BMC Genomics

BioMed Central

Research article

Identification of "pathologs" (disease-related genes) from the RIKEN mouse cDNA dataset using human curation plus FACTS, a new biological information extraction system

Diego G Silva[1,2], Christian Schönbach[3], Vladimir Brusic[4], Luis A Socha[1,2], Takeshi Nagashima[3] and Nikolai Petrovsky*[1,2]

Address: [1]Medical Informatics Centre, University of Canberra, ACT 2601 Australia, [2]John Curtin School of Medical Research, Australian National University, Canberra ACT 2601, Australia, [3]Biomedical Knowledge Discovery Team, Bioinformatics Group, RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan and [4]Institute for Infocomm Research, Singapore 11961

Email: Diego G Silva - diego.silva@anu.edu.au; Christian Schönbach - schoen@postman.riken.jp; Vladimir Brusic - vladimir@i2r.a-star.edu.sg; Luis A Socha - luis.socha@anu.edu.au; Takeshi Nagashima - nagasima@gsc.riken.jp; Nikolai Petrovsky* - nikolai.petrovsky@anu.edu.au

* Corresponding author

Published: 29 April 2004

BMC Genomics 2004, **5**:28

Received: 10 October 2003
Accepted: 29 April 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/28

Abstract

Background: A major goal in the post-genomic era is to identify and characterise disease susceptibility genes and to apply this knowledge to disease prevention and treatment. Rodents and humans have remarkably similar genomes and share closely related biochemical, physiological and pathological pathways. In this work we utilised the latest information on the mouse transcriptome as revealed by the RIKEN FANTOM2 project to identify novel human disease-related candidate genes. We define a new term "patholog" to mean a homolog of a human disease-related gene encoding a product (transcript, anti-sense or protein) potentially relevant to disease. Rather than just focus on Mendelian inheritance, we applied the analysis to all potential pathologs regardless of their inheritance pattern.

Results: Bioinformatic analysis and human curation of 60,770 RIKEN full-length mouse cDNA clones produced 2,578 sequences that showed similarity (70–85% identity) to known human-disease genes. Using a newly developed biological information extraction and annotation tool (FACTS) in parallel with human expert analysis of 17,051 MEDLINE scientific abstracts we identified 182 novel potential pathologs. Of these, 36 were identified by computational tools only, 49 by human expert analysis only and 97 by both methods. These pathologs were related to neoplastic (53%), hereditary (24%), immunological (5%), cardio-vascular (4%), or other (14%), disorders.

Conclusions: Large scale genome projects continue to produce a vast amount of data with potential application to the study of human disease. For this potential to be realised we need intelligent strategies for data categorisation and the ability to link sequence data with relevant literature. This paper demonstrates the power of combining human expert annotation with FACTS, a newly developed bioinformatics tool, to identify novel pathologs from within large-scale mouse transcript datasets.

Background

The majority of common diseases such as cancer, allergy, diabetes or heart disease are characterised by complex genetic traits where genetic and environmental components contribute to disease susceptibility [1]. Unfortunately, our knowledge of genes contributing to the risk of common diseases remains limited. Consequently, a major goal of the post-genomic era is to better identify and characterise disease susceptibility genes and to use this knowledge for improved disease detection, treatment and prevention.

More than 500 genes are conserved across the invertebrate and vertebrate genomes [2]. Because of gene conservation, various organisms including yeast [3], fruitfly [4], zebrafish [5], rat [6], and mouse [7] have been used as genetic models for the study of human disease. Whilst the basic housekeeping genes such as those involved in metabolism, intracellular signalling, transcription/translation, DNA replication and repair are highly conserved in eukaryotes making them useful for the study of basic cellular processes and related diseases, these organisms do not share with humans many genes such as those involved in homeostasis, immunity, and cellular interactions [2]. Rodents and humans have remarkably similar genomes and share closely related biochemical, physiological and pathological pathways making the mouse the most important model organism for the study of human disease genetics and development of new treatments. This is reflected in the fact that approximately 80% of all mouse cDNA clones have matches in the human genome [2].

Genetic manipulations that can be performed in the mouse include point mutations, gene disruptions, insertions, deletions, or chromosomal rearrangements [8]. Random genome-wide mutagenesis can also be used for identification of gene function [9]. Specific genetic manipulations and alterations in the mouse often produce clinical features that are remarkably similar to human disease [10]. For example, targeted mutation of the transferrin receptor-2 gene was shown to induce haemochromatosis in mice [11]. The recent explosion of genomic data has, however, overwhelmed researchers with tens of thousands of novel genes making it difficult to know where to start in order to identify those most relevant to human disease. This has led to the use of comparative genomics as a strategy to identify promising candidates warranting further study from amongst all these novel genes.

Rubin *et al.* [12] compiled a list of 289 human disease genes and compared them to the fruitfly genome, finding 177 fruitfly orthologues to human disease genes. A more recent study [4] focused on identification of a subset of human disease genes that represent good candidates for study in the fruitfly model. Starting from the 929 entries of known human disease genes listed in the OMIM database [13], they identified 548 fruitfly genes with sequence homology to human disease genes. Of these, 56 genes belonged to well-known signalling pathways (such as *BMP*, *Hedgehog* or *Notch*). These strategies starting from known human disease-related genes are directed at the identification of orthologs in non-human species of known human disease genes.

The FANTOM2 project [14] focused on the functional annotation of 60770 cDNA RIKEN clones by large-scale, computerised annotation followed by manual curation. Being the most complete picture of the mouse transcriptome to date, the FANTOM2 dataset provides an ideal opportunity for the identification of novel pathologs thereby leading to the identification of novel human disease-related genes or disease-related gene products (transcript, anti-sense or protein), including candidates not listed in the OMIM Morbidmap database.

Recently, FANTOM2 cDNA clones were searched with TBLASTN (e-value: E-50) against a set of human disease-related genes and mouse orthologs were identified for 807 human disease-related genes [15]. Of these, 67 were novel mouse orthologs for known human disease-related genes [15]. However, this BLAST strategy starting from known human disease-related genes and then searching for orthologs in the mouse is only able to identify mouse genes corresponding to already known human disease-related genes. Consequently, for the present study we developed an alternative strategy for gene discovery from the FANTOM2 cDNA dataset aimed at identifying potential novel human disease-related genes. By comparison to previous reports [15,12,4], we started from novel mouse transcripts with similarity but not identity to human disease-related genes and then mapped these sequences back to the human genome to identify novel potential human disease-related genes.

