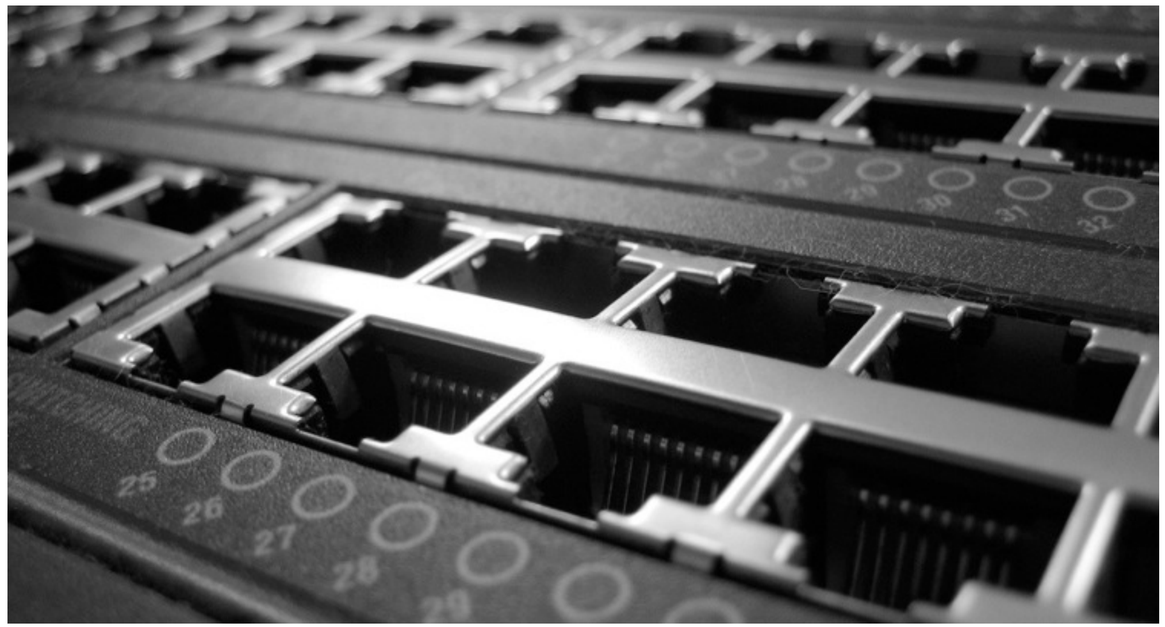


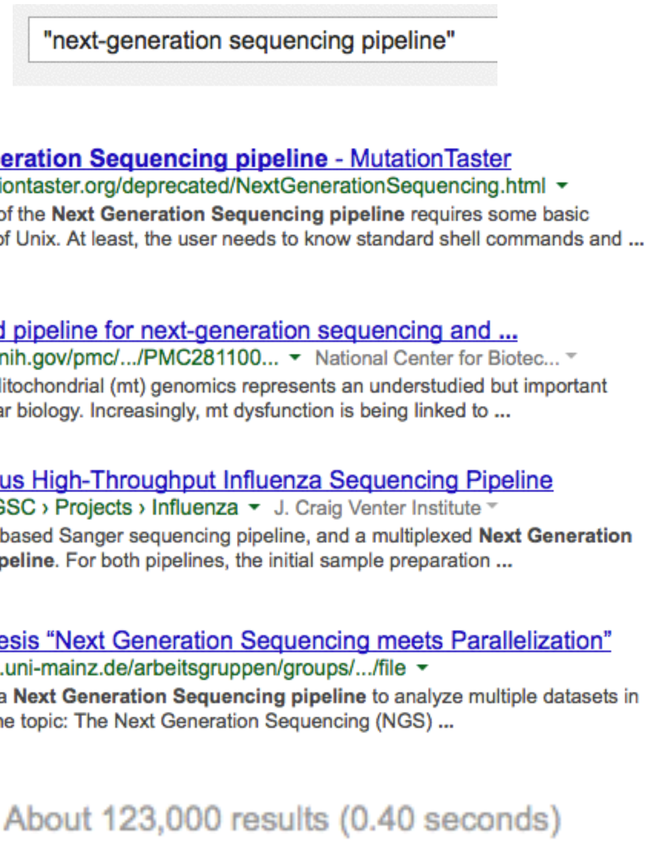
A Framework for RNA-seq DE comparisons

Center for Health Bioinformatics, Harvard School of Public Health

Health Sciences and Technology, Harvard-MIT

Pipeline proliferation

- No consensus on best practices for the majority of analyses
- Low interoperability between tools
- Many different types of experiments
- Analyses are very complex
- End users are often not programmers



