

Analysis of genome-wide, cancer-associated mutation datasets in mouse and human



Jenny Mattison

Wellcome Trust Sanger Institute

Trinity College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

September 2008

Declaration

This thesis describes work carried out between April 2005 and September 2008 under the supervision of Dr Tim Hubbard and Dr David Adams at the Wellcome Trust Sanger Institute, while member of Trinity College, University of Cambridge. This dissertation is the result of my own work and contains nothing that is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. No part of this dissertation nor anything substantially the same has been, or is being, submitted for any qualification at this or any other university. This dissertation does not exceed the page limit set by the Biology Degree Committee.

Jenny Mattison

September 2008

Abstract

The complexity of human cancer genomes complicates the identification of those mutations that drive the tumourigenic process. Integrative analyses, particularly cross-species comparisons, provide a means of distinguishing likely driver mutations from the background of passenger mutations that arise in unstable cancer genomes. This thesis describes the analysis of human and mouse experimental datasets to identify human cancer gene candidates.

In mice, candidate cancer genes can be ‘tagged’ using insertional mutagens such as retroviruses and transposons. The analysis of more than 1,000 mouse tumours generated by insertional mutagenesis is described. Insertion sites are mapped to the mouse genome and are used to identify candidate cancer genes. The distribution of insertions within and around candidate genes is analysed to predict the likely mechanisms of mutagenesis and, therefore, the possible structure and function of the mutated gene products. Candidates are also characterised by comparison with other human and mouse cancer-associated mutation datasets, and co-operating cancer genes are identified in an attempt to better understand cancer gene pathways.

The mouse insertional mutagenesis results are then compared to genome-wide copy number data for human cancers. The Wellcome Trust Sanger Institute has generated comparative genomic hybridisation (CGH) data for ~700 human cancer cell lines using the Affymetrix 10K SNP array and, more recently, for ~600 human cancer cell lines using the high resolution Affymetrix SNP 6.0 array. Regions of copy number change in human cancers often encompass many genes, and it can be difficult to determine which genes contribute to the cancer phenotype. In this thesis, the human CGH data are processed into regions of copy number change and the mouse candidate cancer genes identified by retroviral insertional mutagenesis are used to narrow down the candidates in amplicons and deletions. The over-representation of mouse candidate oncogenes in regions of copy number gain suggests that a significant proportion of genes contributing to retrovirus-induced tumourigenesis in the mouse are also amplified in, and contribute to the development of, human cancers. Candidate oncogenes and tumour suppressor genes that are recurrently mutated in both human tumours and murine lymphomas are identified as strong candidates for a role in tumourigenesis.

Acknowledgements

First and foremost I would like to thank my supervisors Tim Hubbard and David Adams for giving me the opportunity to carry out this project, and for all their advice, support and encouragement along the way. Many thanks also to my thesis committee, Paul Edwards, Andy Futreal and Julian Parkhill, for their guidance and advice. The project would not have been possible without the work of Anton Berns' lab at the Netherlands Cancer Institute (NKI), who generated the mouse tumours in the retroviral insertional mutagenesis screen, and the Wellcome Trust Sanger Institute Cancer Genome Project (CGP), led by Mike Stratton and Andy Futreal, who generated all of the copy number data used in this thesis. At the NKI, I would like to thank Anthony Uren, Jaap Kool, Jos Jonkers, Lodewyk Wessels, Maarten van Lohuizen and Anton Berns for enabling me to participate in their exciting study, and for providing me with the data that I needed for my project. I am particularly grateful to Anthony Uren and Jaap Kool for helping me to understand the intricacies of insertional mutagenesis, and Jeroen de Ridder for performing the statistical analysis required to identify common insertion sites in the data. Also thanks to Anthony Cox and Daoud Sie for discussions and advice on mapping the insertions. Among the CGP team, I am particularly grateful to Richard Wooster and Adam Butler for giving me access to the copy number data, and Graham Bignell for helping me to understand it. Thanks also to Chris Greenman for his helpful statistical advice, and Helen Davies, who provided the *TP53* and *CDKN2A* mutation statuses for the human cancer cell lines. For the *Sleeping Beauty* data, I extend my gratitude to Lara Collier at the University of Minnesota, who performed the screen. Many thanks also to Fanni Gergely at the Cambridge Research Institute, who carried out the functional validation of *QSK*. I would also like to thank all members of the Hubbard Research Group, past and present, who have opened my eyes to a strange new world and have been a pleasure to work with. Thanks particularly to Matias Piipari for extracting the information I needed from TRANSFAC, and Andreas Prlic for general IT support! On a personal note, I thank my parents and my sister Laura for lavishing affection on me, and my grandma Ethel, whose enthusiasm for science and research is as strong as ever at the age of 96. I would also like to acknowledge the little person-to-be who has made thesis writing something of a challenge, but whom I love already. Finally, I thank my husband Simon for supporting me in my endeavour, and particularly for putting up with me over the last few months. I guess it is my turn to vacuum.

