







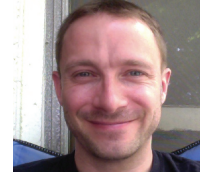




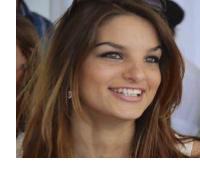
Abstract

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

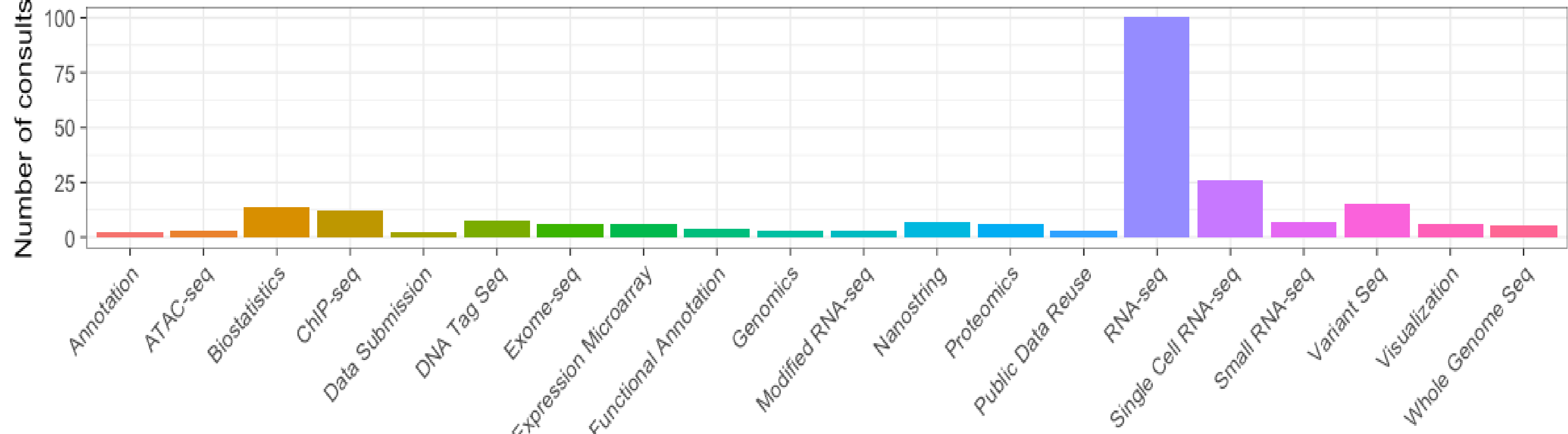
	Leadership	Infrastructure	Analysis	Training	RNA-seq	Single Cell RNA-seq	Small RNA-seq	ChIP-seq/ATAC-seq	Bisulfite Seq	Data Management	Variant Seq (WGS, exome)	Functional Annotation	Data Integration
 Dr. Peter Kraft Faculty Director	●		●		●	●	●	●	●	●	●	●	●
 Dr. Shannan Ho Sui Core Director	●		●		●								●
 Dr. John Hutchinson Associate Core Director	●		●		●			●				●	●
 Dr. Radhika Khetani Training Director	●		●	●	●					●	●		●
 Dr. Brad Chapman		●	●							●	●	●	●
 Dr. Lorena Pantano		●	●		●	●	●	●	●	●	●	●	●
 Dr. Rory Kirchner		●	●		●	●				●	●		
 Dr. Victor Barrera			●		●	●		●		●	●		
 Dr. Meeta Mistry			●	●	●		●		●	●	●	●	●
 Dr. Mary Piper			●	●	●	●					●		
 Dr. Michael Steinbaugh		●	●		●	●							
 Kayleigh Rutherford			●		●					●	●	●	●

