MIT 6.8700/6.8701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution

Lecture 13 (part 2)
Single-cell genomics applications

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

unbiased description of cellular state

bulk genomics

single-cell genomics

gel beads

single cells

oil

adapted from Shalek & Regev & G Zheng (10X Genomics)

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

1. Foundations: Why single-cell profiling

Why single cells

Bendall et al. (2011), Science

Dalerba et al. (2011), Nature Biotech

TLR Signaling

IRF3 Protein Levels - 4h LPS

fightcolorectalcancer.org/

Circulating Tumor Cells

Gore et al. (2005), Nature

Zebrafish early embryo

Cellular responses can vary substantially between "identical" cells.

Overcome low input

Single-cell RNA-Seq

Avital et al., *Genome Biology* (2014).

Shalek et al., *Nature* (2014).

Whole-sample analysis can lead to misleading views

The average may not represent the population

Rare events can be lost …

scRNA data looks like RNA-seq

RNA FISH Validates scRNA-Seq cell-cell variability

Key single cell results can be validated using amplification-free methods.

Shalek*, Satija* et al, *Nature*, **498**, (2013)

Single-Cell RNA-Seq captures inter-cellular variability

18 Single Dendritic Cells Stimulated with LPS

We can profile the full transcriptomes of cells with SMART-Seq.

Shalek*, Satija* et al, *Nature*, **498**, (2013)

Housekeeping vs. variable genes

Least Variable Genes

Term	P-Value
Structural constituent of ribosome	1.66e-51
Structural molecule activity	1.16e-33

Most Variable Genes

Term	P-Value
Defense response	2.23e-29
Inflammatory response	7.99e-25

Screenshots are log scale in IGV

Housekeeping & ribosomal genes are among the least variable, while immune response elements are among the most variable.

Shalek*, Satija* et al, *Nature*, **498**, (2013)

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

2. Scaling up scRNA-seq technology

Exponential scaling of single-cell RNA-seq in the past decade

Cell numbers reported in representative publications by publication date. Key technologies are indicated.

Svensson V, Vento-Tormo R, Teichmann SA. Nat Protoc. 2018 Apr;13(4):599-604.

Evolution of methods for isolating single cells for profiling

All methods seek to: separate cells, amplify RNA, sequence

Paplexi et al. 2017

scRNA-seq technologies vary in cost and sensitivity

Paplexi et al. 2017	FACS	CyTOF	qPCR	Plate-based protocols (STRT-seq, SMART-seq, SMART-seq2)	Fluidigm C1	Pooled approaches (CEL-seq, MARS-seq, SCRB-seq, CEL-seq2)	Massively parallel approaches (Drop-seq, InDrop)
Cell capture method	Laser	Mass cytometry	Micropipettes	FACS	Microfluidics	FACS	Microdroplets
Number of cells per experiment	Millions	Millions	300–1,000	50–500	48–96	500–2,000	5,000–10,000
Cost	$0.05 per cell	$35 per cell	$1 per cell	$3–6 per well	$35 per cell	$3–6 per well	$0.05 per cell
Sensitivity	Up to 17 markers	Up to 40 markers	10–30 genes per cell	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	6,000–9,00 genes per cell for cell lines; 1,000–5,000 genes per cell for primary cells	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	5,000 genes per cell for cell lines; 1,000–3,000 genes per cell for primary cells

CEL-seq, cell expression by linear amplification and sequencing; CyTOF, cytometry by time of flight (mass cytometry); FACS, fluorescence-activated cell sorting; InDrop, indexing droplets sequencing; MARS-seq, massively parallel single-cell RNA sequencing; qPCR, quantitative PCR; SCRB-seq, single-cell RNA barcoding and sequencing; STRT-seq, single-cell tagged reverse transcription sequencing.

