Accessible and Scalable Bioinformatics for the Harvard Community

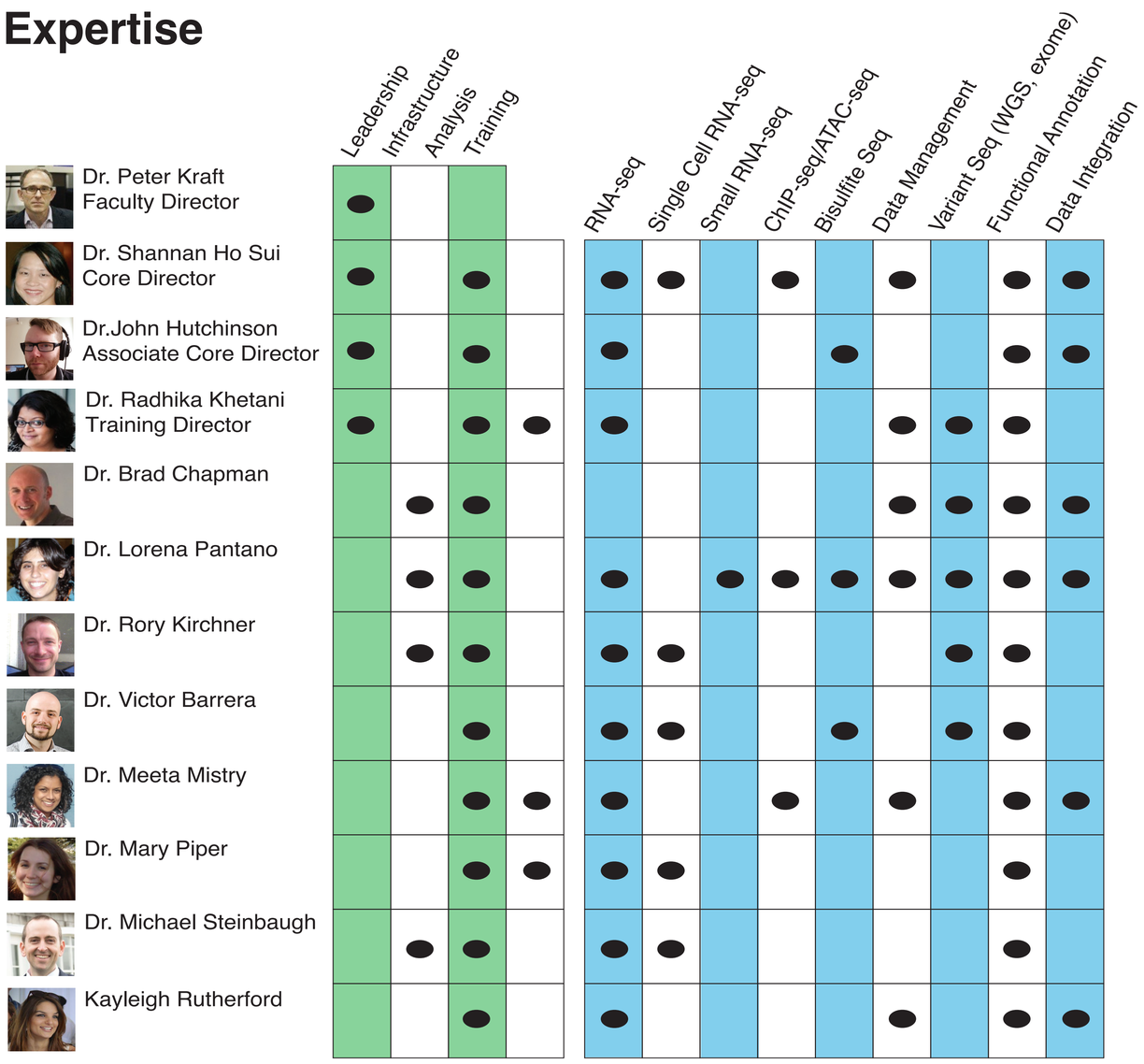
Victor Barrera, Harvard Chan Bioinformatics Core

Department of Biostatistics, Harvard Chan School of Public Health, Boston, MA, USA

Abstract

Objective:
Rapid technological development has enabled Harvard researchers to generate large data sets that require complex analyses . This presents a challenge: expertise may not be available within experimental labs, making it difficult to ensure accurate and reproducible results. The objective of the Harvard Chan Bioinformatics Core (HBC) is to provide best practice bioinformatics support to these researchers in a scalable and collaborative fashion.
Methods and conduct:
Employing a team of 11 bioinformaticians with diverse biological domain expertise and skill sets, the HBC tries to support common bioinformatics issues. To do so, the HBC initiated the development of an open source, community developed set of workflows called bcbio, whose development is steered by continual re-assessment of new methods and bioinformatics needs at Harvard. In this process, the HBC follows best practices and uses documented tools wherever possible.
Impact:
Through its bcbio infrastructure, the HBC can support the majority of next generation sequencing related research requests, with workflows in place for the analysis of bulk RNA-seq, single cell RNA-seq, small RNA-seq, variant sequencing, bisulfite sequencing, and ChIP-seq, as well as functional analysis by gene set enrichment. The HBC has supported hundreds of grants and analyses of all sizes, from small expression studies to studies involving thousands of whole genomes. Many of these consults have resulted in high profile publications.