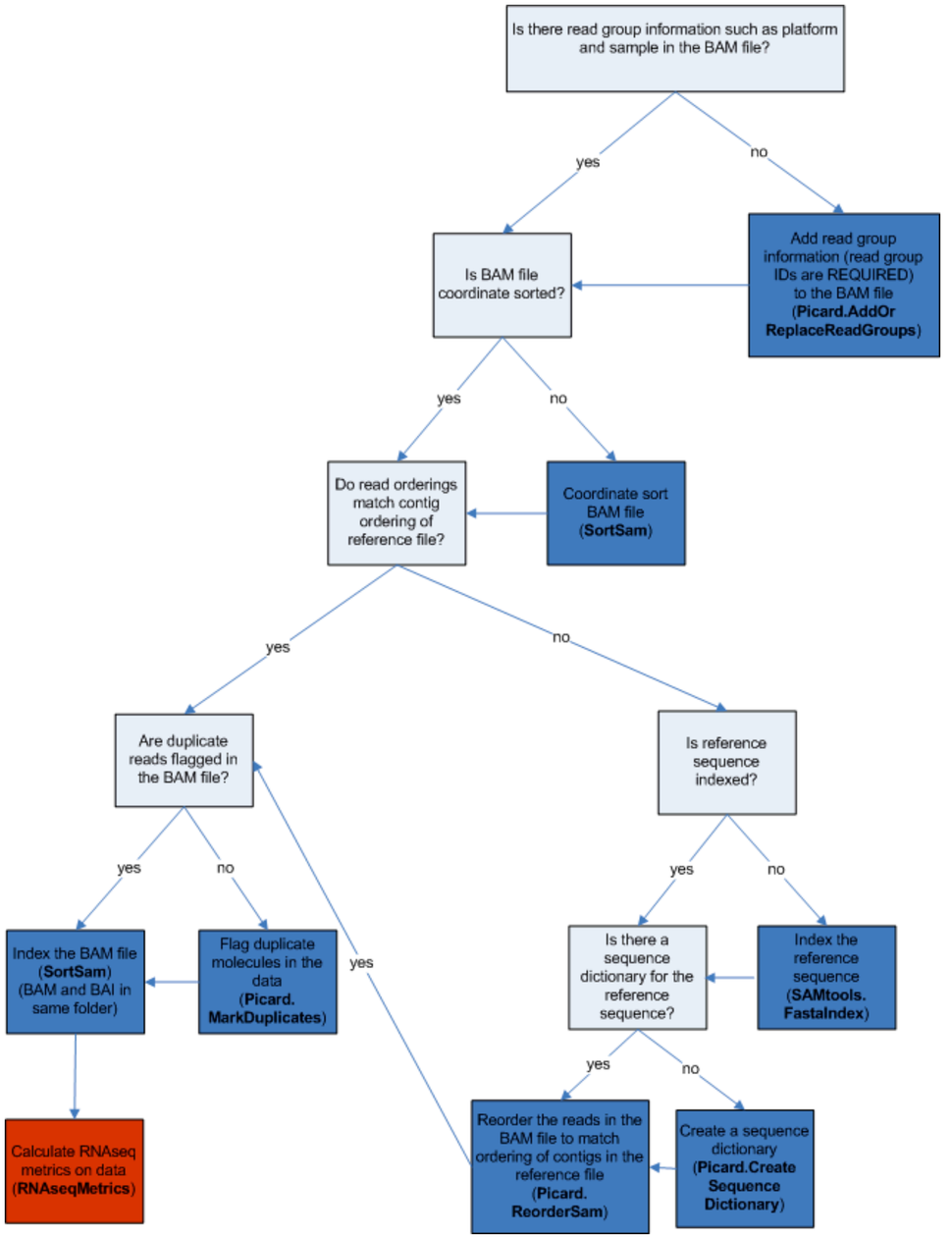
Complexity

Inherent

- Essential to task at hand
- Challenging
- Fun

Incidental

- Caused by our solutions
- Challenging
- Very frustrating
- Not science

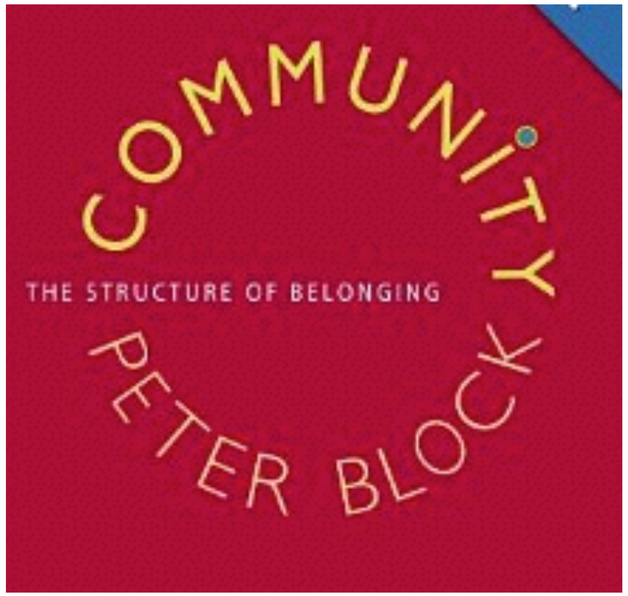


Incidental complexity

- Installation
  - Tools
  - Data
- Choice
- Parameters

Development goals of bcbio-nextgen

- Community developed and driven
- Quantifiable
- Scalable
- Easy to install. Easy to use.
- Well-documented
- Easy to extend



Nick Loman
@pathogenomenick

Follow

Loman's law of bioinformatics: If you haven't found at least one bug in someone's pipeline then you don't understand it properly yet.

Community

PUBLIC chapmanb / **bcbio-nextgen**

Unwatch 22 | Unstar 74 | Fork 36

Best-practice pipelines for fully automated high throughput sequencing analysis
https://bcbio-nextgen.readthedocs.org

2,082 commits | **1** branch | **12** releases | **12** contributors

branch: **master** **bcbio-nextgen** /

Reduce duplicated 'algorithm' section in each data item to have small… ···

chapmanb authored 10 hours ago — latest commit 2288dd263b

