Open Workflows to Enable and Understand Single-Cell RNA-Seq Data

Nathan Salomonis

Department of Biomedical Informatics, CCHMC

Goals for this session

1. Discuss what is needed to enable scientists to analyze and interpret single-cell genomics data.

2. Overview the technologies, best practices and infrastructure needed to accomplish the emerging questions afforded by these new methods.

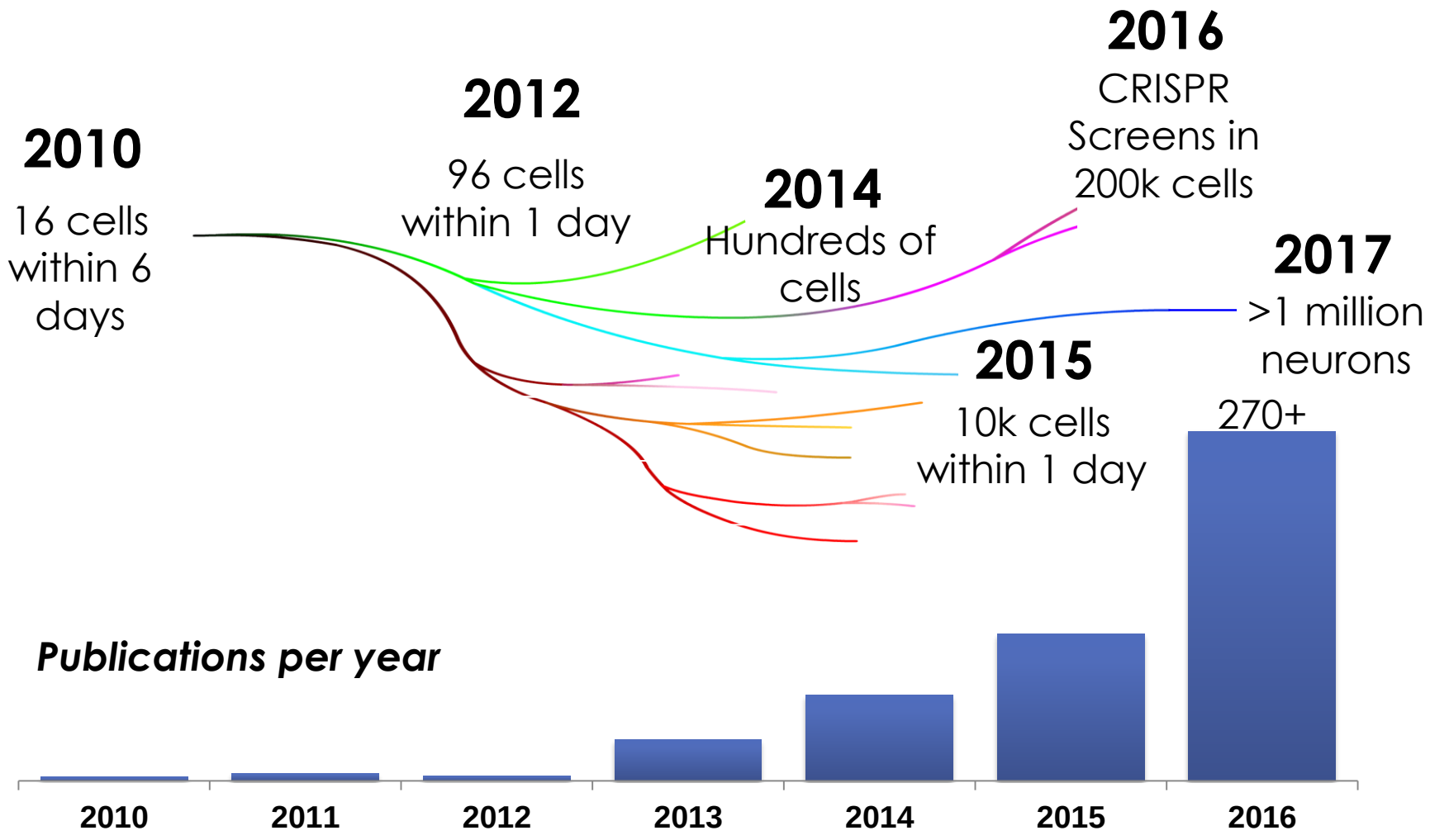
3. Provide workflows which you and your users can apply on their own or in collaboration.

About Me

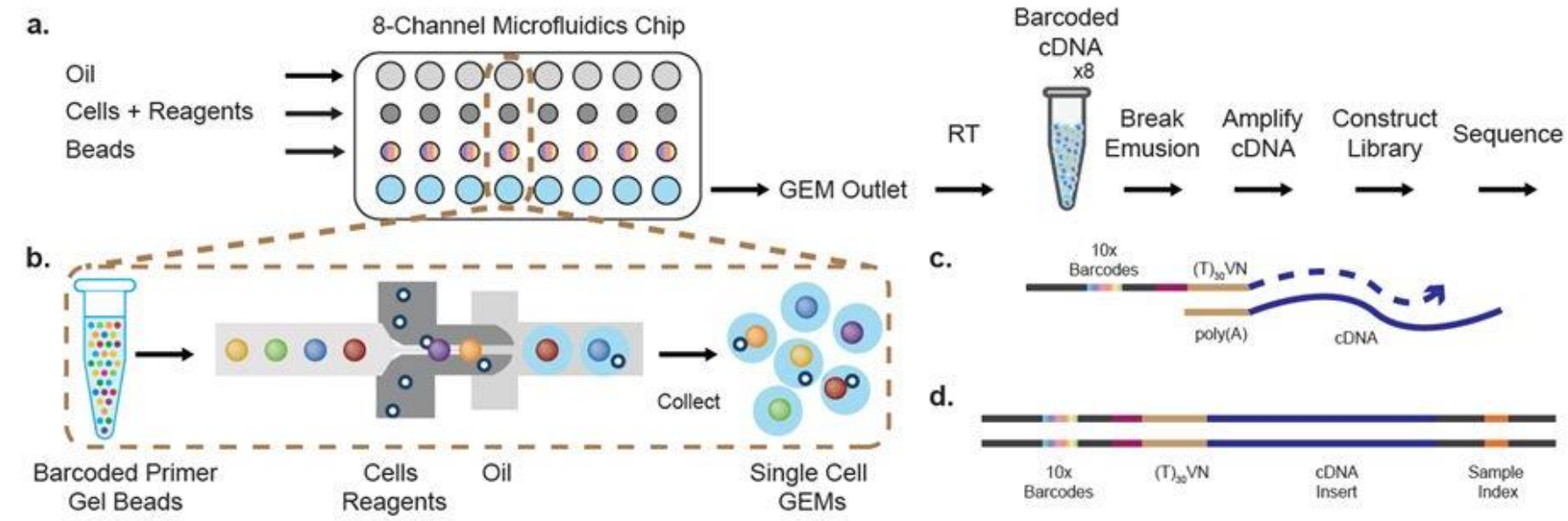
- Computational genomicist – our lab focused on software development, outreach and analysis.

- We automate initial single-cell RNA-Seq analyses for folks at our institution at no cost, but encourage labs to perform downstream analyses on their own (*AltAnalyze, ICGS, cellHarmony, DoubletDecon, Schrödinger*).

