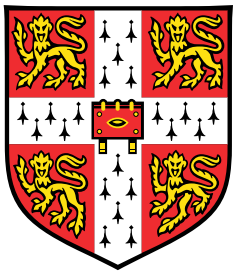
The Intra-Tumour Heterogeneity Landscape of Human Cancers



Stefan Christiaan Dentro

Christ's College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2017

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and 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 diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Stefan Christiaan Dentro
December 2017

Abstract

The Intra-Tumour Heterogeneity Landscape of Human Cancers

Stefan Christiaan Dentro

Tumours accumulate many somatic mutations in their lifetime. Some of these mutations, drivers, convey a selective advantage and can induce clonal expansions. Incomplete clonal expansions give rise to intra-tumour heterogeneity. Somatic mutations can be measured through massively parallel sequencing, where mutations that are supporting incomplete expansions will appear as subclonal. These mutations can be used as a marker of the existence of the expansion and allow for a window into the clonal and subclonal architecture of the tumour at diagnosis.

During my Ph.D. I have developed computational methods to infer intra-tumour heterogeneity from massively parallel sequencing data and applied these to the 2,778 tumour whole genome sequences in the International Cancer Genome Consortium Pan-Cancer Analysis of Whole Genomes initiative to paint the pan-cancer landscape of intra-tumour heterogeneity.

I will first introduce the methods; a method to call somatic copy number alterations (Battenberg) and a method to infer subclones from single nucleotide variants (DPClust). Both are extensively validated on simulated and on real data, and I describe a rigorous quality control procedure. The methods are then applied to a single sample to showcase what can be learned about the life history of a cancer, before introducing additional computational methods for a pan-cancer study of heterogeneity. Finally, I describe the findings.

I find that nearly all cancers, for which there is sufficient power, contain at least one subclone (96.7% of 1,801 primary tumours). The subclones contain driver mutations that are under positive selection, and known cancer genes contain subclonal driver mutations in low proportions. 9.5% of tumours contain only subclonal drivers that are clinically actionable, suggesting that heterogeneity could inform treatment choices. Finally, the analysis reveals that activity of smoking and UV-light associated mutational signatures goes down as the tumour evolves, while activity of the APOBEC associated signatures goes up.

Acknowledgements

I would like to thank my supervisors, Peter Van Loo, David Wedge and David Adams. I'm very grateful to have had the opportunity to work with you and on this very exciting project! Thank you very much for your guidance, thoughts, ideas and your patience. It has been an enormous pleasure to learn from you and to work with you! Thank you for the great discussions and for continuously pushing me to be a better scientist in a friendly and positive environment.

During the PCAWG project I've been in the fortunate position to closely collaborate with members of the Van Loo lab. Thank you Maxime Tarabichi, Kerstin Haase, Clem Jolly, Jonas Demeulemeester and Matt Fittall for the fantastic collaboration and the many in-depth discussions on tumour evolution and heterogeneity. I'd like to also thank the Van Loo lab for being my 'home' for the last few years. Annelien, Clem, Jonas, Kerstin, Lilly, Matt, Maxime and Peter: it's been an enormous pleasure!

Most of the work in this thesis was done as part of the international PCAWG collaboration, I'd like to thank everyone of the PCAWG collaborators and in particular Quaid Morris, Moritz Gerstung, Jeff Wintersinger, Ignaty Leshchiner, Amit Deshwar, Yulia Rubanova and Peter Campbell.

I would like to extend my thanks to Jason J. Pitt for the fruitful collaboration on the West-African breast cancer study and to members of the Adams lab, the Wedge lab and of the Cancer Genome Project for stimulating discussions and continuous helpful feedback.

This thesis could not have happened without the love and support from my parents and sister. Thank you for supporting me and believing in me!

Finally, I would like to thank the Wellcome Trust for generously funding this Ph.D.

Preface

During my Ph.D. I have been in the very fortunate position to heavily collaborate with colleagues close by and far away. Nearly all of the work reported in this thesis was performed as part of an international collaboration project to jointly analyse the cancer whole genome sequencing samples that are part of the International Cancer Genome Consortium (ICGC) The Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative. As this data set is referenced throughout this thesis I have included below a high level description of the project and data set.

