Supplement

* Apply the qualified libraries to 2 × 150 bp paired-end sequencing on the HiSeq X-ten platform.

Two library preparation methods were used:

– Cerba modified protocol: Cerba used NEBNext dsDNA

Fragmentase to shear input DNA and prepared

libraries with NEBNext Ultra II DNA Library Prep Kit for

Illumina. Libraries were then enriched following the

SureSelectQXT target enrichment for Illumina multiplexed

sequencing protocol: https://www.agilent.com/cs/library/

usermanuals/Public/G9681-90000.pdf.

– SureSelectXT low input protocol coupled with upfront

enzymatic shearing using the SureSelect XT HS and XT

low input enzymatic fragmentation kit (SureSelect XT

HS/LI ENZ): with the protocols outlined in: https:// www.

agilent.com/cs/library/usermanuals/public/G9703-

90000.pdf and https://www.agilent.com/cs/library/

usermanuals/public/G9702-90050.pdf

* Align FASTQ files to the human reference genome (hg19/GRCh37) with BWA v0.7.1318,19. Sort the aligned files (sam/bam format files) by samtools and then flag duplicates by using Picard.

**Analysis of the code. (R software)**

①bwa index human.fasta

②$ bwa mem -t 4 -R '@RG\tID:foo\_lane\tPL:illumina\tLB:library\tSM: ASD001\_1' /path/to/human.fasta read\_1.fq.gz read\_2.fq.gz | samtools view -S -b - > ASD001\_1.bam

③Usage: samtools sort [options...] [in.bam]

④java -jar picard.jar MarkDuplicates \

 I= ASD001\_1.sorted.bam \

 O= ASD001\_1.sorted.markdup.bam \

 M= ASD001\_1.markdup\_metrics.txt

⑤$ samtools index ASD001\_1.sorted.markdup.bam

⑥samtools merge <out.bam> <in1.bam> [<in2.bam> ... <inN.bam>]

* Use GATK, realign reads locally and recalibrate base qualities.

**Analysis of the code. (R software)**

1. ##RealignerTargetCreator and IndelRealigner

java -jar /path/to/GenomeAnalysisTK.jar \

 -T RealignerTargetCreator \

 -R /path/to/human.fasta \

 -I ASD001\_1.sorted.markdup.bam \

 -known /path/to/gatk/bundle/1000G\_phase1.indels.b37.vcf \

 -known /path/to/gatk/bundle/Mills\_and\_1000G\_gold\_standard.indels.b37.vcf \

 -o ASD001\_1.IndelRealigner.intervals

java -jar /path/to/GenomeAnalysisTK.jar \

 -T IndelRealigner \

 -R /path/to/human.fasta \

 -I ASD001\_1.sorted.markdup.bam \

 -known /path/to/gatk/bundle/1000G\_phase1.indels.b37.vcf \

 -known /path/to/gatk/bundle/Mills\_and\_1000G\_gold\_standard.indels.b37.vcf \

 -o ASD001\_1.sorted.markdup.realign.bam \

 --targetIntervals ASD001\_1.IndelRealigner.intervals

②##BaseRecalibrator and PrintReads

java -jar /path/to/GenomeAnalysisTK.jar \

 -T BaseRecalibrator \

 -R /path/to/human.fasta \

 -I ASD001\_1.sorted.markdup.realign.bam \

 --knownSites /path/to/gatk/bundle/1000G\_phase1.indels.b37.vcf \

 --knownSites /path/to/gatk/bundle/Mills\_and\_1000G\_gold\_standard.indels.b37.vcf \

 --knownSites /path/to/gatk/bundle/dbsnp\_138.b37.vcf \

 -o ASD001\_1.recal\_data.table

java -jar /path/to/GenomeAnalysisTK.jar \

 -T PrintReads \

 -R /path/to/human.fasta \

 -I ASD001\_1.sorted.markdup.realign.bam \

 --BQSR ASD001\_1.recal\_data.table \

 -o ASD001\_1.sorted.markdup.realign.BQSR.bam

③##GATK HaplotypeCaller

java -jar /path/to/GenomeAnalysisTK.jar \

 -T HaplotypeCaller \

 -R /path/to/human.fasta \

 -I ASD001\_1.sorted.markdup.realign.BQSR.bam \

 -D /path/to/gatk/bundle/dbsnp\_138.b37.vcf \

 -stand\_call\_conf 50 \

 -A QualByDepth \

 -A RMSMappingQuality \

 -A MappingQualityRankSumTest \

 -A ReadPosRankSumTest \

 -A FisherStrand \

 -A StrandOddsRatio \

 -A Coverage \

 -o ASD001\_1.HC.vcf

④## SNP Recalibrator

java -jar /path/to/GenomeAnalysisTK.jar \

 -T VariantRecalibrator \

 -R reference.fasta \

 -input ASD001\_1.HC.vcf \

 -resource:hapmap,known=false,training=true,truth=true,prior=15.0 /path/to/gatk/bundle/hapmap\_3.3.b37.vcf \

 -resource:omini,known=false,training=true,truth=false,prior=12.0 /path/to/gatk/bundle/1000G\_omni2.5.b37.vcf \

 -resource:1000G,known=false,training=true,truth=false,prior=10.0 /path/to/gatk/bundle/1000G\_phase1.snps.high\_confidence.b37.vcf \

 -resource:dbsnp,known=true,training=false,truth=false,prior=6.0 /path/to/gatk/bundle/dbsnp\_138.b37.vcf \

 -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -an DP \

 -mode SNP \

 -recalFile ASD001\_1.HC.snps.recal \

 -tranchesFile ASD001\_1.HC.snps.tranches \

 -rscriptFile ASD001\_1.HC.snps.plots.R

java -jar /path/to/GenomeAnalysisTK.jar -T ApplyRecalibration \

 -R human\_g1k\_v37.fasta \

 -input ASD001\_1.HC.vcf \

 --ts\_filter\_level 99.5 \

 -tranchesFile ASD001\_1.HC.snps.tranches \

 -recalFile ASD001\_1.HC.snps.recal \

 -mode SNP \

 -o ASD001\_1.HC.snps.VQSR.vcf

## Indel Recalibrator

java -jar /path/to/GenomeAnalysisTK.jar -T VariantRecalibrator \

 -R human\_g1k\_v37.fasta \

 -input ASD001\_1.HC.snps.VQSR.vcf \

 -resource:mills,known=true,training=true,truth=true,prior=12.0 /path/to/gatk/bundle/Mills\_and\_1000G\_gold\_standard.indels.b37.vcf \

 -an QD -an DP -an FS -an SOR -an ReadPosRankSum -an MQRankSum \

 -mode INDEL \

 -recalFile ASD001\_1.HC.snps.indels.recal \

 -tranchesFile ASD001\_1.HC.snps.indels.tranches \

 -rscriptFile ASD001\_1.HC.snps.indels.plots.R

java -jar /path/to/GenomeAnalysisTK.jar -T ApplyRecalibration \

 -R human\_g1k\_v37.fasta\

 -input ASD001\_1.HC.snps.VQSR.vcf \

 --ts\_filter\_level 99.0 \

 -tranchesFile ASD001\_1.HC.snps.indels.tranches \

 -recalFile ASD001\_1.HC.snps.indels.recal \

 -mode INDEL \

 -o ASD001\_1.HC.snps.indels.VQSR.vcf