Table of Contents

Chapter 1 Introduction.....	1
1.1 Outline of introduction	1
1.2 An introduction to cancer	1
1.2.1 Definition and classification.....	1
1.2.2 Epidemiology	2
1.2.3 The multi-stage theory of carcinogenesis.....	4
1.2.4 The hallmarks of cancer	7
1.2.5 Cancer genes	7
1.2.6 Pathways in cancer	10
1.2.7 Treatment of cancer.....	13
1.3 Genome-wide approaches for human cancer gene discovery.....	14
1.3.1 Gene resequencing	14
1.3.2 Gene expression profiling	17
1.3.3 Copy number analysis	18
1.3.4 Epigenetic profiling.....	29
1.3.5 Genome-wide mapping of transcription factor binding sites.....	31
1.4 Cancer gene discovery in the mouse	34
1.4.1 The mouse as a model for studying cancer	34
1.4.2 Forward genetic screens in the mouse.....	39
1.5 Cross-species comparative analysis for cancer gene discovery	58
1.6 Aims of this thesis	61
Chapter 2 Identifying insertion sites and candidate cancer genes by insertional mutagenesis in the mouse	63
2.1 Introduction	63
2.2 Description of the datasets.....	64
2.2.1 The retroviral dataset.....	64
2.2.2 The Sleeping Beauty dataset	69
2.2.3 Known cancer genes in the Cancer Gene Census	70
2.3 Mapping the sequence reads using SSAHA2.....	70
2.4 Accounting for unmapped reads.....	73
2.5 Filtering the mapped reads	79

2.6	Identification and filtering of insertion sites.....	82
2.7	Estimating the coverage of the mutagenesis screens.....	87
2.8	Analysis of the distribution of insertions around mouse genes.....	90
2.9	Assigning insertions to genes.....	96
2.10	Identifying statistically significant common insertion sites.....	96
2.10.1	Monte Carlo simulations.....	97
2.10.2	Kernel convolution.....	99
2.10.3	Final set of candidate genes.....	102
2.11	Discussion.....	104

Chapter 3 Analysis of mouse candidate cancer genes identified by insertional mutagenesis..... 107

3.1	Introduction.....	107
3.2	Comparative analyses between the insertional mutagenesis data and other cancer-related datasets.....	108
3.2.1	Description of the datasets.....	108
3.2.2	Comparison with insertional mutagenesis data.....	111
3.3	Comparison of candidate cancer genes in the MuLV and Sleeping Beauty datasets.....	119
3.4	Determining the mechanisms of MuLV insertional mutagenesis.....	125
3.4.1	Analysing the distribution of intragenic insertions.....	125
3.4.2	Analysing co-occurring insertions in candidate genes disrupted by MuLV and T2/Onc.....	132
3.4.3	Identification of tumour suppressor genes inactivated by MuLV.....	138
3.4.4	Identifying retroviral insertions in regulatory features.....	142
3.4.5	Expression analysis of MuLV-induced tumours.....	145
3.5	Identification of co-operating cancer genes in the MuLV dataset.....	148
3.5.1	Genotype-specific cancer genes.....	148
3.5.2	Co-occurrence and mutual exclusivity of disrupted genes.....	153
3.6	Discussion.....	157

Chapter 4 Using mouse candidate cancer genes to narrow down the candidates in regions of copy number change in human cancers..... 161

