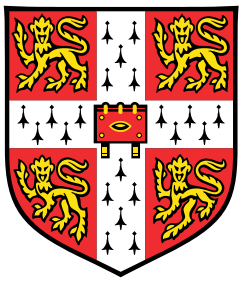
Graphical pangenomics



Erik Peter Garrison

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Fitzwilliam College

September 2018

for E_2 & E_3

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 and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Erik Peter Garrison
September 2018

Graphical pangenomics

Erik Peter Garrison

Completely sequencing genomes is expensive, and to save costs we often analyze new genomic data in the context of a reference genome. This approach distorts our image of the inferred genome, an effect which we describe as reference bias. To mitigate reference bias, I repurpose graphical models previously used in genome assembly and alignment to serve as a reference system in resequencing. To do so I formalize the concept of a *variation graph* to link genomes to a graphical model of their mutual alignment that is capable of representing any kind of genomic variation, both small and large. As this model combines both sequence and variation information in one structure it serves as a natural basis for resequencing. By indexing the topology, sequence space, and haplotype space of these graphs and developing generalizations of sequence alignment suitable to them, I am able to use them as reference systems in the analysis of a wide array of genomic systems, from large vertebrate genomes to microbial pangenomes. To demonstrate the utility of this approach, I use my implementation to solve resequencing and alignment problems in the context of *Homo sapiens* and *Saccharomyces cerevisiae*. I use graph visualization techniques to explore variation graphs built from a variety of sources, including diverged human haplotypes, a gut microbiome, and a freshwater viral metagenome. I find that variation aware read alignment can eliminate reference bias at known variants, and this is of particular importance in the analysis of ancient DNA, where existing approaches result in significant bias towards the reference genome and concomitant distortion of population genetics results. I validate that the variation graph model can be applied to align RNA sequencing data to a splicing graph. Finally, I show that a classical pangenomic inference problem in microbiology can be solved using a resequencing approach based on variation graphs.

Acknowledgements

This work responds to ideas that arose in conversation with Deniz Kural. Our friendship is the first reason that I became a biologist, and his exploration of graphical models for genomes inspired my own. It is thanks to Alexander Wait Zaranek that we had the opportunity to work in George Church's lab, which pulled us both into biology from our previous fields. There I met Madeline Price Ball, who guided me during an immersive and engaging introduction to biology and genomics.

Deniz introduced me to Gabor Marth, with whom I apprenticed in the art of bioinformatics. Gabor encouraged me to contribute extensively to the 1000 Genomes Project, whose objective captured my imagination and whose participants, in particular the members of the analysis group, taught me many lessons in the way of science. I can thank Hyun Min Kang, Goncalo Abecasis, Adam Auton, Laura Clarke, Gerton Lunter, Mark DePristo, Lisa Brooks, Ryan Poplin, Zamin Iqbal, and Heng Li, for always motivating me, and for helping me to understand and correct the many mistakes I made. Meanwhile, Mengyao Zhao and Wan-Ping Lee gave me my first look inside the alignment algorithms that are such an important part of this thesis.

During those years I had the pleasure of living with Benjamin "Mako" Hill and Mika Matsuzaki, who showed me what it means to work as a scientist for the commons. Not only did I learn from them, but from the many thinkers, dreamers, and travelers who they brought into our life in Somerville. These include Hanna Wallach, who helped me to understand the theory and practice of the learning problems I first encountered in genomics, and Nicolás Della Penna, who continues to shape my understanding of many aspects of the scientific artifice, in particular the fuzzy boundary between the social and the technical. I thank my friends Nathan Trachimowicz and Barbara Eghan, with whom I passed so much time in those years, for not letting me lose touch with the beautiful natural and human world in which this work lives and from which it derives its purpose.

