ADAM/Avocado Architecture

Jon Deaton

Thursday, July 5, 2018

Topics

- ADAM/Avocado Schemas
- ADAM/Avocado Class heigherarchy
- Biallectic Genotyper Execution (variant calling)
- Cannonical SNP caller algorithm
- Read Alignment Execution (read mapping)

bdg-formats schemas

ADAM provides several schemas convenient for representing genomic data

- *AlignmentRecord schema* - represents a genomic read & that read's alignment to a reference genome.
- *Feature schema* - represents a generic genomic feature. Annorate a genomic region annotation, (e.g. coverage observed over that region, or the coordinates of an exon) .
- *Fragment schema* - represents a set of read alignments that came from a single sequenced fragment.
- *Genotype schema* - represents a genotype call, along with annotations/quality/read support of called genotype.
- *NucleotideContigFragment schema* represents a section of a contig's sequence.
- *Variant schema* - represents a sequence variant & statistics across samples (indivisuals) and annoration on effect.

Avocado SNP Algorithm

- *Biallelic Variant Calling*
  - **biallelic** genomic locus - site where only two alleles are observed
  - **multiallelic** genomic locus - site where many alleles are observed
- The statistical algorithm used to "call variants" in Avocado (i.e. the business-end of Avocado)
- Originally implemented and used in GATK and SAMtools
- First presented in:"A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data" Heng Li, Bioinformatics 2011 Nov 1;27(21):2987-93. doi: 10.1093/bioinformatics/btr509.
- Deals with calling variants under sequencing error rates

Variant calling theoretical foundations

- *Site independency*: Data at different sites in the genome are independent.
- *Error independency and sample independency*: For a given genomic site, sequencing and mapping errors from different reads are independent.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathcal{L}_i(\theta)$$

- $\mathcal{L}(\theta) =$ likelihood of all n individuals/samples
- $\mathcal{L}_i(\theta) =$ likelihood of the i'th sample

Computing genotype likelihoods

Let d_i be the sequencing data for sample i (array of bases on sequencing reads + quality scores)