1000 Genomes Processing README

This README contains information relating to data associated with the 1000 Genomes resequencing done at New York Genome Center.

Alignment, post-processing and variant calling

Alignment and post-processing are performed exactly as outlined by the Center for Common Disease Genomics project: https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md .

Programs and reference data

The data was aligned to the reference genome using the following programs and reference datasets:

1. BWA-MEM
2. Samtools-1.3.1
3. Picard-2.4.1
4. GATK-3.5-0
5. Resource files
   - All the resource files used in the analysis can be obtained here: https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/ .

Reference genome: GRCh38 with alternative sequences, plus decoys and HLA

The reference genome that the data was aligned to can be obtained here: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa

Command lines

1. Alignment at lane level

```
bwa mem  -Y \
-K 100000000 \
-t 16 \
-R $rg_string \
$reference_fasta_file \
$fastq_file(1) \
$fastq_file(2) | samtools view -Shb -o $bam_file -
```

2. Fix mate information in the BAM

```
java $jvm_args -jar picard.jar \
FixMateInformation \
MAX_RECORDS_IN_RAM=2000000 \
VALIDATION_STRINGENCY=SILENT \
ADD_MATE_CIGAR=True \
ASSUME_SORTED=true \
I=$bam_file \
O=$bam_file_fixedmate
```

3. Merging lane level bam files to Sample level bam files

```
java $jvm_args -jar picard.jar \
MergeSamFiles \
USE_THREADING=true \
MAX_RECORDS_IN_RAM=2000000 \
VALIDATION_STRINGENCY=SILENT \
SORT_ORDER=queryname \
INPUT=$bam1 \
INPUT=$bam2 \
OUTPUT=$bam_merged
```

4. Mark duplicates and coordinate sort BAM

```
java $jvm_args -jar picard.jar \
MarkDuplicates \
MAX_RECORDS_IN_RAM=2000000 \
VALIDATION_STRINGENCY=SILENT \
M=$dedup_metrics \
I=$bam_sorted \
O=$bam_dedup
```

```
java $jvm_args -jar picard.jar \
SortSam \
MAX_RECORDS_IN_RAM=2000000 \
VALIDATION_STRINGENCY=SILENT \
SORT_ORDER=coordinate \
CREATE_INDEX=true \
I=$bam_merged \
O=$bam_sorted
```

5. Recalibrate base quality scores using known SNPs

```
java $jvm_args -jar GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-downsample_to_fraction 0.1 \
-nct 4 \
--preserve_qscores_less_than 6 \
-L $autosomes \
-R $reference_fasta \
-o $recal_data.table \
-I $bam_sorted \
-knownSites $known_snps_from_dbSNP138 \
-knownSites $known_indels \
-knownSites $known_indels_from_mills_1000genomes
```

```
java $jvm_args -jar GenomeAnalysisTK.jar \
-T PrintReads \
-nct 4 \
--disable_indel_quals \
--preserve_qscores_less_than 6 \
-SQQ 10 \
-SQQ 20 \
-SQQ 30 \
-rf BadCigar \
-R $reference_fasta \
-o $recalibrated_bam \
-I $bam_sorted \
-BQSR $recal_data.table
```

6. Creating CRAM files

```
samtools view \
-C \
-T $reference_fasta \
-o $cram \
$recalibrated_bam

samtools index $cram
```

7. Raw variant calls using HaplotypeCaller on single sample

```
java $jvm_args -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \
--genotyping_mode DISCOVERY \
-A AlleleBalanceBySample \
-A DepthPerAlleleBySample \
-A DepthPerSampleHC \
-A InbreedingCoeff \
-A MappingQualityZeroBySample \
-A StrandBiasBySample \
-A Coverage \
-A FisherStrand \
-A HaplotypeScore \
-A MappingQualityRankSumTest \
-A MappingQualityZero \
-A QualByDepth \
-A RMSMappingQuality \
-A ReadPosRankSumTest \
-A VariantType \
-l INFO \
--emitRefConfidence GVCF \
-rf BadCigar \
--variant_index_parameter 128000 \
--variant_index_type LINEAR \
-R $reference_fasta \
-nct 1 \
-I $recalibrated_bam \
-o $gvcf
```

