From the provide website, we can find the related files are

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data\_collections/1000\_genomes\_project/data/CEU/NA12878/alignment/NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.cram

<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/CEU/NA12878/alignment/NA12878.alt_bwamem_GRCh38DH.20150718.CEU.low_coverage.cram.crai>

1. Down load the data to local or cloud (assume wget is installed and there is public access)

wget –c <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/CEU/NA12878/alignment/NA12878.alt_bwamem_GRCh38DH.20150718.CEU.low_coverage.cram>

wget –c <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/CEU/NA12878/alignment/NA12878.alt_bwamem_GRCh38DH.20150718.CEU.low_coverage.cram.crai>

1. Transform cram to bam (assume samtools is installed). reference\_file is the file path for reference genome GRCH38 which is already indexed by “samtools faidx” and index the bam file

Samtools view –T reference\_file –b –o \ NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam \

NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.cram \

Samtools index NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam

1. Check the quality of bam file (assume Picard is installed and represented by pseudo path $PICARD)

3.1 insert size checking

java -Xmx24g -jar $PICARD CollectInsertSizeMetrics \

I= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam \

O= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_insert\_size\_metrics.txt \

H= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_insert\_size\_metrics.pdf

3.2 alignment summary

java -Xmx24g -jar $PICARD CollectAlignmentSummaryMetrics \

I= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam \

O= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_alignment\_metrics.txt \

R= reference\_file

* 1. GC bias checking

java -Xmx24g -jar $PICARD CollectGcBiasMetrics \ I=NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam \ O=NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_gc\_bias\_metrics.txt \ R=/workspace/inputs/references/genome/ref\_genome.fa \ CHART= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage \_gc\_bias\_metrics.pdf\ S= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage \_gc\_bias\_summary.txt

1. Deduplication (PCR process could involve some duplicate reads. we need to remove these reads) also index the new bam file

java -jar $PICARD MarkDuplicates \

I= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam \

O= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped.bam \

M= NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_metrics.txt \

REMOVE\_DUPLICATES=true

Samtools index NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped.bam

1. Local realignment (There could be misalignment around indels. Local realignment can help us to correct these potential misalignment. Assume GATK is installed)

gatk --java-options "-Xmx4g" RealignerTargetCreator \

-R reference\_file \

-I NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped.bam \

-O NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_ realignment\_targets.list

gatk --java-options "-Xmx4g" IndelRealigner \

-R reference\_file \

-I input.bam \

-targetIntervals NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_ realignment\_targets.list \

-O NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.bam

1. Base quality recalibration

gatk BaseRecalibrator \

-I input.bam \

-R reference\_file \

--known-sites ALL.wgs.phase3\_shapeit2\_mvncall\_integrated\_v5b.20130502.sites.vcf \

-O recal\_data.table

gatk ApplyBQSR \

-R reference\_file \

-I NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.bam \

--bqsr-recal-file recal\_data.table \

-O NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.recalibrated.bam

7. clean temporary files not useful anymore.

rm NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.bam

rm NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped.bam

rm NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.bam

rm NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage.cram

…

8. report the coverage (assume bedtools is already installed)

bedtools genomecov –ibam NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.recalibrated.bam –d > NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.recalibrated.bed

if we are interested in a specific region: target.bed

bedtools intersect –a \ NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.recalibrated.bed –b target.bed > \ NA12878.alt\_bwamem\_GRCh38DH.20150718.CEU.low\_coverage\_deduped\_realigned.recalibrated\_target.bed