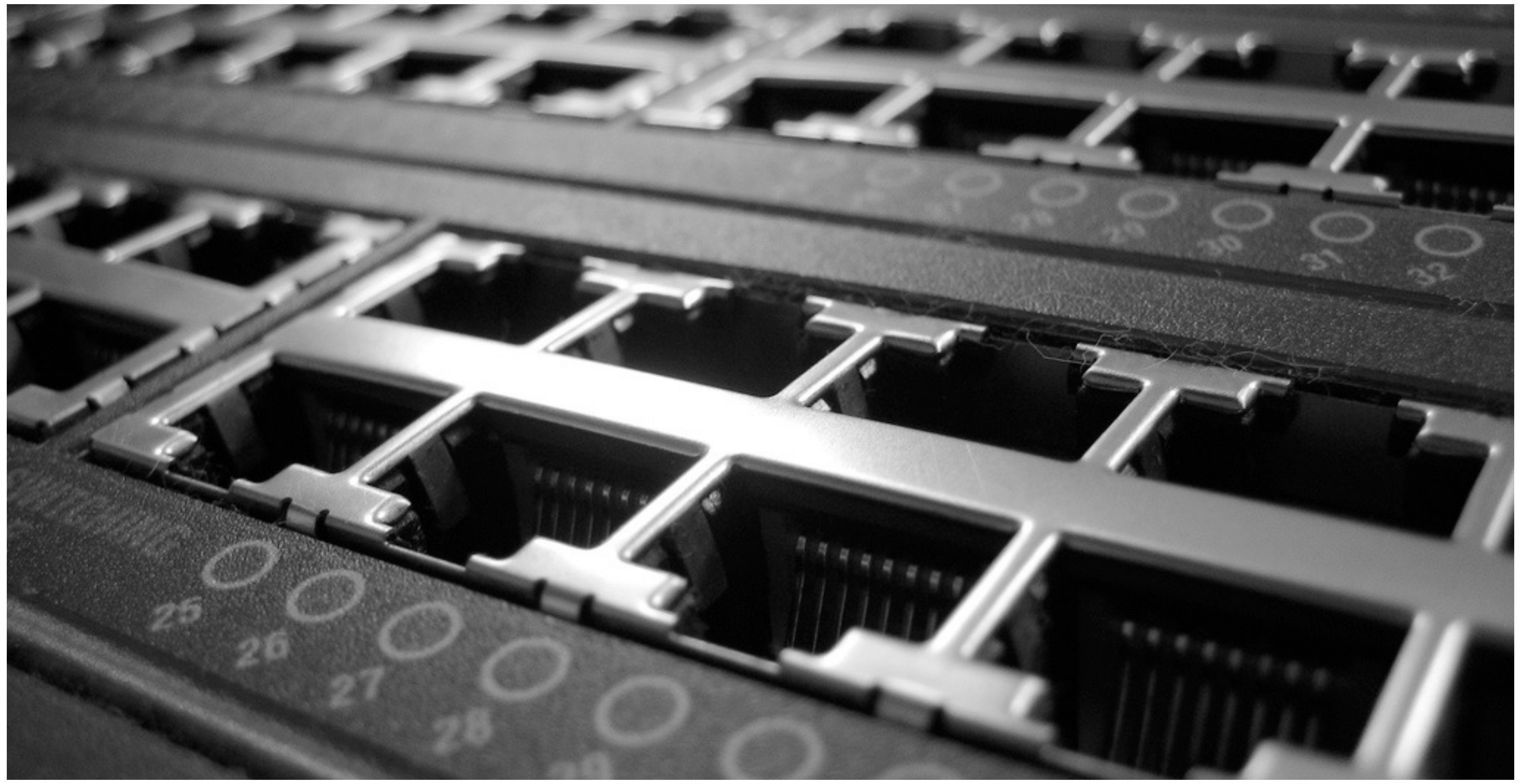
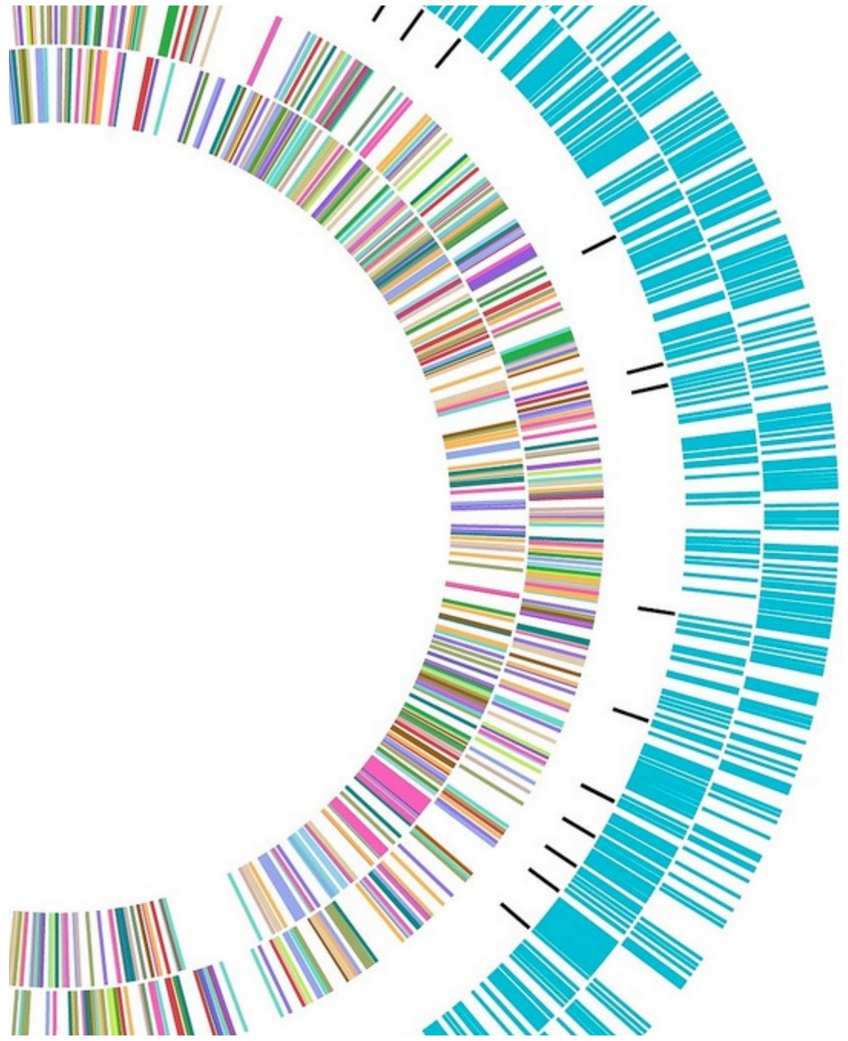




The Harvard Chan Bioinformatics Core

Shannan Ho Sui

February 14, 2020

Webpage: http://bioinformatics.sph.harvard.edu

Email: bioinformatics@hsph.harvard.edu

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Shannan Ho Sui
Director

John Hutchinson
Associate Director

Victor Barrera

Rory Kirchner

Zhu Zhuo

Peter Kraft
Faculty Advisor

Preetida Bhetariya

Meeta Mistry

Mary Piper

Jihe Liu

Radhika Khetani
Training Director

Ilya Sytchev

James Billingsley

Sergey Naumenko

Joon Yoon

Maria Simoneau

Consulting services

- Transcriptomics: RNA-seq, small RNA-seq, single cell RNA-seq
- DNA accessibility and binding: ChIP-seq, ATAC-seq
- Genetic variation: WGS, re-sequencing, exome-seq, structural variation, CNV
- Genome-wide methylation: 450k methylation arrays, RRBS, WGBS
- Data integration
- Functional enrichment analysis
- Experimental design and grant support

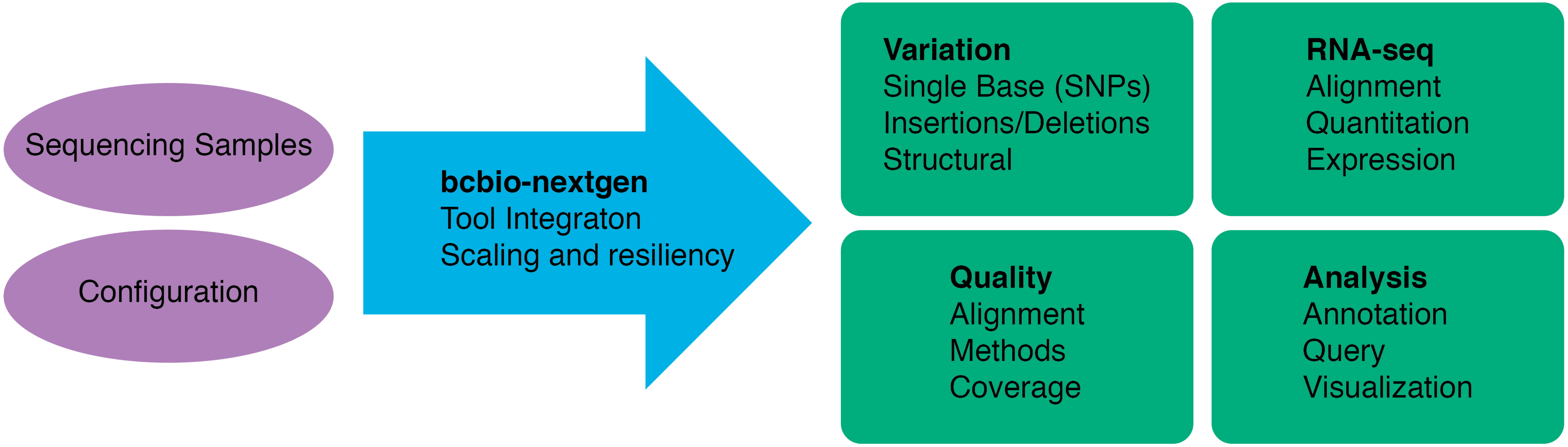
Training

- Basic Data Skills
- Advanced NGS Data Analysis
- In-depth courses (8- to 12-days)
- Free, Monthly, 2 - 3 hour long workshops on various bioinformatics topics
- Lessons: https://hbctraining.github.io/main