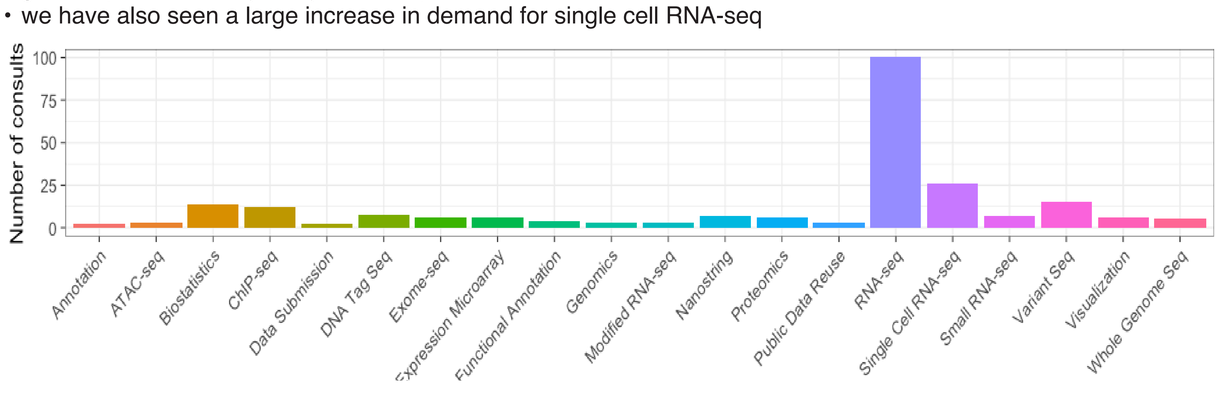
Expertise



Projects

- since the start of 2017, the HBC has worked with over 154 labs at Harvard on 215 projects

- gene expression analysis by RNA-seq is the most common bioinformatics focus since 2017
- we have also seen a large increase in demand for single cell RNA-seq



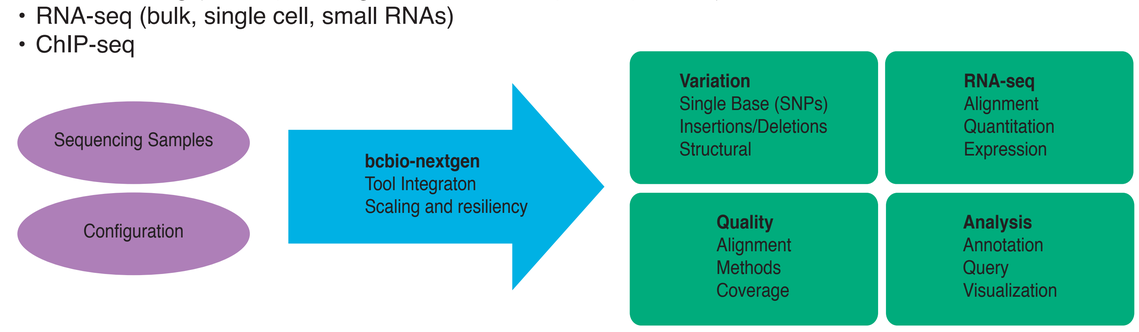
Scaleable Infrastructure

Pipelines

bcbio https://bcbio-nextgen.readthedocs.io

Approach
- Reproducible, Scaleable, Automated, Documented, Self-contained, Interoperable
- Open Source and Community Driven

Functions
- Variant Calling (exome, whole genome, structural, CNVs, cancer)
- RNA-seq (bulk, single cell, small RNAs)
- ChIP-seq



R Packages

DEGreport

platforms all | downloads top 20% | posts 0 | in Bioc 3 years | build ok

DOI: 10.18129/B9.bioc.DEGreport

Report of DEG analysis

Bioconductor version: Release (3.6)

Creation of a HTML report of differential expression analyses of count data. It integrates some of the code mentioned in DESeq2 and edgeR vignettes, and report a ranked list of genes according to the fold changes mean and variability for each selected gene.

Author: Lorena Pantano [aut, cre], John Hutchinson [ctb], Victor Barrera [ctb], Mary Piper [ctb], Kenneth Daily [ctb], Thanneer Malai Perumal [ctb], Rory Kirchner [ctb], Michael Steinbaugh [ctb]

Maintainer: Lorena Pantano <lorena.pantano at gmail.com>

Citation (from within R, enter `citation("DEGreport")`):

Pantano L (2017). *DEGreport: Report of DEG analysis*. R package version 1.14.0.

