Round Table I

Representing Data in BAMs, VCFs and other formats

Aaron Quinlan & Michael Schatz

Problems

While the addition of the alternate loci provides additional sequence that should improve read alignment it is difficult to use these sequences with the current tool chain. This is because the addition of the alternate loci introduces sequence duplication due to homology between the allelic sequences. In addition, many users do not distinguish between alt loci (alternate sequence representations) and fix patches (corrections to existing assembly sequences) when defining map positions.

- Most aligners cannot distinguish between allelic duplication and segmental duplication.
- It is unclear what affect the additional sequences have on variant callers.
- Aligners need to operate on fix patches and alternate loci differently.
- Communicating information about the nature of the non-chromosomal assembly sequences (alt loci and patches) to users is still a challenge.

Questions / Notes

- What are the advantages?
  - Alt-sequences mismap to other loci
  - Alt-sequences contain novel genes (158?)
  - "Perfect" de novo assemblies of individuals becoming possible
  - Are these advantages enough to push the community to use alt-sequences & graphs
  - Recalling known variants (N+1st sample problem)
  - Representations of haplotypes, especially of clinical value

- Who has been successful using the fixes & alternate sequences?
  - What were the challenges using those sequences?
  - How did you overcome the "depressed" mapping scores?
    - Alignments either have depressed MAPQ or miss the second loci
    - Collect the data, but how do we interpret the relationships
  - Who/why are people recalling the alt-sequences

Questions / Notes

- How do we encode sequences, variants, & annotations on alternate loci coordinates?
  - "Bubble sequences", Compressed de Bruijn graphs, string graphs, variant graphs
  - GRC coordinates and other systems
  - Primary path, gluing on additional pieces in a recursive structure
  - SNPs on a haplotype (VCF) miss the largest context
    - Database of relationships between alternate loci, list of SNPs to distinguish
    - Input is Database + VCF, output is the updated VCF with haplotype predictions
    - Have to define alt-relationships in the VCF files; need additional tags: alt-locs, alt-haps
  - "Meta-site" in VCF format of a block substitution

- Tools
  - What alt-aware algs exist?
    - BWA, NCBI-Prism, glia, Novoalign
  - What are the most important algorithms we need to update to solve the chicken and egg problem
    - Mappers, variant callers, RNA-seq, annotation pipelines, genome browsers]
    - graph-BLAST, -BWA, -SAMTools, -TopHat/Cufflinks, -IGV, -UCSC, -MAKER, …
  - Who is going to write and fund all of these? How can we coordinate efforts?
  - Short term stop-gap solutions versus long term solutions

- Other issues
  - Updating legacy analysis: recompute vs project coordinates, other approaches?
  - APIs and data models for storing, querying, and accessing alternate sequences