Evolution of Single-Cell Genomics



Droplet Bead Sequencing -10X Genomics



http://core-genomics.blogspot.com/2016/07/10x-genomics-single-cell-3mrna-seq.html

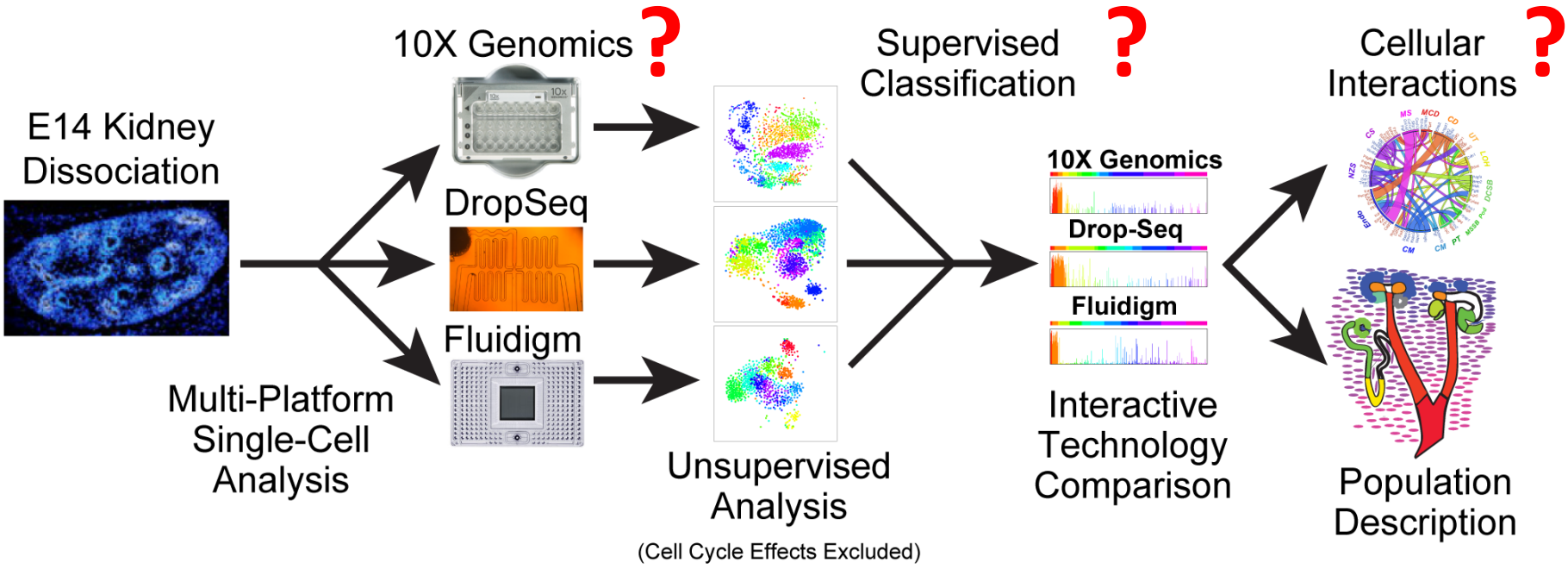
Applications

- 100-15,000 cells measured per experiment.
- Unique molecular indexes or UMI, 3' gene biased.
- Typically low depth-seq (<200k reads) and lower confidence gene exp.
- 1-10% doublets (QC challenging to detect).
- No splicing, polyA containing genes only.

Emerging Technologies for Single-Cell Genomics

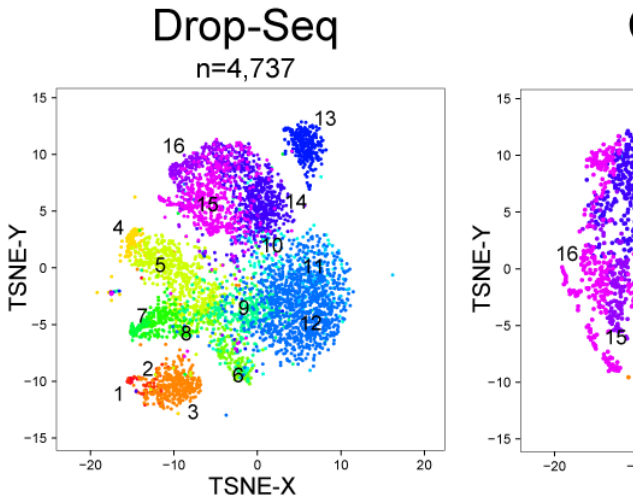
- Single-Cell RNA-Seq
  - Smarter low throughput (Fluidigm)
  - High-throughput 3' or 5' biased (10x, Drop-Seq, Microwell-Seq)
- Single-Cell ATAC-Seq
  - Low resolution, TF activities inferred
- CITE-Seq, BCR/TCR Single-Cell Analysis
  - Couple scRNA-Seq with the detection of specific RNAs or labelled proteins

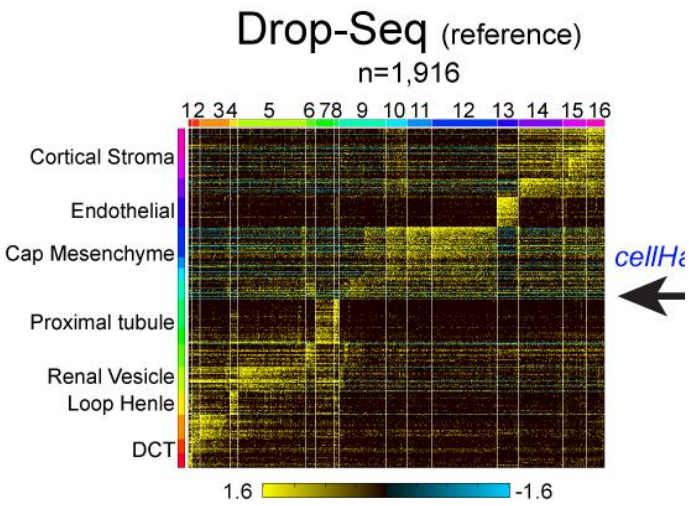
Which Platforms to Choose and Which Give Comparable Results

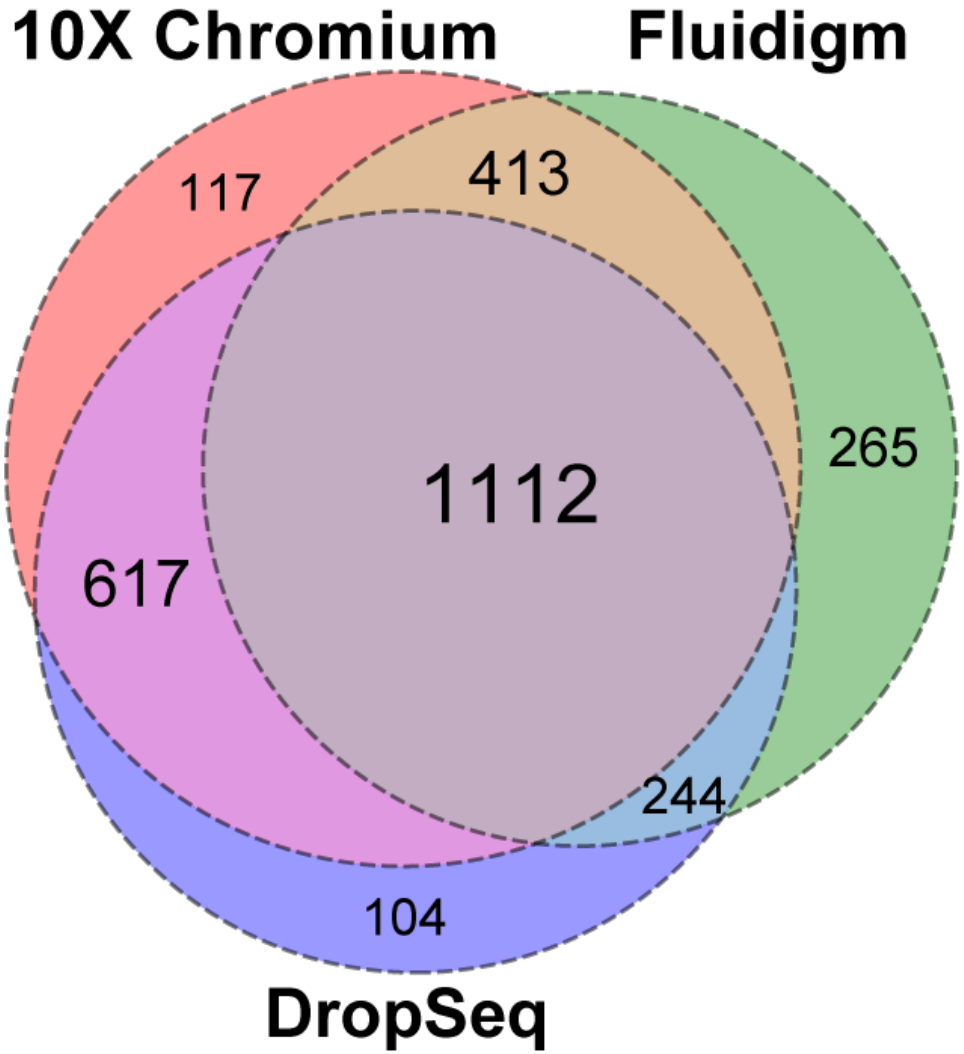


Magela et al. 2017 Developmental Biology

Different Technologies Provide Quantitative and Qualitatively Different Results







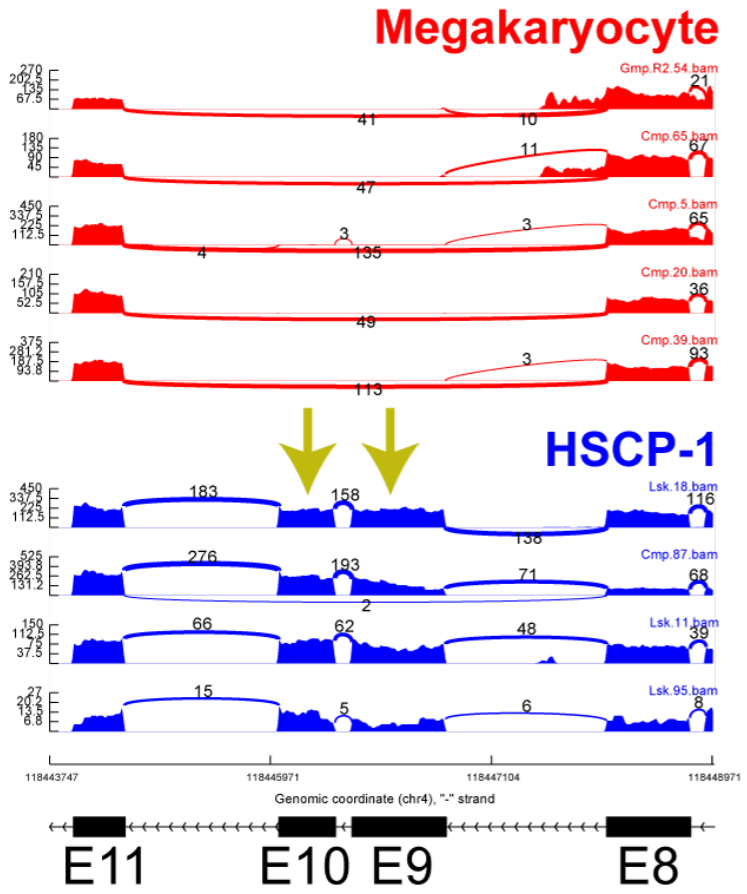
Magela et al. 2017 Developmental Biology

Not all scRNA-Seq Analyses Are the Same

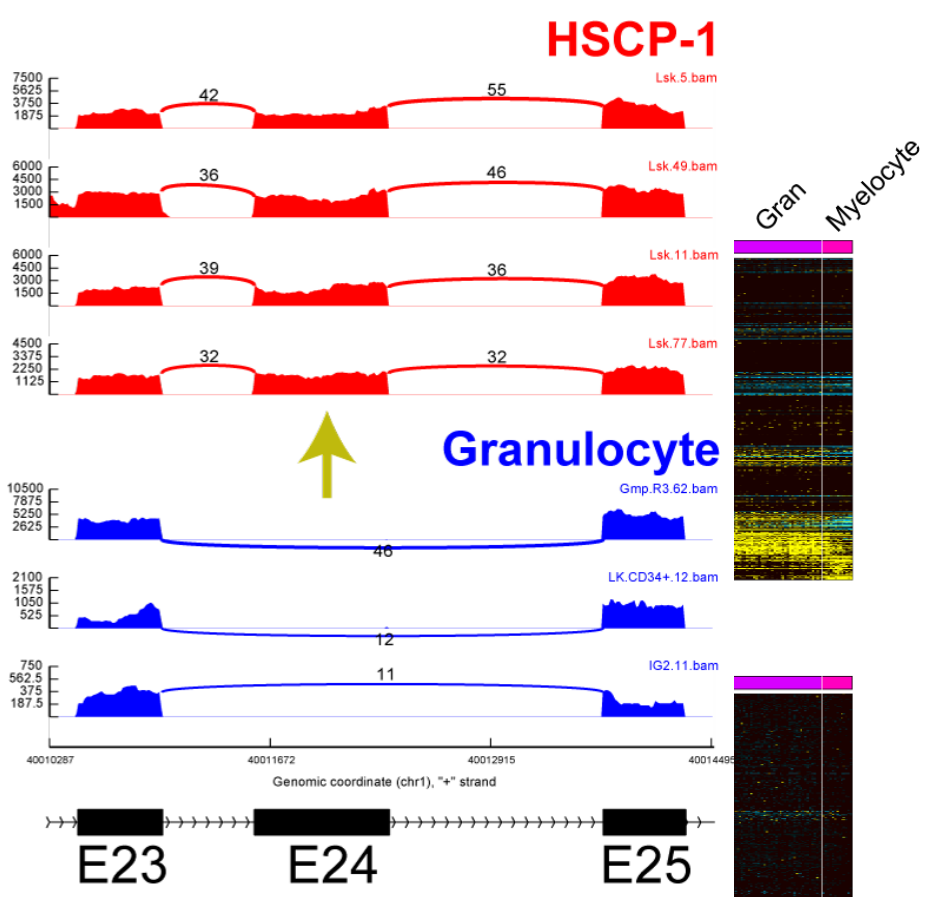
Each experiment has it's own technical challenges and needed bioinformatics expertise.

- Defining notably different cell states.
  - Numerous well-established automated workflows exist.
- Defining subtle heterogeneity or transitional cell states.
  - Testing of many tools likely required (no perfect methods).
- Comparing different scRNA-Seq Datasets
  - Requires prior bioinformatics experience and novel tools.
- Evaluating BCR/TCR subsets
  - Requires immunology and informatics expertise.
- CRISPR Screens
  - Significant bioinformatics and investment expertise required.