Sign up for our mailing list at - https://tinyurl.com/HBC-list-subscribe

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Standardizing using reproducible, scalable, validated best practice workflows



Sequencing Samples

Configuration

bcbio-nextgen
Tool Integraton
Scaling and resiliency

Variation
Single Base (SNPs)
Insertions/Deletions
Structural

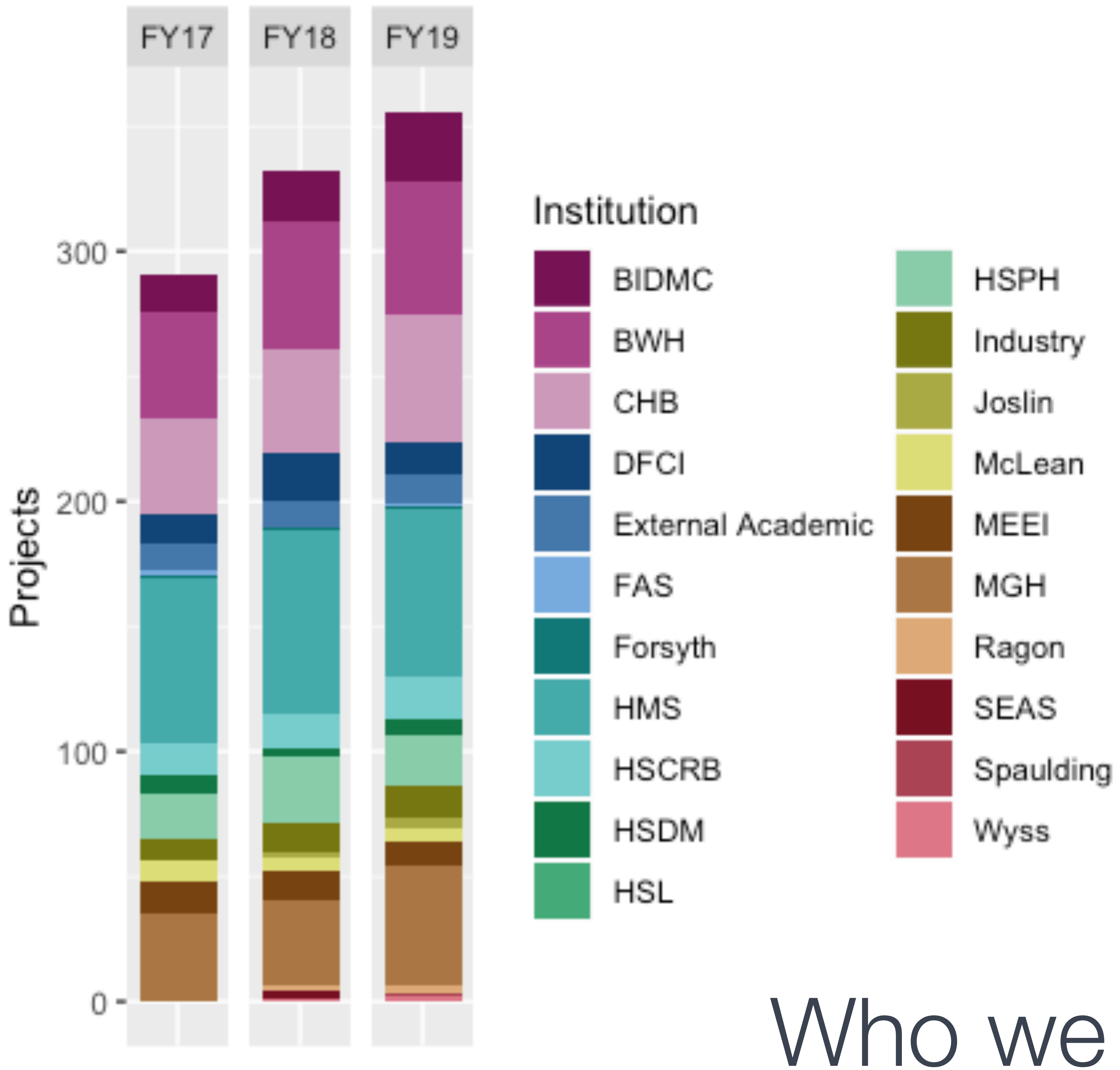
RNA-seq
Alignment
Quantitation
Expression

