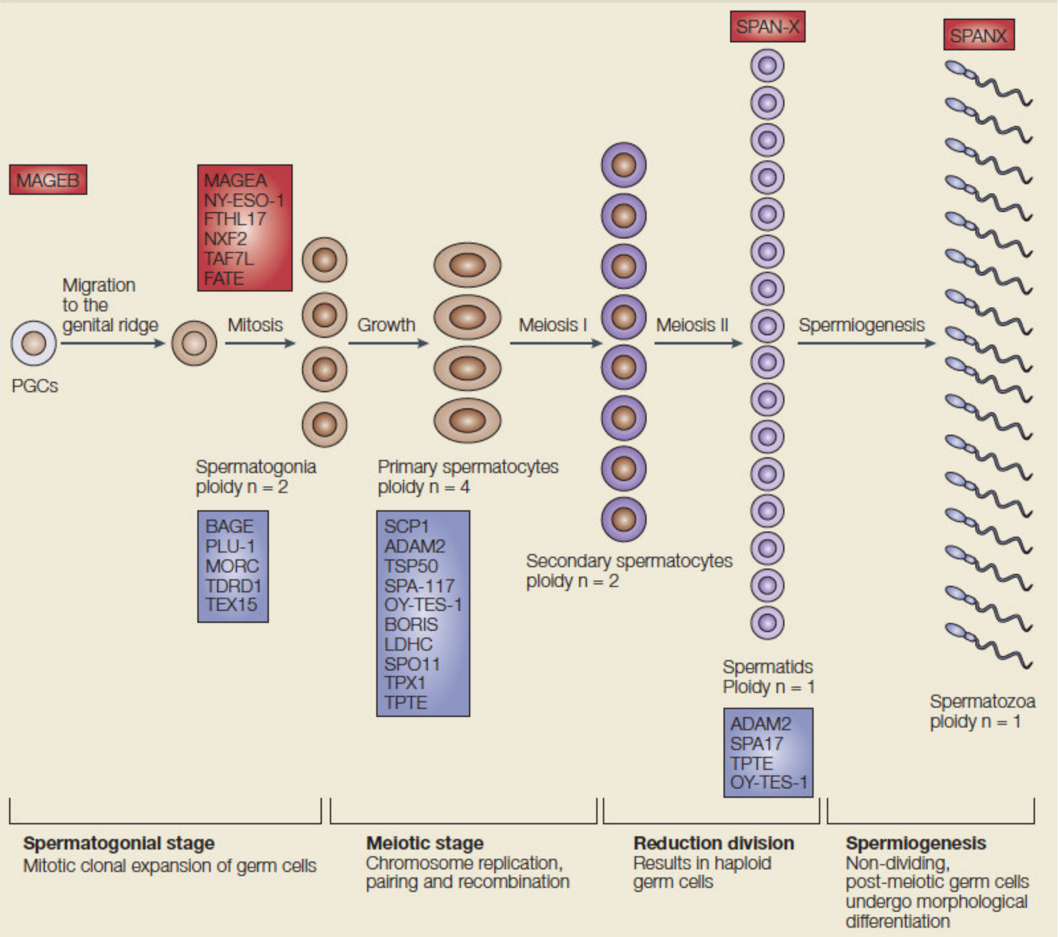


Figure 1.2: Scheme of spermatogenesis and expression of cancer testis (CT) genes in human germ cells [31]. Male primordial germ cells (PGCs) migrate to the genital ridge, proliferate and arrest until puberty. At puberty, spermatogenesis is initiated, which continuously maintains spermatogonia through 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 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].

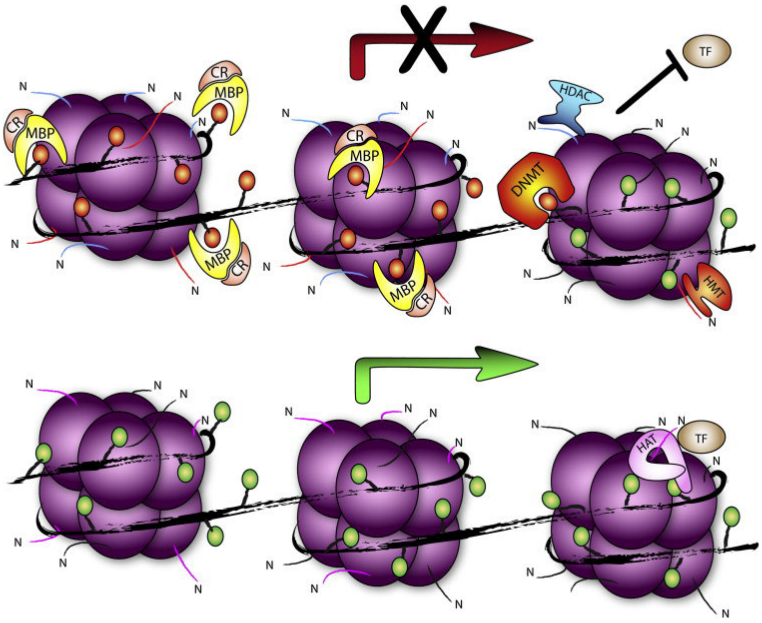


Figure 1.3: Epigenetic events regulating cancer testis (CT) gene expression [32]. Hypermethylation of CT gene promoters causes gene silencing in normal somatic cells by means of two mechanisms. First, methylation represents a physical barrier to transcription factors (TFs). Methylation predominately 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.

- Prophase 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

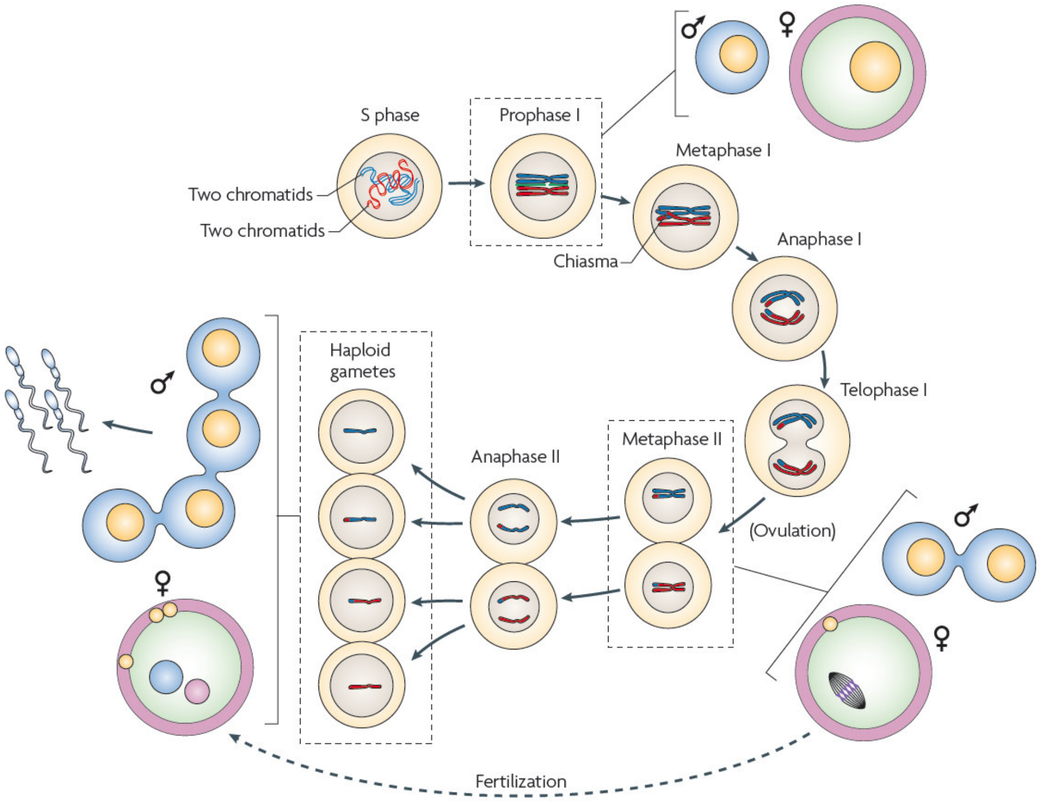


Figure 1.4: Overview of the meiotic process during mammalian gametogenesis [96]. The distinct meiotic phases of the first and second meiotic division are schematically illustrated (Interphase, Prophase, Metaphase, Anaphase and Telophase). Meiosis includes two rounds of cell division resulting in four 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].

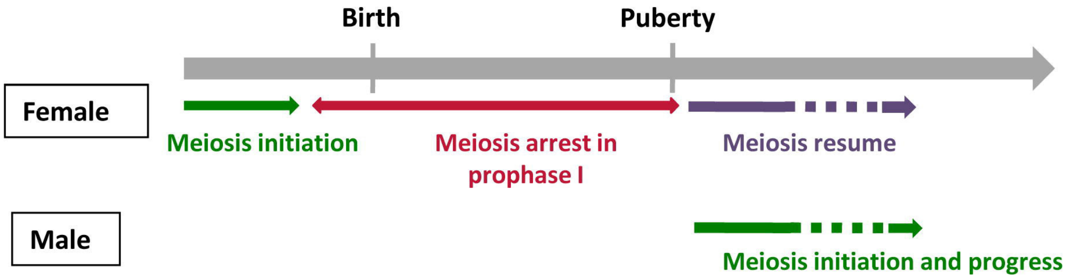


Figure 1.5: Timing of gametogenesis/meiosis in mammalian males and females [109]. There are crucial differences in regulation and timing of gametogenesis in distinct sexes. In males, meiosis is initiated at puberty and occurs continuously, whereas in females, meiosis begins already in the foetal ovaries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

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].

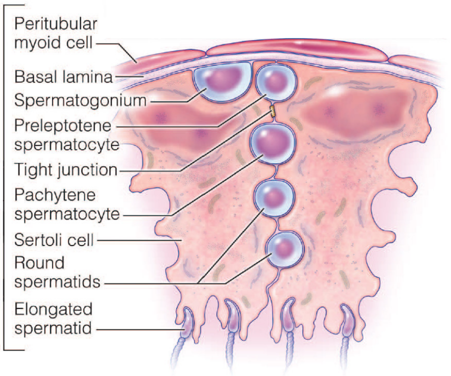


Figure 1.6: Schematic composition of the seminiferous epithelium in the mammalian testes [122]. The illustration shows the developing germ cells at different stages of differentiation surrounded by supporting Sertoli cells [96, 117]. Tight junctions forming the blood-testis-barrier (BTB) separate the adluminal compartment with the differentiating spermatocytes from the rest of the body, providing an immunologically privileged site for meiotic and postmeiotic cells [49, 122].

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].

1.2.4 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].

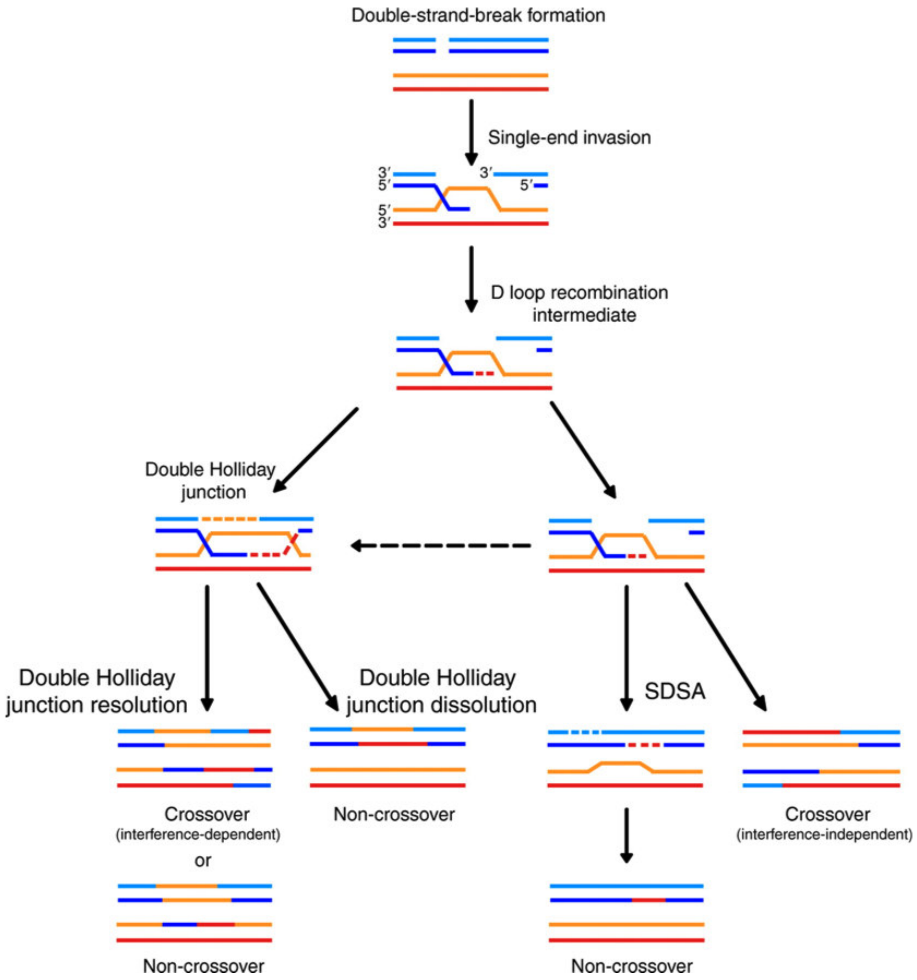

Double-strand-break formation
Single-end invasion
3′
5′
5′
3′
3′
5′
D loop recombination intermediate
Double Holliday junction
Double Holliday junction resolution
Double Holliday junction dissolution
SDSA
Crossover
(interference-dependent)
or
Non-crossover
Non-crossover
Crossover
(interference-independent)
Non-crossover

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].

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

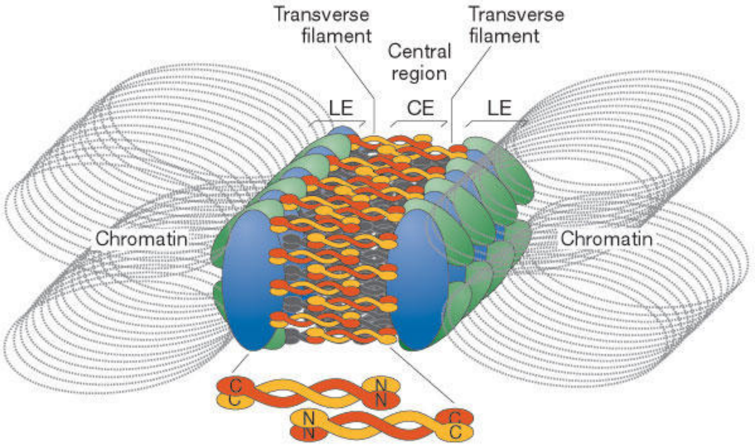


Figure 1.8: Schematic composition of the synaptonemal complex (SC) [99]. The SC is a zipper-like structure consisting of two lateral elements (LEs), which are linked by transverse elements (TVFs) and a central element (CE). It forms and runs between paired homologues and is thought to mediate and maintain pairing as well as to facilitate recombination events at least in some organisms [99, 100].

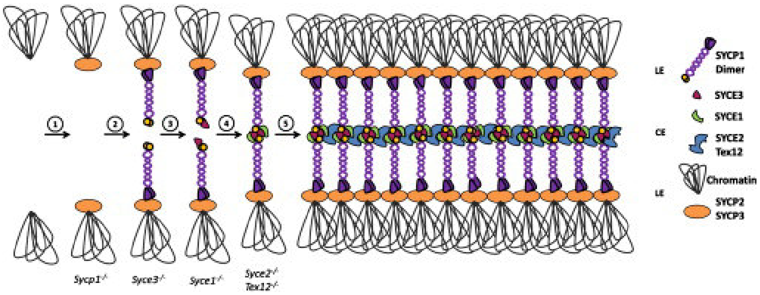


Figure 1.9: Assembly of the synaptonemal complex (SC) [98]. Axial elements (AEs)/lateral elements (LEs) consist of SYCP2, SYCP3, HORMA-domain proteins as well as of cohesin proteins including STAG3, SMC1α, SMC1β, SMC3 and REC8. Transverse elements (TVFs) represent a bridge between LEs and the central 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].

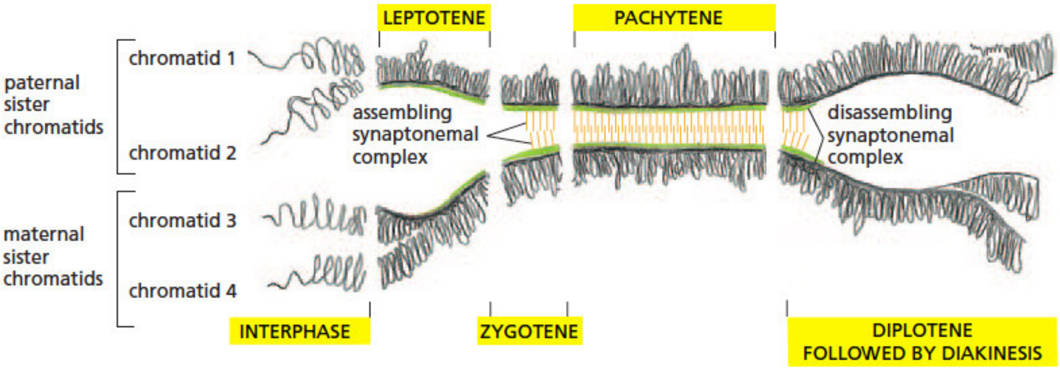

LEPTOTENE
PACHYTENE
paternal sister chromatids
chromatid 1
chromatid 2
assembling synaptonemal complex
disassembling synaptonemal complex
maternal sister chromatids
chromatid 3
chromatid 4
INTERPHASE
ZYGOTENE
DIPLOTENE FOLLOWED BY DIAKINESIS

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: `http://link.springer.com/chapter/10.1007/978-3-7091-0947-2_4` (chapter), `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*†[1,2], Gerhard G. Thallinger‡[2,3], Ramsay J. McFarlane§[1], and Lee D. Larcombe¶[4]

[1]North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK

[2]Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria

[3]Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology (ACIB GmbH), Petersgasse 14, 8010 Graz, Austria

[4]Cranfield 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 Agendia (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 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.

3 Meta-Analysis: "The Analysis of Analyses"

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 (Campain 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).

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 M^{2}DB (Cheng *et al.* 2010) and M^{3D} (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.

Name	**Type**	**Availability**
ArrayExpress R package	R/Bioconductor package to access ArrayExpress	`http://www.bioconductor.org/`
ArrayExpress*	Repository for raw and/or processed data	`http://www.ebi.ac.uk/arrayexpress/`
CIBEX*	Repository for raw and/or processed data	`http://cibex.nig.ac.jp/index.jsp`
GEO*	Repository for raw and/or processed data	`http://www.ncbi.nlm.nih.gov/geo/`
GEOmetadb	Web-based search interface to GEO and R/Bioconductor package	`http://gbnci.abcc.ncifcrf.gov/geo/`
GEOquery R package	R/Bioconductor package to access GEO	`http://www.bioconductor.org/`
L2L	Repository for published gene lists	`http://depts.washington.edu/l2l/`
M^{2}DB	Repository for raw and preprocessed Affymetrix data	`http://metadb.bmes.nthu.edu.tw/m2db/`
M^{3D}	Repository for raw and preprocessed Affymetrix data for three microbial species	`http://m3d.bu.edu/`
MaRe	Web-based search interface to GEO and ArrayExpress	`http://www.lgtc.nl/MaRe/`
SMD*	Repository for raw, preprocessed and processed data and analysis tools for microarray data	`http://smd.stanford.edu/`

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

Name	Type	Availability
Affymetrix annotation files	Annotation files of the manufacturer	`http://www.affymetrix.com/support/technical/annotationfilesmain.affx`
Bioconductor annotation packages	R packages for annotation	`http://www.bioconductor.org/packages/release/data/annotation/`
DAVID	ID conversion, functional annotation and classification	`http://david.abcc.ncifcrf.gov/`
IDconverter	ID converter	`http://idconverter.bioinfo.cnio.es/`
MADGene	ID converter	`http://cardioserve.nantes.inserm.fr/madtools/`
Resourcerer	Annotation for common platforms including comparisons within and across species	`http://compbio.dfci.harvard.edu/cgi-bin/magic/r1.pl`
Source	Mapping of feature identifiers and annotation with additional information from various databases	`http://smd.stanford.edu/cgi-bin/source/sourceSearch`

Table 3: A selection of available meta-analysis tools

Name	Strategy	Implementation	Availability
A-MADMAN	Absolute method	Web platform	`http://compgen.bio.unipd.it/bioinfo/amadman/`
Gene Expression Atlas	Vote counting	Web platform	`http://www.ebi.ac.uk/gxa/`
GeneMeta	Effect size combination	R package	`http://www.bioconductor.org/packages/2.8/bioc/html/GeneMeta.html`
GeneSapiens	Absolute method	Web platform	`http://www.genesapiens.org/`
Genevestigator	Absolute method	Web platform	`https://www.genevestigator.com/`

MAMA	9 in 1 package includes e.g., RankProd and metaMA	R package	`http://cran.r-project.org/src/contrib/Archive/MAMA/`
metaArray	Integrative correlation strategy	R package	`http://www.bioconductor.org/packages/2.10/bioc/html/metaArray.html`
metaMA	Effect size and p-value combination	R package	`http://cran.r-project.org/web/packages/metaMA/`
METRADISC	Rank aggregation	Compaq Visual Fortan90 software	`http://biomath.med.uth.gr`
Oncomine	Vote counting	Web platform	`https://www.oncomine.org/`
RankAggreg	Rank aggregation	R package	`http://cran.r-project.org/web/packages/RankAggreg/`
RankProd	Rank aggregation	R package	`http://www.bioconductor.org/packages/release/bioc/html/RankProd.html`

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 τ_1 and τ_2 (Fagin *et al.* 2003; Kendall 1938), described as:

$$K^{(p)}(\tau_1,\tau_2) = \sum_{i,i' \in P(\tau_1,\tau_2)} \bar{K}^{(p)}_{ii'}(\tau_1,\tau_2) \text{ where } i \neq i' \qquad (1)$$

$P(\tau_1,\tau_2)$ is the set of all paired items i and i' and $\bar{K}^{(p)}_{ii'}(\tau_1,\tau_2)$ 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 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 π of the Markov chain. The stationary distribution π is the principal left eigenvector of the transition matrix M associated with an eigenvalue of 1. π reflects a natural order for G, where the highest value in π 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_i 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):

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

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

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:

$$r_i^* = \frac{\sum_{j=1}^{k} r_{ij}}{k}$$

$$q_i^* = \sum_{j=1}^{k} (r_{ij} - r_i^*)^2$$

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 χ^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 Φ 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:

$$d_{ij} = \frac{\bar{x}_{A_{ij}} - \bar{x}_{B_{ij}}}{s_{ij}}$$

$$g_{ij} = \frac{\bar{x}_{A_{ij}} - \bar{x}_{B_{ij}}}{s^*_j}$$

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

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 μ_i for each gene i across all k datasets can be extracted from the following model, where g^*_{ij} is the effect size and θ_{ij} the study-specific mean of study j. ϵ_{ij} represents the within-study effect with the corresponding variance s^2_{ij}, whereas δ_{ij} is the between-study effect with the corresponding variance τ^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, τ^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 $\sqrt{\tilde{n}}$, which accounts for the sample size:

$$t_{ij} = d_{ij} \cdot \sqrt{\tilde{n}_{ij}}$$

$$\widetilde{n}_{ij} = \frac{n_{A_{ij}} \cdot n_{B_{ij}}}{n_{A_{ij}} + n_{B_{ij}}}$$

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:

$$E_j = g \times n$$

$$e_{ia} \begin{cases} -1 & \text{underexpressed} \\ 0 & \text{not differentially expressed} \\ 1 & \text{overexpressed} \end{cases}$$

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 (Campain 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 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.

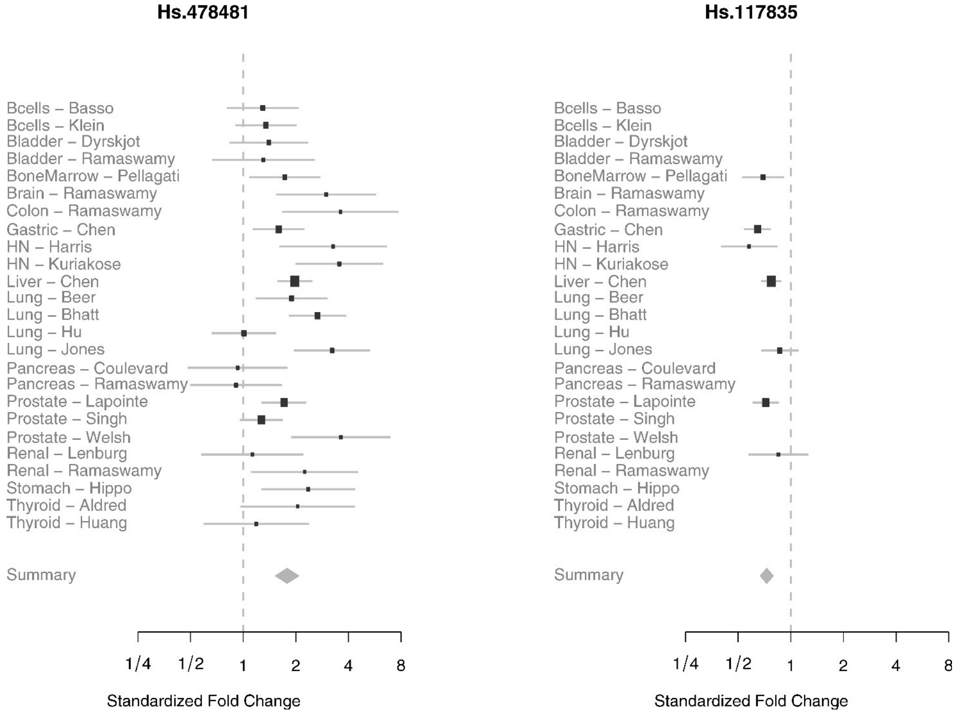


Figure 1: Example Forest plot. The Forest plots illustrate the variations in FC values for the genes *ALG3* (Hs.478481, most significantly upregulated) and *FGD4* (Hs.117835, most significantly down-regulated) between various cancer studies (Ramasamy *et al.* 2008).

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252(5013):1651–1656

Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs. BMC Bioinformatics 8:9

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769):503–511

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Anders M, Fehlker M, Wang Q, Wissmann C, Pilarsky C, Kemmner W, Hocker M (2011) Microarray meta-analysis defines global angiogenesis-related gene expression signatures in human carcinomas. Mol Carcinog (Epub ahead of print) Arasappan D, Tong W, Mummaneni P, Fang H, Amur S (2011) Meta-analysis of microarray data using a pathway-based approach identifies a 37-gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells. BMC Med 9:65

Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N (2004) Submission of microarray data to public repositories. PLoS Biol 2(9):E317

Baron D, Bihouee A, Teusan R, Dubois E, Savagner F, Steenman M, Houlgatte R, Ramstein G (2011) MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. Bioinformatics 27(5):725–726

Barrett T, Troup DB,Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res 39(Database issue):D1005–D1010

Bisognin A, Coppe A, Ferrari F, Risso D, Romualdi C, Bicciato S, Bortoluzzi S (2009) A-MADMAN: annotation-based microarray data meta-analysis tool. BMC Bioinformatics 10:201

Blanchard AP, Kaiser RJ, Hood LE (1996) High-density oligonucleotide arrays. Biosens Bioelectron 11(6/7):687–690

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185–193

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29(4):365–371

Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573(1–3):83–92

Brors B (2005) Microarray annotation and biological information on function. Methods Inf Med 44(3):468–472

Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 350(16):1605–1616

Burgess JK (2001) Gene expression studies using microarrays. Clin Exp Pharmacol Physiol 28(4):321–328

Bushman BJ (1994) Vote-counting procedures in meta-analysis. In: Cooper H, Hedges LV (eds) The handbook of research synthesis, vol 236, 1st edn. Russell Sage, New York, pp 193–213

Cahan P, Rovegno F, Mooney D, Newman JC, St Laurent G 3rd, McCaffrey TA (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. Gene 401(1–2):12–18

Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res 10(12):2022–2029

Campain A, Yang YH (2010) Comparison study of microarray meta-analysis methods. BMC Bioinformatics 11:408

Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, Huang CL, Hsu IC (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS One 6(7):e22859

Cheng WC, Tsai ML, Chang CW, Huang CL, Chen CR, Shu WY, Lee YS, Wang TH, Hong JH, Li CY, Hsu IC (2010) Microarray meta-analysis database (M(2)DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. BMC Bioinformatics 11:421

Cheng WC, Chang CW, Chen CR, Tsai ML, Shu WY, Li CY, Hsu IC (2011) Identification of reference genes across physiological states for qRT-PCR through microarray meta-analysis. PLoS One 6(2):e17347

Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G (1999) Making and reading microarrays. Nat Genet 21(Suppl 1):15–19

Choi JK, Yu U, Kim S, Yoo OJ (2003) Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 19(Suppl 1):i84–i90

Choi H, Shen R, Chinnaiyan AM, Ghosh D (2007) A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. BMC Bioinformatics 8:364

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale

Cui X, Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biol 4(4):210

Dan S, Tsunoda T, Kitahara O, Yanagawa R, Zembutsu H, Katagiri T, Yamazaki K, Nakamura Y, Yamori T (2002) An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. Cancer Res 62(4):1139–1147

Daves MH, Hilsenbeck SG, Lau CC, Man TK (2011) Meta-analysis of multiple microarray datasets reveals a common gene signature of metastasis in solid tumors. BMC Med Genomics 4:56

Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23(14):1846–1847

de Leon J, Susce MT, Murray-Carmichael E (2006) The AmpliChip CYP450 genotyping test: Integrating a new clinical tool. Mol Diagn Ther 10(3):135–151

DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R (2006) Combining results of microarray experiments: a rank aggregation approach. Stat Appl Genet Mol Biol 5(1):Article 15

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4(5):P3

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278(5338):680–686

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7(3):177–188

R Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/

Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412(6849):822–826

Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 31(1):219–223

Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12(1):111–139

Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99(2):147–157

Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the Web. In: Shen VY, Saito N, Lyu MR, Zurko ME (eds) The tenth international world wide web conference, Hong Kong, 1–5 May 2001, pp 613–622

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25):14863–14868

Fagin R, Kumar R, Sivakumar D (2003) Comparing top k lists. SIAM J Discr Math 17(1):134

Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS (2008) Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Res 36(Database issue):D866–D870

Fisher RA (1932) Statistical methods for research workers, 4th edn. Oliver & Boyd, Edinburgh

Furlong EE, Andersen EC, Null B, White KP, Scott MP (2001) Patterns of gene expression during Drosophila mesoderm development. Science 293(5535):1629–1633

Gagarin D, Yang Z, Butler J, Wimmer M, Du B, Cahan P, McCaffrey TA (2005) Genomic profiling of acquired resistance to apoptosis in cells derived from human atherosclerotic lesions: potential role of STATs, cyclinD1, BAD, and Bcl-XL. J Mol Cell Cardiol 39(3):453–465

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11(12):4241–4257

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004a) Affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20(3):307–315

Gautier L, Moller M, Friis-Hansen L, Knudsen S (2004b) Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics 5:111

Gentleman R (2005) Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York

Gershon D (2002) Microarray technology: an array of opportunities. Nature 416(6883):885–891

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Griffith OL, Melck A, Jones SJ, Wiseman SM (2006) Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. J Clin Oncol 24(31):5043–5051

Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, Kloppel G, Saeger HD, Pilarsky C (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. Oncogene 24(32):5079–5088

Guo L, Lobenhofer EK, Wang C, Shippy R, Harris SC, Zhang L, Mei N, Chen T, Herman D, Goodsaid FM, Hurban P, Phillips KL, Xu J, Deng X, Sun YA, Tong W, Dragan YP, Shi L (2006) Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nat Biotechnol 24(9):1162–1169

Harbig J, Sprinkle R, Enkemann SA (2005) A sequencebased identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. Nucleic Acids Res 33(3):e31

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic, New York

Hedges LV, Cooper H, Bushman BJ (1992) Testing the null hypothesis in meta-analysis: a comparison of combined probability and confidence interval procedures. Psychol Bull 111(1):188–194

Hong F, Breitling R (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics 24(3):374–382

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22(22):2825–2827

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 2008:420747

Hu P, Greenwood CMT, Beyene J (2006) Statistical methods for meta-analysis of microarray data: a comparative study. Inform Syst Front 8(1):9–20

Hubbell E, Liu WM, Mei R (2002) Robust estimators for expression analysis. Bioinformatics 18(12):1585–1592

Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TB, Wymore F, Zachariah ZK, Sherlock G, Ball CA (2009) Implementation of GenePattern within the Stanford microarray database. Nucleic Acids Res 37(Database issue):D898–D901

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(Suppl 1):S96–S104

Ikeo K, Ishi-i J, Tamura T, Gojobori T, Tateno Y (2003) CIBEX: center for information biology gene expression database. C R Biol 326(10–11):1079–1082

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4(2):249–264

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W (2005) Multiplelaboratory comparison of microarray platforms. Nat Methods 2(5):345–350

Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR (2002) A stem cell molecular signature. Science 298(5593):601–604

Ivliev AE, t Hoen PA, Villerius MP, den Dunnen JT, Brandt BW (2008) Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. Nucleic Acids Res 36(Web Server issue):W327–W331

Jafari P, Azuaje F (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak 6:27

Jelier R, t Hoen PA, Sterrenburg E, den Dunnen JT, van Ommen GJ, Kors JA, Mons B (2008) Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. BMC Bioinformatics 9:291

Kauffmann A, Rayner TF, Parkinson H, Kapushesky M, Lukk M, Brazma A, Huber W (2009) Importing ArrayExpress datasets into R/bioconductor. Bioinformatics 25(16):2092–2094

Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1–2):81–93

Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce- Jacino MT, Fodor SP, Jones KW (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21(10):1233–1237

Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, Sara H, Pisto T, Saarela M, Skotheim RI, Bjorkman M, Mpindi JP, Haapa-Paananen S, Vainio P, Edgren H, Wolf M, Astola J, Nees M, Hautaniemi

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

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639–1645

LaCroix-Fralish ML, Austin JS, Zheng FY, Levitin DJ, Mogil JS (2011) Patterns of pain: meta-analysis of microarray studies of pain. Pain 152(8):1888–1898

Lalkhen AG, McCluskey A (2008) Statistics V: Introduction to clinical trials and systematic reviews. Continuing Education in Anaesthesia Critical Care Pain 8(4):143–146

Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J (2005) Independence and reproducibility across microarray platforms. Nat Methods 2(5):337–344

Larsson O, Sandberg R (2006) Lack of correct data format and comparability limits future integrative microarray research. Nat Biotechnol 24(11):1322–1323

Larsson O, Wennmalm K, Sandberg R (2006) Comparative microarray analysis. OMICS 10(3):381–397

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14(6):1085–1094

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res 33(Database issue):D71–D74

Lewis S, Clarke M (2001) Forest plots: trying to see the wood and the trees. BMJ 322(7300):1479–1480

Li C, Hung Wong W (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2(8):RESEARCH0032

Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 98(1):31–36

Li X, Quigg RJ, Zhou J, Gu W, Nagesh Rao P, Reed EF (2008) Clinical utility of microarrays: current status, existing challenges and future outlook. Curr Genomics 9(7):466–474

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet 21(Suppl 1):20–24

Lo K, Gottardo R (2007) Flexible empirical Bayes models for differential gene expression. Bioinformatics 23(3):328–335

Loughin T (2004) A systematic comparison of methods for combining p-values from independent tests. Comput Stat Data Anal 47(3):467–485

Lu T, Pan Y, Kao SY, Li C, Kohane I, Chan J, Yankner BA (2004) Gene regulation and DNA damage in the ageing human brain. Nature 429(6994):883–891

Lusa L, Gentleman RC, Ruschhaupt M(2006) GeneMeta: meta-analysis for high throughput experiments. `http://www.bioconductor.org/packages/2.8/bioc/html/Gene-Meta.html`

Marot G, Foulley JL, Mayer CD, Jaffrezic F (2009) Moderated effect size and P-value combinations for microarray meta-analyses. Bioinformatics 25(20):2692–2699

Marshall E (2004) Getting the noise out of gene arrays. Science 306(5696):630–631

McDonald MJ, Rosbash M (2001) Microarray analysis and organization of circadian gene expression in Drosophila. Cell 107(5):567–578

Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365(9458):488–492

Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet 19(10):570–577

Nadon R, Shoemaker J (2002) Statistical issues with microarrays: processing and analysis. Trends Genet 18(5):265–271

Newman JC, Weiner AM (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol 6(9):R81

Normand SL (1999) Meta-analysis: formulating, evaluating, combining, and reporting. StatMed 18(3):321–359

Ntzani EE, Ioannidis JP (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 362(9394):1439–1444

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12:385

Owzar K, Barry WT, Jung SH, Sohn I, George SL (2008) Statistical challenges in preprocessing in microarray experiments in cancer. Clin Cancer Res 14(19):5959–5966

Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, Frey BJ, Blencowe BJ (2004) Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell 16(6):929–941

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A (2009) ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37(Database issue):D868–D872

Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E (2002) A statistical framework for expression-based molecular classification in cancer. J R Stat Soc Ser B Stat Methodol 64(4):717–736

Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clin Cancer Res 10(9):2922–2927

Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, Fang H, Kawasaki ES, Hager J, Tikhonova IR, Walker SJ, Zhang L, Hurban P, de Longueville F, Fuscoe JC, Tong W, Shi L, Wolfinger RD (2006) Performance comparison of one-color and two-color platforms within the MicroArray quality control (MAQC) project. Nat Biotechnol 24(9):1140–1150

Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 96(16):9212–9217

Pihur V, Datta S, Datta S (2008) Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach. Genomics 92(6):400–403

Pihur V, Datta S, Datta S (2009) RankAggreg, an R package for weighted rank aggregation. BMC Bioinformatics 10:62

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23(1):41–46

Pruitt KD, Tatusova T, KlimkeW, MaglottDR (2009) NCBI reference sequences: current status, policy and new initiatives. Nucleic Acids Res 37(Database issue):D32–D36

Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA (2002) "Stemness": transcriptional profiling of embryonic and adult stem cells. Science 298(5593):597–600

Ramasamy A, Mondry A, Holmes CC, Altman DG (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med 5(9):e184

Ramaswamy S, Ross KN, Lander ES, Golub TR (2003) A molecular signature of metastasis in primary solid tumors. Nat Genet 33(1):49–54

Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 62(15):4427–4433

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 101(25):9309–9314

Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, Barrette TR, Anstet MJ, Kincead-Beal C, Kulkarni P, Varambally S, Ghosh D, Chinnaiyan AM (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 9(2):166–180

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J (2006) TM4 microarray software suite. Methods Enzymol 411:134–193

Sahai H, Ageel MI (2000) The analysis of variance: fixed, random, and mixed models. Birkhäuser, Boston

Schadt EE, Li C, Ellis B, Wong WH (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem Suppl 37:120–125

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470

Shalon D, Smith SJ, Brown PO (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6(7):639–645

Shen R, Ghosh D, Chinnaiyan AM (2004) Prognostic meta-signature of breast cancer developed by twostage mixture modeling of microarray data. BMC Genomics 5(1):94

Shi L, Tong W, Fang H, Scherf U, Han J, Puri RK, Frueh FW, Goodsaid FM, Guo L, Su Z, Han T, Fuscoe JC, Xu ZA, Patterson TA, Hong H, Xie Q, Perkins RG, Chen JJ, Casciano DA (2005) Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. BMC Bioinformatics 6(Suppl 2):S12

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, Fan XH, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, LeClerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqsodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, Philips KL, Pine PS, Pusztai L, Qian F, Ren H, Rosen M, Rosenzweig BA, Samaha RR, Schena M, Schroth GP, Shchegrova S, Smith DD, Staedtler F, Su Z, Sun H, Szallasi Z, Tezak Z, Thierry-Mieg D, Thompson KL, Tikhonova I, Turpaz Y, Vallanat B, Van C, Walker SJ, Wang SJ, Wang Y, Wolfinger R, Wong A, Wu J, Xiao C, Xie Q, Xu J, Yang W, Zhang L, Zhong S, Zong Y, Slikker W Jr (2006) The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24(9):1151–1161

Shimizu D, Ishikawa T, Ichikawa Y, Togo S, Hayasizaki Y, Okazaki Y, Shimada H (2004) Current progress in the prediction of chemosensitivity for breast cancer. Breast Cancer 11(1):42–48

Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, Pine PS, Boysen C, Guo X, Chudin E, Sun YA, Willey JC, Thierry-Mieg J, Thierry-Mieg D, Setterquist RA, Wilson M, Lucas AB, Novoradovskaya N, Papallo A, Turpaz Y, Baker SC, Warrington JA, Shi L, Herman D (2006) Using RNA sample titrations to assess microarray platform performance and normalization techniques. Nat Biotechnol 24(9):1123–1131

Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR, Cerrina F (1999)Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 17(10):974–978

Slodkowska EA, Ross JS (2009) MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn 9(5):417–422

Smith DD, Saetrom P, Snove O Jr, Lundberg C, Rivas GE, Glackin C, Larson GP (2008) Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. BMC Bioinformatics 9:63

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article 3

Southern E, Mir K, Shchepinov M (1999) Molecular interactions on microarrays. Nat Genet 21(Suppl 1):5–9

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, ReinholdWO, Weinstein JN, Mesirov JP, Lander ES, Golub TR (2001) Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci USA 98(19):10787–10792

Stouffer SA (1949) The American soldier, vol 2. Princeton University Press, Princeton

Stuart JM, Segal E, Koller D, Kim SK (2003) A genecoexpression network for global discovery of conserved genetic modules. Science 302(5643):249–255

Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18(1):207–208

Suarez E, Burguete A, McLachlan GJ (2009) Microarray data analysis for differential expression: a tutorial. P R Health Sci J 28(2):89–104

Suarez-Farinas M, Noggle S, Heke M, Hemmati- Brivanlou A, Magnasco MO (2005) Comparing independent microarray studies: the case of human embryonic stem cells. BMC Genomics 6:99

Teh MT, Blaydon D, Chaplin T, Foot NJ, Skoulakis S, Raghavan M, Harwood CA, Proby CM, Philpott MP, Young BD, Kelsell DP (2005) Genomewide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key somatic event. Cancer Res 65(19):8597–8603

Thurstone LL (1931) Rank order as a psycho-physical method. J Exp Psychol 14(3):187–201

Troyanskaya OG (2005) Putting microarrays in a context: integrated analysis of diverse biological data. Brief Bioinform 6(1):34–43

Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J (2001) RESOURCERER: a database for annotating and linking microarray resources within and across species. Genome Biol 2(11):SOFTWARE0002

Tukey JW (1977) Exploratory data analysis, vol 2. Addison Wesley, Boston

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98(9):5116–5121

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270(5235):484–487

Venn J (1880) On the diagrammatic and mechanical representation of propositions and reasonings. Philos Mag J Sci 9(59):1–18

Vierlinger K, Mansfeld MH, Koperek O, Nohammer C, Kaserer K, Leisch F (2011) Identification of SERPINA1 as singlemarker for papillary thyroid carcinoma through microarray meta analysis and quantification of its discriminatory power in independent validation. BMC Med Genomics 4:30

Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics 20(17):3166–3178

Wang Y, Miao ZH, Pommier Y, Kawasaki ES, Player A (2007) Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips. Bioinformatics 23(16):2088–2095

Wennmalm K, Wahlestedt C, Larsson O (2005) The expression signature of in vitro senescence resembles mouse but not human aging. Genome Biol 6(13):R109

White KP, Rifkin SA, Hurban P, Hogness DS (1999) Microarray analysis of Drosophila development during metamorphosis. Science 286(5447):2179–2184

Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol 18(5):1368–1373

Wilkinson L, Friendly M (2009) The history of the cluster heat map. Am Stats 63(2):179–184

Wren JD (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. Bioinformatics 25(13):1694–1701

Zembutsu H, Ohnishi Y, Tsunoda T, Furukawa Y, Katagiri T, Ueyama Y, Tamaoki N, Nomura T, Kitahara O, Yanagawa R, Hirata K, Nakamura Y (2002) Genomewide cDNA microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. Cancer Res 62(2):518–527

Zhou XJ, Kao MC, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat Biotechnol 23(2):238–243

Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics 24(23):2798–2800

Zintzaras E, Ioannidis JP (2008) Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. Comput Biol Chem 32(1):38–46

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 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) increases 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, 227]. 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 α, 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:

$$f_{t,g} = a_g \cdot p_t$$
$$p_t = \frac{n_t}{\sum_{t=0}^{k} n_t}$$

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 immunodiagnostic 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.

