|  |  |
| --- | --- |
| **{ UH LAB NGS SOP-3010 RNA Analysis}** | |
| **Responsible Department: NGS** | **Effective Date: { Publication Date }** |
| **Policy Basis for Procedure**  This SOP describes the analysis pipeline of RNA sequencing data. This pipeline includes (1) quality control, (2) variant calling analysis, (3) identification of fusion genes, and (4) statistical analyses of gene expression and isoform expression | |
| **Applicability**  This applies to the bioinformatics personnel of the NGS Lab.  **Description of Standard Procedure**  **DATA REQUIREMENTS**  The pipeline requires as input FASTQ files generated by RNA sequencing. The pipeline also requires the following reference data: (1) the FASTA sequence of a reference genome, currently Human GRCh38 <https://github.com/lh3/bwa/tree/master/bwakit>, and its prebuilt indexes for HISAT2, STAR, and STAR-Fusion; (2) a set of reference single nucleotide polymorphisms (SNPs) from the NCBI SNP database dbSNP\_BUILD\_ID=150; <https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=150> and (3) a set of somatic mutations from COSMIC database <http://cancer.sanger.ac.uk/cosmic/download>.  **QUALITY CONTROL**  Key quality metric thresholds for evaluating rna sample will be: 1) greater 6 million raw reads.  Sample will be automatically be flagged as failed if Map Rate falls below 90%  An exception report will be created for any metrics that does not reach threshold. Corrective actions, further processing or re-processing will be determined by the medical advisor. Corrective action could include anlalysis error log evaluations, re-sequencing of sample or examination of laboratory quality metrics. | |
| **Description of Standard Procedure**  **PROCEDURE**  **Alignment Procedure**    Figure 1. RNASeq Alignment Analysis Overview  1. Trim sequences  Trim the ends of sequences with remaining adapter or quality scores < 25. Remove any sequence less than 35bp after trimming, then generate a file for capturing information about how many sequences were trimmed.   |  |  | | --- | --- | | executable | trimgalore.sh trimgalore/0.4.1 cutadapt/1.9.1 | | input | gzip fastq | | output | trimmed gzip fastq | | command | [] trim\_galore --paired -q 25 --illumina --gzip --length 35 ${fq1} ${fq2}  [] mv ${r1base}\_val\_1.fq.gz ${pair\_id}.trim.R1.fastq.gz  [] mv ${r2base}\_val\_2.fq.gz ${pair\_id}.trim.R2.fastq.gz |   2. Gene fusion detection  Identify gene fusions or fused transcripts using STAR-Fusion.   |  |  | | --- | --- | | executable | starfusion.sh singularity/3.0.2 python/2.7.x-anaconda star/2.5.2b bedtools/2.26.0 trinity/1.6.0 | | input | STAR-Fusion index path, gzip R1 and R2 fastq files | | output | fusion candidate list | | command | [] trinity /usr/local/src/STAR-Fusion/STAR-Fusion --min\_sum\_frags 3 --CPU $SLURM\_CPUS\_ON\_NODE --genome\_lib\_dir ${index\_path} --left\_fq ${fq1} --right\_fq ${fq2} --examine\_coding\_effect --output\_dir ${pair\_id}\_star\_fusion  [] cp ${pair\_id}\_star\_fusion/star-fusion.fusion\_predictions.abridged.coding\_effect.tsv ${pair\_id}.starfusion.txt  export PYENSEMBL\_CACHE\_DIR="/project/shared/bicf\_workflow\_ref/singularity\_images"  [] cut -f 5-8 ${pair\_id}.starfusion.txt |perl -pe 's/\^|:/\t/g' | awk '{print "singularity exec /project/shared/bicf\_workflow\_ref/singularity\_images/agfusion.simg agfusion annotate -db /project/shared/bicf\_workflow\_ref/singularity\_images/pyensembl/GRCh38/ensembl92/agfusion.homo\_sapiens.92.db -g5", $1,"-j5",$4,"-g3",$6,"-j3",$9,"-o",$1"\_"$4"\_"$6"\_"$9}' |grep -v 'LeftGene' |sh  [] cut -f 6,8 ${pair\_id}.starfusion.txt |grep -v Breakpoint |perl -pe 's/\t/\n/g' |awk -F ':' '{print $1"\t"$2-1"\t"$2}' > temp.bed  [] bedtools intersect -wao -a temp.bed -b /project/shared/bicf\_workflow\_ref/human/GRCh38/cytoBand.txt |cut -f 1,2,7 > cytoband\_pos.txt  [] perl $baseDir/filter\_genefusions.pl -p ${pair\_id} -f ${pair\_id}.starfusion.txt |   3. Read mapping to human reference genome  Align RNA sequencing reads onto human reference genome and estimate the sequencing quality and mapping quality.  