Quality
Alignment
Methods
Coverage

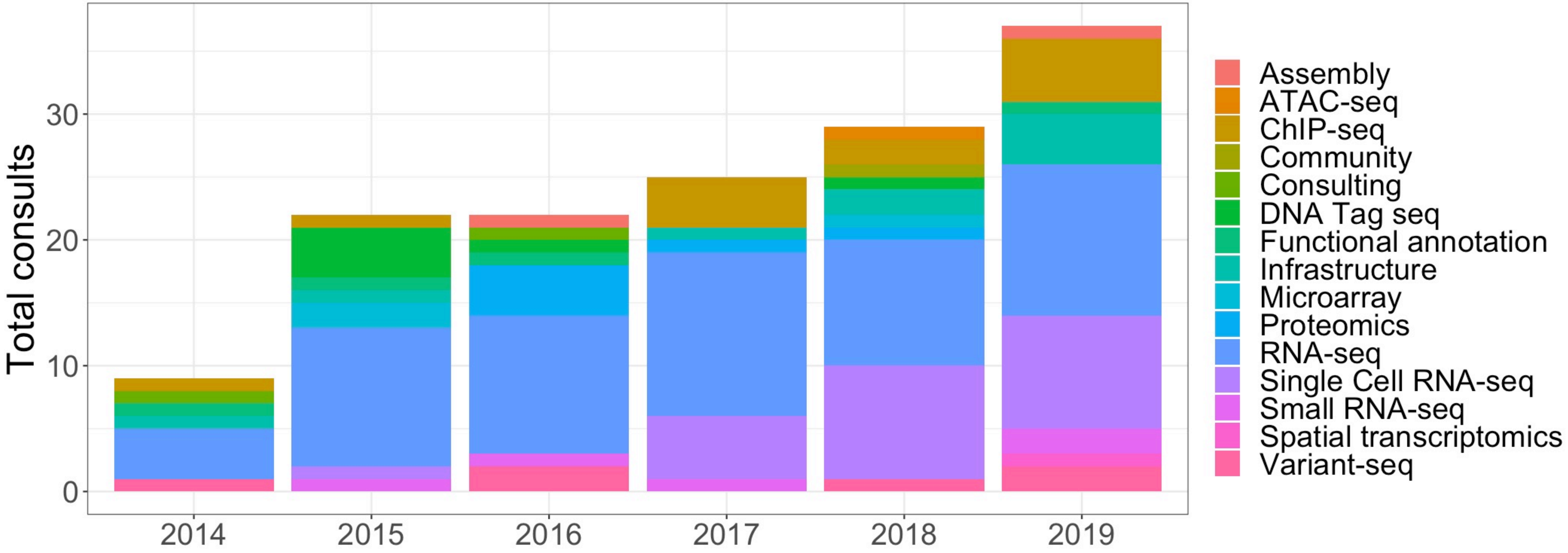
Analysis
Annotation
Query
Visualization

bcbio-nextgen

Python toolkit to automate best practice NGS pipelines



Who we support

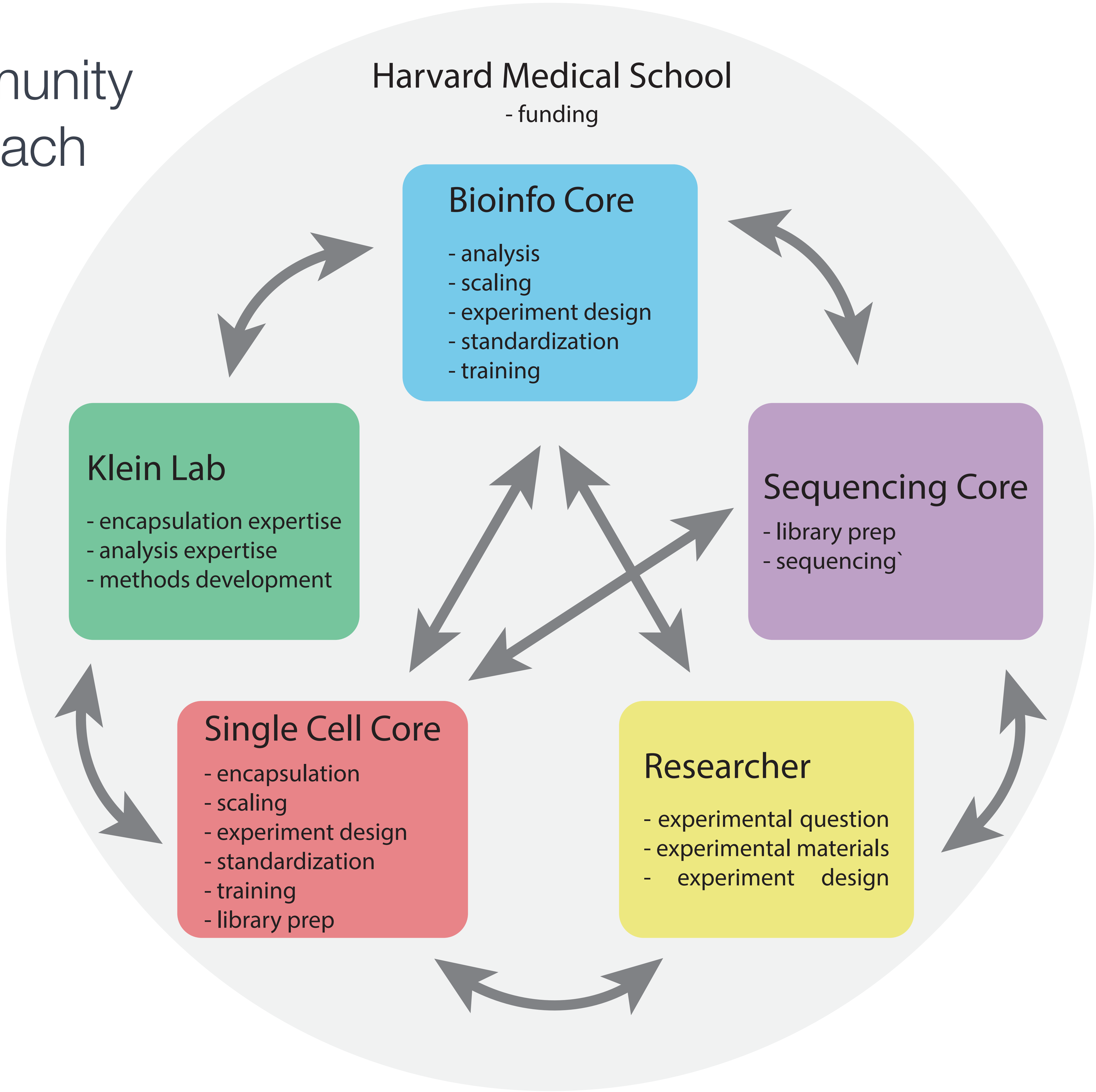


Growing demand for service and new techologies

Common applications of scRNA-seq

- Explore which cell types are present in a tissue
- Identify unknown/rare cell types or states
- Elucidate the changes in gene expression during differentiation or across time or states
- Identify genes that are differentially expressed in particular cell types between conditions (e.g. treatment or disease)
- Explore changes in expression among a cell type while incorporating spatial, regulatory, and/or protein information

A Community Approach



Harvard Medical School

- funding

Bioinfo Core

- analysis
- scaling
- experiment design
- standardization
- training

Sequencing Core

- library prep
- sequencing`

Researcher

- experimental question
- experimental materials
- experiment design

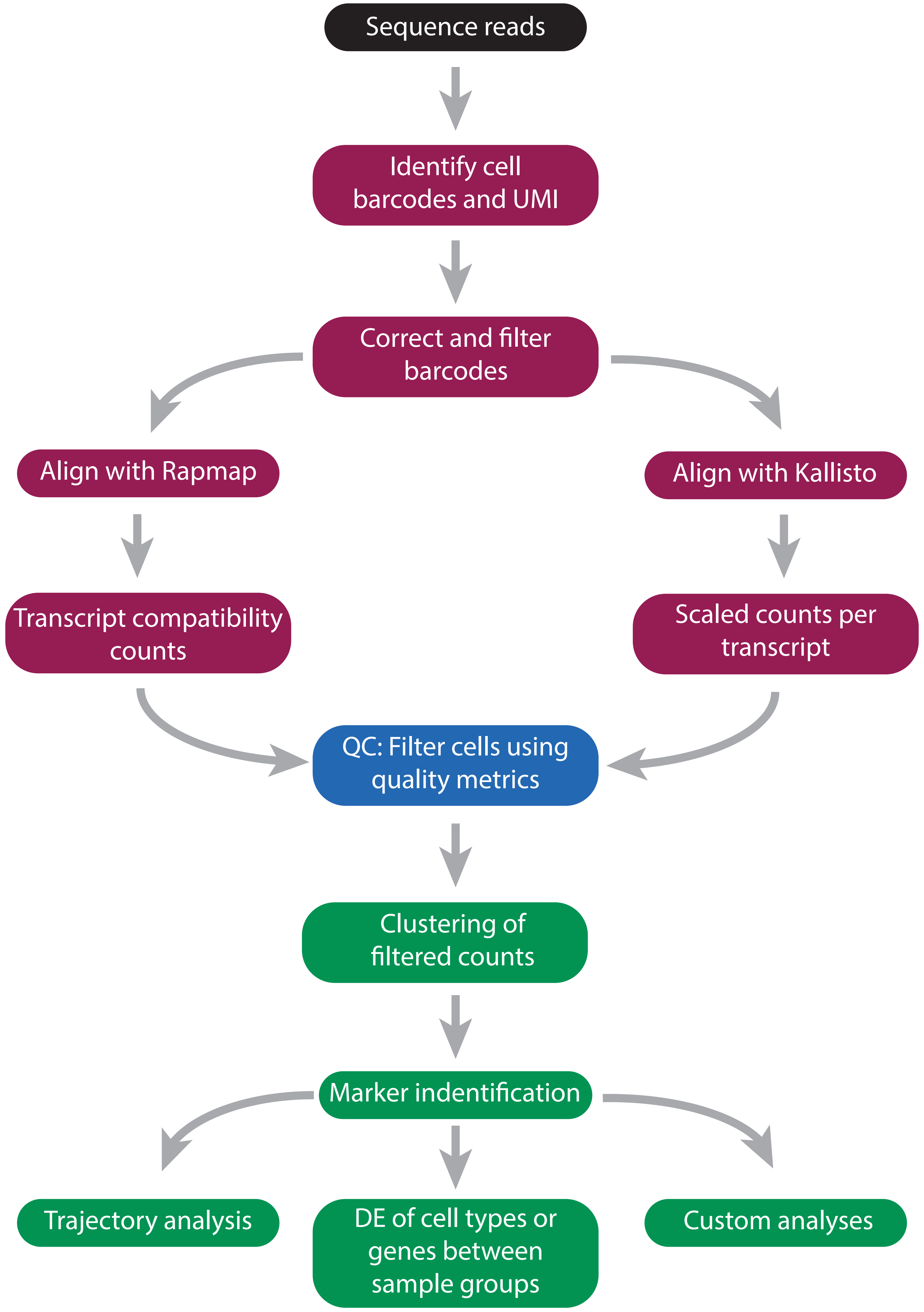
Single Cell Core

- encapsulation
- scaling
- experiment design
- standardization
- training
- library prep