3. Generate and refine a putative human meiosis-specific gene set.

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 Aldeailej, Rebecca Anderson, Mikhlid Almutairi, Ahmed Almatrafi 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`) [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 Feichtinger[1], Ibrahim Aldeailej[1,*], Rebecca Anderson[1,*], Mikhlid Almutairi[1], Ahmed Almatrafi[1], Naif Alsiwiehri[1], Keith Griffiths[2], Nicholas Stuart[1,3,5], Jane A. Wakeman[1], Lee Larcombe[4] and Ramsay J. McFarlane[1,5]

[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 testis [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) and these are referred to as cancer/testis-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 non-allelic 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 pachytene 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 candidature

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, HORMAD1, 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, *REC8* and *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 *REC8* and *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 *REC8* and *STAG3* expression in normal tissue, the other two meiosis cohesin genes tested, *RAD21L* and *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

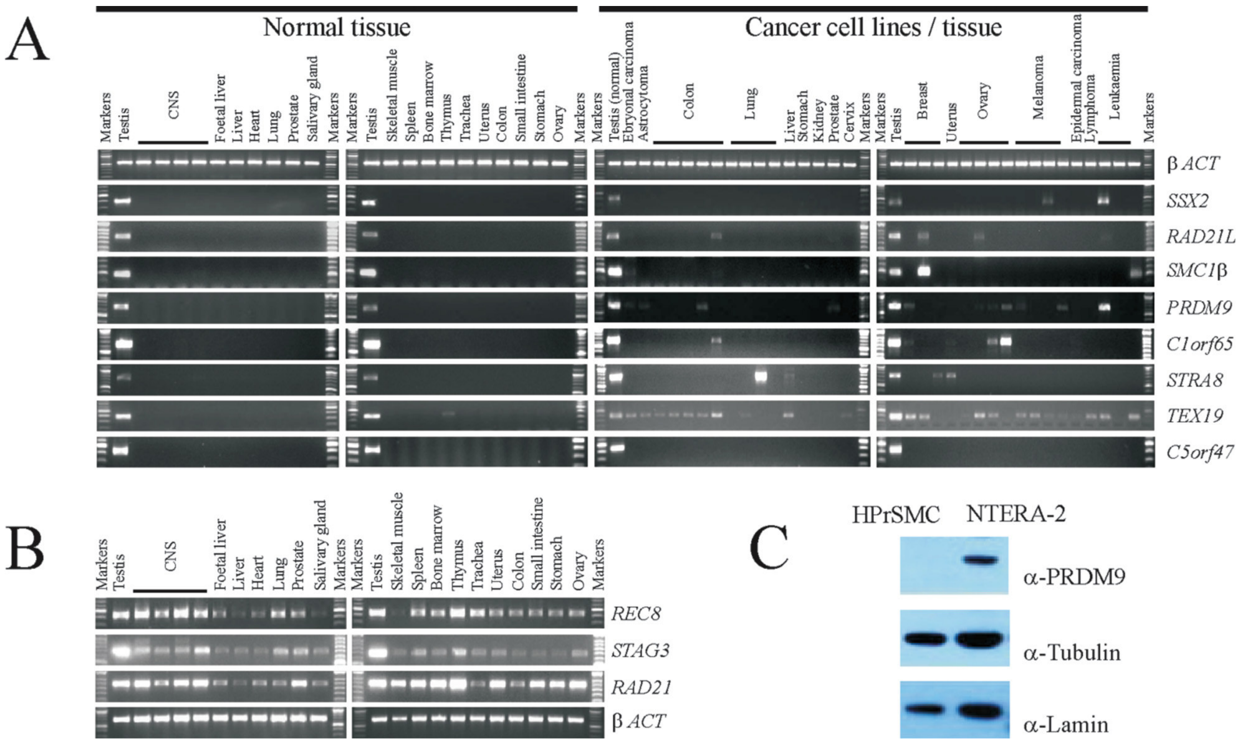


Figure 1: Examples of gene expression and protein production profiles for predicted meiCT genes. A. Agarose gels showing RT-PCR profiles generated from a range of normal human tissues obtained *post mortem* (left two panels) and cDNA generated from a range of cancer cell lines or solid tumours (right hand two panels). The expression profile for *betaACT* is a positive control (top row). The profile for *SSX2* provides an example of a previously characterised X-CT gene. Five examples of testis-restricted meiCT genes are shown (*RAD21L*/*SMC1beta*/*PRDM9*/*C1orf65*/*STRA8*) along with the expression profile of one testis selective meiCT gene, *TEX19*. The *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 *RAD21* and the two cohesin genes *REC8* and *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

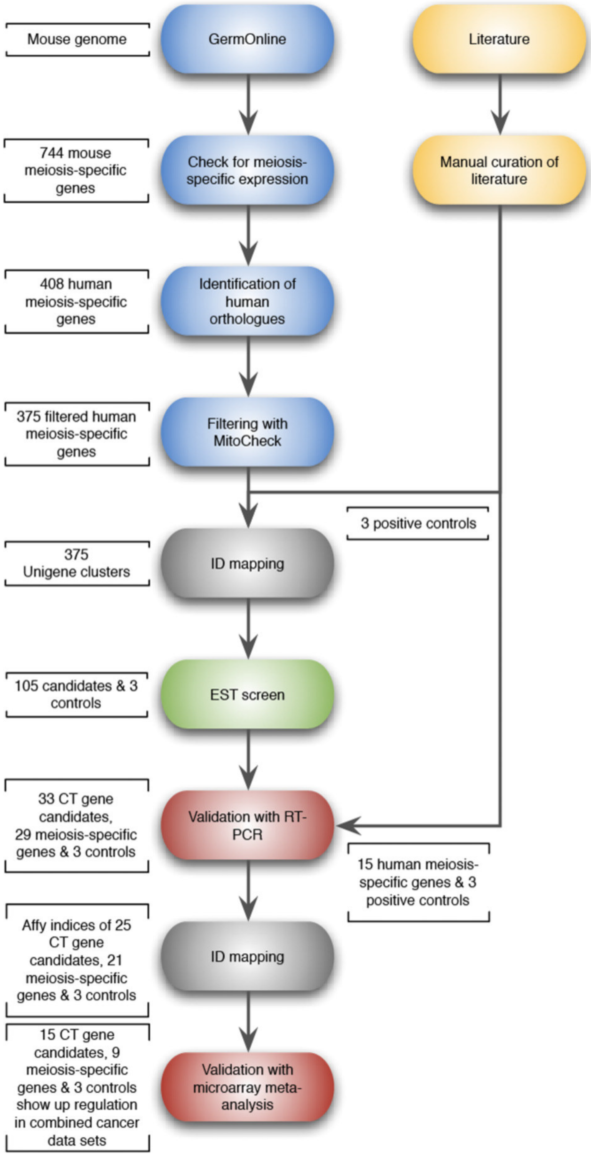

Mouse genome
GermOnline
Literature
744 mouse meiosis-specific genes
Check for meiosis-specific expression
Manual curation of literature
408 human meiosis-specific genes
Identification of human orthologues
375 filtered human meiosis-specific genes
Filtering with MitoCheck
3 positive controls
375 Unigene clusters
ID mapping
105 candidates & 3 controls
EST screen
33 CT gene candidates, 29 meiosis-specific genes & 3 controls
Validation with RT-PCR
15 human meiosis-specific genes & 3 positive controls
Affy indices of 25 CT gene candidates, 21 meiosis-specific genes & 3 controls
ID mapping
15 CT gene candidates, 9 meiosis-specific genes & 3 controls show up regulation in combined cancer data sets
Validation with microarray meta-analysis

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

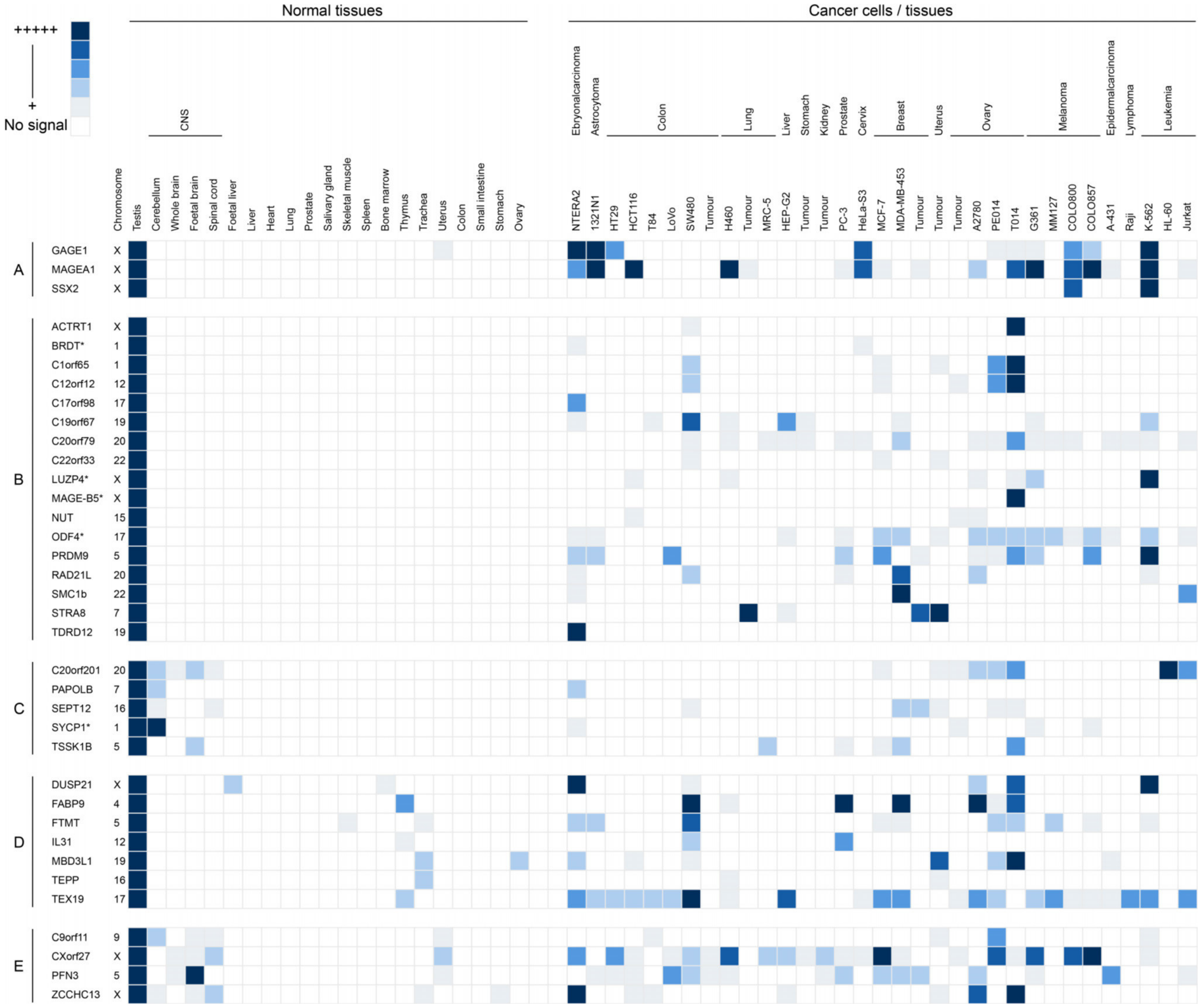


Figure 3: Grid representation of gene expression profiles for the 33 meiCT genes identified. Each gene has a lane allocation on the grid; the presence of a blue square within any column in a given lane represents the presence of an RT-PCR product indicating gene expression. The shade of blue is a qualitative representation of the RT-PCR product intensity on agarose gels. The meiCT genes have been separated into distinct classes based on those of Hofmann *et al.* [25]: A. Examples of known X-CT genes (positive controls); B. Testis-restricted meiCT genes (17 genes); C. Testis/CNS-restricted meiCT genes (5 genes); D. Testis-selective meiCT genes (7 genes); E. Testis/CNS-selective meiCT genes (4 genes). The chromosomal location of all genes is given following the gene name. Genes marked with an asterisk are genes we identified which have previously reported as CT genes [15].

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 up regulated 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.

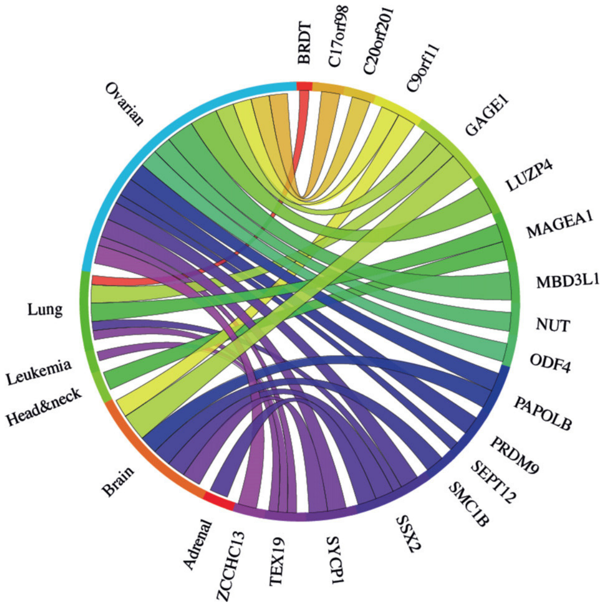


Figure 4: The Circos plot showing the meta-change in gene expression in relation to corresponding cancer types (ascribed by tissue type) for the 25 meiCT genes and the 3 known X-CT genes (*MAGE-A1, GAGE1, SSX2*) covered by array sets. 15 of the represented meiCT genes exhibit an up regulation in combined data set meta-analyses. Each connection between a gene and a cancer type indicates a statistically significant mean up regulation for that cancer type derived from a number of combined array studies for cancer tissue *vs.* normal tissue. The weight of the connection corresponds to the magnitude of the meta-change in gene expression.

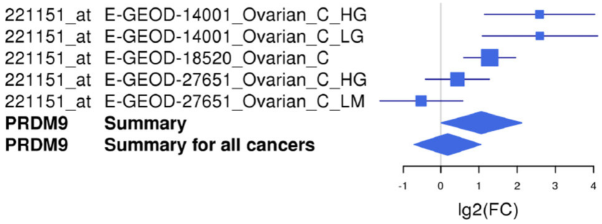


Figure 5: An example of a Forest plot for a meiCT gene, *PRDM9*. *PRDM9* is up regulated in one cancer type, ovarian cancer, according to the microarray meta-analysis. The Forest plot shows the log 2-fold change values for the individual studies as well as the total values for ovarian cancer and for 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.

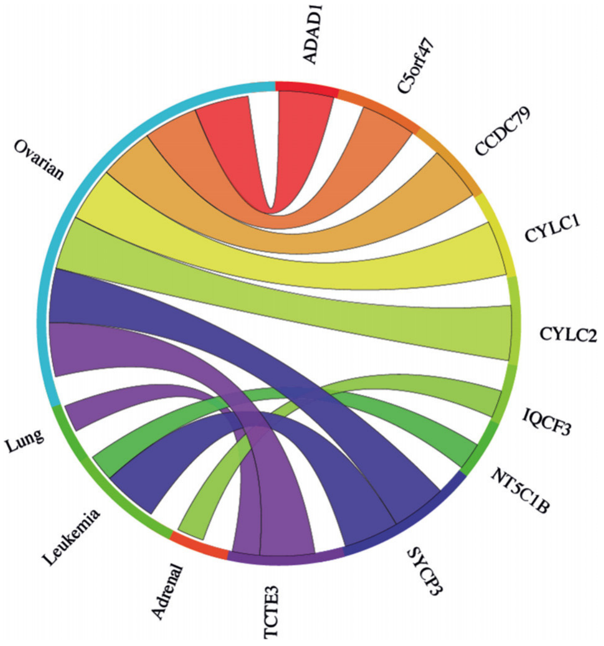


Figure 6: The Circos plot showing the meta-change in gene expression in relation to corresponding cancer types for the 21 genes which gave a testis-only expression profile following RT-PCR analysis and are represented on microarrays (Supplementary Table S2). 9 of the 21 genes show significant up regulation for combined cancer data sets. Each connection between a gene and a cancer type indicates a statistically significant mean up regulation for that cancer type derived from a number of combined array studies for cancer tissue *vs.* normal tissue. The weight of the connection corresponds to the magnitude of the meta-change in gene expression.

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 Invitrogens Dubeco's modified Eagle's medium (DMEM + GLATAMAX™) supplemented with 10% foetal bovine serum (FBS). COLO800, COLO857 and H460 cell lines were cultured in Invitrogens 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% CO_2, with the exception of the NTERA-2 (clone D1) cell line which was grown at 37°C with 10% CO_2.

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 precycling 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 pH7.4, 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 visulisation (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

1. Jensen-Jarolim E, Singer J. Cancer vaccines inducing antibody production: more pros than cons. Expert Rev Vaccines. 2011; 10: 1281-1289.

2. Klebanoff CA, Acquavella N, Yu Z, Restifo NP. Therapeutic cancer vaccines: are we there yet? Immunol Rev. 2011; 239: 27-44.

3. Lesterhuis WJ, Haanen JB, Punt CJ. Cancer immunotherapy—revisited. Nat Rev Drug Discov. 2011; 10: 591-600.

4. Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. Nature. 2011; 480: 480-489.

5. Palucka K, Ueno H, Banchereau J. Recent developments in cancer vaccines. J Immunol. 2011; 186: 1325-1331.

6. Postow M, Callahan MK, Wolchok JD. Beyond cancer vaccines: a reason for future optimism with immunomodulation therapy. Cancer J. 2011; 17: 372-378.

7. Rosenberg SA. Cell transfer immunotherapy for metastatic solid cancer – what clinicians need to know. Nat Rev Clin Oncol. 2011; 8: 577-585.

8. Topalian SL, Weiner GJ, Pardoll DM. Cancer immunotherapy comes of age. J Clin Oncol. 2011; 29: 4828-4836.

9. Tuma RS. Enthusiasm for antibody-drug conjugates. J Natl Cancer Inst. 2011; 103: 1493-1494.

10. Weiner LM, Murray JC, Shuptrine CW. Antibody-based immunotherapy of cancer. Cell. 2012; 148: 1081-1084.

11. Straten P, Andersen MH. The anti-apoptotic members of the Bcl-2 family are attractive tumour associated antigens. Oncotarget 2010; 1: 239-245.

12. Scrimieri F, Calhoun ES, Patel K, Gupta R, Huso RH, Kern SE. FAM190A rearrangements provide a multitude of individualized tumour signatures and neo-antigens in cancer. Oncotarget 2011; 2: 69-75.

13. Ionov Y. A high throughput method for identifying personalized tumour-associated antigens. Oncotarget 2010; 1: 148-155.

14. Costa FF, Le Blanc K, Brodin B. Concise review: cancer/testis antigens, stem cells, and cancer. Stem Cells. 2007; 25: 707-711.

15. Almeida LG, Sakabe NJ, de Oliveira AR, Silva MC, Mundstein AS, Cohen T, Chen YT, Chua R, Gurung S, Gnjatic S, Jungbluth AA, Caballero OL, Bairoch A, Kiesler E, White SL, Simpson AJ, et al. CT database: a knowledge-base of high throughput and curated data on cancer-testis antigens. Nucleic Acids Res. 2009; 37: D816-819.

16. Caballero OL, Chen YT. Cancer/testis (CT) antigens: potential targets for immunotherapy. Cancer Sci. 2009; 100: 2014-2021.

17. Cheng YH, Wong EW, Cheng CY. Cancer/testis (CT) antigens, carcinogenesis and spermatogenesis. Spermatogenesis. 2011; 1: 209-220.

18. Fratta E, Coral S, Covre A, Parisi G, Colizzi F, Danielli R, Marie Nicolay HJ, Sigalotti L, Maio M. The biology of cancer testis antigens: putative function, regulation and therapeutic potential. Mol Oncol. 2011; 5: 164-182.

19. Mirandola L, Cannon M, Cobos E, Bernardini G, Jenkins MR, Kast WM, Chiriva-Internati M. Cancer testis antigens: novel biomarkers and targetable proteins for ovarian cancer. Int Rev Immunol. 2011; 30: 127-137.

20. Lim SH, Zhang Y, Zhang J. Cancer-testis antigens: the current status on antigen regulation and potential clinical use. Am J Blood Res. 2012; 2: 29-35.

21. Simpson AJ, Caballero OL, Jungbluth A, Chen YT, Old LJ. Cancer/testis antigens, gametogenesis and cancer. Nat Rev Cancer. 2005; 5: 615-625.

22. Fijak M, Meinhardt A. The testis in immune privilege. Immunol Rev. 2006; 213: 66-81.

23. Mruk DD, Cheng CY. Tight junctions in the testis: new perspectives. Philos Trans R Soc Lon B Sci. 2010; 365: 1621-1635.

24. Hunder NN, Wallen H, Cao J, Hendricks DW, Reilly JZ, Rodmyre R, Jungbluth A, Gnjatic S, Thompson JA, Yee C. Treatment of metastatic melanoma with autologous CD4+ T cells against NY-ESO-1. N Eng J Med. 2008; 358: 2698-2703.

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

26. Chen YT, Scanlan MJ, Venditti CA, Chua R, Theiler G, Stevenson BJ, Iseli C, Gure AO, Vasicek T, Strausberg RL, Jongeneel CV, Old LJ, Simpson AJ. Identification of cancer/testis-antigen genes by massive parallel signature sequencing. Proc Natl Acad Sci USA. 2005; 102: 7940-7945.

27. Chomez P, De Backer O, Bertrand M, De Plaen E, Boon T, Lucas S. An overview of the MAGE gene family with the identification of all human members of the family. Cancer Res. 2001; 61: 5544-5551.

28. Grigoriadis A, Caballero OL, Hoek KS, da Silva L, Chen YT, Shin SJ, Jungbluth AA, Miller LD, Clouston D, Cebon J, Old LJ, Lakhani SR, Simpson AJ, Neville AM. CT-X antigen expression in human breast cancer. Proc Natl Acad Sci USA. 2009; 106: 13493-13498.

29. Monte M, Simonatto M, Peche LY, Bublik DR, Gobessi S, Pierotti MA, Rodolfo M, Schneider C. MAGE-A tumour antigens target p53 transactivation function through histone deacetylase recruitement and confer resistance to chemotherapeutic agents. Proc Natl Acad Sci USA. 2006; 103: 11160-11165.

30. Duan Z, Duan Y, Lamendola DE, Yusurf RZ, Naeem R, Penson RT, Seiden MV. Over expression of MAGE/GAGE genes in paclitaxel/doxorubicin-resistant human cancer cell lines. Clin Cancer Res. 2003; 9: 2778-2785.

31. Chalmel F, Lardenois A, Primig M. Toward understanding the core meiotic transcriptome in mammals and its implications for somatic cancer. Ann N Y Acad Sci. 2007; 1120: 1-15.

32. Wang J, Emadali A, Le Bescont A, Callanan M, Rousseaux S, Khochbin S. Induced malignant genome reprogramming in somatic cells by testis-specific factors. Biochim Biophys Acta. 2011; 1809: 221-225.

33. Yanowitz J. Meiosis: making a break for it. Curr Opin Cell Biol. 2010; 22: 744-751.

34. Zickler D, Kleckner N. Meiotic chromosomes: integrating structure and function. Annu Rev Genet. 1999; 33: 603-754.

35. Zhao T, Zhang ZN, Rong Z, Xu Y. Immunogenicity of induced pluripotent stem cells. Nature. 2011; 474: 212-215.

36. Turner JM. Meiotic sex chromosome inactivation. Development. 2007; 134: 1823-1831.

37. Jessberger R. Cohesin complexes get more complex: the novel kleisin RAD21L. Cell Cycle. 2011; 10: 2053-2054.

38. Uhlmann F. Cohesin subunit Rad21L, the new kid on the block has new ideas. EMBO Rep. 2011; 12: 183-184.

39. Ishiguro K, Kim J, Fujiyama-Nakamura S, Kato S, Watanabe Y. A new meiosis-specific cohesion complex implicated in the cohesion code for homologous pairing. EMBO Rep. 2011; 12: 267-275.

40. Lee J, Hirano T. RAD21L, a novel cohesion subunit implicated in linking homologous chromosomes in mammalian meiosis. J Cell Biol. 2011; 192: 263-276.

41. Noguchi T, Kato T, Wang L, Maeda Y, Ikeda H, Sato E, Knuth A, Gnjatic S, Ritter G, Sakaquchi S, Old LJ, Shiku H, Nishikawa H. Intracellular tumour-associated antigens represent effective targets for passive immunotherapy. Cancer Res. 2012; 72: 1672-1682.

42. Guo K, Tang JP, Tan CPB, Hong CW, Al-Aidaroos AQO, Varghese L, Huang C, Zeng Q. Targeting intracellular oncoproteins with antibody therapy or vaccination. Sci Transl Med. 2011; 3: ra85.

43. Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SS, Demougin P, Gattiker A, Moore J, Patard JJ, Wolgemuth DJ, Jéqou B, Primig M. The conserved transcriptome in human and rodent male gametogenesis. Proc Natl Acad Sci USA. 2007; 104: 8346-8351.

44. Odunsi K, Qian F, Matsuzaki J, Mhawech-Fauceglia P, Andrews C, Hoffmann EW, Pan L, Ritter G, Villella J, Thomas B, Rodabaugh K, Lele S, Shrikant P, Old LJ, Gnjatic S. Vaccination with an NY-ESO-1 peptide of HLA class I/II specificities induces integrated humoral and T cell responses in ovarian cancer. Proc Natl Acad Sci USA. 2007; 104: 12837-12842.

45. Janic A, Mendizabal L, Llamazares S, Rossell D, Gonzalez C. Ectopic expression of germline genes drives malignant brain tumor growth in *Drosophila*. Nature. 2010; 330: 1824-1827.

46. Hochwagen A, Marais GA. Meiosis: a PRDM9 guide to the hotspot of recombination. Curr Biol. 2010; 20: R271-R274.

47. McVean G, Myers S. PRDM9 marks the spot. Nat Genet. 2010; 42: 821-822.

48. Hayashi K, Yoshida K, Matsui Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature. 2005; 438: 374-378.

49. Harigaya Y, Tanaka H, Yamanaka S, Tanaka K, Watanabe Y, Tsutsumi C, Chikashige Y, Hiraoka Y, Yamashita A, Tamamoto M. Selective elimination of messenger RNA prevents an incidence of untimely meiosis. Nature. 2006; 442: 45-50.

50. Hiriart E, Vavasseur A, Touat-Todeschini L, Yamashita A, Gilquin B, Lambert E, Perot J, Shichino Y, Nazaret N, Boyault C, Lachuer J, Perazza D, Yamamoto M, Verdel A. MmiRNA surveillance machinery directs RNAi complex RITS to specific meiotic genes in fission yeast. EMBO J. 2012; 31: 2296-2308.

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

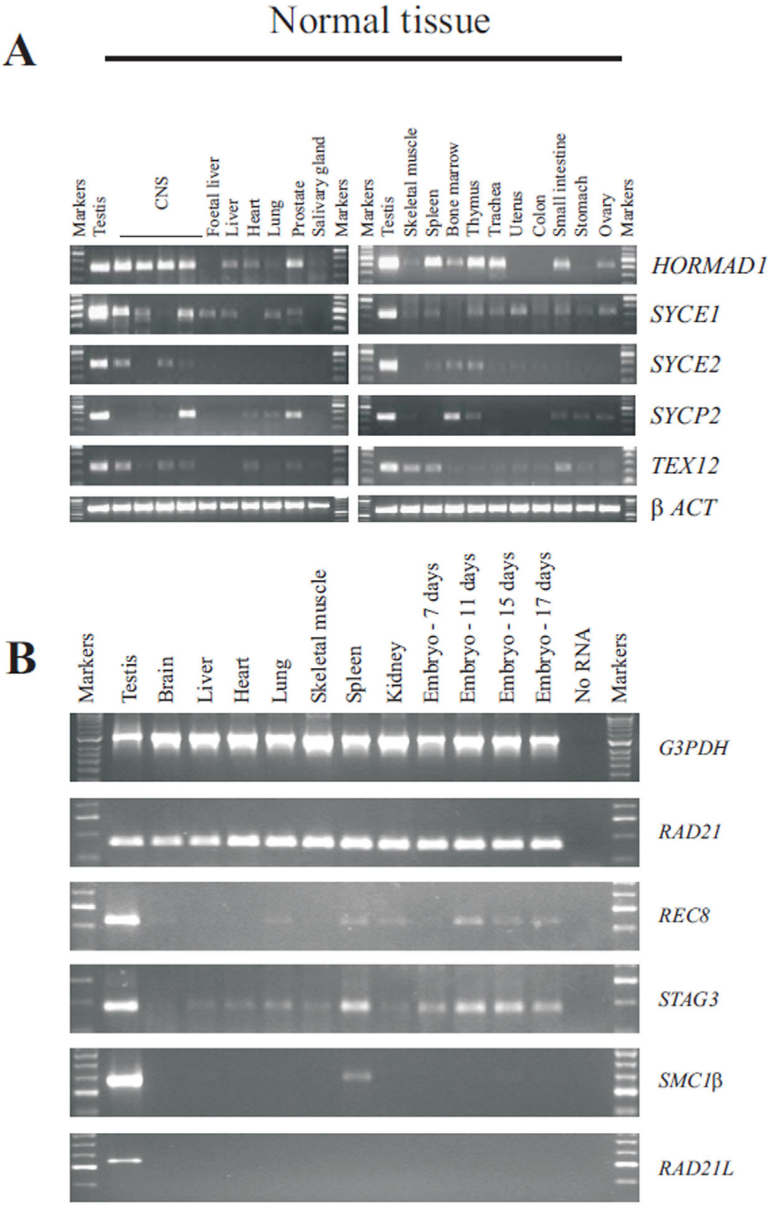


Figure S1: RT-PCR analysis of selected human and mouse meiosis-associated genes. A. Expression profiles for five human genes originally predicted to be testis-specific. The image shows agarose gels of RT-PCR products for the five genes (*HORMAD1, SYCE1, SYCE2, SYCP2, TEX12*) obtained from cDNA derived from normal human tissue RNA (obtained *post mortem*). *β*ACT 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 cohesin 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.

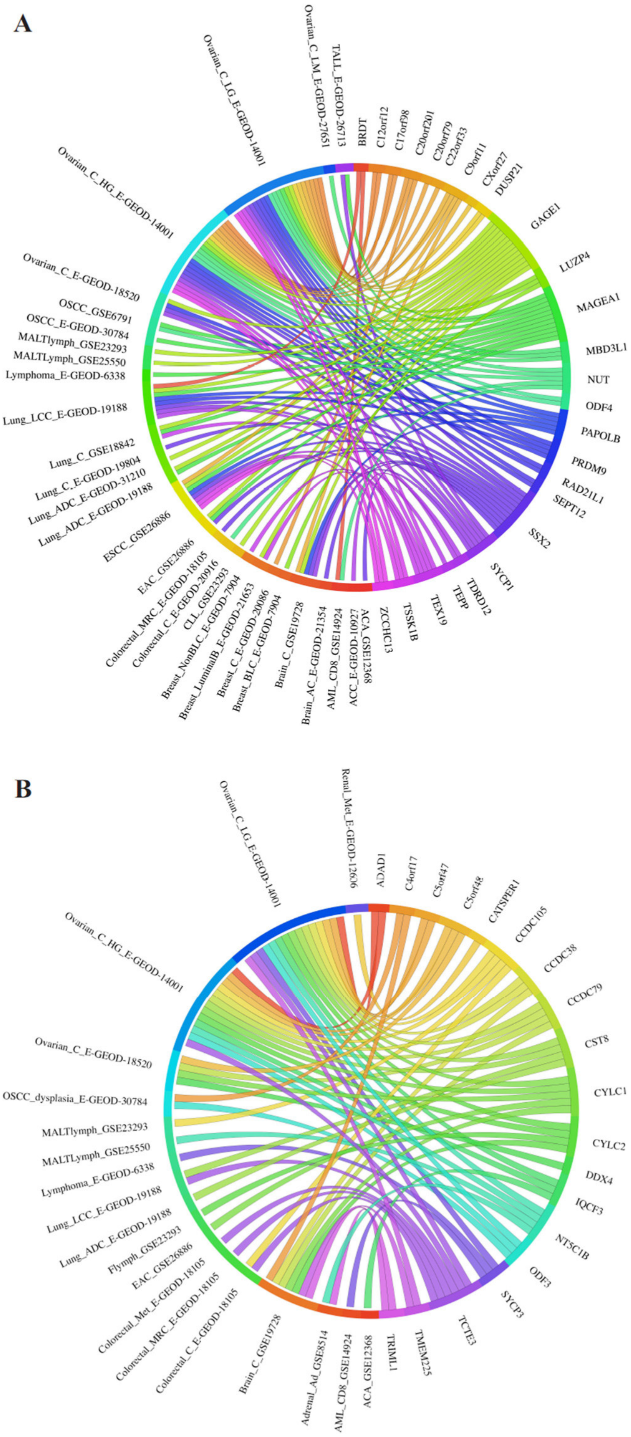


Figure S2: Circus plots for single microarray analyse. A. The Circos plot showing single microarray analysis in relation to corresponding cancer types for the 25 meiCT genes and the 3 known X-CT genes (*MAGE-A1, GAGE1, SSX2*) covered by array sets (Supporting Information Table 4). Each connection between 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.

Gene name	Functional role	Reference	Classification following validation
HORMAD1	Recombination partner choice regulation	Wojtasz *et al.* (2009) *PLoS Genetics* 5: e1000702 Shin *et al.* (2010) *PLoS Genetics* 6: e1001190	Dismissed
NUT (C15orf55)	Unknown	French (2012) *Annu Rev Patho l* 7: 247-265	Restricted CT gene
PRDM9	Meiotic hotspot regulation	Hochwagen & Marais (2010) *Curr Biol* 20: R271-274 Neale (2010) *Genome Biol* 11: 104	Restricted CT gene
RAD21L	Meiotic cohesin subunit	Lee & Hirano (2011) *J Cell Biol* 192:263-276 Ishiguro *et al.* (2011) *EMBO Rep* 12: 267-275	Restricted CT gene
REC8	Meiotic cohesin subunit	Bardhan (2010) Chromosome Res 18: 909-924	Dismissed
SMC1β	Meiotic cohesin subunit	Bardhan (2010) *Chromosome Res* 18: 909-924	Restricted CT gene
STAG3	Meiotic cohesin subunit	Bardhan (2010) Chromosome Res 18: 909-924	Dismissed
STRA8	Retinoic acid induced meiotic regulator	Anderson *et al.* (2008) *Proc Natl Acad Sci USA* 105: 14976-14980	Restricted CT gene
SYCE1	Synaptonemal complex component	Bolcun-Filas *et al.* (2009) PLoS Genetics 5: e1000393	Dismissed
SYCE2	Synaptonemal complex component	Bolcun-Filas *et al.* (2007) *J Cell Biol* 176: 741-747	Dismissed
SYCP1	Synaptonemal complex component	Pousette *et al.* (1997) *Hum Reprod* 12: 2414-2417 Tarsounas *et al.* (1999) *J Cell Sci* 112: 423-434	Restricted CT / CNS gene
SYCP2	Synaptonemal complex component	Schalk *et al.* (1998) *Chromosoma* 107: 540-548 Yang *et al.* (2006) *J Cell Biol* 173: 497-507	Dismissed
TEX12	Meiotically up regulated	Hamer *et al.* (2006) *J Cell Sci* 119: 4025-4032	Dismissed
TEX19	Meiotically up regulated	Kuntz *et al.* (2008) *Stem Cells* 26: 734-744 Ollinger *et al.* (2008) *PLoS Genetics* 4: e1000199	Selective CT gene
TSSK1	Meiotic serine/threonine kinase	Li *et al.* (2011) *Mol Hum Reprod* 17: 42-56	Restricted CT / CNS gene

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

Gene name	Ensembl ID	Unigene cluster ID	Array coverage
ADAD1	ENSG00000164113	Hs.518957	231448_at, 240299_at
ARL13A	ENSG00000174225	Hs.147237	Not on array
ARRDC5	ENSG00000205784	Hs.574574	Not on array
C4orf17	ENSG00000138813	Hs.97501	223990_at
C4orf51	ENSG00000237136	Hs.452865	Not on array
C5orf47	ENSG00000185056	Hs.131469	1557056_at, 1557057_a_at
C5orf48	ENSG00000196900	Hs.177983	237428_at
C5orf50	ENSG00000185662	Hs.591740	Not on array
C7orf72	ENSG00000164500	Hs.99248	Not on array
CATSPER1	ENSG00000175294	Hs.189105	1552335_at
CCDC38	ENSG00000165972	Hs.210377	1553893_at
CCDC79	ENSG00000177461	Hs.376505	1557620_a_at
CCDC105	ENSG00000160994	Hs.375985	1553451_at
CST8	ENSG00000125815	Hs.121602	220627_at
CYLC1	ENSG00000183035	Hs.444230	216778_s_at, 216779_at, 216809_at
CYLC2	ENSG00000155833	Hs.3232	207780_at
DDX4	ENSG00000152670	Hs.223581	221630_s_at
EFCAB9	ENSG00000214360	Hs.716824	Not on array
GLT6D1	ENSG00000204007	Hs.522491	Not on array
IQCF3	ENSG00000229972	Hs.729443	1555235_s_at, 236871_s_at
KCNU1	ENSG00000215262	Hs.13861	237273_at,
NT5C1B	ENSG00000185013	Hs.120319	1554368_at, 222203_s_at, 243100_at
ODF3	ENSG00000177947	Hs.350949	1553051_s_at, 233795_at
SYCP3	ENSG00000139351	Hs.506504	231618_s_at,
SUNC1 (SUN3)	ENSG00000164744	Hs.406711	1553599_a_at, 241861_at
TCTE3	ENSG00000184786	Hs.733746	1554400_at, 1554401_a_at, 1557945_at, 232258_at
TMEM202	ENSG00000187806	Hs.446069	Not on array
TMEM225	ENSG00000204300	Hs.98377	244460_at
TRIML1	ENSG00000184108	Hs.348618	1557677_a_at

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.

