Patterns of somatic genome rearrangement in human cancer

Nicola Diane Roberts
Trinity College
University of Cambridge

January, 2018

Dissertation submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted, for a degree or other qualification at the University of Cambridge or any other university. It does not exceed the prescribed limit of 60,000 words.

Summary

Cancer development is driven by somatic genome alterations, ranging from single point mutations to larger structural variants (SV) affecting kilobases to megabases of one or more chromosomes. Studies of somatic rearrangement have previously been limited by a paucity of whole genome sequencing data, and a lack of methods for comprehensive structural classification and downstream analysis. The ICGC project on the Pan-Cancer Analysis of Whole Genomes provides an unprecedented opportunity to analyse somatic SVs at base-pair resolution in more than 2500 samples from 30 common cancer types.

In this thesis, I build on a recently developed SV classification pipeline to present a census of rearrangement across the pan-cancer cohort, including chromoplexy, replicative two-jumps, and templated insertions connecting as many as eight distant loci. By identifying the precise structure of individual breakpoint junctions and separating out complex clusters, the classification scheme empowers detailed exploration of all simple SV properties and signatures.

After illustrating the various SV classes and their frequency across cancer types and samples, Chapter 2 focuses on structural properties including event size and breakpoint homology. Then, in Chapter 3, I consider the SV distribution across the genome, and show patterns of association with various genome properties. Upon examination of rearrangement hotspot loci, I describe tissue-specific fragile site deletion patterns, and a variety of SV profiles around known cancer genes, including recurrent templated insertion cycles affecting *TERT* and *RB1*.

Turning to co-occurring alteration patterns, Chapter 4 introduces the Hierarchical Dirichlet Process as a non-parametric Bayesian model of mutational signatures. After developing methods for consensus signature extraction, I detour to the domain of single nucleotide variants to test the HDP method on real and simulated data, and to illustrate its utility for simultaneous signature discovery and matching. Finally, I return to the PCAWG SV dataset, and extract SV signatures delineated by structural class, size, and replication timing.

In Chapter 5, I move on to the complex SV clusters (largely set aside throughout Chapters 2–4), and develop an improved breakpoint clustering method to subdivide the complex rearrangement landscape. I propose a raft of summary metrics for groups of five or more breakpoint junctions, and explore their utility for preliminary classification of chromothripsis and other complex phenomena.

This comprehensive study of somatic genome rearrangement provides detailed insight into SV patterns and properties across event classes, genome regions, samples, and cancer types. To extrapolate from the progress made in this thesis, Chapter 6 suggests future strategies for addressing unanswered questions about complex SV mechanisms, annotation of functional consequences, and selection analysis to discover novel drivers of the cancer phenotype.

Acknowledgements

With sincere gratitude, I thank my doctoral supervisor Dr Peter Campbell for his steadfast patience, support, and advice over four transformative years. Plumbing the depths and idiosyncrasies of cancer genome rearrangement proved to be both a rewarding and exasperating endeavour, with Peter's insight and good humour reliably tilting the balance in favour of scientific inspiration.

I was fortunate to benefit from many talented colleagues in the Wellcome Sanger Institute's Cancer Genome Project, absorbing illuminating discussion as well as practical guidance in all matters biological, computational, statistical, and philosophical. In particular, I thank my office companions who helped kick-start my research in the early stages, including Dr Moritz Gerstung, Dr Inigo Martincorena, Dr David Wedge, Dr Kevin Dawson, and Dr Yilong Li. Yilong was especially instrumental as the original architect of the SV classifications underpinning much of this thesis, and I offer particular gratitude for his unfailing generosity in explaining the minutiae of genome rearrangement.

The ICGC PCAWG dataset analysed in this thesis was the result of many years of work by hundreds of researchers across several continents, and I thank the donors and consortium members—especially the structural variation working group—for the privilege of accessing this vast collection of high-quality cancer genome data.

On a personal note, this thesis would not have been possible without the love and understanding of my friends and family in England and Australia. Special thanks to Katie, Ellese, Mariel, and Alice, and to my parents—Barb and Tony—for their tireless support and encouragement, especially in the final six months!

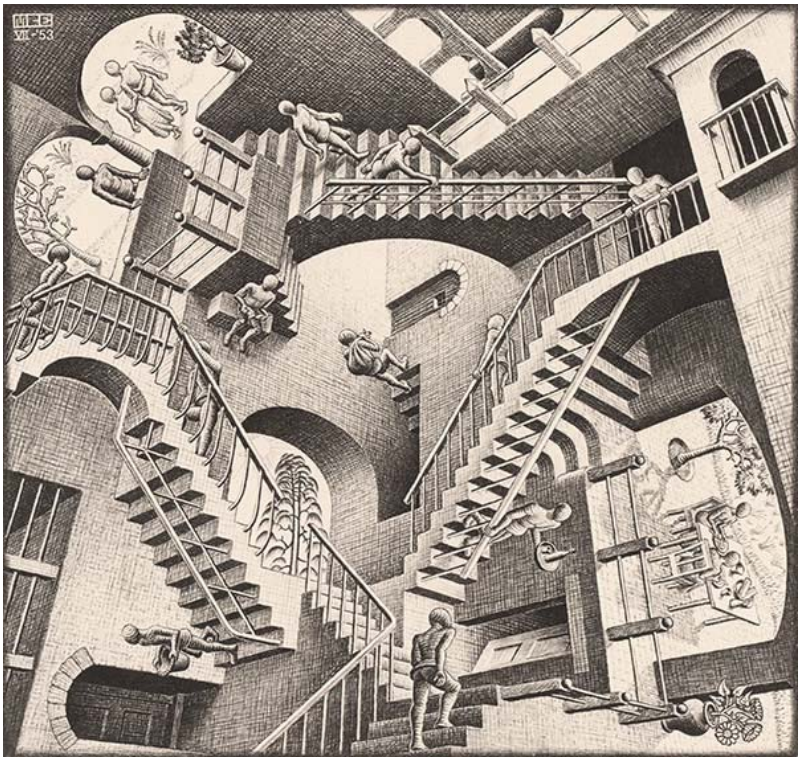
I also thank the Wellcome Trust for their financial support, and Trinity College for providing the social nexus of my studentship in Cambridge.

Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line... Nature exhibits not simply a higher degree but an altogether different level of complexity.

— *Benoit Mandelbrot*

I am progressing very slowly, for nature reveals herself to me in very complex forms; and the progress needed is incessant.

— *Paul Cézanne*



Contents