Klein Lab

- encapsulation expertise
- analysis expertise
- methods development

Our single cell workflow



Sequence reads

Identify cell barcodes and UMI

Correct and filter barcodes

Align with Rapmap

Align with Kallisto

Transcript compatibility counts

Scaled counts per transcript

QC: Filter cells using quality metrics

Clustering of filtered counts

Marker indentification

Trajectory analysis

DE of cell types or genes between sample groups

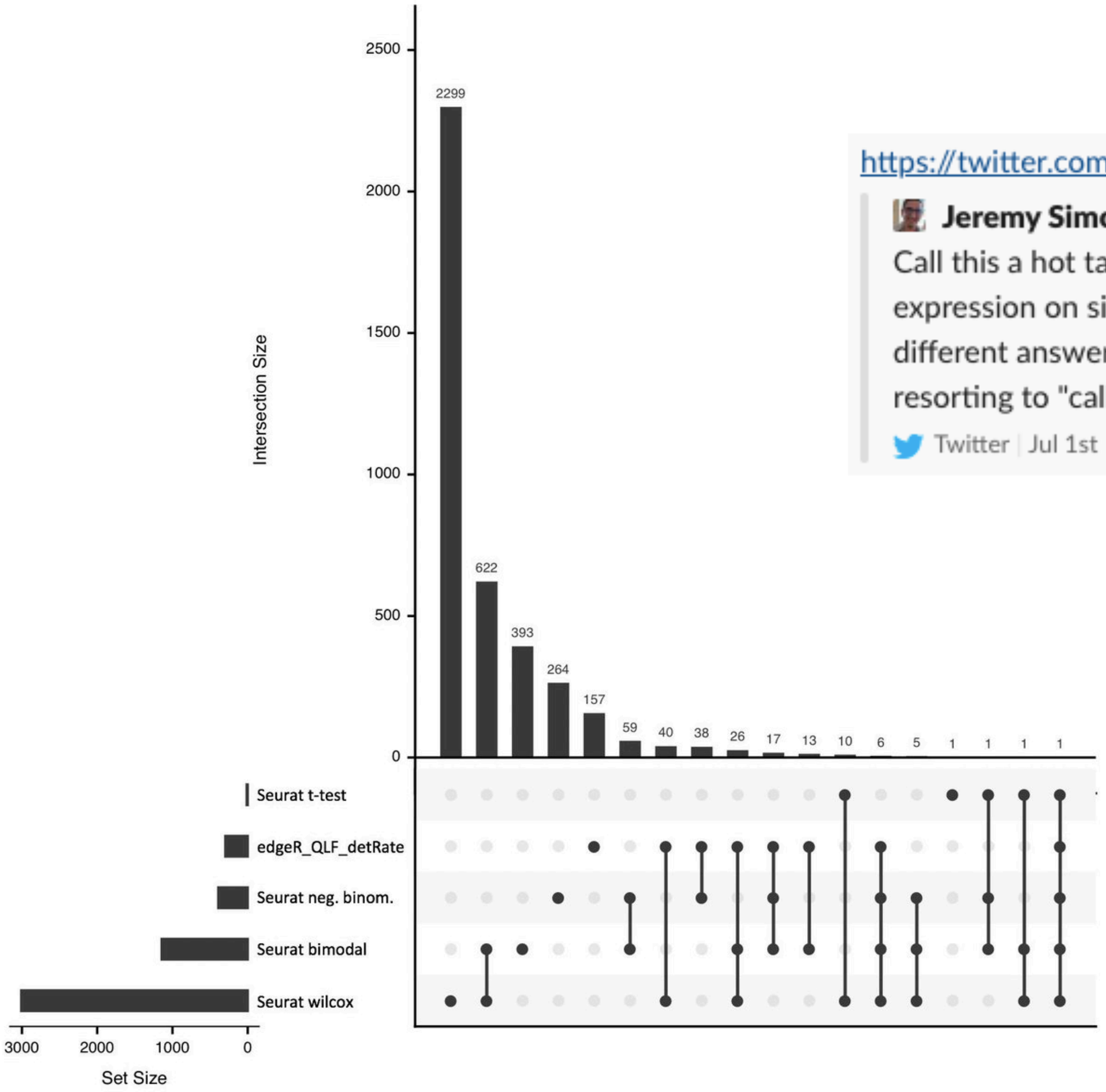
Custom analyses

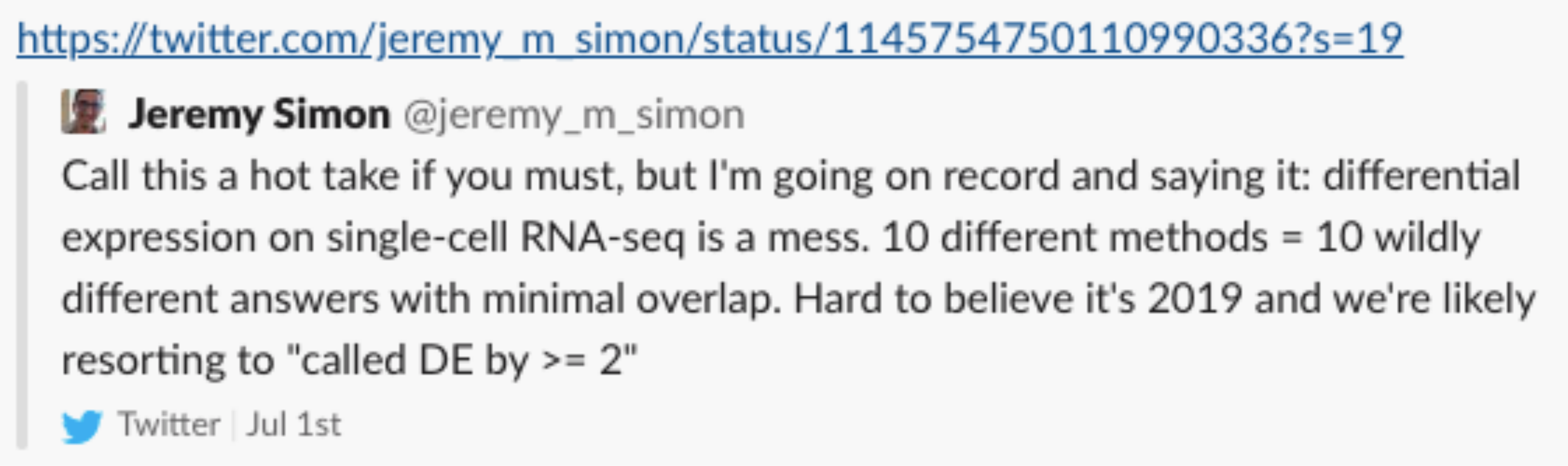
Quality control

- Focus on getting high quality data, messy data is EXTREMELY hard to work with
- Cells must be alive, before running the experiment shoot for viability > 95%
- When filtering, it is better to start out too strict than too lax
- Essential to work closely with the biologists, many, many judgement calls need to be made based on expert knowledge
- Before beginning analysis, have a good list of marker genes in hand for each cell type that could be present in the sample
- Often markers used for FACS sorting are not good for single-cell RNA-seq

Challenges and Opportunities

- Complex designs - replicates, batches, technologies
- Close collaborations to allow for rapid, iterative analyses
- Rapidly emerging methods and evolving tools
- Which ones to use?
- Keeping versions consistent/synchronized (esp. among computing environments)
- Different results from different methods
- Lots of open questions





https://twitter.com/jeremy_m_simon/status/1145754750110990336?s=19

Jeremy Simon @jeremy_m_simon

Call this a hot take if you must, but I'm going on record and saying it: differential expression on single-cell RNA-seq is a mess. 10 different methods = 10 wildly different answers with minimal overlap. Hard to believe it's 2019 and we're likely resorting to "called DE by >= 2"

Twitter | Jul 1st

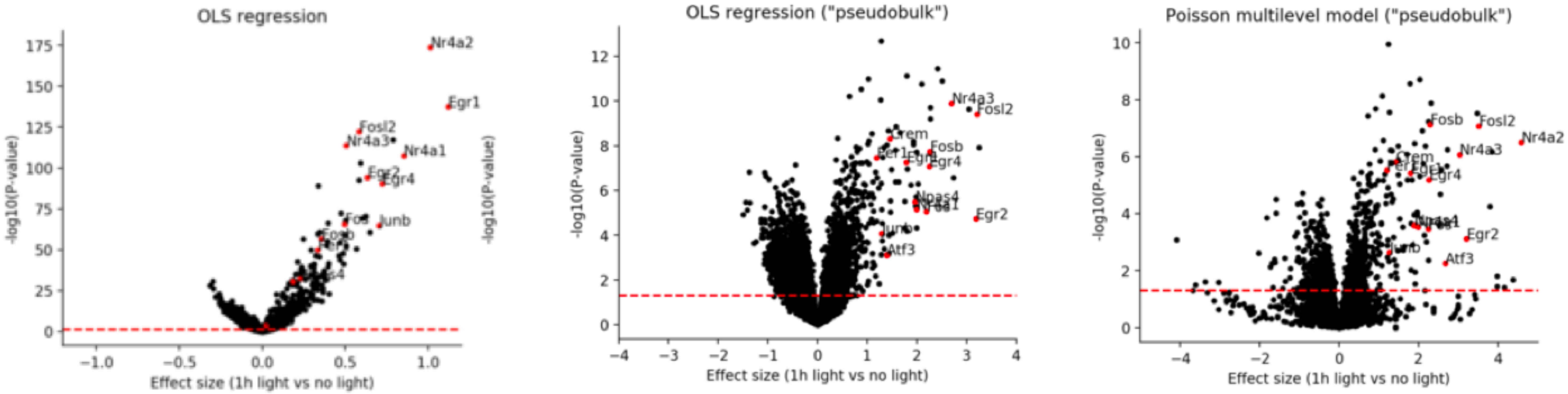
What is N in single cell experiments?

- Three treated patients, three control patients.
- Extract PBMC and want to look at the effect of the treatment on B cells.
- Identified B-cell clusters in the treated and non-treated patients via marker genes and found 300 B cells in each patients for a total of 900 B cells in each treatment condition
- If I want to ask what the effect of the treatment is on B-cells, what is my N here? Is it 900 for each condition? No. But almost all single-cell papers to date (including ours!) treat it as if it is.
- N should be 3, not 900.

How to get to N=3?

- Pseudobulk: sum all B-cells for each sample, and treat it like an *in-silico* FACS sorted experiment
- multilevel model: model the patient level data in a multilevel model so you can account for the non-independence of measurements from B-cells of the same patient

pseudobulk is simple and works well



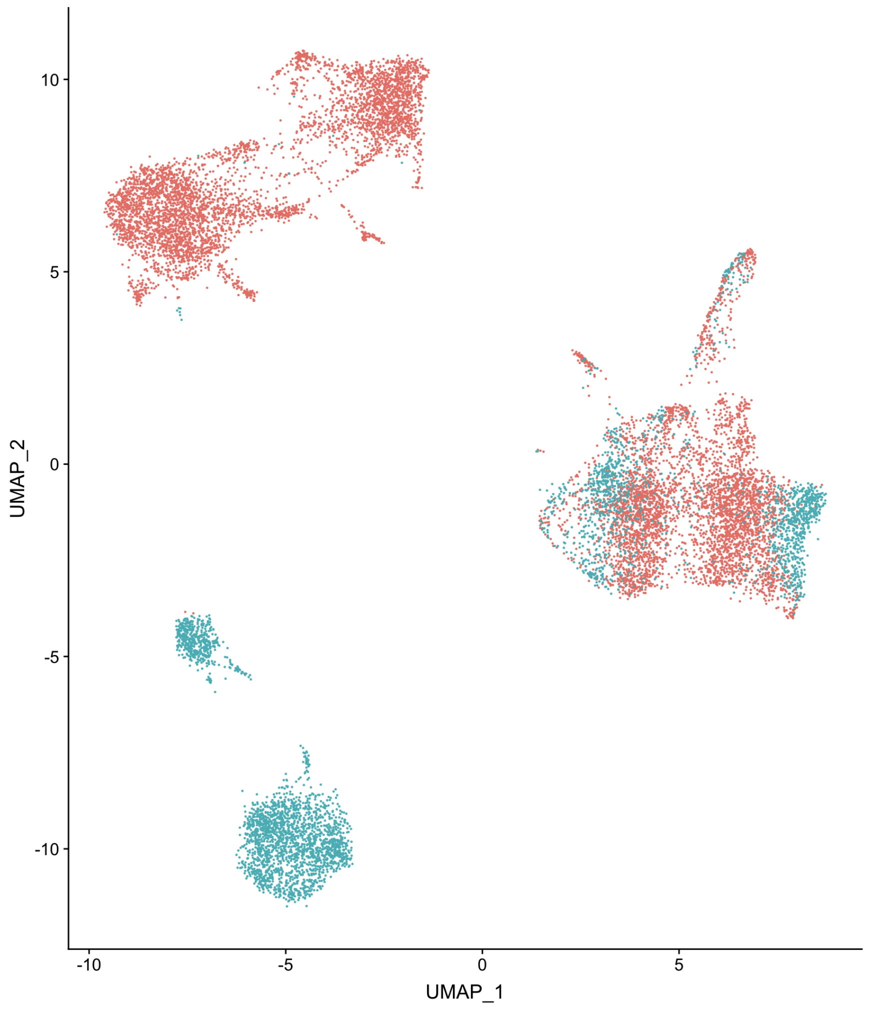
http://www.nxn.se/valent/2019/2/15/handling-confounded-samples-for-differential-expression-in-scrna-seq-experiments (Valentine Svensson, via twitter discussion with Rory Kirchner)

Challenges and Opportunities

- Projects take longer to complete
- Practical approach to training
- Internal training through retreats, development of materials, group discussions
- Community training through our Bioinformatics Training Program