3.1. Use STAR   |  |  | | --- | --- | | executable | rnaseqalign.sh star/2.4.2a samtools/gcc/1.6 picard/2.10.3 | | input | STAR index path, gzip fastq | | output | bam, quality report | | command | [] STAR --genomeDir ${index\_path}/star\_index/ --readFilesIn $fq1 $fq2 --readFilesCommand zcat --genomeLoad NoSharedMemory --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMheaderCommentFile COfile.txt --outSAMheaderHD @HD VN:1.4 SO:coordinate --outSAMunmapped Within --outFilterType BySJout --outSAMattributes NH HI AS NM MD --outSAMstrandField intronMotif --outSAMtype BAM SortedByCoordinate --quantMode TranscriptomeSAM --sjdbScore 1 --limitBAMsortRAM 60000000000 --outFileNamePrefix out  [] mv outLog.final.out ${pair\_id}.alignerout.txt  [] mv outAligned.sortedByCoord.out.bam output.bam  [] samtools sort -@ $SLURM\_CPUS\_ON\_NODE -O BAM -n -o output.nsort.bam output.bam  [] java -jar $PICARD/picard.jar FixMateInformation ASSUME\_SORTED=TRUE SORT\_ORDER=coordinate ADD\_MATE\_CIGAR=TRUE I=output.nsort.bam O=${pair\_id}.bam  [] samtools index -@ $SLURM\_CPUS\_ON\_NODE ${pair\_id}.bam |   3.2. Use HISAT2   |  |  | | --- | --- | | executable | rnaseqalign.sh hisat2/2.1.0-intel samtools/1.6 picard/2.10.3 | | input | HISAT2 index path, gzip fastq | | output | Aligned bam, quality report | | command | [] hisat2 -p $SLURM\_CPUS\_ON\_NODE --rg-id ${pair\_id} --rg LB:tx --rg PL:illumina --rg PU:barcode --rg SM:${pair\_id} --add-chrname --no-unal --dta -x ${index\_path}/hisat\_index/genome -1 $fq1 -2 $fq2 -S out.sam --summary-file ${pair\_id}.alignerout.txt  [] samtools view -1 --threads $SLURM\_CPUS\_ON\_NODE -o output.bam out.sam  [] samtools sort -@ $SLURM\_CPUS\_ON\_NODE -O BAM -n -o output.nsort.bam output.bam  [] java -jar $PICARD/picard.jar FixMateInformation ASSUME\_SORTED=TRUE SORT\_ORDER=coordinate ADD\_MATE\_CIGAR=TRUE I=output.nsort.bam O=${pair\_id}.bam  [] samtools index -@ $SLURM\_CPUS\_ON\_NODE ${pair\_id}.bam |   4. Generate counts of bam file  Use bam-read count to generate per base read counts   |  |  | | --- | --- | | executable | samtools/1.6 | | input | Aligned bam file | | output | Count text file | | command | [] /project/shared/bicf\_workflow\_ref/seqprg/bam-readcount/bin/bam-readcount -w 0 -q 0 -b 25 -f ${index\_path}/genome.fa $rbam > ${pair\_id}.bamreadct.txt |   5. Calculate Quality Metrics of Alignment  Gather workflow quality metrics, run information and reference. Output summary file per sample.   |  |  | | --- | --- | | executable | bamqc.sh sequenceqc\_rnaseq.pl samtools/1.6 fastqc/0.11.5 git/gcc/v2.12.2 | | input | alignment flagstat.txt output file | | output | alignment and workflow quality metrics | | command | [] samtools flagstat ${sbam} > ${pair\_id}.flagstat.txt  [] fastqc -f bam ${sbam}  [] perl sequenceqc\_rnaseq.pl -r ${index\_path} \*.flagstat.txt |   6. Gene Abundance  Gather count reads per gene using featureCounts. Transcript assembly and quantification using StringTie   |  |  | | --- | --- | | executable | geneabundance.sh subread/1.5.0-intel stringtie/1.1.2-intel | | input | Deduped bam file | | output | Gene read count table, gtf of potential transcripts (isoforms) | | command | [] featureCounts -s $stranded -T $SLURM\_CPUS\_ON\_NODE -p -g gene\_name -a ${gtf} -o ${pair\_id}.cts ${sbam}  [] mkdir ${pair\_id}\_stringtie  [] cd ${pair\_id}\_stringtie  [] stringtie ../${sbam} -p $SLURM\_CPUS\_ON\_NODE -G ../${gtf} -B -e -o denovo.gtf -A ../${pair\_id}.fpkm.txt  [] perl cBioPortal\_documents.pl -p $pair\_id -l ${pair\_id}.cts -f ${pair\_id}.fpkm.txt |   **RNASeq Alignment**  The Workflow used in this SOP can be downloaded here: <https://git.biohpc.swmed.edu/brandi.cantarel/clinseq_workflows>  **Usage**  >nextflow -C nextflow.config run rnaseq.nf -w work\_directory --design samplesheet.txt --input input\_directory --output output\_directory –markdups picard >nextflow.log  where the work\_directory is the folder where the work files for each step will be deposited, the input directory is the name of the folder where the fastq files are located, the output directory is the name of the folder where the output files will be written, the samplesheet.txt is a tab delimited file with the following headers:SampleMergeName,sampleID,SubjectID,FullPathToFqR1,FullPathToFqR2,BAM,OntargetBAM and the nextflow\_rnaseq.log is the file that keeps the log of the run and tracks workflow errors.  **Variant Detection Procedure**    Figure 2: Variant Detection Overview  1. Index BAM files   |  |  | | --- | --- | | executable | indexbams.sh samtools/1.6 | | input | tumor BAM file | | output | tumor BAM index file (bai) | | command | [] samtools index -@ $SLURM\_CPUS\_ON\_NODE ${i} |   2. Index final BAM files   |  |  | | --- | --- | | executable | indexbams.sh samtools/1.6 | | input | tumor Final BAM file | | output | tumor BAM index file (bai) | | command | [] samtools index -@ $SLURM\_CPUS\_ON\_NODE ${i} |   3. Variant Calling and Filtering using Freebayes  Call variants with and base calling quality >= 20, AF >0.1, DP >10, and filter them based on the following cutoffs: DP >= 10.   |  |  | | --- | --- | | executable | germline\_vc.sh uni\_norm\_annot.sh freebayes/gcc/1.2.0 parallel/20150122 bedtools/2.26.0 samtools/1.6 bcftools/1.6 snpeff/4.3q python/2.7.x-anaconda picard/2.10.3 samtools/gcc/1.8 bcftools/gcc/1.8 vcftools/0.1.14 | | Input | reference genome fasta, bam | | output | Freebayes variant call file | | command | [] cut -f 1 ${index\_path}/genomefile.5M.txt | parallel --delay 2 -j $SLURM\_CPUS\_ON\_NODE "freebayes -f ${index\_path}/genome.fa --min-base-quality 20 --min-coverage 10 --min-alternate-fraction 0.01 -C 3 --use-best-n-alleles 3 -r {} ${bamlist} > fb.{}.vcf"  [] vcf-concat fb.\*.vcf | vcf-sort | vcf-annotate -n --fill-type | bcftools norm -c s -f ${reffa} -w 10 -O z -o ${pair\_id}.freebayes.vcf.gz –  [] bcftools norm -m - -Oz $j -o ${pair\_id}.norm.vcf.gz  [] bash annotvcf.sh -p ${pair\_id} -r $index\_path -v ${pair\_id}.norm.vcf.gz  /project/shared/bicf\_workflow\_ref/seqprg/vt/vt decompose\_blocksub ${pair\_id}.annot.vcf.gz -p -a -o ${pair\_id}.vcf  [] bgzip -f ${pair\_id}.vcf |   8. Generate union vcf  Integrate result variant lists generated by GATK, SAMtools, Hotspot algorithm, SpeedSeq, and Platypus variant callers. The priority order for ambiguous variants is SAMtools - SpeedSeq - GATK - Platypus - Hotspot. Annotate VCF with publicly available genes   |  |  | | --- | --- | | executable | union.sh bedtools/2.26.0 samtools/1.6 bcftools/1.6 snpeff/4.3q unionvcf.pl vcfsorter.pl | | input | reference genome fasta, GATK, Strelka, Freebayes, Platypus gzipped and annotated vcf | | output | Union vcf | | command | perl unionvcf.pl ${index\_path}/union.header.vcf $list2  perl vcfsorter.pl ${index\_path}/genome.dict int.vcf |bgzip > ${pair\_id}.union.vcf.gz  mv ${pair\_id}.union.vcf.gz ${pair\_id}.dna.vcf.gz |   **Variant Detection** **Implementation**  The Workflow used in this SOP can be downloaded here: <https://git.biohpc.swmed.edu/brandi.cantarel/clinseq_workflows>  **Usage**  >nextflow -C nextflow.config run tumoronly.nf -w work\_directory --design samplesheet.txt –genome hisat\_index –nuctype rna –callsvs skip --input input\_directory --projectid ${prjid} --output output\_directory >nextflow\_tumoronly.log  where the work\_directory is the folder where the work files for each step will be deposited, hisat\_index is the GRCh38 hisat prebuilt index, the input directory is the name of the folder where the bam files are located, the output directory is the name of the folder where the output files will be written, the samplesheet.txt is a tab delimited file with the following headers: SampleID, BAM, OntargetBAM, and the nextflow\_tumoronly.log is the file that keeps the log of the run and tracks workflow errors. Structural variant calling is skipped with the command –svcalling and –nuctype is defined as rna for the workflow.  **CALCULATIONS**  N/A  **INTERPRETATION/RESULTS/ALERT VALUES**  N/A  **REFERENCE INTERVALS (NORMAL REFERENCE RANGE)**  N/A  **ANALYTICAL MEASURING RANGE (LINEARITY)**  N/A  **RESULT REPORTING CRITERIA**  N/A | |
| **Definitions**   |  |  | | --- | --- | | Term | Definition | | Indel | Insertion and Deletions | | SNP | Single Nucleotide Polymorphism | | |
| **Applicable Forms**   |  |  | | --- | --- | | Document ID# | Title of Document | | 3010 | RNA\_Analysis | |  |  | |  |  |   **Related Documents**   |  |  | | --- | --- | | Document ID# | Title of Document | | 3012 | Overview\_Bioinformatics\_SOP | | 3011 | DNA\_Analysis | |  |  | | |
| **References**   * <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>   <http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/>   * <https://github.com/STAR-Fusion/STAR-Fusion> * <https://github.com/alexdobin/STAR>   + Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. * <https://ccb.jhu.edu/software/hisat2/index.shtml>   + Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357-60. * HISAT, StringTie and Ballgown protocol   + Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016 Sep;11(9):1650-67. * <http://samtools.sourceforge.net/>   + Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.   + Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. * <http://lomereiter.github.io/sambamba/>   + Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015 Jun 15;31(12):2032-4. * <https://broadinstitute.github.io/picard/> * <https://github.com/Illumina/strelka>   + Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Doruk Beyter, Peter Krusche, Christopher T Saunders. * <https://github.com/hall-lab/speedseq>   + Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015 Oct;12(10):966-8. * <http://cancer.sanger.ac.uk/cosmic> * <http://bedtools.readthedocs.io/en/latest/>   + Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. * <http://vcftools.sourceforge.net/index.html>   + Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8. * <http://snpeff.sourceforge.net/>   + Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35. * <https://ccb.jhu.edu/software/stringtie/>   + Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar;33(3):290-5. * Ballgown   + <http://biorxiv.org/content/biorxiv/early/2014/09/05/003665.full.pdf>   + <http://bioconductor.org/packages/release/bioc/html/ballgown.html>   + <https://github.com/alyssafrazee/ballgown> * [http://subread.sourceforge.net](http://subread.sourceforge.net/)   + Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. * <https://bioconductor.org/packages/release/bioc/html/edgeR.html>   + Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40.   + McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012 May;40(10):4288-97. * Freebayes: https//github.com/ekg/frebayes | |
| **Review, Revision, and Approval History**   |  |  |  |  | | --- | --- | --- | --- | | **History of Document** | | | | | Version | Date | Description of Change | Authored/Revised by | | 3 | 2019/07 | * Freebayes, starfusion update, Pipeline update | Brandi Cantarel, Erika Villa | |  |  |  |  | | |
|  | |
| **Contact for Further Information** | |

**Title:** { Title } **Version:** { Version Number } **Status:** { Status }

**Writer(s)**: { Writers > All Users > Name & Title }

**Approver(s)**: { Approvers > All Users > Name & Title } **Approved Date**: { Date Approved }

**Effective Date**: { Publication Date }

**Next Review Date**: { Next Periodic Review Date }