bcbio	Reduce duplicated 'algorithm' section in each data item to have small…	10 hours ago
conda-recipe	Update import tests for conda package with new parallel structure	24 days ago
config	Use AlienTrimmer instead of cutadapt.	7 days ago
docs	Add more in depth description of the RNA-seq output.	8 days ago
scripts	Avoid issues with numpy 1.7/1.8 discrepency in automated conda instal…	15 days ago
tests	Ensure pipeline tests work on installations in non-standard locations…	6 days ago
.gitignore	Automatically write bcbio-nextgen version during install, including g…	6 months ago
.travis.yml	Merge slurm and SciLifeLab introduction from Roman. Fixes #64	7 months ago

Code | Issues 21 | Pull Requests 2 | Pulse | Graphs | Network

SSH clone URL

git@github.com:chap

You can clone with HTTPS, SSH, or Subversion.

Clone in Desktop

Download ZIP

Community

Installation

Tools

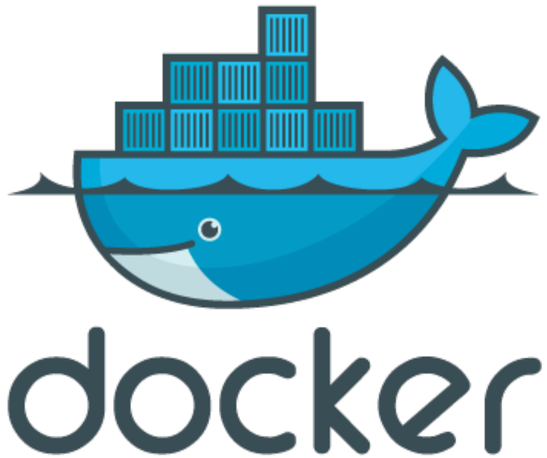
- compatible
- versioned
- no sudo, no problem
- sandboxed

Data

- coherent
- versioned

Current target environment

- Cluster scheduler
  - Torque
  - SLURM
  - SGE
  - LSF
- Shared filesystem
  - NSF
  - Lustre
- Local temporary disk
  - SSD