Developing Reliable Methods to Interface Multi-Omics Single-Cell Data



Mpl



Map4k4

How Can Core Facilities Aid in the Processing and Interpretation of Massive Datasets?

- Communicating Quality Control
- Establishing Workflows that Automate Primary Analyses
- Partnering with Computational Genomics Collaborators/Vendors
- Providing Documentation on User-Friendly Workflows Users Can Run (e.g., AltAnalyze)

Analysis of 10x Genomics Datasets

- Ideally requires a cluster computing environment.
- Software from 10x Genomics is open source and is run on the command-line (Cell Ranger).
- Works with input bcl2 sequence files directly from the sequencing core.
- Runs in 1-2 days, requiring 200GB.
- Provides an interactive viewer to explore their results (Loupe Browser).
- Starting point for deeper analysis with various tool kits.

Analysis of 10x Genomics Datasets

file:///Volumes/salomonis2/Bliss-Magella-10x/10x_WT/outs/web_summary.html

10x GENOMICS

Cell Ranger · 10x_WT ·

SUMMARY ANALYSIS

Estimated Number of Cells
2,295

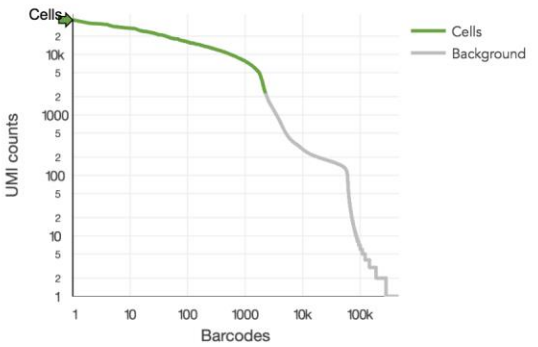
Mean Reads per Cell
41,916

Median Genes per Cell
2,556

Sequencing

Number of Reads	96,199,414
Valid Barcodes	97.7%
Reads Mapped Confidently to Transcriptome	67.0%
Reads Mapped Confidently to Exonic Regions	71.1%
Reads Mapped Confidently to Intronic Regions	16.4%
Reads Mapped Confidently to Intergenic Regions	3.3%
Sequencing Saturation	48.5%
Q30 Bases in Barcode	99.2%
Q30 Bases in RNA Read	95.8%
Q30 Bases in Sample Index	98.9%
Q30 Bases in UMI	99.2%

Cells

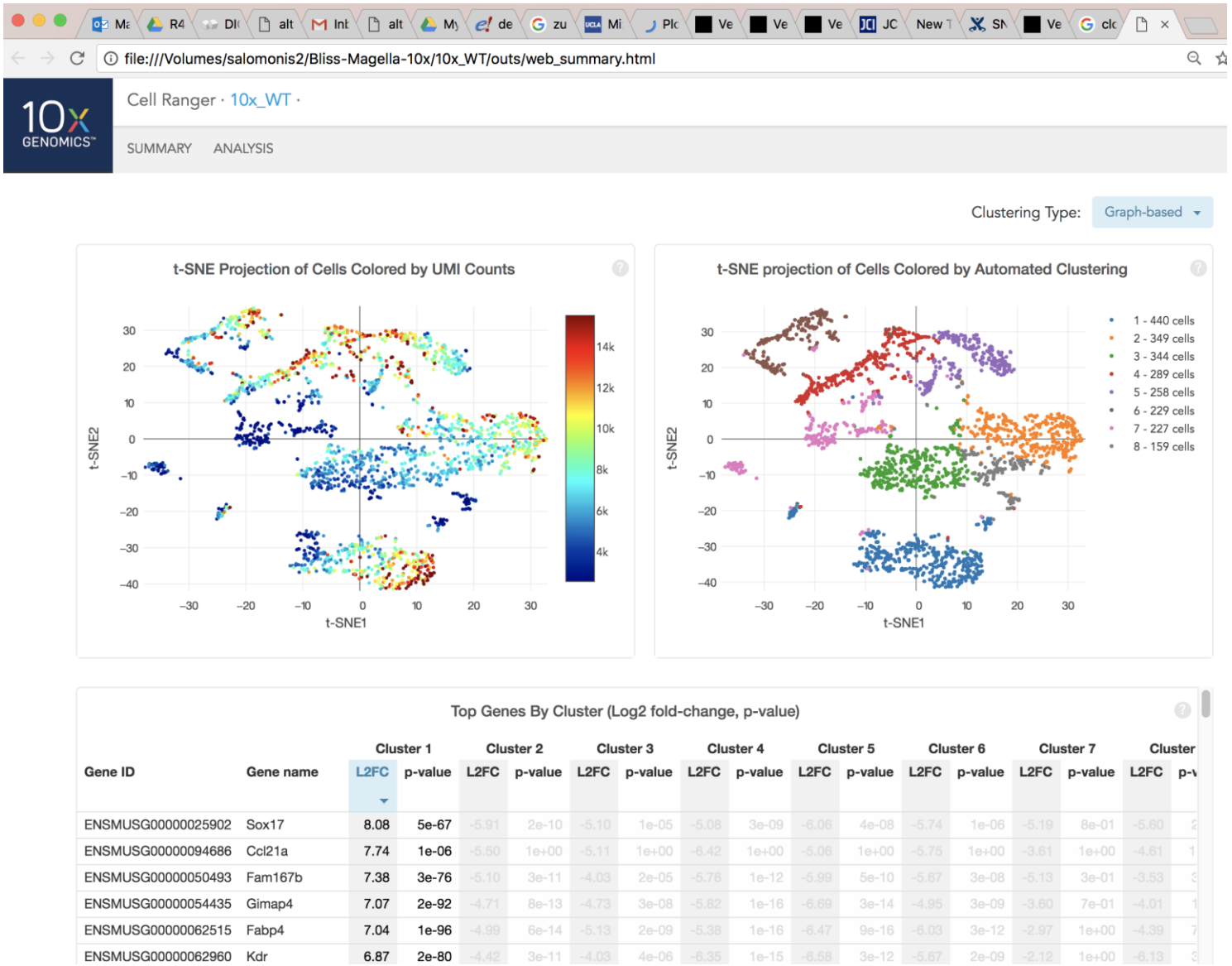


Estimated Number of Cells	2,295
Fraction Reads in Cells	55.6%
Mean Reads per Cell	41,916
Median Genes per Cell	2,556
Total Genes Detected	17,964
Median UMI Counts per Cell	7,121

Sample

Name	10x_WT
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v2
Cell Ranger Version	1.3.0

Analysis of 10x Genomics Datasets



Top Genes By Cluster (Log2 fold-change, p-value)

Gene ID	Gene name	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7		Cluster
		L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-v
ENSMUSG00000025902	Sox17	8.08	5e-67	-5.91	2e-10	-5.10	1e-05	-5.08	3e-09	-6.06	4e-08	-5.74	1e-06	-5.19	8e-01	-5.60	2
ENSMUSG00000094686	Ccl21a	7.74	1e-06	-5.50	1e+00	-5.11	1e+00	-6.42	1e+00	-5.06	1e+00	-5.75	1e+00	-3.61	1e+00	-4.61	1
ENSMUSG00000050493	Fam167b	7.38	3e-76	-5.10	3e-11	-4.03	2e-05	-5.76	1e-12	-5.99	5e-10	-5.67	3e-08	-5.13	3e-01	-3.53	3
ENSMUSG00000054435	Gimap4	7.07	2e-92	-4.71	8e-13	-4.73	3e-08	-5.82	1e-16	-6.69	3e-14	-4.95	3e-09	-3.60	7e-01	-4.01	1
ENSMUSG00000062515	Fabp4	7.04	1e-96	-4.99	6e-14	-5.13	2e-09	-5.38	1e-16	-6.47	9e-16	-6.03	3e-12	-2.97	1e+00	-4.39	7
ENSMUSG00000062960	Kdr	6.87	2e-80	-4.42	3e-11	-4.03	4e-06	-6.35	1e-15	-6.58	3e-12	-5.67	2e-09	-2.12	1e+00	-6.13	3

