Chapter 1

Introduction

1.1 Genomic Variation

Genomic variation is one of the fundamental aspects of biology. Difference in the DNA-sequence between two individuals can lead to a change in the translated RNA sequence and further in the sequence, and thereby the form and function, of the expressed protein. Or it can lead to less drastic changes such as changes in the RNA-structure or the shape of the DNA molecule itself. In this way genomic variation can determine differences between specimens of the same species and also differences between the species themselves.

Within a species, the genomic variation between individuals are often limited by evolutionary conservative pressure, meaning that the difference in DNA-sequence between viable specimens of the same species are often small and does not lead to big changes in neither the structure of the DNA or the translated proteins. This limitation has made possible the use of *reference genome*s for a species, a DNA-sequence though to represent the generic sequence for that species, where individual variation from the reference sequence are though to be small.

Reference genomes have helped make sense of sequencing data, that have been dominated by large numbers of small sequence fragments. Without a good reference genome, one would need to fit all those sequence fragments together like a puzzle, in a process called *assembly*. Using a reference genome, one can instead find the best sequence match for each read in the reference sequence, and thus find both where each read belongs in the genome, and also how the sequence differs from the reference. Finding the position of each read can give context to them, since the location of biologically important parts of the genome can be represented on the genome. For instance, if the sequencing of an individual maps to a location within a the known location of a protein coding gene and differs from the reference sequence, it's possible that that individual has a genomic variant that alters the function of that protein.

Since this process of *mapping* sequence reads to the reference is such a

fundamental step in many biological analyses, the quality of the reference sequence has been of high importance. Thus, since the dawn of human genome sequencing in *year*, the Genome Reference Consortium has released n versions of the human reference genome alone. The latest version of the human reference genome (GRCh38), highlights some shortcomings of representing the reference of a species as a single sequence. GRCh38 includes, in addition to the traditional linear reference sequence, a set of alternative sequences from areas of the genome where there are known variations of the DNA-sequences which makes it problematic to map reads from those areas to the reference sequences. Secondly the new version included changes that disrupted the coordiante system of the reference. This has lead to a backward incompatibility that has prevented a widespread adoption of the new reference.

1.2 Mapping Bias

Mapping sequencing reads to the reference entails finding subsequences in the reference sequence that are highly similar to the sequenced reads. The match can be inexact due to either the actual DNA-sequence being different or sequencing errors that can substitute one nucleotide with another. It is necessary to set a limit to how dissimilar the reference subsequence can be in order to produce a match due to computational complexity and that allowing a high level of divergence can lead to a large number of matches. For instance, *BWA-mem* by default requires a shared subsequence of at least 19 bp in order to produce a match. This means that reads from regions with much variation from the reference can be unmappable using standard mapping software, and be susceptible to small amount of sequenceing error yielding them unmappable. This means that the resemblence of the sample to the reference genome will lead to better mapping quality, and that some regions of the genome will have a lower mapping rate than others.

1.3 Geometry of the reference genome

As well as being an indexable lookup table for sequencing reads, a reference genome provides a coordinate system and a geometry for sequence data that allows us to look at different sequence elements in conjunction. Most important is the analysis of overlap and distances between subsequences. For instance, the location of a potential transcription factor binding site, predicted from a ChIP-seq experiment, can be compared to the positions of known genes, predicted from amongst others RNA-seq experiments, to determine which gene is most likely regulated by the binding of a TF to the binding site. The distances between subsequences is also used in some mapping tools themselves, as a way to evaluate the match of a read to a reference subsequence. For instance Minimap2 [**?**] uses the relative positioning of kmer matches to the reference to find which chain of kmer-anchors match the read sequence best. The distance be-

tween two subsequences on the reference is invariant to SNPs, since they do not affect distances. However indels and especially structural variants affect the distances between subsequences, and can thus change the outcome of any analysis involving distances. For instance, a structural variant that changes which gene a regulatory element affects.

1.4 Sequence Graphs

A DNA-sequence can be represented by a sequence of letters from the alpahbet (A, C, G, T). In nature it is common for DNA-sequences to be highly similar to each other, only separated by small variations caused by mutations. The most common such mutations are single base-pair mutations, commonly called single-nucleotide-polymorphism (SNP), and insertions and deletions of short subsequences. The most common format for representing similiar DNA-sequences are a block format where deletions are represented by a 'blank' symbol '_'. This format has a redundancy in the representation of the shared parts of the sequences. This works fine for small sets of short sequences, but for sequences on the genomic scale the redundancy gets significant, and the number of '-' symbols needed gets two big. Sequence graphs are an alternative to the block format for representing multiple similar sequences. In its basic form a sequence graph is a collection of nodes, representing nucleotides, and a set of edges representing neighbouring pairs of nucleotides. A single nucleotide sequence can then be represented as the nucleotides of the sequence connected by edges. Similarily two sequences that are separated by a single SNP can be represented as such... Any set of sequences represented in block format can be uniquely represented by a sequence graph. In addition to the graph, a representation of which paths each sequence takes through the graph is needed in order to contain the same information as the block format. Even without thes specific paths, the graph representation of the sequences contain meaningful information. They succinclty sum up the variation between the sequences. Also any path through the sequence graph is a possible combination of the sequences present that can represent a sequence obtained by recombination events from the available sequences. Also each path through the graph represents a possible common ancestor from the graph.