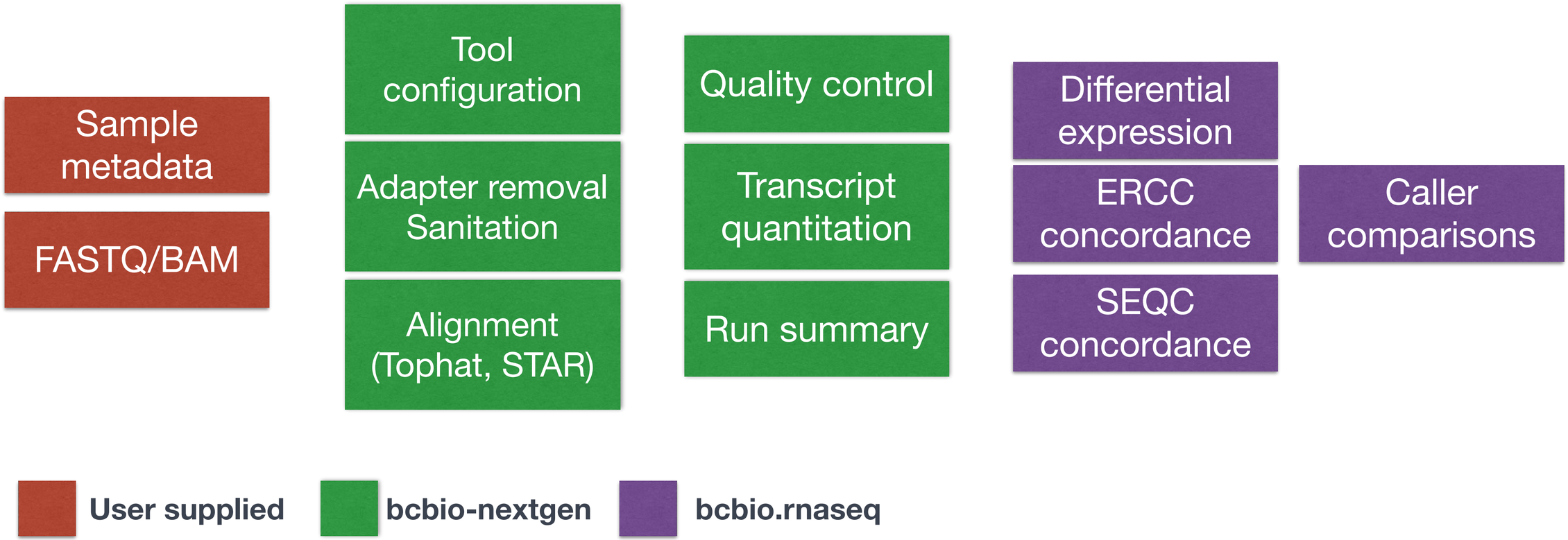
Virtualization and reproducibility

Ease of use

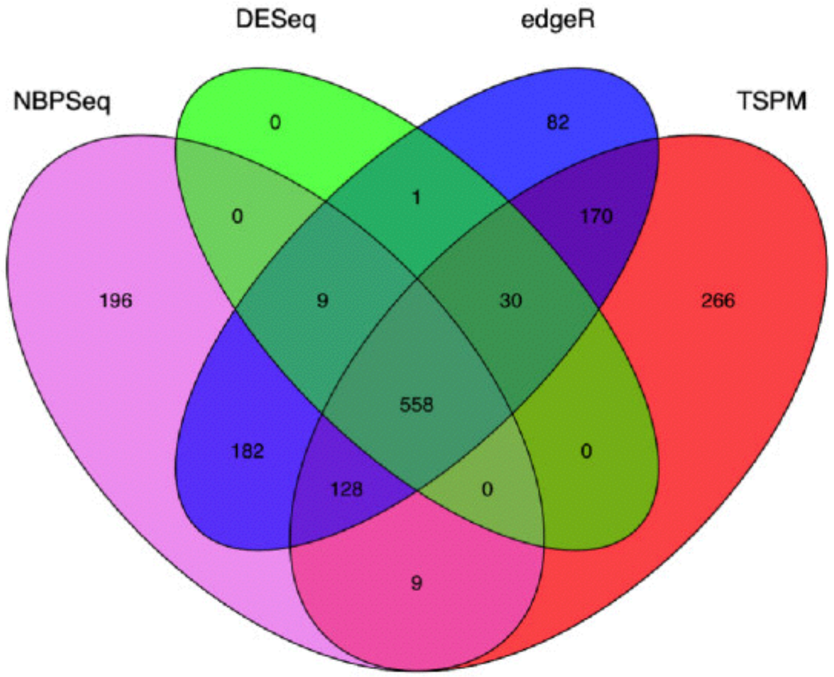
- Tools come pre-configured
- Analysis involves
  - Putting FASTQ/BAM files in a directory
  - Creating a CSV metadata file describing the samples
  - Editing a small configuration file