4.1	Introduction.....	161
4.2	Description of the datasets.....	162

4.2.1	Mouse candidate cancer genes identified by retroviral insertional mutagenesis.....	162
4.2.2	Copy number data for human cancer cell lines.....	164
4.2.3	Copy number variants (CNVs).....	166
4.3	Processing the copy number data.....	166
4.4	Characterising gains and losses in cancer genomes.....	173
4.5	Comparative analysis of mouse candidate cancer genes and CGH data from human cancers.....	175
4.5.1	Global comparison.....	175
4.5.2	Identification of individual candidates for a role in human cancer.....	190
4.6	Comparison of methods for calling gains and losses.....	208
4.7	Global comparison of mouse candidate cancer genes and human CNVs.....	215
4.8	Discussion.....	218
Chapter 5 Identifying human cancer genes in high-resolution copy number data		223
5.1	Introduction.....	223
5.2	Description and processing of the datasets.....	224
5.2.1	High-resolution copy number data.....	224
5.2.2	Additional datasets.....	229
5.3	Comparative analysis of human high-resolution CGH data versus mouse insertional mutagenesis data.....	229
5.3.1	Global comparison.....	229
5.3.2	Identifying individual cancer gene candidates.....	232
5.4	Comparison between high-resolution and 10K CGH data.....	253
5.5	Identification of co-operating cancer genes.....	264
5.5.1	Genotype-specific cancer genes.....	264
5.5.2	Co-occurrence of amplified and deleted candidate cancer genes.....	267
5.6	Discussion.....	273
Chapter 6 Summary and Conclusions.....		276
References.....		284
Appendices.....		337
Publication.....		365

List of Figures

1.1	Summary of cancer incidence in 2004 and deaths from cancer in 2005 for the most common sites of cancer in males and females in the UK.	3
1.2	The clonal evolution of cancer.	5
1.3	Mutations in different genes in the same pathway can have an equivalent effect.	11
1.4	Array design and whole-genome sampling assay for the Affymetrix SNP array.	20
1.5	End sequence profiling of tumour DNA.	28
1.6	Overview of ChIP-PET for mapping transcription factor binding sites.	32
1.7	Generation of a conditional knockout allele in ES cells.	37
1.8	Structure of a retroviral provirus.	41
1.9	The mechanisms of mutagenesis of murine leukaemia virus include enhancer mutation, promoter mutation and premature termination of gene transcription.	43
1.10	Isolation of retroviral insertion sites by inverse PCR and splinkerette PCR.	47
1.11	Structure of the <i>Sleeping Beauty</i> transposon.	54
2.1	Workflow for identifying mouse candidate cancer genes from sequencing reads generated in a retroviral insertional mutagenesis screen.	65
2.2	The number of sequence reads per tumour before mapping.	68
2.3	The lengths of retroviral reads that are unambiguously mapped, unmapped, and unmapped and uncharacterised .	74
2.4	BLAST scores for uncharacterised unmapped reads.	78
2.5	The filtering process for single mapping reads.	81
2.6	Determining the exact insertion site and orientation of retroviral and transposon insertions in the mouse genome.	83
2.7	Insertions in the mouse aminoadipate-semialdehyde synthase (<i>Aass</i>) gene are PCR artefacts that map to an LTR-like sequence in the mouse genome	85
2.8	A high proportion of insertions in control samples map to the <i>Myc</i> gene.	86
2.9	The number of insertions per tumour and reads per insertion.	88
2.10	The number of genes with insertions in 100 bp intervals up to 20 kb upstream in the sense and antisense orientation and downstream in the sense and antisense orientation with respect to the gene.	91
2.11	Insertions around known cancer genes <i>Pim1</i> , <i>Kit</i> , <i>Gata1</i> and <i>Blm</i> .	94

