Structural Variation in 1000 Genomes Data

Background

Genome sequencing is generally done using rather short *reads*, fragments of DNA where the individual genetic letters (A, C, G, T) have been detected. For technical reasons, read lengths beyond a few tens or a few hundreds of letters are frequently not reliable or used. The exact location of a specific read is also unknown. Thus, the reads need to be *aligned* to a reference genome, under the assumption that a specific individual (or, specifically, the *two* copies of the genome carried by that individual), do not deviate significantly in structure from the reference. Rather, all differences are assumed to be local, so almost any valid read from the genome of an individual should have a well-defined location in the reference genome. So-called single-nucleotide polymorphisms (SNP) are easy-to-detect archetypical examples of such variation.

In recent years with the advent of current sequencing technologies, the interest in minor and major structural variation has increased. Structural variation can described as instances where individual genome contains copies, novel sequences, or deletions, that can be far longer than typical read lengths.

Some sequencing technologies produce so-called paired reads, where a unique DNA fragment is amplified and then read in both directions. Paired reads were common in the main Illumina sequencing runs in the 1000 Genomes Project, which actually provides low-coverage reads for around 2,500 individuals from different human populations across the globe. The fragments are typically 500 base pairs in length, with each read being approximately 70 bp.

Using paired reads to find structural variation

If a structural variant exists, this could mean that some reads are hard to map to the reference genome. However, due to contamination and other factors, such reads will always exist. Instead, we can rely on the known length of paired reads. If the two paired ends map to different locations in the reference genome, this indicates that the two locations are close together in this individual, but further apart in the reference genome. Naturally, filtering will then be needed to find actual structural variation, rather than simple alignment errors.

The 1000 genome project reads are available in the “bam” file format. For this project, we have restricted ourselves to the subset of chromosome 20 for all individuals. For just a single individual (in this case individual NA21144), that bam contains 4670424 unique read entries (each paired end can appear multiple times under some circumstances). If we only look for reads parts of alignment pairs where the two ends are more than 1000 bps apart, rather than the expected count of 500, just 938 entries are retrieved – a data reduction of more than a factor of 1000.

A text-formatted BAM entry can look like the following:

```
ERR251691.33096641      145     20      2806669 37      100M    =       2803068 -3701
AAATTGAACTGCTCTTCACAGATATCTTTGACTTGATATTGATTTAAAATGAACTGCTTGCTTCTAAGTTTCCACCATTAAGAATTATGTA
CATTGTTGT
FGIIKCCJHJIEGIMKJIGJHJFJJFHGGHIIAJHHFIKJCHJCDIIIKJKFFIIFBMJFKKKIECJHLKIEFDGGHKHHJIIGKGGGEDE
GHDHFHH@@    X0:i:1  X1:i:0  MD:Z:100        RG:Z:ERR251691  AM:i:37 NM:i:0  SM:i:37 XT:A:U
BQ:Z:EED@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@
```

Describing the structure of these entries is beyond the scope of this task. It suffices to say that column 9 is the total length of the fragment, according to the alignment that this entry is part of. The value here is negative, indicating that the other end of the read is upstream of the current end. The third column is the chromosome number and the fourth column is the position of the current entry. Other columns list read quality and alignment quality, among other things.

For further processing, we would like to filter out all reads part of alignments resulting in a fragment length of more than 1000 bp. Using old-style UNIX piping, this can be accomplished using the following command.

```
samtools view  *chrom20*.bam | awk -F'\t' 'function abs(x){return ((x < 0.0) ? -x : x)} {if (abs($9) > 1000) print $0}'
```

If need be, samtools could repackage the output as a bam file again. We would instead want to store the resulting tuples in an efficient way for further cloud processing, also including the individual ID in the final tuple. The original folder structure is simple enough, with one folder per individual, and one file with a name on the form `NA21144.chrom20.ILLUMINA.bwa.GIH.low_coverage.20130415.bam` in each folder.

There are libraries to interact with bam files for both Python and Scala, so there should be no need to call samtools explicitly from the command line – and the use of awk should certainly be seen as nothing more than a proof of concept.

The resulting total tuple set is expected to amount to approximately 1 GB of expanded text data, which is small enough for repeated interactive querying to explore the data. For example, one could visualize the approximate chromosome locations where putative structural variation is common, or even pairwise links between locations that are often coaligned to pairs. Colors could be used to indicate the different main ancestries (African, European, East Asian etc) within the dataset to better understand structural variation that are only common to some subpopulations, just like some SNPs are only found in some populations.

While the solution is tested for chromosome 20, it is expected to be able to scale to all 22 autosomal human chromosomes and the sex chromosomes. That dataset is approximately 50 TB in size for all 2,500 individuals.

It is frequent to need to refer back to original read files to validate bioinformatics conclusions drawn on more condense data, such as VCF files only including the called variations.

Task

Extract all tuples from chromosome 20 for all individuals in the 1000 Genome Project low-coverage data where the fragment length indicated by the alignment exceeds 1000 bp. The tuples should identify the originating individual and all other data present in the BAM source file for that fragment alignment.

Possible extensions

Exclude all fragments where there are alternative alignments of the same fragment that would put the total length below 1000 bp. This is not possible to do on the reduced dataset, since that data is lost, so it needs to be done in the filtering step. Other criteria, such as only selecting “CIGAR strings” of 100M (indicating a good alignment), can be applied a posteriori.

Visualize the resulting sequence variation hotspots, possibly with some encoding of population origin.