Analysis of 10x Genomics Datasets

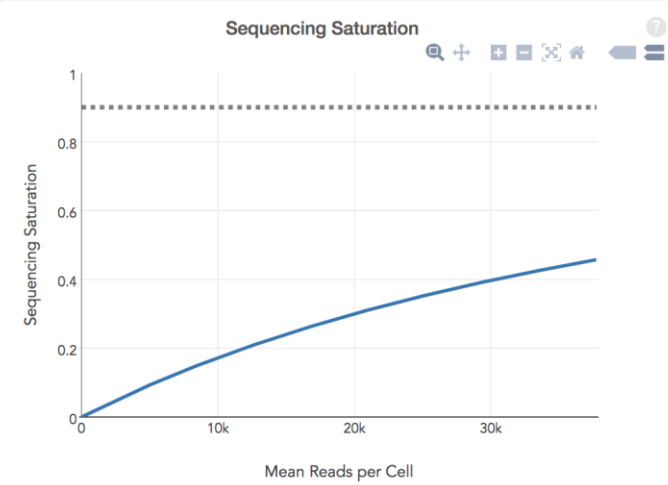
file:///Volumes/salomonis2/Bliss-Magella-10x/10x_WT/outs/web_summary.html

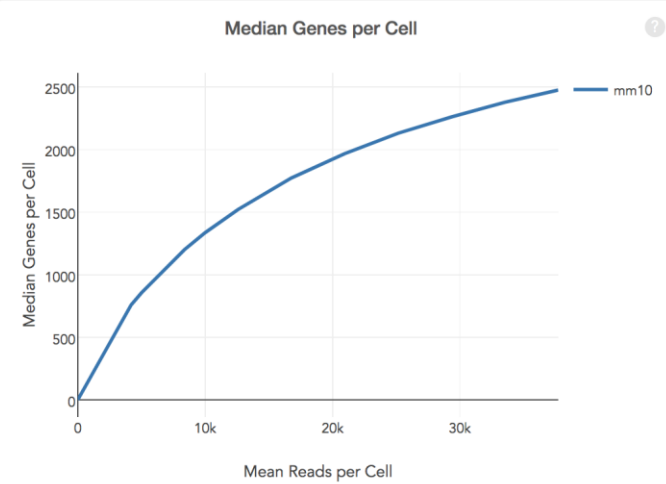
10x GENOMICS

Cell Ranger · 10x_WT ·

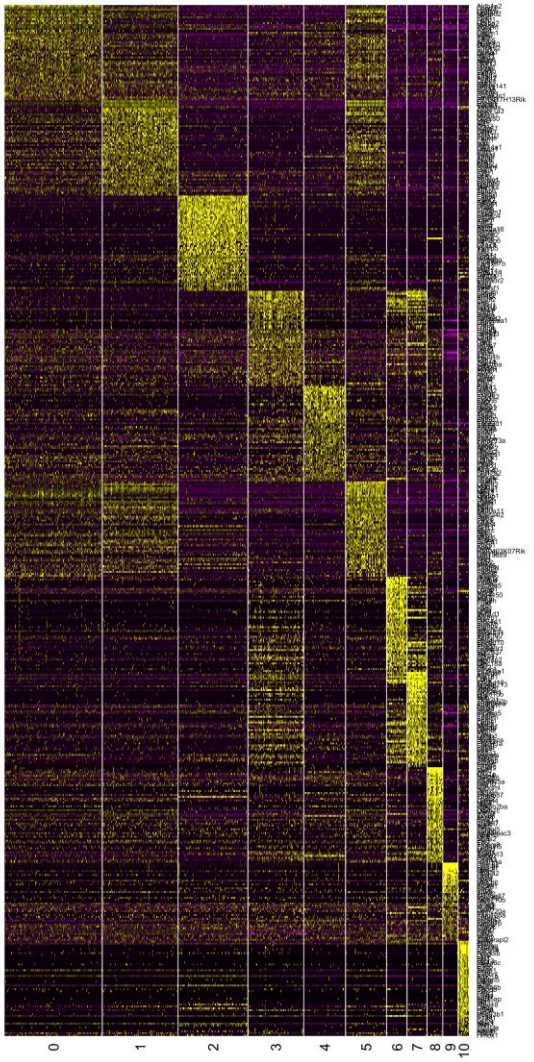
SUMMARY ANALYSIS

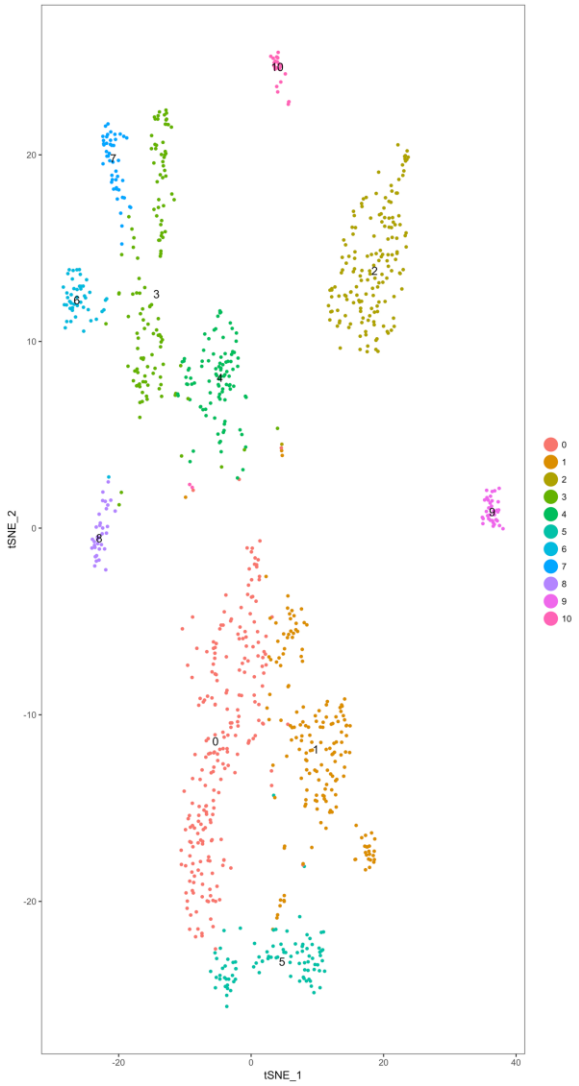
ENSMUSG00000062960	Kdr	6.87	2e-80	-4.42	3e-11	-4.03	4e-06	-6.35	1e-15	-6.58	3e-12	-5.67	2e-09	-2.12	1e+00	-6.13	3
ENSMUSG00000029648	Flt1	6.85	5e-72	-4.82	7e-11	-4.60	2e-06	-5.74	1e-12	-4.97	1e-08	-5.07	2e-07	-1.93	1e+00	-5.52	5
ENSMUSG00000044338	Aplnr	6.70	3e-73	-4.15	1e-09	-4.83	6e-07	-5.29	1e-12	-6.52	2e-11	-7.21	3e-10	-1.64	1e+00	-6.07	5
ENSMUSG00000090019	Gimap1	6.66	1e-79	-5.19	8e-13	-5.38	2e-08	-4.09	1e-10	-5.75	3e-11	-5.02	1e-08	-1.87	1e+00	-5.89	3
ENSMUSG00000044258	Ctla2a	6.66	9e-73	-5.05	1e-11	-4.02	1e-05	-4.97	8e-12	-5.61	5e-10	-4.88	1e-07	-2.00	1e+00	-4.75	9
ENSMUSG00000001029	Icam2	6.29	6e-79	-4.37	9e-12	-3.90	2e-06	-5.22	1e-14	-6.60	8e-14	-4.70	1e-08	-1.55	1e+00	-4.73	5
ENSMUSG00000016494	Cd34	6.27	1e-79	-4.34	8e-12	-3.92	2e-06	-4.74	2e-13	-5.49	4e-12	-4.52	2e-08	-2.11	1e+00	-4.53	5
ENSMUSG00000027435	Cd93	6.26	1e-70	-4.64	3e-11	-3.53	4e-05	-5.14	1e-12	-6.37	2e-11	-3.88	3e-06	-1.79	1e+00	-4.33	1
ENSMUSG00000020717	Pecam1	6.20	1e-72	-4.80	5e-12	-3.40	5e-05	-5.46	4e-14	-5.69	3e-11	-4.56	9e-08	-1.35	1e+00	-4.65	7
ENSMUSG00000022579	Gpihbp1	6.17	1e-43	-5.39	3e-08	-5.16	9e-05	-5.30	4e-08	-4.95	1e-05	-4.63	1e-04	-0.46	1e+00	-4.08	7
ENSMUSG00000044562	Rasip1	6.05	4e-74	-3.92	2e-10	-4.11	1e-06	-5.16	3e-14	-4.22	1e-08	-5.17	3e-09	-1.85	1e+00	-3.55	2