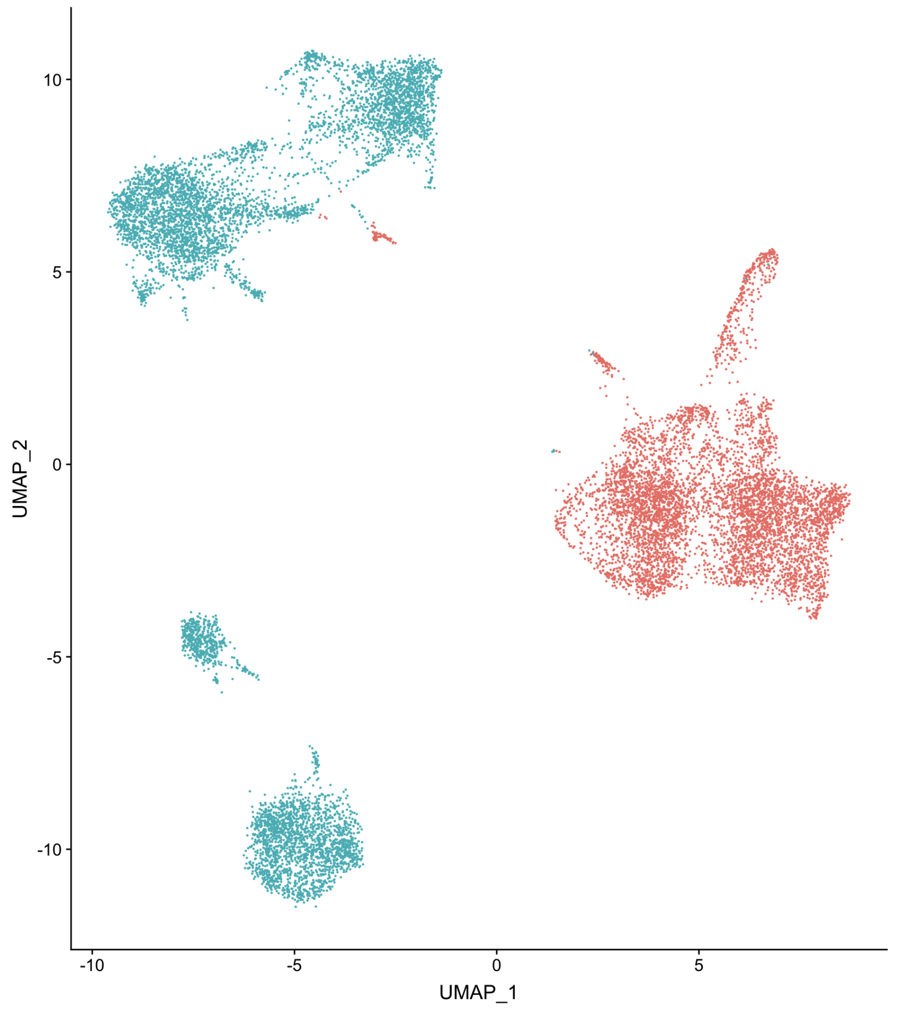
Seurat v.3.0

State



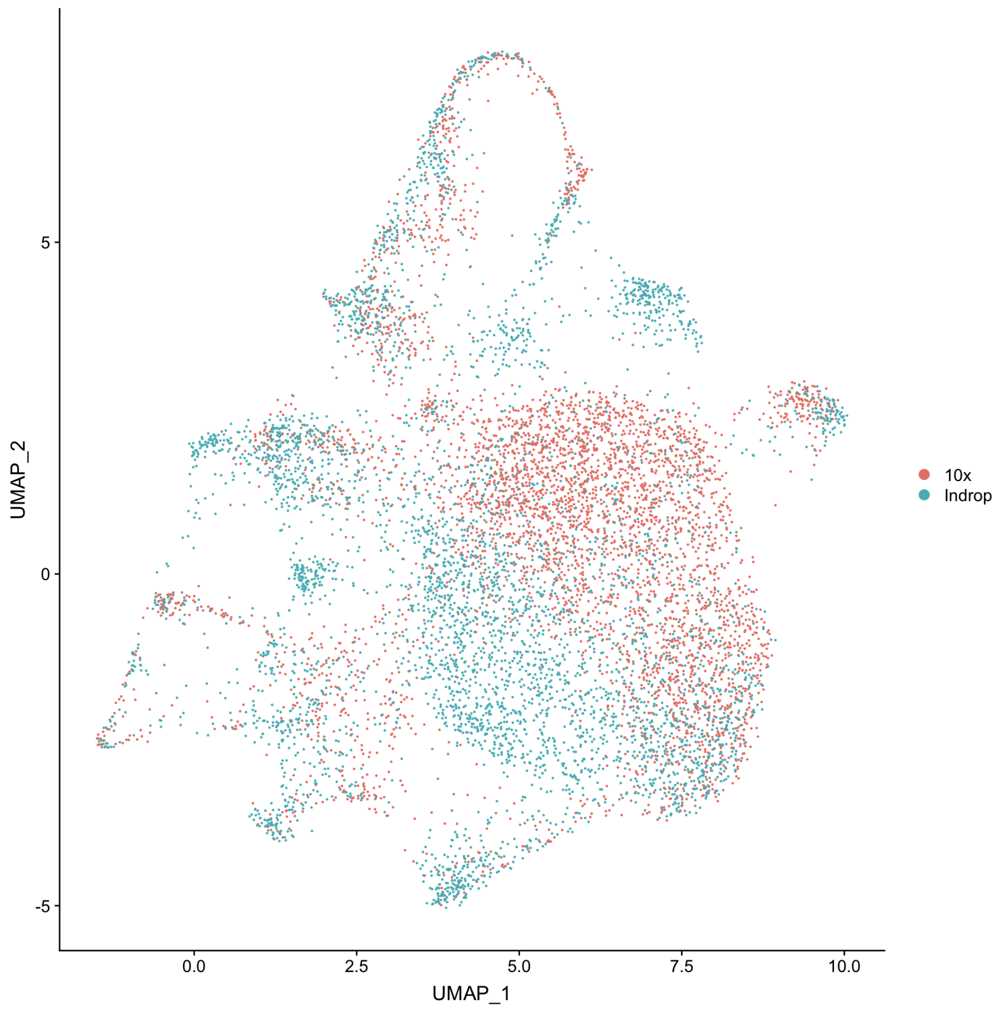
🟥 injured
🟦 uninjured

Technology



🟥 10X
🟦 InDrop

Multi-CCA

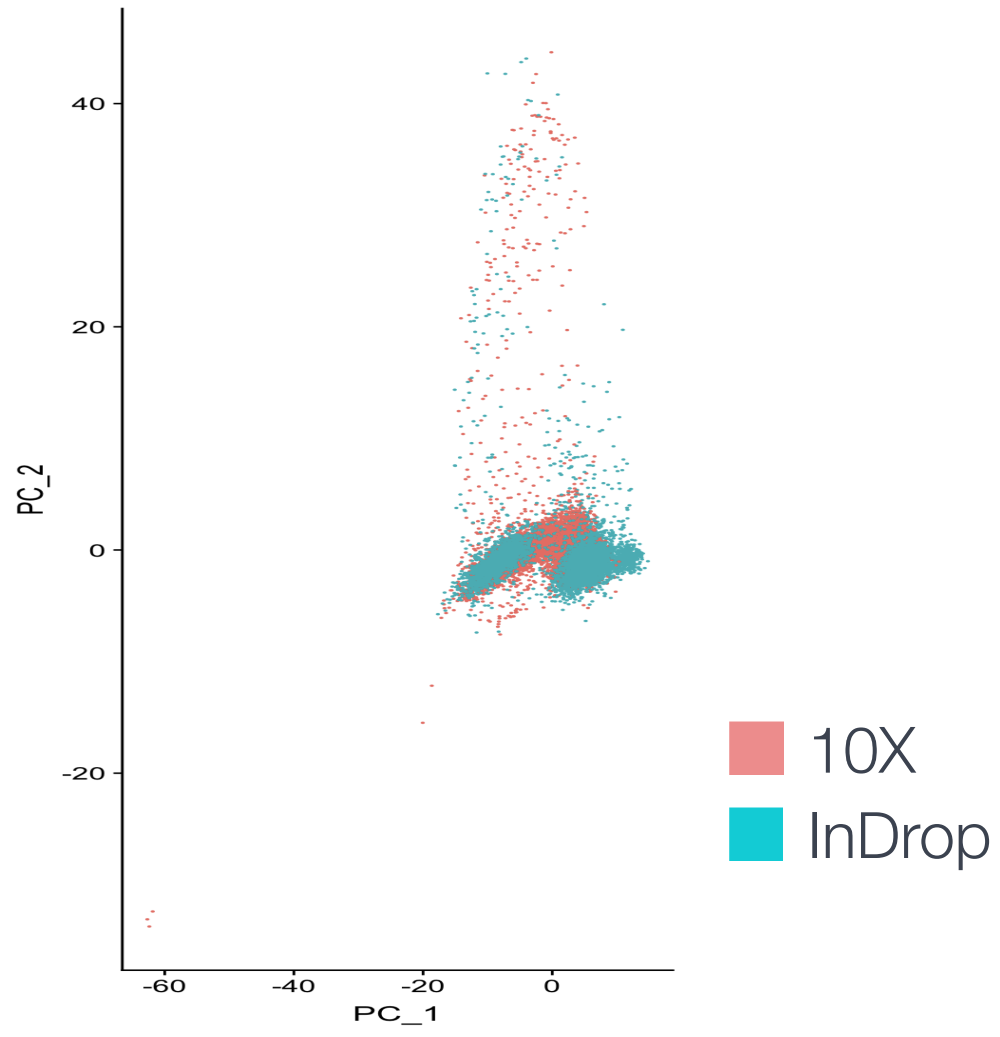


