rnaseq-methods

HiSeq Illumina sequencing will be performed on our behalf by ###. All samples will be indexed so that ### RNA pools can be run on a single lane of an 8-laned Illumina flow cell, providing an estimated ### million reads per sample [Note: aim for 20-30 million reads, 100bp paired-end].

All samples will be processed using RNA-seq pipeline implemented in the bcbio-nextgen project. Raw reads will be examined for quality issues using FastQC to ensure library generation and sequencing are suitable for further analysis. Adapter sequences, other contaminant sequences such as polyA tails and low quality sequences with PHRED quality scores less than five will be trimmed from reads using cutadapt[7]. Trimmed reads will be aligned to build XX of the XX genome, augmented with transcript information from Ensembl release XX using STAR[3].

Alignments will be checked for evenness of coverage, rRNA content, genomic context of alignments (for example, alignments in known transcripts and introns), complexity and other quality checks using a combination of FastQC, RNA-SeQC[2] and custom tools. Counts of reads aligning to known genes and isoforms will be generated by a combination of featureCounts[5], eXpress[9] and DEXSeq[1].

Novel transcripts will be identified via reference-guided assembly with Cufflinks[10], with novel transcripts filtered for coding potential agreement[11] with known genes to reduce false positive assemblies. Variant calls and RNA-editing events will be called from alignments using the GATK HaplotypeCaller, filtered with custom tools to remove false positive events due to alignment errors and other artifacts. RNA-editing events will be separated from SNPs by a combination of looking clusters of A->G or T->C events and known edit sites from RADAR[8].

Differential expression at the gene level will be called with DESeq2[6], which has been shown to be a robust, conservative differential expression calller. Isoform and exon-level calls are much more prone to false positives, and so an ensemble method combining calls that agree from both DEXSeq and EBSeq[4] will be used to call differential isoforms.

Gene-level differntial RNAseq results will be validated by qRT-PCR on a subset of genes. Isoform and event-level differential expression calls will be validated using semi-quantitative PCR to flag percent spliced in levels for the event.

[Add in whatever particular downstream stuff is going to happen].

References

[1] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, October 2012.

[2] David S DeLuca, Joshua Z Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler, and Gad Getz. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics (Oxford, England)*, 28(11):1530–1532, June 2012.

[3] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

[4] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics (Oxford, England)*, 29(8):1035–1043, April 2013.

[5] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–930, April 2014.

[6] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014.

[7] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, February 2011.

[8] Gokul Ramaswami and Jin Billy Li. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, 42(Database issue):D109–13, January 2014.

[9] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, January 2013.

[10] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010.

[11] Liguo Wang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, 41(6):e74–e74, April 2013.