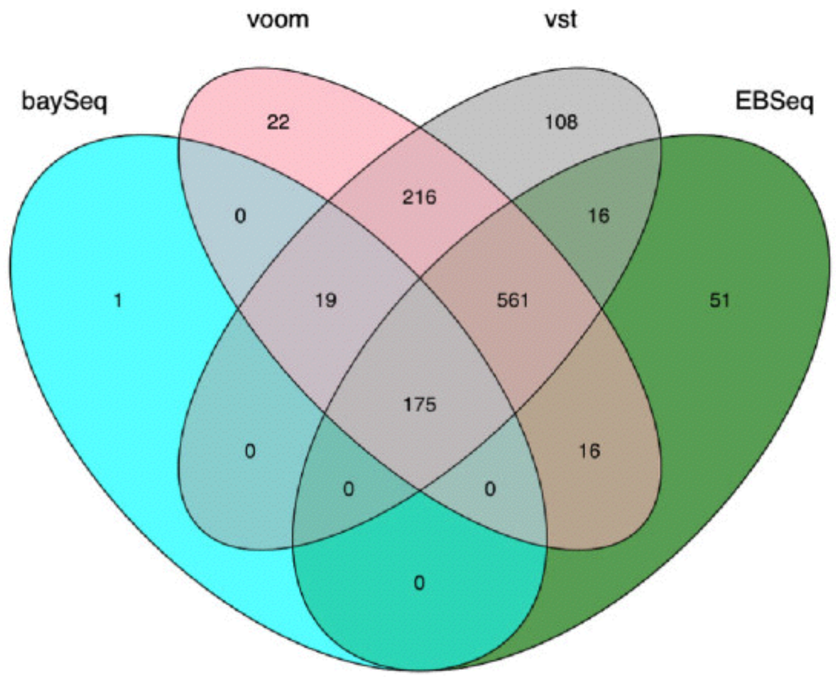
```
samplename,description,panel
SRR950078,UHRR_rep1,UHRR
SRR950079,HBRR_rep1,HBRR
SRR950080,UHRR_rep2,UHRR
SRR950081,HBRR_rep2,HBRR
SRR950082,UHRR_rep3,UHRR
SRR950083,HBRR_rep3,HBRR
SRR950084,UHRR_rep4,UHRR
SRR950085,HBRR_rep4,HBRR
SRR950086,UHRR_rep5,UHRR
SRR950087,HBRR_rep5,HBRR
```

```
details:
  - analysis: RNA-seq
    genome_build: GRCh37
    algorithm:
      aligner: star
      quality_format: Standard
      trim_reads: read_through
      adapters: [truseq, polya]
      strandedness: unstranded
```



RNA-seq pipeline overview





A comparison of methods for differential expression analysis of RNA-seq data
Charlotte Soneson[1*] and **Mauro Delorenzi**[12]

Varying DE calls between methods

Differential expression callers

edgeR

DESeq

DESeq2

BaySeq

voom + limma

Cuffdiff

NOISeq*

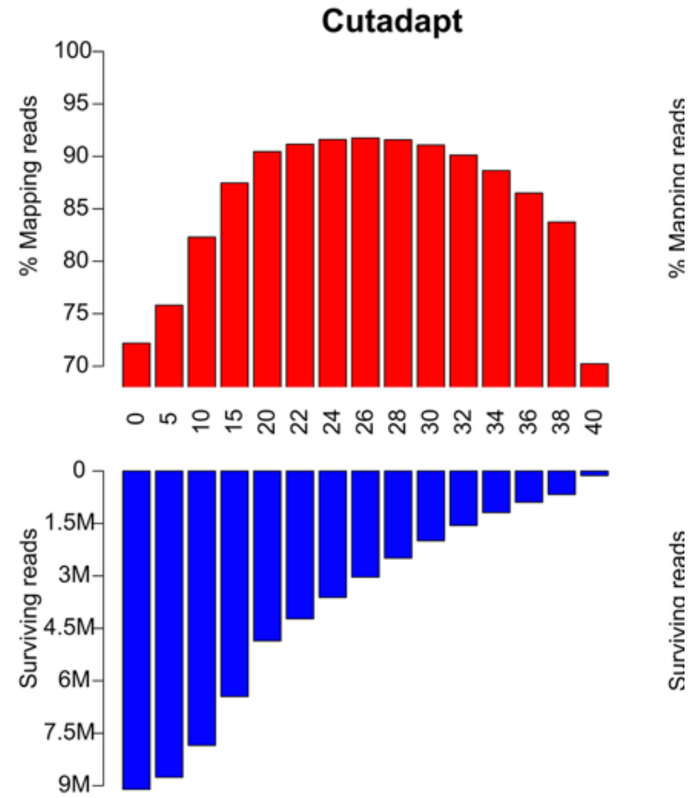
DERFinder*

```
# deseq analysis
# Soneson, C. & Delorenzi, M. A comparison of methods for differential expression
# analysis of RNA-seq data. BMC Bioinformatics 14, 91 (2013).

library(DESeq)
library(limma)
library(HTSFilter)
library(tools)
count_file = {{{count-file}}}
out_file = {{{out-file}}}
class = {{{class}}}
project = {{{project}}}
normalized_file = paste(strsplit(out_file, file_ext(out_file)[[1]][[1]]),
    "counts", sep="")
counts = read.table(count_file, header=TRUE, row.names="id")
DESeq.cds = newCountDataSet(countData = counts, conditions = class)
DESeq.cds = estimateSizeFactors(DESeq.cds)
DESeq.cds = estimateDispersions(DESeq.cds, method = "per-condition",
                                fitType = "local")
#DESeq.cds <- HTSFilter(DESeq.cds, s.len=25)$filteredData
res = nbinomTest(DESeq.cds, levels(class)[1], levels(class)[2])

comparison = paste(levels(class)[1], "_vs_", levels(class)[2], sep="")
out_table = data.frame(id=res$id, expr=res$baseMean, logFC=res$log2FoldChange,
         pval=res$pval, padj=res$padj, algorithm="deseq", project=project)
out_table$pval[is.na(out_table$pval)] = 1
out_table$padj[is.na(out_table$padj)] = 1
write.table(out_table, file=out_file, quote=FALSE, row.names=FALSE,
           sep="\t")
write.table(counts(DESeq.cds, normalized=TRUE), file=normalized_file,
           quote=FALSE, sep="\t")
```