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

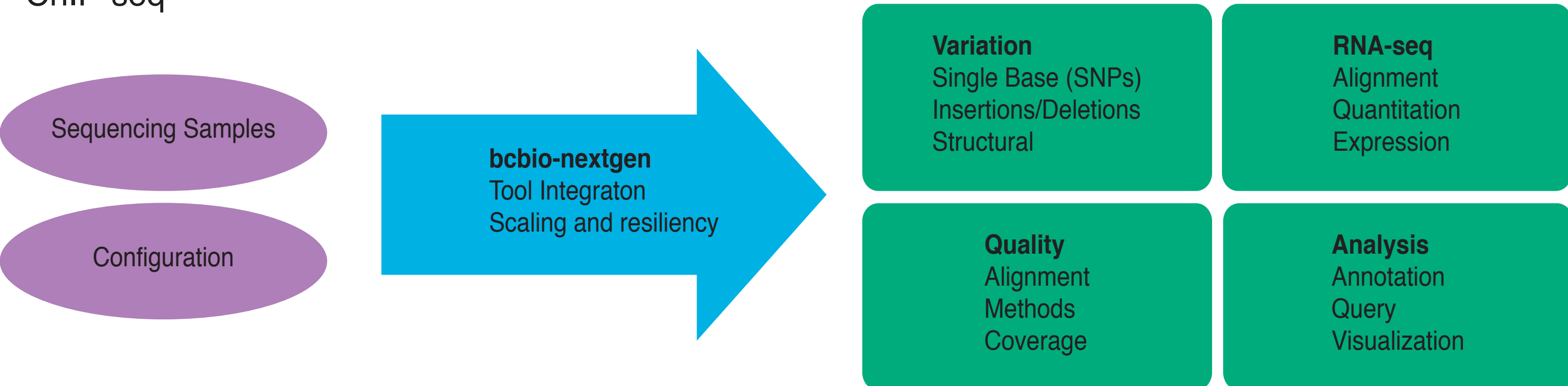
<https://bcbio-nextgen.readthedocs.io>

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: [10.18129/89.bioc.DEGreport](https://doi.org/10.18129/89.bioc.DEGreport) [f](#) [t](#) [b](#)

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], Thanaseer Melai Perumal [ctb], Rory Kirchner [ctb], Michael Steinbaugh [ctb]

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

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

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

SOFTWARE TOOL ARTICLE

Check for updates

Metrics

1085 VIEWS

250 DOWNLOADS

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

✉ Michael J. Steinbaugh [@mjsteinbaugh](#), ✉ Lorena Pantano^{1*}, Rory D. Kirchner¹, Victor Barrera¹, Brad A. Chapman¹, Mary E. Piper¹, Meeta Mistry¹, Radhika S. Khetani¹, Kayleigh D. Rutherford¹, Oliver Hofmann², John N. Hutchinson [@jnhutchinson](#), Shannan Ho Sui¹

* Equal contributors

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,0.
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-gc30c3ee
gatk-framework,3.5.21
gemin,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
novaalign,3.04.04
novosort,V3.00.02
oncofuse,1.1.0
phyloWS,20150714
picard,1.141
platypus-variant,0.8.1
qualimap,2.1.3
rna-star,2.4.1d
rtg-tools,3.6
salmon,0.9.0
```

Rmarkdown report with code

3.1 Over-representation analysis

- for differentially expressed genes and top fold change genes

3.1.1 gprofiler

- gprofiler will look for overrepresentation of a group of genes among multiple functional gene groups derived from databases including the Gene Ontologies, KEGG pathways, Reactome and others
- did a first pass with just the DE genes as defined by log2fold change (1.5) and adjusted pvalue cutoff (0.2)
- pvalues for the gprofiler results are all adjusted for multiple testing

