Bioinformatics Pipelines

CSI NGS Portal currently has 14 bioinformatics pipelines implemented, covering 10 different types of NGS data from DNA, RNA, smallRNA, 4C, ChIP, RIP, SHAPE, circRNA, eCLIP and Bisulfite-treated DNA sequencing libraries:

1. DNA-Seq

This pipeline identifies and annotates somatic mutations (single nucleotide variations and indels) in the DNA of tumour samples. Both whole genome (WGS) and whole exome (WES) sequencing data can be both used as the input. Use of matched normal sample is highly recommended to increase the confidence to call the somatic events although this is not required. If matched normal sample is provided, filtering of the germline mutations present in the normal sample will be performed. Otherwise "tumour-only mode" is used which may include many false positives and should be used with caution. Tumor-only mode is useful only for specific purposes, which and further details about the somatic mutation calling pipeline are described on the GATK [1] website (https://software.broadinstitute.org/gatk/documentation/article?id=11136). Currently utilisation of Panel of Normals (PoN) is not supported. After mutation calling by Mutect2 [2], a comprehensive annotation of the mutations is performed by ANNOVAR [3] including its genomic location, predicted functional impact of the mutation, implication as a clinically-associated mutation, presence in copy number variations from healthy individuals, availability in public databases such as 1000 Genomes [4] and COSMIC [5], and so on.

2. RNA-Seq

This pipeline performs gene and isoform expression quantification from RNA-Seq data, as well as comparison of alternative splicing events across samples. In the case of strand-specific RNA-Seq data, all the analyses are done on both strands separately by using the strand information and the output data are provided for both forward and reverse strands (sense and antisense for gene expression). For gene expression, raw read counts are provided, which is the input for many downstream analyses such as differential gene expression (DESeq2 [6], EdgeR [7], etc.) and raw read counts are not directly comparable between different samples alone. A separate pipeline for differential gene expression analysis with DESeq2 [6] as well as gene set enrichment analysis with GSEA [8] is also available named as "Diff-Exp", which is described next. For the isoform expression, both read counts and Transcripts Per Million (TPM) are provided in a strand-specific manner (if available), and TPM values allow comparison between different samples. Significant alternative splicing events are provided in the "SPLICING" folder as pairwise comparison across all the samples submitted under the same job for 5 different types as described on the "Docs" page of the website.

3. Diff-Exp

This pipeline performs differential gene expression analysis by using DESeq2 [6] starting from raw read counts using the output of an RNA-Seq job. Therefore, it is required to run an RNA-Seq job first on the batch of samples, which will automatically be available to the Diff-Exp pipeline once finished. The differentially expressed genes are identified by comparing two groups of samples specified by the user and the samples under the same group are collapsed as replicates. Note that replicates are required for the estimation of dispersion, as treating single samples as

replicates is no longer supported by DESeq2 since v1.22. The group assignment of the samples is imported from the "Diff-Exp Group" column on the "Annotate" page, and can be changed to submit a new job for a different comparison of interest. Normalization, PCA and clustering analyses are still performed by regionReport [9] for all the samples together under the same RNA-Seq job, even though a subset of samples ("contrast" parameter in DESeq2) are used for the differential gene expression analysis. Therefore, for accurate results, all RNA-Seq samples under the same job should come from the same library/batch. Note that for strand-specific RNA-Seq, read counts only from the sense strand are used. Also note that to decrease the memory usage, only expressed genes above a certain threshold are retained for the differential expression analysis (total read counts more than 2 times of the number of samples), hence the number of genes in the output table may vary from one batch to another.

Once the differentially expressed genes are identified, pathway enrichment analysis by Reactome is performed on the up- and down-regulated genes separately, which is described in the next section. In addition, gene set enrichment analysis (GSEA) [8] is optionally available under this pipeline. Gene sets from the Molecular Signatures Database [10] (MSigDB) can be selected as the input, including all gene sets, 8 major collections (H: hallmark [11] as the default) and several sub-collections. For GSEA, the normalised read counts processed by DESeq2 and filtered for lowly expressed genes are used, and the same grouping of the samples are applied as above (Group1 vs Group2 as the "phenotype" parameter). Different options for the "permute" and "metric" parameters are used depending on the sample size, i.e. if the number of samples in either group is less than 3, "log2_Ratio_of_Classes" is applied rather than the default "Signal2Noise", similarly

if the number of samples in either group is less than 7, “gene_set” is applied rather than the default “phenotype”. The analysis results as a comprehensive report are directly viewable on the browser.

