Variant calling pipeline on a toy-example dataset

Data

This toy dataset represent a small genomic region from the human genome (7:158876691-158884496 in hg19). It's only about 8 Kbp long. The reads are paired-end Illumina short-reads from a real sample (HG002).

We know of two structural variants (SVs) in this region that the sample should have. They were originally found by integrating a multitude of sequencing data (including long-read sequencing), so they might not be detectable from short-read only. We could try nonetheless.

The dataset is very small so that the analysis can be performed in less than a minute on a laptop.

Files

- `ref.fa`: reference genome in FASTA format
- `reads.fastq.gz`: short sequencing reads in interleaved FASTQ format (gzipped)
- `svs.vcf`: known SVs in VCF format

Goal

The goal is to have you implement a tiny variant calling pipeline. The motivation is really to have a support for the discussion during the interview. It's not really about how you do it, but more about understanding and discussing what you tried, implemented and would adapt in the future. So you are not expected to write the best and most comprehensive variant calling pipeline! A minimal pipeline will be enough to chat about during the interview.

Even if you don't manage to have a pipeline working, it's still interesting to see what didn't work and "debug" together.

In practice, on the toy dataset provided, the pipeline could align the reads to the reference genome, then call small variants or genotype the known structural variants.

Hints/suggestions

You can use any tools you want. The tools suggested below are just suggestions in case you are not sure where to start.

- A minimal pipeline could have ~2 steps:
    1. Align the reads to the reference genome
    2. Call variants from theses aligned reads
- The pipeline could be written in a BASH script or, better, with a workflow language (e.g. snakemake, WDL, nextflow)
- Docker or Singularity could be helpful to avoid having to install tools.

- `bwa mem` is a typical option to align reads.
  - docker: `quay.io/biocontainers/bwa:0.7.17--he4a0461_11`
- `bcftools` is a simple option to quickly call variants (`mpileup` and `call` subcommands)
  - docker: `quay.io/biocontainers/bcftools:1.19--h8b25389_0`
- `svtyper` is an option to genotype known structural variants
  - docker: `quay.io/biocontainers/svtyper:0.7.1--py_0`