Single-Cell Expression Profiling Pipeline

1. Cell Harvest

- Harvest cells in media
- Pre-enrich (FACS)

Cell work

2. Single Cell Preparation

(a) C_1: Cells → Whole Transcriptome Amplification
(b) Multiwell Plates
(c) Next Gen Technologies

Single cell processing

3. Expression Profiling

Library construction, validation, HiSeq

We can go from single cells to aligned reads in less than a day

w/ Fluidigm, E. Macosko, S. McCarroll, A. Regev, D. Weitz, C. Love, T. Gierahn, & Others

Drop-seq: Droplets as reaction chambers (10x)

SPLIT-seq/sci-RNA-seq: Sequential combinatorial barcoding

- Single cells never individually isolated
- Instead: fixed, and mRNA is manipulated in situ inside each cell
- Split cells into ~100 wells (e.g. 96 or 384-well plate) with unique barcodes in each well
- Labels all cells with a first barcode, for that well. Chance of same barcode: 1/100
- Pool cells, shuffle, split again, **randomly** re-assorting into same set of ~100 wells
- Add second barcode. Chance of same 2 barcodes: 1/10,000.
- Repeat: pool, shuffle, split, add 3rd barcode. Chance of same 3 barcodes: 1/1,000,000
- Can scale number of cells exponentially by number of barcoding rounds

Single-cell Profiling technologies

1. Cells in wells, traps, and valves (nanowell, Flow sorting, Fluidigm C1)

- Screen for and retrieve single cells of interest
- Enrich for rare cells with desired properties
- Control the cellular microenvironment
- Monitor or control cell-cell interactions
- Precise/extensive manipulation of single cells

Passive wells

Active pumps and valves

2. Droplets (Drop-seq, ddPCR)

- Introduce distinct “packets” of reagents to single cells
  - e.g., primers, barcodes
- Perform amplification on individual cells
- Sort large populations of single cells

3. Combinatorial indexing (SCI-seq, SPLiT-seq)

- Economic use of reagents for cell separation
- Efficiency of handling larger populations than Drop-seq
- Maintain complexities of population without bias from droplet or well.

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

3. Case studies: Single-cell Dissection of Brain Disorders

Dissect mechanisms of disease-associated regions

1. Disease genetics reveals common + rare variants/regions

2. Profile RNA + Epigenome in healthy + disease samples

3. Integrate data to predict driver genes, regions, cell types

4. Validate predictions in human cells + mouse models

5. Disseminate results

Single-cell profiling across 1500+ brain samples, 20M+ cells

- Alzheimer's
- FrontoTemporal
- LewyBodyDem
- ALS amyotr.lat.scl
- Huntington's
- AD+Psychosis
- Schizophrenia
- Bipolar Disorder
- Down Syndrome
- Autism
- Depression
- Suicide
- PTSD
- Aging

Across traits (N=12, each ~48)

Across individuals (N=1500)

Across cell types (n=12+)

Across individual cells (n=20,000,000+)

Frontal cortex
Cognition, personality, social behavior, psychosis in AD

Hippocampus
Information consolidation, memory, social behavior, psychosis in AD

Across brain regions (N=7+)

Across scRNA (N=1000), scATAC (N=500)

Single-cell AD dissection: sex diffs, early/late, myelin/WML

scRNA-seq in 48 individuals: cell type diversity

• All cell types affected • Female tx burden

Phenotypic diversity across individuals

Hansruedi Mathys
Jose Davila-Velderrain
Li-Huei Tsai
Nature, 2019

• Early: cell-type-specific
• Late: common, damage response

• Sex-specific & cell-type sp.
• Myelin: white-matter loss

Spatio-temporal AD progression across brain regions, cells, genes, pathology

1.6 M cells across 7 regions (48 indiv./region)

AD progression across regions, time, cells, genes

Markers for 30 Exc+23 Inh neuronal subtypes

Inhibitory neurons vary

Glial cells vary (eg.astrocytes)

Carles Boix,
Hans Mathys,
Kiki Galani,
Li-Huei Tsai

Deep Learning for Spatial Transcriptomics + Single-cell Integration

Jose Davila, Hansruedi Mathys, Li-Huei Tsai

snRNA profiles capture hippocampal sub-regions

Davila *et al*, In submission

Na Sun, Djuna Von Maydell, Guillaume Leclerc, Fatima Gunter-Rahman

Kai Kang, Kiki Galani, Myriam Heiman

Spatial transcriptomics single-cell deconvolution

Kang *et al*, In preparation

Spatial + single-cell possible, but slow, scarce, expensive

- **Train deep-learning model to predict spatial positioning information for single-cell data from expression patterns**
- **Recover neuronal layer information, well-recognized information**
- **Astrocytes and oligodendrocytes also show spatial positioning**

