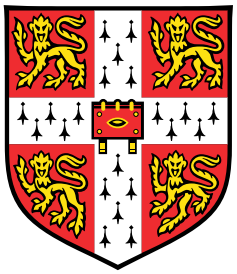
Lineage tracing of normal human development and childhood cancers



Tim Coorens

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare Hall

September 2020

I would like to dedicate this dissertation to family and friends close by and far away, who have never stopped making me smile and laugh.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This dissertation does not exceed 60,000 words in length.

Tim Coorens
September 2020

Lineage tracing of normal human development and childhood cancers

Tim Coorens

Summary

From fertilisation onwards, the cells of the human body continuously experience damage to their genome, either from intrinsic causes or from exposure to mutagens. While the vast majority of DNA damage is repaired and the genome is replicated with extremely high fidelity, cells steadily acquire single nucleotide variants throughout life. Since cells pass these genetic changes on to their descendants, mutations shared between any two cells therefore imply a shared developmental path. In essence, these somatic mutations connect all cells together into one large phylogenetic tree of human development with the zygote at the root.

Reconstructing phylogenies of human development requires readouts of somatic mutations present in single cells. Recently, low-input whole-genome sequencing following laser-capture microdissection has allowed us to reliably call somatic mutations in distinct single-cell derived physiological units, such as colonic crypts and endometrial glands, while retaining spatial information on a microscopic level. In this way, I reconstructed large-scale phylogenies of cells from many different organs of three individuals. These phylogenetic trees recapitulate the early stages of embryonic development and asymmetric cell allocation in the blastocyst, as well as later clonal expansions such as benign prostatic hyperplasia and neoplastic polyp formation.

In a similar way, I also used somatic mutations to investigate the emergence of paediatric cancer, which is thought to be closely linked to aberrations in development. In the context of phylogenetic analyses of tumours, mutations shared between childhood cancers and different normal tissues can shed light on the embryonic lineage of tumours and may reveal the precise juncture at which tumours began to form. Accordingly, I studied the origin of Wilms tumour, the most common childhood cancer of the kidney. I discovered that these tumours often arise from large tissue-resident precursor clones residing in the normal kidney. These embryonal precursors represent an early clonal expansion driven by *H19* hypermethylation.

Lastly, using somatic mutations I discovered that the human placenta is made up of large clonal patches of closely related trophoblast cells. Comparing early embryonic mutations between placental lineages and umbilical cord DNA, which is derived from the inner cell mass, revealed that in approximately half of the cases, a trophectodermal lineage shares no somatic mutations with the umbilical cord. Furthermore, in a quarter of cases, the umbilical

cord is entirely derived from a progenitor later than the zygote. This indicates a natural early segregation between these lineages and a pathway to generate confined placental mosaicism.

This dissertation as a whole provides a new framework to study normal and aberrant human development from whole-genome sequencing. The ability to reconstruct developmental lineages retrospectively can answer fundamental questions about human development and carcinogenesis.

Acknowledgements

The work described in this thesis would have been impossible without the support of many people. First and foremost, I would like to thank my *de facto* triumvirate of supervisors: Professor Mike Stratton, for his deep insights, endless patience and delightful deliberations on Gibbon's *Decline and Fall* and Procopius' *Secret Histories*; Dr Iñigo Martincorena, for his invaluable analytic mind, steadfast support and imbuing me with a strong affinity for the beta-binomial distribution; Dr Sam Behjati, for his unwavering enthusiasm, endless ideas and the lengthy dialogues about right and wrong in the use of colour in figures.

I would also like to thank the other members of my thesis advisory committee, Dr Peter Campbell, Professor Elizabeth Murchison, Professor Magdalena Zernicka-Goetz, Dr Moritz Gerstung and Dr Nick Goldman, for all the helpful insights, discussions and comments along the road.

I would never have been able to perform this work without the support of the Wellcome Trust Mathematical Genomics and Medicine PhD programme, and I am very grateful to the directors and my fellow students over the years for making this programme so special.

I am especially indebted to the administrative staff, IT team and lab support that make the Sanger Institute run smoothly every day and made this research possible on numerous fronts. A special thanks to Jo Jones, Carl Logan and Wendy McLaughlin, the true unsung heroes of this work, whose impeccable handling of chaotic agendas and enquiries small and large helped me enormously. I am very grateful to all the colleagues I closely worked with over the years, especially Dr Luiza Moore, Dr Taryn Treger, Dr Thomas Oliver, Dr Raheleh Rahbari, Dr Roser Vento-Tormo and Dr Philip Robinson.

I am extremely fortunate to have to thank a small legion of friends, without whom this PhD would not nearly have been so much fun. At the risk of committing the heinous crime of omission, I would like to give a few friends a special thanks for extensively proofreading this dissertation, providing me with much-needed distractions and helping with the crosswords: Jannat Ijaz, Dr Pantelis Nicola, Dr Grace Collord, Dr Alex Cagan, Dr Timothy Butler, Dr Heather Machado, Andrew Lawson, Luke Harvey, Dr Patrick McClanahan, Dr Sophie Allcock, Timo Haber and Natalie Shatashvili.

It almost goes without saying that I want express my deepest gratitude to my family back in the Netherlands, to whom I owe everything and who have supported me every step along the way, especially my parents, Peter and Sandra, my siblings Jodie and Jim, my step-parents, Paul and Angelique, and of course, my grandparents.

And of course I would like to thank you, dear reader, for opening this dissertation and having a look at this work. With all sincerity, I hope you will enjoy this as much as I did.

Contents

List of Figures

List of Tables

Nomenclature

Greek Symbols

ρ The overdispersion parameter in a beta-binomial distribution

Acronyms / Abbreviations

AIC Akaike information criterion

AML Acute myeloid leukaemia

BIC Bayesian information criterion

ccRCC Clear cell renal cell carcinoma

CMN Congenital mesoblastic nephroma

cnn-LOH Copy number neutral loss of heterozygosity

CNV Copy number variant

DNV Double nucleotide variant

Indel Insertion and deletion

LCM Laser capture microscopy/microdissection

LOH Loss of heterozygosity

LOI Loss of imprinting

MDS Myelodysplastic syndrome

MNV Multinucleotide variant

MRT Malignant rhabdoid tumour

PGC Primordial germ cell

RNA Ribonucleic acid

SBS Single base substitution

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

SV Structural variant

TGS Targeted genome sequencing

VAF Variant allele frequency

WES Whole-exome sequencing

WGS Whole-genome sequencing