IRB BARCELONA

INSTITUTE FOR RESEARCH IN BIOMEDICINE

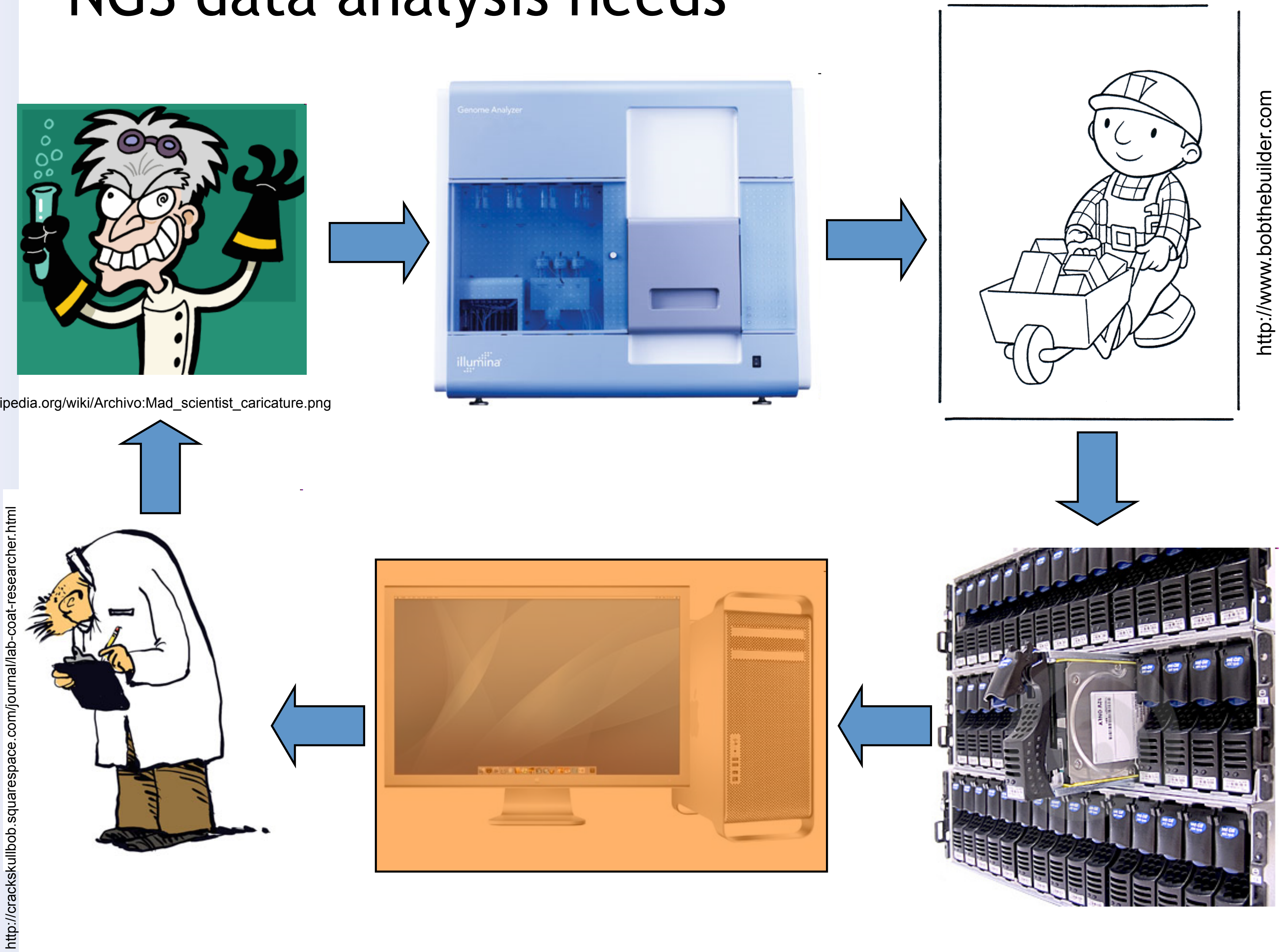
High-Throughput Sequencing data analysis Tools (htSeqTools)

