RNASeq Pipeline

UT Southwestern Medical Center

Author: Erika Villa

Version: 1.0

Effective Date: 06/09/2017

1. Abstract

2. Introduction

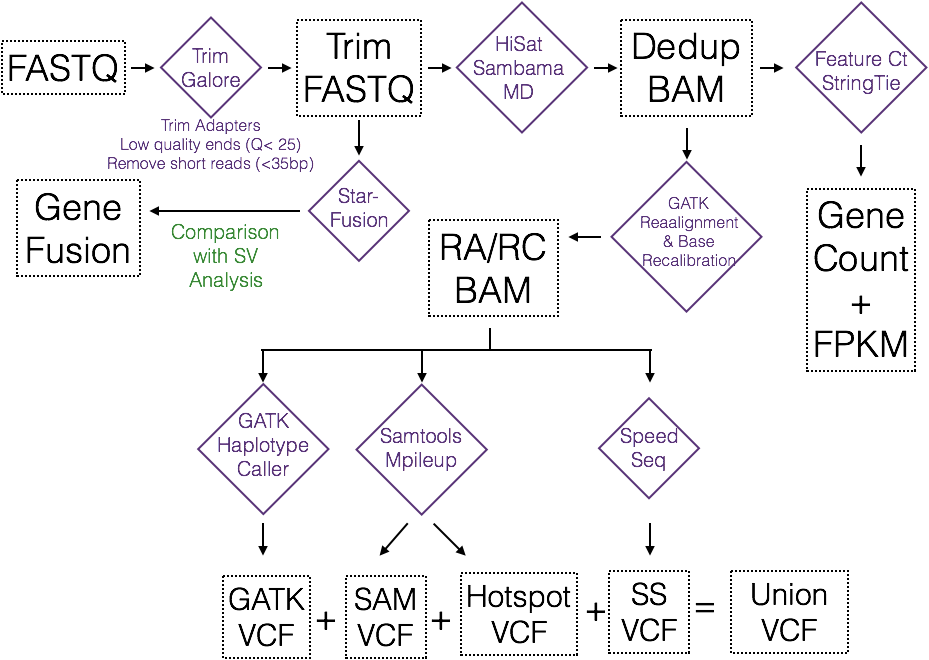
This SOP describes the analysis pipeline of RNA sequencing data. This pipeline includes (1) quality control, (2) variant calling analysis, (3) identification of fusion genes, and (4) statistical analyses of gene expression and isoform expression. The result R data of the statistical analysis can be visualized using a custom R shiny service.

3. Requirements

3.1. 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, and its prebuilt indexes for HISAT2, STAR, and STAR-Fusion; (2) a set of reference single nucleotide polymorphisms (SNPs) from the NCBI SNP database and (3) a set of somatic mutations from COSMIC database.

4. Procedures



4.1. 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 | trim\_galore |
| input | gzip fastq |
| output | trimmed gzip fastq |
| command | trim\_galore --paired -q 25 --illumina --gzip --length 35 ${read1} ${read2} |

4.2. Gene fusion detection

Identify gene fusions or fused transcripts using STAR-Fusion.

|  |  |
| --- | --- |
| executable | STAR-Fusion |
| input | STAR-Fusion index path, gzip fastq |
| output | fusion candidate list |
| command | STAR-Fusion --genome\_lib\_dir ${index\_path}/CTAT\_lib/ --left\_fq ${fq1} --right\_fq ${fq2} --output\_dir star\_fusion &> star\_fusion.err mv star\_fusion/star-fusion.fusion\_candidates.final.abridged ${pair\_id}.starfusion.txt |

4.3. Read mapping to human reference genome

Align RNA sequencing reads onto human reference genome and estimate the sequencing quality and mapping quality.