Data set name	Cancer type	Cancer sub-type	Tissue
E-GEOD-10927_ACA	Adrenal cancer	Adenoma	Adrenal gland
E-GEOD-10927_ACC	Adrenal cancer	Carcinoma	Adrenal gland
GSE12368_ACA	Adrenal cancer	Adenoma	Adrenal gland
GSE12368_ACC	Adrenal cancer	Carcinoma	Adrenal gland
GSE8514_Adrenal_Ad	Adrenal cancer	Adenoma	Adrenal gland
E-GEOD-21354_Brain_AC	Brain cancer	Sarcoma	Brain
E-GEOD-21354_Brain_EM	Brain cancer	Sarcoma	Brain
E-GEOD-21354_Brain_OG	Brain cancer	Sarcoma	Brain
E-MEXP-2351_Brain_AC	Brain cancer	Sarcoma	Brain
GSE19728_Brain_C	Brain cancer	Sarcoma	Brain
E-GEOD-20086_Breast_C	Breast cancer	Carcinoma	Breast
E-GEOD-21653_Breast_BLC	Breast cancer	Carcinoma	Breast
E-GEOD-21653_Breast_ERBB2	Breast cancer	Carcinoma	Breast
E-GEOD-21653_Breast_LuminalA	Breast cancer	Carcinoma	Breast
E-GEOD-21653_Breast_LuminalB	Breast cancer	Carcinoma	Breast
E-GEOD-22544_Breast_C	Breast cancer	Carcinoma	Breast
E-GEOD-5764_Breast_IDC	Breast cancer	Carcinoma	Breast
E-GEOD-5764_Breast_ILC	Breast cancer	Carcinoma	Breast
E-GEOD-7904_Breast_BLC	Breast cancer	Carcinoma	Breast
E-GEOD-7904_Breast_BRCA1	Breast cancer	Carcinoma	Breast
E-GEOD-7904_Breast_NonBLC	Breast cancer	Carcinoma	Breast
E-GEOD-18105_Colorectal_C	Colorectal cancer	Carcinoma	Colon
E-GEOD-18105_Colorectal_Met	Colorectal cancer	Metastasis	Colon
E-GEOD-18105_Colorectal_MRC	Colorectal cancer	Metastasis	Colon
E-GEOD-20916_Colorectal_Ad	Colorectal cancer	Adenoma	Colon
E-GEOD-20916_Colorectal_ADC	Colorectal cancer	Carcinoma	Colon
E-GEOD-20916_Colorectal_C	Colorectal cancer	Carcinoma	Colon
E-GEOD-20916_Colorectal_Ep_Ad	Colorectal cancer	Adenoma	Colon
E-GEOD-20916_Colorectal_Ep_C	Colorectal cancer	Carcinoma	Colon
E-GEOD-20916_Colorectal_Muc_Ad	Colorectal cancer	Adenoma	Colon
E-GEOD-20916_Colorectal_Muc_C	Colorectal cancer	Carcinoma	Colon
E-GEOD-23878_Colorectal_C	Colorectal cancer	Carcinoma	Colon
E-GEOD-4183_Colorectal_C	Colorectal cancer	Carcinoma	Colon
E-GEOD-4183_Colorectal_PreAd	Colorectal cancer	Adenoma	Colon
E-GEOD-12452_NPC	Head and neck cancer	Carcinoma	Nasopharynx
E-GEOD-17351_ESCC	Head and neck cancer	Carcinoma	Esophagus
E-GEOD-30784_OSCC	Head and neck cancer	Carcinoma	Oral tissue
E-GEOD-30784_OSCC_dysplasia	Head and neck cancer	Carcinoma	Oral tissue
GSE6791_OSCC	Head and neck cancer	Carcinoma	Oral tissue
GSE26886_EAC.txt	Head and neck cancer	Carcinoma	Esophagus
GSE26886_ESCC.txt	Head and neck cancer	Carcinoma	Esophagus
E-GEOD-24739_CML	Leukemia	Hematological malignancy	Blood/bone marrow
E-GEOD-26713_TALL	Leukemia	Hematological malignancy	Blood/bone marrow
GSE14924_AML_CD4	Leukemia	Hematological malignancy	Blood/bone marrow
GSE14924_AML_CD8	Leukemia	Hematological malignancy	Blood/bone marrow
E-GEOD-19188_Lung_ADC	Lung cancer	Carcinoma	Lung
E-GEOD-19188_Lung_LCC	Lung cancer	Carcinoma	Lung
E-GEOD-19188_Lung_SCC	Lung cancer	Carcinoma	Lung
E-GEOD-19804_Lung_C	Lung cancer	Carcinoma	Lung
E-GEOD-31210_Lung_ADC	Lung cancer	Carcinoma	Lung
GSE18842_Lung_C	Lung cancer	Carcinoma	Lung
E-GEOD-35331_Flymph	Lymphoma	Hematological malignancy	Blood/bone marrow
E-GEOD-6338_Lymphoma	Lymphoma	Hematological malignancy	Lymph node
E-MEXP-2957_CLL	Lymphoma	Hematological malignancy	Blood/bone marrow
GSE23293_CLL	Lymphoma	Hematological malignancy	Blood/bone marrow
GSE23293_Flymph	Lymphoma	Hematological malignancy	Blood/bone marrow
GSE23293_MALTLymph	Lymphoma	Hematological malignancy	Blood/bone marrow
GSE25550_MALTLymph	Lymphoma	Hematological malignancy	Spleen
GSE26725_CLL	Lymphoma	Hematological malignancy	Blood/bone marrow
E-GEOD-14001_Ovarian_C_HG	Ovarian cancer	Carcinoma	Ovary
E-GEOD-14001_Ovarian_C_LG	Ovarian cancer	Carcinoma	Ovary
E-GEOD-18520_Ovarian_C	Ovarian cancer	Carcinoma	Ovary
E-GEOD-27651_Ovarian_C_HG	Ovarian cancer	Carcinoma	Ovary
E-GEOD-27651_Ovarian_C_LM	Ovarian cancer	Carcinoma	Ovary
E-GEOD-22780_Pancreatic_ADC	Pancreatic cancer	Carcinoma	Pancreas

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)

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

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.

Gene	Ensembl ID	Array Coverage	Cancer type	log2FC	CI left	CI right
ACTRT1	ENSG00000123165	Not on array				
BRDT	ENSG00000137948	206787_at	Lung cancer	0.97	0.22	1.73
C12orf12	ENSG00000197651	236968_at				
C15orf55	ENSG00000184507	1564603_at, 231338_at	Ovarian cancer	1.40	0.33	2.47
C17orf98	ENSG00000214556	244316_at	Ovarian cancer	1.14	–0.17	2.44
C19orf67	ENSG00000188032	Not on array				
C1orf65	ENSG00000178395	1552391_at				
C20orf201	ENSG00000171695	1554977_at	Ovarian cancer	1.05	0.11	1.99
C20orf79	ENSG00000132631	231134_at				
C22orf33	ENSG00000185264	231617_at				
C9orf11	ENSG00000120160	1554981_at, 1554982_a_at, 232868_at	Ovarian cancer	1.80	0.56	3.05
			Brain cancer	1.83	–0.30	3.95
CXorf27	ENSG00000187516	215142_at				
DUSP21	ENSG00000189037	220515_at				
FABP9	ENSG00000205186	Not on array				
FTMT	ENSG00000181867	Not on array				
GAGE1	ENSG00000205777	207086_x_at, 207739_s_at, 208155_x_at, 208283_at	Ovarian cancer	1.23	–0.30	2.76
			Lung cancer	1.13	0.30	1.95
			Brain cancer	2.38	0.30	4.47
IL31	ENSG00000204671	Not on array				
LUZP4	ENSG00000102021	220665_at	Ovarian cancer	2.73	1.38	4.08
MAGEA1	ENSG00000198681	207325_x_at	Head and neck cancer	1.21	0.15	2.28
			Lung cancer	1.48	0.67	2.30
MAGEB5	ENSG00000188408	Not on array				
MBD3L1	ENSG00000170948	1552458_at, 1552459_a_at	Ovarian cancer	2.56	1.45	3.66
ODF4	ENSG00000184650	1552408_at, 1552409_a_at	Ovarian cancer	1.09	–0.06	2.23
PAPOLB	ENSG00000218823	208271_at, 242158_at	Ovarian cancer	1.70	0.64	2.76
			Brain cancer	1.36	0.32	2.40
PFN3	ENSG00000196570	Not on array				
PRDM9	ENSG00000164256	221151_at	Ovarian cancer	1.06	0.01	2.12
RAD21L1	ENSG00000244588	215917_at, 234662_at				
SEPT12	ENSG00000140623	230947_at	Ovarian cancer	0.75	0.04	1.45
SMC1β	ENSG00000077935	1553249_at	Brain cancer	1.39	–0.46	3.24
SSX2	ENSG00000241476	207493_x_at, 210497_x_at, 215881_x_at, 216471_x_at	Adrenal cancer	1.09	0.08	2.10
			Ovarian cancer	1.07	–0.19	2.34
			Lung cancer	0.89	0.28	1.51
			Brain cancer	1.50	0.33	2.67
STRA8	ENSG00000146857	Not on array				
SYCP1	ENSG00000198765	206740_x_at, 216917_s_at	Ovarian cancer	1.64	0.51	2.77
			Brain cancer	1.78	0.12	3.45
TDRD12	ENSG00000173809	215356_at				
TEPP	ENSG00000159648	240119_at				
TEX19	ENSG00000182459	241367_at	Ovarian cancer	0.96	0.14	1.78
			Lung cancer	0.47	0.01	0.93
			Leukemia	0.47	0.01	0.93
TSSK1B	ENSG00000212122	211694_at				
ZCCHC13	ENSG00000187969	1554210_at	Ovarian cancer	1.41	0.40	2.42

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.

Gene name	Chromosome	Original EST class	Method of meiCT designation	CT class
ACTRT1	X	1	R, M	CT restricted
ADAD1	4	3	M	CT restricted*
BRDT	1	3	R, M	CT restricted
C1orf65	1	3	R	CT restricted
C4orf17	4	2	S	CT restricted*
C5orf47	5	2	M	CT restricted*
C5orf48	5	2	S	CT restricted*
C9orf11	9	2	R, M	CT/CNS selective
C12orf12	12	2	R, S	CT restricted
C17orf98	17	1	R, M	CT restricted
C19orf67	19	2	R	CT restricted
C20orf79	20	2	R, S	CT restricted
C20orf201	20	2	R, M	CT/CNS restricted
C22orf33	22	2	R, S	CT restricted
CXorf27	X	2	R, S	CT/CNS selective
CATSPER1	11	3	S	CT restricted*
CCDC38	12	2	S	CT restricted*
CCDC79	16	2	M	CT restricted*
CCDC105	19	2	S	CT restricted*
CST8	20	2	S	CT restricted*
CYCL1	X	2	M	CT restricted*
CYCL2	9	2	M	CT restricted*
DDX4	5	2	S	CT restricted*
DUSP21	X	2	R, S	CT selective
FABP9	4	2	R	CT selective
FTMT	5	2	R	CT selective
IL31	12	2	R	CT selective
IQCF3	3	2	M	CT restricted*
LUZP4	X	2	R, M	CT restricted
MAGE-B5	X	2	R	CT restricted
MBD3L1	19	2	R, M	CT selective
NT5C1B	2	2	M	CT restricted*
NUT	15	Manual selection	R, M	CT restricted
ODF3	11	2	S	CT restricted*
ODF4	17	1	R, M	CT restricted
PAPOLB	7	2	R, M	CT/CNS restricted
PFN3	5	2	R	CT/CNS selective
PRDM9	5	Manual selection	R, M	CT restricted
RAD21L	20	Manual selection	R, S	CT restricted
SEPT12	16	3	R, M	CT/CNS restricted
SMC1β	22	Manual selection	R, M	CT restricted
STRA8	7	Manual selection	R	CT restricted
SYCP1	1	Manual selection	R, M	CT/CNS restricted
SYCP3	12	2	M	CT restricted*
TCTE3	6	2	M	CT restricted*
TDRD12	19	2	R, S	CT restricted
TEPP	16	2	R, S	CT selective
TEX19	17	3	R, M	CT selective
TMEM225	11	2	S	CT restricted*
TRIML1	4	3	S	CT restricted*
TSSK1B	5	Manual selection	R, S	CT/CNS restricted
ZCCHC13	X	2	R, M	CT/CNS selective

* 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, spermatogonia, 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’:mid ratio of 3 and a GAPDH 3’:mid 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 identifier 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 journall *Database* (available at: `http://database.oxfordjournals.org/`) [234]. The content structure, layout, language and reference style follow the specifications of *Database*.

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2012, Article ID bas055, doi:10.1093/database/bas055

Database Tool

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

Julia Feichtinger[1,2,]*, Ramsay J. McFarlane[1,3,]* and Lee D. Larcombe[4]

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

Citation details: Julia Feichtinger, Ramsay J. McFarlane and Lee D. Larcombe. CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data. *Database* (2012) Vol. 2012: article ID bas055; doi:10.1093/database/bas055.

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.

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.

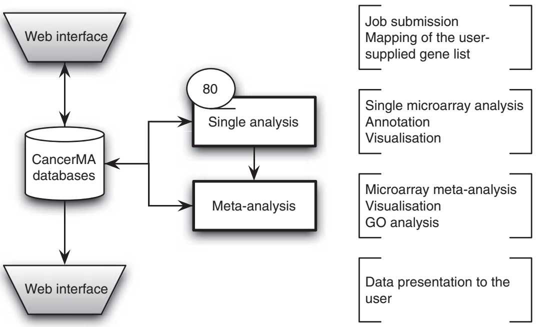


Figure 1. CancerMA workflow. The web interface box indicates the areas where the user provides input and/or can view the mapping or analysis results. The analysis is carried out automatically without any user input. The single analysis determines the differential expression for 80 cancer microarray datasets individually, whereas the meta-analysis combines the results form the individual analyses to a differential meta-expression profile.

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 (HGNC) database (16) and the annotation files provided by Affymetrix (available at: http://www.affymetrix.com/support/technical/annotationfilesmain.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 (20) and weighted linear combination (21), 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 downregulation separately (22). In case of multiple probes mapping to the same gene identifier, the most extreme log 2-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 'GOstats' 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 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). For the relational databases, MySQL 5.0.77 (available at: http://www.mysql.com) was used. The CancerMA web interface was implemented using: HTML/CSS, Twitter Bootstrapp (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 CancerMA pipeline was implemented using: R 2.12.1 (available at: http://www.cran.r-project.org) (27); the Bioconductor package (available at: http://www.bioconductor.org) (28) and Perl 5.8.8 (available at: http://www.perl.org). CancerMA is freely available online at http://www.cancerma.org.uk.

Use of CancerMA

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

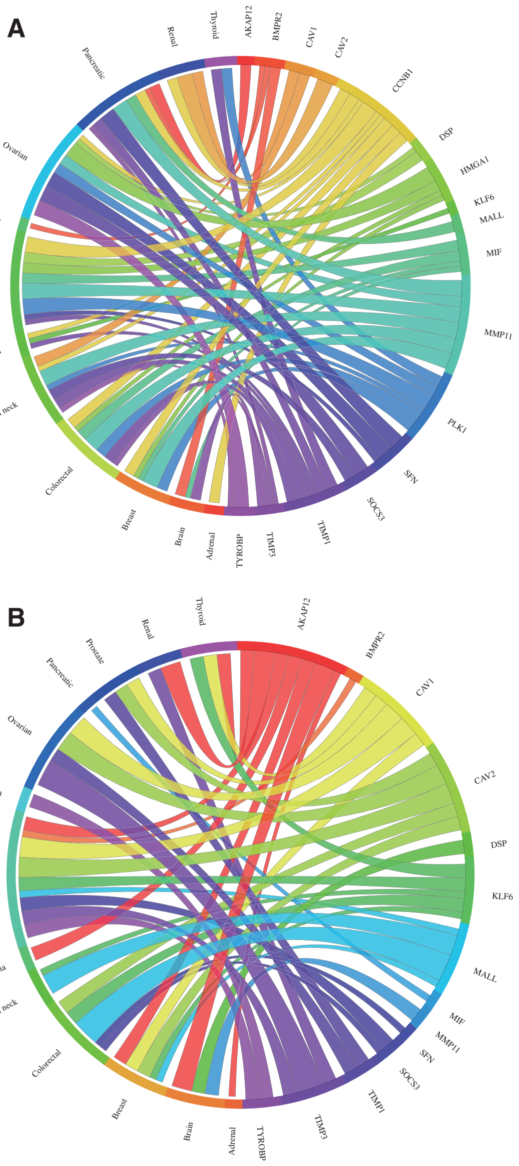


Figure 2. Circos plots showing the meta-change in gene expression in relation to corresponding cancer types. The plot shows the meta-up- and meta-downregulated genes of the validation dataset from Kettunen *et al.* (29): (**A**) The expression of the genes *DSP*, *CCNB1*, *PLK1*, *MIF*, *HMGA1*, *SFN*, *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.

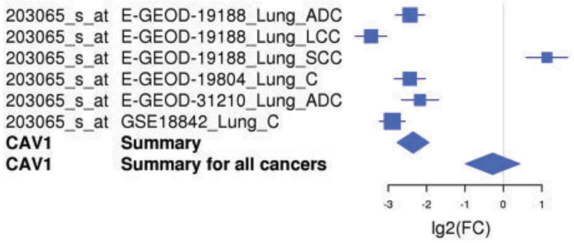


Figure 3. An example of a forest plot showing the expression of gene CAV1 downregulated in lung cancer. The expression of the *CAV1* gene is downregulated in five of six microarray studies and upregulated in one study. The forest plot shows the meta-log 2-fold change values for the individual studies 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 (lg2FC ratio), the size proportional to the weight of the study and the horizontal line through it reflecting the confidence interval of the estimate.

analysis and partially validated by RT–PCR (29) and (ii) 13 upregulated genes in ovarian cancer validated by RT–PCR (30).

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

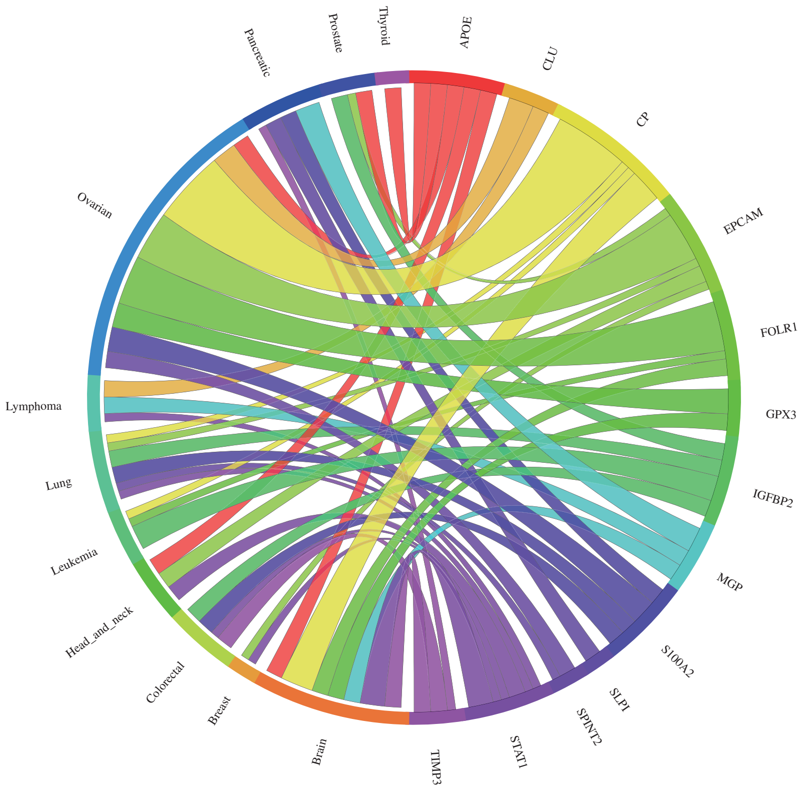


Figure 4. Circos plot showing the meta-change in gene expression in relation to corresponding cancer types. The plot shows the meta-upregulated genes of the validation dataset from Hough *et al.* (30): The expression of the genes *GPX3*, *CLU*, *EPCAM*, *SPINT2*, *FOLR1*, *S100A2*, *APOE* and *CP* was found to be upregulated in ovarian 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.

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 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 applications 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 M^{2}DB (57) and M^{3D} (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

1. Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.
2. Parkinson,H., Sarkans,U., Kolesnikov,N. *et al.* (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
3. Barrett,T., Troup,D.B., Wilhite,S.E. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
4. Moreau,Y., Aerts,S., De Moor,B. *et al.* (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.*, **19**, 570–577.
5. Feichtinger,J., Thallinger,G.G., McFarlane,R.J. *et al.* (2012) Microarray meta-analysis: From data to expression to biological relationships. In: Trajanoski,Z. (ed). *Computational Medicine*. Springer, New York, NY, pp. 59–77.
6. Ramasamy,A., Mondry,A., Holmes,C.C. *et al.* (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.*, **5**, 13.
7. Campain,A. and Yang,Y.H. (2010) Comparison study of microarray meta-analysis methods. *BMC Bioinformatics*, **11**, 408.
8. Normand,S.L. (1999) Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.*, **18**, 321–359.
9. Arasappan,D., Tong,W., Mummaneni,P. *et al.* (2011) Meta-analysis of microarray data using a pathway-based approach identifies a 37-gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells. *BMC Med.*, **9**, 65.
10. Vierlinger,K., Mansfeld,M.H., Koperek,O. *et al.* (2011) Identification of SERPINA1 as single marker for papillary thyroid carcinoma through microarray meta analysis and quantification of its discriminatory power in independent validation. *BMC Med. Genomics*, **4**, 30.
11. Feichtinger,J., Aldeailej,I., Anderson,R. *et al.* (2012) Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes. *Oncotarget*, **3**, 843–53.
12. Wilson,C.L. and Miller,C.J. (2005) Simpleaffy: A BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*, **21**, 3683–3685.
13. Hubbell,E., Liu,W.-M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
14. Gautier,L., Cope,L., Bolstad,B.M. *et al.* (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
15. Flicek,P., Amode,M.R., Barrell,D. *et al.* (2011) Ensembl 2012. *Nucleic Acids Res.*, **40**, 84–90.
16. Seal,R.L., Gordon,S.M., Lush,M.J. *et al.* (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.
17. Scholtens,D. and von Heydebreck,A. (2005) Analysis of differential gene expression studies. In: Gentleman,R., Carey,V., Huber,W. *et al.* (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 229–248.
18. Smyth,G.K. (2005) Limma: Linear models for microarray data. In: Gentleman,R., Carey,V., Huber,W. *et al.* (eds), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, NY, pp. 397–420.
19. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300.
20. Stouffer,S.A. (1949) *The American Soldier. Studies in Social Psychology in World War II*. Princeton University Press, Princeton, NJ.
21. Morgan,A.A., Khatri,P., Jones,R.H. *et al.* (2010) Comparison of multiplex meta analysis techniques for understanding the acute rejection of solid organ transplants. *BMC Bioinformatics*, **11**, S6.
22. Zaykin,D.V. (2011) Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J. Evol. Biol.*, **24**, 1836–1841.
23. Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
24. Krzywinski,M., Schein,J., Birol,I. *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
25. Lewis,S. and Clarke,M. (2001) Forest plots: trying to see the wood and the trees. *BMJ*, **322**, 1479–1480.
26. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinf.*, **12**, 385.
27. R Development Core Team,R. (2011) R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, p. 409.
28. Gentleman,R.C., Carey,V.J., Bates,D.M. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
29. Kettunen,E., Anttila,S., Seppänen,J.K. *et al.* (2004) Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genet. Cytogenet.*, **149**, 98–106.
30. Hough,C.D., Cho,K.R., Zonderman,A.B. *et al.* (2001) Coordinately up-regulated genes in ovarian cancer. *Cancer Res.*, **61**, 3869–3876.
31. Boelens,M.C., Van Den Berg,A., Vogelzang,I. *et al.* (2007) Differential expression and distribution of epithelial adhesion molecules in non-small cell lung cancer and normal bronchus. *J. Clin. Pathol.*, **60**, 608–614.
32. Soria,J.C., Jang,S.J., Khuri,F.R. *et al.* (2000) Overexpression of cyclin B1 in early-stage non-small cell lung cancer and its clinical implication. *Cancer Res.*, **60**, 4000–4004.
33. Wang,Z.-X., Xue,D., Liu,Z.-L. *et al.* (2012) Overexpression of polo-like kinase 1 and its clinical significance in human non-small cell lung cancer. *Int. J. Biochem. Cell Biol.*, **44**, 200–210.
34. Campa,M.J., Wang,M.Z., Howard,B. *et al.* (2003) Protein expression profiling identifies macrophage migration inhibitory factor and cyclophilin a as potential molecular targets in non-small cell lung cancer. *Cancer Res.*, **63**, 1652–1656.
35. Hillion,J., Wood,L.J., Mukherjee,M. *et al.* (2009) Upregulation of MMP-2 by HMGA1 promotes transformation in undifferentiated, large-cell lung cancer. *Mol. Cancer Res.*, **7**, 1803–1812.
36. Aljada,I.S., Ramnath,N., Donohue,K. *et al.* (2004) Upregulation of the tissue inhibitor of metalloproteinase-1 protein is associated with progression of human non-small-cell lung cancer. *J. Clin. Oncol.*, **22**, 3218–3229.

37. Coussens,L.M., Fingleton,B. and Matrisian,L.M. (2002) Matrix metalloproteinase inhibitors and cancer: trials and tribulations. *Science*, **295**, 2387–2392.

38. Jo,U., Whang,Y.M., Kim,H.K. *et al*. (2006) AKAP12α is associated with promoter methylation in lung cancer. *Cancer Res. Treat.*, **38**, 144–151.

39. Ito,G., Uchiyama,M., Kondo,M. *et al*. (2004) Krüppel-like factor 6 is frequently down-regulated and induces apoptosis in non-small cell lung cancer cells. *Cancer Res.*, **64**, 3838–3843.

40. He,B., You,L., Uematsu,K. *et al*. (2003) SOCS-3 is frequently silenced by hypermethylation and suppresses cell growth in human lung cancer. *Proc. Natl Acad. Sci. USA*, **100**, 14133–14138.

41. Garofalo,M., Di Leva,G., Romano,G. *et al*. (2009) miR-221&222 regulate TRAIL resistance and enhance tumorigenicity through PTEN and TIMP3 downregulation. *Cancer Cell*, **16**, 498–509.

42. Wikman,H., Seppänen,J.K., Sarhadi,V.K. *et al*. (2004) Caveolins as tumour markers in lung cancer detected by combined use of cDNA and tissue microarrays. *J. Pathol.*, **203**, 584–593.

43. Cahan,P., Rovegno,F., Mooney,D. *et al*. (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, **401**, 12–18.

44. Lee,H.J., Do,J.H., Bae,S. *et al*. (2010) Immunohistochemical evidence for the over-expression of Glutathione peroxidase 3 in clear cell type ovarian adenocarcinoma. *Med. Oncol.*, **28**, S522–S527.

45. Xie,D., Lau,S.H., Sham,J.S.T. *et al*. (2005) Up-regulated expression of cytoplasmic clusterin in human ovarian carcinoma. *Cancer*, **103**, 277–283.

46. Kim,J.-H., Herlyn,D., Wong,K. *et al*. (2003) Identification of epithelial cell adhesion molecule autoantibody in patients with ovarian cancer. *Clin. Cancer Res.*, **9**, 4782–4791.

47. Foulkes,W.D., Campbell,I.G., Stamp,G.W. *et al*. (1993) Loss of heterozygosity and amplification on chromosome 11q in human ovarian cancer. *Br. J. Cancer*, **67**, 268–273.

48. Santin,A.D., Zhan,F., Bellone,S. *et al*. (2004) Gene expression profiles in primary ovarian serous papillary tumors and normal ovarian epithelium: identification of candidate molecular markers for ovarian cancer diagnosis and therapy. *Int. J. Cancer*, **112**, 14–25.

49. Chen,Y.-C., Pohl,G., Wang,T.-L. *et al*. (2005) Apolipoprotein E is required for cell proliferation and survival in ovarian cancer. *Cancer Res.*, **65**, 11424–11431.

50. Hough,C.D., Sherman-Baust,C.A., Pizer,E.S. *et al*. (2000) Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. *Cancer Res.*, **60**, 6281–6287.

51. Lee,C.M., Lo,H.-W., Shao,R.-P. *et al*. (2004) Selective activation of ceruloplasmin promoter in ovarian tumors: potential use for gene therapy. *Cancer Res.*, **64**, 1788–1793.

52. Baeuerle,P.A. and Gires,O. (2007) EpCAM (CD326) finding its role in cancer. *Br. J. Cancer*, **96**, 417–423.

53. Fisher,R.E., Siegel,B.A., Edell,S.L. *et al*. (2008) Exploratory study of 99mTc-EC20 imaging for identifying patients with folate receptor-positive solid tumors. *J. Nucl. Med.*, **49**, 899–906.

54. Benetkiewicz,M., Wang,Y., Schaner,M. *et al*. (2005) High-resolution gene copy number and expression profiling of human chromosome 22 in ovarian carcinomas. *Genes, Chromosomes Cancer*, **42**, , 228–237.

55. Meinhold-Heerlein,I., Bauerschlag,D., Hilpert,F. *et al*. (2005) Molecular and prognostic distinction between serous ovarian carcinomas of varying grade and malignant potential. *Oncogene*, **24**, 1053–1065.

56. Vogelstein,B., Lane,D. and Levine,A.J. (2000) Surfing the p53 network. *Nature*, **408**, 307–310.

57. Cheng,W.-C., Tsai,M.-L., Chang,C.-W. *et al*. (2010) Microarray meta-analysis database (M2DB): a uniformly pre-processed, quality controlled, and manually curated human clinical microarray database. *BMC Bioinformatics*, **11**, 421.

58. Faith,J.J., Driscoll,M.E., Fusaro,V.A. *et al*. (2008) Many microbe microarrays database: Uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

59. Rhodes,D.R., Kalyana-Sundaram,S., Mahavisno,V. *et al*. (2007) Oncomine 3.0: Genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.

60. Zoubarev,A., Hamer,K.M., Keshav,K.D. *et al*. (2012) Gemma: A resource for the re-use, sharing and meta-analysis of expression profiling data. *Bioinformatics*, **28**, 2272–2273.

61. Kilpinen,S., Autio,R., Ojala,K. *et al*. (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**, R139.

62. Hruz,T., Laule,O., Szabo,G. *et al*. (2008) Genevestigator V3: A reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008**, 420747.

63. Marot,G., Foulley,J.-L., Mayer,C.-D. *et al*. (2009) Moderated effect size and P-value combinations for microarray meta-analyses. *Bioinformatics*, **25**, 2692–2699.

64. Hong,F., Breitling,R., McEntee,C.W. *et al*. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825–2827.

65. Choi,H., Shen,R., Chinnaiyan,A.M. *et al*. (2007) A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, **8**, 364.

Supplemental Material

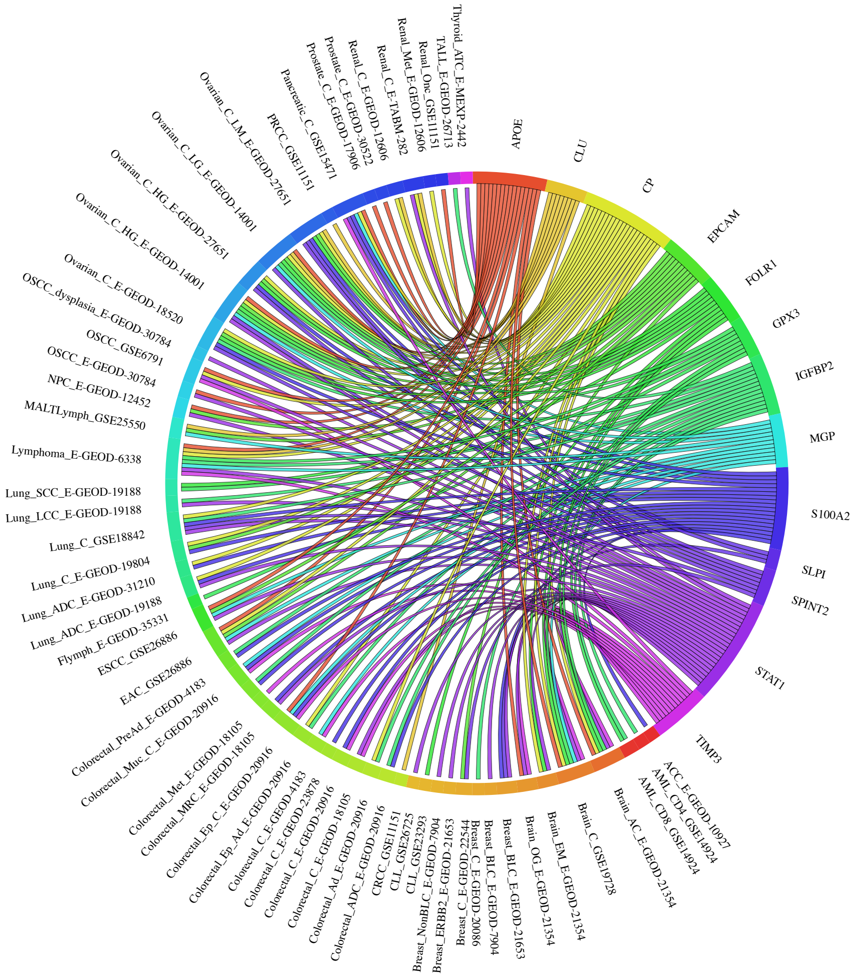


Figure S1: Circos plot showing single array gene expression analysis in relation to corresponding cancer types. The plot shows the upregulated genes of the validation dataset from Hough *et al.* (30). The expression of the genes *GPX3*, *CLU*, *EPCAM*, *SPINT2*, *FOLR1*, *S100A2*, *APOE* and *CP* is 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.

Cancer type	**Microarray experiments**	**Microarray datasets**
Adrenal cancer	3	5
Brain cancer	3	5
Breast cancer	5	11
Colorectal cancer	4	13
Head and neck cancer	5	7
Leukemia	3	4
Lung cancer	4	6
Lymphoma	6	8
Ovarian cancer	3	5
Pancreatic cancer	2	2
Prostate cancer	2	2
Renal cancer	3	7
Thyroid cancer	2	5
Total	45	80

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]

[1]North West Cancer Research Fund Institute, Bangor University, Bangor, Gwynedd LL57 2UW, UK
[2]Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
[3]NISCHR Cancer Genetics Biomedical Research Unit, Bangor University, Bangor, Gwynedd LL57 2UW, UK
[4]Cranfield 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.macfarlane@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 focussing 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.

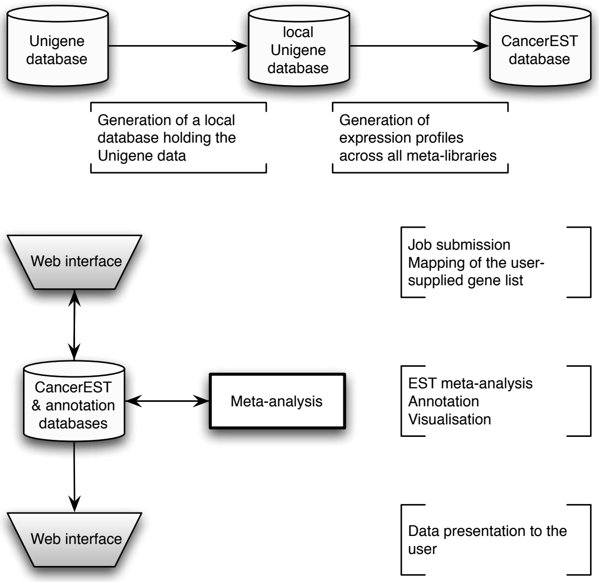

Unigene database
local Unigene database
CancerEST database
Generation of a local database holding the Unigene data
Generation of expression profiles across all meta-libraries
Web interface
Job submission Mapping of the user-supplied gene list
CancerEST & annotation databases
Meta-analysis
EST meta-analysis Annotation Visualisation
Web interface
Data presentation to the user

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 web site (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

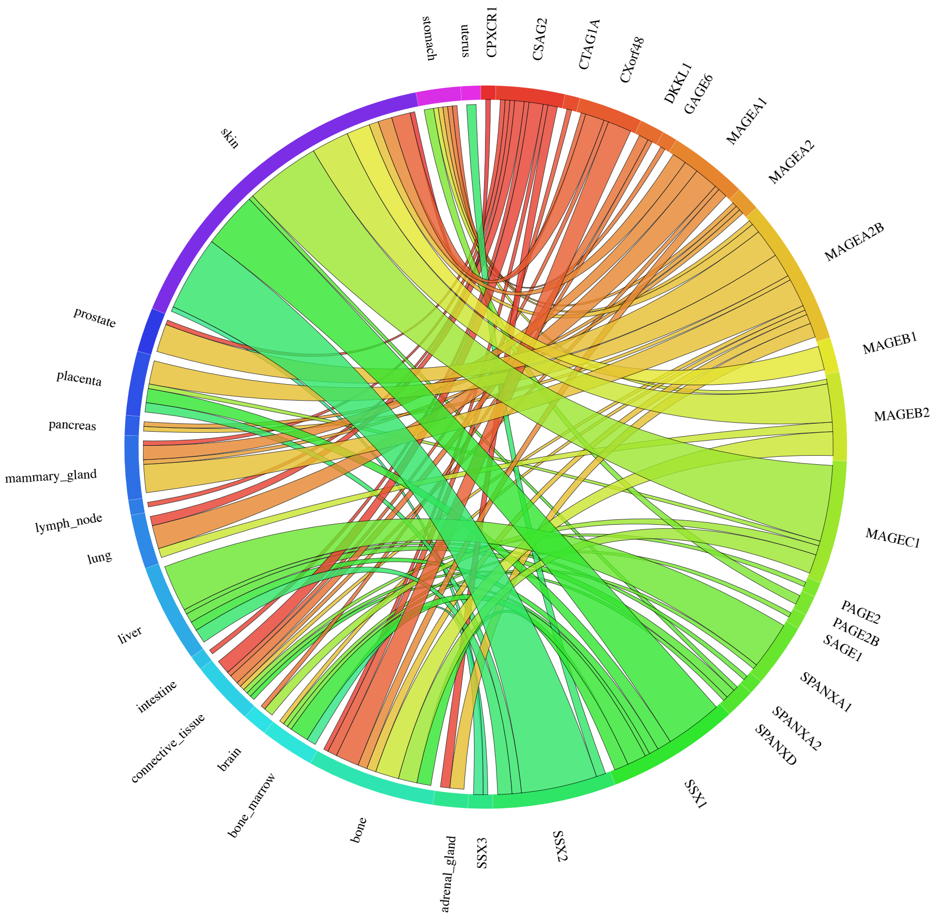


Figure 2: Circos plot showing the gene expression in relation to the corresponding cancer types for the 39 testis-restricted genes determined by Hofmann *et al.* (9). 21 of the 39 testis-restricted genes exhibit expression in various cancer types, in particular in melanoma. Each connection between a 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.

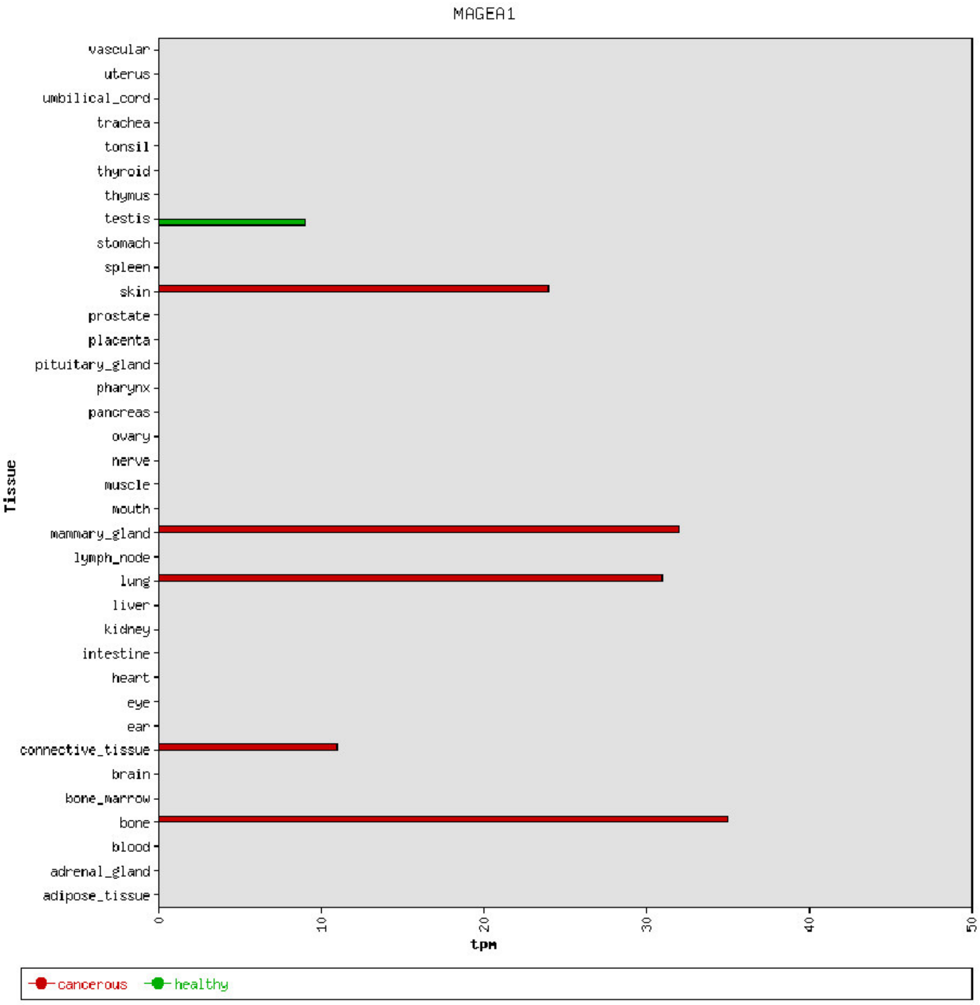


Figure 3: An example of a bar chart showing the integrated expression profile of the *MAGEA1* gene. *MAGEA1* exhibits a testis-restricted gene expression profile, but is aberrantly expressed in a number of cancer types. The expression is given in transcripts per million (*tpm*).