Bioconductor Developer meeting

EMBL. Heidelberg. November 2010

Oscar Reina – Biostatistics and Bioinformatics Unit

NGS data analysis needs



http://es.wikipedia.org/wiki/Archivo:Mad_scientist_caricature.png

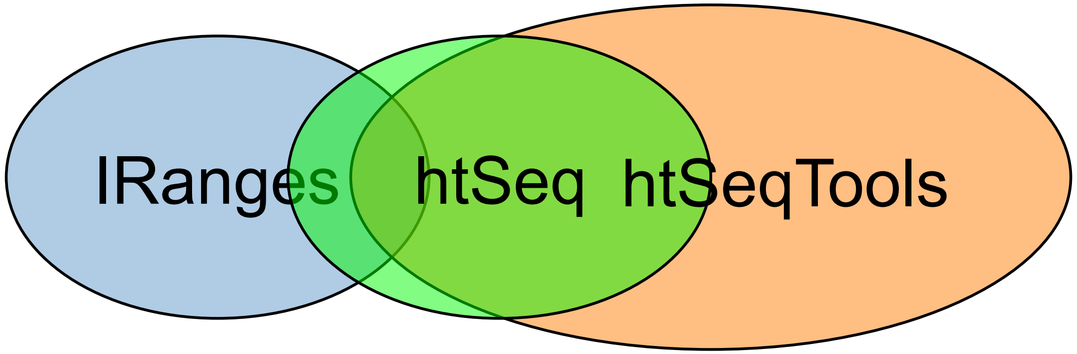
http://www.bobthebuilder.com

http://crackskullbob.squarespace.com/journal/lab-coat-researcher.html

htSeq and htSeqTools

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

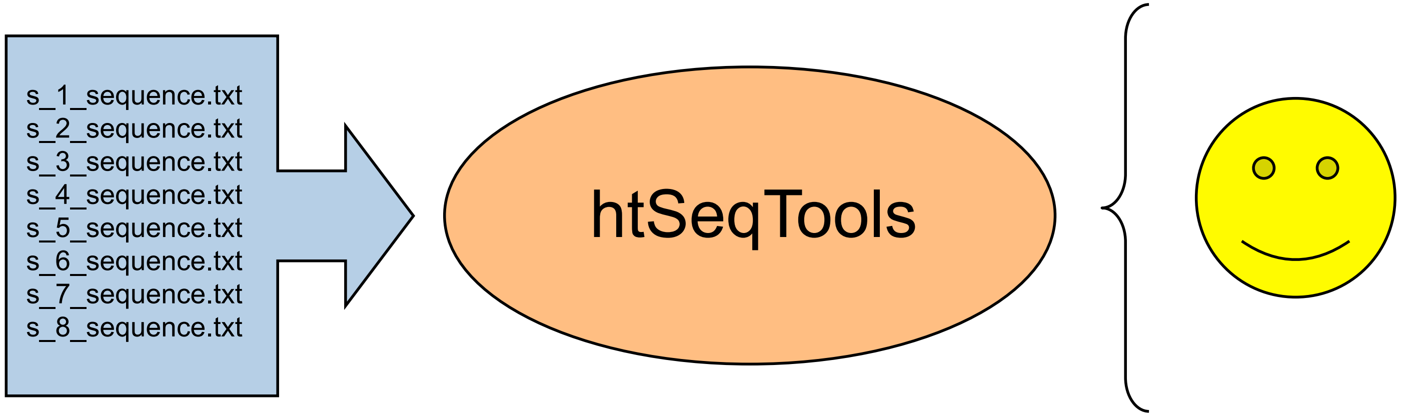
- **htSeq**: David Rossell. Functions for NGS data analysis. Extensive use of **RangedData** objects (**IRanges** package).
- **htSeqTools**: Convenience wrapper around htSeq and other NGS data processing functions to implement a **NGS pre-processing pipeline.**