Van Maydell, Sun, *et al*, In preparation

Module DEG analysis in AD: correlated vulnerabilities in connected regions

Layer-selective AD vulnerability in entorhinal cortex & hippocampus

Coordinated DEGs across cells ➔ cell-type-specific module changes in AD

Communicating regions show correlated AD-assoc. changes

Microglia modules: M1: phagosome/immune. M18: Mitochondrial genes. M5: Glycolysis.
M8: Interferon-alpha. M2+M11: chaperone, heat shock (NFT, cog, not amyloid). M20: DNA-damage responding genes

Carles Boix,
Hans Mathys,
Kiki Galani,
Li-Huei Tsai

Single-cell insights into schizophrenia genetics/dysregulation

800k cells, 20+ cell states
SZTR cell type: transcr. resilience

PRS➔PTS: Polygenic-Transcriptome Score for Sz
SZTR individuals show rescued tx signatures

Dissect Schizophrenia GWAS loci: target genes, cell types of action

Ruzicka, *et al*, medRxiv 20225342, 2020

Master regulators of Sz DEGs: GWAS hits, devel/synaptic func.

SZ DEGs ➔ synaptic signaling/organization + devlpmt

Master regulators of SZ DEGs ➔ neurodevelopment

Coexpression of upstream master TF regulators ➔ five distinct modules
Core ➔ connects developmental and synaptic regulation

Stage-specific enrichments
SZTR: up+down, fetal+adult

Ruzicka, *et al*, medRxiv 20225342, 2020

Strong-effect APOE/ABCA7 variants converge on common lipid pathways

Strongest-effect AD variants: APOE, ABCA7

APOE,ABCA7: Lipid transport

Convergence into common mechanisms

scRNA profiling of ABCA7, APOE mutation carriers

Fatty acid metabolism dysregulation

Cholesterol/lipid regulation DEGs

Djuna Von Maydell,
Carles Boix,
Li-Huei Tsai

Psychosis in AD: broad **repression** of neuronal activity

231k cells, 96 samples (24 AD+P/AD-P, 24 M/F, 24 PFC/Hipp)

AD**+P** neuronal/glial changes+**ExcN Loss**

PFC: Ex/In **dn**, Ex/Ast **up.** **Hip**: Ex/In **dn**

Astroglia up: cilium, repress of apoptosis, signaling, migr

Neuron repression of ribosome biogenesis, neuronal projection, autophagy/apoptosis, secretion, signaling, energy

Na Sun, Kiki Galani, Li-Lun Ho, Robert Sweet, Julia Kofler, Li-Huei Tsai, Brad Ruzicka

Psychosis in AD (repression) appears opposite of Psychosis in Schizophrenia (overactivation)

Pathogen-associated changes in AD

Na Sun, Will Eimer, Rudy Tanzi

Viral pathogens in AD (HSV-1 and CMV): 189k cells, 75 individuals

- CMV+AD+: OPC increased in HSV1+
- AD-: Vascular increased in CMV+HSV1+
- CMV-HSV1-: microglia increased in AD+

Virus: increase in OPC, vasculature, microglia cell fractions

- nonAD
  1. CMV- vs CMV+
  2. CMV+: HSV1- vs HSV1+
- AD
  3. CMV- vs CMV+
  4. CMV+: HSV1- vs HSV1+

Virus-associated changes, with and without AD

CMV changes

CMV changes

HSV-1 vs. CMV: similar changes in non-AD, distinct in AD

Viral changes without AD

1

CMV+AD-

- Astro: regulation of endothelial cell migration-> blood vessel function
- Microglia: inflammatory response
- Oligo: cellular response to stress

CMV-AD-

- ExN: synaptic transmission, ion transport, axon development
- Oligo: regulation of neuron projection
- Microglia: cytokine-mediated signaling

2

HSV1+.CMV+AD-

- Astro: response to cytokine stimulus, regulation of angiogenesis
- Microglia: response to stress,
- response to cytokine stimulus
- Oligo: regulation of protein binding

HSV1-.CMV+AD-

- ExN: synaptic transmission, regulation of synapse assembly
- Oligo: regulation of cell migration
- Microglia: synapse pruning
- Astro: response to ion

Non-AD: Infl up, vessel up, synaptic transm./pruning down.