2.12	Insertions in <i>En2</i> and <i>Foxf2</i> are located at the splice junctions used to construct the T2/Onc transposon and are contaminating sequences.	103
3.1	MuLV and T2/Onc insertions across the mouse genome.	120
3.2	Known and putative tumour suppressor genes identified in the <i>Sleeping Beauty</i> screen.	122
3.3	The distribution of MuLV insertions within candidate cancer genes.	126
3.4	Intragenic MuLV insertions in candidate cancer genes.	128
3.5	Co-occurring MuLV and T2/Onc insertions help to identify the mechanism of mutagenesis of genes <i>Notch1</i> , <i>Rasgrp1</i> and <i>Etv6</i> .	133
3.6	Variation in the distribution of MuLV and T2/Onc insertions in <i>Myb</i> , <i>Fli1</i> and <i>Erg</i> may reflect differences in the mechanisms of mutagenesis.	137
3.7	<i>Smg6</i> and <i>Foxp1</i> are putative tumour suppressor genes identified by MuLV insertional mutagenesis.	141
3.8	Knockdown of <i>QSK</i> in human HeLa cells is associated with increased chromosome lagging at anaphase.	159
4.1	The distance between the genomic coordinates of adjacent SNPs on the 10K and 10K 2.0 SNP arrays.	167
4.2	The number of SNPs per human protein-coding gene on the 10K and 10K 2.0 SNP arrays.	167
4.3	Altering the values for parameters in DNACopy leads to differences in the regions of copy number change detected by the algorithm, as demonstrated for ovarian cancer cell line 41M-CISR.	169
4.4	Graphical output from MergeLevels for human cancer cell lines 786-0 and AN3-CA.	171
4.5	The number of human cancer cell lines with segments of varying log ₂ -ratio following processing with DNACopy and DNACopy plus MergeLevels.	172
4.6	The number of human cancer cell lines with segments of varying copy number ratio following processing with DNACopy plus MergeLevels.	172
4.7	Distribution of the lengths of amplicons, deletions and homozygous deletions in 713 human cancer cell lines.	174
4.8	Overview of the method for identifying over-representation of the human orthologues of mouse candidate cancer genes in regions of human copy number change.	176
4.9	Over-representation of human orthologues of genes nearest to CISs in amplicons with boundaries extended beyond the first and last amplified SNP by a maximum	

distance of 0 kb, 200 kb, 500 kb, 1 Mb, 3 Mb, 5 Mb and up to the adjacent, non-amplified SNPs.	178
4.10 Over-representation of human orthologues of genes nearest to CISs in full-length human amplicons and shuffled full-length amplicons.	181
4.11 Over-representation of human orthologues of candidate cancer genes in regions of copy number change.	182
4.12 Over-representation of known oncogenes and known tumour suppressor genes in regions of copy number change in human cancer cell lines.	183
4.13 Over-representation of human orthologues of genes nearest to CISs with a <i>P</i> -value of <0.001 in regions of copy number change in human cancer cell lines derived from solid tumours and haematopoietic and lymphoid cancers.	183
4.14 Over-representation of human orthologues of candidate cancer genes in regions of copy number change in cancer cell lines derived from the upper aerodigestive tract, autonomic ganglia, breast, large intestine, oesophagus and stomach.	185
4.15 Over-representation of human orthologues of genes nearest to CISs and genes further from CISs in amplicons and deletions, where CISs have a <i>P</i> -value of <0.001 and <0.05.	187
4.16 Insertions appear to be associated with the gene nearest to the CIS, i.e. <i>1600014C10Rik</i> and <i>Slamf6</i> , even though adjacent genes are also amplified.	189
4.17 Known human oncogenes <i>EVII</i> , <i>FGFR2</i> and <i>KIT</i> are amplified in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas.	193
4.18 Candidate oncogenes <i>MMP13</i> , <i>SLAMF6</i> and <i>RREB1</i> are amplified in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas.	195
4.19 Insertions assigned to <i>Heatr5a</i> may be associated with <i>Hectd1</i> or <i>EG544864</i> .	198
4.20 Candidate tumour suppressor genes <i>WWOX</i> and <i>ARFRP2</i> are deleted in human cancer cell lines and are disrupted by retroviral insertional mutagenesis in mouse lymphomas.	206
4.21 Under- and over-representation of human orthologues of candidate cancer genes in regions of copy number variation.	216
5.1 Distance between adjacent copy number probes across the human genome.	226
5.2 Characterisation of amplicons and deletions in 598 human cancer cell lines analysed using high-resolution SNP array CGH.	228