The FANTOM2 dataset contains 2578 cDNA clones annotated as "similar to" known genes or proteins, comprising 1114 members of the representative transcript protein set 6.3 (RTPS6.3) FANTOM2 clusters [14]. The sequences of each of these "similar to" clones have 70–85% identity over more than 70% length to known reference protein or gene sequences. By searching the publication abstracts databases PubMed/MEDLINE [16], we identified 182 mouse mRNA transcripts that we called "potential pathologs" as they had sequence similarity to human disease-related genes or proteins. We defined "patholog" as a non-human gene with homology to a human gene that encodes a product (transcript, anti-sense or proteins) involved in human disease. A disease-related gene has a role in a patho-physiological pathway, or is relevant to the

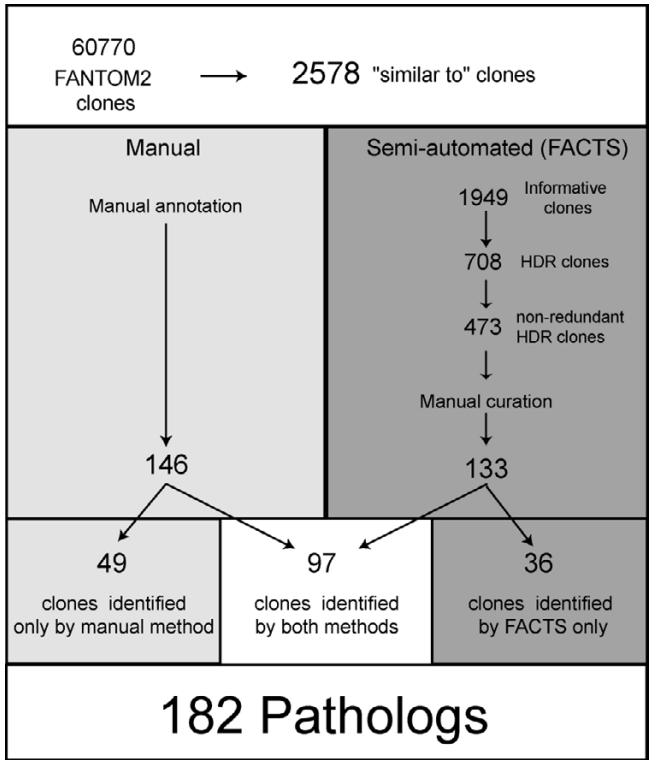

Figure 1
Flow chart of method used for the identification of "pathologs". Obtained from the FANTOM2 dataset, "similar to" clones were analysed using a manual (left) and a semi-automated approach (right) to identify "patholog" genes. HDR clones: clones with Human Disease Relationship.

diagnosis or treatment of a human disease. The most common disease classifications to which these potential pathologs corresponded were neoplastic, hereditary, immune, cardiovascular or neurological diseases. Each patholog represents a potential target for creating novel mouse models of human disease.

One of the bottlenecks in the use of genomic data to search for potential disease genes, is the time required to search the literature and assess the significance of the search results. Semi-automated knowledge extraction tools offer the potential to dramatically accelerate this process, albeit at the risk of some loss of information as a result of misqueries and ambiguous data. An important aspect of this project was a comparison of the performance of FACTS, a newly developed semi-automated knowledge extraction tool, against expert human annotators to determine whether in the future it will be feasible to automate the process of disease gene identification.

Results

Identification of novel pathologs

We identified 182 candidate pathologs from amongst 2578 FANTOM2 mouse "similar to" cDNA transcripts (Figure 1). Each of the transcripts representing these targets shows 70–85% identity over more than 70% of its length to a known human disease related gene or protein found by sequence similarity comparisons (see Methods). Of these, 146 were identified by manual and 133 by a semi-automated approach with 97 (53.3%) of targets being detected by both methods. The manual approach uniquely detected 49 (26.9%) human disease-related gene targets and semi-automated approach uniquely detected 36 (19.8%) targets.

Classification of pathologs by disease

The 182 "pathologs" were classified by the disease they relate to. The majority of the clones were related to neoplastic disorders (53%), followed by hereditary (24%), immunological (5%), cardio-vascular (4%), and other (14%), disorders (Table 1).

Cancer-related pathologs

The cancer-related category comprised the largest number of potential pathologs, with 96 unique targets (Table 2). Cancers represented in this list cover a variety of systems including the central nervous system, gastro-intestinal tract, breast, and prostate, among others. These potential pathologs related to cancer pathophysiology 61 (63.5%), diagnosis 19 (19.8%) and treatment 16 (16.7%).

Each of the 96 cancer-related potential pathologs was associated with one of the molecular circuits that maintain normal cell proliferation and homeostasis. Defects in these circuits often induce dysregulation of cell growth and apoptosis, or contribute to tissue invasion, metastasis, or angiogenesis. Defects in these pathways are thus central to cancer development [17]. Amongst the pathologs we identified genes encoding proteins involved in the SOS-Ras-Raf-MAPK cascade that has a key role in normal cell growth, and proteins linked with gene expression and cell proliferation, including Wnt-β Catenin, CdC42-Rac-Rho, and the pRb-E2F transcription factors [17]. A number of matrix metalloproteinases and cell adhesion molecules involved in cell invasion and metastasis were also identified.

Hereditary-disease pathologs

Potential pathologs related to hereditary diseases were the second largest group in this study (Table 3). We found 43 transcripts related to hereditary diseases, of which 39 (90.7%) were described to be defective or deleted in hereditary disorders and 4 (9.3%) were related to diagnosis. Defective pathways contributing to the pathogenesis of these diseases included metabolic pathways (e.g.

Table 1: Novel potential "pathologs" classified by type of human disorder and relationship to the disease process.

Disorder	Pathophysiology	Diagnosis	Treatment	Total
Cancer	61	19	16	**96**
Hereditary	39	4	0	**43**
Immunological	5	5	0	**10**
Cardio-vascular	8	0	0	**8**
Reproductive	6	0	0	**6**
Other	17	0	2	**19**
Total	**136**	**28**	**18**	**182**

Table 2: Cancer related pathologs. Representative disease is shown for each clone. * RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM.