Seurat Analysis Workflow

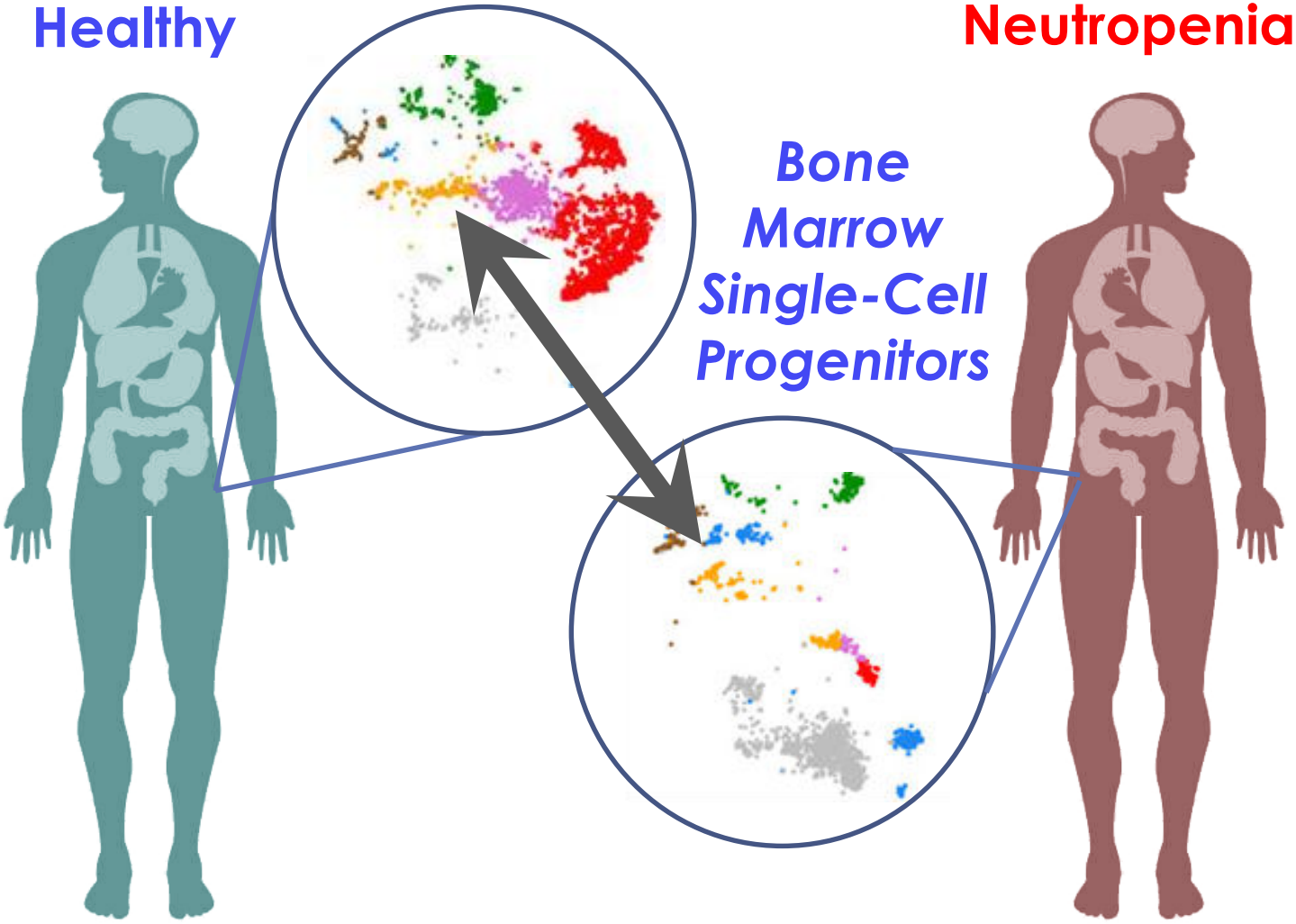




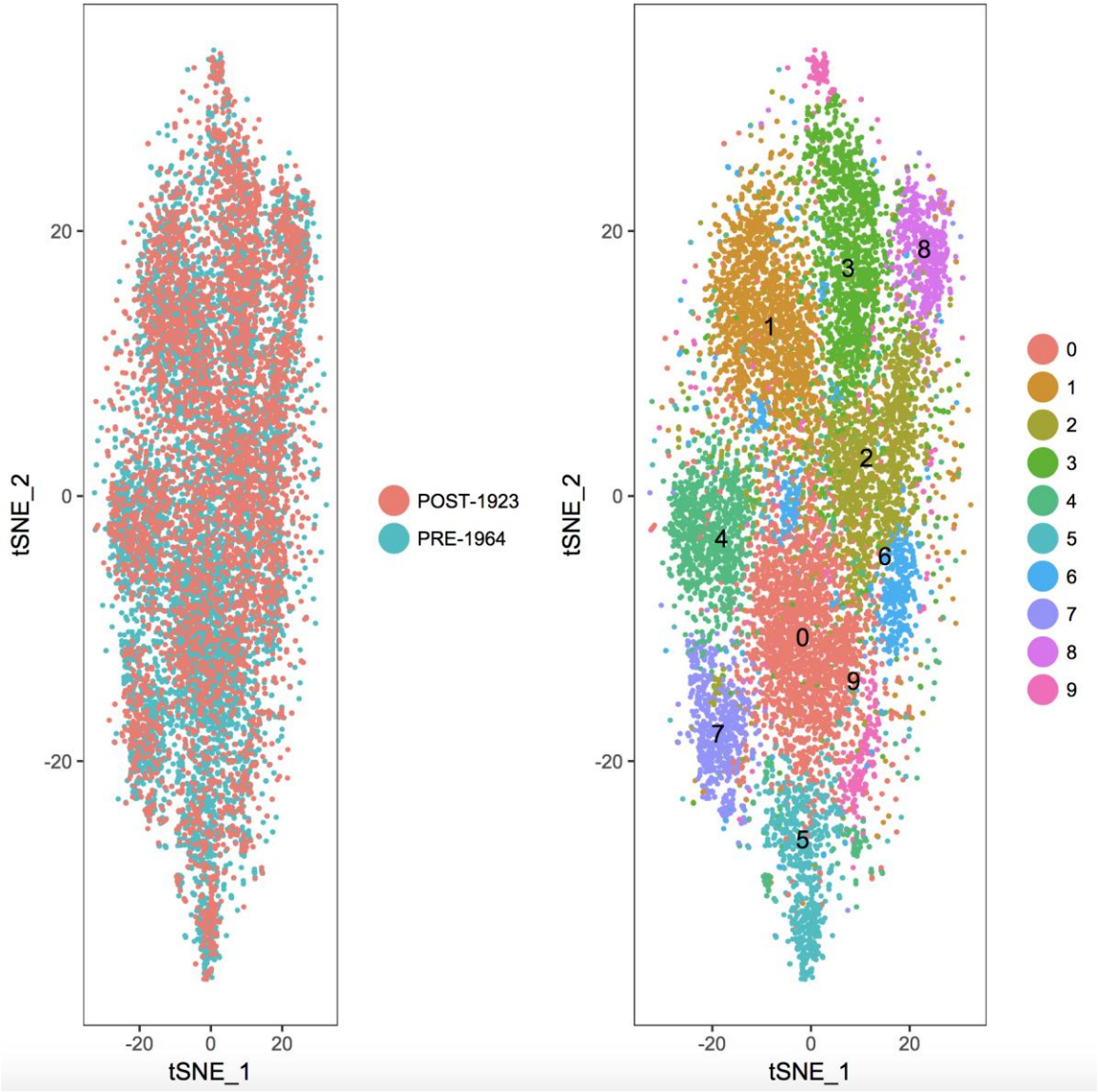
- Fast
- Handles >100k cells
- Command-Line
- Well-supported
- Many options
- Can visualize the expression of genes on the t-SNE plot.

Satija R. et al. Nat Biotechnol. 2015 May;33(5):495-502