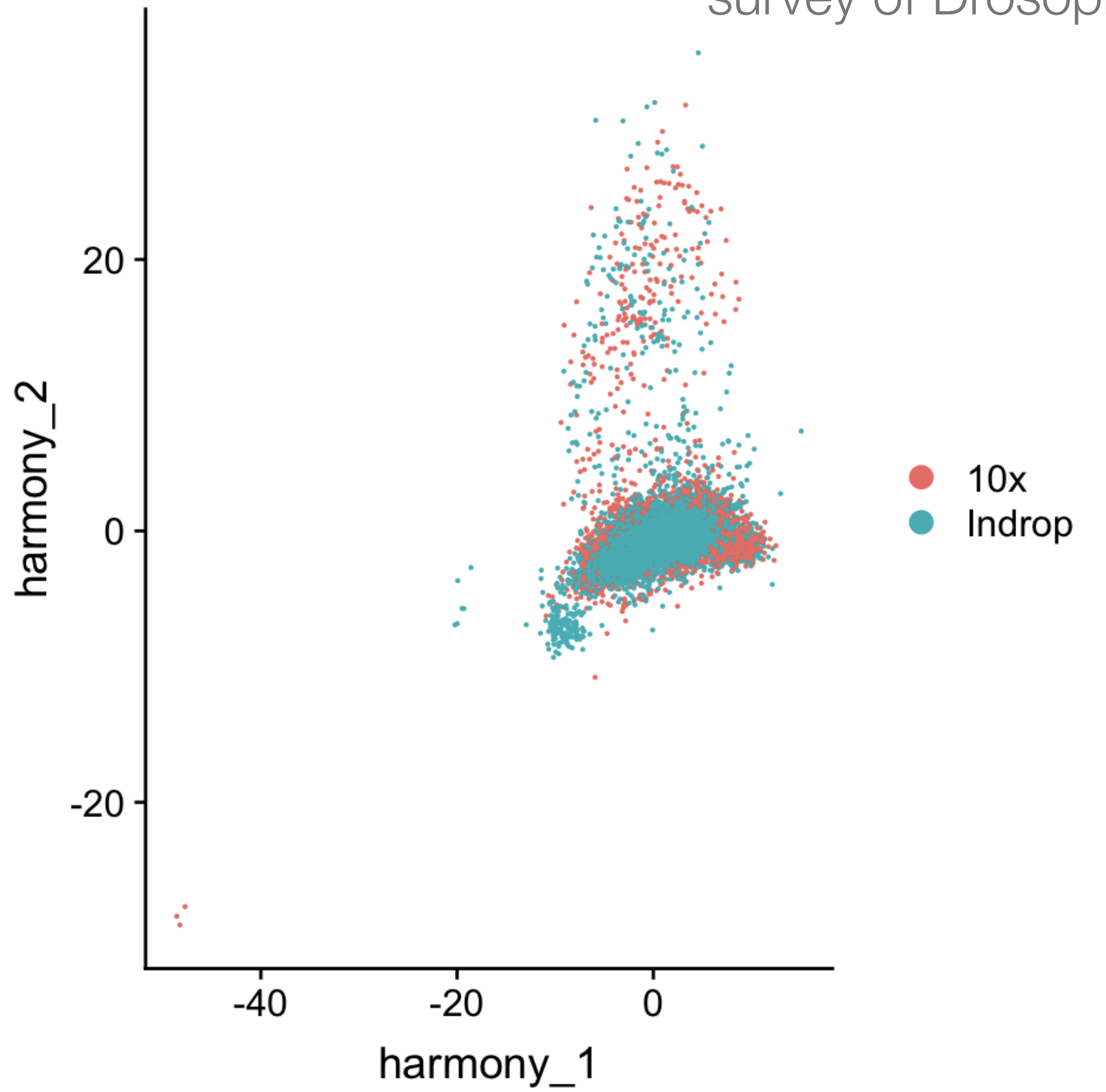
🟥 10X
🟦 InDrop

Harmony

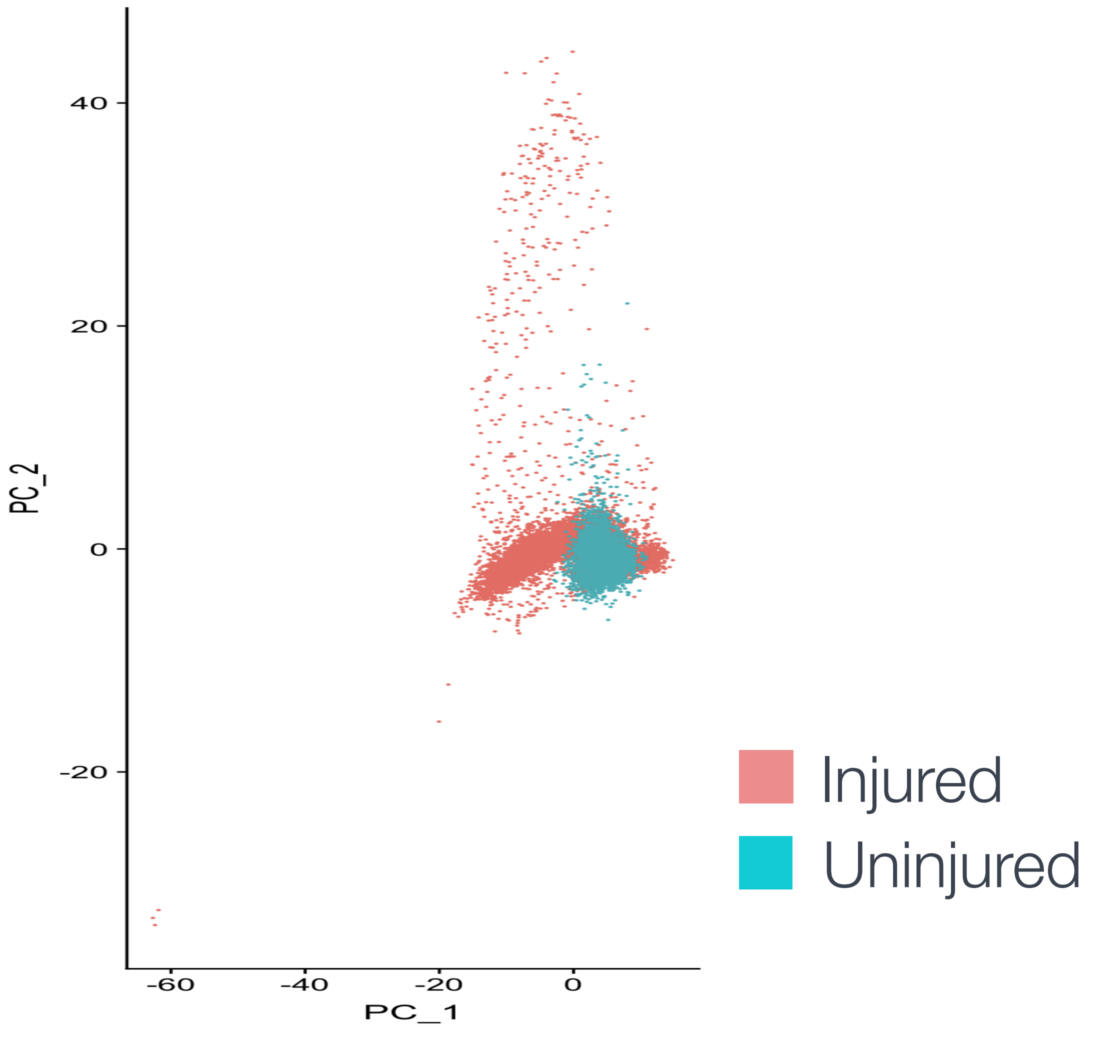
Tattikota et al., 2019, A single cell survey of Drosophila blood, BioRxiv.