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

1. Ludwig,J.A. and Weinstein,J.N. (2005) Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer*, 5, 845–856.

2. Backus,J., Laughlin,T., Wang,Y., *et al.* (2005) Identification and characterization of optimal gene expression markers for detection of breast cancer metastasis. *J. Mol. Diagn.*, 7, 327–336.

3. Adams,M.D., Kelley,J.M., Gocayne,J.D., *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252, 1651–1656.

4. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for "expressed sequence tags". *Nat. Genet.*, 4, 332–3.

5. Pontius,J.U., Wagner,L. and Schuler,G.D. (2003) UniGene: a unified view of the transcriptome. In McEntyre,J. and Ostell,J. (ed). *The NCBI Handbook.* National Center for Biotechnology Information, pp. 1–12..

6. Fierro,A.C., Vandenbussche,F., Engelen,K., *et al.* (2008) Meta Analysis of Gene Expression Data within and Across Species. *Curr. Genomics*, 9, 525–534.

7. Kim,B., Lee,H.J., Choi,H.Y., *et al.* (2007) Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Res.*, 67, 7431–7438.

8. Campagne,F. and Skrabanek,L. (2006) Mining expressed sequence tags identifies cancer markers of clinical interest. *BMC Bioinf.*, 7, 481.

9. Hofmann,O., Caballero,O.L., Stevenson,B.J., *et al.* (2008) Genome-wide analysis of cancer/testis gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 105, 20422–20427.

10. Simpson,A.J.G., Caballero,O.L., Jungbluth,A., *et al.* (2005) Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer*, 5, 615–25.

11. Feichtinger,J., Aldeailej,I., Anderson,R., *et al.* (2012) Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes. *Oncotarget*, 3, 843–53.

12. Fijak,M. and Meinhardt,A. (2006) The testis in immune privilege. *Immunol. Rev.*, 213, 1–121.

13. Scanlan,M.J., Gordon,C.M., Williamson,B., *et al.* (2002) Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int. J. Cancer*, 98, 485–492.

14. Flicek,P., Amode,M.R., Barrell,D., *et al.* (2011) Ensembl 2012. *Nucleic Acids Res.*, 40, 84–90.

15. Seal,R.L., Gordon,S.M., Lush,M.J., *et al.* (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, 39, D514–D519.

16. Fisher,R.A. (Oliver and Boyd: Edinburgh, 1954). *Statistical methods for research workers. Biological monographs and manuals.* Edinburgh, Oliver and Boyd.

17. Krzywinski,M., Schein,J., Birol,I., *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, 19, 1639–1645.

18. Almeida,L.G., Sakabe,N.J., deOliveira,A.R., *et al.* (2009) CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.*, 37, D816–D819.

19. De Backer,O., Arden,K.C., Boretti,M., *et al.* (1999) Characterization of the GAGE genes that are expressed in various human cancers and in normal testis. *Cancer Res.*, 59, 3157–3165.

20. van der Bruggen,P., Traversari,C., Chomez,P., *et al.* (1991) A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science*, 254, 1643–1647.

21. Brasseur,F., Rimoldi,D., Liénard,D., *et al.* (1995) Expression of MAGE genes in primary and metastatic cutaneous melanoma. *Int. J. Cancer*, 63, 375–380.

22. Jang,S.J., Soria,J.C., Wang,L., *et al.* (2001) Activation of melanoma antigen tumor antigens occurs early in lung carcinogenesis. *Cancer Res.*, 61, 7959–7963.

23. Otte,M., Zafrakas,M., Riethdorf,L., *et al.* (2001) MAGE-A gene expression pattern in primary breast cancer. *Cancer Res.*, 61, 6682–6687.

24. Sudo,T., Kuramoto,T., Komiya,S., *et al.* (1997) Expression of MAGE genes in osteosarcoma. *J. Orthop. Res.*, 15, 128–132.

25. Antonescu,C.R., Busam,K.J., Iversen,K., *et al.* (2002) MAGE antigen expression in monophasic and biphasic synovial sarcoma. *Hum. Pathol.*, 33, 225–229.

26. Chen,Y.-T., Scanlan,M.J., Venditti,C.A., *et al.* (2005) Identification of cancer/testis-antigen genes by massively parallel signature sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 102, 7940–7945.

27. Scanlan,M.J., Simpson,A.J.G. and Old,L.J. (2004) The cancer/testis genes: review, standardization, and commentary. *Cancer Immun.*, 4, 1.

28. Su,A.I., Wiltshire,T., Batalov,S., *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.*, 101, 6062–6067.

29. Chikina,M.D., Huttenhower,C., Murphy,C.T., *et al.* (2009) Global Prediction of Tissue-Specific Gene Expression and Context-Dependent Gene Networks in Caenorhabditis elegans. *PLoS Comput. Biol.*, 5, 13.

30. Ramaswamy,S., Tamayo,P., Rifkin,R., *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U. S. A.*, 98, 15149–15154.

31. Vogelstein,B. and Kinzler,K.W. (2004) Cancer genes and the pathways they control. *Nat. Med.*, 10, 789–99.

32. Skrabanek,L. and Campagne,F. (2001) TissueInfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res.*, 29, e102.

33. Liu,X., Yu,X., Zack,D.J., *et al.* (2008) TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinf.*, 9, 271.

34. Wang,J. and Liang,P. (2003) DigiNorthern, digital expression analysis of query genes based on ESTs. *Bioinformatics*, 19, 653–654.

35. Chen,Y.-C., Hsiao,C.-D., Lin,W.-D., *et al.* (2006) ZooDDD: a cross-species database for digital differential display analysis. *Bioinformatics*, 22, 2180–2182.

36. Wu,X., Walker,M.G., Luo,J., *et al.* (2005) GBA server: EST-based digital gene expression profiling. *Nucleic Acids Res.*, 33, W673–W676.

Supplemental Material

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.

#	**Tissue type**	#	**Tissue type**
1	Adipose tissue	19	Nerve
2	Adrenal gland	20	Ovary
3	Blood	21	Pancreas
4	Bone	22	Pharynx
5	Bone marrow	23	Pituitary gland
6	Brain	24	Placenta
7	Connective tissue	25	Prostate
8	Ear	26	Skin
9	Eye	27	Spleen
10	Heart	28	Stomach
11	Intestine	29	Testis
12	Kidney	30	Thymus
13	Liver	31	Thyroid
14	Lung	32	Tonsil
15	Lymph node	33	Trachea
16	Mammary gland	34	Umbilical cord
17	Mouth	35	Uterus
18	Muscle	36	Vascular

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.

Gene	**Hofmann *et al.* classification**	**CancerEST class**	**CancerEST state**	**Expression in healthy tissues**	**Expression in cancer**
CPXCR1	Testis-restricted, cancer expression	1	Specific	Testis	Skin
CSAG2	Testis-restricted, cancer expression	1a	Specific	Placenta	Prostate, mammary gland, bone, lung, connective tissue, intestine, lymph node, adrenal gland
CTAG1A	Testis-restricted, cancer expression	1	Specific	Testis	Bone
MAGEA1	Testis-restricted, cancer expression	1	Specific	Testis	Skin, bone, mammary gland, lung, connective tissue
MAGEA2	Testis-restricted, cancer expression	1	Specific	Testis	Stomach, pancreas, mammary gland
MAGEC1	Testis-restricted, cancer expression	1	Specific	Testis	Brain, placenta, bone, skin
PAGE2	Testis-restricted, cancer expression	1	Specific	Testis	Connective tissue
PAGE2B	Testis-restricted, cancer expression	1	Specific	Testis	Stomach
SPANXA1	Testis-restricted, cancer expression	1	Specific	Testis	Liver, bone marrow
SPANXA2	Testis-restricted, cancer expression	1a	Specific		Liver
SPANXD	Testis-restricted, cancer expression	1	Specific	Testis	Connective tissue, skin, liver
SSX3	Testis-restricted, cancer expression	1	Specific	Testis	Skin, bone marrow
DDX53	Testis-restricted	3	Specific	Testis	None

FTHL17	Testis-restricted	3a	Specific		None
MAGEB4	Testis-restricted	3	Specific	Testis	None
MAGEB5	Testis-restricted	3a	Specific		None
MAGEB6	Testis-restricted	3	Specific	Testis	None
SPANXB1	Testis-restricted	3a	Specific		None
SPANXC	Testis-restricted	3	Specific	Testis	None
SPANXN3	Testis-restricted	3	Specific	Testis	None
SPANXN4	Testis-restricted	3	Specific	Testis	None
SPANXN5	Testis-restricted	3a	Specific		None
XAGE5	Testis-restricted	3	Specific	Testis	None
GAGE6	Testis-restricted	1a	Specific		Stomach
DKKL1	Testis-restricted, cancer expression	2	Specific	Testis, brain	Brain, connective tissue
MAGEB1	Testis-restricted, cancer expression	2	Specific	Testis, brain	Skin
MAGEB2	Testis-restricted, cancer expression	2	Specific	Testis, brain	Skin, stomach, bone, lung
SAGE1	Testis-restricted, cancer expression	2	Specific	Testis, brain	Bone marrow
SSX1	Testis-restricted, cancer expression	2	Specific	Testis, brain	Connective tissue, skin, liver, bone, bone marrow, placenta
SSX2	Testis-restricted, cancer expression	2	Specific	Testis, brain	Liver, uterus, skin, placenta
CTAG1B	Testis-restricted, cancer expression	3a	Specific		None
TSPY1	Testis-restricted	3	Specific	Testis, brain	None
CXorf48	Testis-restricted, cancer expression	4		Testis, brain, eye	Connective tissue, bone, skin, stomach

MAGEA2B	Testis-restricted, cancer expression	4		Testis, muscle, skin	Mammary gland, connective tissue, prostate, bone marrow, skin, pancreas, bone, adrenal gland, placenta, stomach
MAGEB3	Testis-restricted	4	Enriched	Testis, prostate	None
CT69[na]	Testis-restricted	NA	NA	NA	NA
CT70[na]	Testis-restricted	NA	NA	NA	NA
LOC203413[na]	Testis-restricted	NA	NA	NA	NA
SPANXE[na]	Testis-restricted, cancer expression	NA	NA	NA	NA

6 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/`). 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*[1,2], Lee D. Larcombe†[3], and Ramsay J. McFarlane‡§[1,4]

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

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

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 *Drosophila melanogaster* ectopically express germline genes; even a quarter of the upregulated 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 *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 *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 *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 $|\text{log2-fold change}| > 1$ 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 $|\text{meta-log2-fold change}| > 1$ 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).

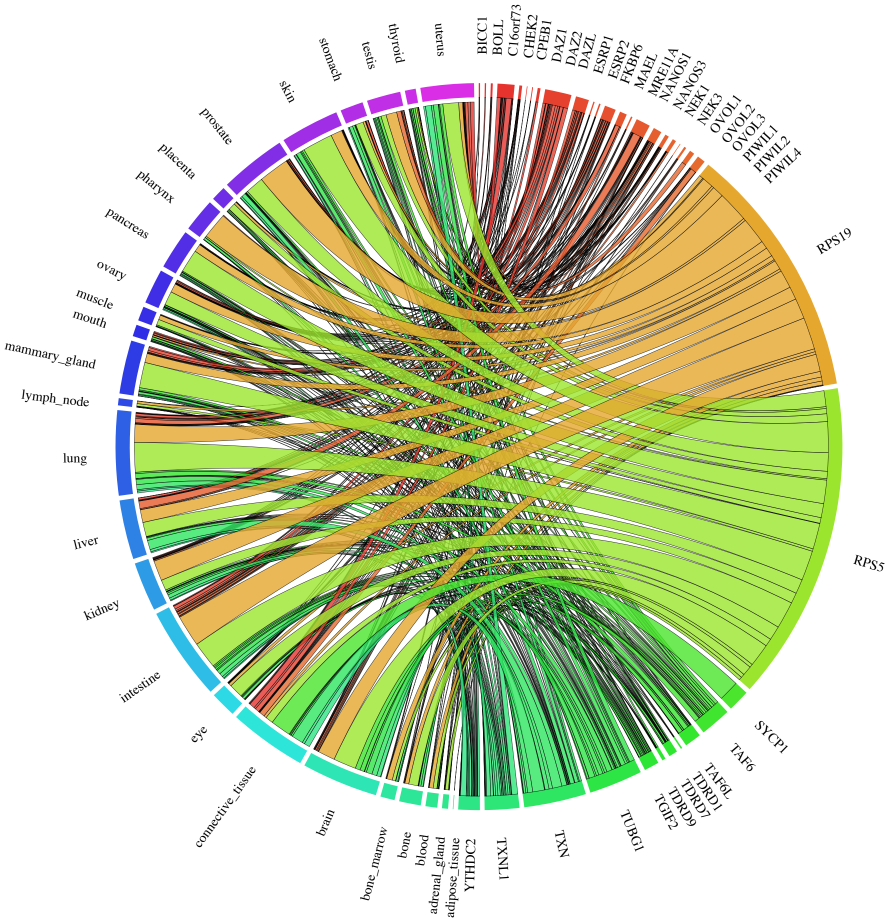


Figure 1: Circos plot showing the gene expression in relation to the corresponding cancer types for the 43 human orthologues based on the EST meta-analysis. 35 of these 43 human genes present in the Unigene database exhibit ectopic expression or are upregulated in a wide range of cancers according to the EST meta-analysis. Each connection between a gene and a cancer type indicates found expression in cancer. The magnitude of the connection corresponds to the transcripts per million (*tpm*) for the given gene in a given tissue.

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).

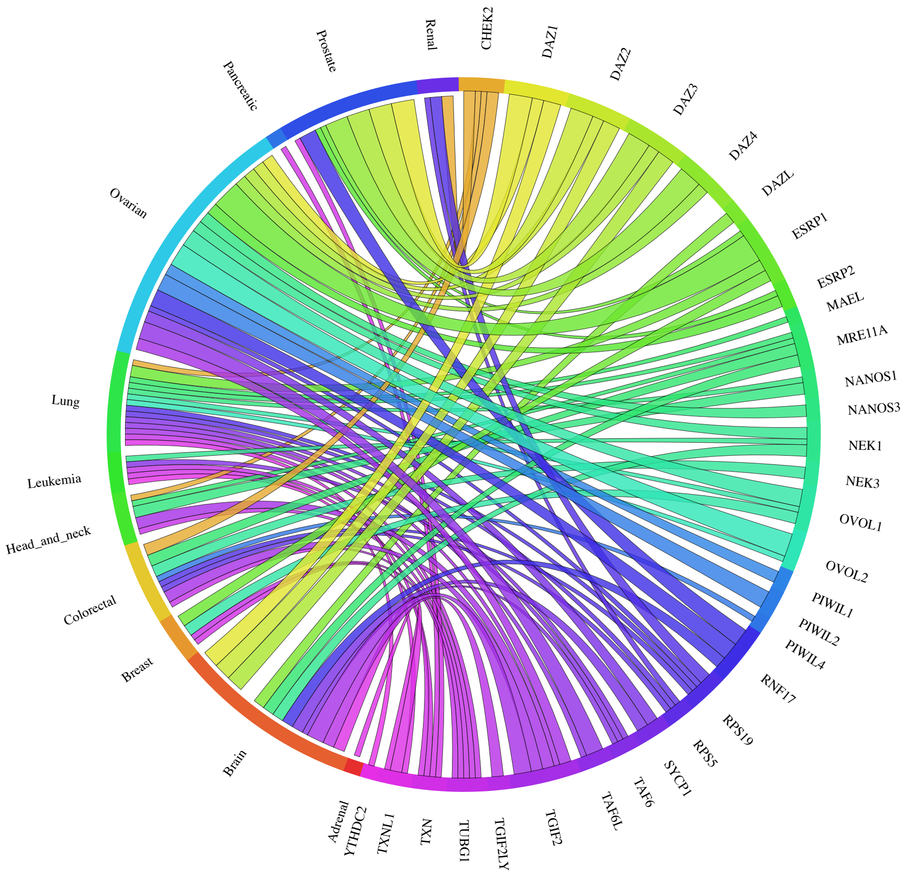


Figure 2: Circos plot showing the meta-change in gene expression in relation to corresponding cancer types for the 41 human homologues based on the microarray meta-analysis. 31 of these 41 human genes covered by the arrays exhibit an upregulation according to the microarray meta-analysis. Each connection 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-analysis, 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 belonging 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 support the acquisition of malignant attributes such as rapid proliferation, undifferentiated 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) complex [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

[1] Wu X and Ruvkun G (2010). Germ cell genes and cancer. *Science* **330**, 1761-1762.

[2] Simpson AJG, Caballero OL, Jungbluth A, Chen Y-T, and Old LJ (2005). Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer* **5**, 615-625.

[3] Janic A, Mendizabal L, Llamazares S, Rossell D, and Gonzalez C (2010). Ectopic expression of germline genes drives malignant brain tumor growth in Drosophila. *Science* **330**, 1824-1827.

[4] Boccuni P, MacGrogan D, Scandura JM, and Nimer SD (2003). The human L(3)MBT polycomb group protein is a transcriptional repressor and interacts physically and functionally with TEL (ETV6). *J Biol Chem* **278**, 15412-15420.

[5] Lewis PW, Beall EL, Fleischer TC, Georlette D, Link AJ, and Botchan MR (2004). Identification of a Drosophila Myb-E2F2/RBF transcriptional repressor complex. *Genes Dev* **18**, 2929-2940.

[6] Georlette D, Ahn S, MacAlpine DM, Cheung E, Lewis PW, Beall EL, Bell SP, Speed T, Manak JR, and Botchan MR (2007). Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells. *Genes Dev* **21**, 2880-2896.

[7] Unhavaithaya Y, Shin TH, Miliaras N, Lee J, Oyama T, and Mello CC (2002). MEP-1 and a homolog of the NURD complex component Mi-2 act together to maintain germline-soma distinctions in C. elegans. *Cell* **111**, 991-1002.

[8] Wang D, Kennedy S, Conte D, Jr., Kim JK, Gabel HW, Kamath RS, Mello CC, and Ruvkun G (2005). Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants. *Nature* **436**, 593-597.

[9] Mruk DD and Cheng CY (2010). Tight junctions in the testis: new perspectives. *Philos Trans R Soc Lond B Biol Sci* **365**, 1621-1635.

[10] Hunder NN, Wallen H, Cao J, Hendricks DW, Reilly JZ, Rodmyre R, Jungbluth A, Gnjatic S, Thompson JA, and Yee C (2008). Treatment of metastatic melanoma with autologous CD4+ T cells against NY-ESO-1. *N Engl J Med* **358**, 2698-2703.

[11] Sang M, Lian Y, Zhou X, and Shan B (2011). MAGE-A family : Attractive targets for cancer immunotherapy. *Vaccine* **29**, 8496-8500.

[12] Hofmann O, Caballero OL, Stevenson BJ, Chen Y-T, Cohen T, Chua R, Maher CA, Panji S, Schaefer U, Kruger A, et al. (2008). Genome-wide analysis of cancer/testis gene expression. *Proc Natl Acad Sci U S A* **105**, 20422-20427.

[13] Scanlan MJ, Gordon CM, Williamson B, Lee S-Y, Chen Y-T, Stockert E, Jungbluth A, Ritter G, Jäger D, Jäger E, et al. (2002). Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int J Cancer* **98**, 485-492.

[14] Feichtinger J, Aldeailej I, Anderson R, Almutairi M, Almatrafi A, Alsiwiehri N, Griffiths K, Stuart N, Wakeman JA, Larcombe L, et al. (2012). Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes. *Oncotarget* **3**, 843-853.

[15] Türeci Ö, Sahin U, Zwick C, Koslowski M, Seitz G, and Pfreundschuh M (1998). Identification of a meiosis-specific protein as a member of the class of cancer/testis antigens. *Proc Natl Acad Sci U S A* **95**, 5211-5216.

[16] Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**, D555-D559.

[17] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, and Bryant SH (2010). The NCBI BioSystems database. *Nucleic Acids Res* **38**, D492-D496.

[18] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Denise C-S, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. (2011). Ensembl 2012. *Nucleic Acids Res* **40**, 84-90.

[19] Huang DW, Sherman BT, and Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57.

[20] Pontius JU, Wagner L, and Schuler GD. (2003). UniGene: a unified view of the transcriptome. *The NCBI Handbook*. McEntyre J, and Ostell J, Eds. National Center for Biotechnology Information. pp. 1-12.

[21] Fisher RA. (1954). Statistical methods for research workers, 12th Edition (Edinburgh: Oliver and Boyd).

[22] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, and Marra MA (2009). Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645.

[23] Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, et al. (2011). ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res* **39**, D1002-D1004.

[24] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, et al. (2011). NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res* **39**, D1005-D1010.

[25] Wilson CL and Miller CJ (2005). Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* **21**, 3683-3685.

[26] Hubbell E, Liu W-M, and Mei R (2002). Robust estimators for expression analysis. *Bioinformatics* **18**, 1585-1592.

[27] Scholtens D and von Heydebreck A. (2005). Analysis of Differential Gene Expression Studies. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R, Carey V, Huber W, Irizarry R, and Dudoit S, Eds. Springer, New York, Cambridge. pp. 229-248.

[28] Smyth GK. (2005). Limma : Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R, Carey V, Dudoit S, Irizarry R, and Huber W, Eds. Springer, New York, Cambridge. pp. 397-420.

[29] Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol* **57**, 289-300.

[30] Stouffer SA. (1949). The American soldier (Princeton: Princeton University Press).

[31] Morgan AA, Khatri P, Jones RH, Sarwal MM, and Butte AJ (2010). Comparison of multiplex meta analysis techniques for understanding the acute rejection of solid organ transplants. *BMC Bioinformatics* **11**, S6.

[32] Lewis S and Clarke M (2001). Forest plots: trying to see the wood and the trees. *BMJ* **322**, 1479-1480.

[33] R Development Core Team R. (2011). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

[34] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80.

[35] Loriot A, Boon T, and De Smet C (2003). Five new human cancer-germline genes identified among 12 genes expressed in spermatogonia. *Int J Cancer* **105**, 371-376.

[36] Lee JH, Schütte D, Wulf G, Füzesi L, Radzun H-J, Schweyer S, Engel W, and Nayernia K (2006). Stem-cell protein Piwil2 is widely expressed in tumors and inhibits apoptosis through activation of Stat3/Bcl-XL pathway. *Hum Mol Genet* **15**, 201-211.

[37] Xiao L, Wang Y, Zhou Y, Sun Y, Sun W, Wang L, Zhou C, Zhou J, and Zhang J (2010). Identification of a novel human cancer/testis gene MAEL that is regulated by DNA methylation. *Mol Biol Rep* **37**, 2355-2360.

[38] Chen YT, Scanlan MJ, Venditti CA, Chua R, Theiler G, Stevenson BJ, Iseli C, Gure AO, Vasicek T, Strausberg RL, et al. (2005). Identification of cancer/testis-antigen genes by massively parallel signature sequencing. *Proc Natl Acad Sci U S A* **102**, 7940-7945.

[39] Feichtinger J, Thallinger GG, McFarlane RJ, and Larcombe L. (2012). Microarray meta-analysis: from data to expression to biological relationships. In *Computational medicine : tools and challenges.* Trajanoski Z, Ed. Springer Berlin Heidelberg, New York, NY. pp. 59-77.

[40] Kusz KM, Tomczyk L, Sajek M, Spik A, Latos-Bielenska A, Jedrzejczak P, Pawelczyk L, and Jaruzelska J (2009). The highly conserved NANOS2 protein: testis-specific expression and significance for the human male reproduction. *Mol Hum Reprod* **15**, 165-171.

[41] Reijo RA, Dorfman DM, Slee R, Renshaw AA, Loughlin KR, Cooke H, and Page DC (2000). DAZ family proteins exist throughout male germ cell development and transit from nucleus to cytoplasm at meiosis in humans and mice. *Biol Reprod* **63**, 1490-1496.

[42] Hansen CN, Ketabi Z, Rosenstierne MW, Palle C, Boesen HC, and Norrild B (2009). Expression of CPEB, GAPDH and U6snRNA in cervical and ovarian tissue during cancer development. *APMIS* **117**, 53-59.

[43] Leontieva OV and Ionov Y (2009). RNA-binding motif protein 35A is a novel tumor suppressor for colorectal cancer. *Cell Cycle* **8**, 490-497.

[44] Fay DS and Yochem J (2007). The SynMuv genes of Caenorhabditis elegans in vulval development and beyond. *Dev Biol* **306**, 1-9.

[45] Passannante M, Marti C-O, Pfefferli C, Moroni PS, Kaeser-Pebernard S, Puoti A, Hunziker P, Wicky C, and Müller F (2010). Different Mi-2 Complexes for Various Developmental Functions in Caenorhabditis elegans. *PLoS One* **5**, 15.

[46] Andersen EC, Lu X, and Horvitz HR (2006). C. elegans ISWI and NURF301 antagonize an Rb-like pathway in the determination of multiple cell fates. *Development* **133**, 2695-2704.

[47] Cui M, Kim EB, and Han M (2006). Diverse Chromatin Remodeling Genes Antagonize the Rb-Involved SynMuv Pathways in C. elegans. *PLoS Genet* **2**, 14.

[48] Strome S and Lehmann R (2007). Germ versus soma decisions: lessons from flies and worms. *Science* **316**, 392-393.

[49] Nevins JR (2001). The Rb/E2F pathway and cancer. *Hum Mol Genet* **10**, 699-703.

[50] Sherr CJ and McCormick F (2002). The RB and p53 pathways in cancer. *Cancer Cell* **2**, 103-112.

[51] Wang J, Emadali A, Le Bescont A, Callanan M, Rousseaux S, and Khochbin S (2011). Induced malignant genome reprogramming in somatic cells by testis-specific factors. *Biochim Biophys Acta* **1809**, 221-225.

[52] Chalmel F, Lardenois A, and Primig M (2007). Toward understanding the core meiotic transcriptome in mammals and its implications for somatic cancer. *Ann N Y Acad Sci* **1120**, 1-15.

[53] Kalejs M, Ivanov A, Plakhins G, Cragg MS, Emzinsh D, Illidge TM, and Erenpreisa J (2006). Upregulation of meiosis-specific genes in lymphoma cell lines following genotoxic insult and induction of mitotic catastrophe. *Bmc Cancer* **6**, 6.

Supplemental Material

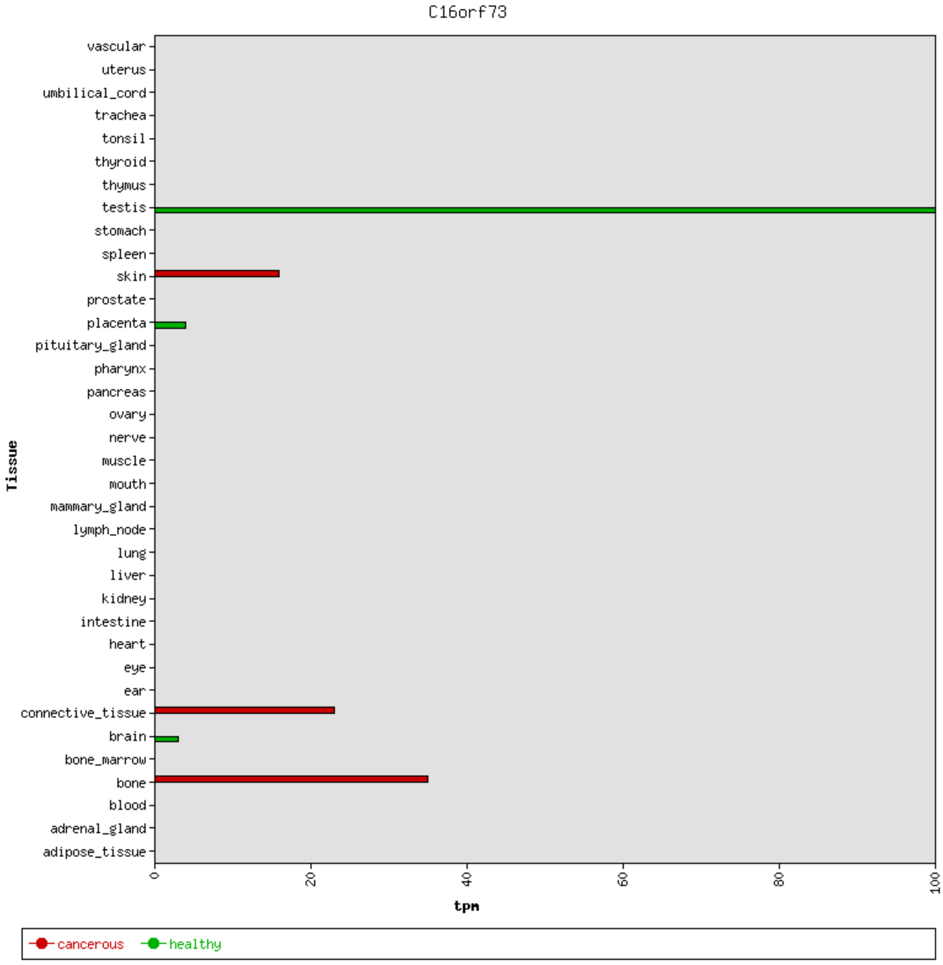


Figure S1: An example of a bar chart showing the integrated expression profile of the *C16orf73* gene. *C16orf73* exhibits expression restricted to the brain, placenta and testis, but is aberrantly expressed in melanoma and sarcomas of the bone and of the connective tissue. The expression is given in transcripts per million (*tpm*).

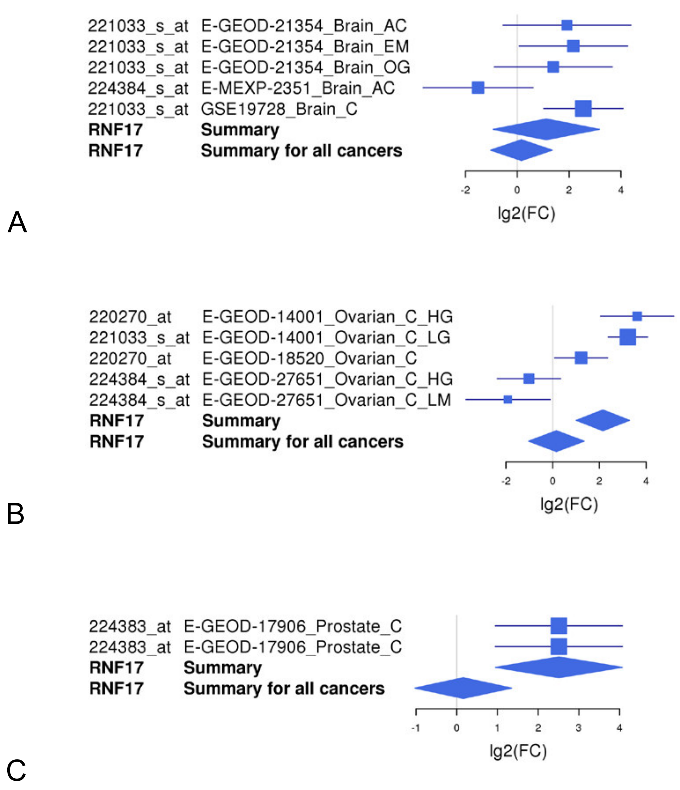


Figure S2: Forest plots for the gene *RN17*. *RN17* is upregulated in (A) ovarian, (B) prostate and (C) brain cancer, according to the microarray meta-analysis. A Forest plot shows the log 2-fold change (lg2FC) values for the individual studies as well as the total values for the given cancer type and 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.

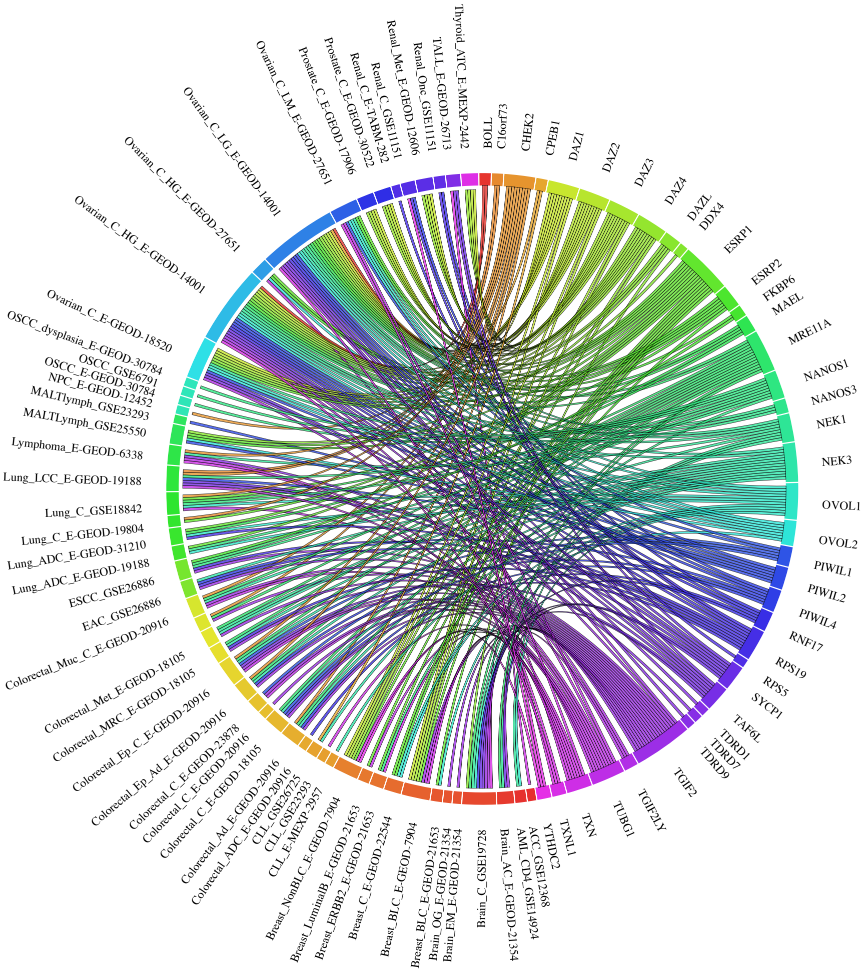


Figure S3: Circos plot for the 41 human homologues based on the single microarray analysis. The Circos plot shows that 39 of the total 41 human homologues covered by the arrays are upregulated in a wide range of cancer types. Each connection between a gene and an individual cancer type indicates a statistically 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.

Gene	**Flybase ID**	**Human orthologues**	**Human Ensembl IDs**	**Source**
AGO3	FBgn0250816	*PIWIL4, PIWIL2, PIWIL3, PIWIL1*	ENSG00000134627, ENSG00000197181, ENSG00000184571, ENSG00000125207	Flybase, Ensembl Biomart
aub	FBgn0000146	*PIWIL4, PIWIL2, PIWIL3, PIWIL1*	ENSG00000134627, ENSG00000197181, ENSG00000184571, ENSG00000125207	Flybase
bam	FBgn0000158			
bgcn	FBgn0004581	*YTHDC2*	ENSG00000047188	Ensembl Biomart
BicC	FBgn0000182	*BICC1*	ENSG00000122870	Flybase, Ensembl Biomart
bol	FBgn0011206	*DAZ2, DAZ3, DAZ4, BOLL, DAZL, DAZ1*	ENSG00000205944, ENSG00000187191, ENSG00000205916, ENSG00000152430, ENSG00000092345, ENSG00000188120	Flybase, Ensembl Biomart
c(3)G	FBgn0000246	*SYCP1*	ENSG00000198765	[5]
CG15930	FBgn0029754			
CG31755	FBgn0051755			
CG32313	FBgn0052313			
CG40115	FBgn0058115			

CG7194	FBgn0035868			
CG7795	FBgn0262598			
CG9925	FBgn0038191			
cona	FBgn0038612			
del	FBgn0086251			
dhd	FBgn0011761	*TXN, TXNL1*	ENSG00000136810, ENSG00000091164	Flybase
fs(1)Yb	FBgn0000928			
fus	FBgn0023441	*ESRP1, ESRP2*	ENSG00000104413, ENSG00000103067	Flybase, Ensembl Biomart
gnu	FBgn0001120			
hdm	FBgn0029977	*C16orf73*	ENSG00000162039	Flybase, Ensembl Biomart
krimp	FBgn0034098	*TDRD1, RNF17*	ENSG00000095627, ENSG00000132972	[6] (potential functional homologues)
loki	FBgn0019686	*CHEK2*	ENSG00000183765	Flybase, Ensembl Biomart
mael	FBgn0016034	*MAEL*	ENSG00000143194	[7,8]
mia	FBgn0014342	*TAF6, TAF6L*	ENSG00000106290, ENSG00000162227	Flybase, Ensembl Biomart
mre11	FBgn0020270	*MRE11A*	ENSG00000020922	Flybase, Ensembl Biomart
Mst57Db	FBgn0011669			
Mst77F	FBgn0086915			
nos	FBgn0002962	*NANOS3, NANOS2, NANOS1*	ENSG00000187556, ENSG00000188425, ENSG00000188613	Flybase, Ensembl Biomart

orb	FBgn0004882	*CPEB1, RP11-152F13.10*	ENSG00000214575, ENSG00000260836	Flybase, Ensembl Biomart
ovo	FBgn0003028	*OVOL1, OVOL2, OVOL3*	ENSG00000172818, ENSG00000125850, ENSG00000105261	Flybase, Ensembl Biomart
piwi	FBgn0004872	*PIWIL4, PIWIL2, PIWIL3, PIWIL1*	ENSG00000134627, ENSG00000197181, ENSG00000184571, ENSG00000125207	Flybase
png	FBgn0000826	*NEK1, NEK5, NEK3*	ENSG00000137601, ENSG00000197168, ENSG00000136098	Ensembl Biomart
Pxt	FBgn0261987			
RpS19b	FBgn0039129	*RPS19*	ENSG00000105372	Flybase, Ensembl Biomart
RpS5b	FBgn0038277	*RPS5*	ENSG00000083845	Flybase, Ensembl Biomart
shu	FBgn0003401	*FKBP6*	ENSG00000077800	Flybase, Ensembl Biomart
spn-E	FBgn0003483	*TDRD9*	ENSG00000156414	Flybase, Ensembl Biomart
squ	FBgn0002652			
stil	FBgn0003527			
swa	FBgn0003655			
tej	FBgn0033921	*TDRD7*	ENSG00000196116	Ensembl Biomart
topi	FBgn0037751			
tor	FBgn0003733			
TrxT	FBgn0029752			

vasa	FBgn0262526	*DDX4*	ENSG00000152670	[9]
vis	FBgn0033748	*TXN, TXNL1, TGIF2LX, TGIF2LY*	ENSG00000136810, ENSG00000091164, ENSG00000153779, ENSG00000176679	Flybase, Ensembl Biomart
zpg	FBgn0024177	*TGIF2*	ENSG00000118707	Flybase
γTub37C	FBgn0010097	*TUBG1*		Homologene

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.

