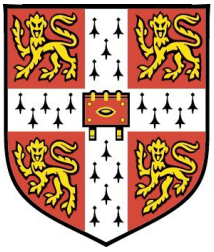
Analysis of genome-wide, cancer-associated mutation datasets in mouse and human



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

List of Figures

List of Tables

Appendices

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	*Sleeping Beauty*
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