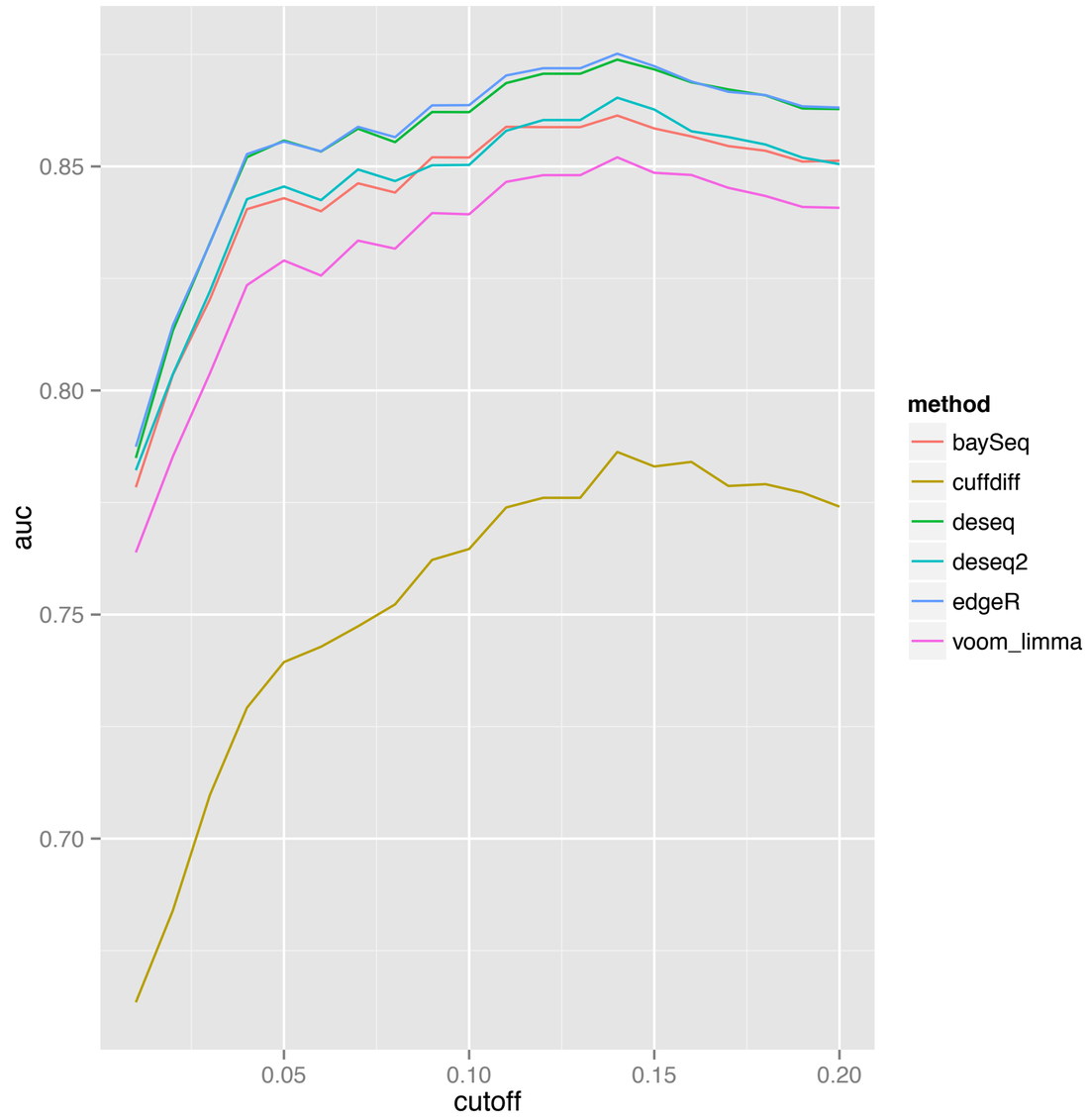
Is trimming beneficial in RNA-seq?