Viral changes with AD

3

CMV+AD+

- Microglia: response to cytokine, response to iron ion, regulation of immune response
- Oligo: myelination, metabolic response
- Astro: response to calcium ion

CMV-AD+

- ExN: synaptic transmission, neurotransmitter secretion, axon guidance
- Oligo: regulation of neurotransmitter receptor activity
- Astro: regulation of transcription, neuron migration

4

HSV1+.CMV+AD+

- Astro: regulation of local adhesion, regulation of axon extension
- Microglia: regulation of endothelial cell migration, protein stabilization
- Oligo: protein localization

HSV1-.CMV+AD+

- ExN: chemical synaptic transmission, synapse assembly
- Oligo: modulation of synaptic transmission
- Microglia: response to cytokine, cell junction
- Astro: response to ion

AD: Changes in Virus infection effects

Single-cell dissection of ALS/FTD in frontal/motor cortex

Cell states, Exc/Inh cortical layer distinctions

ALS/FTD show opposite directions in motor cortex

Female individuals more diverse expression

Age is a major driver of overall variability

Pineda, *et al*, In Submission

Distinct pathways change in ALS and FTD in motor cortex

cell_t	directi	Term	Category	-log10P	Genes	Overlap	OddsR	Score
Oligo	down	Protein processing in endoplasmic reticulum	Endo	24.0	ERO1B;UBE2D4;TUSC3;UBE2D2;UBE2D3;UBE2D1;FBXO6;C/	130/165	2.1	127
Astro	down	Protein processing in endoplasmic reticulum	Endo	23.3	ERO1A;ERO1B;UBE2D2;UBE2D3;UBE2D1;FBXO6;SEC62;SIL1	138/165	1.9	114
OPC	down	Protein processing in endoplasmic reticulum	Endo	12.7	TRAM1;HSP90AB1;PRKCSH;UBE2D3;UBE2D1;HERPUD1;SEC	61/165	3.0	104
Astro	down	Endocytosis	Endo	11.4	WASHC4;VPS29;WASHC5;SH3GL3;ZFYVE9;WIPF1;WIPF2;W/	164/244	1.5	47
Oligo	down	Endocytosis	Endo	6.7	WASHC4;VPS29;WASHC5;ZFYVE9;WIPF1;WIPF3;WASHC2C;	137/244	1.5	29
Astro	down	Protein export	Endo	5.5	SRP19;HSPA5;SRP54;SRP68;SRP14;SRP9;SEC61A2;OXA1L;IM	22/23	2.2	35
Ex1	up	Endocytosis	Endo	5.3	VPS29;SH3GL3;ZFYVE9;WIPF2;CLTC;CLTA;ARRB1;EEA1;KIF5	52/244	2.3	40
In4b	down	Protein processing in endoplasmic reticulum	Endo	5.0	PPP1R15A;HSP90AA1;HSP90AB1;PRKCSH;HSPA6;RAD23A;H	16/165	5.3	89
Astro	down	Autophagy	Autophagy	12.1	IRS1;IRS2;ZFYVE1;IGF1R;STK11;DEPTOR;LAMP1;AKT2;AKT3;	99/128	1.8	58