Disease	*FANTOM ID*	DDBJ accession	Gene name	Disease relationship	OMIM status
CANCER					
Diagnosis					
	1110013A16*	AK003650	*Squamous cell carcinoma antigen 2*	Squamous cell carcinoma	1
	1300012C15*	AK004970	*Cargo selection protein tip47*	Gynecologic malignancies	2
	1600002M22*	AK005400	*Pregnancy-specific beta 1-glycoprotein*	Trophoblastic disease, tumour marker	1
	1810011L16*	AK007436	*Adamts-9 precursor*	Hereditary renal tumors	3
	2310046E09	AK009843	*Pancreatic secretory granule membrane major glycoprotein gp2 precursor*	Chronic lymphocytic leukemia	2
	2600013I19*	AK011199	*Diphthamide biosynthesis protein-2*	Prognosis of neoplastic diseases	2
	4631401E18*	AK019470	*Udp-n-acetyl-alpha-d-galactosamine:polypeptide n-acetylgalactosaminyl Transferase 7*	Colorectal carcinoma	2
	4930435F02	AK019597	*3-oxo-5-alpha-steroid 4-dehydrogenase 1*	Breast cancer	3
	4932411D20*	AK029960	*Ctcl tumor antigen se2-2*	T-cell based immunotherapy	3
	5330423N11*	AK077331	*Melastatin 2*	Cutaneous malignant melanoma	3
	9130023H10*	AK033603	*Mlze*	Melanoma	3
	9130413M24*	AK078964	*Ctcl tumor antigen se57-1*	T-cell based immunotherapy	3
	9330156N18*	AK034104	*Desmoglein 1 precursor*	Paraneoplastic pemphigus	2
	A030012E10*	AK037235	*Serine protease desc1 precursor*	Squamous cell carcinoma	3
	A230060D07*	AK038754	*Reverse transcriptase-like protein*	Chronic myelogenous leukemia	3
	A430037M23	AK039975	*Dipeptidyl-peptidase iii*	Endometrial neoplasms	2
	A630051G17	AK080312	*Meningioma-expressed antigen 6/11*	Meningioma and glioma	1
	E130307D12*	AK087504	*Scaffold attachment Factor b*	Breast cancer	1
	G430124K07*	AK090101	*Restricted expression proliferation associated protein 100*	Lung carcinoma	3
Pathophysiology					
	0610006O14	AK002260	*Vacuolar proton-atpase subunit atp6h*	Melanoma	2
	0610008P16*	AK002360	*Hp33 protein*	Hepatocellular carcinoma	3
	1110068E08*	AK004405	*Ku70-binding protein*	Gliomas	3
	1200003O15*	AK004587	*Proto-oncogene tyrosine-protein kinase fes/fps*	Leukaemia	1
	1500012D09	AK005230	*Ras-related protein rab-2*	Nasopharyngeal carcinoma	1
	1700001P03*	AK005620	*Homeobox transcription factor*	Colon cancer	1
	1700006L01		*Smac protein, mitochondrial precursor*	Multiple myeloma	2
	1700012B18*	AK005892	*Pregnancy-induced growth inhibitor*	Breast cancer	3
	1700045I19*	AK006700	*Hsd-4 protein*	Prostate cancer	2
	1810017F10*	AK007525	*Beta-casein-like protein*	Tumour-associated antigen	3
	2010003F10*	AK008064	*Transmembrane 4 superfamily, member 5*	Pancreatic cancer	1
	2210006M16*	AK008663	*Gasdermin*	Human gastric cancer cells	3

Table 2: Cancer related pathologs. Representative disease is shown for each clone. * RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM. *(Continued)*

2210007N23	AK019050	*Ca11 protein homolog*	Gastric carcinogenesis	1
2210412D05	AK008907	*Rho guanine nucleotide exchange factor 5*	Acute myeloid leukaemia	1
2210414K06	AK008928	*Nesh*	Cell metastasis and malignant transformation	2
2310016C08*	AK009377	*Hypoxia-inducible protein 2*	Cervix cancer	3
2510002J07*	AK010891	*1-acyl-sn-glycerol-3-phosphate acyltransferase beta*	Cancers and inflammation-associated diseases.	2
2610002I10	AK011289	*C-myc target jpo1*	Tumourigenesis	3
2610005L07	AK011323	*Cadherin*	Cancer development	1
2700048G21*	AK012392	*Antigen ny-co-8*	Colon cancer antigen	3
2810411G23	AK013085	*Tumor protein d54*	Breast cancer	1
2810425C21*	AK013153	*Death domain of death-associated protein kinase 1*	Non-small cell lung cancer	2
2810449N18*	AK013316	*Sirtuin 6*	Thyroid carcinoma	2
2900009D12	AK013503	*Succinate dehydrogenase*	Hereditary paraganglioma	1
4432412E01*	AK014491	*X-ray repair cross-complementing protein 3*	Bladder-cancer	2
4631410F01	AK028459	*Adamts-12 precursor*	Gastric carcinomas	3
4732440A06*	AK028704	*Calcium-activated chloride channel-2*	Breast cancer, metastasis	2
4930500E24*	AK019661	*Gas-2 related protein on chromosome 22*	Central nervous system tumours	3
4932436B18*	AK030081	*Pms1 protein homolog 1*	Prostate cancer	1
4933405E14*	AK016662	*Serologically defined colon cancer antigen 1*	Tumour suppressor	3
4933409E02*	AK016751	*Retinoblastoma-associated protein rap140*	Colon cancer cell line	3
5430417G24*	AK030670	*Headpin serine proteinase inhibitor*	Squamous cell carcinoma	1
5630400A09*	AK030708	*P63 protein*	Basal cell and squamous cell carcinomas	1
5730484M20*	AK077629	*Cell cycle checkpoint protein chfr*	Lung cancer	1
6430531D06*	AK032385	*Elks*	Capillary thyroid carcinoma	1
6430587E11	AK032541	*Copine vii*	Breast cancer	1
6820401K01*	AK033012	*Npat*	Cancer development	2
9330137N20	AK079068	*Hepatic leukemia factor*	Lymphoblastic leukaemia	1
9330200A01*	AK034492	*Ubiquitin specific protease*	Squamous non-small cell lung carcinoma	2
A130080E24*	AK038118	*C-myb protein*	Colon tumour	3
A430025E01*	AK039888	*P68 RNA helicase*	Colorectal tumours	2
A430075F05	AK040185	*Lipoma preferred partner*	Acute myeloid leukaemia	1
A730042E07	AK042943	*Serine/threonine protein phosphatase 2a, 72/130 kda regulatory subunit B*	Melanoma	2
A830098D13*	AK044188	*Megacaryocytic acute leukemia protein, isoform i*	Acute megakaryoblastic leukemias	1
A930033B01	AK020920	*Graf*	Hematopoeitic disorders	1
B130017M24	AK044984	*Hepatocellular carcinoma autoantigen*	Hepatocellular carcinoma	3
B230314N17	AK045848	*Deleted in lung and esophageal cancer 1*	Carcinogenesis	1
B930026D14*	AK047144	*Myc box dependent interacting protein 1*	Prostate carcinoma	3
B930095M03*	AK081171	*Frat2*	Gastric cancer	1
C130050F24*	AK048341	*Ranbp7/importin 7*	Colorectal carcinoma	2
C130062I06*	AK048462	*Fibrillarin*	Hepatocellular carcinoma	1
C230012L01	AK048706	*Androgen-induced prostate proliferative shutoff associated protein*	Prostate cancer	1
C630001O15	AK049821	*Malt1*	MALT lymphoma	1
D330038I09	AK052361	*10-formyltetrahydrofolate dehydrogenase*	Tumour cells	2
E030001H09*	AK086788	*Phd finger protein 3*	Glioblastoma multiforme	1
E030027H19*	AK087108	*Cub domain containing Protein 1*	Human colorectal cancer	3