4.3.1. Use STAR

|  |  |
| --- | --- |
| executable | STAR, sambamba, fastqc |
| 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}.hisatout.txt  sambamba sort -t 30 -o ${pair\_id}.bam outAligned.sortedByCoord.out.bam  sambamba flagstat -t 30 ${pair\_id}.bam > ${pair\_id}.flagstat.txt  fastqc -f bam ${pair\_id}.bam |

4.3.2. Use HISAT2

|  |  |
| --- | --- |
| executable | hisat2, sambamba, fastqc |
| input | HISAT2 index path, gzip fastq |
| output | bam, quality report |
| command | hisat2 -p 30 --no-unal --dta -x ${index\_path}/${index\_name} -1 ${fq1} -2 ${fq2} -S out.sam 2> ${pair\_id}.hisatout.txt  sambamba view -t 30 -f bam -S -o output.bam out.sam  sambamba sort -t 30 -o ${pair\_id}.bam output.bam  sambamba flagstat -t 30 ${pair\_id}.bam > ${pair\_id}.flagstat.txt  fastqc -f bam ${pair\_id}.bam |

4.4. Alignment Statistics

Gather workflow quality metrics, run information and reference. Output summary file per sample.

|  |  |
| --- | --- |
| executable | sequenceqc\_rnaseq.pl |
| input | alignment flagstat.txt output file |
| output | alignment and workflow quality metrics |
| command | perl $baseDir/scripts/sequenceqc\_rnaseq.pl -r ${index\_path} \*.flagstat.txt |

4.5. Post alignment processing of BAM file

4.5.1. Mark PCR duplicates

Detect PCR duplicate alignments and edit their flag values.

|  |  |
| --- | --- |
| executable | sambamba, samtools |
| input | bam |
| output | duplicate-marked BAM file |
| command | sambamba markdup -t 20 -r ${sbam} ${pair\_id}.dedup.bam  samtools index ${pair\_id}.dedup.bam |

4.5.2. Preprocessing pipeline of GATK

(1) Repair the format problem of input BAM file, (2) add read group information, (3) handle N-split mapping of reads caused by RNA splicing, (4) base quality score recalibration (BQSR): detect systematic errors made by the sequencer when it estimates the quality score of each base call and build a model of covariation based on the data and a set of known variants, then adjust the base quality scores in the data based on the model.

|  |  |
| --- | --- |
| executable | picard.jar, sambamba, GenomeAnalysisTK.jar |
| input | bam, indexed reference genome file, dbSNP vcf |
| output | processed bam file (N-split, base-quality-recalibrated) |
| command | java -Xmx4g -jar \$PICARD/picard.jar CleanSam INPUT=${rbam} O=${pair\_id}.clean.bam  java -Xmx4g -jar \$PICARD/picard.jar AddOrReplaceReadGroups INPUT=${pair\_id}.clean.bam O=${pair\_id}.rg\_added\_sorted.bam SO=coordinate RGID=${pair\_id} RGLB=tx RGPL=illumina RGPU=barcode RGSM=${pair\_id}  sambamba index ${pair\_id}.rg\_added\_sorted.bam  java -Xmx4g -jar \$GATK\_JAR -T SplitNCigarReads -R ${index\_path}/hisat\_genome.fa -I ${pair\_id}.rg\_added\_sorted.bam -o ${pair\_id}.split.bam -rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 -U ALLOW\_N\_CIGAR\_READS  java -Xmx4g -jar \$GATK\_JAR -l INFO -R ${index\_path}/hisat\_genome.fa --knownSites ${dbsnp} -I ${pair\_id}.split.bam -T BaseRecalibrator -cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov ContextCovariate -o ${pair\_id}.recal\_data.grp -nt 1 -nct 30  java -Xmx4g -jar \$GATK\_JAR -T PrintReads -R ${index\_path}/hisat\_genome.fa -I ${pair\_id}.split.bam -BQSR ${pair\_id}.recal\_data.grp -o ${pair\_id}.final.bam -nt 1 -nct 8 |

4.6. Variant calling using SAMtools

Call variants with mapping quality >= 50 and base calling quality >= 20. Left-align and normalize indels; check if reference alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows.