Astro	down	Ubiquitin mediated proteolysis	Autophagy	12.0	UBE3C;UBE2D2;UBE2D3;KEAP1;UBE2D1;UBE3A;CBLB;UBE3	104/137	1.7	56
Oligo	down	Autophagy	Autophagy	8.5	MTMR14;PTEN;IRS2;PIK3CB;HMGB1;IGF1R;STK11;DEPTOR;	85/128	1.8	42
Micro	down	Phagosome	Autophagy	7.7	MSR1;ATP6V0B;ITGB2;HLA-B;HLA-C;CYBA;HLA-A;TCIRG1;C	23/152	5.1	116
Astro	down	Mitophagy	Autophagy	6.5	CITED2;CALCOCO2;TBK1;FUNDC1;USP30;HRAS;USP8;BNIP3	51/65	1.8	33
Oligo	down	Ubiquitin mediated proteolysis	Autophagy	5.8	DET1;UBE2D4;UBE2D2;UBE2D3;KEAP1;UBE2D1;UBE3A;CBL	83/137	1.6	27
Oligo	down	Mitophagy	Autophagy	5.7	BECN1;USP15;AMBRA1;FOXO3;HIF1A;RELA;GABARAP;MAP	46/65	1.9	31
Oligo	down	Ribosome	TxRegulation	11.0	RPL4;RPL30;RPL3;RPL32;RPL34;RPL8;RPL10A;RPL6;RPL7;RPS	103/153	1.8	53
Oligo	down	Spliceosome	TxRegulation	10.9	RBM25;DDX46;EIF4A3;HNRNPU;PRPF19;USP39;PQBP1;SNR	93/134	1.8	55
Astro	down	Spliceosome	TxRegulation	7.7	TCERG1;RBM25;DDX46;DDX42;HNRNPU;PRPF19;USP39;PQ	94/134	1.6	35
Astro	down	N-Glycan biosynthesis	TxRegulation	6.7	DPAGT1;B4GALT2;B4GALT3;B4GALT1;RPN2;RPN1;FUT8;GA	42/50	1.9	37
Oligo	down	RNA transport	TxRegulation	6.6	CYFIP2;EIF4A2;POP5;EIF4A1;POP7;NUP188;GEMIN2;FMR1;	99/165	1.6	30
Astro	down	RNA degradation	TxRegulation	5.8	BTG2;DIS3L;ENO1;TENT4A;ENO2;TENT4B;EDC3;MTREX;NU	58/79	1.7	28
Astro	down	RNA transport	TxRegulation	5.0	CYFIP2;EIF4A2;CYFIP1;POP5;EIF4A1;NUP107;NUP188;POP1	103/165	1.4	20
Astro	down	Longevity regulating pathway	Differ/Cancer	7.9	ATF2;IRS1;EHMT2;PRKAG1;PRKAG2;PIK3CD;IRS2;PIK3CB;IG	76/102	1.7	38
Oligo	down	Longevity regulating pathway	Differ/Cancer	6.4	ATF2;CLPB;PRKAG1;EHMT1;PRKAG2;IRS2;PIK3CB;IGF1R;STK	67/102	1.7	32
Micro	down	Osteoclast differentiation	Differ/Cancer	6.2	CSF1R;SPI1;JUND;MITF;CYBA;TREM2;NFKB1;FOSL2;TNFRSF	19/127	5.0	95
Astro	down	Endometrial cancer	Differ/Cancer	5.5	GSK3B;TCF7;ARAF;PTEN;ILK;PIK3CD;PIK3CB;PIK3R1;FOXO3;	45/58	1.8	28
Astro	down	Colorectal cancer	Differ/Cancer	5.3	GSK3B;ARAF;PIK3CD;PIK3CB;CASP9;AKT2;CASP3;AKT3;AKT1	61/86	1.6	25
Oligo	down	Chronic myeloid leukemia	Differ/Cancer	5.2	RB1;CDKN1B;ARAF;PIK3CB;CRKL;MECOM;AKT2;AKT1;POLK;	51/76	1.8	28
Astro	down	Chronic myeloid leukemia	Differ/Cancer	5.2	RB1;CDKN1B;ARAF;PIK3CD;PIK3CB;CRKL;IKBKB;AKT2;AKT3;	55/76	1.7	25
Astro	down	Pancreatic cancer	Differ/Cancer	5.0	RB1;ARAF;PIK3CD;PIK3CB;CASP9;IKBKB;AKT2;AKT3;AKT1;RA	54/75	1.7	24