Nearly all of the work reported in this thesis is performed in collaboration with, or building on top of the work done by others. Throughout this thesis I therefore systematically refer to work that is solely to my credit with "I", and work that was done by others (with my involvement) as "we". There is also the occasional reference to "a collaborator", where work was done that I had no involvement with.

To help the story line I have spread the introduction across Chapters 1 and 2. Chapter 1 contains a brief review of the relevant literature, but descriptions of algorithms which were already in advanced development when I started are described in Chapter 2. I have worked on and with these algorithms throughout my Ph.D. and have made numerous improvements and adjustments (of which improvements on computational resource requirements are not reported as they are not of direct scientific interest). I felt it would make this thesis more easily readable when the algorithms and updates are described in one chapter.

This setup was chosen to paint a comprehensive story that can hopefully be understood from this thesis alone.

The ICGC Pan-Cancer Analysis of Whole Genomes initiative

The International Cancer Genome Consortium (ICGC) was created to coordinate cancer genome sequencing projects spanning 50 different types of cancer, with the aim to sequence over 25,000 cancer genomes (ICGC Consortium, 2010). ICGC is organised as a series of projects based in countries spread all over the world that focus on analysis of a single cancer

type. Over 17,000 cancers have now been sequenced, of which the majority through whole exome sequencing.

The Pan-Cancer Analysis of Whole Genomes (PCAWG) project was launched to comprehensively characterise those samples for which whole genome sequencing (WGS) is available as a single data set (Campbell et al., 2017). The advantage of focussing on samples for which WGS is available is that the full genome can be interrogated, including the full array of single nucleotide variants (SNV), indels (short insertions and deletions) and structural variants (SVs).

The project consists of 16 working groups with each their own distinct theme. I have been a member of the working group that focusses on tumour evolution and heterogeneity, which is a collaboration of about 60 scientists representing 12 different laboratories.

The tumours that are part of ICGC PCAWG had to meet a series of criteria to be included: a minimal set of clinical annotations should be available, both tumour and normal samples have to be paired-end sequenced from an Illumina machine to a coverage of at least 30x and 25x respectively (Campbell et al., 2017). The data set consists of primarily treatment-naive primary tumours and nearly all matched normals are generated from blood samples.

Data from 2,834 donors was selected to be included in ICGC PCAWG, of which data from 2,658 donors passed quality control procedures (Whalley et al., 2017). The analysed data set consists of 2,778 tumours, of which 2,605 are primary tumours and 173 from a metastasis or relapse case, and spreads 39 histologically distinct types of cancer. Each sequencing sample was processed using a standardised set of primary analysis pipelines, that includes alignment of the sequencing reads and variant calling and filtering from pipelines provided by the Sanger, Broad and EMBL/DKFZ (Yung et al., 2017). These pipelines were extended by one additional SNV and one additional indel caller to further increase the reliable detection of low allele frequency variants (Campbell et al., 2017). Clinical data was systematically collected and standardised (Campbell et al., 2017).

Output from the three primary variant calling pipelines was combined into a high quality set of somatic consensus SNVs (Campbell et al., 2017), indels (Campbell et al., 2017), SVs (Campbell et al., 2017) and copy number alterations (CNAs) (Dentro et al., 2017, manuscript in preparation), of which the latter is described in this thesis. The optimal strategy to find consensus SNVs and indels was found by first running 19 different variant callers across a selected set of 64 tumours. 250.000 calls were selected for validation through deep targeted capture sequencing such that every combination of variant callers is represented and were stratified by allele frequency, after which consensus strategy that maximises precision and recall was then developed to generate the final PCAWG calls (Campbell et al., 2017).

These steps have created the largest set of whole genome cancer sequences to date, spreading a broad range of cancer types. The data set is uniformly processed and the variant calling pipelines have been extensively validated. It therefore provides a unique opportunity for a high quality, in-depth study of tumour heterogeneity.

Table of contents