Table 2: Cancer related pathologs. Representative disease is shown for each clone. * RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM. *(Continued)*

	E230037B21	AK054229	*Vault poly(adp-ribose) polymerase*	Several tumour types	3
	F630110I03	AK089273	*Matrix metalloproteinase-25 precursor*	Colon carcinomas or brain tumours	3
	F730035A01	AK089461	*Swi/snf complex 170 kda subunit*	Malignant rhabdoid tumours	1
	G630018E19*	AK090207	*Prostate cancer overexpressed gene 1*	Prostate cancer	1
	G630034H08*	AK090280	*Transcriptional repressor scratch*	Small cell lung cancer	3
Treatment					
	1100001P14	AK075618	*Beta-tubulin class iva isotype*	Human colon adenocarcinoma	3
	1200012D01*	AK004723	*Magic roundabout*	Angiogenesis	3
	1810010L20	AK075773	*Pituitary tumor-transforming gene 1 protein-interacting protein*	Pituitary adenomas	2
	2310039D24*	AK009689	*Carboxylesterase*	Solid tumours	1
	2600009M07	AK011174	*Polyamine modulated factor-1*	Antineoplastic activity	3
	2810459H04	AK013366	*Thrombospondin*	Angiogenesis	1
	3010033I09*	AK019405	*Alex1*	Tumours originating from epithelial tissue	1
	5730405M22*	AK077421	*Phosphoprotein enriched in astrocytes 15*	Glioma	2
	6030493E19	AK031701	*Melanoma antigen p15*	Melanoma	1
	6230424I22*	AK031785	*NUCLEAR MATRIX PROTEIN p84*	Tumor suppression	2
	A730016J02	AK042696	*Acetyltransferase tubedown-1*	Vascular and haematopoietic development	3
	B130023J22	AK045057	*Greb1b*	Breast cancer	3
	C920008O22	AK050594	*Retinoblastoma-binding protein 1*	Breast cancer	2
	D030050C19	AK083587	*Chronic myelogenous leukemia tumor antigen 66*	Leukemias and tumour cell lines	1
	D630010E08*	AK052639	*Carbonic anhydrase xii precursor*	Cancer tumour cells	1
	D830007E07*	AK052857	*Inositol hexakisphosphate kinase 3*	Ovarian cancer	2

peroxisomal biosynthesis and oxidation, mitochondrial respiratory chain, and phospholipid biosynthesis), cytoskeleton synthesis and organization and ion transport. Interestingly, 11 gene products from this group were previously described in the literature but their function was unknown or putative.

Other pathologs

All other potential pathologs (immunological, neurological, reproductive, cardiovascular, and others) have been summarised in table 4. The immunological disorders-related group comprised 10 transcripts. The majority of them represent genes involved in autoimmune diseases including systemic lupus erythematosus, rheumatoid arthritis, Sjogren's syndrome, sarcoidosis and Crohn's disease. Eight of the transcripts in this group encode proteins that have homology to known autoantigens. For neurological disorders, four potential pathologs related to Alzheimer's or Huntington's disease. Some neurological pathologs were also classified as hereditary disorder pathologs because of their Mendelian inheritance pattern. Six pathologs were related to reproductive disorders, eight to cardiovascular disorders (mainly hypertension), and 15 to other diseases.

Identification of novel pathologs

These 182 transcripts were further analysed to find those that by sequence comparison and conserved gene synteny corresponded to potential new pathologs, classified as "ortholog candidates" or "novel sequences" (Figure 2). We found 137 pathologs that represented the most similar mouse sequences to known human genes. Of these, 72 (52.5%) were found in public databases (NCBI and SPTR) as previously described mouse orthologs and their function is known or inferred. The remaining 65 (47.5%) represent the best mouse to human match by sequence similarity but their function is not known, making them excellent candidate mouse orthologs.

Of the 72 potential pathologs known to be mouse orthologs, 33 (45.9%) were related to neoplastic disorders, 23 (31.9%) to hereditary disorders and 16 (22.2%) corresponded to immunological, cardio-vascular, reproductive and other disorders. The majority of the 65 pathologs representing candidate mouse orthologs were related to cancer (37 or 57%). The remaining transcripts were related to the following disease categories: hereditary disorders 13 transcripts (20%), immunological 4 (6%), cardio-vascular 4 (6%), and other disease classification 7

Table 3: Pathologs related to hereditary disorders. Representative disease is shown for each clone. * RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM.

Disease	FANTOM ID	DDBJ accession	Gene name	Disease relationship	OMIM status
HEREDITARY					
Diagnosis					
	1700026F24	AK006381	*Neuronal protein 15.6*	Neurogenetic disorders	3
	2300002L19*	AK009012	*Chitotriosidase precursor*	Gaucher's disease	1
	2700028P07	AK012300	*14-3-3 protein tau*	Creutzfeldt-Jakob disease	3
	4732420G08*	AK028628	*Methionine synthase reductase*	Methionine synthase reductase deficiency	1
Pathophysiology					
	1010001M04*	AK003132	*Nadh-ubiquinone oxidoreductase 20 kda subunit, mitochondrial precursor*	Mitochondrial complex I deficiency	1
	1110019I12*	AK003819	*Selenoprotein n precursor*	Congenital muscular dystrophy	1
	4930414M06	AK005847	*Sterol carrier protein 2*	Peroxisomal D-hydroxyacyl-CoA dehydrogenase deficiency	1
	1810064C02	AK007951	*Sedlin*	Spondyloepiphyseal dysplasia tarda	1
	2310057L06	AK075908	*Tubulin-specific chaperone d*	Retinitis pigmentosa	2
	2410004F01*	AK010385	*Protoheme ix farnesyltransferase, mitochondrial precursor*	Charcot-marie-tooth disease	1
	2610205J09	AK011891	*Periodic tryptophan Protein 1*	Progressive myoclonus epilepsy	1
	2900072D10*	AK013765	*Sco2 protein homolog, mitochondrial precursor*	Cardioencephalomyopathy and a severe COX deficiency	1
	3110031I02*	AK014104	*N-wasp protein*	Wiskott-Aldrich syndrome	1
	4832440C16	AK029338	*Apical-like protein*	Ocular albinism type 1	1
	4930430B17	AK076748	*Machado-joseph disease protein 1*	Machado-joseph disease	1
	5830404H04	AK017896	*Protein c21orf2*	Autoimmune polyglandular disease type I	1
	6030476O14	AK031666	*Myoferlin*	Muscular dystrophy and cardiomyopathy	1
	6430516P20	AK032293	*Ceroid-lipofuscinosis neuronal protein 5*	Late infantile neuronal ceroid lipofuscinosis	1
	6430560A18	AK078275	*Caltractin, isoform 2*	Barth syndrome and chondrodysplasia punctata	2
	8030487I16	AK033295	*Gdp-fucose transporter 1*	Leukocyte adhesion deficiency II	1
	9330166I04*	AK034236	*Sialidase*	Sialidosis	1
	6720416P20*	AK032725	*Zinc finger protein 25*	MEN2a MEN2b	1
	9630046L06	AK036225	*Glycogen debranching enzyme*	Glycogen storage disease type III	1
	9930121L06*	AK037126	*Artemis protein*	Athabascan SCID	3
	A130054J05*	AK037846	*Nuclear localization signal protein absent in velo-cardio-facial patients*	Velo-cardio-facial syndrome	1
	A230074J06*	AK038912	*Nyctalopin*	X-linked congenital stationary night blindness	1
	A230090N11*	AK039054	*Cyld protein*	Cylindromatosis	1
	A630004L17	AK041354	*Transmembrane protein vezatin*	Deafness	3
	A730020L24	AK042745	*Alkyl-dihydroxyacetonephosphate synthase*	Zellweger syndrome	1
	A830020B12	AK043682	*Peroxisome assembly protein 10*	Peroxisome-biogenesis disorders	1
	A930007F16	AK044320	*Inositol polyphosphate 5-phosphatase ocrl-1*	Lowe syndrome	1
	A930014F04	AK044460	*Mitochondrial intermediate peptidase, mitochondrial precursor*	Friedreich's ataxia	1
	B230307C21*	AK045712	*Epilepsy holoprosencephaly candidate-1 protein*	Progressive myoclonus epilepsy	1
	B230311E17	AK045797	*Monocarboxylate transporter 5*	Mitochondrial myopathies	2