4. Pathway-Enrichment

This pipeline performs standalone pathway enrichment analysis based on Reactome [12] starting from a list of input gene ids (Entrez Gene ID [13] and/or HUGO Gene Symbol [14]). Several plots are generated for different representations of the enriched genes and the pathways, including barplot, dotplot, cnetplot, upsetplot, heatplot, emapplot and pmcplot available in the enrichplot package. This pipeline is also available as part of the “Diff-Exp” pipeline described above, where the input genes are the differentially expressed genes identified in the RNA-Seq samples provided by the user.

5. RNA-Editing

This pipeline, adopted from a previously published method [15], identifies specific nucleotide changes that occur on the RNA caused by a post- and/or co-transcriptional modification known as RNA editing. In mammals, RNA editing predominantly results in A->I(G) changes due to the deamination activity of the ADAR enzymes. This pipeline runs on RNA-Seq data alone without matched genomic DNA sequence by using a set of stringent filters to exclude potential false positives, such as known single nucleotide polymorphisms (SNPs) and spurious sites. In case of cell lines, a pre-compiled list of cell line specific DNA mutations can be optionally excluded from the editing sites, which should be done in case cell line mutations for the input samples are available. The final list of candidate editing sites are reported per sample in repetitive (Alu and Non-Alu) and non-repetitive (Unique) genomic regions. In addition, a single merged table is also

provided which allows to compare the editing sites (only A->G and C->T changes) across all the samples submitted under the same job, and includes annotation with ANNOVAR [3] for the genomic location and the predicted functional consequence of each variant. The reported editing sites should be further filtered by the coverage and the editing frequency with a proper cut-off for the downstream analyses, for example Coverage >= 20 and Mutation_Frequency > 0.1.

6. smallRNA

This pipeline, developed in-house, quantifies the expression of the smallRNA family including miRNAs, snoRNAs, tRNAs, rRNAs and piRNAs. The annotations of the smallRNAs are obtained as follows:

i. miRNAs

- downloaded from miRBase v21 [16], which is originally in hg38, the coordinates were then converted to hg19 using UCSC LiftOver [17] tool in Galaxy [18].

ii. snoRNAs, tRNAs and rRNAs

- downloaded from UCSC hg19 sno/miRNA, tRNA and rmsk tracks [19], respectively.

iii. piRNA

- downloaded from piRNABank [20].

Based on the above annotations, a smallRNA reference genome was prepared to which the raw reads are mapped by using NovoAlign (http://www.novocraft.com/products/novoalign/). The expression of the smallRNAs are quantified based on the Concise Idiosyncratic Gapped Alignment Report (CIGAR) string in the alignment bam file by using an in-house perl script. For the miRNAs, because the hairpin and the mature miRNAs share identical sequence but are of different length,

the reads are assigned according to the mapped read length. The reads which map to a sequence longer and shorter than 30bp are counted as hairpin and mature miRNA, respectively. To reduce the output file size, raw read counts only for the expressed smallRNAs are provided in the expression output file, i.e. those with 0 read count are omitted.

7. 4C-Seq

This pipeline identifies long-range genomic interaction regions generated by 4C-Seq experiment using the R package r3Cseq [21]. Briefly, for each sample/replicate the raw reads are aligned to the masked version of the reference genome (masked for the gap, repetitive and ambiguous sequences) for human (hg19) or mouse (mm10, mm9) species as downloaded from the R Bioconductor repository (BSgenome.Hsapiens.UCSC.hg19.masked, BSgenome.Mmusculus.UCSC.mm10.masked, BSgenome.Mmusculus.UCSC.mm9.masked). The viewpoint chromosome, the restriction enzyme (first cutter) to digest the genome, the reads count method and the primers (forward and reverse) are the required inputs from the user. The primers must be minimum 20 bases long and uniquely mapped to the reference genome on the specified viewpoint chromosome. To count the number of reads per region, in addition to the default method "Fragment" where the restriction fragments are considered, a non-overlapping window size in the range of 2-100kb can also be selected. The number of mapped reads for each fragment/window are then counted and normalised to obtain RPM (the reads per million per fragment/window) values to perform the statistical analysis. This pipeline works with or without control samples, and also with or without replicates, however, if replicates are provided, it is compulsory also to provide control samples. The output is a text file containing the interaction regions along with the statistics

and the overlapping genes, and a pdf report which provides plots for the visualisation of the interactions.

