Notes for RNAseq analysis course 20111129

Robert.Lyle@medisin.uio.no

Where to get help:

Google

Type software name with no options will often give a help page. Or try options: -h or -help

Files for course are organised as follows:

There are four directories

ref	contains reference files (sequence, and gene models)
reads	contains the two fastq sequence files
results	where results will be written to in the examples
precomputed	contains above directories, results contains precomputed results

The steps are:

1. Prepare an index file for the reference sequence for alignment
2. Run TopHat to align sequence reads to the reference, considering provided gene models
3. View coverage and splice junctions on the genome
4. Count number of reads for each gene, a measure of gene expression level

1. Prepare bowtie index

```
bowtie-build chr21.fa chr21
```

2. Run tophat to do alignment

There are a lot of options here, check the manual online, or type tophat at the command line, for help

tophat -r 50 -G ref/Homo_sapiens_chr21.GRCh37.64.gtf --no-novel-juncs --library-type=fr-unstranded -p 1 -o results ref/chr21 reads/sample_1_hsa21_R1.fastq reads/sample_1_hsa21_R2.fastq

3. View coverage (bedtools package)

genomeCoverageBed -split -bg -ibam accepted_hits_np.bam > accepted_hits_coverage.bed

Sort hits, and make sam file

samtools sort -n accepted_hits.bam accepted_hits_n

samtools view -h accepted_hits_n.bam > accepted_hits_n.sam

View results: UCSC, IGV

http://genome.ucsc.edu/

http://genome.ucsc.edu/FAQ/FAQformat.html

http://www.broadinstitute.org/igv/

4. Count expression levels

htseq-count -m intersection-strict --stranded=no -t exon -i gene_id results/accepted_hits_n.sam ref/Homo_sapiens_chr21.GRCh37.64.gtf > results/sample_1_hsa21.count

Info and links to software

Basic unix tutorial (there are many - use Google!)

http://www.ee.surrey.ac.uk/Teaching/Unix/

Software - sequence alignment/manipulation

http://tophat.cbcb.umd.edu/

http://bowtie-bio.sourceforge.net/index.shtml

http://samtools.sourceforge.net/

http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html

http://code.google.com/p/bedtools/

For those interested in looking at differential expression

http://www.bioconductor.org/packages/release/bioc/html/DESeq.html

Info on file formats (very important)

http://en.wikipedia.org/wiki/FASTQ_format

http://samtools.sourceforge.net/SAM1.pdf

http://genome.ucsc.edu/FAQ/FAQformat.html

General

http://seqanswers.com/

http://seqanswers.com/wiki/SEQanswers