illumina®

DNA-Seq Data Processing

An overview of the BaseSpace® Correlation Engine DNA sequencing pipeline.

Introduction

The DNA-Seq pipeline in the BaseSpace Correlation Engine processes raw DNA next-generation sequencing (NGS) data from various platforms, produces refined sequence alignment, and reports potential variants. Figure 1 illustrates the overall pipeline workflow. This technical note provides details of the DNA-Seq pipeline.

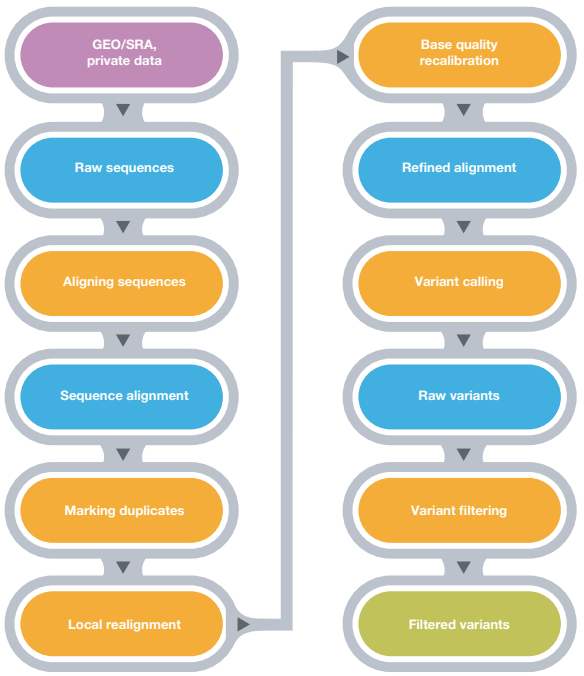
Extraction of Raw Sequences

The pipeline supports the FASTQ format for the input raw NGS data. The Sequence Read Archive (SRA)[1] is a publicly available raw NGS data repository, storing data from many published studies. The archived sequence data from SRA are decompressed and properly split while converting to FASTQ format.

Sequence Alignment

The Burrows-Wheeler Aligner (BWA) is used to align DNA sequences to a reference genome.[2] BWA allows gapped alignment, works for sequences with lengths up to around 100 kb, and also supports the alignment of paired-end reads. The alignment result from BWA is reported in the SAM format.[3]

Figure 1: DNA-Seq Pipeline Workflow



Alignment Refining

To improve alignment quality and reduce potential false variant calls, the alignment output from BWA goes through several refining steps before variant calling. The pipeline uses the Picard[4] and GATK[5,6] suite of software libraries to refine the initial sequence alignment.

Duplicate Marking

Raw NGS data frequently includes PCR and optical duplicates introduced during PCR amplification and sequencing. These duplicates lead to biased variant calls and therefore need to be properly filtered. The pipeline uses the MarkDuplicates tool from Picard to identify and mark alignments of potential duplicate sequences. The marked alignments will be ignored in the downstream analysis.

Local Realignment

The presence of insertion/deletion (indel) sites near the end of a sequence often leads to misalignments that, in turn, contribute to false positive variant calls around this site. The pipeline uses the IndelRealigner tool from GATK to realign reads locally around known indels, which minimizes the number of mismatching bases across all reads covering the target site.

Base Quality Recalibration

The pipeline uses the tools CountCovariates and TableRecalibration from GATK to recalibrate all base quality scores to reflect the true effective base error rates in the alignment file. In particular, all the aligned bases are first categorized based on their characteristics, eg, the reported quality score, the position in the read, and the dinucleotide composition of the preceding and current bases. Next, an empirical error rate is estimated for each category by counting the mismatched bases compared to the reference sequence after excluding the known variants (eg, variants from dbSNP). The empirical base error rates are then converted to the recalibrated base quality scores in the new alignment file.

Variant Calling

The pipeline uses SAMtools/BCFtools[3] to make variant calls for single-nucleotide variants (SNVs) and short indels on the refined alignment file. SAMtools goes over all read-covered genomic positions and reports normalized likelihood values for the observed alignment. BCFtools is used to estimate the allele frequency in the SAMtools output. It then computes the posterior probability of each genotype and makes variant calls. The raw variants are reported in the Variant Call Format (VCF) format.[7]

Variant Filtering

Variant calls are filtered on the following attributes, using vcfutil.pl, to reduce the false positive rate.

- Minimum read depth
- Maximum read depth
- Minimum root mean square mapping quality for SNVs
- Minimum number of alternate bases
- Minimum distance of an SNV to a nearby indel
- Minimum distance between indels
- Minimum p-value for strand bias
- Minimum p-value for baseQ bias
- Minimum p-value for mapQ bias
- Minimum p-value for end distance bias
- Minimum p-value for Hardy-Weinberg Equilibrium (HWE)

Variant Reporting

The VCF files are converted to a bioset format that includes information about position, reference allele, Alleles 1 and 2, read depth, and variant and genotype quality score:

Input Requirements

Currently, the ideal input for the pipeline is high-coverage, targeted exome sequencing data for a few individuals, which is expected for many practical individual-based studies. The output variants from the current pipeline are SNVs and short indels with one alternate form.

References

1. Leinonen R, Sugaware H, Shumway M (2010) The sequence read archive. Nucleic Acids Res 39: D19–D21.
2. Li H, Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754–1760
3. Li H, Handsaker B, Wysoker A. Fennell T, Ruan J, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078–2079.
4. picard.sourceforge.net
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9): 1297–1303.
6. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5): 491–498.
7. vcftools.sourceforge.net/specs.html

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

 Pub. No. 970-2014-011 Current as of 28 March 2016

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA