Mapping to a Reference Genome Structure

Abstract:

To support comparative genomics, population genetics, and medical genetics, we propose that a reference genome should come with a scheme for mapping each base in any DNA string to a position in that reference genome. We refer to a collection of one or more reference genomes and a scheme for mapping to their positions as a reference structure. Here we describe the desirable properties of reference structures and give examples. To account for natural genetic variation, we consider the more general case in which a reference genome is represented by a graph rather than a set of phased chromosomes; the latter is treated as a special case.

Introduction:

A genome assembly is typically represented as a set of strings over the nucleotide alphabet {A,C,G,T}, termed contigs, partitioned into a set of scaffolds, each of which is the concatenation of a sequence of contigs, interspersed with runs of wildcard characters (typically 'N') that represent uncertainty about the DNA sequence between the contigs.

A reference genome assembly is a genome assembly used to represent a species. The first draft of the human reference genome assembly (Lander et al., 2001) was monoploid, meaning that each chromosome was (essentially) represented by a single scaffold. Some polymorphic (variable) regions of the genome in the population were poorly represented by the chosen scaffold (Levy et al., 2007) (Wheeler et al., 2008) (Kidd et al., 2010). To better represent this variation, the current reference human genome assembly, GRCh38, while still primarily a chosen monoploid assembly, contains a substantial number of variant scaffolds, termed alternative haplotypes (Church et al., 2011).

These scaffolds are mapped at both ends to the chosen monoploid assembly while deviating from it in the middle. GRCh38 can therefore be viewed as a graph, consisting of nodes representing reference DNA bases connected by edges that represent the main linear path of bases along the reference chromosome plus branching paths of polymorphism. In this paper we formalize this notion of a reference genome as a graph.

A central function of a reference genome is as a target for mapping DNA bases from other sequenced human genomes. For any position in an input string (typically a short “read” output directly from a sequencing experiment) a mapping to the reference genome is the identification of a position in the reference genome that is considered homologous. Such mappings provide 2 information about the organization of the bases within the newly sequenced input genome, and, through analysis of the structure of the mapping, the variations that it contains.

Results:

Sequence graphs We model both reference and individual genome assemblies as sequence graphs, which allow for a more general class of assembly representation, and which incorporate phased contig/scaffold representations as a special case. Base positions within a reference structure should have an aspect of permanence and universality, so that we can maintain canonical names and aliases for them indefinitely into the future.

One possibility is to assign UUIDs to them for this purpose (http://en.wikipedia.org/wiki/Universally_unique_identifier), but other more compact identifiers can be used as well. A base instance is a pair (b,P) consisting of a labeling base b in {A,C,T,G} and a position P (represented, e.g., as a UUID).

All positions are globally unique, so given a position P, one may determine the base at that position, i.e. the unique b such that (b,P) is a base instance. The nucleotides in all genomes are described as base instances. In order to define a sequence graph, we need to specify how base instances are connected to form sequences. To do this generally, as DNA is double stranded, we must distinguish the forward and reverse complement orientations of base instances. Each base instance (b,P) has a left side, denoted Pl, and a right side, denoted Pr.

An adjacency is an unordered pair of two sides; 4 an adjacency {Ps, Qt} asserts that the s side of the base at position P is connected to the t side of the base at position Q. A sequence graph G = (VG,EG) is a bidirected graph (Medvedev & Brudno, 2009) in which each node in the set VG of nodes is a base instance and each edge in the set EG of edges is an adjacency connecting the sides of two base instances. The forward label of a node (b,P) is the base b, and the reverse label is the reverse complement base b*, where A* = T, T* = A, G* = C, and C* =

G. Using its sides for orientation, for a base instance (b, P) we write b(Pl) = b and b(Pr) = b* to denote the base label oriented by the given side.

A linear thread is a special kind of path in a sequence graph composed of a sequence of oriented nodes and edges terminated by oriented nodes, such that each node other than the first and last node on the path is entered on one side and exited on the other.

Nodes can be visited more than once in a thread. The traversal of a thread specifies a sequence of nucleotides, decoded by enumerating the labels of base instances in the order and orientation specified by the thread, such that if a base instance (b, P) is oriented from Pl to Pr then b(Pl)=b is incorporated into the traversal, and if oriented from Pr to Pl then b(Pr)=b* is incorporated into the traversal.

A circular thread is a circular path of oriented nodes and edges in which each node is entered on one side and exited on the other. Its traversal is a circular sequence of nucleotides, e.g. a mitochondrial sequence. A contig (graph) (we drop the word "graph" when it is clear from the context) is a sequence graph that consists of a single linear or circular thread with no node repetitions. A phased sequence graph is a sequence graph consisting of a set of disjoint contig subgraphs (Fig. 2).

In Appendix A we discuss extension of (phased) sequence graphs and the mapping scheme presented below to represent complete linear chromosomes (sequences terminated at each end by special nodes called telomeres), and scaffolds (sequences of contig subgraphs interspersed with runs of Ns).

Discussion:

We've introduced a scheme to define reference structures with positions that have persistent identifiers, and with the ability to both represent a wide spectrum of human genetic variation and provide an integral method for mapping. This scheme avoids the problem of ill-defined alignment to the reference. In addition, we've shown how the multi-mapping problem can be dealt with by creating a hierarchy of reference structures. We've defined reference structures using sequence graphs, and described context-driven mapping schemes that employ simple exact string matching. This is concordant with the indexing schemes for (directed acyclic) graphs, which build upon the

Burrows Wheeler Transform (M Burrows, 1994), and future implementation of the reference structures we introduce here is likely to use such schemes.

A reference hierarchy built from human genomes organized by populations and subpopulations would be a good way to create a rigorous representation of segregating human genetic variation for both population and medical genomics. Such a reference hierarchy could start from many individual human genomes at the bottom level, each in a separate context-driven reference structure.

These could then be grouped into larger and larger subpopulations at higher levels, each such subpopulation being represented by a merged sequence graph containing all the variation present in the subpopulation's bottom level genomes. The depth-first search process for rapidly recovering mappings can be used in a hierarchy of this type to find haplotypes in subpopulation-specific reference genomes that match a position in any given human input genome.

A de Bruijn or similarly merged reference, perhaps at a higher level in the hierarchy, can similarly be used to map positions in the repetitive areas of a reference genome from shorter contexts.

If a phased reference genome contains a piece of DNA larger than a read size p that is repeated multiple times as identical paralogs, then these will be merged into a single subgraph the poverlap merged graph. While it would require a very large context to map uniquely to a position in this repetitive region in an unmerged reference, in a p-overlap merged reference a position would be uniquely mapped with a typical read-sized context.

Once we map to the unique position in the merged reference using the short context, we can use the depth-first search procedure to efficiently recover all the separate paralogs in the unmerged genomes at lower levels in the hierarchy that were merged to form this single position in the merged reference.

Summary :

Reference hierarchies of sequence graphs and accompanying context-driven mapping schemes combine the strengths of the GRC reference genome with those of the dbSNP variation catalog to provide a single unified approach to human genomic reference variation. As the uses of genomics in science and

medicine rapidly expand in the coming decade, it is vital that we take time to re-examine our methodology for defining human genomic reference sequences and variants, so that we can have a system that is both comprehensive and efficiently extensible, while remaining computationally scalable.

References:

hierarchies offer an attractive approach. Acknowledgements We thank Heng Li, Gad Getz, Ewan Birney and Richard Durbin and other members of the Global Alliance for Genomics and Health Data Working Group for helpful conversations and providing examples that motivated the described approach. We in particular thank Heng Li for providing the alignment.