Using Single-Cell Genomics to Understand the Cellular Origins of Disease



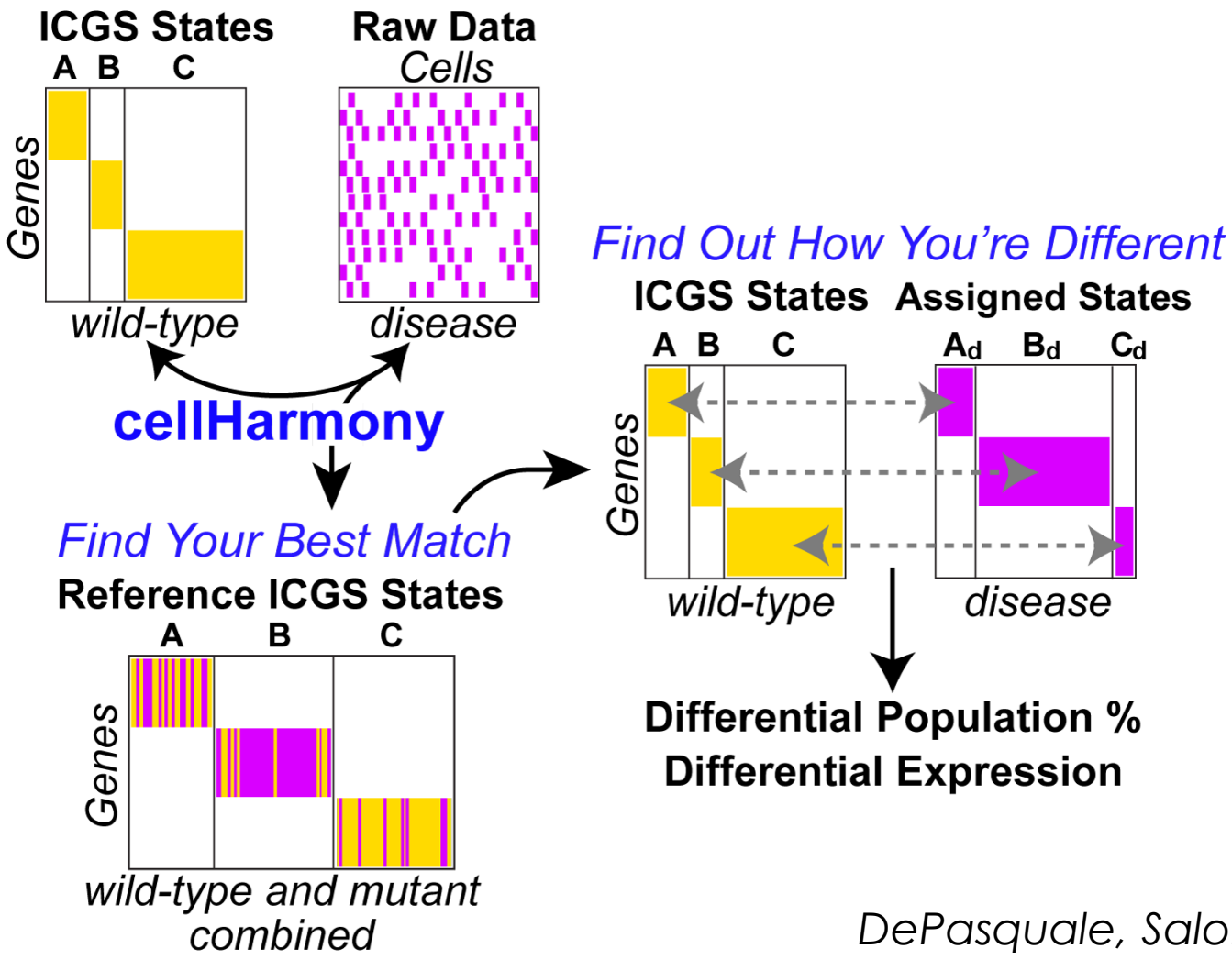
Seurat Canonical Correlation Analysis



- Can align cells independent of batch
- Can align cells between species
- Requires similar cell types to align.

Butler A et al. Nature Biotechnol. 2018 Apr 2.