htSeqTools overview

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

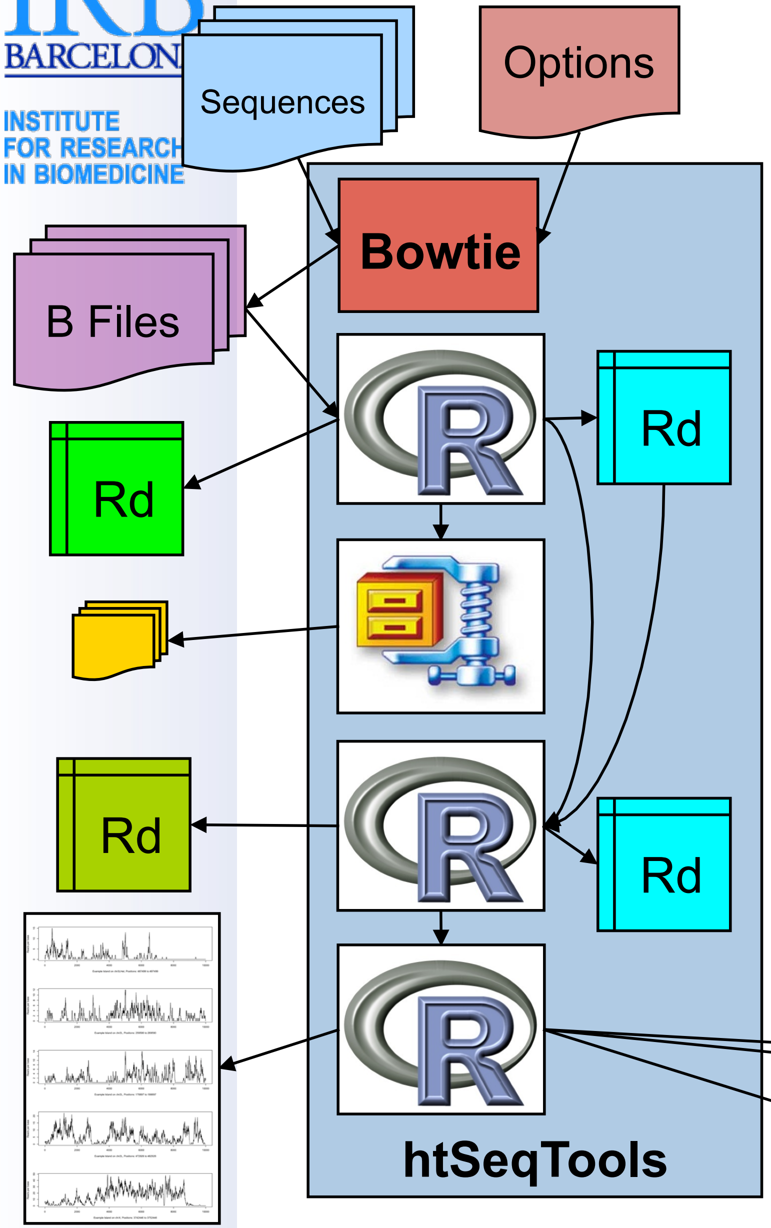
- **Input**: FASTQ sequence ASCII files (s_x_sequence.txt) as delivered by last step of Illumina GA pipeline (GERALD).
- **Output**: Processed data for further specific analysis (ChIP-Seq, RNA-Seq, etc) and report set to assess for experiment quality control



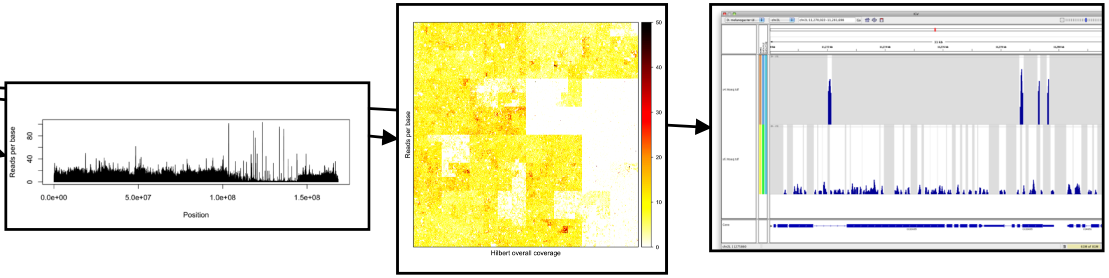
s_1_sequence.txt
s_2_sequence.txt
s_3_sequence.txt
s_4_sequence.txt
s_5_sequence.txt
s_6_sequence.txt
s_7_sequence.txt
s_8_sequence.txt