|  |  |
| --- | --- |
| executable | samtools, bcftools, vcf-concat, vcf-sort, vcf-annotate |
| input | bam, indexed reference genome |
| output | variant calls vcf |
| command | samtools mpileup -t 'AD,DP,INFO/AD' -ug -Q20 -C50 -f ${index\_path}/hisat\_genome.fa ${gbam} | bcftools call -vmO z -o ${pair\_id}.sam.ori.vcf.gz  vcf-concat ${pair\_id}.sam.ori.vcf.gz | vcf-sort |vcf-annotate -n --fill-type | bcftools norm -c s -f ${index\_path}/hisat\_genome.fa -w 10 -O z -o ${pair\_id}.sam.vcf.gz - |

4.7. Hotspot variant identification

List COSMIC variants with variant allele fraction > 0.01 and depth > 50. Left-align and normalize indels; check if reference alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows.

|  |  |
| --- | --- |
| executable | samtools, bcftools, vcf-annotate, SnpSift.jar, bgzip |
| input | bam, indexed reference genome FASTA file, cosmic.vcf.gz |
| output | hospot variant list vcf |
| command | samtools mpileup -d 99999 -t 'AD,DP,INFO/AD' -uf ${index\_path}/hisat\_genome.fa ${gbam} > ${pair\_id}.mpi  bcftools filter -i "AD[1]/DP > 0.01" ${pair\_id}.mpi | bcftools filter -i "DP > 50" | bcftools call -m -A | vcf-annotate -n --fill-type | bcftools norm -c s -f /project/shared/bicf\_workflow\_ref/GRCh38/hisat\_genome.fa -w 10 -O z -o ${pair\_id}.lowfreq.vcf.gz -  java -jar SnpSift.jar annotate ${index\_path}/cosmic.vcf.gz ${pair\_id}.lowfreq.vcf.gz | java -jar SnpSift.jar filter "(CNT[\*] > 0)" - | bgzip > ${pair\_id}.hotspot.vcf.gz |

4.8. Variant calling using SpeedSeq

Detect variants using SpeedSeq. Left-align and normalize indels; check if reference alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows.

|  |  |
| --- | --- |
| executable | speedseq, tabix, bcftools, vcf-annotate, bgzip |
| input | bam, indexed reference genome |
| output | variant calls in VCF |
| command | speedseq var -t \$SLURM\_CPUS\_ON\_NODE -o ssvar ${index\_path}/hisat\_genome.fa ${gbam}  tabix -f ssvar.vcf.gz  bcftools norm -c s -f /project/shared/bicf\_workflow\_ref/GRCh38/hisat\_genome.fa -w 10 -O v ssvar.vcf.gz 2> bcftools.err | vcf-annotate -n --fill-type |bgzip > ${pair\_id}.ssvar.vcf.gz |

4.9. Variant calling using GATK

List variants and confident non-variant sites and genotype variants. Left-align and normalize indels; check if reference alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows.

|  |  |
| --- | --- |
| executable | GenomeAnalysisTK.jar, vcf-annotate, bcftools, tabix |
| input | BAM file, indexed reference genome, dbSNP vcf |
| output | variant calls in VCF |
| command | java -Xmx64g -jar \$GATK\_JAR -R ${index\_path}/hisat\_genome.fa -D ${dbsnp} -T HaplotypeCaller -stand\_call\_conf 30 -stand\_emit\_conf 10.0 -A FisherStrand -A QualByDepth -A VariantType -A DepthPerAlleleBySample -A HaplotypeScore -A AlleleBalance -variant\_index\_type LINEAR -variant\_index\_parameter 128000 --emitRefConfidence GVCF -I ${gbam} -o ${pair\_id}.gatk.g.vcf -nct 2  java -Xmx64g -jar \$GATK\_JAR -R ${index\_path}/hisat\_genome.fa -D ${dbsnp} -T GenotypeGVCFs -o gatk.vcf -nt 4 --variant ${pair\_id}.gatk.g.vcf  vcf-annotate -n --fill-type gatk.vcf | bcftools norm -c s -f ${index\_path}/hisat\_genome.fa -w 10 -O z -o ${pair\_id}.gatk.vcf.gz - &> bcftools.out  tabix ${pair\_id}.gatk.vcf.gz |