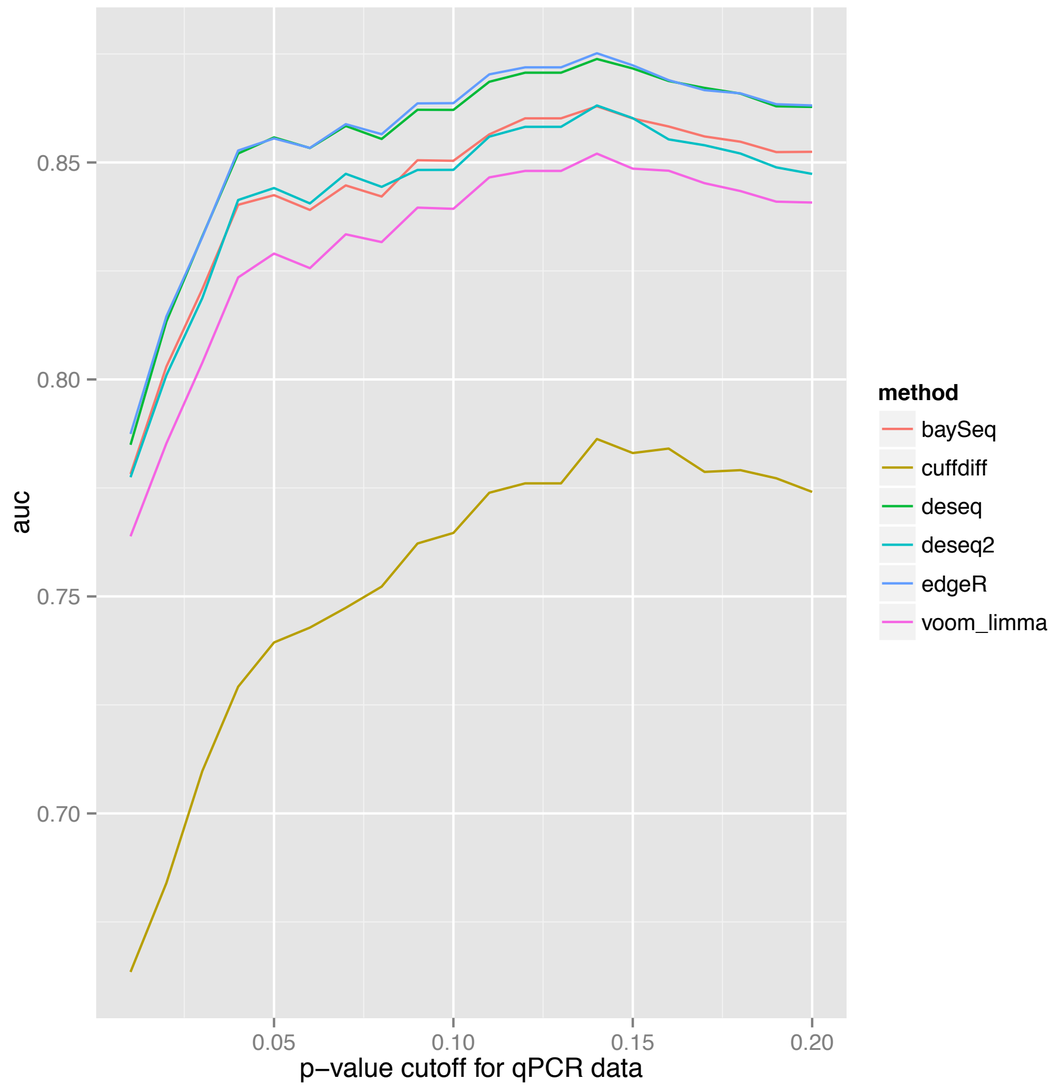


An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro, Simone Scalabrin, Michele Morgante, Federico M. Giorgi

Published: December 23, 2013 • DOI: 10.1371/journal.pone.0085024


auc
0.85
0.80
0.75
0.70
0.05
0.10
0.15
0.20
cutoff
method
baySeq
cuffdiff
deseq
deseq2
edgeR
voom_limma