htSeqTools

htSeqTools workflow structure

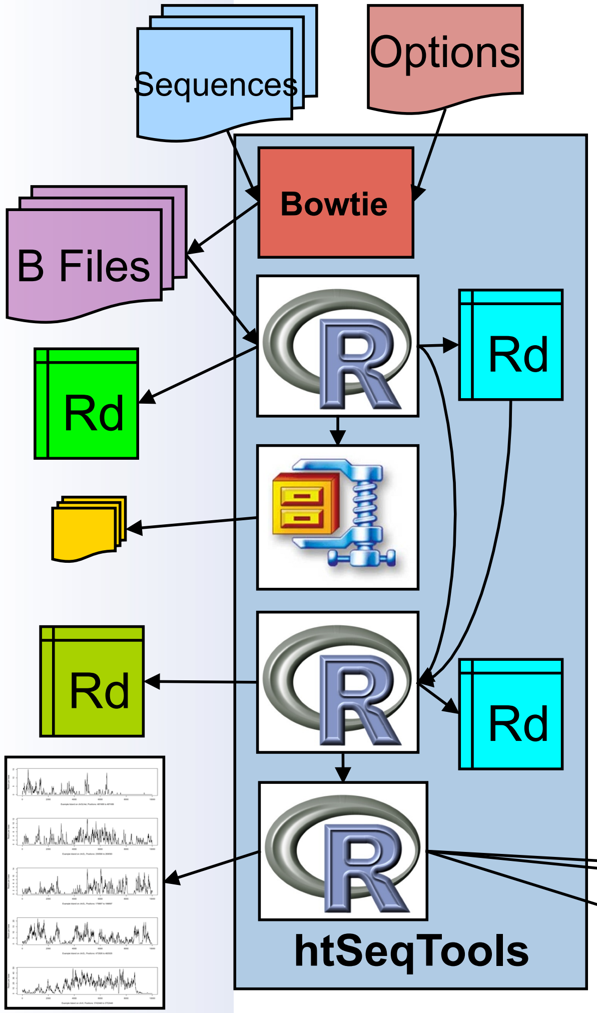


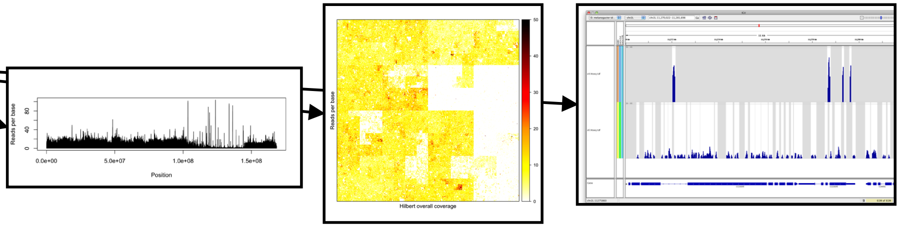
1. Setting **workflow options** and loading **run information.**
2. **Alignment** of sequence files against reference genome.
3. **Import** of aligned files and storing **genomic interval data** in R.
4. **Compression** of original data files
5. **Pre-processing steps**: common tasks usually performed 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.**
3. **Paralellized Import** of Bowtie aligned data into R and storing genomic interval information as **RangedDataList** object. Raw sequences .Rdata saving.
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.
7. Generation of **.TDF** binary track files for Broad's **IGV Genomics Viewer.**





```
> setwd('/Volumes/biostats/routines/R/htSeq_wrapper/exampleData')
> ###############################################################################
> htSeqPars <- loadDefaultOptions()
> htSeqPars$bowtieoptions$bowtiepath <- '~/soft/biostats/bowtie/bowtie-0.12.1/bowtie‘
> htSeqTools(htSeqPars)
> ### [1] "### HTSEQ: START of htSeq analysis at 2010-10-20 10:48:47 ###"
> ### [1] "*** Files and folders OK ***"
> ### [1] "*** Performing Bowtie alignment on s_2_sequence.txt ***"
> ### [1] "*** Bowtie parameters: -n 2 -p 6 -m 1 --solexa1.3-quals  ***"
> ### [1] "*** Bowtie versions is ~/soft/biostats/bowtie/bowtie-0.12.1/bowtie ***"
> ### [1] "*** 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)
> ### [1] "*** Reading Bowtie Aligned files s2_bowtie.txt ***"
> ### Reading s2_bowtie.txt...
> ### [1] "*** Saving all seqs Interval info as RangedDataList in seqs.RData ***"
> ### [1] "*** Compressing Sequence files ***"
> ### [1] "*** Compressing Bowtie files ***"
> ### [1] "*** 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
> ### [1] "*** Removing duplicate reads from Seqs s2 using 6 CPU cores ***"
> ###               s2
> ### 4133715, 4132894
> ### [1] "*** rangeExtension is set to Zero, so no extension is done ***"
> ### [1] "*** Saving Clean Seqs in seqsProcessed.RData ***"
> ### [1] "*** Exporting TDF IGV files for seqs s2. Options: -z 7 -w 25 -e 0 -f
p10,p90,min,max,mean,median ***"
> ### [1] "*** Writing s2.htseq.aligned... ***"
> ### [1] "*** Generating .TDF files for seqs s2. May take a moment... ***"
> ### [1] "*** Removing temporary .aligned files... ***"
> ### [1] "*** .TDF files available at /Volumes/biostats/routines/R/htSeq_wrapper/exampleData/
> ### [1] "### HTSEQ: END of htSeq analysis at 2010-10-20 10:55:10 ###"
```