A reading list for next generation sequencing

Cavan Reilly

October 28, 2019

Early Publication

Wang, Sandberg, Luo et al. (2008) "Alternative isoform regulation in human tissue transcriptomes", *Nature*, 456, 470–476.

Useful web resource

SEQanswers: `seqanswers.com`

A typical RNA-seq example: the cow data set

McCabe MS, Waters SM, Morris DG, et al. (2012), "RNA-seq analysis of differential gene expression in liver from lactating dairy cows divergent in negative energy balance", *BMC Genomics*, 13.

Relation to microarrays

Marioni, Mason, Mane, et al. (2008), "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays", *Genome Research*, 18, 1509–1517.

Su, Li, Chen, et al. (2011) "Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys", *Chemical Research in Toxicology*, 2011, 1486-93.

Bowtie

Burrows, M. and Wheeler, D.J. (1994), "A block-sorting lossless data compression algorithm"

Ferragina, P. and Manzini, G. (2000), "Opportunistic data structures with applications"

Langmead B, Trapnell C, Pop M, et al. (2009), "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome", *Genome Biology*, 10:R25.

Langmead B, Salzberg SL (2012), "Fast gapped-read alignment with Bowtie 2", *Nature Methods*, 9(4):357–359.

SAM file specification

Li H, Handsaker B, Wysoker A, et al. (2009), "The sequence alignment/map format and SAMtools", *Bioinformatics*, 25, 2078–2079.

SAMtools website: `http://samtools.sourceforge.net/` and `http://www.htslib.org/`.

detailed reference: `http://samtools.sourceforge.net/SAM1.pdf`

RNA-Seq data analysis: TopHat, Cufflinks and more

Trapnell C, Pachter L, Salzberg SL (2009), "TopHat: discovering splice junctions with RNA-Seq", *Bioinformatics*, 25, 1105–1111.

Trapnell C, Williams BA, Pertea G, et al. (2010), "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation", *Nature Biotechnology*, 28, 511–515.

Trapnell C, Roberts A, Goff L, et al. (2012), "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", *Nature Protocols*, 7, 562–578.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and Salzberg SL (2013), "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions", *Genome Biology*, 14:R36.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT and Salzberg SL (2013), "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads", *Nature Biotechnology*, doi:10.1038/nbt.3122.

Kim D, Langmead B and Salzberg SL (2015), "HISAT: a fast spliced aligner with low memory requirements", *Nature Methods*, 12, 357-360.

Pertea M, Kim D, Pertea GM, Leek JT and Salzberg SL (2016), "Transcript-level expression analysis of RNA-seq experiments with HISAT, Stringtie and Ballgown", *Nature Protocols*, 11, 1650–1667.

Bray N, Pimentel H,Melsted P and Pachter L (2016), "Near-optimal probabilistic RNA-seq quantification", *Nature Biotechnology*, 34, 525-527.

Fu J, Frazee AC, Collado-Torres L, Jaffe AE and Leek JT (2017). ballgown: Flexible, isoform-level differential expression analysis. R package version 2.10.0.

RNA-Seq data analysis with negative binomial based methods: edgeR, DESeq and DEXSeq

Robinson MD, Smyth GK (2007), "Moderated statistical tests for assessing differences in tag abundance", *Bioinformatics*, 23, 2881–2887.

Robinson MD, McCarthy DJ, Smyth GK (2009), "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics*, 26, 139–140.

Anders S, Huber W (2010), "Differential expression analysis for sequence count data", *Genome Biology*, 11:R106.

McCarthy DJ, Chen Y, Smyth GK (2012), "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation", *Nucleic Acids Research*, 40, 4288–4297.

Anders S, Reyes A, Huber W (2012), "Detecting differential usage of exons from RNA-seq data", *Genome Research*, 22, 2008–2017.

Love MI, Huber W, Anders S (2014), "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", *Genome Biology*, 15:550.

DNA-Seq data analysis: GATK and Genome STRiP

McKenna A, Hanna M, Banks E et al. (2010), "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequence data", *Genome Research*, 20, 1297–1303.

DePristo M, Banks E, Poplin R, et al (2010), "A framework for variation discovery and genotyping using next-generation DNA sequencing data", *Nature Genetics*, 43, 491–498.

Handsaker RE, Korn JM, Nemesh J, et al. (2011), "Discovery and genotyping of genome structural polymorphism by sequencing on a population scale", *Nature Genetics*, 43, 269–276.

Nielsen R, Paul JS, Albrechtsen A et al. (2011), "Genotype and SNP calling from next-generation sequencing data", *Nature Reviews: Genetics*, 12, 443–451.

Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, et al. (2013), "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline", *Current Protocols in Bioinformatics*, 43:11.10.1-11.10.33.

Microbiomics

Wang Q, Garrity GM, Tiedje JM, et al. (2007), "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy", *Applied and Environmental Microbiology*, 73, 5261–5267.

Cole JR, Wang Q, Cardenas E, et al. (2008), "The ribosomal database project: improved alignments and new tools for rRNA analysis", *Nucleic Acids Research*, 37, D141–D145.

Edgar, R.C. (2013), "UPARSE: Highly accurate OTU sequences from microbial amplicon reads", *Nature Methods*, 10(10):996–998.