Table 3: Pathologs related to hereditary disorders. Representative disease is shown for each clone. * RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM.

B430307M20*	AK046679	*Ataxin 7*	Spinocerebellar ataxia type 7	1
C130020P08	AK047906	*Lowe oculocerebrorenal syndrome protein*	Oculocerebrorenal syndrome of Lowe	1
C330001M22*	AK049106	*Ubash3a protein*	Autosomal recessive deafness	2
C330016K18*	AK049248	*Sodium bicarbonate cotransporter isoform 1*	Proximal renal tubular acidosis associated with ocular abnormalities	1
C430015N23*	AK049478	*Y+l amino acid transporter 1*	Lysinuric protein intolerance	1
D030003E11	AK050684	*Beta-1,4-galactosyltransferase 7*	Progeroid type Ehlers-Danlos syndrome	1
E130016P05	AK084312	*T-box transcription factor tbx22*	Cleft palate	1
D630003K02*	AK085272	*Cytochrome b5 reductase b5r.2*	Methemoglobinemia	1
D630025L11*	AK052697	*Chorein*	Chorea-acanthocytosis	1

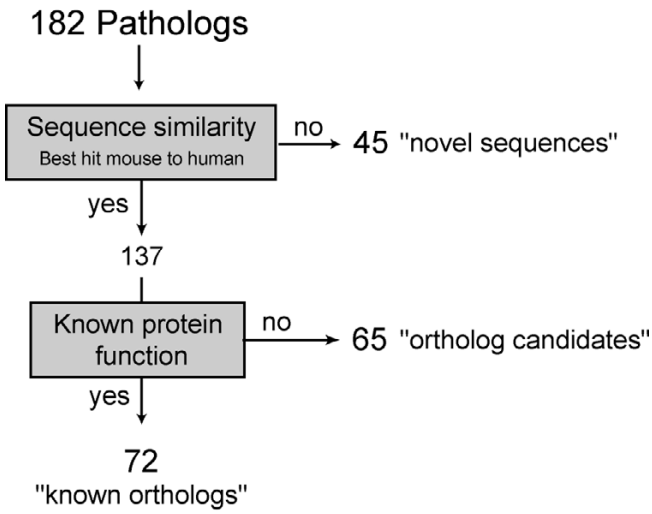

Figure 2
Flow chart of method used to classify "pathologs". To identify pathologs that correspond to already known mouse orthologs or potential new orthologs, cDNA sequences were compared to known human sequences and conservation of synteny assessed using mouse to human mapping information. If the patholog corresponded to best mouse to human hit the reported function of the gene product was checked. Mouse sequences with reported human ortholog and known function were classified as "known ortholog", sequences reported as best mouse to human hit with unknown function were classified as "ortholog-candidate" and sequences with unknown function that did not correspond to the best mouse to human hit were classified as "novel sequences".

(11%) which included neurological, haematological, reproductive, endocrine, and respiratory disorders.

Classification of candidate orthologs and novel homologs

We also located 45 potential pathologs not representing mouse orthologs of human genes, as there was a better mouse transcript match for the human gene they share sequence homology with. However, they may represent novel mouse homologs as deducted from sequence analysis and conservation of synteny.

These 45 potential pathologs with novel sequences, corresponded to cancer 26 (58%), hereditary disorders 7 (15%), cardiovascular disease 3 (7%) and other diseases 9 (20%). Nine of these targets had short sequences (less than 1000 bp) and might correspond to pseudogenes (based on gene synteny).

Comparison to Online Mendelian Inheritance in Man (OMIM) database entries

Genome-wide studies of pathologs in other organisms [2,4,18] focused on systematic searches for paralogs or orthologs of human disease genes in the respective genomes. In those studies pathologs in different organisms were detected using the OMIM database [19] that contains entries on hereditary human disorders. We were interested, however, in identifying potential pathologs involved in both inherited and non-inherited diseases and consequently elected to use a broader search strategy focusing on sequence analysis combined with key-word searching of literature abstracts based on annotated gene names and MeSH terms. Scientific abstracts listed in PubMed were searched to identify human disease-related genes or proteins related to our dataset of "similar to" FANTOM2 clones. In a comparison between the disease genes listed in OMIM and those detected using PubMed, we found that out of the 182 potential pathologs we identified in this project, 128 (70.3%) were listed in OMIM, but only 89 were listed as having a disease relationship (Table 5). For the remaining 39 pathologs either

Table 4: Pathologs related to other disorders. Representative disease is shown for each clone. *RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM.