On a larger scale the block format is unconvenient due to the redundancy in the sequences. It is then more common to use a single sequence as a reference sequence and then represent the different sequences only by the places they are different from the reference. This is common for example when representing mulitple genomes from the same species. The variant common format? (VCF) format is a common such format that represents the. It contains a list of variants represented by a position in the reference sequence, a subsequence from the refernce from that positions, and the alternative sequence that replaces the reference sequence. In addition it can contain extra columns for known haplotypes, specifying which variants are present in each haplotype.

Such vcf representations can also be represented uniquely by a sequence

graph, by first representing the reference sequence as a linear sequence graph and then adding to the graph the nodes and edges to represent the variants. It is also here necessary to represent the haplotypes as paths through the graph in order to keep the haplotype information.

The most common variations (SNPs and indels), leads to directed acyclic graphs (DAG) when converted to a sequence graphs. However other types of variants does not have this property and thus leads to more complicated graphs.

Other types of variants Large scale variants that affects the ordering of the nucleotides are not well suited to being represented by DAGs. An example of this is transpositions, where a subsequence of DNA is moved to another location in the genome. This can be represented as a DAG by adding a new variant in the graph covering the whole sequence from covering both the old an new postions of the subsequence. However this can lead to much redundancy since the sequence between the two positions will be represented twice. The other alternative is to only add new edges to the graph, representing the sequence when the substring has been moved. This will however break the acyclisity of the graph, and thus complicate most operations one would do on the graph.

Another case which will lead to redundancy if represented as a graph is reversals. Here a piece of the DNA is reversed leading to the subsequence being substituded with its reverse complement. Adding a subsequence in the graph representing the reverse compliment is inderictly redundant. Even though the reverse compliment is not included as a path in the graph, it is directly deducible from a sequence in the graph. In order to represent suche reversals, and other things needing the reverse compement, the concept of a side graph has been introduced. In a side graph, all the nodes representing nucleotides have to sides and an edge is defined as going from one side of a node to a side of another node. One node-side represents the nucleotide, while the other side represents the complement in the other reading direction.

Applications Multiple sequence alignment An early application of sequence graphs was in sequence alignment. The application made a sequence graph of the pairwise alignments and also made it possible to align a sequence graph to another sequence graph. A central theme in this application was that if an an alignment of two sequences minimizes some edit distance between them, then any path through the sequence graph representing this alignment will represent a possible common ancestor of the two sequences that minimizes the combined edit distance to the two sequences. Thus in an iterative pairwise joining scheme, one can in each step achieve a sequence graph that represents the most likely common ancestor and then align these sequence graphs to each other. Here, the graph representation is clearly intuitive and benefitial. The fact that each possible path through the graph represents something meaningful makes the graph format very succinct and benefitial for this application.

Mapping In recent years, mapping reads to such sequence graphs have gathered much attention. The process of mapping reads to a reference sequence is trying to find out where in the reference sequence a read is from and usually involves an index that can quickly look up subsequences from the reference. Common such indexes for linear reference sequences are the FM-index, that uses

the burrows wheeler transform to look up subsequences in linear time, and kmer based indexes that can look up subsequences of constant length in constant time. Both of these types of indexes encounters problems when applied to a graph, due to the combinatorial growth of possible subsequences. The GCSA2 index is a kmer-based index able to index a generic sequence graph, but needs to prune out variants in complex regions in order to restrain the combinatiorial growth in kmers. It also requires significantly more computaional time and memory to create the index than for linear references. Another problem with mapping to a graph based references is that combining several subsequence matches is also hard. The early attempts at graph-mapping had the goal of finding any path in the graph. It is however doubful that this is the right approach due to two facts. First when including variants from many individuals in the graph, a region can be filled up with variants from many different samples where none of them appear in the same individual. Thus a region of the graph can include very many paths, where just a marginal fraction of them actually represents sequences that are seen in the samples or could be attained by a small set of recombinations. This is problematic since sequences will have an increased chance of mapping to such areas. The mismapping in itself is an issue, but this will also lead to a bias toward such regions, so any downstream analysis of the data might be severly compromised. A possible remedy for this is to only search in sequences represented by a real haplotype, that is using the haplotype path information in the mapping. This avoids the problem both of combinatorial growth of subsequences and also that of mapping sinks. It is however a question of whether the graph format is meaningful in this context, as the mapping is in reality linear.

Downstream analysis The possiblity to map reads to a variation graph leads to possibilities for downstream analysis. The most natural is to call variants. The process of variant calling is using mapped reads to determine in which parts of the sequence of the sample differs from the sequence in the reference. In this case, graph mapped reads can be advantageos