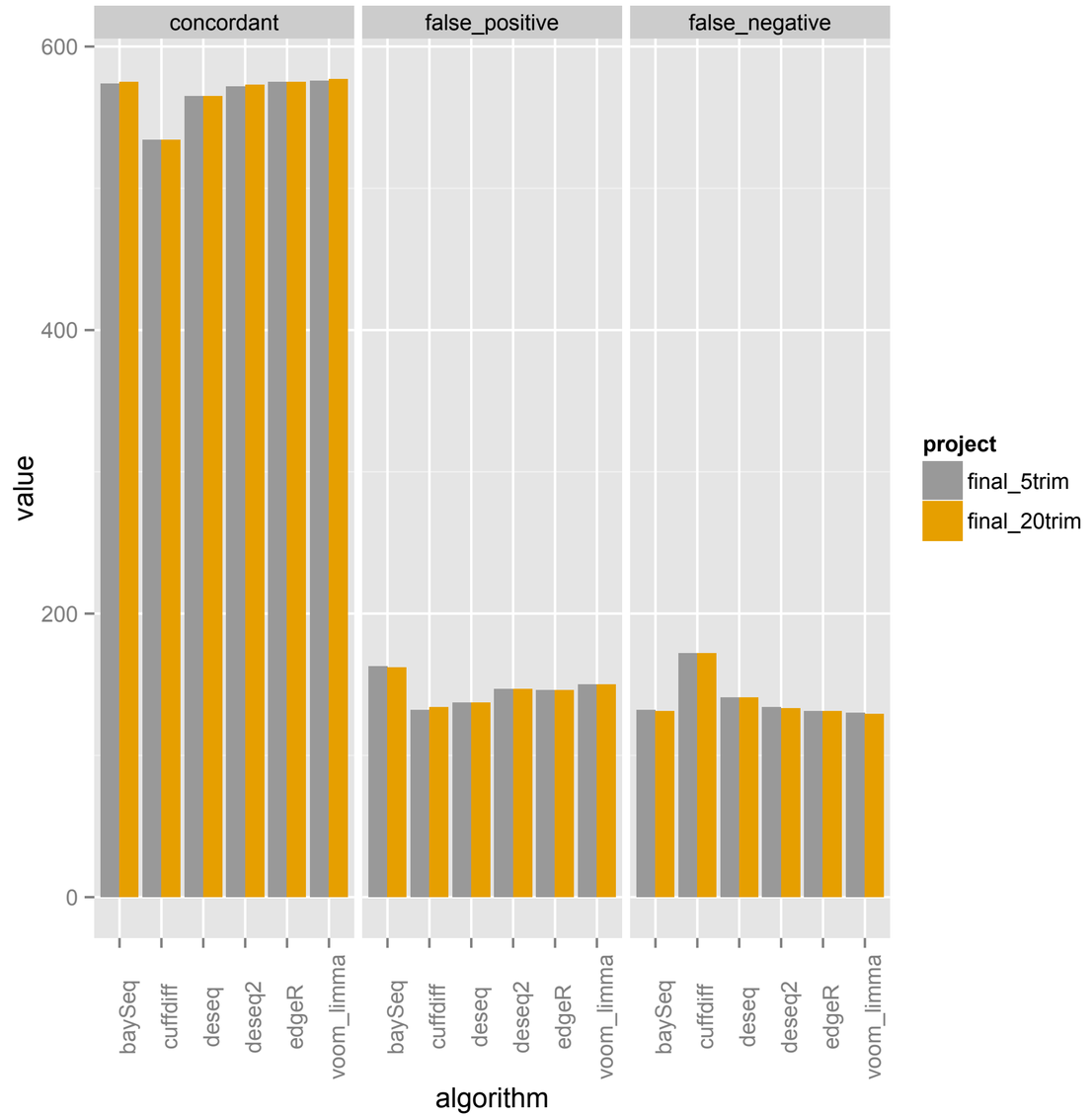

auc
0.85
0.80
0.75
0.70
0.05
0.10
0.15
0.20
p−value cutoff for qPCR data
method
baySeq
cuffdiff
deseq
deseq2
edgeR
voom_limma

Concordance

concordant/false positive/false negative

Jaccard index

Fold change



Simulation

- SEQC data set not a great set
- Count based simulation
  - More complicated models
  - Model biological variability
- Which algorithm is best?
- Plug in and go

Get, install, develop

Get

wget https://raw.github.com/chapmanb/bcbio-nextgen/master/scripts/bcbio_nextgen_install.py

Install

python bcbio_nextgen_install.py /usr/local/share/bcbio-nextgen —tooldir=/usr/local

Develop

https://github.com/chapmanb/bcbio-nextgen (Python)

https://github.com/roryk/bcbio.rnaseq (Clojure, R)

Free pizza, free beer and free compute

Arvados Demo and Hackathon

3-11-2014, 6pm-8pm

Curoverse Offices
51 Melcher St
Boston, MA 02210

Open source informatics platform running on top of Amazon, has nice solutions for provenance tracking and managing compute resources.

Free pizza, free beer and a free beta account.