Disease	FANTOM ID	DDBJ accession	Gene name	Disease relationship	OMIM status
IMMUNOLOGICAL					
Diagnosis					
	1500019E10*	AK005285	*Replication protein a 14 kda subunit*	Systemic lupus erythematosus	1
	1810019E15	AK007546	*Dek protein*	LES and Juvenile RA	2
	2700059D02*	AK012454	*Uveal autoantigen*	Behecet's Disease, sarcoidosis and Vogt-Koyanagi-Harada disease	3
	4632415P04*	AK028516	*Golgi complex autoantigen golgin-97*	Sjogren's syndrome	1
	D030032G01*	AK050903	*Neuroblast differentiation associated protein ahnak*	Systemic lupus erythematosus	2
Pathophysiology					
	A630006E02*	AK041380	*Minor histocompatibility antigen ha-1*	Graft-versus-host disease	1
	B230303A05*	AK045681	*U1 small nuclear ribonucleoprotein c*	Autoimmunity to U1 snrnps	1
	B230397K24*	AK046480	*L1 retroposon, orf2 mrna*	Rheumatoid arthritis	3
	C330006A15*	AK049133	*Ribonuclease p protein subunit p38*	Scleroderma autoimmune antigens	1
	F830032C23*	AK089843	*Caspase recruitment domain protein 15*	Crohn's disease	1
NEUROLOGICAL					
Pathophysiology					
	2210420D18		*Serine/threonine kinase rick*	Alzheimer's disease	2
	4833420A15	AK014731	*Huntingtin-interacting protein-1 protein interactor*	Huntington's disease	1
	9330170I02*	AK034263	*Metabotropic glutamate receptor 2 precursor*	Alzheimer's disease	2
	B230307E07	AK045716	*Excitatory amino acid transporter 1*	Alzheimer's disease	3
CARDIOVASCULAR					
Pathophysiology					
	1700127D06*	AK007298	*Tisuee kallilrein*	Hypertension	1
	2510048K03*	AK011112	*Prolylcarboxypeptidase*	Essential hypertension	1
	4833405G23	AK029362	*Rtp801-like protein*	Ischemic diseases	3
	9630044I02	AK036182	*Mitochondrial isoleucine trna synthetase*	Cardiomyopathy	3
	B430208E24*	AK046620	*Ras gtpase-activating Protein 1*	Neuropathology of ischemia	2
	D130064D17*	AK051677	*Lysosomal pro-x carboxypeptidase precursor*	Essential hypertension	3
	E030024D09*	AK087056	*Angiotensin converting enzyme*	Hypertension	1
	2310063B19*	AK010021	*Epoxide hydrolase*	Hhypertension	2
REPRODUCTIVE					
Pathophysiology					
	0610007H07*		*Z-protein*	Fetal loss	3
	1700010P14*	AK005852	*Nyd-sp6*	Spermatogenesis	3
	4631410O16*	AK028461	*Sumo-1-specific protease 1*	Reproduction	1
	4930406H24*	AK029590	*Adam 26 precursor*	Spermatogenesis	3
	B130010I06*	AK044891	*Dmrt2/terra-like protein*	Sex differentiation disorders	3
	D030049L20	AK050982	*Sperm antigen*	Immunologic infertility	1
OTHERS					
Pathophysiology					
	1110004H01*	AK003421	*Mitochondrial import inner membrane translocase subunit tim9 b*	Fracture healing	3
	1810015E19	AK007508	*Slp-1*	Autism	1
	1810015M19*	AK019013	*Lw glycoprotein*	Sickle cell disease	1
	4932434G09*	AK016547	*Ribonucleases p/mrp protein subunit pop1*	Connective tissue diseases	3
	5730465G20	AK077598	*N-acetyllactosaminide beta-1,6-n-acetylglucosaminyl Transferase*	Blood group I gene	3
	9830131G07*	AK036537	*Bomapin*	Hematopoiesis	2
	9930031F20	AK036967	*Vp165*	Type 2 diabetes	3

Table 4: Pathologs related to other disorders. Representative disease is shown for each clone. *RTPS6.3 (representative transcript protein set 6.3) cluster representative transcriptional unit (TU) of the FANTOM2 clone set. OMIM status: 1 = gene present in OMIM with a reported disease; 2 = gene present in OMIM with different disease association or without disease; 3 = gene not present in OMIM.

	A130042M24*	AK037730	*Mucin*	COPD	1
	A730076H11	AK043253	*T-cell receptor alpha chain precursor v-j region*	Leishmania major	3
	A830007N20	AK043557	*Wd-repeat protein 3*	Triple-A syndrome	1
	B830002B15*	AK046772	*Polycystic kidney disease 2-like protein*	Cystic diseases	1
	C230099M23*	AK082743	*Vesicular glutamate transporter 2*	Schizophrenia	3
	E130216C05	AK087459	*Amyloid beta precursor-like protein 2*	Healing corneal epithelium	2
Treatment					
	5033405N08*	AK017155	*Agmatinase*	Chronic pain, addictive states and brain injury	3
	A930028N13	AK044634	*Ankyrin-2*	Human hemolytic anemias	2

no disease relationship was listed in OMIM or the disease association listed was not the same as the one found by our manual expert curation. Furthermore, through our search strategy of PubMed abstracts we identified 93 additional potential pathologs not identified through OMIM.

Discussion

The mouse is the most important animal model of human disease, hence the importance of the FANTOM project to characterise the mouse transcriptome, complete with functional annotation and human genome mapping. The FANTOM2 cDNA dataset represents the most complete set of mouse transcripts to date, and it was utilised by us to identify potential novel pathologs. The identification of pathologs was assisted by integrating the FANTOM2 mouse data with all scientific literature referenced by PubMed, which is currently the most comprehensive literature source of molecular and clinical data. The problem is that most relevant data in medical literature databases is embedded in the free text and searching by automated methods often results in the loss of information. Therefore, to more thoroughly screen for potential pathologs, we employed two approaches in parallel; one relying on semi-automated sequence analysis and text searching (FACTS) and the other relying on human expert manual searching. The results of this study clearly indicate the importance of using multiple parallel approaches to identify all potential pathologs.

The semi-automated approach detected 133 (73%) of the potential pathologs compared to 146 (80%) using manual search. Interestingly the overlap between the two methods was only 97 (53%), suggesting that both approaches are required for identification of all potential pathologs. Although the semi-automated approach utilises less than one third of the time required by manual searching, three quarters of the hits detected by this system were classified as false positives, only 134 transcripts out of the initial 708 produced by automated search meeting the criteria for potential pathologs. This is not unusual when using computerized systems. Problems were caused by retrieval of irrelevant abstracts, misconstructed queries, queries containing ambiguous gene symbols or synonyms, wrong disease MeSH term associations in the abstracts or because the abstract did not meet the human expert's criteria for a potential patholog.

Several reasons contribute to a better performance using the FACTS system compared to expert annotation. The coverage and specificity of abstract retrieval from MEDLINE depends on how queries are constructed. Manual searches were performed using gene names and symbols from the FANTOM2 database, while FACTS constructed queries from an automated QueryMaker program that extracts gene/protein names, symbols and synonym accessions of their annotation sources (e.g., MGI, SwissProt). This information is integrated according to query rules and then used to perform MEDLINE searches. The FACTS system also combines a MeSH TermMatcher program with a Sentence Splitter system to identify disease associations from MEDLINE abstracts and OMIM morbidmap database (for detailed explanation of the FACTS system see [20]). These programs enhance the accuracy of automated queries and searches used for the identification of pathologs. The high frequency of false positive hits makes manual curation an important step when using computational screening. Whilst manual searching produces more true positive hits, it is less efficient than the semi-automated approach. Expert analysis identified 49 clones that were missed by the automated system as FACTS-derived results are based on MEDLINE whilst the expert annotators used PubMed for abstract searches.

