Chapter 4

Conclusions

Genomics is driven by comparison. It is rare that we have reason to consider a single genome in isolation. We apply alignment algorithms to infer the sequence-level relationship between genomes, or between additional sequence data and genomes. As data scales have increased, we have required incomplete methods to determine these relationships among many individuals. Contemporary resequencing techniques focus on the placement of new sequencing information into a reference system which is typically linear and representative of only a single copy of each genomic locus. A pangenomic reference system allows us to represent multiple versions of each locus, but until recently such techniques have been difficult to apply at scales commonly reached in current analyses.

I propose the use of variation graphs as reference systems in resequencing. These path-labeled, bidirectional DNA sequence graphs allow us to represent collections of genomes in a single, coherent structure which fully capture the sequences and variation between them. They support the direct representation of all kinds of genomic variation. By building a software system supporting the construction, manipulation, indexing, and alignment of new read sets and genomes to variation graphs, I am able to show that this model reduces bias towards the reference in alignment in a wide array of genomic contexts. These methods achieve a level of performance that will make them usable for large-scale resequencing analyses. As I have shown, their modular implementation, based around a handful of core data models, enables the rapid construction of novel graph-based analysis processes that provide conceptual unity to alignment, assembly, and variant calling. Although other methods for aligning sequences against pangenome data structures exist, vg is the first set of tools that does so in a completely coherent manner against arbitrary bidirectional sequence graphs. This is also the first framework to provide graph based analogs of many of the data types standardly used in resequencing.

In addition to developing methods to support alignment to variation graphs, we have explored a variety of related analyses. We developed new techniques for visualizing variation graphs that will help to build the genome browsers necessary to navigate data placed in the context of the graph. To record and interface with annotations embedded in a variation graphs, we have linked the variation graph data model to RDF. To simplify their construction, I provide a method to losslessly induce variation graphs from a set of aligned sequences. We built systems that allow for the efficient summarization of alignment data sets against variation graph. And we worked on methods to support genotyping known and novel variation in graphs. Throughout my work I have supported and worked with a growing group of researchers focused on these techniques, collaborating in the development of graph sequence and haplotype indexing techniques, the evaluation of diverse variation graph models, and the study of ancient DNA using variation graphs.

Graphical models are often regarded with apprehension by members of the bioinformatics community who are accustomed to working with linear reference genomes. I show that arbitrary variation graphs may be consistently linearized for visualization and analysis. Variation graphs built from related sequences tend to have a regional linear property despite the frequent presence of large scale variation. I show that this holds for graphs constructed from a variety of sources using alignment or assembly techniques. They retain relatively linear structures locally, and as such can be used for efficient alignment. The linearization of the graph suggests a projection of sequencing information in the graph into a basis vector space defined by the graph itself. Such an approach may greatly simplify genomic analyses by removing the complicated variant calling step. If the variation we want to consider is already embedded in the graph, we do not need to genotype novel variation or engage in filtering our results. As variation is now embedded in the graph, we can perhaps avoid variant calling altogether where downstream it is possible to work with a normalized coverage model across this graph basis vector. Doing so practically will require the development of techniques that can scale genetic analyses to the large matrix representations implied by such maps.

It is not clear how to build the best graph for a given analysis context. The results I present show that the addition of variation to a graph does not necessarily improve alignment performance in all contexts. Additional variation increases graph complexity, and this can make results more ambiguous. One important step is likely to be the use of haplotype information at the level of alignment. Ongoing work suggests that doing so may mitigate scaling issues that will occur as we build graphs from tens and hundreds of thousands of genomes, but there is still much work to be done. We can expect that, with time, practices will arise that capture the ideal patterns for constructing variation graphs. I found a number of potential input sources unreliable in their current form, and I hope to explore them as variation graph analysis techniques mature. Progressive and multiple whole genome alignment algorithms look to be the most promising way to merge haplotype resolved genome assemblies that new genome inference technologies are enabling. However, as they have difficulty scaling to more than single human chromosomes, I am interested in exploring ways of building variation graphs from networks of pairwise alignments. Given improvement of the input alignment process, this technique could also serve as a scalable way to construct variation graphs in any context where collections of sequenced genomes exist.

I believe that reference genomes should be replaced with pangenomic structures. This is the clearest way to resolve representational issues that arise as we collect large collections of genomes in the species we examine. The variation graph is a natural model with which to do this. Its adoption is now a social as well as a technical question. Can the community generate a unified set of data structures that encapsulate the ideas I have presented here? Large distributed projects like the 1000GP gave rise to the current generation of genomic data formats. It seems natural that the next, graphical, pangenomic phase will require the same. At present, it is not clear what project might support this. Top-down approaches like that presented by the GA4GH have not proven as capable of promoting standards as analysis-oriented projects like the 1000GP, although they have served a coordinating role for the community of researchers interested in these topics. One obvious target for the widespread introduction of variation graph data models would be in the generation of a new reference genome system based on a collection of fully-resolved genomes. Motivation for such an advance increases as evidence mounts that a substantial and important fraction of genetic variation is neither small nor simple. I am hopeful that my work may support such an effort, and that the ideas which arise therein may follow at least in part from the generic graphical pangenomic models I have proposed and demonstrated here.