```
> top.results.df.annot <- subset(results.df.annot, padj < qval.cutoff & abs(log2FoldChange) >
  life.cutoff)
> # order for life
> top.results.df.annot <- top.results.df.annot[order(abs(top.results.df.annot$log2FoldChange)
  , decreasing = TRUE), ]
>
> # run gprofiler with ordered query and background set of genes
> gprofiler.results <- gprofiler(query = as.vector(top.results.df.annot$logi_symbol), organism
  = "musmusculus", ordered_query = TRUE, exclude_ies = F, correction_method = "gsc")
> knitr::kable(gprofiler.results, rownames = FALSE, caption = "gprofiler results for DE gene
  s")
```

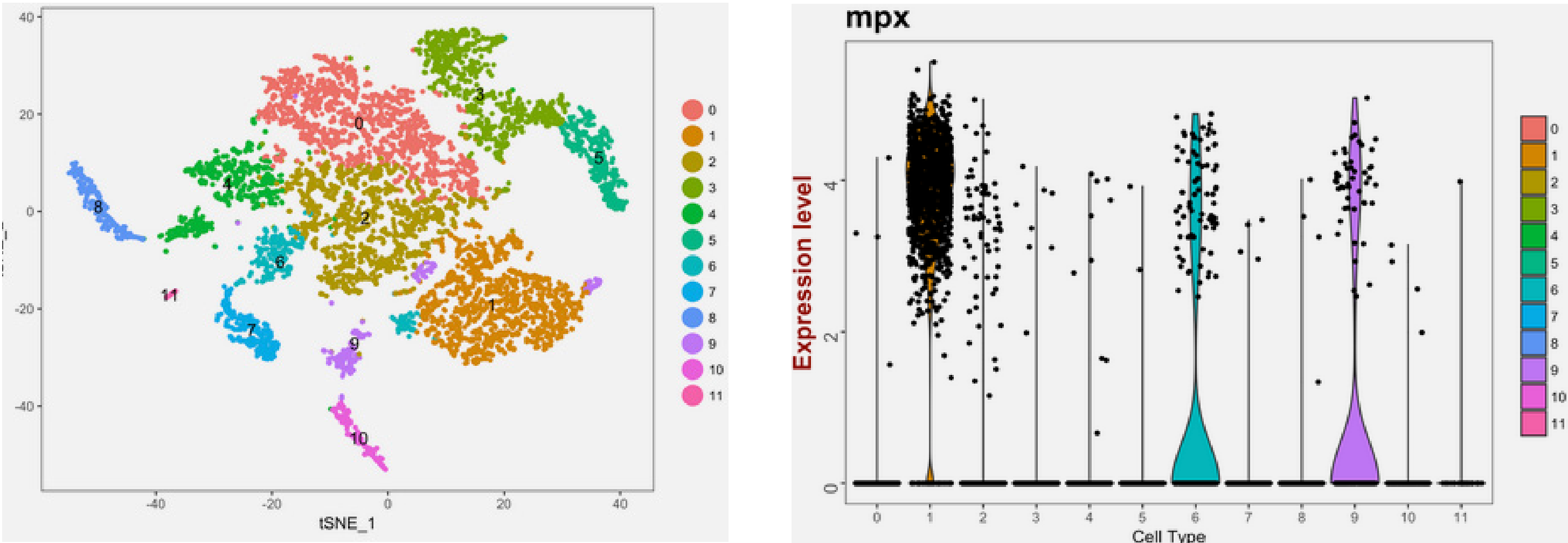
gprofiler results for DE genes

query.number	significant	pvalue	term.size	query.size	overlap.size	recall	precision	term.id	domain	subgraph.number	term.name	relative.depth	intersection
1	TRUE	0.032	7023	75	37	0.493	0.005	TFM0368.1	df	1	factor:CP2		1
											motif:WFO007G06	match class:1	

- this first pass shows very few enrichments of any categories for the differentially expressed genes
- for the next pass I used the top 200 genes as determined by sorting by log2fold change

```
> top.results.df.annot <- results.df.annot
> # subset to genes with actual adjusted values
> top.results.df.annot <- subset(top.results.df.annot, is.finite(padj))
> # order for life
> top.results.df.annot <- top.results.df.annot[order(abs(top.results.df.annot$log2FoldChange)
  , decreasing = TRUE), ][1:200, ]
>
> # run gprofiler with ordered query and background set of genes
> gprofiler.results <- gprofiler(query = as.vector(top.results.df.annot$logi_symbol), organism
  = "musmusculus", ordered_query = TRUE, exclude_ies = F, correction_method = "gsc")
> knitr::kable(gprofiler.results, rownames = FALSE, caption = "gprofiler results for DE gene
  s")
```

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

