High-Throughput Sequencing data analysis Tools (htSeqTools)

Bioconductor Developer meeting
EMBL. Heidelberg. November 2010

Oscar Reina – Biostatistics and Bioinformatics Unit
NGS data analysis needs
htSeq and htSeqTools

Bioconductor package(s) (expected 2011). Intended as workflow processing pipeline for our Solexa-Illumina Genome Analyzer experiments data.

- **htSeq**: David Rossell. Functions for NGS data analysis. Extensive use of RangedData objects (**IRanges** package).
- **htSeqTools**: Convenience wrapper around htSeq and other NGS data processing functions to implement a NGS pre-processing pipeline.
htSeqTools overview

**Linear workflow** with the most common tasks involved in Solexa-Illumina GA data processing after delivered by the Illumina pipeline.

- **Input**: FASTQ sequence ASCII files (s_x_sequence.txt) as delivered by last step of Illumina GA pipeline (GERALD).

- **Output**: Processed data for further specific analysis (ChIP-Seq, RNA-Seq, etc) and report set to assess for experiment quality control
htSeqTools workflow structure

1. Setting **workflow options** and loading run information.
2. **Alignment** of sequence files against reference genome.
3. **Import** of aligned files and storing genomic interval data in R.
4. **Compression** of original data files
5. **Pre-processing steps**: common tasks usually performed to prepare data for further analysis (*duplicate reads, strand shift bias, read extension...*)
6. Generation of sequence coverage and quality control reports
7. Generation of track files for visualization in external genome browser
htSeqTools workflow structure (details...)

1. **Nested list of pre-configured parameters** (setwd(‘…’) and go) and array-like sampleinfo.txt file sets it up for running. That is all.

2. **Parallelized Alignment** of FASTQ sequence files against target genome with Bowtie.


4. **Parallelized background compression** of original sequence and bowtie files (.tar.bz2).

5. **Parallel pre-processing tasks**: duplicate reads policy, peak shift bias correction (ChIP-Seq), read extension. Wrapping around htSeq functions. Rdata saving of pre-processed RangedDataList object (clean sequences).

6. Generation of **Linear/Hilbert sequence coverage** and Gini/Lorenz QC PNG/PDF reports.

setwd('/Volumes/biostats/routines/R/htSeq_wrapper/exampleData')

htSeqPars <- loadDefaultOptions()
htSeqPars$bowtieoptions$bowtiepath <- '~/soft/biostats/bowtie/bowtie-0.12.1/bowtie'
htSeqTools(htSeqPars)

### HTSEQ: START of htSeq analysis at 2010-10-20 10:48:47 ###
*** Files and folders OK ***
*** Performing Bowtie alignment on s_2_sequence.txt ***
*** Bowtie parameters: -n 2 -p 6 -m 1 --solexa1.3-quals ***
*** Bowtie versions is ~/soft/biostats/bowtie/bowtie-0.12.1/bowtie ***
*** Bowtie reference genome used: /Volumes/biostats/databases/bowtie_indexes/dm3 ***
# reads processed: 6219895
# reads with at least one reported alignment: 4133715 (66.46%)
# reads that failed to align: 195739 (3.15%)
# reads with alignments suppressed due to -m: 1890441 (30.39%)
Reported 4133715 alignments to 1 output stream(s)
*** Reading Bowtie Aligned files s2_bowtie.txt ***
*** Saving all seqs Interval info as RangedDataList in seqs.RData ***
*** Compressing Sequence files ***
*** Compressing Bowtie files ***
*** Aligning Peaks for Seqs s2 for +/- strand bias with 1000 Peaks for shift estimation and 150 Bandwidth using 6 CPU cores ***
Estimated shift size is 18.96383
*** Removing duplicate reads from Seqs s2 using 6 CPU cores ***
*** rangeExtension is set to Zero, so no extension is done ***
*** Saving Clean Seqs in seqsProcessed.RData ***
*** Exporting TDF IGV files for seqs s2. Options: -z 7 -w 25 -e 0 -f p10,p90,min,max,mean,median ***
*** Writing s2.htseq.aligned... ***
*** Generating .TDF files for seqs s2. May take a moment... ***
*** Removing temporary .aligned files... ***
.TDF files available at /Volumes/biostats/routines/R/htSeq_wrapper/exampleData/
*** HTSEQ: END of htSeq analysis at 2010-10-20 10:55:10 ***