GO ID	GO term	Gene count	Percentage	p value	Genes
GO:0006417	Regulation of translation	12	28.6	4.61E-14	*DAZ3, DAZ4, NANOS3, DAZ1, DAZ2, NANOS2, NANOS1, CPEB1, RPS5, BOLL, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4*
GO:0051321	Meiotic cell cycle	11	26.2	8.06E-14	*MRE11A, OVOL1, MAEL, PIWIL1, PIWIL2, PIWIL3, PIWIL4, TUBG1, SYCP1, BOLL, TDRD1*
GO:0048232	Male gamete generation	14	33.3	6.55E-13	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*
GO:0007283	Spermatogenesis	14	33.3	6.55E-13	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*
GO:0007276	Gamete generation	15	35.7	7.54E-13	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*

GO:0007126	Meiosis	10	23.8	3.56E-12	*MRE11A, OVOL1, MAEL, PIWIL1, PIWIL2, PIWIL3, PIWIL4, SYCP1, BOLL, TDRD1*
GO:0051327	M phase of meiotic cell cycle	10	23.8	3.56E-12	*MRE11A, OVOL1, MAEL, PIWIL1, PIWIL2, PIWIL3, PIWIL4, SYCP1, BOLL, TDRD1*
GO:0010608	Post-transcriptional regulation of gene expression	12	28.6	5.42E-12	*DAZ3, DAZ4, NANOS3, DAZ1, DAZ2, NANOS2, NANOS1, CPEB1, RPS5, BOLL, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4*
GO:0019953	Sexual reproduction	15	35.7	5.58E-12	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*
GO:0048609	Reproductive process in a multicellular organism	15	35.7	1.27E-11	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*
GO:0032504	Multicellular organism reproduction	15	35.7	1.27E-11	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*

GO:0000279	M phase	13	31.0	3.31E-11	*NEK3, NEK1, MRE11A, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1*
GO:0034587	piRNA metabolic process	5	11.9	2.23E-10	*MAEL, PIWIL1, PIWIL2, PIWIL4, TDRD1*
GO:0022403	Cell cycle phase	13	31.0	4.69E-10	*NEK3, NEK1, MRE11A, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1*
GO:0022414	Reproductive process	15	35.7	4.56E-09	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*
GO:0000003	Reproduction	15	35.7	4.96E-09	*DAZ3, NANOS3, DAZ4, RNF17, DAZ1, DAZ2, TDRD7, NANOS2, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4, TDRD1*
GO:0022402	Cell cycle process	13	31.0	1.57E-08	*NEK3, NEK1, MRE11A, MAEL, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1*

GO:0031047	Gene silencing by RNA	6	14.3	2.42E-08	*MAEL, PIWIL1, PIWIL2, PIWIL3, PIWIL4, TDRD1*
GO:0032268	Regulation of cellular protein metabolic process	12	28.6	2.89E-08	*DAZ3, DAZ4, NANOS3, DAZ1, DAZ2, NANOS2, NANOS1, CPEB1, RPS5, BOLL, PIWIL1, PIWIL2, PIWIL3, DAZL, PIWIL4*
GO:0007049	Cell cycle	14	33.3	5.70E-08	*NEK3, NEK1, MRE11A, MAEL, CHEK2, SYCP1, BOLL, OVOL1, PIWIL1, PIWIL2, PIWIL3, TUBG1, PIWIL4, TDRD1*

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.

UniGene ID	**Gene**	**Expression in immunologically privileged areas [*tpm*]**	**Upregulation/ectopic expression in cancer**	**Number of other healthy tissues were expression was found**
Hs.729604	*C16orf73*	Brain (3), placenta (4), testis (100)	Bone, skin, connective tissue	
Hs.661266	*FKBP6*	Brain (9), testis (174)	Blood, ovary, lung	
Hs.592257	*DAZ2*	Testis (13)	Lung, stomach	
Hs.112743	*SYCP1*	Brain (1), testis (182)	Connective tissue	
Hs.169797	*BOLL*	Brain (6), testis (330)	Connective tissue	
Hs.223581	*DDX4*	Brain (3), testis (578)		
Hs.97464	*RNF17*	Placenta (12), testis (165)		
Hs.112148	*TGIF2LY*	Testis (35)		
Hs.448343	*PIWIL3*	Testis (13)		
Hs.592220	*TGIF2LX*	Brain (1), testis (9)		
Hs.434218	*NANOS2*	Brain (1), testis (4)		
Hs.661013	*OVOL2*	-	Placenta, prostate, uterus, ovary, intestine, thyroid, stomach	1 (Lung)
Hs.333132	*TDRD1*	Brain (2), testis (113)	Prostate, liver, pancreas	1 (Trachea)
Hs.127982	*NANOS3*	Brain (1)	Mouth, connective tissue, lung	1 (Eye)
Hs.131179	*DAZL*	Brain (6), testis (426)	Lung, connective tissue	1 (Muscle)
Hs.143920	*OVOL3*	Testis (9)	Connective tissue	1 (Ovary)

Hs.405659	*PIWIL1*	Brain (2), testis (512)	Intestine, uterus	2 (Muscle, eye)
Hs.614809	*PIWIL2*	Brain (3), placenta (4), testis (100)	Testis	2 (Muscle, intestine)
Hs.672144	*NEK5*	Placenta (4), testis (4)		2 (Pharynx, trachea)
Hs.70936	*DAZ1*	Testis (39)	Stomach	3 tissues
Hs.591918	*NANOS1*	Brain (1), placenta (29)	Brain, intestine, lung, muscle, ovary	3 tissues
Hs.651245	*MAEL*	Brain (31), placenta (8), testis (1107)	Liver, mammary gland	3 tissues
Hs.158745	*BICC1*	Brain (3)		4 tissues
Hs.660188	*PIWIL4*	Brain (2), placenta (8), testis (178)	Pancreas, connective tissue, blood, ovary, eye, intestine, liver	5 tissues
Hs.21454	*TDRD9*	Brain (9), testis (122)	Connective tissue, bone marrow, ovary, thyroid	6 tissues
Hs.134434	*OVOL1*	Placenta (17), testis (96)	Mammary gland, uterus, prostate, pancreas, stomach, eye	7 tissues
Hs.547988	*CPEB1*	Brain (50), testis (35)	Skin, liver, adrenal gland, bone marrow, pancreas, intestine	7 tissues
Hs.714400	*TAF6L*	Brain (16), placenta (21), testis (26)	Pancreas, thyroid, stomach, ovary, intestine, mouth, blood, brain, lung, mammary gland, bone	10 tissues

Hs.632264	*TGIF2*	Brain (4), testis (9)	Muscle, mammary gland, mouth, thyroid, lung, ovary, eye, placenta, intestine, testis, skin, liver, brain	11 tissues
Hs.409989	*NEK3*	Brain (8), placenta (12), testis (39)	Stomach, mouth, liver, bone, ovary, connective tissue	11 tissues
Hs.291363	*CHEK2*	Brain (2), placenta (8), testis (22)	Bone marrow, thyroid, connective tissue, skin, mammary gland, uterus	11 tissues
Hs.487471	*ESRP1*	Brain (1), placenta (25), testis (9)	Eye, stomach, mammary gland, mouth, uterus, connective tissue, ovary, thyroid, liver, pharynx	12 tissues
Hs.193842	*TDRD7*	Brain (19), placenta (8), testis (87)	Stomach, lymph node, mammary gland, ovary, pharynx, connective tissue, intestine	13 tissues
Hs.481181	*NEK1*	Brain (17), testis (9)	Eye, ovary, adrenal gland, intestine, bone, liver	14 tissues
Hs.592053	*ESRP2*	Placenta (70), testis (17)	Mouth, connective tissue, pancreas, ovary, adrenal gland	14 tissues
Hs.231942	*YTHDC2*	Brain (17), placenta (17), testis (48)	Ovary, mouth, eye, bone marrow, thyroid, adrenal gland, lung, connective tissue	18 tissues

Hs.192649	*MRE11A*	Brain (9), placenta (41), testis (4)	Mouth, stomach, thyroid, testis, bone, prostate, mammary gland	18 tissues
Hs.489309	*TAF6*	Brain (40), placenta (54), testis (30)	Eye, stomach, testis, liver, pancreas, thyroid, brain, connective tissue	19 tissues
Hs.435136	*TXN*	Brain (21), placenta (83), testis (35)	Brain, adrenal gland, mouth, ovary, stomach, connective tissue, kidney	21 tissues
Hs.279669	*TUBG1*	Brain (45), placenta (66), testis (287)	Brain, mouth, thyroid, skin, pharynx, liver, eye, lung, connective tissue, ovary	22 tissues
Hs.114412	*TXNL1*	Brain (40), placenta (25), testis (30)	Thyroid, stomach, uterus, liver, connective tissue, brain, mammary gland	25 tissues
Hs.438429	*RPS19*	Brain (48), placenta (75), testis (43)	Placenta, prostate, bone, kidney, skin, connective tissue, testis, thyroid, lymph node, pharynx, mammary gland, mouth, lung, brain, pancreas	28 tissues
Hs.378103	*RPS5*	Brain (111), placenta (199), testis (69)	Bone, pancreas, brain, uterus, muscle, kidney, liver, eye, placenta, prostate, skin, mammary gland, testis, lung, intestine, lymph node	31 tissues

Table S4: List the human homologues of the *Drosophila* genes ectopically expressed in *l(3)mbt* 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.

Gene	Ensembl ID	Cancer type	Weighted $log2FC$	Weighted CI left	Weighted CI right
CHEK2	ENSG00000183765	Colorectal cancer	1.27	0.93	1.6
CHEK2	ENSG00000183765	Head and neck cancer	0.48	0.1	0.86
CHEK2	ENSG00000183765	Renal cancer	1.1	0.29	1.9
CHEK2	ENSG00000183765	Lung cancer	0.68	0.39	0.97
DAZ1	ENSG00000188120	Ovarian cancer	1.87	0.69	3.05
DAZ1	ENSG00000188120	Brain cancer	2.42	0.94	3.9
DAZ1	ENSG00000188120	Prostate cancer	3.3	2	4.59
DAZ2	ENSG00000205944	Ovarian cancer	1.87	0.69	3.05
DAZ2	ENSG00000205944	Brain cancer	2.42	0.94	3.9
DAZ2	ENSG00000205944	Prostate cancer	3.3	2	4.59
DAZ3	ENSG00000187191	Ovarian cancer	1.87	0.69	3.05
DAZ3	ENSG00000187191	Brain cancer	2.42	0.94	3.9
DAZ3	ENSG00000187191	Prostate cancer	3.3	2	4.59
DAZ4	ENSG00000205916	Ovarian cancer	1.87	0.69	3.05
DAZ4	ENSG00000205916	Brain cancer	2.42	0.94	3.9
DAZ4	ENSG00000205916	Prostate cancer	3.3	2	4.59
DAZL	ENSG00000092345	Brain cancer	1.78	-0.27	3.83
ESRP1	ENSG00000104413	Breast cancer	1.24	0.62	1.86
ESRP1	ENSG00000104413	Ovarian cancer	4.06	3.23	4.89
ESRP1	ENSG00000104413	Lung cancer	1.13	0.74	1.52

ESRP1	ENSG00000104413	Prostate cancer	0.84	0.34	1.34
ESRP2	ENSG00000103067	Ovarian cancer	1.85	1.03	2.67
ESRP2	ENSG00000103067	Prostate cancer	0.95	0.55	1.35
MAEL	ENSG00000143194	Lung cancer	0.74	0.15	1.34
MRE11A	ENSG00000020922	Colorectal cancer	1.82	1.45	2.18
MRE11A	ENSG00000020922	Head and neck cancer	0.59	0.09	1.1
MRE11A	ENSG00000020922	Lung cancer	0.35	0.1	0.61
MRE11A	ENSG00000020922	Brain cancer	1.94	0.74	3.15
NANOS1	ENSG00000188613	Head and neck cancer	1.05	0.27	1.83
NANOS1	ENSG00000188613	Lung cancer	0.86	0.45	1.26
NANOS3	ENSG00000187556	Ovarian cancer	1.13	0.27	1.99
NEK1	ENSG00000137601	Ovarian cancer	1.71	0.67	2.75
NEK1	ENSG00000137601	Brain cancer	1.97	0.25	3.69
NEK1	ENSG00000137601	Leukemia	0.78	0.15	1.41
NEK3	ENSG00000136098	Colorectal cancer	1.12	0.79	1.44
OVOL1	ENSG00000172818	Breast cancer	1.01	0	2.02
OVOL1	ENSG00000172818	Ovarian cancer	2.16	1.32	3.01
OVOL1	ENSG00000172818	Lung cancer	0.84	0.45	1.22
OVOL2	ENSG00000125850	Ovarian cancer	3.06	2.26	3.85
OVOL2	ENSG00000125850	Lung cancer	0.46	0.15	0.76
PIWIL1	ENSG00000125207	Ovarian cancer	1.44	0.31	2.56
PIWIL2	ENSG00000197181	Ovarian cancer	2.3	1.45	3.16
PIWIL4	ENSG00000134627	Colorectal cancer	0.54	0.13	0.95
RNF17	ENSG00000132972	Ovarian cancer	2.15	1.01	3.29
RNF17	ENSG00000132972	Brain cancer	1.12	-0.93	3.17

RNF17	ENSG00000132972	Prostate cancer	2.51	0.95	4.07
RPS19	ENSG00000105372	Colorectal cancer	0.37	0.05	0.68
RPS19	ENSG00000105372	Renal cancer	1.05	0.52	1.58
RPS19	ENSG00000105372	Ovarian cancer	0.78	0.16	1.41
RPS19	ENSG00000105372	Lung cancer	0.66	0.34	0.98
RPS5	ENSG00000083845	Colorectal cancer	0.39	0.05	0.72
RPS5	ENSG00000083845	Renal cancer	0.79	0.21	1.37
RPS5	ENSG00000083845	Lung cancer	0.25	0.08	0.42
SYCP1	ENSG00000198765	Ovarian cancer	1.64	0.55	2.72
SYCP1	ENSG00000198765	Brain cancer	1.77	0.18	3.36
TAF6	ENSG00000106290	Lung cancer	0.51	0.29	0.73
TAF6	ENSG00000106290	Brain cancer	0.86	0.01	1.71
TAF6	ENSG00000106290	Leukemia	0.39	0.06	0.72
TAF6L	ENSG00000162227	Ovarian cancer	2.2	1.12	3.27
TAF6L	ENSG00000162227	Lung cancer	0.51	0.02	0.99
TGIF2	ENSG00000118707	Colorectal cancer	1.67	1.27	2.06
TGIF2	ENSG00000118707	Head and neck cancer	1.12	0.6	1.64
TGIF2	ENSG00000118707	Ovarian cancer	1.49	0.84	2.13
TGIF2	ENSG00000118707	Lung cancer	0.31	0.08	0.53
TGIF2	ENSG00000118707	Brain cancer	2.09	1.08	3.1
TGIF2LY	ENSG00000176679	Brain cancer	1.04	-0.52	2.59
TUBG1	ENSG00000131462	Colorectal cancer	0.86	0.52	1.2
TUBG1	ENSG00000131462	Breast cancer	0.61	0.15	1.06
TUBG1	ENSG00000131462	Head and neck cancer	0.6	0.23	0.97
TUBG1	ENSG00000131462	Lung cancer	0.88	0.64	1.11

TUBG1	ENSG00000131462	Leukemia	0.63	0.21	1.05
TXN	ENSG00000136810	Breast cancer	0.61	0.14	1.09
TXN	ENSG00000136810	Pancreatic cancer	0.7	0.17	1.24
TXN	ENSG00000136810	Leukemia	0.64	0.24	1.04
TXN	ENSG00000136810	Prostate cancer	0.49	0.19	0.79
TXNL1	ENSG00000091164	Adrenal cancer	0.39	0.03	0.75
TXNL1	ENSG00000091164	Lung cancer	0.2	0.04	0.35
TXNL1	ENSG00000091164	Brain cancer	1.11	-0.03	2.25
YTHDC2	ENSG00000047188	Leukemia	0.62	0.1	1.14

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.

Gene	**Chrom.**	**Ensembl ID**	**UniGene ID**	**Array**	**Expression in immunolog. privileged tissues**	**Expression in normal tissues**	**Upregulation/ectopic expression in cancer (method)**
BOLL	2	ENSG00000152430	Hs.169797	On array	Testis, brain	0 tissues	E, S
C16orf73	16	ENSG00000162039	Hs.729604	On array	Testis, brain, placenta	0 tissues	E, S
DAZ2	Y	ENSG00000205944	Hs.592257	On array	Testis	0 tissues	E, M, S
DDX4	5	ENSG00000152670	Hs.223581	On array	Testis, brain	0 tissues	S
FKBP6	7	ENSG00000077800	Hs.661266	On array	Testis, brain	0 tissues	E, S
NANOS2	19	ENSG00000188425	Hs.434218	On array	Testis, brain	0 tissues	-
PIWIL3	22	ENSG00000184571	Hs.448343	NA	Testis	0 tissues	-
RNF17	13	ENSG00000132972	Hs.97464	On array	Testis, placenta	0 tissues	M, S
*SYCP1**	1	ENSG00000198765	Hs.112743	On array	Testis, brain	0 tissues	E, M, S
TGIF2LX	X	ENSG00000153779	Hs.592220	NA	Testis, brain	0 tissues	-
TGIF2LY	Y	ENSG00000176679	Hs.112148	On array	Testis	0 tissues	M, S
DAZL	3	ENSG00000092345	Hs.131179	On array	Testis, brain	1 tissue (Muscle)	E, M, S
NANOS3	19	ENSG00000187556	Hs.127982	On array	Brain	1 tissue (Eye)	E, M, S
OVOL2	20	ENSG00000125850	Hs.661013	On array	-	1 tissue (Lung)	E, M, S
OVOL3	19	ENSG00000105261	Hs.143920	NA	Testis	1 tissue (Ovary)	E
*TDRD1**	10	ENSG00000095627	Hs.333132	On array	Testis, brain	1 tissue (Trachea)	E, S

NEK5	13	ENSG00000197168	Hs.672144	NA	Testis, placenta	2 tissues (Pharynx, trachea)	-
PIWIL1	12	ENSG00000125207	Hs.405659	On array	Testis, brain	2 tissues (Muscle, eye)	E, M, S
*PIWIL2**	8	ENSG00000197181	Hs.614809	On array	Testis, brain, placenta	2 tissues (Muscle, intestine)	E, M, S
DAZ1	Y	ENSG00000188120	Hs.522868	On array	Testis	3 tissues	E, M, S
*MAEL**	1	ENSG00000143194	Hs.651245	On array	Testis, brain, placenta	3 tissues	E, M, S
NANOS1	10	ENSG00000188613	Hs.591918	On array	Brain, placenta	3 tissues	E, M, S
BICC1	10	ENSG00000122870	Hs.158745	On array	Brain	4 tissues	-
PIWIL4	11	ENSG00000134627	Hs.660188	On array	Testis, brain, placenta	5 tissues	E, M, S
TDRD9	14	ENSG00000156414	Hs.21454	On array	Testis, brain	6 tissues	E, S
CPEB1	15	ENSG00000214575	Hs.547988	On array	Testis, brain	7 tissues	E, S
OVOL1	11	ENSG00000172818	Hs.134434	On array	Testis, placenta	7 tissues	E, M, S
TAF6L	11	ENSG00000162227	Hs.714400	On array	Testis, brain, placenta	10 tissues	E, M, S
CHEK2	22	ENSG00000183765	Hs.291363	On array	Testis, brain, placenta	11 tissues	E, M, S
NEK3	13	ENSG00000136098	Hs.409989	On array	Testis, brain, placenta	11 tissues	E, M, S
TGIF2	20	ENSG00000118707	Hs.632264	On array	Testis, brain	11 tissues	E, M, S
ESRP1	8	ENSG00000104413	Hs.487471	On array	Testis, brain, placenta	12 tissues	E, M, S
TDRD7	9	ENSG00000196116	Hs.193842	On array	Testis, brain, placenta	13 tissues	E, S
ESRP2	16	ENSG00000103067	Hs.592053	On array	Testis, placenta	14 tissues	E, M, S
NEK1	4	ENSG00000137601	Hs.481181	On array	Testis, brain	14 tissues	E, M, S
MRE11A	11	ENSG00000020922	Hs.192649	On array	Testis, brain, placenta	18 tissues	E, M, S

YTHDC2	5	ENSG00000047188	Hs.231942	On array	Testis, brain, placenta	18 tissues	E, M, S
TAF6	7	ENSG00000106290	Hs.489309	On array	Testis, brain, placenta	19 tissues	E, M, S
TXN	9	ENSG00000136810	Hs.435136	On array	Testis, brain, placenta	21 tissues	E, M, S
TUBG1	17	ENSG00000131462	Hs.279669	On array	Testis, brain, placenta	22 tissues	E, M, S
TXNL1	18	ENSG00000091164	Hs.114412	On array	Testis, brain, placenta	25 tissues	E, M, S
RPS19	19	ENSG00000105372	Hs.438429	On array	Testis, brain, placenta	28 tissues	E, M, S
RPS5	19	ENSG00000083845	Hs.378103	On array	Testis, brain, placenta	31 tissues	E, M, S
DAZ3	Y	ENSG00000187191	NA	On array	NA	NA	M, S
DAZ4	Y	ENSG00000205916	NA	On array	NA	NA	M, S
RP11-152F13.10	15	ENSG00000260836	NA	NA	NA	NA	NA

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

[1] Janic A, Mendizabal L, Llamazares S, Rossell D, and Gonzalez C (2010). Ectopic expression of germline genes drives malignant brain tumor growth in Drosophila. *Science* **330**, 1824-1827.

[2] Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**, D555-D559.

[3] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, and Bryant SH (2010). The NCBI BioSystems database. *Nucleic Acids Res* 38, D492-D496.

[4] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Denise C-S, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. (2011). Ensembl 2012. *Nucleic Acids Res* **40**, 84-90.

[5] Türeci Ö, Sahin U, Zwick C, Koslowski M, Seitz G, and Pfreundschuh M (1998). Identification of a meiosis-specific protein as a member of the class of cancer/testis antigens. *Proc Natl Acad Sci* U S A **95**, 5211-5216.

[6] Lim AK and Kai T (2007). Unique germ-line organelle, nuage, functions to repress selfish genetic elements in Drosophila melanogaster. *Proc Natl Acad Sci* U S A **104**, 6714-6719.

[7] Costa Y, Speed RM, Gautier P, Semple CA, Maratou K, Turner JM, and Cooke HJ (2006). Mouse MAELSTROM: the link between meiotic silencing of unsynapsed chromatin and microRNA pathway? *Hum Mol Genet* **15**, 2324-2334.

[8] Findley SD, Tamanaha M, Clegg NJ, and Ruohola-Baker H (2003). Maelstrom, a Drosophila spindle-class gene, encodes a protein that colocalizes with Vasa and RDE1/AGO1 homolog, Aubergine, in nuage. *Development* **130**, 859-871.

[9] Castrillon DH, Quade BJ, Wang TY, Quigley C, and Crum CP (2000). The human VASA gene is specifically expressed in the germ cell lineage. *Proc Natl Acad Sci* U S A **97**, 9585-9590.

[10] Huang DW, Sherman BT, and Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57.

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]. Micorarrays 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].

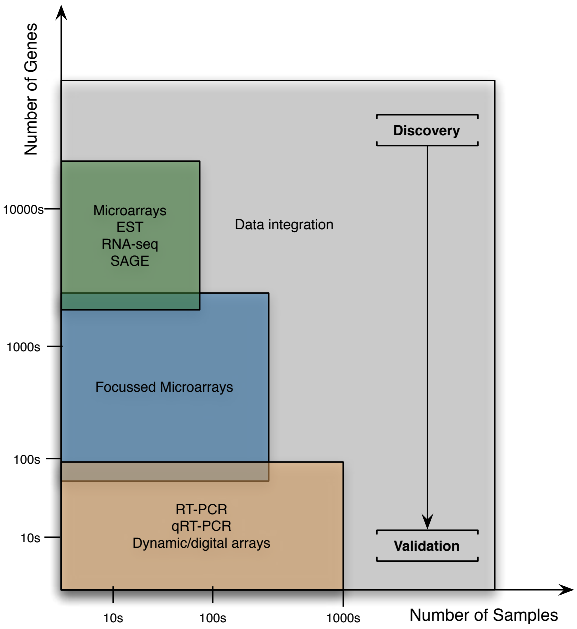


Figure 7.1: High-throughput expression profiling techniques in comparison with high-sensitivity methods. High-throughput methods are used for screening to discover new target genes and yield a snapshot of a certain tissue- or condition-specific transcriptome. In contrast, high-sensitivity methods are 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/`), 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

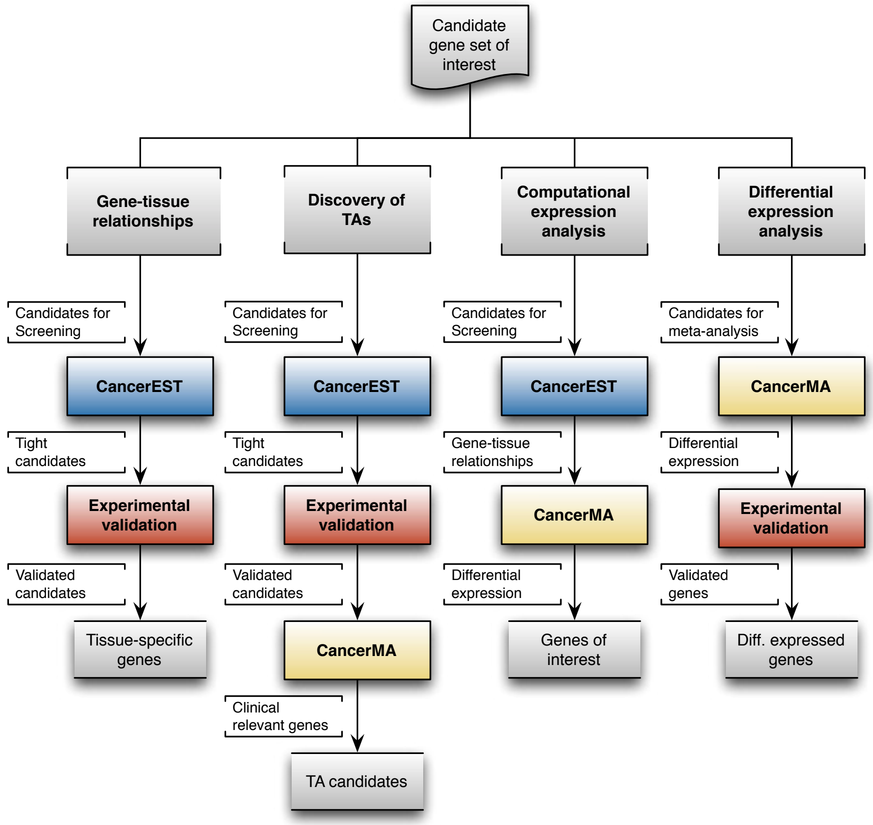


Figure 7.2: Examples of applications and workflows for both pipelines. CancerMA and CancerEST might be used separately or in combination. CancerEST allows inferring tissue-specificity and gene-tissue relationships. CancerMA, in contrast, allows analysing the differential expression across 13 cancer types 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.

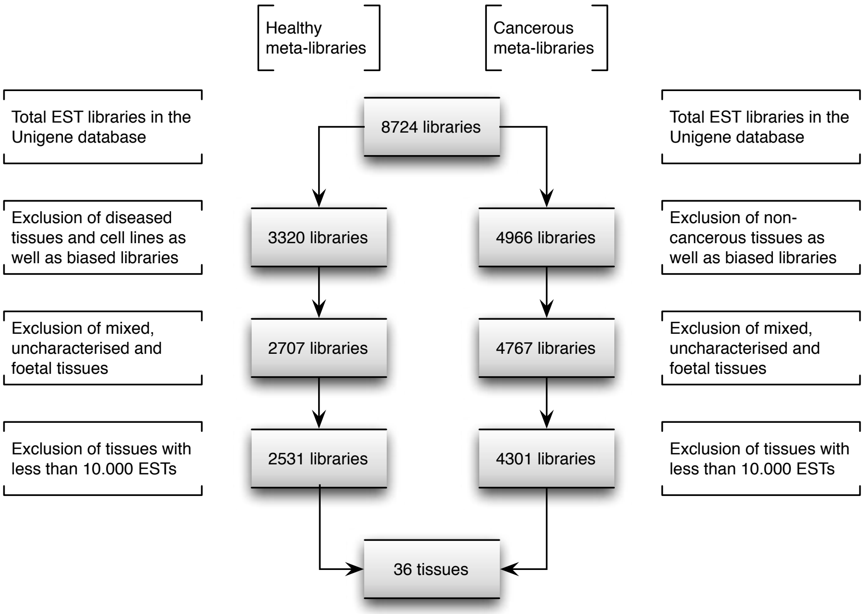

Healthy meta-libraries
Cancerous meta-libraries
Total EST libraries in the Unigene database
8724 libraries
Total EST libraries in the Unigene database
Exclusion of diseased tissues and cell lines as well as biased libraries
3320 libraries
4966 libraries
Exclusion of non-cancerous tissues as well as biased libraries
Exclusion of mixed, uncharacterised and foetal tissues
2707 libraries
4767 libraries
Exclusion of mixed, uncharacterised and foetal tissues
Exclusion of tissues with less than 10.000 ESTs
2531 libraries
4301 libraries
Exclusion of tissues with less than 10.000 ESTs
36 tissues

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 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 based on 4301 and 2531 libraries, respectively.

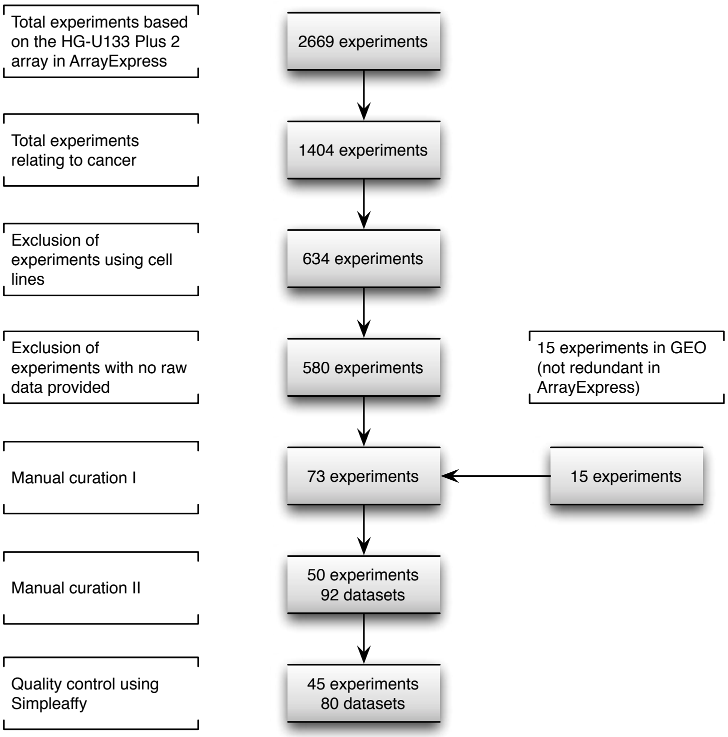

Total experiments based on the HG-U133 Plus 2 array in ArrayExpress
2669 experiments
Total experiments relating to cancer
1404 experiments
Exclusion of experiments using cell lines
634 experiments
Exclusion of experiments with no raw data provided
580 experiments
15 experiments in GEO (not redundant in ArrayExpress)
Manual curation I
73 experiments
15 experiments
Manual curation II
50 experiments 92 datasets
Quality control using Simpleaffy
45 experiments 80 datasets

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, Unigene 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 identification 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 address. 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 manipulation 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.

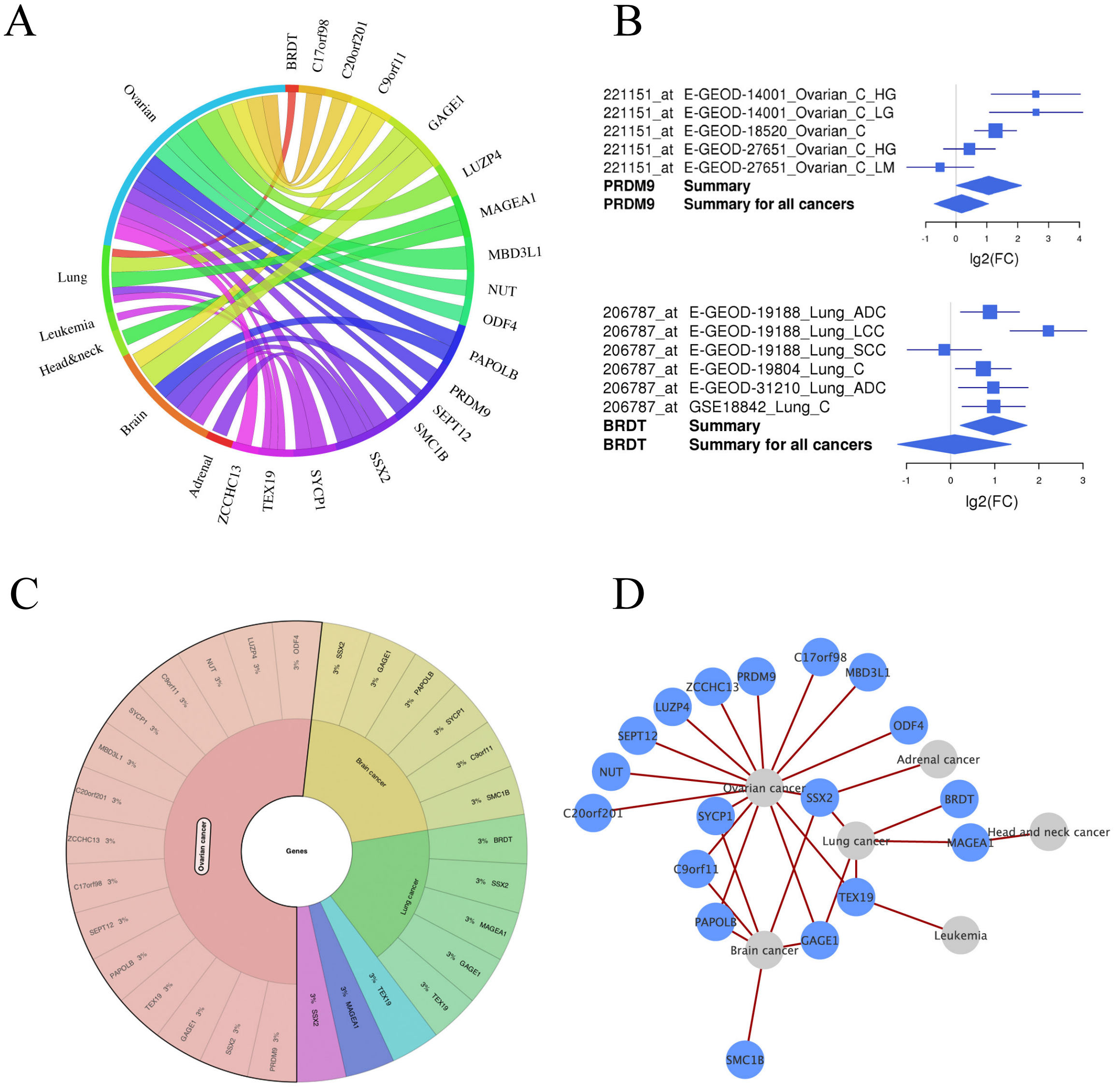


Figure 7.5: Visualisations of the results generated by CancerMA. Visualisations are a key feature of the CancerMA tool and help the user to grasp the available information. (A & C) CancerMA creates Circos [259] and Krona [260] plots to present the single and meta-analysis results in their entirety and 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 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 the both screening pipelines to identify novel CT genes successfully meets project objective 4.

meiCT genes are unique is 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.

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.

Appendix A

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.

Ensembl Gene ID	Protein	NLStradamus	Nucleo	Sequence Analysis	Uniprot	Summary
ENSG00000123165	ACTRT1				C	
ENSG00000137948	BRDT	N	N	N	N*	N
ENSG00000197651	C12orf12	N	N			N
ENSG00000184507	C15orf55	N	N	N	N/C*	N/C
ENSG00000214556	C17orf98					
ENSG00000188032	C19orf67			N		
ENSG00000178395	C1orf65	N	N	N		N
ENSG00000171695	C20orf201	N				
ENSG00000132631	C20orf79					
ENSG00000120160	C9orf11				C/ME	
ENSG00000187516	CXorf27		N	N		N
ENSG00000189037	DUSP21		N		N/C*	N/C
ENSG00000205186	FABP9				C	
ENSG00000181867	FTMT		N		M*	M
ENSG00000204671	IL31				S	
ENSG00000102021	LUZP4	N	N	N	N*	N
ENSG00000188408	MAGEB5					
ENSG00000170948	MBD3L1		N		N*	N
ENSG00000184650	ODF4			N	ME	
ENSG00000218823	PAPOLB				N*	N

ENSG00000196570	PFN3				N/C*	N/C
ENSG00000164256	PRDM9		N	N	N	N
ENSG00000244588	RAD21L1	N	N	N	N*	N
ENSG00000140623	SEPT12				C	C
ENSG00000077935	SMC1B			N	N	N
ENSG00000146857	STRA8				C	C
ENSG00000198765	SYCP1	N		N	N*	N
ENSG00000173809	TDRD12			N		
ENSG00000159648	TEPP				S	
ENSG00000182459	TEX19	N			N*	N
ENSG00000185264	TEX33		N			
ENSG00000212122	TSSK1B				C*	C
ENSG00000187969	ZCCHC13			N		

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`).

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 paralogue, 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).

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

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`).

Appendix C

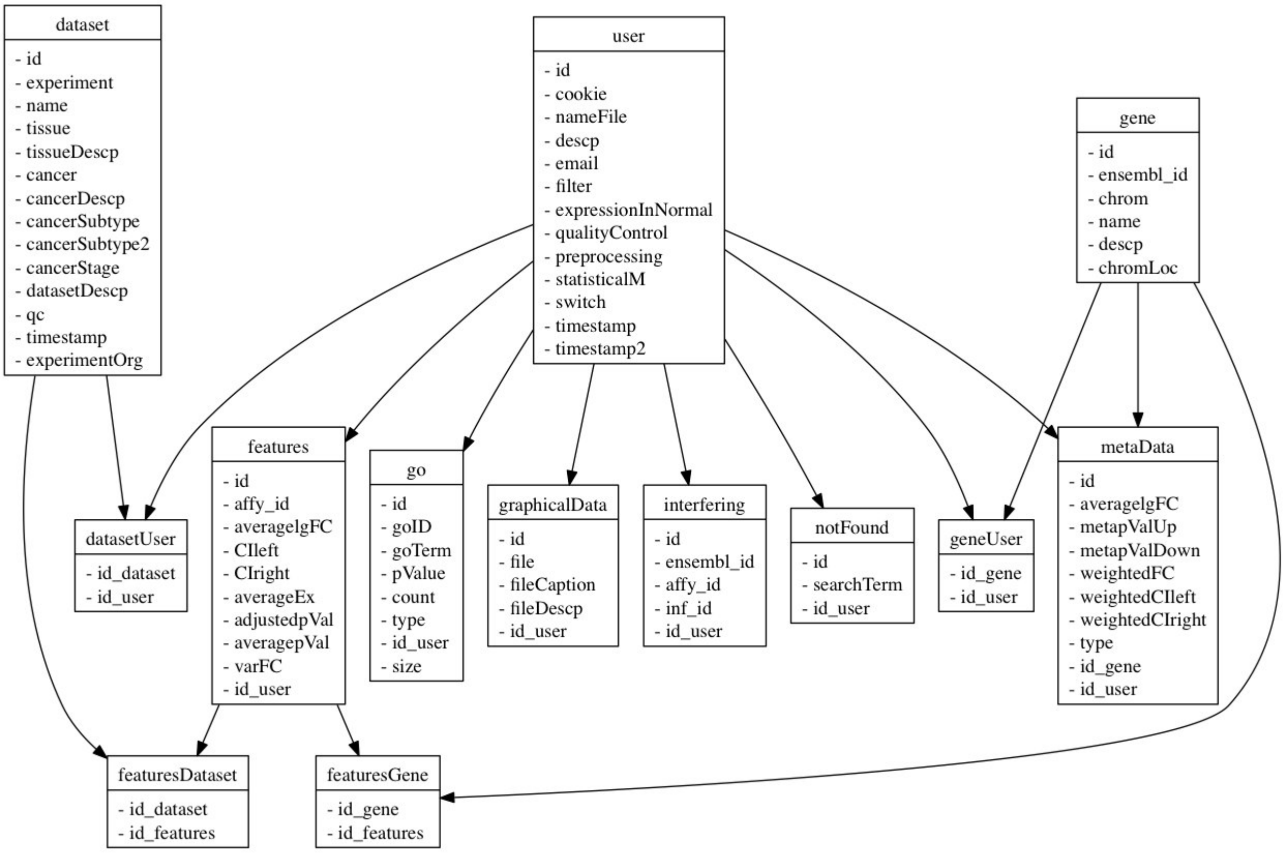


Figure C.1: Database schema of CancerMA. Visualisation of the database schema of the underlying database used by CancerMA.

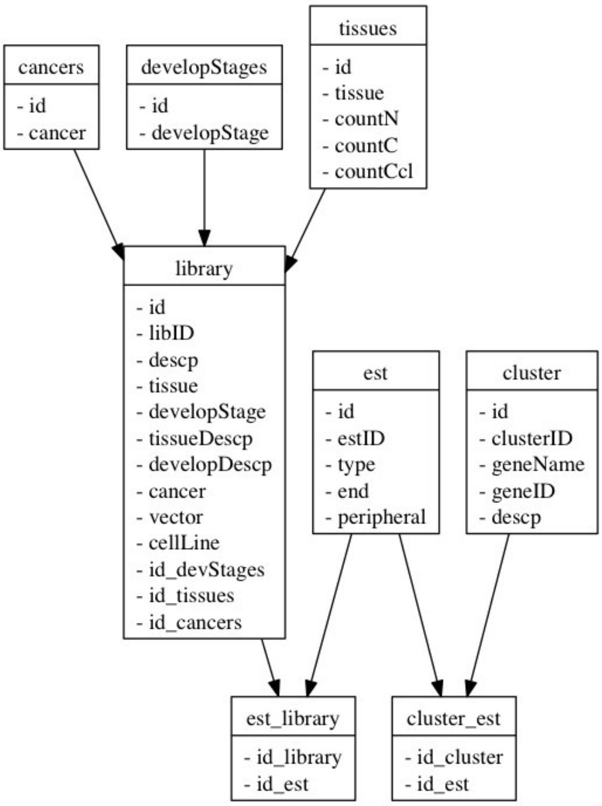


Figure C.2: Database schema of Unigene. Visualisation of the database schema of the local Unigene database.