8. ChIP-Seq

This pipeline identifies and characterises genome-wide binding sites for DNA-protein interactions. Peak regions are called by using MACS2 [22] with default parameters, allowing user an option to choose for “regular” or “broad” peaks. Peak annotation and motif analysis are done by HOMER [23]. The visualisation of the data is provided by a UCSC track hub [24]. The tracks in the genome browser are normalised by shrinking the larger data set to fit the smaller one (same as the default behaviour of MACS2). The output is a text file containing the peak regions with gene annotations, motif enrichment analysis results and a link to the UCSC Genome Browser [17] to display the peaks as custom tracks. The peaks and the custom tracks are provided in both hg19 and hg38 by default.

9. RIP-Seq

This pipeline, developed in-house (unpublished work), identifies and characterises genome-wide binding sites for RNA-protein interactions. Reads from the RIP-Seq sample and its control are mapped against specified reference genome by STAR [25] with GENCODE [26] transcriptome annotation. The resulting alignments are separated into two parts: (1) Exonic part consisting of alignments belonging to GENCODE annotated transcripts. (2) Non-exonic part consisting of the other alignments. Based on the number of reads mapped to the transcriptome, the larger one of the RIP and the control is shrunk down linearly to fit the size of the smaller one, thereby producing the normalised read count. The read coverage of each position is estimated as the average of

normalised read counts within surrounding 150 bases. Based on the comparison of the read coverage between the IP and the control, sites with ≥ 2-fold enrichment and Poisson distribution p-value ≤ 10-5 are defined as peaks. Each peak is extended to surrounding areas until the fold enrichment dropped below 2 (note: a peak from the exonic part could span across multiple exons). Overlapping peaks are merged, and those ≤ 300bp in size are ignored. The summit of a peak is defined as the position of highest fold enrichment in the peak. This pipeline accepts only paired-end reads as the input. For each job, 1 experiment up to 10 replicates may be submitted, and each experiment targeting a different protein can be submitted as a different job. The output is text files containing the peak regions in exonic and non-exonic regions and a link to the UCSC track hub [24] to visualise the peaks.

10. SHAPE-Seq

This pipeline provides secondary structure information on RNA based on experimental constraints. The analysis is performed by using icSHAPE pipeline [27] and RNAfold [28] from the ViennaRNA package [29] with the default parameters unless otherwise selected. Briefly, after trimming for adapter sequences and removal of PCR duplicates, the reads are mapped to the selected human transcriptome by using bowtie2 [30]. The transcript abundance is estimated by using Reads Per Kilobase of transcript per Million (RPKM) values and reverse transcription (RT) stops are calculated in each transcript. The background and the target RT stops from the control (DMSO) and the treated (NAI) samples, respectively, are normalised to calculate the enrichment reactivity scores for all the transcripts. These enrichment scores are further filtered to select the candidate transcripts with valid scores as well as high hit coverage and base density, where the enrichment threshold can be set by the user. The secondary structures of the substrates are then

predicted with the SHAPE [31] reactivity scores as constraints to guide the structure prediction. The output is text files containing the reactivity scores before and after filtering, and pdf files depicting the secondary structures of the target RNAs.

11. rMATS

This pipeline identifies differential alternative splicing events from RNA-Seq data between test and control samples by using replicate Multivariate Analysis of Transcript Splicing (rMATS) [32] (http://rnaseq-mats.sourceforge.net/). The identified events are categorised as skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site, mutually exclusive exons (MXE) and retained intron (RI). This pipeline requires replicates as the input, and the output is text files for each category. This is a standalone tool for splicing analysis in addition to the in-house developed pipeline available under the RNA-Seq pipeline.

12. circRNA

This pipeline, developed in-house, identifies circRNAs based on chimeric junction reads from STAR [25] alignment and quantifies their expression as read counts. Both RNA-Seq and circRNA-Seq data can be used as the input. However, circRNA enriched libraries are strongly recommended for easier detection, i.e. polyA(-), rRNA-depleted, RNase R treated for linear RNA digestion etc. PolyA selected RNA-Seq data are not useful for circRNA detection, as circRNAs do not possess polyA tails. At the end of the pipeline, circRNAs identified from all the samples under the same job are merged into one file allowing easy comparison and filtering.

13. eCLIP-Seq

This pipeline identifies genomic locations of RNA-bound proteins. The output is a text file containing the normalised peak regions annotated with the overlapping genes. The peaks are identified by eCLIP [33, 34] pipeline and annotated by using ANNOVAR [3] for the genes and the genomic locations. For the annotation, the mid-point (i.e. the base at the centre) of the peaks are used rather than the entire region for simplicity.