In4b	down	Estrogen signaling pathway	Signaling	7.7	HSP90AA1;MAP2K2;HSP90AB1;HSPA6;ITPR2;FOS;HSP90B1;	18/137	7.2	170
Astro	down	Insulin signaling pathway	Signaling	6.6	GSK3B;PYGB;IRS1;PDE3B;INPPL1;ARAF;PRKAG1;PRKAG2;PIK	93/137	1.6	29
Ex3e	down	Oxytocin signaling pathway	Signaling	6.6	RYR1;CAMK2B;GUCY1B1;SRC;NPR2;GNAI3;ITPR2;ADCY3;AD	46/153	2.5	50
Astro	down	mTOR signaling pathway	Signaling	6.1	IRS1;IGF1R;RPS6KA3;STK11;DEPTOR;AKT2;AKT3;SESN2;AKT	100/152	1.5	27
Ex3e	down	MAPK signaling pathway	Signaling	6.1	ATF2;HSPB1;ARRB2;ECSIT;ELK1;CRKL;RPS6KA4;PPP3CA;RPS	70/295	2.0	36
Astro	down	Neurotrophin signaling pathway	Signaling	5.8	YWHAE;GSK3B;PRDM4;IRS1;PIK3CD;PIK3CB;CRKL;IKBKB;RP	81/119	1.6	26
Astro	down	ErbB signaling pathway	Signaling	5.1	GSK3B;CDKN1B;ARAF;PIK3CD;CBLB;PIK3CB;ELK1;CRKL;AKT2	60/85	1.6	24
Micro	down	Estrogen signaling pathway	Signaling	5.1	NCOA2;GABBR2;HSPA6;ITPR1;OPRM1;KRT10;ESR1;HSP90B	18/137	4.4	70
Astro	down	Thyroid hormone signaling pathway	Signaling	5.1	GSK3B;THRB;THRA;SLC2A1;PIK3CD;ATP2A2;PIK3CB;MED16	77/116	1.5	22
Micro	down	Antigen processing and presentation	Immune	7.6	PDIA3;CD74;HSPA5;HSPA6;HLA-B;HLA-C;HLA-A;TAPBP;HLA	17/77	7.4	173
Micro	down	Staphylococcus aureus infection	Immune	6.1	C1QB;C1QA;ITGB2;FPR3;KRT10;ITGAL;FCGR3A;C3AR1;HLA-	14/68	6.9	128
Micro	down	Human T-cell leukemia virus 1 infection	Immune	5.1	NRP1;CRTC2;SPI1;IL15;ITGB2;HLA-B;HLA-C;HLA-A;ITGAL;ET	23/219	3.5	55
Ex3e	down	Cholinergic synapse	Synapse	6.5	CAMK2B;ACHE;CHRM1;CHRNA4;GNAI3;ITPR2;ADCY3;ADCY	38/112	2.8	58
Ex3e	down	Dopaminergic synapse	Synapse	6.0	CAMK2B;ATF2;GSK3A;MAOB;GNAI3;ITPR2;ARRB2;GNAI1;G	40/131	2.5	46
In4b	down	Dopaminergic synapse	Synapse	5.0	GSK3A;ITPR2;FOS;GNG3;CALY;PPP2R1A;GNB2;GNAS;AKT1;	14/131	5.9	94
Ex3e	down	Circadian entrainment	Other	5.9	RYR1;CAMK2B;GUCY1B1;GNAI3;ADCY3;ADCY2;GNAI1;CACI	33/97	2.8	52
Ex3e	down	Cushing syndrome	Other	5.4	RB1;CAMK2B;ATF2;FH;WNT2B;CDKN1B;GNAI3;ITPR2;ADCY	43/155	2.3	38
Micro	down	Viral myocarditis	Other	5.2	CD86;ITGB2;HLA-B;HLA-C;HLA-DRA;HLA-A;ITGAL;ACTB;HLA	12/59	6.8	108
Oligo	down	Thermogenesis	Other	5.1	ATF2;SMARCB1;NDUFA12;NDUFA10;ACTB;ACTG1;LIPE;RPS	125/231	1.4	22
In4b	down	Circadian entrainment	Other	4.9	CACNA1I;GNG3;GUCY1B1;GNB2;GNAS;CALM3;FOS;PLCB1;F	12/97	6.8	105