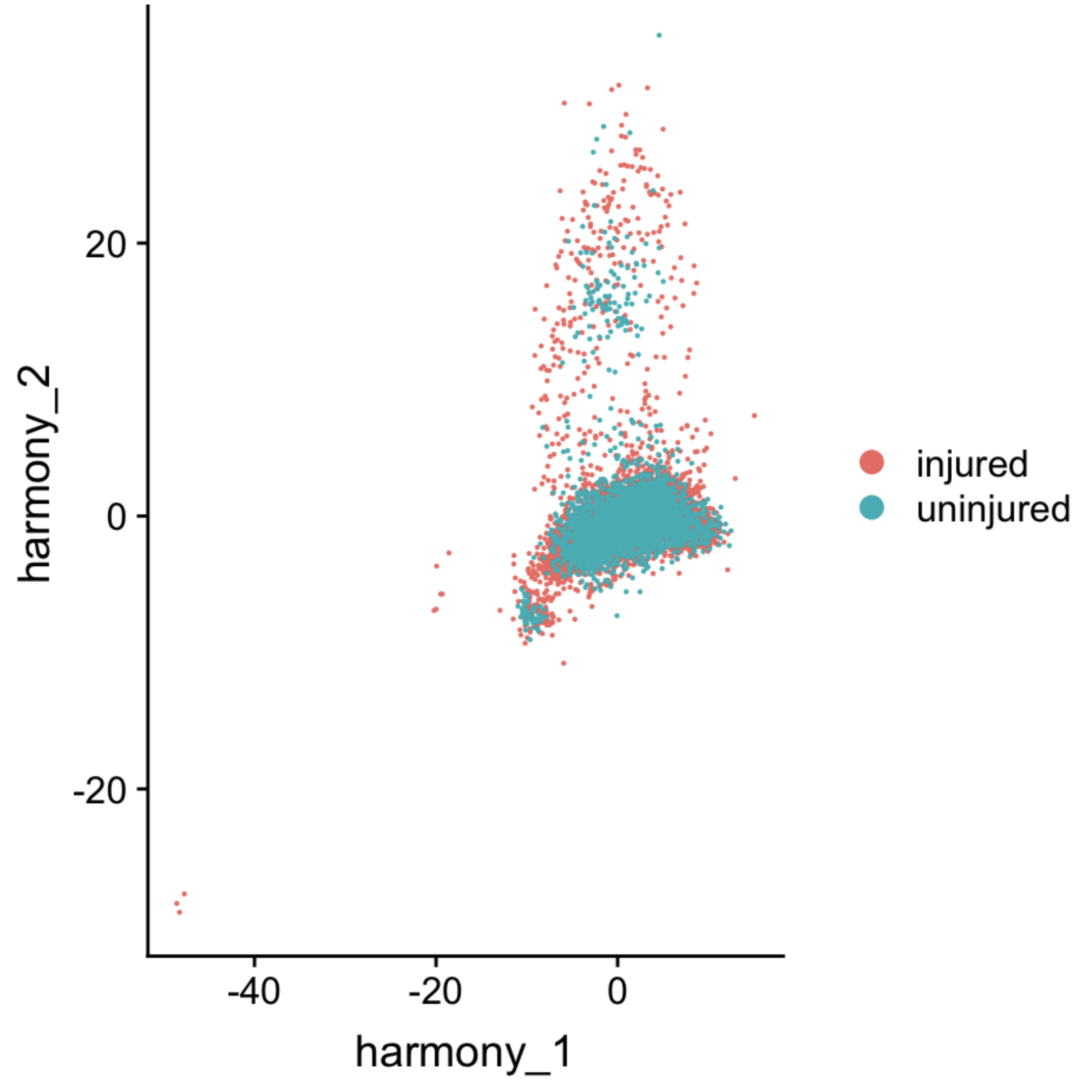
Technology



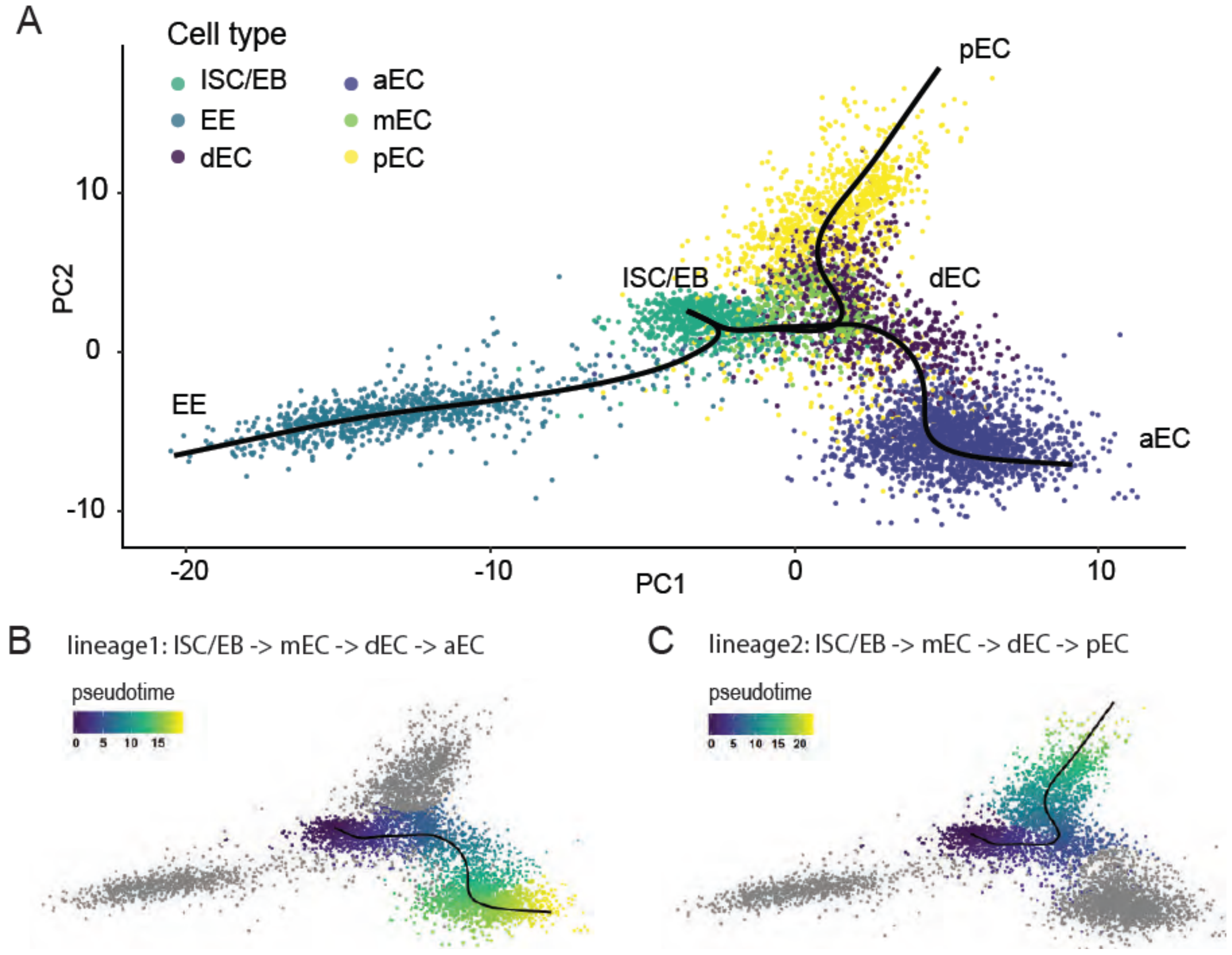


Injury state





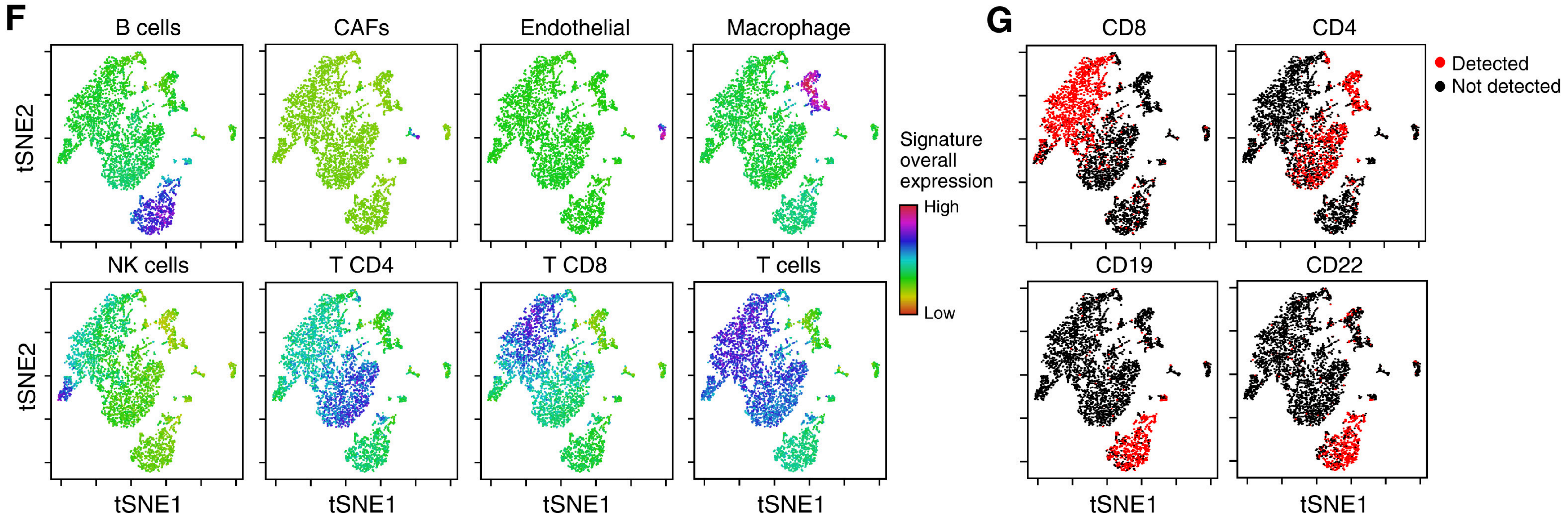
Trajectory analysis using Slingshot



Hung et al., 2019, A cell atlas of the Drosophila midgut, PNAS.