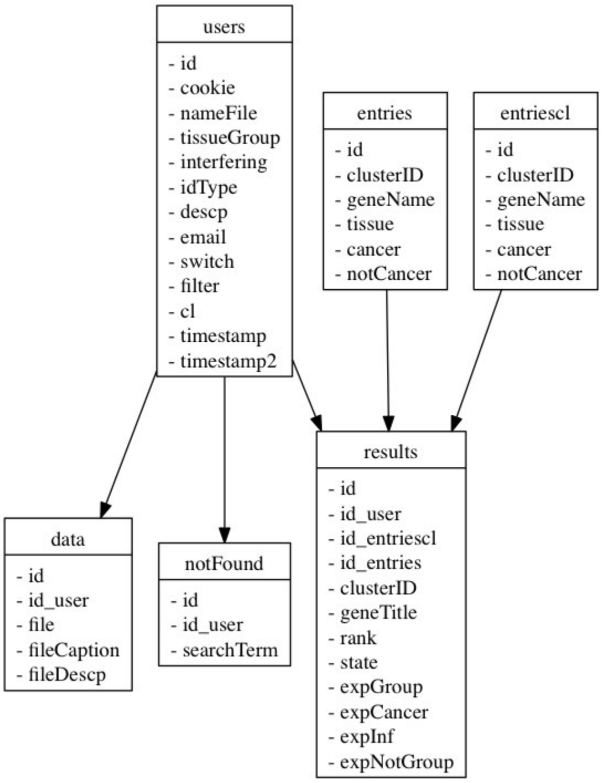


Figure C.3: Database schema of CancerEST. Visualisation of the database schema of the underlying database used by CancerEST.

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

- 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

- 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/`).

Input and Output

Input: Text file with data to be visualised.
Output: Image file.

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.
Output: Image file.

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

- 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/`). 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
- Spreadsheet::WriteExcel
- List::MoreUtils
- List::Util
- File::Remove
- XML::DOM
- XML::DOM::XPath
- Switch
- TryCatch
- Cwd
- Mail::Sender
- File::Copy
- POSIX
- Math::Round

- 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/`).

Circos Plots

Description

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

Input and Output

Input: Text file with data to be visualised.
Output: Image file.

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/`). The Unigene database was downloaded (UniGene Build #230) and read into a local MySQL database.

Citation

Pontius *et al.* [222].

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/`). 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
- List::MoreUtils
- LWP::UserAgent
- File::Remove
- Mail::Sender
- Switch
- Math::Round
- Data::Dumper
- JSON::XS
- Scalar::Util

- 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

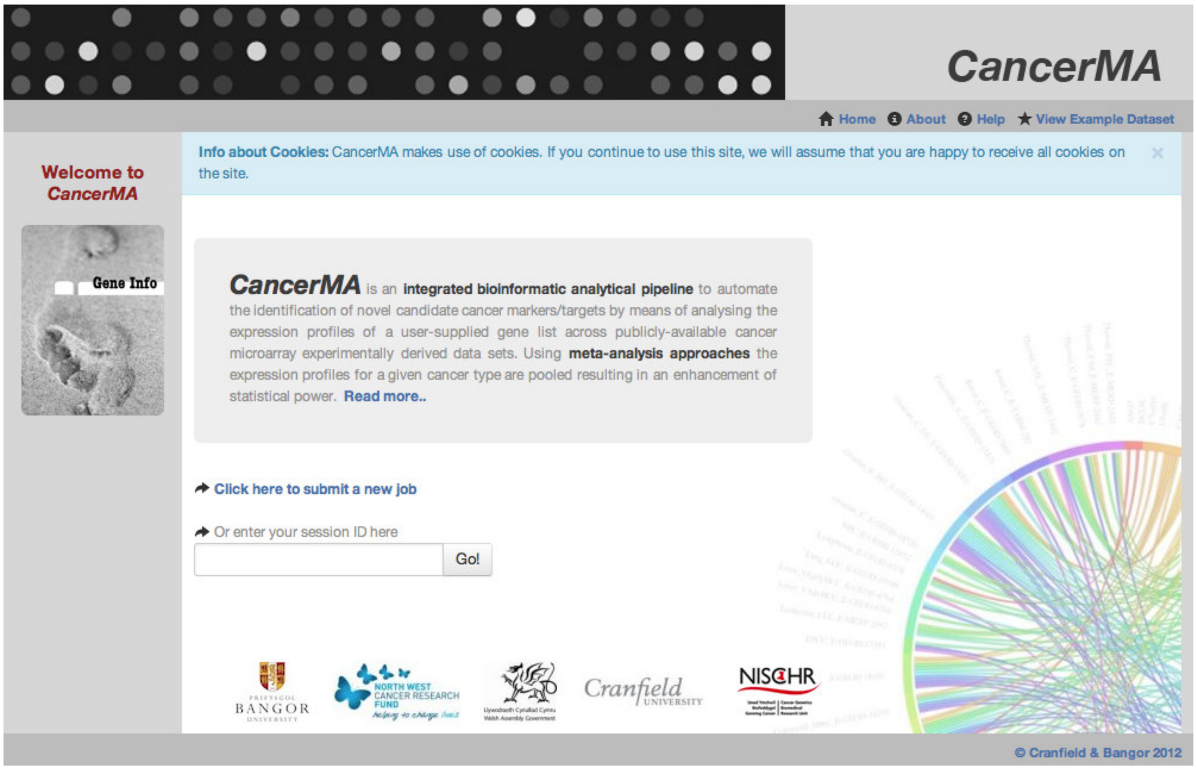
Javascript was employed to enhance the web pages. Javascript/jQuery libraries were used among others for displaying dynamic tables and for validation. This project made use of the Javascript libraries listed below.

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



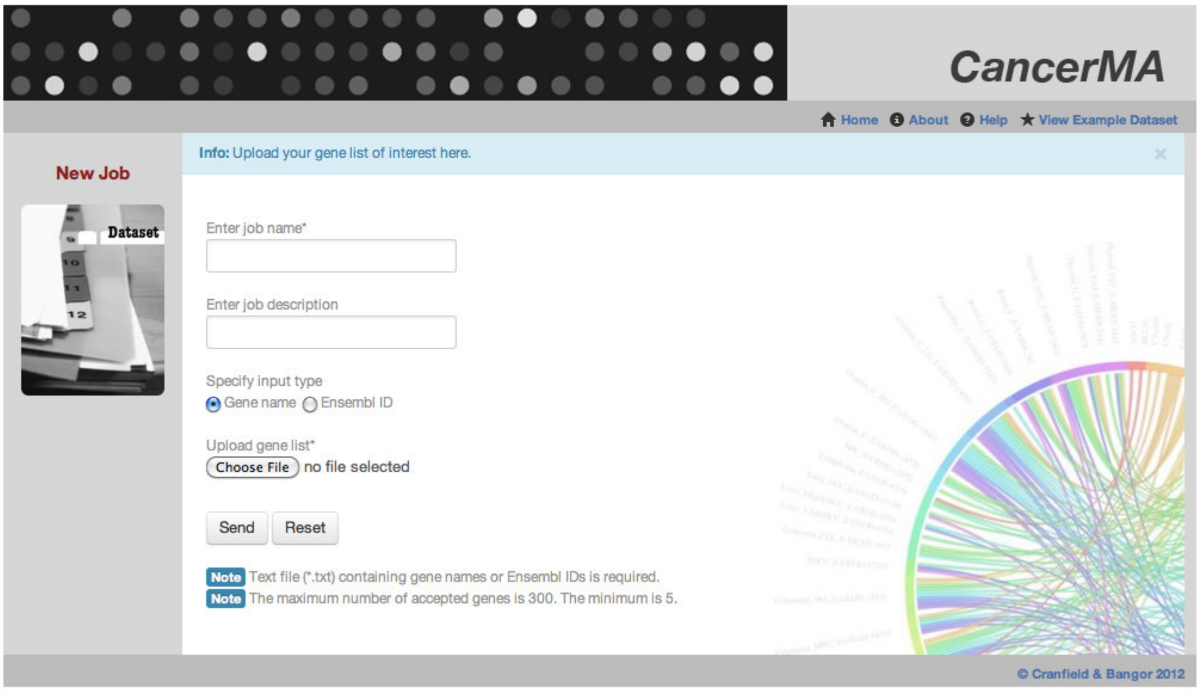
Info about Cookies: CancerMA makes use of cookies. If you continue to use this site, we will assume that you are happy to receive all cookies on the site.

CancerMA is an **integrated bioinformatic analytical pipeline** to automate the identification of novel candidate cancer markers/targets by means of analysing the expression profiles of a user-supplied gene list across publicly-available cancer microarray experimentally derived data sets. Using **meta-analysis approaches** the expression profiles for a given cancer type are pooled resulting in an enhancement of statistical power. Read more..

Click here to submit a new job

Or enter your session ID here

Figure E.1: Homepage of CancerMA. The screenshot shows the homepage of CancerMA (available at `http://www.cancerma.org.uk/`).



Info: Upload your gene list of interest here.

Enter job name*

Enter job description

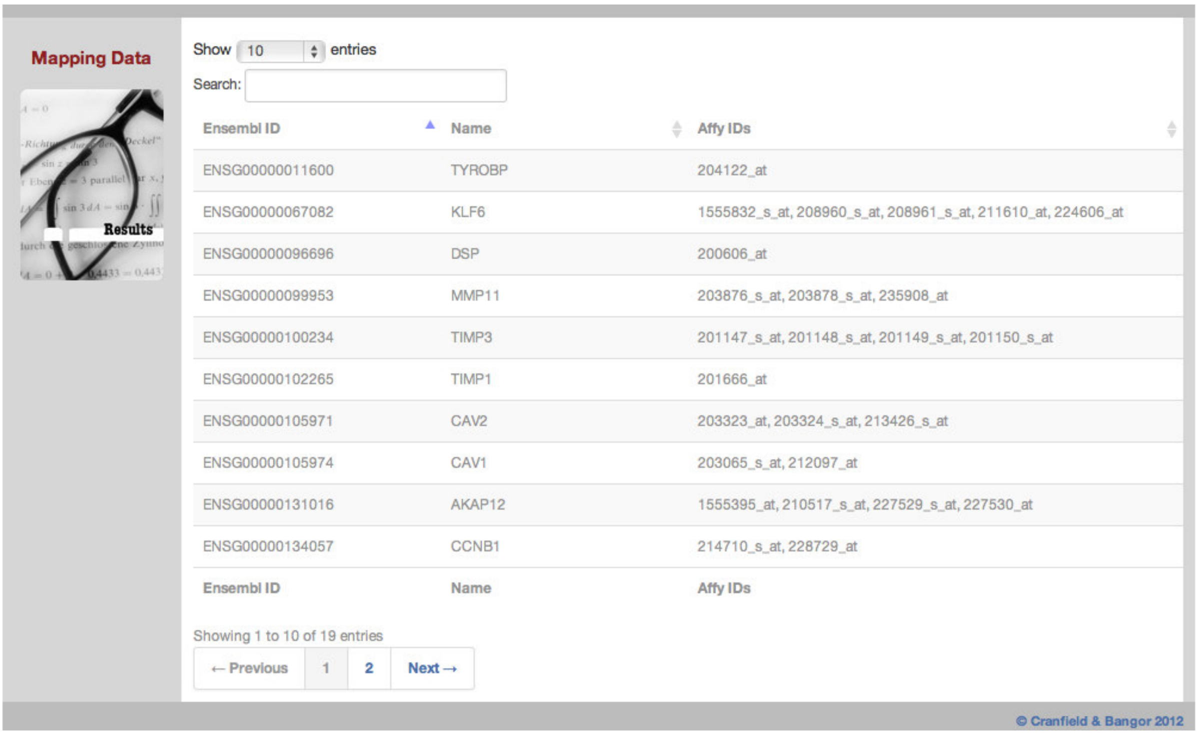
Specify input type
● Gene name ○ Ensembl ID

Upload gene list*
Choose File no file selected

Send Reset

Note Text file (*.txt) containing gene names or Ensembl IDs is required.
Note The maximum number of accepted genes is 300. The minimum is 5.

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.



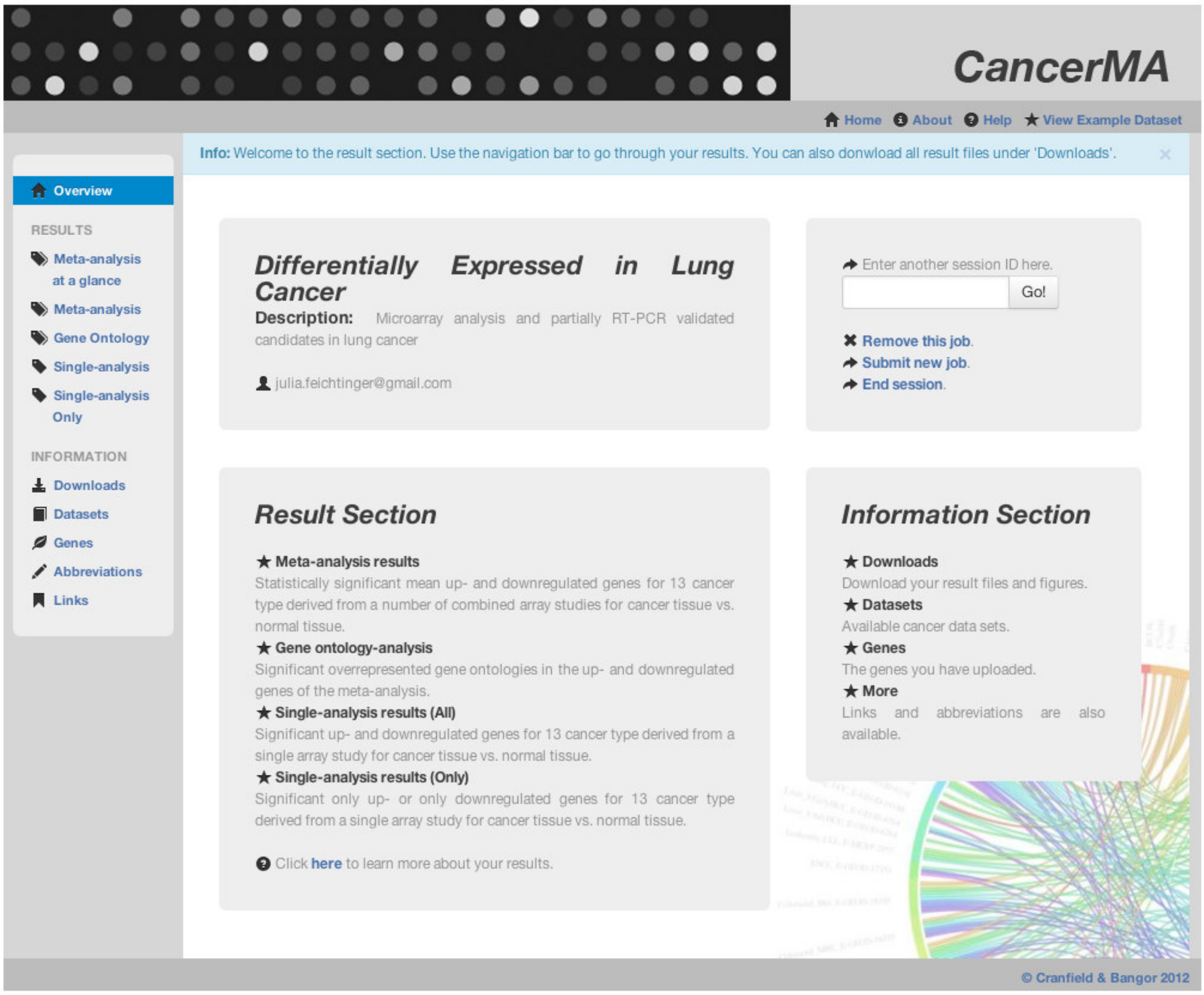
Show 10 entries

Search:

Ensembl ID	Name	Affy IDs
ENSG00000011600	TYROBP	204122_at
ENSG00000067082	KLF6	1555832_s_at, 208960_s_at, 208961_s_at, 211610_at, 224606_at
ENSG00000096696	DSP	200606_at
ENSG00000099953	MMP11	203876_s_at, 203878_s_at, 235908_at
ENSG00000100234	TIMP3	201147_s_at, 201148_s_at, 201149_s_at, 201150_s_at
ENSG00000102265	TIMP1	201666_at
ENSG00000105971	CAV2	203323_at, 203324_s_at, 213426_s_at
ENSG00000105974	CAV1	203065_s_at, 212097_at
ENSG00000131016	AKAP12	1555395_at, 210517_s_at, 227529_s_at, 227530_at
ENSG00000134057	CCNB1	214710_s_at, 228729_at
Ensembl ID	**Name**	**Affy IDs**

Showing 1 to 10 of 19 entries

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.



Info: Welcome to the result section. Use the navigation bar to go through your results. You can also donwload all result files under 'Downloads'.

Differentially Expressed in Lung Cancer

Description: Microarray analysis and partially RT-PCR validated candidates in lung cancer

julia.feichtinger@gmail.com

Enter another session ID here.

Remove this job.
Submit new job.
End session.

Result Section

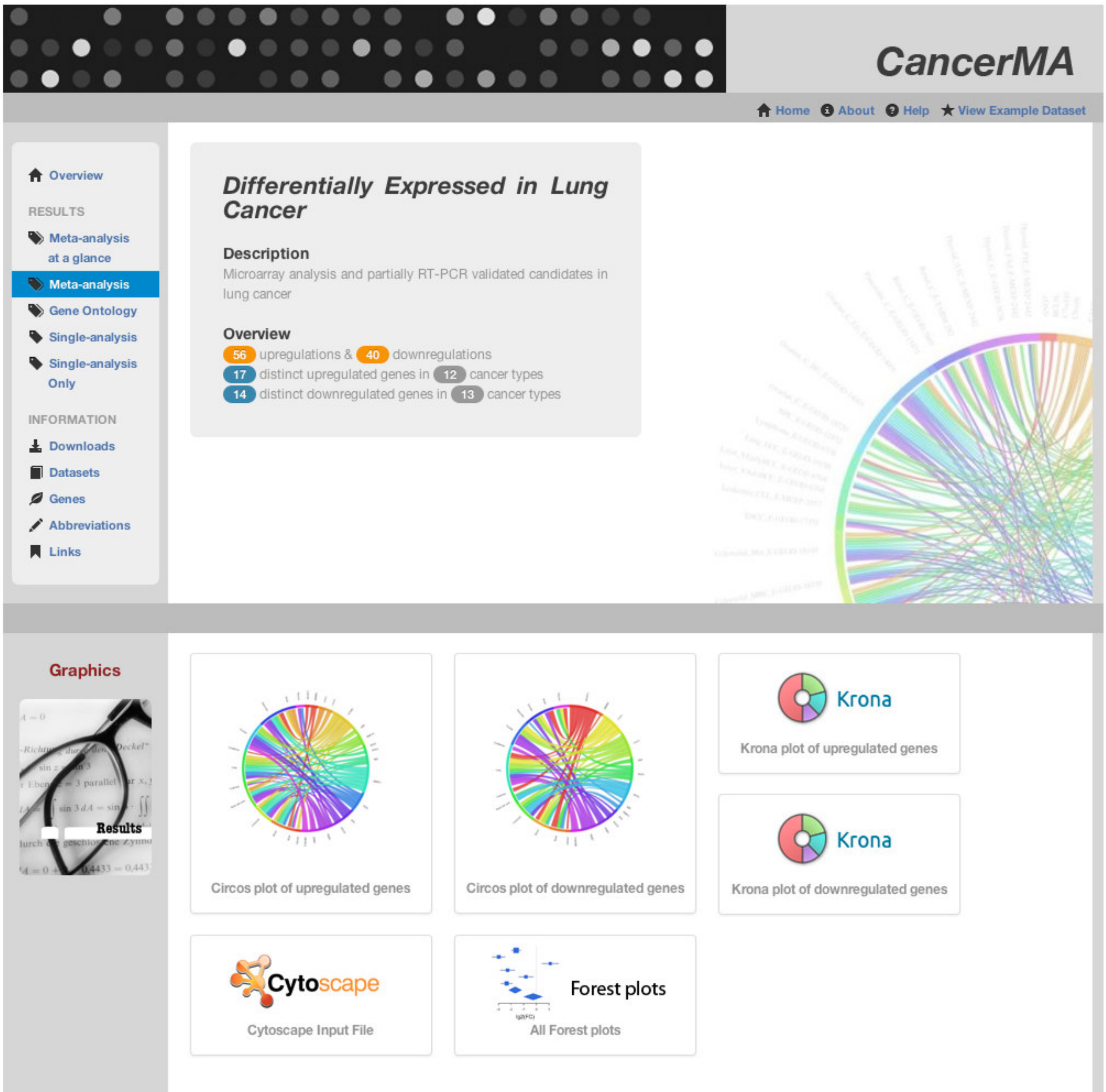
★ **Meta-analysis results**
Statistically significant mean up- and downregulated genes for 13 cancer type derived from a number of combined array studies for cancer tissue vs. normal tissue.
★ **Gene ontology-analysis**
Significant overrepresented gene ontologies in the up- and downregulated genes of the meta-analysis.
★ **Single-analysis results (All)**
Significant up- and downregulated genes for 13 cancer type derived from a single array study for cancer tissue vs. normal tissue.
★ **Single-analysis results (Only)**
Significant only up- or only downregulated genes for 13 cancer type derived from a single array study for cancer tissue vs. normal tissue.

Click **here** to learn more about your results.

Information Section

★ **Downloads**
Download your result files and figures.
★ **Datasets**
Available cancer data sets.
★ **Genes**
The genes you have uploaded.
★ **More**
Links and abbreviations are also available.

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.



Differentially Expressed in Lung Cancer

Description

Microarray analysis and partially RT-PCR validated candidates in lung cancer

Overview

56 upregulations & 40 downregulations

17 distinct upregulated genes in 12 cancer types

14 distinct downregulated genes in 13 cancer types

Circos plot of upregulated genes

Circos plot of downregulated genes

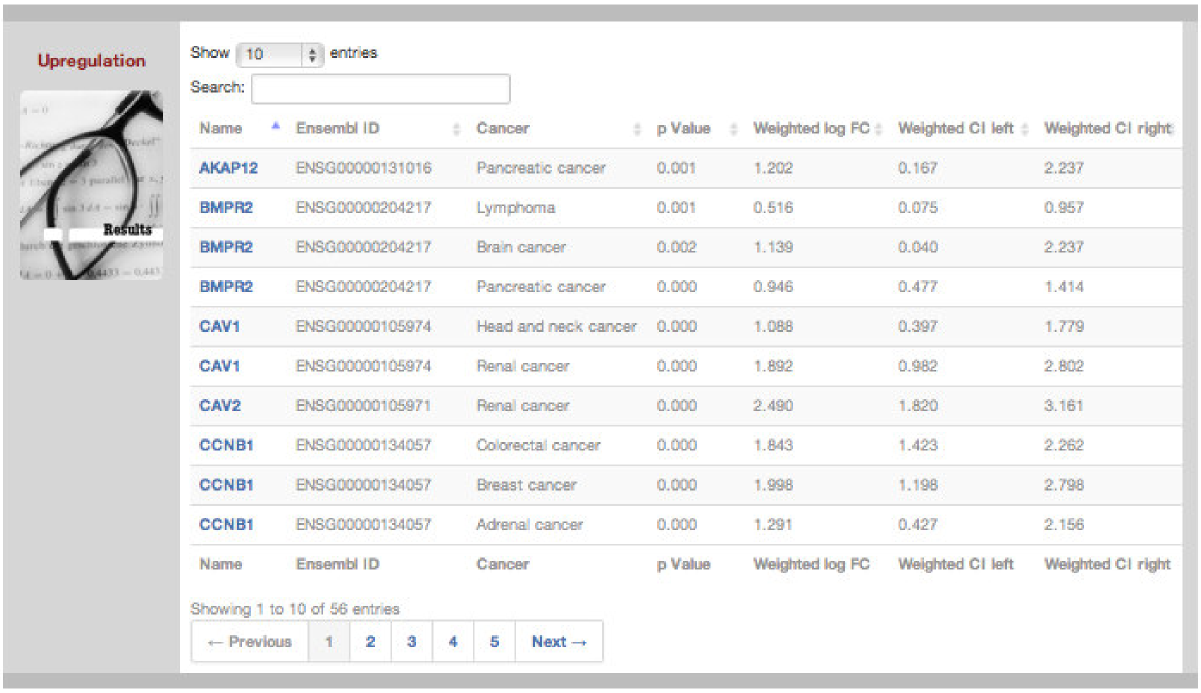
Krona plot of upregulated genes

Krona plot of downregulated genes

Cytoscape Input File

All Forest plots

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.



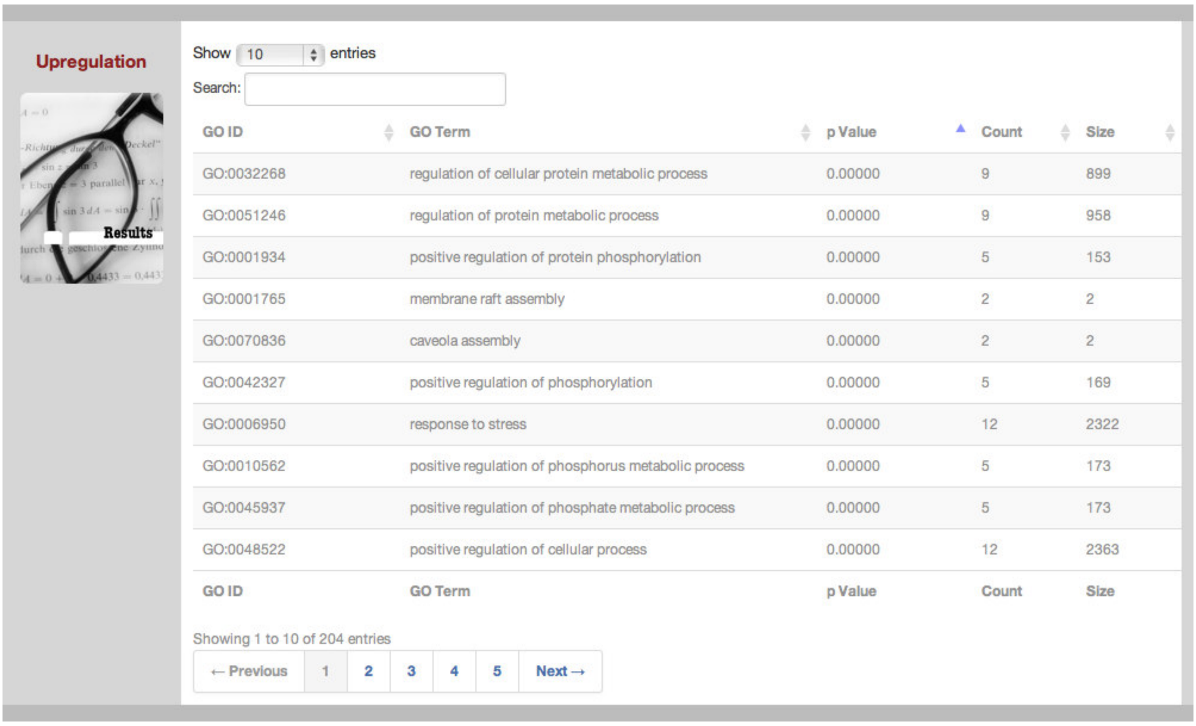
Show 10 entries

Search:

Name	Ensembl ID	Cancer	p Value	Weighted log FC	Weighted CI left	Weighted CI right
AKAP12	ENSG00000131016	Pancreatic cancer	0.001	1.202	0.167	2.237
BMPR2	ENSG00000204217	Lymphoma	0.001	0.516	0.075	0.957
BMPR2	ENSG00000204217	Brain cancer	0.002	1.139	0.040	2.237
BMPR2	ENSG00000204217	Pancreatic cancer	0.000	0.946	0.477	1.414
CAV1	ENSG00000105974	Head and neck cancer	0.000	1.088	0.397	1.779
CAV1	ENSG00000105974	Renal cancer	0.000	1.892	0.982	2.802
CAV2	ENSG00000105971	Renal cancer	0.000	2.490	1.820	3.161
CCNB1	ENSG00000134057	Colorectal cancer	0.000	1.843	1.423	2.262
CCNB1	ENSG00000134057	Breast cancer	0.000	1.998	1.198	2.798
CCNB1	ENSG00000134057	Adrenal cancer	0.000	1.291	0.427	2.156
Name	Ensembl ID	Cancer	p Value	Weighted log FC	Weighted CI left	Weighted CI right

Showing 1 to 10 of 56 entries

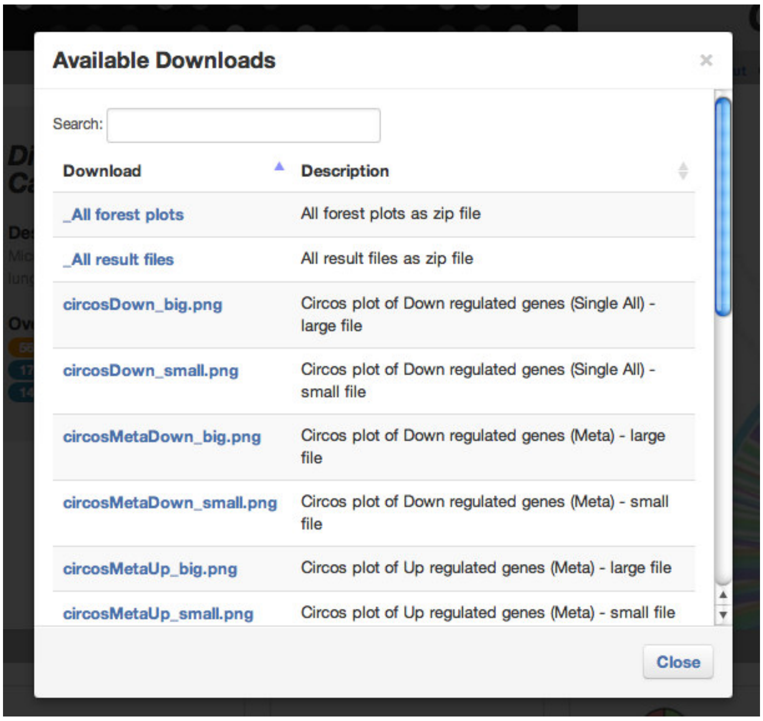
Figure E.6: Meta-analysis results of CancerMA (continued). The screenshot shows the meta-analysis results of CancerMA (continued). The meta-analysis results comprise tables with statistical values and visualisations for the meta-upregulated as well as for the meta-downregulated genes of interest.



GO ID	GO Term	p Value	Count	Size
GO:0032268	regulation of cellular protein metabolic process	0.00000	9	899
GO:0051246	regulation of protein metabolic process	0.00000	9	958
GO:0001934	positive regulation of protein phosphorylation	0.00000	5	153
GO:0001765	membrane raft assembly	0.00000	2	2
GO:0070836	caveola assembly	0.00000	2	2
GO:0042327	positive regulation of phosphorylation	0.00000	5	169
GO:0006950	response to stress	0.00000	12	2322
GO:0010562	positive regulation of phosphorus metabolic process	0.00000	5	173
GO:0045937	positive regulation of phosphate metabolic process	0.00000	5	173
GO:0048522	positive regulation of cellular process	0.00000	12	2363
GO ID	GO Term	p Value	Count	Size

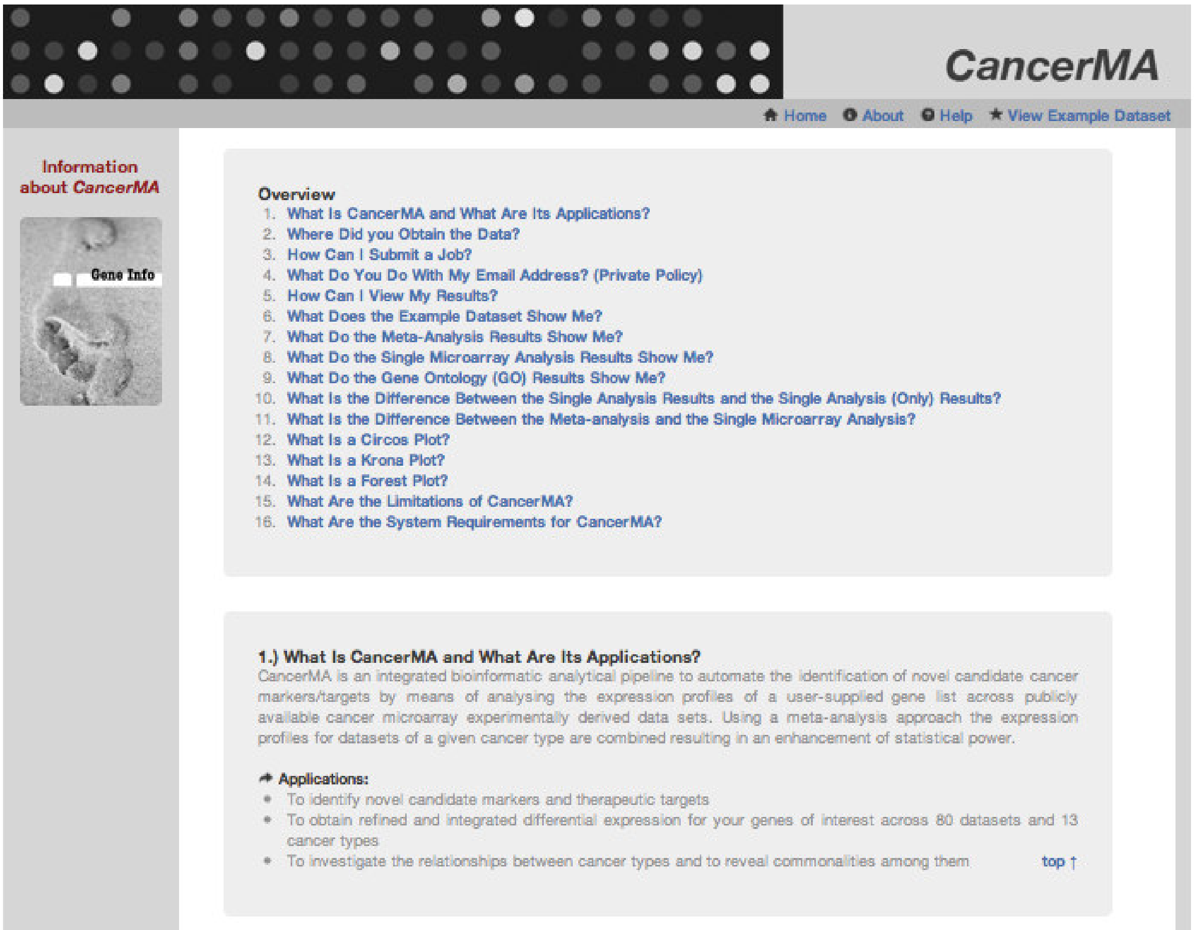
Showing 1 to 10 of 204 entries

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.



Download	Description
_All forest plots	All forest plots as zip file
_All result files	All result files as zip file
circosDown_big.png	Circos plot of Down regulated genes (Single All) - large file
circosDown_small.png	Circos plot of Down regulated genes (Single All) - small file
circosMetaDown_big.png	Circos plot of Down regulated genes (Meta) - large file
circosMetaDown_small.png	Circos plot of Down regulated genes (Meta) - small file
circosMetaUp_big.png	Circos plot of Up regulated genes (Meta) - large file
circosMetaUp_small.png	Circos plot of Up regulated genes (Meta) - small file

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.



Overview

1. What Is CancerMA and What Are Its Applications?
2. Where Did you Obtain the Data?
3. How Can I Submit a Job?
4. What Do You Do With My Email Address? (Private Policy)
5. How Can I View My Results?
6. What Does the Example Dataset Show Me?
7. What Do the Meta-Analysis Results Show Me?
8. What Do the Single Microarray Analysis Results Show Me?
9. What Do the Gene Ontology (GO) Results Show Me?
10. What Is the Difference Between the Single Analysis Results and the Single Analysis (Only) Results?
11. What Is the Difference Between the Meta-analysis and the Single Microarray Analysis?
12. What Is a Circos Plot?
13. What Is a Krona Plot?
14. What Is a Forest Plot?
15. What Are the Limitations of CancerMA?
16. What Are the System Requirements for CancerMA?

1.) What Is CancerMA and What Are Its Applications?

CancerMA is an integrated bioinformatic analytical pipeline to automate the identification of novel candidate cancer markers/targets by means of analysing the expression profiles of a user-supplied gene list across publicly available cancer microarray experimentally derived data sets. Using a meta-analysis approach the expression profiles for datasets of a given cancer type are combined resulting in an enhancement of statistical power.

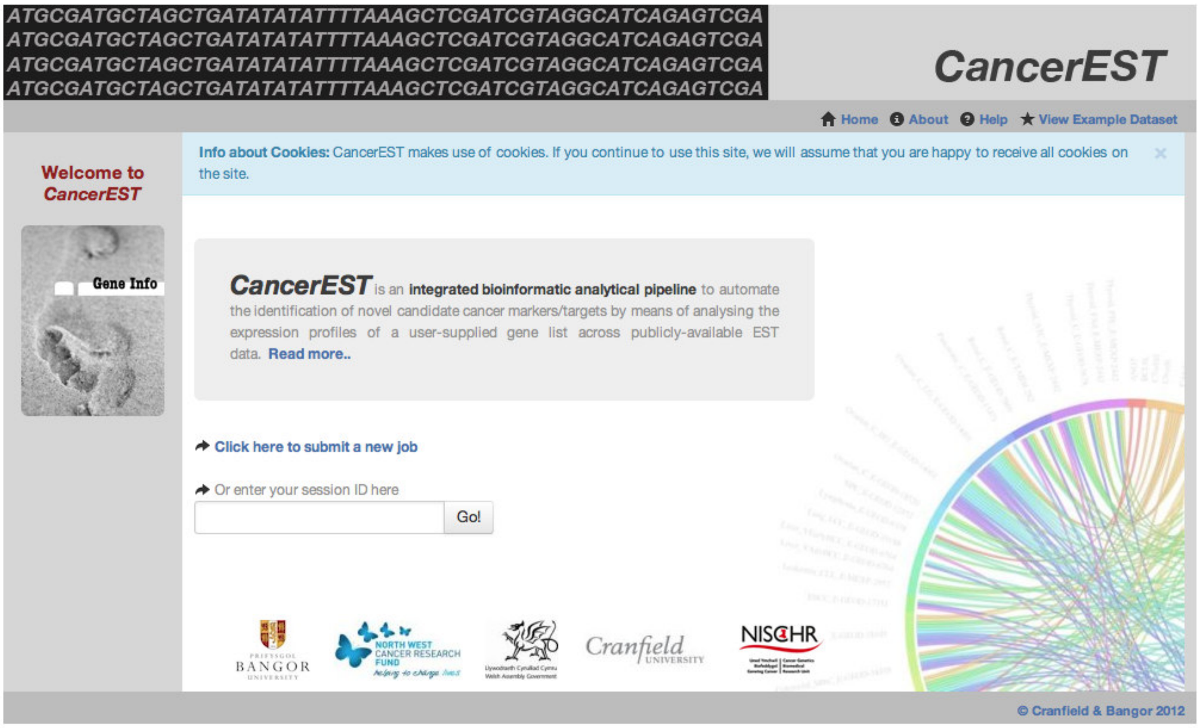
Applications:

- To identify novel candidate markers and therapeutic targets
- To obtain refined and integrated differential expression for your genes of interest across 80 datasets and 13 cancer types
- To investigate the relationships between cancer types and to reveal commonalities among them

top ↑

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`).