In previous reports, pathologs in non-human organisms were identified using the OMIM database, the problem

Table 5: Comparison of the OMIM entries (July 2003) with pathologs. "OMIM NDA" stands for pathology entries that are in OMIM, but disease association was not specified, or it was not consistent with the disease specified in PubMed abstracts. "OMIM DA" stands for pathologs that match both OMIM entries and disease association.

Disease	OMIM NDA	Not in OMIM	OMIM DA	Total
Cancer	26	33	37	96
Hereditary disorders	4	4	35	43
Other	9	17	17	43
Total	39	54	89	182

with this approach being that it requires the human disease gene to be already known. Our approach produced 93 potential pathologs that were identified through the scientific literature but were not in the OMIM database, suggesting that the true number of pathologs is far higher than those with strictly Mendelian inheritance. Furthermore, given that this study only focused on the subset of "similar to" cDNA clones and did not cover those annotated as "weakly similar to" (see methods) we anticipate that there are many more pathologs in the mouse that are yet to be identified. Our analysis also suggests that the field of disease-related molecular databases is underserved, other than the Mendelian disorders covered by OMIM.

The pathologs identified in this study were selected from a group of FANTOM2 mouse cDNA clones similar to, but not identical to, other known genes. As expected, sequence comparison revealed that the majority of pathologs (137 or 75%) corresponded to the best mouse to human match although many of them (65) remain to be confirmed as orthologs as no function for them has yet been described. We also located an extra 45 potential pathologs that may represent mouse homologs to novel human disease-related genes as deduced from sequence analysis and conservation of synteny. It is likely that at least some of the potential pathologs with function unknown (110) will represent non-functional transcripts, or gene products with different function. Those pathologs that are experimentally validated as orthologs can be used as targets for genetic manipulation and development of mouse models of human disease.

Conclusions

This paper demonstrates the power of combining human expert annotation with FACTS, a newly developed bioinformatics tool, to identify novel pathologs from within large-scale mouse transcript datasets. Those pathologs can be used as targets for genetic manipulation and development of mouse models of human disease. The similarity between mouse and human genomes and their closely-related biochemical, physiological, and pathological pathways makes the mouse an invaluable model organism for the study of human disease.

Methods

FANTOM2 system

The FANTOM2 set of full-length mouse cDNA clones contains 60770 sequences. The FANTOM2 clones were functionally annotated using automated computational annotation followed by expert human curation [14].

Accession numbers

Accession numbers in the manuscript refer to FANTOM accessions submitted to the DNA data bank of Japan (DDBJ), or public accessions.

Sequence analysis

Pre-computed results of sequence similarity comparisons were retrieved from the FANTOM2 database [21]. The method used for detection of sequence similarity has been explained by Okasaki et. al. [14]. Briefly, according to the percentage of DNA sequence identity and the length of the similarity region to known genes the FANTOM2 clones were annotated as: "identical-to" or "homolog", "similar-to", or "weakly-similar-to". "Identical to" (>95%) and "homolog" (85–95%) were clones with more than 85% identity over more than 90% of their length to known genes. "Similar to" were clones with identity of 70–85% over more than 70% of their length to known genes. "Weakly similar to" were clones with identity between 50–70% over more than 70% of their length to known genes.

The clones grouped as "similar to" and "weakly similar to" could represent novel mouse transcripts whose function may be inferred because of their similarity to known proteins. This study focused on the analysis of "similar to" clones, which are referred here to as the "target set". The clones classified as "weakly similar to" require further bioinformatic characterisation and therefore will be matter of a different study. The target set was comprised of 2578 annotated clones, representing a workable size subset for this study. Using the RIKEN clone ID number of each potential human disease related target, we identified the

representative transcript from RTPS 6.3 [22] to indicate the FANTOM2 cluster representative transcriptional unit associated with disease (see Tables 2, 3, 4).

Human disease-related genes

We defined "patholog" as a non-human gene with homology to a human gene that encodes a product (transcript, anti-sense or proteins) involved in human disease. In this study, to be classified as a disease-related gene, there must be at least one scientific publication providing evidence linking a gene (or the related protein) to a disease phenotype (such as protein mutation or up/down regulation), to a diagnostic test, or to a disease treatment. *In vitro* studies using human cells (fresh tissue, cell lines or tumour cell cultures) or clinical studies were all accepted as evidence for a human disease relationship. Scientific publications where experiments were done using non-human organisms or where results were not tested directly in humans were discarded from the analysis.

All potential pathologs from the target set were used for identification of the corresponding human gene by mapping to the human genome sequence. We used a semi-automated and a manual approach for data searching and identification of pathologs. The manual approach involved searching literature abstracts from the PubMed database [23] using protein names for each clone in the target set, to identify potential human disease relationships. The gene or protein name was searched via the PubMed interface for keyword search and the retrieved abstracts were analysed by medical experts. Queries that returned one or more abstracts and that met the patholog definition criteria were noted: clone ID, clone name, PubMed ID, and disease-relationship were recorded. In the semi-automated approach we used the FACTS (Functional Association/Annotation of cDNA clones from Text/Sequence Sources) system to query MEDLINE abstracts (described in detail by Nagashima *et al.* [20]. Briefly, we constructed MEDLINE queries from RIKEN cDNA clone annotations using 205 query construction rules and the FACTS MeSH TermMatcher program. Of 2578 similar to annotated clones 1,949 clones had gene names considered informative for MEDLINE abstract searches that were clustered into 639 queries. 522 queries corresponding to 708 clones yielded 17,051 abstracts with 2637 disease MeSH terms. As FACTS extracts both abstract and sequence-derived based information using accession mapping, from the 708 clones we obtained 109 that had 92 disease associations in OMIM Morbimap. From 629 clones without informative names we extracted 47 OMIM Morbidmap associations for 57 clones. In total FACTS provided 27% of all and 36% of informative disease candidate associations. The MEDLINE and OMIM inferred disease associations can be annotated upon registration through a FACTS annotation interface. The interface displays basic clone information (symbol names, protein motifs and RTPS cluster information) and links to tissue expression information in READ [24] and GNF gene expression atlas [25] together with the automatically constructed query. The computationally inferred human disease MeSH terms and OMIM Morbidmap titles are listed in a table containing a hyperlinked MEDLINE identifier, MeSH term and check boxes to delete or confirm the MeSH term and assign a confidence value. The confidence values low, medium, high, and unknown indicate whether the MeSH assignment is based on direct (e.g. mutation in gene) or indirect (pathway component in disease gene pathway) evidence. A comment field provided the possibility of entering evidence and decision-supporting comments. Automated results were obtained in 48 hours and manual curation required approximately 60 man hours.