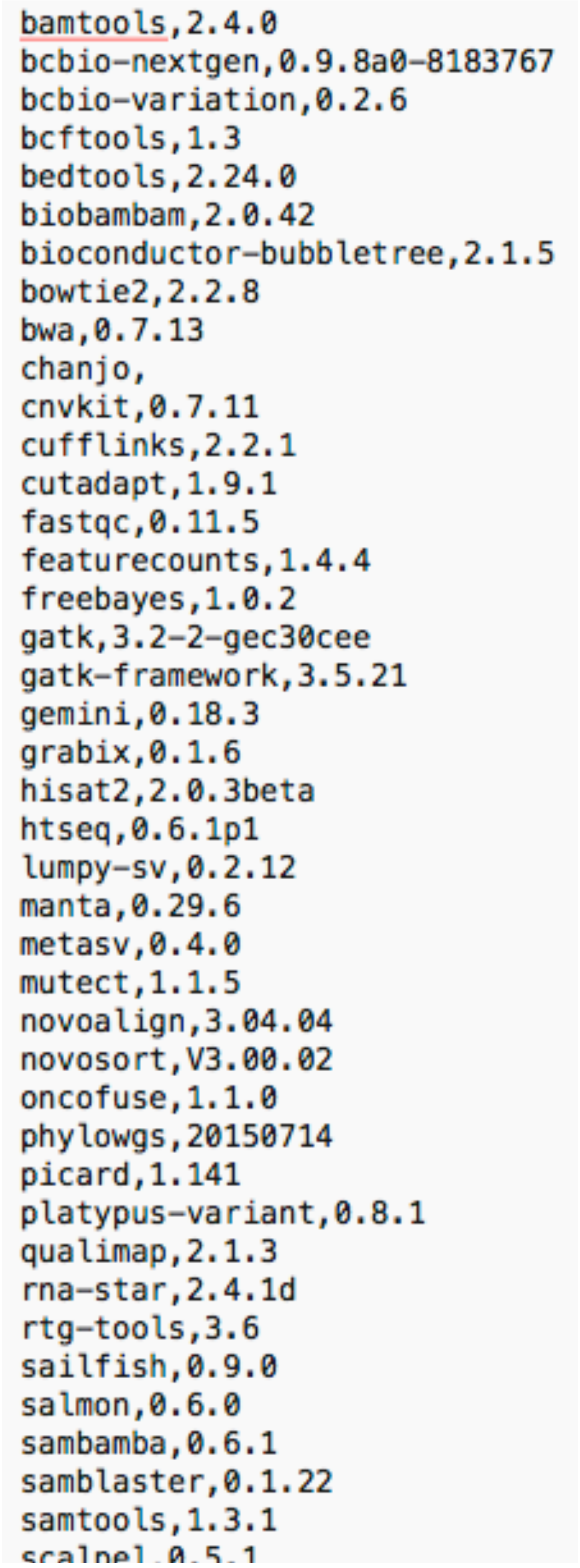
New technologies we're preparing for

- CITE-seq: Cellular Indexing of Transcriptomes and Epitopes by Sequencing
  - scRNAseq with quantitative and qualitative information on surface proteins with available antibodies
- SLAMseq: High throughput metabolic sequencing of transcripts
  - Analyze transcriptome-wide kinetics of RNA synthesis and turnover
  - Measure nascent RNA expression and transcript stability
  - Enhance the temporal resolution of differential expression
- CyCIF: Highly Multiplexed Cell and Tissue Imaging
  - Spatial profiling using multiplexed antibody staining



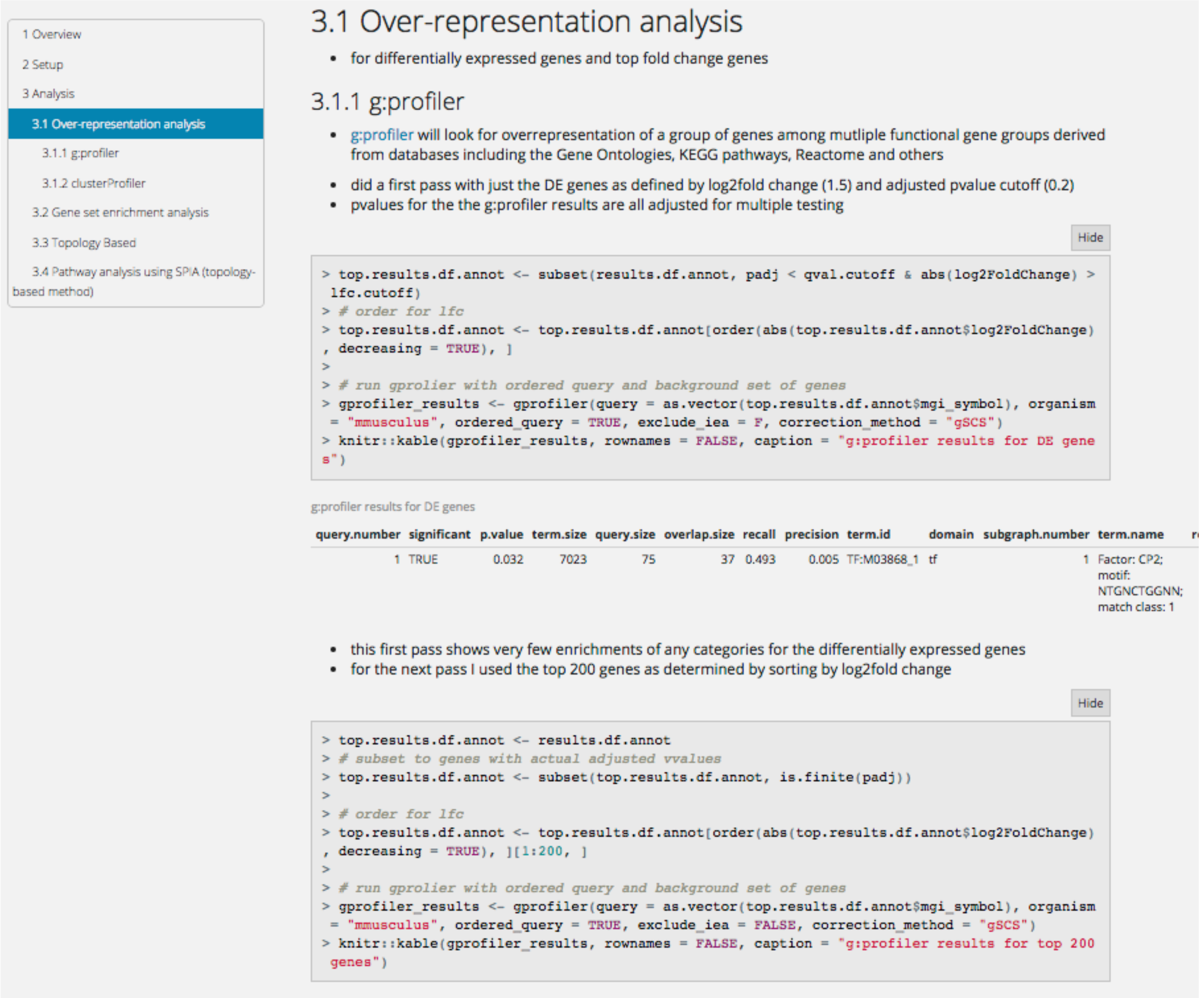
Reproducibility and re-use

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
salmon,0.6.0
sambamba,0.6.1
samblaster,0.1.22
samtools,1.3.1
scalpel,0.5.1
```

Rmarkdown report with code



3.1 Over-representation analysis

- for differentially expressed genes and top fold change genes

3.1.1 g:profiler

- g:profiler will look for overrepresentation of a group of genes among mutliple 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 the g:profiler results are all adjusted for multiple testing

```
> top.results.df.annot <- subset(results.df.annot, padj < qval.cutoff & abs(log2FoldChange) >
 lfc.cutoff)
> # order for lfc
> top.results.df.annot <- top.results.df.annot[order(abs(top.results.df.annot$log2FoldChange)
, decreasing = TRUE), ]
>
> # run gprolier with ordered query and background set of genes
> gprofiler_results <- gprofiler(query = as.vector(top.results.df.annot$mgi_symbol), organism
 = "mmusculus", ordered_query = TRUE, exclude_iea = F, correction_method = "gSCS")
> knitr::kable(gprofiler_results, rownames = FALSE, caption = "g:profiler results for DE gene
s")
```

g:profiler results for DE genes

query.number	significant	p.value	term.size	query.size	overlap.size	recall	precision	term.id	domain	subgraph.number	term.name	r
1	TRUE	0.032	7023	75	37	0.493	0.005	TF:M03868_1	tf	1	Factor: CP2; motif: NTGNCTGGNN; 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 vvalues
> top.results.df.annot <- subset(top.results.df.annot, is.finite(padj))
>
> # order for lfc
> top.results.df.annot <- top.results.df.annot[order(abs(top.results.df.annot$log2FoldChange)
, decreasing = TRUE), ][1:200, ]
>
> # run gprolier with ordered query and background set of genes
> gprofiler_results <- gprofiler(query = as.vector(top.results.df.annot$mgi_symbol), organism
 = "mmusculus", ordered_query = TRUE, exclude_iea = FALSE, correction_method = "gSCS")
> knitr::kable(gprofiler_results, rownames = FALSE, caption = "g:profiler results for top 200
 genes")
```

Pipeline development

- Updating the code base to meet the Python3 standard
- Supporting background inputs for copy number variant (CNV) calling to allow for a pre-computed panel of normals for tumor-only or single sample variant calling
- Implementing tumor-only variant calling with duplex barcodes
- Whole-genome bisulfite-seq pipeline based on bismark2 for DNA methylation analysis with the Illumina Truseq Methyl Capture platform
- Making it multiple orders of magnitudes faster to set up runs with thousands of inputs.
- For bulk transcriptomics data, added support for gene fusion calling with Arriba
- Set up the hg38 reference in STAR
- Enabled 2-pass STAR alignment, which performs sequential alignment, genome indexing and re-alignment to improve quantification of novel splice junctions
- Notably, we also resolved more than 500 issues on the bcbio github
- Now funded through the CZI

Pipeline development

- ATAC-seq/ChIP-seq pipeline
    - Stabilizing infrastructure
    - Integration of ATACqv
    - Shifting, removing low quality mapping, subsetting in to NFR regions
    - Still problems with experimental design (lack of replicates)
- 5hmC-seq analysis to detect regions with differential methylation
    - Early stages
- Variant calling analysis
    - Exploring cloud options for a large projects
    - Will be learning more about DRAGEN and Terra in the near term

Training in 2019

- 30 workshops focused on basic data analysis skills and NGS analysis
  - spanning 42 training days and training over 900 researchers
- For the third year, members of the training team were involved in the development of the online “Introduction to Omics Research” course offered by Harvard Catalyst
- Continued to collaborate with Harvard’s FAS Research Computing “RC” group to make high performance computing more accessible to Chan School researchers
- Launched a monthly Bioinformatics Breakfast community event
- Explored new models for training and collaborating, including embedding core bioinformaticians within research labs to train lab members in private group settings and to oversee their analyses, and mentoring students from our training program to perform bioinformatics consulting within the community
- We piloted a new program with more workshops every month and have moved away from the intensive course model

Acknowledgements

Harvard Chan School of Public Health

Harvard Medical School

Harvard Stem Cell Institute

Harvard Catalyst

National Institute of Environmental Health Sciences (NIEHS)

Harvard University Center for AIDS Research (CFAR)

AstraZeneca

Boehringer Ingelheim

Chan Zuckerberg Initiative

Perrimon Lab

- Norbert Perrimon
- Sudhir Tattikota
- Ruie-jiun Hung

HMS RC

HMS Single Cell Core

Biopolyers Facility

Molecular Biology Core Facilities

Bauer Core

BWH Single Cell Core