Screenshots of the CancerEST Web Interface



ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
CancerEST
Home
About
Help
View Example Dataset
Welcome to CancerEST
Gene Info
Go!
BANGOR UNIVERSITY
NORTH WEST CANCER RESEARCH FUND
Cranfield UNIVERSITY
NISCHR
© Cranfield & Bangor 2012

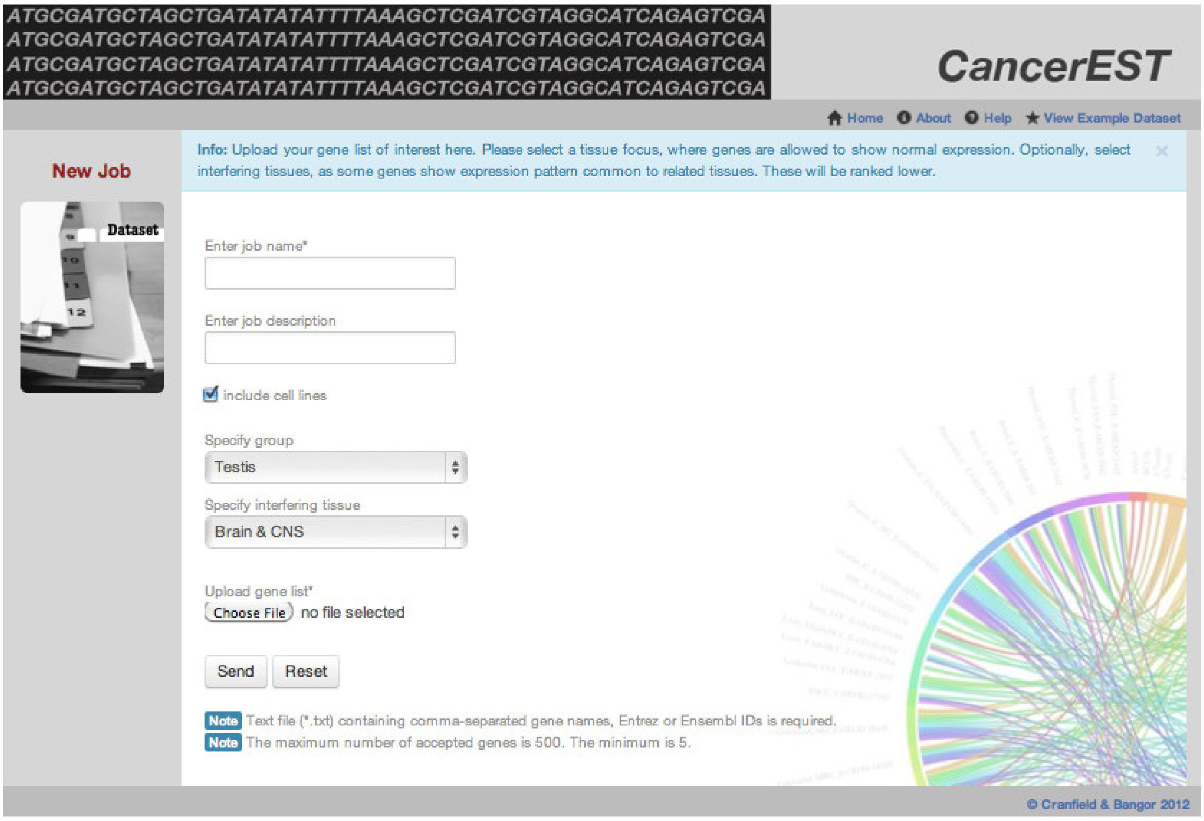
Info about Cookies: CancerEST makes use of cookies. If you continue to use this site, we will assume that you are happy to receive all cookies on the site.

CancerEST is an **integrated bioinformatic analytical pipeline** to automate the identification of novel candidate cancer markers/targets by means of analysing the expression profiles of a user-supplied gene list across publicly-available EST data. **Read more..**

Click here to submit a new job

Or enter your session ID here

Figure E.10: Homepage of CancerEST. The screenshot shows the homepage of CancerEST (available at `http://www.cancerest.org.uk/`).


ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
CancerEST
Home
About
Help
View Example Dataset
New Job
Dataset
Choose File
Send
Reset
© Cranfield & Bangor 2012

Info: Upload your gene list of interest here. Please select a tissue focus, where genes are allowed to show normal expression. Optionally, select interfering tissues, as some genes show expression pattern common to related tissues. These will be ranked lower.

Enter job name*

Enter job description

☑ include cell lines

Specify group: Testis

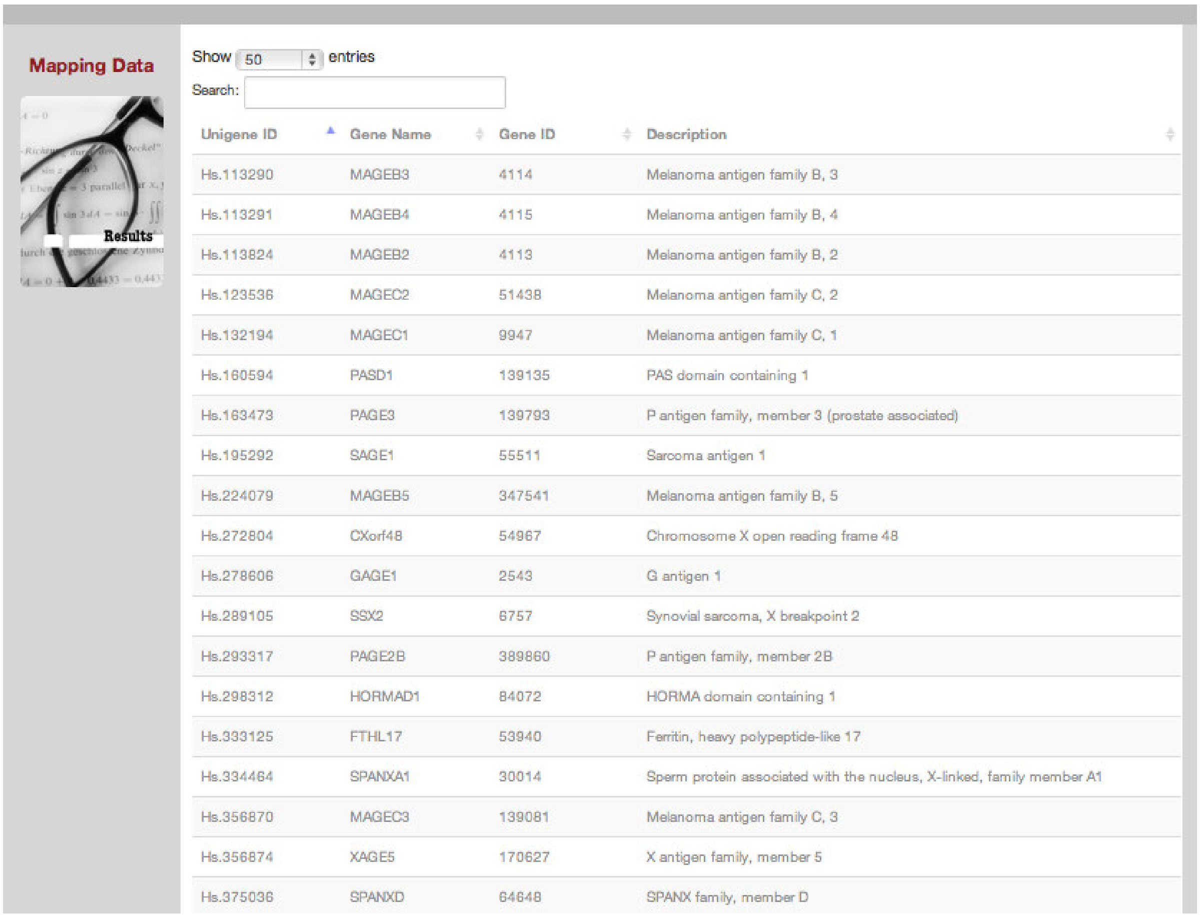
Specify interfering tissue: Brain & CNS

Upload gene list*: no file selected

Note Text file (*.txt) containing comma-separated gene names, Entrez or Ensembl IDs is required.

Note The maximum number of accepted genes is 500. The minimum is 5.

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.

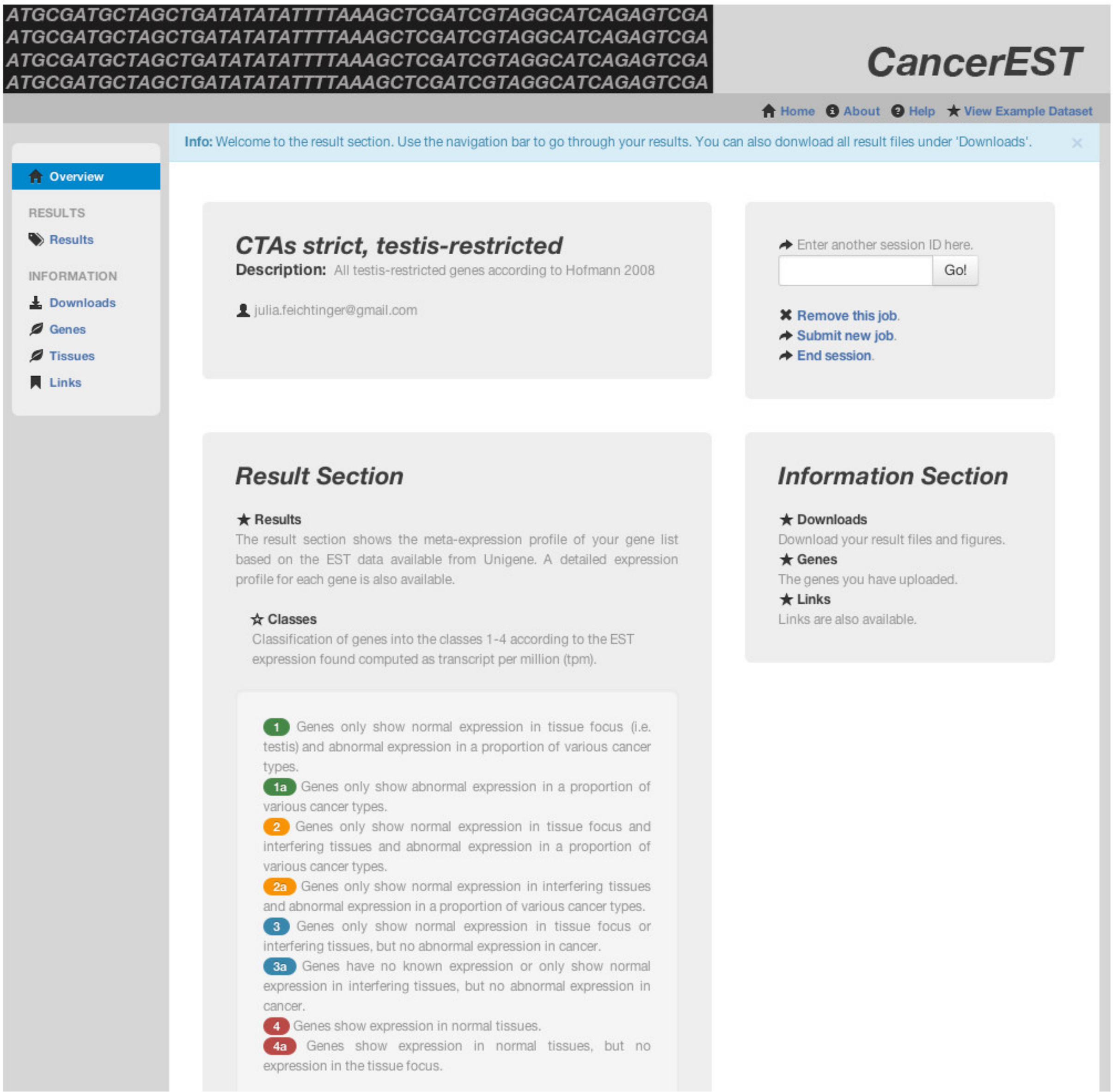


Show 50 entries

Search:

Unigene ID	Gene Name	Gene ID	Description
Hs.113290	MAGEB3	4114	Melanoma antigen family B, 3
Hs.113291	MAGEB4	4115	Melanoma antigen family B, 4
Hs.113824	MAGEB2	4113	Melanoma antigen family B, 2
Hs.123536	MAGEC2	51438	Melanoma antigen family C, 2
Hs.132194	MAGEC1	9947	Melanoma antigen family C, 1
Hs.160594	PASD1	139135	PAS domain containing 1
Hs.163473	PAGE3	139793	P antigen family, member 3 (prostate associated)
Hs.195292	SAGE1	55511	Sarcoma antigen 1
Hs.224079	MAGEB5	347541	Melanoma antigen family B, 5
Hs.272804	CXorf48	54967	Chromosome X open reading frame 48
Hs.278606	GAGE1	2543	G antigen 1
Hs.289105	SSX2	6757	Synovial sarcoma, X breakpoint 2
Hs.293317	PAGE2B	389860	P antigen family, member 2B
Hs.298312	HORMAD1	84072	HORMA domain containing 1
Hs.333125	FTHL17	53940	Ferritin, heavy polypeptide-like 17
Hs.334464	SPANXA1	30014	Sperm protein associated with the nucleus, X-linked, family member A1
Hs.356870	MAGEC3	139081	Melanoma antigen family C, 3
Hs.356874	XAGE5	170627	X antigen family, member 5
Hs.375036	SPANXD	64648	SPANX family, member D

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.


ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
CancerEST
Home
About
Help
View Example Dataset
Overview
RESULTS
Results
INFORMATION
Downloads
Genes
Tissues
Links
Go!
Remove this job.
Submit new job.
End session.

Info: Welcome to the result section. Use the navigation bar to go through your results. You can also donwload all result files under 'Downloads'.

CTAs strict, testis-restricted

Description: All testis-restricted genes according to Hofmann 2008

julia.feichtinger@gmail.com

Enter another session ID here.

Result Section

★ **Results**

The result section shows the meta-expression profile of your gene list based on the EST data available from Unigene. A detailed expression profile for each gene is also available.

☆ **Classes**

Classification of genes into the classes 1-4 according to the EST expression found computed as transcript per million (tpm).

1 Genes only show normal expression in tissue focus (i.e. testis) and abnormal expression in a proportion of various cancer types.

1a Genes only show abnormal expression in a proportion of various cancer types.

2 Genes only show normal expression in tissue focus and interfering tissues and abnormal expression in a proportion of various cancer types.

2a Genes only show normal expression in interfering tissues and abnormal expression in a proportion of various cancer types.

3 Genes only show normal expression in tissue focus or interfering tissues, but no abnormal expression in cancer.

3a Genes have no known expression or only show normal expression in interfering tissues, but no abnormal expression in cancer.

4 Genes show expression in normal tissues.

4a Genes show expression in normal tissues, but no expression in the tissue focus.

Information Section

★ **Downloads**

Download your result files and figures.

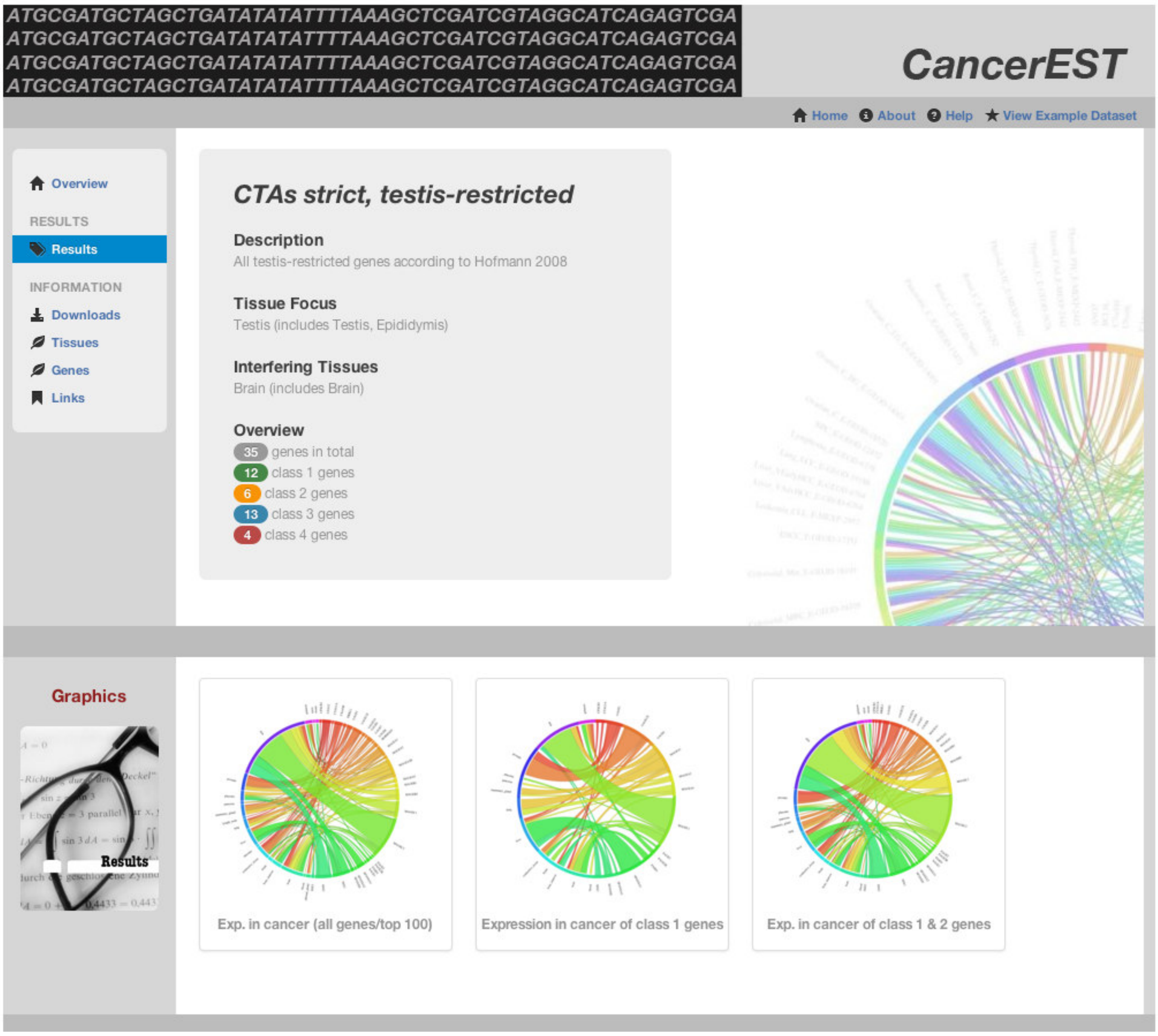
★ **Genes**

The genes you have uploaded.

★ **Links**

Links are also available.

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.


ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
CancerEST
Home
About
Help
View Example Dataset
Overview
RESULTS
Results
INFORMATION
Downloads
Tissues
Genes
Links
Graphics
Results
Exp. in cancer (all genes/top 100)
Expression in cancer of class 1 genes
Exp. in cancer of class 1 & 2 genes

CTAs strict, testis-restricted

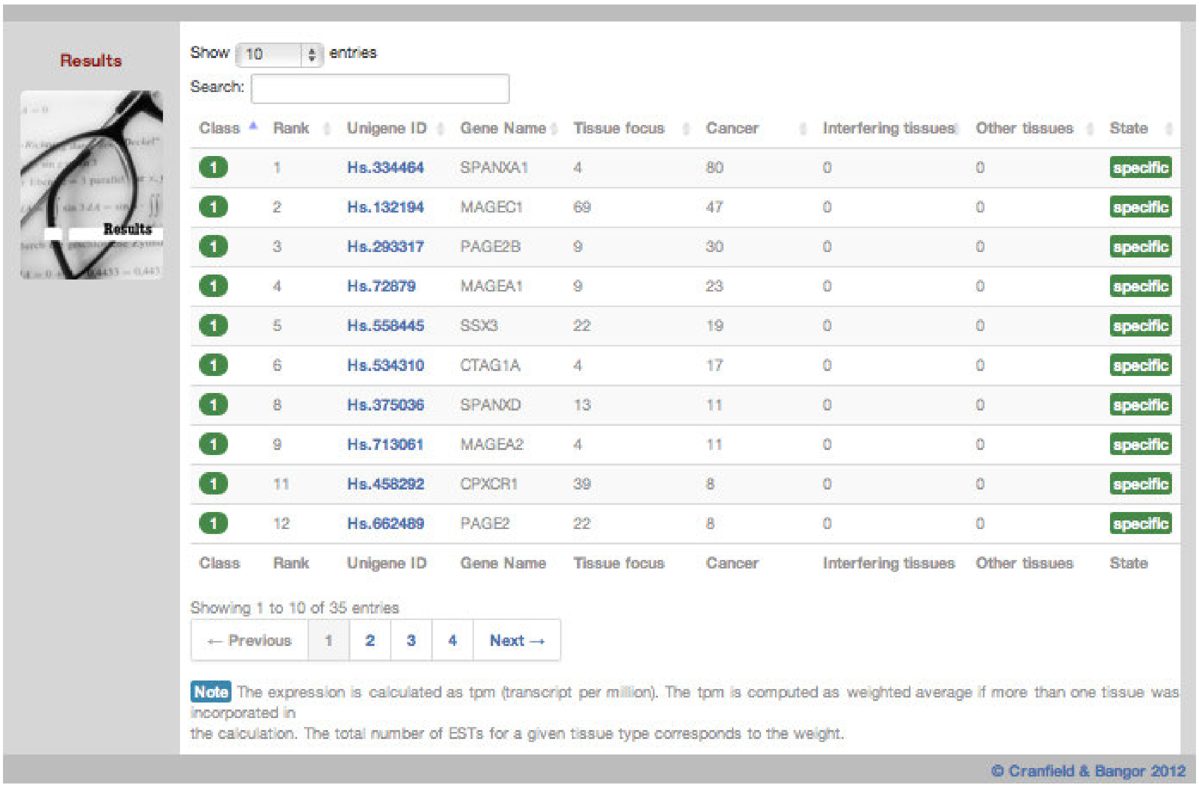
Description
All testis-restricted genes according to Hofmann 2008

Tissue Focus
Testis (includes Testis, Epididymis)

Interfering Tissues
Brain (includes Brain)

Overview
35 genes in total
12 class 1 genes
6 class 2 genes
13 class 3 genes
4 class 4 genes

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.



Show 10 entries

Search:

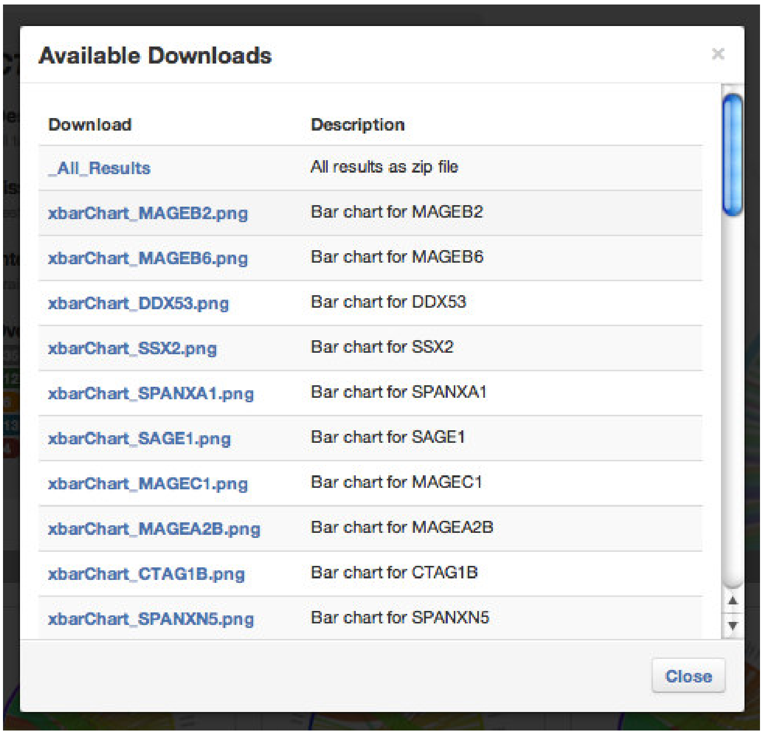
Class	Rank	Unigene ID	Gene Name	Tissue focus	Cancer	Interfering tissues	Other tissues	State
1	1	Hs.334464	SPANXA1	4	80	0	0	specific
1	2	Hs.132194	MAGEC1	69	47	0	0	specific
1	3	Hs.293317	PAGE2B	9	30	0	0	specific
1	4	Hs.72879	MAGEA1	9	23	0	0	specific
1	5	Hs.558445	SSX3	22	19	0	0	specific
1	6	Hs.534310	CTAG1A	4	17	0	0	specific
1	8	Hs.375036	SPANXD	13	11	0	0	specific
1	9	Hs.713061	MAGEA2	4	11	0	0	specific
1	11	Hs.458292	CPXCR1	39	8	0	0	specific
1	12	Hs.662489	PAGE2	22	8	0	0	specific
Class	Rank	Unigene ID	Gene Name	Tissue focus	Cancer	Interfering tissues	Other tissues	State

Showing 1 to 10 of 35 entries

Note The expression is calculated as tpm (transcript per million). The tpm is computed as weighted average if more than one tissue was incorporated in
the calculation. The total number of ESTs for a given tissue type corresponds to the weight.

© Cranfield & Bangor 2012

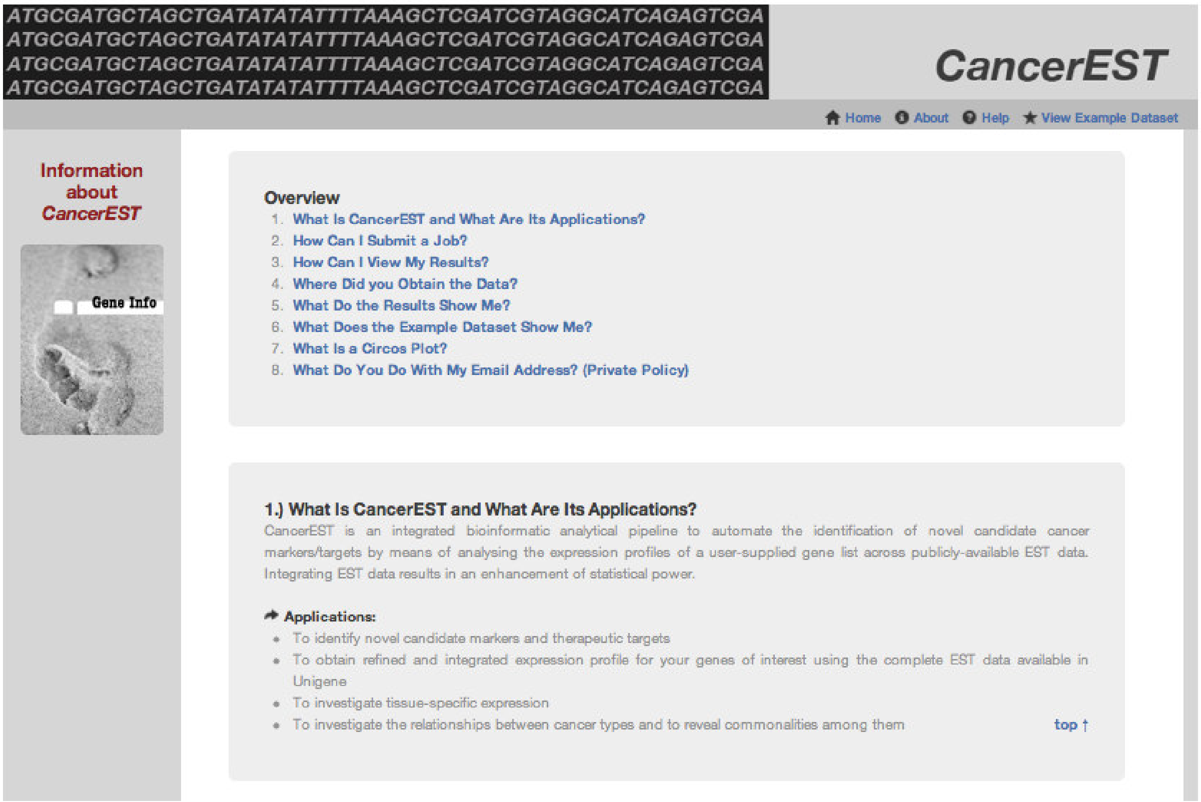
Figure E.18: Meta-analysis results of CancerEST (continued). The screenshot shows the meta-analysis results of CancerEST (continued). The result section includes the EST meta-analysis results comprising of a ranked list of genes according to (i) their cancer marker potential; or to (ii) their tissue-specificity.



Available Downloads

Download	Description
_All_Results	All results as zip file
xbarChart_MAGEB2.png	Bar chart for MAGEB2
xbarChart_MAGEB6.png	Bar chart for MAGEB6
xbarChart_DDX53.png	Bar chart for DDX53
xbarChart_SSX2.png	Bar chart for SSX2
xbarChart_SPANXA1.png	Bar chart for SPANXA1
xbarChart_SAGE1.png	Bar chart for SAGE1
xbarChart_MAGEC1.png	Bar chart for MAGEC1
xbarChart_MAGEA2B.png	Bar chart for MAGEA2B
xbarChart_CTAG1B.png	Bar chart for CTAG1B
xbarChart_SPANXN5.png	Bar chart for SPANXN5

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.



ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
ATGCGATGCTAGCTGATATATATTTTAAAGCTCGATCGTAGGCATCAGAGTCGA
CancerEST
Home
About
Help
View Example Dataset
Information about CancerEST
Gene Info

Overview

1. What Is CancerEST and What Are Its Applications?
2. How Can I Submit a Job?
3. How Can I View My Results?
4. Where Did you Obtain the Data?
5. What Do the Results Show Me?
6. What Does the Example Dataset Show Me?
7. What Is a Circos Plot?
8. What Do You Do With My Email Address? (Private Policy)

1.) What Is CancerEST and What Are Its Applications?

CancerEST is an integrated bioinformatic analytical pipeline to automate the identification of novel candidate cancer markers/targets by means of analysing the expression profiles of a user-supplied gene list across publicly-available EST data. Integrating EST data results in an enhancement of statistical power.

Applications:

- To identify novel candidate markers and therapeutic targets
- To obtain refined and integrated expression profile for your genes of interest using the complete EST data available in Unigene
- To investigate tissue-specific expression
- To investigate the relationships between cancer types and to reveal commonalities among them