5.3	Over-representation of CIS genes in amplicons and deletions of varying copy number threshold and number of cell lines across all cell lines, haematopoietic and lymphoid cancer cell lines, and cell lines derived from solid tumours.	231
5.4	The minimal amplified regions within putative tumour suppressor genes <i>CUGBP2</i> and <i>IKZF3</i> are localised around specific exons .	238
5.5	Amplicons and deletions in the <i>LGALS9</i> paralogue <i>NP_001035167.2</i> overlap with a copy number variant from Redon <i>et al.</i> (2006).	241
5.6	All of the MuLV insertions assigned to the <i>Rcbtb2</i> gene are within a larger, unannotated, EST transcript.	249
5.7	Intragenic deletions within <i>NOTCH1</i> result in the formation of the oncogenic NOTCH-IC protein.	249
5.8	Mutations in the <i>ETS1</i> gene result in removal of the Ets domain.	250
5.9	High-resolution and 10K SNP array CGH data for the entire genome of B-cell lymphoma cell line DOHH-2 and breast cancer cell line HCC1143.	254
5.10	Characterisation of amplicons and deletions in 598 human cancer cell lines analysed using 10K SNP array CGH.	256
5.11	Over-representation of CIS genes in amplicons of varying copy number threshold and number of cell lines in the 10K dataset.	258
5.12	<i>SLA2</i> and <i>NDRG3</i> and <i>ZNF217</i> are amplified in a greater number of cell lines in the 10K dataset than in the high-resolution dataset but do not contain SNPs in the 10K dataset.	261

List of Tables

2.1	Characterisation of the insertional mutagenesis datasets.	66
2.2	The number of MuLV reads mapped using SSAHA2, with varying values for parameters seeds and skip, and BLASTN.	72
2.3	Repeat elements that are over-represented and under-represented among unmapped reads compared with unambiguously mapped reads.	77
2.4	Summary of the proportions of unmapped and unambiguously mapping reads that contain vector sequences, or sequences of low complexity, low quality or short length.	77
2.5	Number of intergenic insertions up to 20 kb upstream and downstream of known oncogenes and tumour suppressor genes from the Cancer Gene Census.	93
2.6	Maximum window sizes in kb for significant CISs for varying numbers of insertions in the retroviral and <i>Sleeping Beauty</i> screens.	98
2.7	Comparison of the methods used to generate candidate cancer genes lists from the retroviral and <i>Sleeping Beauty</i> screens.	98
3.1	The human orthologues of mouse CIS genes can help to identify the critical gene(s) in regions of copy number change in acute lymphoblastic leukaemias from Mullighan <i>et al.</i> (2007).	115
3.2	Over-represented GO terms among CIS genes identified using MuLV.	116
3.3	The predicted mutation types and mechanisms of mutagenesis based on the distribution of MuLV insertions within and around candidate cancer genes.	131
3.4	Gene expression values for candidate cancer genes in insertion-containing tumours compared with tumours that do not contain insertions.	147
3.5	Genes containing an over-representation or under-representation of insertions on a given tumour background compared with all other backgrounds and compared with wild-type insertions only.	150
3.6	Gene pairs in which insertions co-occur more often or less often than expected by chance.	155
4.1	Description of the lists of mouse candidate cancer genes used for comparison with human cancer copy number data.	163
4.2	Tissues of origin of human cancer cell lines used in the 10K SNP array CGH analysis.	165