4.10. Generate union

Integrate result variant lists generated by GATK, SAMtools, SpeedSeq variant callers, and hotspot algorithm. The priority order for ambiguous variants is SAMtools - SpeedSeq - GATK - Hotspot.

|  |  |
| --- | --- |
| executable | bedtools, bgzip, tabix, GenomeAnalysisTK.jar, uniform\_integrated\_vcf.pl, bcftools |
| input | GATK vcf, SAMtools, SpeedSeq variant caller, Hotspot, indexed reference genome |
| output | union gzip vcf |
| command | bedtools multiinter -i ${gatk} ${sam} ${ss} ${hs} -names gatk sam ssvar hotspot |cut -f 1,2,3,5 | bedtools sort -i stdin | bedtools merge -c 4 -o distinct > ${fname}\_integrate.bed  bedtools intersect -header -v -a ${hs} -b ${sam} |bedtools intersect -header -v -a stdin -b ${gatk} | bedtools intersect -header -v -a stdin -b ${ss} | bgzip > ${fname}.hotspot.nooverlap.vcf.gz  tabix ${fname}.hotspot.nooverlap.vcf.gz  tabix ${gatk}  tabix ${sam}  tabix ${ss}  java -Xmx32g -jar \$GATK\_JAR -R ${index\_path}/hisat\_genome.fa -T CombineVariants --filteredrecordsmergetype KEEP\_UNCONDITIONAL --variant:gatk ${gatk} --variant:sam ${sam} --variant:ssvar ${ss} --variant:hotspot ${fname}.hotspot.nooverlap.vcf.gz -genotypeMergeOptions PRIORITIZE -priority sam,ssvar,gatk,hotspot -o ${fname}.int.vcf  perl $baseDir/scripts/uniform\_integrated\_vcf.pl ${fname}.int.vcf  bgzip ${fname}\_integrate.bed  tabix ${fname}\_integrate.bed.gz  bgzip ${fname}.uniform.vcf  tabix ${fname}.uniform.vcf.gz  bcftools annotate -a ${fname}\_integrate.bed.gz --columns CHROM,FROM,TO,CallSet -h ${index\_path}/CallSet.header ${fname}.uniform.vcf.gz | bgzip > ${fname}.union.vcf.gz |

4.11. Annotate

|  |  |
| --- | --- |
| executable | samtools, tabix, bcftools annotate, snpEff, snfSift |
| input | Union gzip vcf, Exac vcf, dbSnp vcf, clinvar vcf, cosmic vcf |
| output | annot gzip vcf, stats txt, stats plot |
| command | samtools/intel/1.3  tabix ${unionvcf}  bcftools annotate -Oz -a ${index\_path}/ExAC.vcf.gz -o ${fname}.exac.vcf.gz --columns CHROM,POS,AC\_Het,AC\_Hom,AC\_Hemi,AC\_Adj,AN\_Adj,AC\_POPMAX,AN\_POPMAX,POPMAX ${unionvcf}  tabix ${fname}.exac.vcf.gz  bcftools annotate -Oz -a ${index\_path}/dbSnp.vcf.gz -o ${fname}.dbsnp.vcf.gz --columns CHROM,POS,ID,RS ${fname}.exac.vcf.gz  tabix ${fname}.dbsnp.vcf.gz  bcftools annotate -Oz -a ${index\_path}/clinvar.vcf.gz -o ${fname}.clinvar.vcf.gz --columns CHROM,POS,CLNSIG,CLNDSDB,CLNDSDBID,CLNDBN,CLNREVSTAT,CLNACC ${fname}.dbsnp.vcf.gz  tabix ${fname}.clinvar.vcf.gz  java -Xmx10g -jar \$SNPEFF\_HOME/snpEff.jar -no-intergenic -lof -c \$SNPEFF\_HOME/snpEff.config ${snpeff\_vers} ${fname}.clinvar.vcf.gz | java -jar \$SNPEFF\_HOME/SnpSift.jar annotate ${index\_path}/cosmic.vcf.gz - | java -Xmx10g -jar \$SNPEFF\_HOME/SnpSift.jar dbnsfp -v -db ${index\_path}/dbNSFP.txt.gz - | java -Xmx10g -jar \$SNPEFF\_HOME/SnpSift.jar gwasCat -db ${index\_path}/gwas\_catalog.tsv - |bgzip > ${fname}.annot.vcf.gz  tabix ${fname}.annot.vcf.gz  bcftools stats ${fname}.annot.vcf.gz > ${fname}.stats.txt  plot-vcfstats -s -p ${fname}.statplot ${fname}.stats.txt |