8. Jointly recalibrate Genotype Quality score of all samples

```
        java $jvm_args -jar GenomeAnalysisTK.jar \
        -T GenotypeGVCFs \
        -R $reference_fasta \
        -nt 5 \
        --disable_auto_index_creation_and_locking_when_reading_rods \
        --variant $gvcf \
        -o $recalibrated_vcf
```

9. Variant Quality Score Recalibration (VQSR) to assign FILTER status

```
        java $jvm_args -jar GenomeAnalysisTK.jar /
        -T VariantRecalibrator /
        -R $reference_fasta /
        -nt 5 /
        -input $recalibrated_vcf /
        -mode SNP /
        -recalFile $vqsr_snp.recal /
        -tranchesFile $vqsr_snp.tranches /
        -rscriptFile $vqsr_snp_plots.R /
        -resource:hapmap,known=false,training=true,truth=true,prior=15.0
$hapmap /
        -resource:omni,known=false,training=true,truth=true,prior=12.0
$kg_omni /
        -resource:1000G,known=false,training=true,truth=false,prior=10.0
$kg_snps /
        -resource:dbsnp,known=true,training=false,truth=false,prior=2.0
$dbsnp /
        -an QD /
        -an MQ /
        -an FS /
        -an MQRankSum /
        -an ReadPosRankSum /
        -an SOR /
        -an DP /
        -tranche 100.0 /
        -tranche 99.8 /
        -tranche 99.6 /
        -tranche 99.4 /
        -tranche 99.2 /
        -tranche 99.0 /
        -tranche 95.0 /
        -tranche 90.0

        java $jvm_args -jar GenomeAnalysisTK.jar /
        -T VariantRecalibrator /
        -R $reference_fasta /
        -nt 5 /
        -input $recalibrated_vcf /
        -mode INDEL /
        -recalFile $recalibrate_indel.recal /
        -tranchesFile $recalibrate_indel.tranches /
        -rscriptFile $recalibrate_indel_plots.R /
        -resource:mills,known=true,training=true,truth=true,prior=12.0
$kg_mills /
        -resource:dbsnp,known=true,training=false,truth=false,prior=2.0
$dbsnp /
        -an QD /
        -an FS /
        -an ReadPosRankSum /
        -an MQRankSum /
        -an SOR /
        -an DP /
        -tranche 100.0 /
        -tranche 99.0 /
        -tranche 95.0 /
        -tranche 92.0 /
        -tranche 90.0 /
        --maxGaussians 4

        java $jvm_args -jar GenomeAnalysisTK.jar /
        -T ApplyRecalibration /
        -R $reference_fasta /
        -nt 5 /
        -input $recalibrated_vcf /
        -mode SNP /
        --ts_filter_level 99.80 /
        -recalFile $recalibrate_SNP.recal /
        -tranchesFile $recalibrate_SNP.tranches /
        -o $vqsr_snp_vcf

        java $jvm_args -jar GenomeAnalysisTK.jar /
        -T ApplyRecalibration /
        -R $reference_fasta /
        -nt 5 /
        -input $vqsr_snp_vcf /
        -mode INDEL /
        --ts_filter_level 99.0 /
        -recalFile $recalibrate_INDEL.recal /
        -tranchesFile $recalibrate_INDEL.tranches /
        -o $vqsr_snp_indel_vcf
```

Definitions of delivered files

1. recalibrated_variants.vcf.gz[.tbi]
    - All variants in Variant Call Format (VCF) file along with index.
2. recalibrated_variants.annotated.vcf.gz[.tbi]

  - Normalized VCF stripped of genotype calls and annotated using snpEff and BCFTools.

3. recalibrated_variants.annotated.txt
   - Variant annotations in a tab-delimited file.
4. recalibrated_variants.annotated.coding.txt
   - All annotated variants with HIGH/MODERATE impact in a tab-delimited file.
   - High impact - The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay. e.g. stop_gained, frameshift_variant.
   - Moderate impact - A non-disruptive variant that might change protein effectiveness. e.g. missense_variant, inframe_deletion
5. recalibrated_variants.annotated.coding_rare.txt
   - All HIGH/MODERATE annotated variants with less than 5% allele frequency in 1000genomes and ExAC in a tab-delimited file.
6. recalibrated_variants.annotated.clinical.txt
   - All low frequency HIGH/MODERATE annotated variants with possible clinical impact from ClinVar in a tab-delimited file.