top ↑

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)
limma_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))
```

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("l","l","l"), 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_delim_collapse
-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

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D: **Global cancer statistics.** *Ca-Cancer J. Clin.* 2011, **61**:69–90.

2. **Comprehensive Cancer Information - National Cancer Institute** [`http://cancer.gov/`].

3. Anand P, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, Sung B, Aggarwal BB: **Cancer is a Preventable Disease that Requires Major Lifestyle Changes**. *Pharm. Res.* 2008, **25**:2097–2116.

4. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis.** *Cell* 1990, **61**:759–67.

5. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome.** *Nature* 2009, **458**:719–24.

6. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat. Med.* 2004, **10**:789–99.

7. Sherr C: **Principles of Tumor Suppression**. *Cell* 2004, **116**:235–246.

8. Croce CM: **Oncogenes and cancer.** *N. Engl. J. Med.* 2008, **358**:502–11.

9. Bos JL: **ras Oncogenes in Human Cancer: A Review**. *Cancer Res.* 1989, **49**:4682–4689.

10. Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D: **RAS oncogenes: weaving a tumorigenic web.** *Nat. Rev. Cancer* 2011, **11**:761–74.

11. Vogelstein B, Lane D, Levine AJ: **Surfing the p53 network.** *Nature* 2000, **408**:307–10.

12. Vazquez A, Bond EE, Levine AJ, Bond GL: **The genetics of the p53 pathway, apoptosis and cancer therapy.** *Nat. Rev. Drug Discov.* 2008, **7**:979–87.

13. Friedberg EC: **DNA damage and repair.** *Nature* 2003, **421**:436–40.

14. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144**:646–74.

15. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57–70.

16. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin M-L, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, et al.: **Massive genomic rearrangement acquired in a single catastrophic event during cancer development.** *Cell* 2011, **144**:27–40.

17. Whiteside TL: **The tumor microenvironment and its role in promoting tumor growth.** *Oncogene* 2008, **27**:5904–12.

18. Kopper L, Hajdú M: **Tumor stem cells.** *Pathol. Oncol. Res.* 2004, **10**:69–73.

19. Gupta PB, Chaffer CL, Weinberg RA: **Cancer stem cells: mirage or reality?** *Nat. Med.* 2009, **15**:1010–2.

20. Valent P, Bonnet D, De Maria R, Lapidot T, Copland M, Melo J V, Chomienne C, Ishikawa F, Schuringa JJ, Stassi G, Huntly B, Herrmann H, Soulier J, Roesch A, Schuurhuis GJ, Wöhrer S, Arock M, Zuber J, Cerny-Reiterer S, Johnsen HE, Andreeff M, Eaves C: **Cancer stem cell definitions and terminology: the devil is in the details.** *Nat. Rev. Cancer* 2012, **12**:767–75.

21. Nguyen L V, Vanner R, Dirks P, Eaves CJ: **Cancer stem cells: an evolving concept.** *Nat. Rev. Cancer* 2012, **12**:133–43.

22. Bjerkvig R, Tysnes BB, Aboody KS, Najbauer J, Terzis AJA: **Opinion: the origin of the cancer stem cell: current controversies and new insights.** *Nat. Rev. Cancer* 2005, **5**:899–904.

23. Pardoll D: **Does the immune system see tumors as foreign or self?** *Annu. Rev. Immunol.* 2003, **21**:807–39.

24. Aly HAA: **Cancer therapy and vaccination.** *J. Immunol. Methods* 2012, **382**:1–23.

25. Franks HA, Wang Q, Patel PM: **New anticancer immunotherapies.** *Anticancer Res.* 2012, **32**:2439–53.

26. Mellman I, Coukos G, Dranoff G: **Cancer immunotherapy comes of age.** *Nature* 2011, **480**:480–9.

27. Topalian SL, Weiner GJ, Pardoll DM: **Cancer immunotherapy comes of age.** *J. Clin. Oncol.* 2011, **29**:4828–36.

28. Van der Bruggen P, Traversari C, Chomez P, Lurquin C, De Plaen E, Van den Eynde B, Knuth A, Boon T: **A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma**. *Science* 1991, **254**:1643–1647.

29. Pejawar-Gaddy S, Finn OJ: **Cancer vaccines: accomplishments and challenges.** *Crit. Rev. Oncol. Hematol.* 2008, **67**:93–102.

30. Buonaguro L, Petrizzo A, Tornesello ML, Buonaguro FM: **Translating tumor antigens into cancer vaccines.** *Clin. Vaccine Immunol.* 2011, **18**:23–34.

31. Simpson AJG, Caballero OL, Jungbluth A, Chen Y-T, Old LJ: **Cancer/testis antigens, gametogenesis and cancer.** *Nat. Rev. Cancer* 2005, **5**:615–25.

32. Fratta E, Coral S, Covre A, Parisi G, Colizzi F, Danielli R, Nicolay HJM, Sigalotti L, Maio M: **The biology of cancer testis antigens: putative function, regulation and therapeutic potential.** *Mol. Oncol.* 2011, **5**:164–82.

33. Houghton AN, Eisinger M, Albino AP, Cairncross JG, Old LJ: **Surface antigens of melanocytes and melanomas. Markers of melanocyte differentiation and melanoma subsets.** *J. Exp. Med.* 1982, **156**:1755–66.

34. Scott AM, Wolchok JD, Old LJ: **Antibody therapy of cancer.** *Nat. Rev. Cancer* 2012, **12**:278–87.

35. Zhou G, Levitsky H: **Towards curative cancer immunotherapy: overcoming posttherapy tumor escape.** *Clin. Dev. Immunol.* 2012, **2012**:124187.

36. Weiner GJ: **Rituximab: mechanism of action.** *Semin. Hematol.* 2010, **47**:115–23.

37. Hudis CA: **Trastuzumab–mechanism of action and use in clinical practice.** *N. Engl. J. Med.* 2007, **357**:39–51.

38. Restifo NP, Dudley ME, Rosenberg SA: **Adoptive immunotherapy for cancer: harnessing the T cell response.** *Nat. Rev. Immunol.* 2012, **12**:269–81.

39. Hunder NN, Wallen H, Cao J, Hendricks DW, Reilly JZ, Rodmyre R, Jungbluth A, Gnjatic S, Thompson JA, Yee C: **Treatment of metastatic melanoma with autologous CD4+ T cells against NY-ESO-1.** *N. Engl. J. Med.* 2008, **358**:2698–2703.

40. Kantoff PW, Higano CS, Shore ND, Berger ER, Small EJ, Penson DF, Redfern CH, Ferrari AC, Dreicer R, Sims RB, Xu Y, Frohlich MW, Schellhammer PF: **Sipuleucel-T immunotherapy for castration-resistant prostate cancer.** *N. Engl. J. Med.* 2010, **363**:411–22.

41. Lesterhuis WJ, Haanen JBAG, Punt CJA: **Cancer immunotherapy–revisited.** *Nat. Rev. Drug Discov.* 2011, **10**:591–600.

42. Handy B: **The Clinical Utility of Tumor Markers**. *Lab. Med.* 2009, **40**:99–103.

43. Hicks DG, Kulkarni S: **Trastuzumab as adjuvant therapy for early breast cancer: the importance of accurate human epidermal growth factor receptor 2 testing.** *Arch. Pathol. Lab. Med.* 2008, **132**:1008–15.

44. Wolff AC, Hammond MEH, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, McShane LM, Paik S, Pegram MD, Perez EA, Press MF, Rhodes A, Sturgeon C, Taube SE, Tubbs R, Vance GH, Van de Vijver M, Wheeler TM, Hayes DF: **American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer.** *Arch. Pathol. Lab. Med.* 2007, **131**:18–43.

45. Scanlan MJ, Simpson AJG, Old LJ: **The cancer/testis genes: review, standardization, and commentary.** *Cancer Immun.* 2004, **4**:1.

46. Caballero OL, Chen Y-T: **Cancer/testis (CT) antigens: potential targets for immunotherapy.** *Cancer Sci.* 2009, **100**:2014–21.

47. Cheng Y-H, Wong EW, Cheng CY: **Cancer/testis (CT) antigens, carcinogenesis and spermatogenesis.** *Spermatogenesis* 2011, **1**:209–220.

48. Lim SH, Zhang Y, Zhang J: **Cancer-testis antigens: the current status on antigen regulation and potential clinical use.** *Am. J. Blood Res.* 2012, **2**:29–35.

49. Mruk DD, Cheng CY: **Tight junctions in the testis: new perspectives.** *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 2010, **365**:1621–1635.

50. Fijak M, Meinhardt A: **The testis in immune privilege.** *Immunol. Rev.* 2006, **213**:1–121.

51. Sang M, Lian Y, Zhou X, Shan B: **MAGE-A family: Attractive targets for cancer immunotherapy**. *Vaccine* 2011, **29**:8496–8500.

52. De Backer O, Arden KC, Boretti M, Vantomme V, De Smet C, Czekay S, Viars CS, De Plaen E, Brasseur F, Chomez P, Van Den Eynde B, Boon T, Van Der Bruggen P: **Characterization of the GAGE genes that are expressed in various human cancers and in normal testis.** *Cancer Res.* 1999, **59**:3157–3165.

53. Boël P, Wildmann C, Sensi ML, Brasseur R, Renauld JC, Coulie P, Boon T, Van der Bruggen P: **BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes.** *Immunity* 1995, **2**:167–75.

54. Chen YT, Scanlan MJ, Sahin U, Türeci O, Gure AO, Tsang S, Williamson B, Stockert E, Pfreundschuh M, Old LJ: **A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening.** *Proc. Natl. Acad. Sci. U. S. A.* 1997, **94**:1914–8.

55. Türeci O, Chen YT, Sahin U, Güre AO, Zwick C, Villena C, Tsang S, Seitz G, Old LJ, Pfreundschuh M: **Expression of SSX genes in human tumors.** *Int. J. Cancer* 1998, **77**:19–23.

56. Chen Y-T, Scanlan MJ, Venditti CA, Chua R, Theiler G, Stevenson BJ, Iseli C, Gure AO, Vasicek T, Strausberg RL, Jongeneel CV, Old LJ, Simpson AJG: **Identification of cancer/testis-antigen genes by massively parallel signature sequencing**. *Proc. Natl. Acad. Sci. U. S. A.* 2005, **102**:7940–7945.

57. Feichtinger J, Aldeailej I, Anderson R, Almutairi M, Almatrafi A, Alsiwiehri N, Griffiths K, Stuart N, Wakeman JA, Larcombe L, McFarlane RJ: **Meta-analysis of clinical data using human meiotic genes identifies a novel cohort of highly restricted cancer-specific marker genes.** *Oncotarget* 2012, **3**:843–53.

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

59. Scanlan MJ, Gordon CM, Williamson B, Lee S-Y, Chen Y-T, Stockert E, Jungbluth A, Ritter G, Jäger D, Jäger E, Knuth A, Old LJ: **Identification of cancer/testis genes by database mining and mRNA expression analysis.** *Int. J. Cancer* 2002, **98**:485–492.

60. Almeida LG, Sakabe NJ, deOliveira AR, Silva MCC, Mundstein AS, Cohen T, Chen Y-T, Chua R, Gurung S, Gnjatic S, Jungbluth AA, Caballero OL, Bairoch A, Kiesler E, White SL, Simpson AJG, Old LJ, Camargo AA, Vasconcelos ATR: **CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens**. *Nucleic Acids Res.* 2009, **37**:D816–D819.

61. Zendman AJW, Ruiter DJ, Van Muijen GNP: **Cancer/testis-associated genes: identification, expression profile, and putative function.** *J. Cell. Physiol.* 2003, **194**:272–88.

62. Ross MT, Grafham D V, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, Frankish A, Lovell FL, Howe KL, Ashurst JL, Fulton RS, Sudbrak R, Wen G, Jones MC, Hurles ME, Andrews TD, Scott CE, Searle S, Ramser J, Whittaker A, Deadman R, Carter NP, Hunt SE, Chen R, Cree A, Gunaratne P, et al.: **The DNA sequence of the human X chromosome.** *Nature* 2005, **434**:325–37.

63. Sahin U, Türeci O, Chen YT, Seitz G, Villena-Heinsen C, Old LJ, Pfreundschuh M: **Expression of multiple cancer/testis (CT) antigens in breast cancer and melanoma: basis for polyvalent CT vaccine strategies.** *Int. J. Cancer* 1998, **78**:387–9.

64. Tajima K, Obata Y, Tamaki H, Yoshida M, Chen Y-T, Scanlan MJ, Old LJ, Kuwano H, Takahashi T, Takahashi T, Mitsudomi T: **Expression of cancer/testis (CT) antigens in lung cancer.** *Lung Cancer* 2003, **42**:23–33.

65. Ghafouri-Fard S, Modarressi M-H: **Cancer-testis antigens: potential targets for cancer immunotherapy.** *Arch. Iran. Med.* 2009, **12**:395–404.

66. De Bruijn DRH, Dos Santos NR, Kater-Baats E, Thijssen J, Van den Berk L, Stap J, Balemans M, Schepens M, Merkx G, Van Kessel AG: **The cancer-related protein SSX2 interacts with the human homologue of a Ras-like GTPase interactor, RAB3IP, and a novel nuclear protein, SSX2IP.** *Genes, Chromosomes Cancer* 2002, **34**:285–98.

67. Pivot-Pajot C, Caron C, Govin J, Vion A, Rousseaux S, Khochbin S: **Acetylation-dependent chromatin reorganization by BRDT, a testis-specific bromodomain-containing protein.** *Mol. Cell. Biol.* 2003, **23**:5354–65.

68. Loukinov DI, Pugacheva E, Vatolin S, Pack SD, Moon H, Chernukhin I, Mannan P, Larsson E, Kanduri C, Vostrov AA, Cui H, Niemitz EL, Rasko JEJ, Docquier FM, Kistler M, Breen JJ, Zhuang Z, Quitschke WW, Renkawitz R, Klenova EM, Feinberg AP, Ohlsson R, Morse HC, Lobanenkov V V: **BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma.** *Proc. Natl. Acad. Sci. U. S. A.* 2002, **99**:6806–11.

69. Laduron S, Deplus R, Zhou S, Kholmanskikh O, Godelaine D, De Smet C, Hayward SD, Fuks F, Boon T, De Plaen E: **MAGE-A1 interacts with adaptor SKIP and the deacetylase HDAC1 to repress transcription.** *Nucleic Acids Res.* 2004, **32**:4340–50.

70. Modarressi MH, Behnam B, Cheng M, Taylor KE, Wolfe J, Van der Hoorn FA: **Tsga10 encodes a 65-kilodalton protein that is processed to the 27-kilodalton fibrous sheath protein.** *Biol. Reprod.* 2004, **70**:608–15.

71. Brown PR, Miki K, Harper DB, Eddy EM: **A-kinase anchoring protein 4 binding proteins in the fibrous sheath of the sperm flagellum.** *Biol. Reprod.* 2003, **68**:2241–8.

72. Lacy HM, Sanderson RD: **Sperm protein 17 is expressed on normal and malignant lymphocytes and promotes heparan sulfate-mediated cell-cell adhesion.** *Blood* 2001, **98**:2160–5.

73. Eto K, Huet C, Tarui T, Kupriyanov S, Liu H-Z, Puzon-McLaughlin W, Zhang X-P, Sheppard D, Engvall E, Takada Y: **Functional classification of ADAMs based on a conserved motif for binding to integrin alpha 9beta 1: implications for sperm-egg binding and other cell interactions.** *J. Biol. Chem.* 2002, **277**:17804–10.

74. Lee JH, Schütte D, Wulf G, Füzesi L, Radzun H-J, Schweyer S, Engel W, Nayernia K: **Stem-cell protein Piwil2 is widely expressed in tumors and inhibits apoptosis through activation of Stat3/Bcl-XL pathway.** *Hum. Mol. Genet.* 2006, **15**:201–211.

75. Cilensek ZM, Yehiely F, Kular RK, Deiss LP: **A member of the GAGE family of tumor antigens is an anti-apoptotic gene that confers resistance to Fas/CD95/APO-1, Interferon-gamma, taxol and gamma-irradiation.** *Cancer Biol. Ther.* , **1**:380–7.

76. Cho B, Lim Y, Lee D-Y, Park S-Y, Lee H, Kim WH, Yang H, Bang Y-J, Jeoung D-I: **Identification and characterization of a novel cancer/testis antigen gene CAGE.** *Biochem. Biophys. Res. Commun.* 2002, **292**:715–26.

77. Xu H, Shan J, Jurukovski V, Yuan L, Li J, Tian K: **TSP50 encodes a testis-specific protease and is negatively regulated by p53.** *Cancer Res.* 2007, **67**:1239–45.

78. Pousette A, Leijonhufvud P, Arver S, Kvist U, Pelttari J, Höög C: **Presence of synaptonemal complex protein 1 transversal filament-like protein in human primary spermatocytes.** *Hum. Reprod.* 1997, **12**:2414–7.

79. Keeney S, Giroux CN, Kleckner N: **Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family.** *Cell* 1997, **88**:375–84.

80. Wojtasz L, Daniel K, Roig I, Bolcun-Filas E, Xu H, Boonsanay V, Eckmann CR, Cooke HJ, Jasin M, Keeney S, McKay MJ, Toth A: **Mouse HORMAD1 and HORMAD2, two conserved meiotic chromosomal proteins, are depleted from synapsed chromosome axes with the help of TRIP13 AAA-ATPase.** *PLoS Genet.* 2009, **5**:e1000702.

81. Ono T, Kurashige T, Harada N, Noguchi Y, Saika T, Niikawa N, Aoe M, Nakamura S, Higashi T, Hiraki A, Wada H, Kumon H, Old LJ, Nakayama E: **Identification of proacrosin binding protein sp32 precursor as a human cancer/testis antigen.** *Proc. Natl. Acad. Sci. U. S. A.* 2001, **98**:3282–7.

82. De Smet C, Lurquin C, Lethé B, Martelange V, Boon T: **DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter.** *Mol. Cell. Biol.* 1999, **19**:7327–35.

83. Weber J, Salgaller M, Samid D, Johnson B, Herlyn M, Lassam N, Treisman J, Rosenberg SA: **Expression of the MAGE-1 tumor antigen is up-regulated by the demethylating agent 5-aza-2'-deoxycytidine.** *Cancer Res.* 1994, **54**:1766–71.

84. Shichijo S, Yamada A, Sagawa K, Iwamoto O, Sakata M, Nagai K, Itoh K: **Induction of MAGE genes in lymphoid cells by the demethylating agent 5-aza-2'-deoxycytidine.** *Jpn. J. Cancer Res.* 1996, **87**:751–6.

85. Wang Z, Zhang Y, Ramsahoye B, Bowen D, Lim SH: **Sp17 gene expression in myeloma cells is regulated by promoter methylation.** *Br. J. Cancer* 2004, **91**:1597–603.

86. Wang Z, Zhang J, Zhang Y, Lim SH: **SPAN-Xb expression in myeloma cells is dependent on promoter hypomethylation and can be upregulated pharmacologically.** *Int. J. Cancer* 2006, **118**:1436–44.

87. De Smet C, De Backer O, Faraoni I, Lurquin C, Brasseur F, Boon T: **The activation of human gene MAGE-1 in tumor cells is correlated with genome-wide demethylation.** *Proc. Natl. Acad. Sci. U. S. A.* 1996, **93**:7149–53.

88. Klenova EM, Morse HC, Ohlsson R, Lobanenkov V V: **The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer.** *Semin. Cancer Biol.* 2002, **12**:399–414.

89. Vatolin S, Abdullaev Z, Pack SD, Flanagan PT, Custer M, Loukinov DI, Pugacheva E, Hong JA, Morse H, Schrump DS, Risinger JI, Barrett JC, Lobanenkov V V: **Conditional expression of the CTCF-paralogous transcriptional factor BORIS in normal cells results in demethylation and derepression of MAGE-A1 and reactivation of other cancer-testis genes.** *Cancer Res.* 2005, **65**:7751–62.

90. Hong JA, Kang Y, Abdullaev Z, Flanagan PT, Pack SD, Fischette MR, Adnani MT, Loukinov DI, Vatolin S, Risinger JI, Custer M, Chen GA, Zhao M, Nguyen DM, Barrett JC, Lobanenkov V V, Schrump DS: **Reciprocal binding of CTCF and BORIS to the NY-ESO-1 promoter coincides with derepression of this cancer-testis gene in lung cancer cells.** *Cancer Res.* 2005, **65**:7763–74.

91. Woloszynska-Read A, James SR, Song C, Jin B, Odunsi K, Karpf AR: **BORIS/CTCFL expression is insufficient for cancer-germline antigen gene expression and DNA hypomethylation in ovarian cell lines.** *Cancer Immun.* 2010, **10**:6.

92. Goelz SE, Vogelstein B, Hamilton SR, Feinberg AP: **Hypomethylation of DNA from benign and malignant human colon neoplasms.** *Science* 1985, **228**:187–90.

93. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat. Genet.* 2003, **33 Suppl**:245–54.

94. Champion MD, Hawley RS: **Playing for half the deck: the molecular biology of meiosis.** *Nat. Cell Biol.* 2002, **4 Suppl**:s50–6.

95. Yanowitz J: **Meiosis: making a break for it.** *Curr. Opin. Cell Biol.* 2010, **22**:744–51.

96. Handel MA, Schimenti JC: **Genetics of mammalian meiosis: regulation, dynamics and impact on fertility.** *Nat. Rev. Genet.* 2010, **11**:124–36.

97. Castro A, Lorca T: **Exploring meiotic division in Cargèse. Meeting on meiotic divisions and checkpoints.** *EMBO Rep.* 2005, **6**:821–5.

98. Fraune J, Schramm S, Alsheimer M, Benavente R: **The mammalian synaptonemal complex: protein components, assembly and role in meiotic recombination.** *Exp. Cell Res.* 2012, **318**:1340–6.

99. Page SL, Hawley RS: **The genetics and molecular biology of the synaptonemal complex.** *Annu. Rev. Cell Dev. Biol.* 2004, **20**:525–58.

100. Hawley RS: **Solving a meiotic LEGO puzzle: transverse filaments and the assembly of the synaptonemal complex in Caenorhabditis elegans.** *Genetics* 2011, **189**:405–9.

101. Costa Y, Cooke HJ: **Dissecting the mammalian synaptonemal complex using targeted mutations.** *Chromosome Res.* 2007, **15**:579–89.

102. Youds JL, Boulton SJ: **The choice in meiosis - defining the factors that influence crossover or non-crossover formation.** *J. Cell. Sci.* 2011, **124**:501–13.

103. Andersen SL, Sekelsky J: **Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes.** *Bioessays* 2010, **32**:1058–66.

104. Phadnis N, Hyppa RW, Smith GR: **New and old ways to control meiotic recombination.** *Trends Genet.* 2011, **27**:411–21.

105. Bosco G: **When segregation hangs by a thread.** *PLoS Genet.* 2009, **5**:e1000371.

106. Petronczki M, Siomos MF, Nasmyth K: **Un ménage à quatre: the molecular biology of chromosome segregation in meiosis.** *Cell* 2003, **112**:423–40.

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

108. Söder O: **Sexual dimorphism of gonadal development.** *Best Pract. Res. Clin. Endocrinol. Metab.* 2007, **21**:381–91.

109. Baillet A, Mandon-Pepin B: **Mammalian ovary differentiation - a focus on female meiosis.** *Mol. Cell. Endocrinol.* 2012, **356**:13–23.

110. Hunt PA, Hassold TJ: **Sex matters in meiosis.** *Science* 2002, **296**:2181–3.

111. De Felici M, Farini D: **The control of cell cycle in mouse primordial germ cells: old and new players.** *Curr. Pharm. Des.* 2012, **18**:233–44.

112. Von Stetina JR, Orr-Weaver TL: **Developmental control of oocyte maturation and egg activation in metazoan models.** *Cold Spring Harb Perspect Biol* 2011, **3**:a005553.

113. Hunt PA, Hassold TJ: **Human female meiosis: what makes a good egg go bad?** *Trends Genet.* 2008, **24**:86–93.

114. Telfer EE, McLaughlin M: **Natural history of the mammalian oocyte**. *Reprod. BioMed. Online* 2007, **15**:288–295.

115. Pepling ME: **Follicular assembly: mechanisms of action.** *Reproduction* 2012, **143**:139–49.

116. Cheng CY, Wong EWP, Yan HHN, Mruk DD: **Regulation of spermatogenesis in the microenvironment of the seminiferous epithelium: new insights and advances.** *Mol. Cell. Endocrinol.* 2010, **315**:49–56.

117. Hess RA, Renato de Franca L: **Spermatogenesis and cycle of the seminiferous epithelium.** *Adv. Exp. Med. Biol.* 2008, **636**:1–15.

118. Lie PPY, Mruk DD, Lee WM, Cheng CY: **Cytoskeletal dynamics and spermatogenesis.** *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 2010, **365**:1581–92.

119. Johnson L, Thompson DL, Varner DD: **Role of Sertoli cell number and function on regulation of spermatogenesis.** *Anim. Reprod. Sci.* 2008, **105**:23–51.

120. Cheng CY, Mruk DD: **The biology of spermatogenesis: the past, present and future.** *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 2010, **365**:1459–63.

121. O'Donnell L, Nicholls PK, O'Bryan MK, McLachlan RI, Stanton PG: **Spermiation: The process of sperm release.** *Spermatogenesis* 2011, **1**:14–35.

122. Smith BE, Braun RE: **Germ Cell Migration Across Sertoli Cell Tight Junctions.** *Science* 2012, **338**:798–802.

123. Pelletier R-M: **The blood-testis barrier: the junctional permeability, the proteins and the lipids.** *Prog. Histochem. Cytochem.* 2011, **46**:49–127.

124. Fijak M, Bhushan S, Meinhardt A: **Immunoprivileged sites: the testis.** *Methods Mol. Biol.* 2011, **677**:459–70.

125. Kauppi L, Jasin M, Keeney S: **The tricky path to recombining X and Y chromosomes in meiosis.** *Ann. N. Y. Acad. Sci.* 2012, **1267**:18–23.

126. Turner JMA: **Meiotic sex chromosome inactivation.** *Development* 2007, **134**:1823–31.

127. De Vries M, Vosters S, Merkx G, D'Hauwers K, Wansink DG, Ramos L, De Boer P: **Human male meiotic sex chromosome inactivation.** *PLoS ONE* 2012, **7**:e31485.

128. Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SSW, Demougin P, Gattiker A, Moore J, Patard J-J, Wolgemuth DJ, Jégou B, Primig M: **The conserved transcriptome in human and rodent male gametogenesis.** *Proc. Natl. Acad. Sci. U. S. A.* 2007, **104**:8346–8351.

129. Chalmel F, Lardenois A, Primig M: **Toward understanding the core meiotic transcriptome in mammals and its implications for somatic cancer.** *Ann. N. Y. Acad. Sci.* 2007, **1120**:1–15.

130. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, De Massy B: **PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice.** *Science* 2010, **327**:836–40.

131. McVean G, Myers S: **PRDM9 marks the spot.** *Nat. Genet.* 2010, **42**:821–2.

132. Steilmann C, Cavalcanti MCO, Bartkuhn M, Pons-Kühnemann J, Schuppe H-C, Weidner W, Steger K, Paradowska A: **The interaction of modified histones with the bromodomain testis-specific (BRDT) gene and its mRNA level in sperm of fertile donors and subfertile men.** *Reproduction* 2010, **140**:435–43.

133. Yamaguchi S, Hong K, Liu R, Shen L, Inoue A, Diep D, Zhang K, Zhang Y: **Tet1 controls meiosis by regulating meiotic gene expression**. *Nature* 2012.

134. Bowles J, Koopman P: **Retinoic acid, meiosis and germ cell fate in mammals.** *Development* 2007, **134**:3401–11.

135. Keeney S, Neale MJ: **Initiation of meiotic recombination by formation of DNA double-strand breaks: mechanism and regulation.** *Biochem. Soc. Trans.* 2006, **34**:523–5.

136. Hunter N, Kleckner N: **The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination.** *Cell* 2001, **106**:59–70.

137. Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW: **The double-strand-break repair model for recombination.** *Cell* 1983, **33**:25–35.

138. Schwacha A, Kleckner N: **Identification of double Holliday junctions as intermediates in meiotic recombination.** *Cell* 1995, **83**:783–91.

139. Zakharyevich K, Tang S, Ma Y, Hunter N: **Delineation of joint molecule resolution pathways in meiosis identifies a crossover-specific resolvase.** *Cell* 2012, **149**:334–47.

140. Schwartz EK, Heyer W-D: **Processing of joint molecule intermediates by structure-selective endonucleases during homologous recombination in eukaryotes.** *Chromosoma* 2011, **120**:109–27.

141. Whitby MC: **Making crossovers during meiosis.** *Biochem. Soc. Trans.* 2005, **33**:1451–5.

142. Lynn A, Soucek R, Börner GV: **ZMM proteins during meiosis: crossover artists at work.** *Chromosome Res.* 2007, **15**:591–605.

143. Osman F, Dixon J, Doe CL, Whitby MC: **Generating crossovers by resolution of nicked Holliday junctions: a role for Mus81-Eme1 in meiosis.** *Mol. Cell* 2003, **12**:761–74.

144. Bachrati CZ, Borts RH, Hickson ID: **Mobile D-loops are a preferred substrate for the Bloom's syndrome helicase.** *Nucleic Acids Res.* 2006, **34**:2269–79.

145. Barber LJ, Youds JL, Ward JD, McIlwraith MJ, O'Neil NJ, Petalcorin MIR, Martin JS, Collis SJ, Cantor SB, Auclair M, Tissenbaum H, West SC, Rose AM, Boulton SJ: **RTEL1 maintains genomic stability by suppressing homologous recombination.** *Cell* 2008, **135**:261–71.

146. Wu L, Hickson ID: **The Bloom's syndrome helicase suppresses crossing over during homologous recombination.** *Nature* 2003, **426**:870–4.

147. Buard J, De Massy B: **Playing hide and seek with mammalian meiotic crossover hotspots.** *Trends Genet.* 2007, **23**:301–9.

148. Serrentino M-E, Borde V: **The spatial regulation of meiotic recombination hotspots: are all DSB hotspots crossover hotspots?** *Exp. Cell Res.* 2012, **318**:1347–52.

149. Hochwagen A, Marais GAB: **Meiosis: a PRDM9 guide to the hotspots of recombination.** *Curr. Biol.* 2010, **20**:R271–4.

150. Moses MJ: **Synaptinemal Complex**. *Annu. Rev. Genet.* 1968, **2**:363–412.

151. Offenberg HH, Schalk JA, Meuwissen RL, Van Aalderen M, Kester HA, Dietrich AJ, Heyting C: **SCP2: a major protein component of the axial elements of synaptonemal complexes of the rat.** *Nucleic Acids Res.* 1998, **26**:2572–9.

152. Lammers JH, Offenberg HH, Van Aalderen M, Vink AC, Dietrich AJ, Heyting C: **The gene encoding a major component of the lateral elements of synaptonemal complexes of the rat is related to X-linked lymphocyte-regulated genes.** *Mol. Cell. Biol.* 1994, **14**:1137–46.

153. Eijpe M, Offenberg H, Jessberger R, Revenkova E, Heyting C: **Meiotic cohesin REC8 marks the axial elements of rat synaptonemal complexes before cohesins SMC1beta and SMC3.** *J. Cell Biol.* 2003, **160**:657–70.

154. Kouznetsova A, Novak I, Jessberger R, Höög C: **SYCP2 and SYCP3 are required for cohesin core integrity at diplotene but not for centromere cohesion at the first meiotic division.** *J. Cell. Sci.* 2005, **118**:2271–8.

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

direct interactions between the N-termini of SCP1 molecules organized head-to-head. *Exp. Cell Res.* 1996, **226**:11–9.

156. De Vries FAT, De Boer E, Van den Bosch M, Baarends WM, Ooms M, Yuan L, Liu J-G, Van Zeeland AA, Heyting C, Pastink A: **Mouse Sycp1 functions in synaptonemal complex assembly, meiotic recombination, and XY body formation.** *Genes Dev.* 2005, **19**:1376–89.

157. Costa Y, Speed R, Ollinger R, Alsheimer M, Semple CA, Gautier P, Maratou K, Novak I, Höög C, Benavente R, Cooke HJ: **Two novel proteins recruited by synaptonemal complex protein 1 (SYCP1) are at the centre of meiosis.** *J. Cell. Sci.* 2005, **118**:2755–62.

158. Schramm S, Fraune J, Naumann R, Hernandez-Hernandez A, Höög C, Cooke HJ, Alsheimer M, Benavente R: **A novel mouse synaptonemal complex protein is essential for loading of central element proteins, recombination, and fertility.** *PLoS Genet.* 2011, **7**:e1002088.

159. Hamer G, Gell K, Kouznetsova A, Novak I, Benavente R, Höög C: **Characterization of a novel meiosis-specific protein within the central element of the synaptonemal complex.** *J. Cell. Sci.* 2006, **119**:4025–32.

160. Bolcun-Filas E, Costa Y, Speed R, Taggart M, Benavente R, De Rooij DG, Cooke HJ: **SYCE2 is required for synaptonemal complex assembly, double strand break repair, and homologous recombination.** *J. Cell Biol.* 2007, **176**:741–7.

161. Bolcun-Filas E, Hall E, Speed R, Taggart M, Grey C, De Massy B, Benavente R, Cooke HJ: **Mutation of the mouse Syce1 gene disrupts synapsis and suggests a link between synaptonemal complex structural components and DNA repair.** *PLoS Genet.* 2009, **5**:e1000393.

162. Peoples TL, Dean E, Gonzalez O, Lambourne L, Burgess SM: **Close, stable homolog juxtaposition during meiosis in budding yeast is dependent on meiotic recombination, occurs independently of synapsis, and is distinct from DSB-independent pairing contacts.** *Genes Dev.* 2002, **16**:1682–95.

163. Baudat F, Manova K, Yuen JP, Jasin M, Keeney S: **Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11.** *Mol. Cell* 2000, **6**:989–98.

164. Bähler J, Wyler T, Loidl J, Kohli J: **Unusual nuclear structures in meiotic prophase of fission yeast: a cytological analysis.** *J. Cell Biol.* 1993, **121**:241–56.

165. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell.* 5th edition. New York: Garland Science; 2007.

166. Old LJ: **Cancer/testis (CT) antigens - a new link between gametogenesis and cancer.** *Cancer Immun.* 2001, **1**:1.

167. Wu X, Ruvkun G: **Germ cell genes and cancer.** *Science* 2010, **330**:1761–1762.

168. Janic A, Mendizabal L, Llamazares S, Rossell D, Gonzalez C: **Ectopic expression of germline genes drives malignant brain tumor growth in Drosophila.** *Science* 2010, **330**:1824–1827.

169. Unhavaithaya Y, Shin TH, Miliaras N, Lee J, Oyama T, Mello CC: **MEP-1 and a homolog of the NURD complex component Mi-2 act together to maintain germline-soma distinctions in C. elegans.** *Cell* 2002, **111**:991–1002.

170. Wang D, Kennedy S, Conte D, Kim JK, Gabel HW, Kamath RS, Mello CC, Ruvkun G: **Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants.** *Nature* 2005, **436**:593–597.

171. Ridley AJ, Schwartz MA, Burridge K, Firtel RA, Ginsberg MH, Borisy G, Parsons JT, Horwitz AR: **Cell migration: integrating signals from front to back.** *Science* 2003, **302**:1704–9.

172. Wang J, Emadali A, Le Bescont A, Callanan M, Rousseaux S, Khochbin S: **Induced malignant genome reprogramming in somatic cells by testis-specific factors.** *Biochim. Biophys. Acta.* 2011, **1809**:221–225.

173. Kalejs M, Ivanov A, Plakhins G, Cragg MS, Emzinsh D, Illidge TM, Erenpreisa J: **Upregulation of meiosis-specific genes in lymphoma cell lines following genotoxic insult and induction of mitotic catastrophe**. *BMC Cancer* 2006, **6**:6.

174. Feichtinger J, Thallinger GG, McFarlane RJ, Larcombe L: **Microarray Meta-Analysis: From Data to Expression to Biological Relationships**. In *Computational Medicine.* edited by Trajanoski Z New York, NY: Springer Berlin Heidelberg; 2012:59–77.

175. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252**:1651–1656.

176. Lee NH, Weinstock KG, Kirkness EF, Earle-Hughes JA, Fuldner RA, Marmaros S, Glodek A, Gocayne JD, Adams MD, Kerlavage AR: **Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment.** *Proc. Natl. Acad. Sci. U. S. A.* 1995, **92**:8303–7.

177. Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ: **Comprehensive EST analysis of tomato and comparative genomics of fruit ripening.** *Plant J.* 2004, **40**:47–59.

178. Ajioka JW, Boothroyd JC, Brunk BP, Hehl A, Hillier L, Manger ID, Marra M, Overton GC, Roos DS, Wan KL, Waterston R, Sibley LD: **Gene discovery by EST sequencing in Toxoplasma gondii reveals sequences restricted to the Apicomplexa.** *Genome Res.* 1998, **8**:18–28.

179. Verdun RE, Di Paolo N, Urmenyi TP, Rondinelli E, Frasch AC, Sanchez DO: **Gene discovery through expressed sequence Tag sequencing in Trypanosoma cruzi.** *Infect. Immun.* 1998, **66**:5393–8.

180. Jongeneel C V: **Searching the expressed sequence tag (EST) databases: panning for genes.** *Brief. Bioinformatics* 2000, **1**:76–92.

181. Brett D, Hanke J, Lehmann G, Haase S, Delbrück S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms.** *FEBS Lett.* 2000, **474**:83–6.

182. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res.* 1999, **9**:1288–93.

183. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M: **Mining SNPs from EST databases.** *Genome Res.* 1999, **9**:167–74.

184. Buetow KH, Edmonson MN, Cassidy AB: **Reliable identification of large numbers of candidate SNPs from public EST data.** *Nat. Genet.* 1999, **21**:323–5.

185. Kim B, Lee HJ, Choi HY, Shin Y, Nam S, Seo G, Son D-S, Jo J, Kim J, Lee J, Kim J, Kim K, Lee S: **Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data.** *Cancer Res.* 2007, **67**:7431–7438.

186. Campagne F, Skrabanek L: **Mining expressed sequence tags identifies cancer markers of clinical interest**. *BMC Bioinf.* 2006, **7**:481.

187. Skrabanek L, Campagne F: **TissueInfo: high-throughput identification of tissue expression profiles and specificity**. *Nucleic Acids Res.* 2001, **29**:e102.

188. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res.* 1999, **9**:1106–15.

189. Ewing RM, Ben Kahla A, Poirot O, Lopez F, Audic S, Claverie JM: **Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression.** *Genome Res.* 1999, **9**:950–9.

190. Kan Z, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** *Genome Res.* 2001, **11**:889–900.

191. Jiang J, Jacob HJ: **EbEST: an automated tool using expressed sequence tags to delineate gene structure.** *Genome Res.* 1998, **8**:268–75.

192. Rudd S: **Expressed sequence tags: alternative or complement to whole genome sequences?** *Trends Plant Sci.* 2003, **8**:321–9.

193. Bonaldo MF, Lennon G, Soares MB: **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genome Res.* 1996, **6**:791–806.

194. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis.** *Brief. Bioinformatics* 2007, **8**:6–21.

195. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res.* 1998, **8**:175–85.

196. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res.* 1998, **8**:186–94.

197. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997, **25**:3389–402.

198. **Phred, Phrap, and Consed** [`http://www.phrap.org/phredphrapconsed.html`].

199. Li S, Chou H-H: **LUCY2: an interactive DNA sequence quality trimming and vector removal tool.** *Bioinformatics* 2004, **20**:2865–6.

200. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet.* 2000, **16**:276–7.

201. Wan H, Li L, Federhen S, Wootton JC: **Discovering simple regions in biological sequences associated with scoring schemes.** *J. Comput. Biol.* 2003, **10**:171–85.

202. Tempel S: **Using and understanding RepeatMasker.** *Methods Mol. Biol.* 2012, **859**:29–51.

203. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res.* 1999, **9**:868–77.

204. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res.* 2004, **14**:1147–59.

205. Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S: **EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments.** *Nucleic Acids Res.* 2006, **34**:W459–62.

206. Mao C, Cushman JC, May GD, Weller JW: **ESTAP–an automated system for the analysis of EST data.** *Bioinformatics* 2003, **19**:1720–2.

207. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res.* 1997, **7**:986–95.

208. Fisher RA: *Statistical methods for research workers.* 12th edition. Edinburgh: Oliver and Boyd; 1954, **354**:356

209. Stekel DJ, Git Y, Falciani F: **The comparison of gene expression from multiple cDNA libraries.** *Genome Res.* 2000, **10**:2055–61.

210. Normand SL: **Meta-analysis: formulating, evaluating, combining, and reporting.** *Stat. Med.* 1999, **18**:321–359.

211. Fierro AC, Vandenbussche F, Engelen K, Van De Peer Y, Marchal K: **Meta Analysis of Gene Expression Data within and Across Species**. *Curr. Genomics* 2008, **9**:525–534.

212. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG: **Global Prediction of Tissue-Specific Gene Expression and Context-Dependent Gene Networks in Caenorhabditis elegans**. *PLoS Comput. Biol.* 2009, **5**:13.

213. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc. Natl. Acad. Sci. U. S. A.* 2004, **101**:6062–6067.

214. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR: **Multiclass cancer diagnosis using tumor gene expression signatures**. *Proc. Natl. Acad. Sci. U. S. A.* 2001, **98**:15149–15154.

215. Stanton JL, Green DP: **Meta-analysis of gene expression in mouse preimplantation embryo development.** *Mol. Hum. Reprod.* 2001, **7**:545–52.

216. Yu X, Lin J, Zack DJ, Qian J: **Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues.** *Nucleic Acids Res.* 2006, **34**:4925–36.

217. Kim B, Lee HJ, Choi HY, Shin Y, Nam S, Seo G, Son D-S, Jo J, Kim J, Lee J, Kim J, Kim K, Lee S: **Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data.** *Cancer Res.* 2007, **67**:7431–7438.

218. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST–database for "expressed sequence tags".** *Nat. Genet.* 1993, **4**:332–3.

219. **NCBI dbEST** [`http://www.ncbi.nlm.nih.gov/dbEST/`].

220. Boguski MS, Schuler GD: **ESTablishing a human transcript map.** *Nat. Genet.* 1995, **10**:369–71.

221. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tomé P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, et al.: **A gene map of the human genome.** *Science* 1996, **274**:540–6.

222. Pontius JU, Wagner L, Schuler GD: **UniGene: a unified view of the transcriptome.** In *The NCBI Handbook.* edited by McEntyre J, Ostell J National Center for Biotechnology Information; 2003:1–12.

223. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic Acids Res.* 2005, **33**:D71–4.

224. Quackenbush J, Liang F, Holt I, Pertea G, Upton J: **The TIGR gene indices: reconstruction and representation of expressed gene sequences.** *Nucleic Acids Res.* 2000, **28**:141–5.

225. Dong Q, Schlueter SD, Brendel V: **PlantGDB, plant genome database and analysis tools.** *Nucleic Acids Res.* 2004, **32**:D354–9.

226. Ptitsyn A, Hide W: **CLU: a new algorithm for EST clustering.** *BMC Bioinf.* 2005, **6 Suppl 2**:S3.

227. Mullan L: **"We are gathered here today" – EST cluster databases**. *Briefings Bioinf.* 2004, **5**:284–286.

228. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J. Comput. Biol.* , **7**:203–14.

229. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets.** *Bioinformatics* 2003, **19**:651–2.

230. Liu X, Yu X, Zack DJ, Zhu H, Qian J: **TiGER: A database for tissue-specific gene expression and regulation**. *BMC Bioinf.* 2008, **9**:271.

231. Wang J, Liang P: **DigiNorthern, digital expression analysis of query genes based on ESTs.** *Bioinformatics* 2003, **19**:653–654.

232. Chen Y-C, Hsiao C-D, Lin W-D, Hu C-M, Hwang P-P, Ho J-M: **ZooDDD: a cross-species database for digital differential display analysis.** *Bioinformatics* 2006, **22**:2180–2182.

233. Wu X, Walker MG, Luo J, Wei L: **GBA server: EST-based digital gene expression profiling**. *Nucleic Acids Res.* 2005, **33**:W673–W676.

234. Feichtinger J, McFarlane RJ, Larcombe LD: **CancerMA: a Web-based Tool for Automatic Meta-analysis of Public Cancer Microarray Data**. *Database* 2012, **in print**.

235. Butterfield LH, Comin-Anduix B, Vujanovic L, Lee Y, Dissette VB, Yang J-Q, Vu HT, Seja E, Oseguera DK, Potter DM, Glaspy JA, Economou JS, Ribas A: **Adenovirus MART-1-engineered autologous dendritic cell vaccine for metastatic melanoma.** *J. Immunother.* 2008, **31**:294–309.

236. Odunsi K, Matsuzaki J, Karbach J, Neumann A, Mhawech-Fauceglia P, Miller A, Beck A, Morrison CD, Ritter G, Godoy H, Lele S, duPont N, Edwards R, Shrikant P, Old LJ, Gnjatic S, Jäger E: **Efficacy of vaccination with recombinant vaccinia and fowlpox vectors expressing NY-ESO-1 antigen in ovarian cancer and melanoma patients.** *Proc. Natl. Acad. Sci. U. S. A.* 2012, **109**:5797–802.

237. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484–7.

238. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467–70.

239. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat. Rev. Genet.* 2009, **10**:57–63.

240. Ottesen EA, Hong JW, Quake SR, Leadbetter JR: **Microfluidic digital PCR enables multigene analysis of individual environmental bacteria.** *Science* 2006, **314**:1464–7.

241. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A: **ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments**. *Nucleic Acids Res.* 2011, **39**:D1002–D1004.

242. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets—10 years on**. *Nucleic Acids Res.* 2011, **39**:D1005–D1010.

243. Leinonen R, Sugawara H, Shumway M: **The sequence read archive**. *Nucleic Acids Res.* 2011, **39**:D19–21.

244. VanGuilder HD, Vrana KE, Freeman WM: **Twenty-five years of quantitative PCR for gene expression analysis.** *BioTechniques* 2008, **44**:619–26.

245. Richards MP: **Techniques for Gene Expression Profiling**. In *Medical Biomethods Handbook.* edited by Walker JM, Rapley R Springer Protocols; 2005:11.

246. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets**. *PLoS Med.* 2008, **5**:13.

247. Scholtens D, Von Heydebreck A: **Analysis of Differential Gene Expression Studies**. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* edited by Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S New York, Cambridge: Springer; 2005:229–248.

248. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martínez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W: **Multiple-laboratory comparison of microarray platforms.** *Nat. Methods* 2005, **2**:345–50.

249. Larsson O, Sandberg R: **Lack of correct data format and comparability limits future integrative microarray research.** *Nat. Biotechnol.* 2006, **24**:1322–3.

250. Smyth GK: **Limma: Linear Models for Microarray Data**. *In Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* edited by Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S New York: Springer; 2005:397–420.

251. Hubbell E, Liu W-M, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18**:1585–92.

252. Stouffer SA: *The American soldier.* Princeton, NJ: Princeton University Press; 1949.

253. Morgan AA, Khatri P, Jones RH, Sarwal MM, Butte AJ: **Comparison of multiplex meta analysis techniques for understanding the acute rejection of solid organ transplants**. *BMC Bioinf.* 2010, **11**:S6.

254. Campain A, Yang YH: **Comparison study of microarray meta-analysis methods**. *BMC Bioinf.* 2010, **11**:408.

255. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, De Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen R V, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, et al.: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat. Biotechnol.* 2006, **24**:1151–61.

256. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis.** *Bioinformatics* 2005, **21**:3683–3685.

257. **Data Protection and Freedom of Information advice - ICO** [`http://www.ico.gov.uk/`].

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

259. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res.* 2009, **19**:1639–1645.

260. Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser**. *BMC Bioinf.* 2011, **12**:385.

261. Lewis S, Clarke M: **Forest plots: trying to see the wood and the trees**. *BMJ* 2001, **322**:1479–1480.

262. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003, **13**:2498–504.

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

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

265. Kettunen E, Anttila S, Seppänen JK, Karjalainen A, Edgren H, Lindström I, Salovaara R, Nissén A-M, Salo J, Mattson K, Hollmén J, Knuutila S, Wikman H: **Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer.** *Cancer Genet. Cytogenet.* 2004, **149**:98–106.

266. Hough CD, Cho KR, Zonderman AB, Schwartz DR, Morin PJ: **Coordinately up-regulated genes in ovarian cancer.** *Cancer Res.* 2001, **61**:3869–3876.

267. Neumann B, Walter T, Hériché J-K, Bulkescher J, Erfle H, Conrad C, Rogers P, Poser I, Held M, Liebel U, Cetin C, Sieckmann F, Pau G, Kabbe R, Wünsche A, Satagopam V, Schmitz MHA, Chapuis C, Gerlich DW, Schneider R, Eils R, Huber W, Peters J-M, Hyman AA, Durbin R, Pepperkok R, Ellenberg J: **Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes.** *Nature* 2010, **464**:721–727.

268. Hiriart E, Vavasseur A, Touat-Todeschini L, Yamashita A, Gilquin B, Lambert E, Perot J, Shichino Y, Nazaret N, Boyault C, Lachuer J, Perazza D, Yamamoto M, Verdel A: **Mmi1 RNA surveillance machinery directs RNAi complex RITS to specific meiotic genes in fission yeast.** *EMBO J.* 2012, **31**:2296–308.

269. Harigaya Y, Tanaka H, Yamanaka S, Tanaka K, Watanabe Y, Tsutsumi C, Chikashige Y, Hiraoka Y, Yamashita A, Yamamoto M: **Selective elimination of messenger RNA prevents an incidence of untimely meiosis.** *Nature* 2006, **442**:45–50.

270. Jessberger R: **Cohesin complexes get more complex: the novel kleisin RAD21L.** *Cell Cycle* 2011, **10**:2053–4.

271. Uhlmann F: **Cohesin subunit Rad21L, the new kid on the block has new ideas.** *EMBO Rep.* 2011, **12**:183–4.

272. Hayashi K, Yoshida K, Matsui Y: **A histone H3 methyltransferase controls epigenetic events required for meiotic prophase.** *Nature* 2005, **438**:374–8.

273. Fay DS, Yochem J: **The SynMuv genes of Caenorhabditis elegans in vulval development and beyond.** *Dev. Biol.* 2007, **306**:1–9.

274. Passannante M, Marti C-O, Pfefferli C, Moroni PS, Kaeser-Pebernard S, Puoti A, Hunziker P, Wicky C, Müller F: **Different Mi-2 Complexes for Various Developmental Functions in Caenorhabditis elegans**. *PLoS ONE* 2010, **5**:15.

275. Wu X, Shi Z, Cui M, Han M, Ruvkun G: **Repression of germline RNAi pathways in somatic cells by retinoblastoma pathway chromatin complexes.** *PLoS Genet.* 2012, **8**:e1002542.

276. Georlette D, Ahn S, MacAlpine DM, Cheung E, Lewis PW, Beall EL, Bell SP, Speed T, Manak JR, Botchan MR: **Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells.** *Genes Dev.* 2007, **21**:2880–2896.

277. Kee K, Angeles VT, Flores M, Nguyen HN, Reijo Pera RA: **Human DAZL, DAZ and BOULE genes modulate primordial germ-cell and haploid gamete formation.** *Nature* 2009, **462**:222–5.

278. Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH: **Classical nuclear localization signals: definition, function, and interaction with importin alpha.** *J. Biol. Chem.* 2007, **282**:5101–5.

279. Marfori M, Mynott A, Ellis JJ, Mehdi AM, Saunders NFW, Curmi PM, Forwood JK, Bodén M, Kobe B: **Molecular basis for specificity of nuclear import and prediction of nuclear localization.** *Biochim. Biophys. Acta* 2011, **1813**:1562–77.

280. Hawkins J, Davis L, Bodén M: **Predicting nuclear localization.** *J. Proteome Res.* 2007, **6**:1402–9.

281. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM: **NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction.** *BMC Bioinf.* 2009, **10**:202.

282. UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res.* 2011, **39**:D214–9.

283. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart–biological queries made easy.** *BMC Genomics* 2009, **10**:22.

284. Bird A: **Perceptions of epigenetics.** *Nature* 2007, **447**:396–8.

285. Hore TA, Deakin JE, Marshall Graves JA: **The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes.** *PLoS Genet.* 2008, **4**:e1000169.

286. De Necochea-Campion R, Ghochikyan A, Josephs SF, Zacharias S, Woods E, Karimi-Busheri F, Alexandrescu DT, Chen C-S, Agadjanyan MG, Carrier E: **Expression of the epigenetic factor BORIS (CTCFL) in the human genome.** *J Transl Med* 2011, **9**:213.

287. Bhan S, Negi SS, Shao C, Glazer CA, Chuang A, Gaykalova DA, Sun W, Sidransky D, Ha PK, Califano JA: **BORIS binding to the promoters of cancer testis antigens, MAGEA2, MAGEA3, and MAGEA4, is associated with their transcriptional activation in lung cancer.** *Clin. Cancer Res.* 2011, **17**:4267–76.

288. Gaykalova D, Vatapalli R, Glazer CA, Bhan S, Shao C, Sidransky D, Ha PK, Califano JA: **Dose-dependent activation of putative oncogene SBSN by BORIS.** *PLoS ONE* 2012, **7**:e40389.

289. Martin-Kleiner I: **BORIS in human cancers – a review.** *Eur. J. Cancer* 2012, **48**:929–35.

290. Ohlsson R, Lobanenkov V, Klenova E: **Does CTCF mediate between nuclear organization and gene expression?** *Bioessays* 2010, **32**:37–50.

291. R Core Team: **R: A Language and Environment for Statistical Computing**. *R Foundation for Statistical Computing* 2011, **1**:409.

292. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Denise C-S, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, et al.: **Ensembl 2012**. *Nucleic Acids Res.* 2011, **40**:84–90.

293. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA: **genenames.org: the HGNC resources in 2011**. *Nucleic Acids Res.* 2011, **39**:D514–D519.