Medical experts performed manual searches of the 2578 target clones through abstract inspection and thereby selected candidate novel mouse pathologs. The time taken to identify the final number of pathologs required approximately 160 man hours.

Classification and interpretation

The results of the manual and the semi-automated approaches were combined in a single final list. Clones on this list were classified into groups in accordance to the physiological system affected by the related disease. Pathologs were subdivided according to the role of the protein in the disease process: pathophysiology, diagnosis, or treatment. Finally, we compared the pathologs identified in this study with entries from the OMIM database to identify pathologs that could be identified by direct searching of the OMIM database.

Identification of mouse known orthologs, ortholog-candidates and novel sequences was based on sequence similarity (FANTOM2 website [21]) and conservation of synteny based on mouse to human mapping information (NCBI Map viewer [26]) and RIKEN-genomapper [27] (July 2003), and reported function (Locus Link [28] search – july 2003). Mouse sequences with reported human ortholog and known function were classified as "known orthologs", sequences reported as best mouse to human match with unknown function were classified as "ortholog-candidates" and sequences with function unknown that did not correspond to the best mouse to human match were grouped as "novel sequences".

Authors' contributions

DS, CS, VB, LS, NP carried out the gene annotation and expert curation and drafted the manuscript. TN and CS designed and created the FACTS system. DS and NP par-

ticipated in the annotation of its entries. All authors read and approved the final manuscript.

Acknowledgements

DS is the recipient of a scholarship from the Canberra Hospital Salaried Specialists Private Practice Fund.

References

1. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies.** *Genet Med* 2002, **4:**45-61.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
3. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW: **Systematic screen for human disease genes in yeast.** *Nat Genet* 2002, **31:**400-404.
4. Reiter LT, Potocki L, Chien S, Gribskov M, Bier E: **A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster.** *Genome Res* 2001, **11:**1114-1125.
5. Penberthy WT, Shafizadeh E, Lin S: **The zebrafish as a model for human disease.** *Front Biosci* 2002, **7:**d1439-53.
6. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A, Eppig J, Maltais L, Maglott D, Schuler G, Jacob H, Tonellato PJ: **Rat Genome Database (RGD): mapping disease onto the genome.** *Nucleic Acids Res* 2002, **30:**125-128.
7. Hamilton BA, Frankel WN: **Of mice and genome sequence.** *Cell* 2001, **107:**13-16.
8. Muller U: **Ten years of gene targeting: targeted mouse mutants, from vector design to phenotype analysis.** *Mech Dev* 1999, **82:**3-21.
9. Zambrowicz BP, Friedrich GA: **Comprehensive mammalian genetics: history and future prospects of gene trapping in the mouse.** *Int J Dev Biol* 1998, **42:**1025-1036.
10. Rossant J, McKerlie C: **Mouse-based phenogenomics for modelling human disease.** *Trends Mol Med* 2001, **7:**502-507.
11. Fleming RE, Ahmann JR, Migas MC, Waheed A, Koeffler HP, Kawabata H, Britton RS, Bacon BR, Sly WS: **Targeted mutagenesis of the murine transferrin receptor-2 gene produces hemochromatosis.** *Proc Natl Acad Sci U S A* 2002, **99:**10653-10658.
12. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S: **Comparative genomics of the eukaryotes.** *Science* 2000, **287:**2204-2215.
13. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30:**52-55.
14. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CA, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
15. Schriml LM, Hill DP, Blake JA, Bono H, Wynshaw-Boris A, Pavan WJ, Ring BZ, Beisel K, Setou M, Okazaki Y: **Human disease genes and their cloned mouse orthologs: exploration of the FANTOM2 cDNA sequence data set.** *Genome Res* 2003, **13:**1496-1500.
16. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of the National Center for Biotechnology Information: 2002 update.** *Nucleic Acids Res* 2002, **30:**13-16.
17. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100:**57-70.
18. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, Gaasterland T, Gissi C, King B, Kochiwa H, Kuehl P, Lewis S, Matsuo Y, Nikaido I, Pesole G, Quackenbush J, Schriml LM, Staubli F, Suzuki R, Tomita M, Wagner L, Washio T, Sakai K, Okido T, Furuno M, Aono H, Baldarelli R, Barsh G, Blake J, Boffelli D, Bojunga N, Carninci P, de Bonaldo MF, Brownstein MJ, Bult C, Fletcher C, Fujita M, Gariboldi M, Gustincich S, Hill D, Hofmann M, Hume DA, Kamiya M, Lee NH, Lyons P, Marchionni L, Mashima J, Mazzarelli J, Mombaerts P, Nordone P, Ring B, Ringwald M, Rodriguez I, Sakamoto N, Sasaki H, Sato K, Schonbach C, Seya T, Shibata Y, Storch KF, Suzuki H, Toyo-oka K, Wang KH, Weitz C, Whittaker C, Wilming L, Wynshaw-Boris A, Yoshida K, Hasegawa Y,

Kawaji H, Kohtsuki S, Hayashizaki Y: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409:**685-690.
19. **Online Mendelian Inheritance in Man (OMIM) database** [http://www.ncbi.nlm.nih.gov/omim/]
20. Nagashima T, Silva DG, Petrovsky N, Socha LA, Suzuki H, Saito R, Kasukawa T, Kurochkin IV, Konagaya A, Schonbach C: **Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS.** *Genome Res* 2003, **13:**1520-1533.
21. **FANTOM2 website** [http://fantom2.gsc.riken.go.jp/]
22. Baldarelli RM, Hill DP, Blake JA, Adachi J, Furuno M, Bradt D, Corbani LE, Cousins S, Frazer KS, Qi D, Yang L, Ramachandran S, Reed D, Zhu Y, Kasukawa T, Ringwald M, King BL, Maltais LJ, McKenzie LM, Schriml LM, Maglott D, Church DM, Pruitt K, Eppig JT, Richardson JE, Kadin JA, Bult CJ: **Connecting sequence and biology in the laboratory mouse.** *Genome Res* 2003, **13:**1505-1519.
23. **PubMed database** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]
24. Bono H, Kasukawa T, Hayashizaki Y, Okazaki Y: **READ: RIKEN Expression Array Database.** *Nucleic Acids Res* 2002, **30:**211-213.
25. **GNF gene expression atlas** [http://expression.gnf.org/]
26. **FANTOM2 website** [http://fantom2.gsc.riken.go.jp/]
27. **NCBI Map viewer** [http://www.ncbi.nlm.nih.gov/mapview/]
28. **RIKEN-genomapper** [http://fantom21.gsc.riken.go.jp/GenoMapperMm/]
29. **Locus Link** [http://www.ncbi.nlm.nih.gov/LocusLink/]