4.12. Count reads per gene using featureCounts

|  |  |
| --- | --- |
| executable | featureCounts |
| input | gtf, bam |
| output | gene read count table |
| command | featureCounts -s params.stranded -T 30 -p -g gene\_name -a ${gtf\_file} -o ${pair\_id}.cts ${dbam} |

4.13. Transcript assembly and quantification using StringTie

|  |  |
| --- | --- |
| executable | stringtie |
| input | BAM file, GTF file |
| output | gtf of potential transcripts (isoforms) |
| command | mkdir ${pair\_id}\_stringtie  cd ${pair\_id}\_stringtie  stringtie ../${dbam} -p 30 -G ../${stringtie\_gtf} -B -e -o denovo.gtf -A ../${pair\_id}.fpkm.txt |

4.14. Statistical analysis of differential gene expression using edgeR

|  |  |
| --- | --- |
| executable | concat\_cts.pl, dea.R, concat\_edgeR.pl |
| input | gene read count table |
| output | differentially expressed gene list |
| command | perl concat\_cts.pl -o ./ \*.cts  cp design.txt design.shiny.txt  cp geneset.gmt geneset.shiny.gmt  Rscript $baseDir/scripts/dea.R  Perl concat\_edgeR.pl \*.edgeR.txt |

4.15. Statistical analysis of differential isoform expression using Ballgown

|  |  |
| --- | --- |
| executable | build\_ballgown.R |
| input | StringTie GTF file |
| output | List of differentially expressed isoforms (transcripts) |
| command | Rscript build\_ballgown.R \*\_stringtie |

5. Implementation

The Workflow used in this SOP can be downloaded here:

<https://git.biohpc.swmed.edu/BICF/Astrocyte/rnaseq>

Usage:

nextflow run –with-timeline –w work\_directory rnaseq.nf –input input\_directory –output output\_directory –design design\_file\_name.txt &>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 reside, the output directory is the name of the folder where the output files will be written, the design\_file\_name.txt is the name of the sample sheet which contains sample identifiers and paths to fastq files. The nextflow.log is the file that keeps the log of the run and tracks workflow errors.

Design file

The design file is a tab delimited file. The following columns are mandatory

SampleMergeName

SampleID

SampleName

SubjectID

FullPathToFqR1

FullPathToFqR2

Post-Script Metrics Thresholds

Assure nextflow.log has no errors. If errors exist direct the command prompt to the folder where error occurred and run sh .command.sh to troubleshoot.

Gather the following metrics from files and assure thresholds are reached. If thresholds are not reached, flag the sample as failed and do not proceed with the sample.

File:

|  |  |
| --- | --- |
| Metric | Lower Threshold Values |
| Total\_Raw\_Count | 5,000,000 |
| Total\_Pairs | 5,000,000 |
| Read1\_Map | 2,000,000 |
| Read2\_Map | 2,000,000 |
| Map\_Rate | 0.90 |
| Concordant\_Rate | 0.90 |
| Protein\_Coding\_Perc |  |
| Rrna\_perc |  |
| Status | Pass |

6. Discussion

7. Related Documents & 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/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.
* <https://software.broadinstitute.org/gatk/>
  + McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303.
  + DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8.
  + Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1-33.
  + <http://gatkforums.broadinstitute.org/gatk>
* <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.

8. Revision History