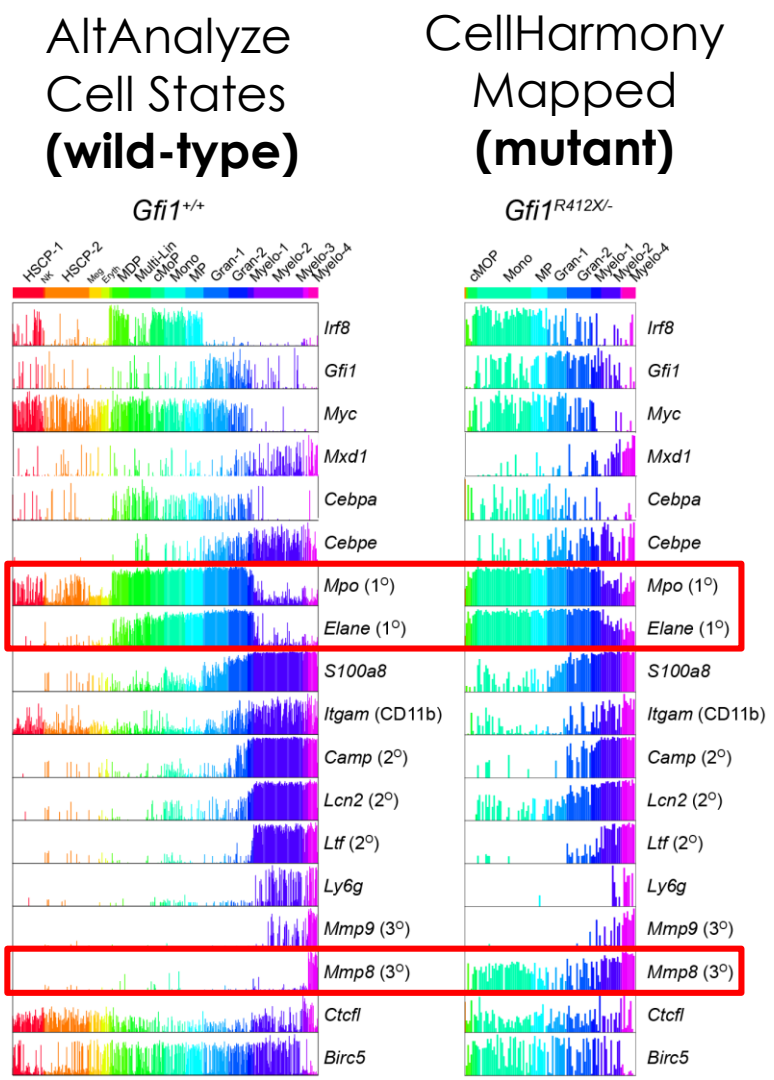
Cell Classification and Molecular Impact Analysis with cellHarmony



DePasquale, Salomonis in preparation

Defining Mis-Expressed Genes with CellHarmony

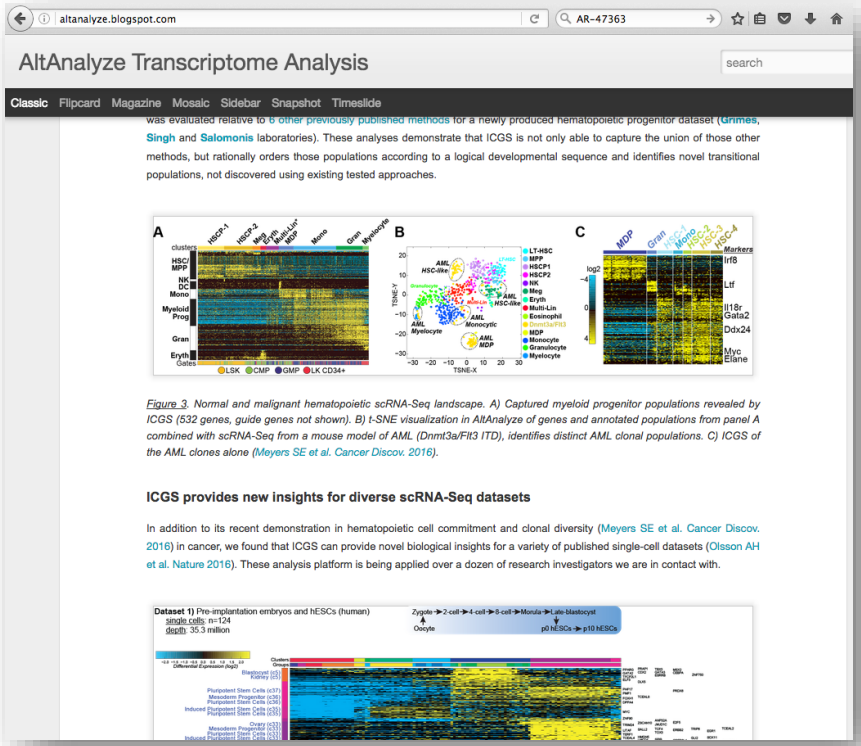
Row = gene
Column = single cell
Height = expression amplitude



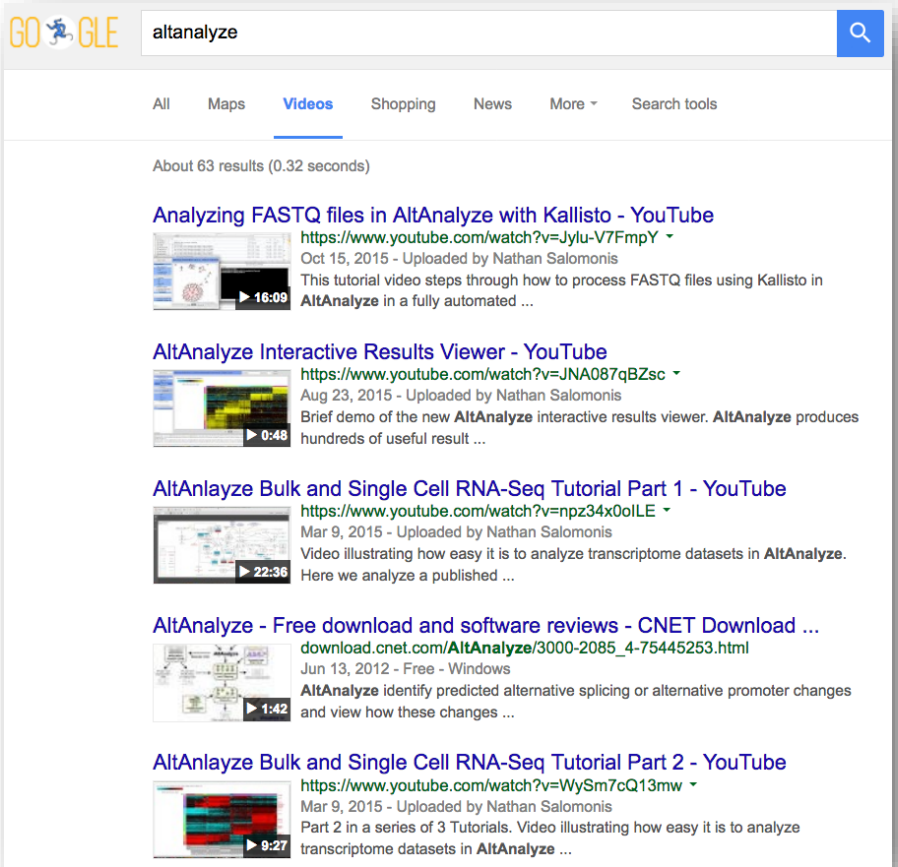
Muench, Salomonis, Grimes in preparation

AltAnalyze: Comprehensive and Intuitive Application for Genome Research

- Among the most easy-to-use freely available bioinformatics tools.
- Extensive documentation, blogs, tutorials, online help.



http://www.altanalyze.org



- Over 18,000 downloads
- >200 peer-reviewed citations
- Available for Mac, Windows, Linux and source-code.

Contact and Tutorials

Nathan Salomonis

Cincinnati Children's Hospital
email: Nathan.Salomonis@cchmc.org
email: altanalyze@gmail.com

website: http://altanalyze.org

tutorials: YouTube AltAnalyze