4.3	The number of amplicons in which known cancer genes among genes nearest to CISs are identified when the amplicon boundaries are altered.	179
4.4	Genes that are nearest to CISs in mouse lymphomas and are also promising candidates for targets of amplification in human cancer cell lines.	192
4.5	miRNA genes that are nearest to CISs in mouse lymphomas and are amplified and/or deleted in human cancer cell lines.	198
4.6	Mouse genes that contain retroviral insertions within the coding region and are also promising candidates for targets of amplification or deletion in human cancer cell lines.	201
4.7	Mouse genes that contain retroviral insertions within the transcribed or translated region and are also promising candidates for targets of deletion in human cancer cell lines.	203
4.8	Comparison of methods for detecting regions of copy number gain in 50 randomly selected cancer cell lines.	212
4.9	<i>P</i> -values for the co-occurrence between genes from each gene list within CNVs and regions of copy number change in human cancer cell lines.	217
5.1	Tissues of origin of human cancer cell lines used in high-resolution copy number analysis.	225
5.2	Number of copy number probes per human autosome.	226
5.3	Lists of CIS genes that are in recurrent amplicons across all cell lines and across haematopoietic and lymphoid cancer cell lines only.	235
5.4	A list of CIS genes for which the maximum copy number across all cell lines is significantly higher than expected by chance.	239
5.5	A list of CIS genes that are in recurrent deletions of copy number less than or equal to 0.6 across all cell lines and across haematopoietic and lymphoid cancer cell lines.	246
5.6	A list of CIS genes that are in recurrent deletions of copy number 0.3 or less across all cell lines and across haematopoietic and lymphoid cancer cell lines only.	251
5.7	Comparison of the high- and low-resolution datasets based on the proportion of CIS genes and known cancer genes that are amplified and deleted.	258
5.8	A list of CIS genes that are in recurrent amplicons, recurrent deletions of copy number 0.6 or less and recurrent deletions of copy number 0.3 or less in the 10K CGH dataset.	260
5.9	A list of amplified and deleted CIS genes that are over- or under-represented in cell lines that contain a mutation in <i>TP53</i> or <i>CDKN2A</i> .	265

- 5.10 A list of CIS genes that are co-amplified or co-deleted across a significant number of human cancer cell lines. 269
- 5.11 A list of genes that are co-amplified, co-deleted or amplified and deleted across human cancer cell lines and are also co-disrupted by MuLV in mouse lymphomas. 271

Appendices

A	Human Ensembl genes and their mouse orthologues for known cancer genes in the Cancer Gene Census.	337
B1	<i>Sleeping Beauty</i> CISs and predicted CIS genes obtained using the kernel convolution-based framework with a kernel width of 30 kb.	342
B2	MuLV CISs and predicted CIS genes obtained using the kernel convolution-based framework with a kernel width of 30 kb.	342
C1	List showing other cancer-associated datasets in which the MuLV CIS genes appear.	348
C2	List showing other cancer-associated datasets in which the <i>Sleeping Beauty</i> CIS genes appear.	353
D	Nearest and further genes from CISs with a <i>P</i> -value of <0.001 or <0.05 in lists supplied by the Netherlands Cancer Institute.	354
E	Human cancer cell lines used in the 10K and SNP 6.0 CGH analyses.	361

Abbreviations

ALL	acute lymphoblastic leukaemia
AML	acute myeloid leukaemia
API	application programming interface
BAC	bacterial artificial chromosome
CGH	comparative genomic hybridisation
ChIP	chromatin immunoprecipitation
CIS	common insertion site
CML	chronic myelogenous leukaemia
CNV	copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer
CRUK	Cancer Research UK
DAS	distributed annotation system
ES	embryonic stem
EST	expressed sequence tag
ESP	end-sequence profiling
HGNC HUGO	Gene Nomenclature Committee
HMM	hidden Markov model
IARC	International Agency for Research on Cancer
IR/DR	inverted repeat/direct repeat
KC	kernel convolution
LINE	long interspersed nuclear element
LOH	loss of heterozygosity
LTR	long terminal repeat
MC	Monte Carlo
MCC	Matthew's Correlation Coefficient
MCR	minimal common region of amplification or deletion
MGI	Mouse Genome Informatics
MMTV _{mouse}	mammary tumour virus
MuLV	murine leukaemia virus
NCBI	National Center for Biotechnology Information
NKI	Netherlands Cancer Institute
PCR	polymerase chain reaction
PET	paired-end ditag sequencing
RTCGD	Retroviral Tagged Cancer Gene Database
SB	<i>Sleeping Beauty</i>
SNP	single nucleotide polymorphism
SINE	short interspersed nuclear element
TFBS	transcription factor binding site
UTR	untranslated region
VISA	viral insertion site amplification
WGSA	whole-genome sampling assay
WHO	World Health Organization
WTSI	Wellcome Trust Sanger Institute