Check for updates

SOFTWARE TOOL ARTICLE

bcbioRNASeq: R package for bcbio RNA-seq analysis [version 1; referees: 2 approved with reservations]

Michael J. Steinbaugh[1*], Lorena Pantano[1*], Rory D. Kirchner[1], Victor Barrera[1], Brad A. Chapman[1], Mary E. Piper[1], Meeta Mistry[1], Radhika S. Khetani[1], Kayleigh D. Rutherford[1], Oliver Hofmann[2], John N. Hutchinson[1], Shannan Ho Sui[1]

[*] Equal contributors

METRICS
1085 VIEWS
250 DOWNLOADS

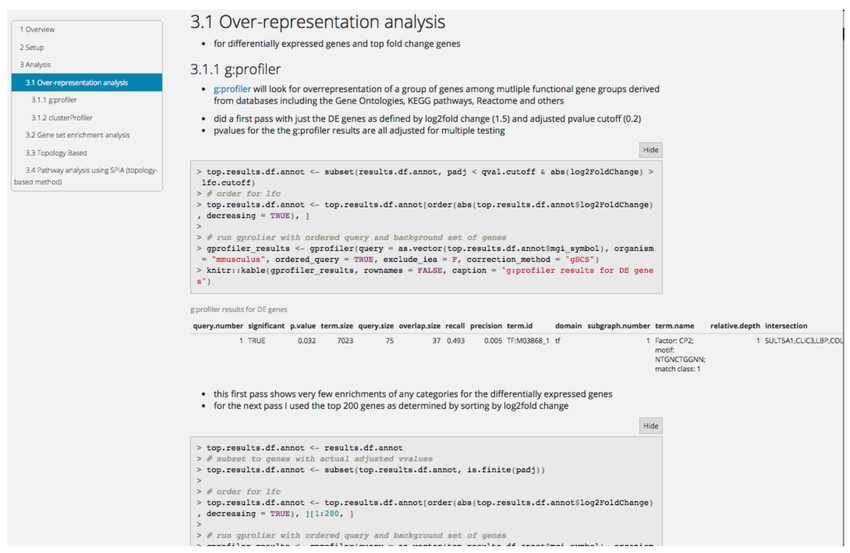
Collaborative Approach

- get involved early to help with experimental design
- schedule an initial meeting, free of charge
- create a timeline with deliverables
- provide a quote covering personnel, data storage and compute costs
- regularly document progress on a secure project site
- share all data sets, results and documentation
- provide methods, publication quality figures and help with GEO submissions

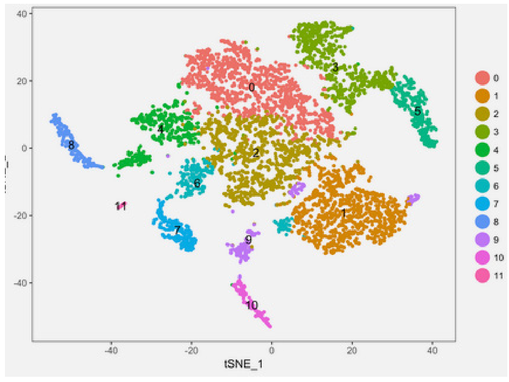
Programs

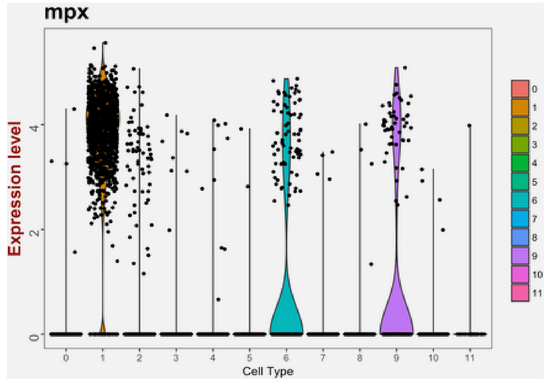
```
bamtools,2.4.0
bcbio-nextgen,0.9.8a0-8183767
bcbio-variation,0.2.6
bcftools,1.3
bedtools,2.24.0
biobambam,2.0.42
bioconductor-bubbletree,2.1.5
bowtie2,2.2.8
bwa,0.7.13
chanjo,
cnvkit,0.7.11
cufflinks,2.2.1
cutadapt,1.9.1
fastqc,0.11.5
featurecounts,1.4.4
freebayes,1.0.2
gatk,3.2-2-gec30cee
gatk-framework,3.5.21
gemini,0.18.3
grabix,0.1.6
hisat2,2.0.3beta
htseq,0.6.1p1
lumpy-sv,0.2.12
manta,0.29.6
metasv,0.4.0
mutect,1.1.5
novoalign,3.04.04
novosort,V3.00.02
oncofuse,1.1.0
phylowgs,20150714
picard,1.141
platypus-variant,0.8.1
qualimap,2.1.3
rna-star,2.4.1d
rtg-tools,3.6
sailfish,0.9.0
```

Rmarkdown report with code



Figures





Contact us:
We are located at the Harvard School of Public Health, SPH2, 2nd floor, Room 215.
Projects: bioinformatics@hsph.harvard.edu
Training: hbctraining@hsph.harvard.edu
Website : bioinformatics.sph.harvard.edu