14. Bisulfite-Seq

This pipeline identifies methylation pattern on bisulfite-treated genomic DNA. The leading bases and the adaptor sequences are trimmed from the reads by TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Then methylation calls are performed by using Bismark [35] with default parameters on the trimmed files, first by removing PCR duplicates by *deduplicate_bismark* script and then extracting the DNA methylation status on every cytosine site by *bismark_methylation_extractor* script. DNA methylation status are converted to bigWig format for the visualization of the data as custom tracks by a UCSC track hub [24]. The processing and summary reports are generated by *bismark2report* and *bismark2summary* scripts.

References

1. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res 2010;20:1297-1303.

2. Cibulskis K, Lawrence MS, Carter SL et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, Nat Biotechnol 2013;31:213-219.

3. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res 2010;38:e164.

4. Genomes Project C, Auton A, Brooks LD et al. A global reference for human genetic variation, Nature 2015;526:68-74.

5. Tate JG, Bamford S, Jubb HC et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Res 2019;47:D941-D947.

6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 2014;15:550.

7. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation, Nucleic Acids Res 2012;40:4288-4297.

8. Subramanian A, Tamayo P, Mootha VK et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 2005;102:15545-15550.

9. Collado-Torres L, Jaffe AE, Leek JT. regionReport: Interactive reports for region-level and feature-level genomic analyses, F1000Res 2015;4:105.

10. Liberzon A, Subramanian A, Pinchback R et al. Molecular signatures database (MSigDB) 3.0, Bioinformatics 2011;27:1739-1740.

11. Liberzon A, Birger C, Thorvaldsdottir H et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection, Cell Syst 2015;1:417-425.

12. Fabregat A, Jupe S, Matthews L et al. The Reactome Pathway Knowledgebase, Nucleic Acids Res 2018;46:D649-D655.

13. Maglott D, Ostell J, Pruitt KD et al. Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res 2007;35:D26-31.

14. Yates B, Braschi B, Gray KA et al. Genenames.org: the HGNC and VGNC resources in 2017, Nucleic Acids Res 2017;45:D619-D625.

15. Ramaswami G, Zhang R, Piskol R et al. Identifying RNA editing sites using RNA sequencing data alone, Nat Methods 2013;10:128-132.

16. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function, Nucleic Acids Res 2019;47:D155-D162.

17. Haeussler M, Zweig AS, Tyner C et al. The UCSC Genome Browser database: 2019 update, Nucleic Acids Res 2019;47:D853-D858.

18. Afgan E, Baker D, Batut B et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic Acids Res 2018;46:W537-W544.

19. Karolchik D, Hinrichs AS, Furey TS et al. The UCSC Table Browser data retrieval tool, Nucleic Acids Res 2004;32:D493-496.

20. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs, Nucleic Acids Res 2008;36:D173-177.

21. Thongjuea S, Stadhouders R, Grosveld FG et al. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data, Nucleic Acids Res 2013;41:e132.

22. Zhang Y, Liu T, Meyer CA et al. Model-based analysis of ChIP-Seq (MACS), Genome Biol 2008;9:R137.

23. Heinz S, Benner C, Spann N et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol Cell 2010;38:576-589.

24. Raney BJ, Dreszer TR, Barber GP et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser, Bioinformatics 2014;30:1003-1005.

25. Dobin A, Davis CA, Schlesinger F et al. STAR: ultrafast universal RNA-seq aligner, Bioinformatics 2013;29:15-21.

26. Frankish A, Diekhans M, Ferreira AM et al. GENCODE reference annotation for the human and mouse genomes, Nucleic Acids Res 2019;47:D766-D773.

27. Flynn RA, Zhang QC, Spitale RC et al. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE, Nat Protoc 2016;11:273-290.

28. Lorenz R, Hofacker IL, Stadler PF. RNA folding with hard and soft constraints, Algorithms Mol Biol 2016;11:8.

29. Lorenz R, Bernhart SH, Honer Zu Siederdissen C et al. ViennaRNA Package 2.0, Algorithms Mol Biol 2011;6:26.

30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2, Nat Methods 2012;9:357-359.

31. Merino EJ, Wilkinson KA, Coughlan JL et al. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE), J Am Chem Soc 2005;127:4223-4231.

32. Shen S, Park JW, Lu ZX et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, Proc Natl Acad Sci U S A 2014;111:E5593-5601.

33. Van Nostrand EL, Pratt GA, Shishkin AA et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), Nat Methods 2016;13:508-514.

34. Van Nostrand EL, Nguyen TB, Gelboin-Burkhart C et al. Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-end Enhanced Crosslinking and Immunoprecipitation (seCLIP), Methods Mol Biol 2017;1648:177-200.

35. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, Bioinformatics 2011;27:1571-1572.