This thesis would cover a considerably narrower set of topics if not for the efforts of the many people with whom I have worked to build the variation graph toolkit, `vg`. These include, but are not limited to: Jouni Sirén, from whom I learned about the world of succinct data structures as I watched him build the graph sequence indexes that brought

vg to a level of quality I never could have achieved on my own; Benedict Paten, who applied his unique expertise on genome graphs to help guide the effort of the ever-growing group of project collaborators without a pause in his own stream of contributions; Eric Dawson, whose ready conversation, energy and persistence buoyed me in the early days of vg, and who laid the foundation for future work on structural variant calling on graphs; Shilpa Garg, who brought new ideas about assembly and diploid genome inference to our project while helping to establish the long read alignment algorithms in vg; Adam Novak, who arrived first and transformed the heart of vg from a weak toy model into a foundation suitable for the work of this wide-ranging team, and who continues to carry it forward; Charles Markello, whose precise work on building resequencing pipelines with vg has ensured it is and will be widely usable; Emily Kobayashi, who further honed vg's topology index and is pushing improvement in the dynamic graph indexes we use; Wolfgang Beyer, Toshiyuki Yokoyama, Orion Buske, and Ryan Wick, whose work on sequence graph visualization gave me eyes to understand my work; William Jones, whose clear-minded experiments on alignment identity and score comparison provide the basis for so many figures in this work; Hajime Suzuki, whose work on alignment acceleration lies at the core of the next phase of graph based mappers; Jordan Eizenga, with whom I explored the deep complexity of string to graph alignment algorithms; Xian Chang, who is driving the next high-performance paradigm for sequence to graph alignment; Mike Lin, whose experience and tact guided me both in the vg project and in extracurricular work for DNAnexus; Yohei Rosen, who showed us that old models of haplotype matching could thrive inside our new pangenomic, graphical world; Glenn Hickey, who, in addition to caring for the project, took vg full circle and built variant calling into the graph model; Jerven Bolleman, who found a way to link vg into the enormous world of semantic biological data; and Toshiaki Katayama, who has explored our work and done so much to bring developers of graphical pangenomic techniques together. These members of the "vgteam" have each shared so much more than I can describe here, and I am deeply grateful to have had the opportunity to work with them. The work that I present here is fundamentally based on the productive collaboration that we have shared, and I could not have completed it without their watchful critique and steadfast exploration of their own research questions.

During my time as a student, I benefited from many long conversations with my fellow Sanger PhD cohort over lunch and coffee at the Genome Campus. In particular I learned from the other students with whom I lived: Ignacio Vázquez García, Martin Fabry, Daniel Bruder, Manuela Carrasquilla, and Girish Nivarti.

Pedro Fernandes gave Tobias Marschall and me the chance to teach a course on pangenomics using `vg`, which motivated many of the applications of variation graphs that I present here. Eppie Jones, Rui Martiniano, and Daniel Wegmann have guided and supported my work on ancient DNA. Remo Sanges and Mariella Ferrante gave me a lab to be part of and a fascinating project to explore in my time in Napoli. My corrections for this thesis drew on Alex Bowe's excellent series of blog posts on succinct data structures, and I thank him for allowing me to exposit some of his examples here.

The final version of this thesis reflects a long and memorable conversation led by my examiners Aylwyn Scally and Gerton Lunter. I thank them for their clear suggestions for its improvement, and will forever be grateful for the time they devoted to this sprawling work. They focused my time on its roughest and most incomplete aspects. I am proud of the result that their critique has encouraged me to achieve.

Working with Richard Durbin has been a singular pleasure. Richard has an expansive vision for genomics, but he is always ready to dig into the details of a problem. He is a true master of his craft, able to support and guide every aspect of our work. The group he leads is motivated by his wide-ranging interests in biology. I owe its former and current members thanks for their encouragement and imagination.

Without my family, it is unlikely I would have ever begun the meandering trip that has led me to this thesis. My parents, Mark Garrison and Diane Garrison, helped me to be independent, and opened my mind to the world of ideas, which set me out on a wonderful trip. Along the way, my brother Nels and sister Astrid have kept me honest and careful of myself.

Much of this trip has been alongside my partner Enza Colonna. Non so come dire quanto mi ha aiutato, o quanti passi ho fatto in questo viaggo secondo le idee che abbiamo condiviso. I also am grateful to her parents, Donato Colonna e Concetta Tummolo, under whose almond and olive trees I wrote many pages of this work. Mi hanno sollevato dai problemi di vita quotidiana, e con il loro aiuto ho potuto scrivere la prima bozza di questa tesi in un solo mese.

Our daughter, Exa, who always convinces me to play, made sure that I was never too tired to keep going. I look forward to sharing this with her.

Table of contents

List of figures

List of tables