ALS: Endocytosis, apoptosis, Regulation, Diff, Sign

cell_type	direction	Term	Term	-log10P	Genes	Overlap	OddsRatic	Score
Ex6b	up	Ribosome	TxRegulation	29.4	RPL4;RPL5;RPL30;RPL3;MRPS16;RPL32;RPL34;MRPS11;F	80/153	4.2	306
Ex1	up	Ribosome	TxRegulation	22.1	RPL4;RPL5;RPL30;RPL3;RPL32;RPL34;RPLP1;RPL8;MRPL	65/153	4.1	234
In6b	up	Ribosome	TxRegulation	20.4	RPL4;RPL30;RPL3;MRPS16;RPL32;RPLP1;RPLP0;RPL8;MF	53/153	4.8	250
Ex6a	up	Ribosome	TxRegulation	19.2	RPL4;RPL3;RPL32;RPL10;RPL34;MRPS11;MRPL19;RPLP0	46/153	5.4	259
Ex3e	up	Ribosome	TxRegulation	14.6	RPL5;RPL30;RPL3;RPL32;MRPS11;RPLP1;RPL8;MRPL34;M	64/153	3.1	120
Ex6b	up	Huntington disease	NeuroD	23.3	COX7B;NDUFA12;COX4I1;CLTC;CLTB;CLTA;AP2A1;COX7	83/193	3.4	201
Ex6b	up	Parkinson disease	NeuroD	23.1	COX7B;NDUFA12;COX4I1;PARK7;COX7C;UBE2L3;UCHL1	69/142	3.9	224
Ex6a	up	Parkinson disease	NeuroD	22.9	COX7B;COX4I1;PARK7;COX7C;UBE2L3;UCHL1;UQCRFS1	49/142	6.2	359
In6b	up	Parkinson disease	NeuroD	22.6	COX7B;NDUFA12;COX4I1;PARK7;COX7C;UBE2L3;UCHL1	54/142	5.3	307
Ex1	up	Parkinson disease	NeuroD	20.5	COX7B;NDUFA12;COX4I1;PARK7;COX7C;UBE2L3;UCHL1	60/142	4.1	214
Ex6b	up	Alzheimer disease	NeuroD	19.7	APP;COX7B;NDUFA12;COX4I1;ATP2A2;COX7C;PPP3CA;	72/171	3.4	166
Ex1	up	Alzheimer disease	NeuroD	19.5	APP;COX7B;NDUFA12;COX4I1;ATP2A2;COX7C;PPP3CB;	65/171	3.7	182
Ex6a	up	Huntington disease	NeuroD	19.3	COX7B;COX4I1;CLTC;CLTB;CLTA;AP2A1;COX7C;UQCRFS	52/193	4.8	235
Ex6a	up	Alzheimer disease	NeuroD	19.3	APP;COX7B;COX4I1;ATP2A2;COX7C;PPP3CB;UQCRFS1;M	49/171	5.1	250
In6b	up	Huntington disease	NeuroD	19.1	COX7B;NDUFA12;COX4I1;CLTC;CLTB;CLTA;AP2A1;COX7	58/193	4.2	203
In6b	up	Alzheimer disease	NeuroD	18.9	COX7B;NDUFA12;COX4I1;ATP2A2;COX7C;PPP3CB;UQCI	54/171	4.4	210
Ex1	up	Huntington disease	NeuroD	16.3	COX7B;NDUFA12;COX4I1;CLTC;CLTB;CLTA;AP2A1;COX7	65/193	3.3	137
Ex3e	up	Huntington disease	NeuroD	8.0	COX7B;NDUFA12;COX4I1;NDUFA10;CLTB;CLTA;COX7C;	61/193	2.3	53
Ex3e	up	Parkinson disease	NeuroD	6.3	COX7B;NDUFB10;UQCRB;NDUFB6;NDUFA12;NDUFB11;	46/142	2.4	44
Ex6b	up	Oxidative phosphorylation	Energy	22.8	COX7B;NDUFA12;COX4I1;COX7C;UQCRFS1;ATP6V1E1;A	66/133	4.0	225
In6b	up	Oxidative phosphorylation	Energy	22.7	COX7B;NDUFA12;COX4I1;COX7C;UQCRFS1;ATP6V1E1;A	52/133	5.5	313
Ex6a	up	Oxidative phosphorylation	Energy	22.7	ATP6V1A;COX7B;NDUFB10;UQCRB;NDUFB11;NDUFB5;	47/133	6.3	360
Ex1	up	Oxidative phosphorylation	Energy	22.2	COX7B;NDUFA12;COX4I1;COX7C;UQCRFS1;ATP6V1E1;A	60/133	4.4	247
Ex6b	up	Thermogenesis	Energy	15.9	ATF2;COX7B;NDUFA12;COX4I1;COX7C;ACTB;ACTG1;DP	80/231	2.8	112
In6b	up	Thermogenesis	Energy	14.5	ATF2;COX7B;NDUFA12;COX4I1;COX7C;ACTB;ACTG1;DP	57/231	3.4	129
Ex1	up	Thermogenesis	Energy	12.6	ATF2;COX7B;NDUFA12;COX4I1;COX7C;ACTB;ACTG1;DP	66/231	2.8	92
Ex1	up	Non-alcoholic fatty liver disease (NAFLD)	Energy	11.1	GSK3A;COX7B;NDUFB7;NDUFB10;UQCRB;IRS1;NDUFA1	48/149	3.1	92
Ex6a	up	Thermogenesis	Energy	9.8	ATF2;COX7B;NDUFB10;UQCRB;NDUFB11;NDUFB5;COX	43/231	3.3	88
Ex6a	up	Non-alcoholic fatty liver disease (NAFLD)	Energy	9.5	GSK3A;COX7B;NDUFB10;UQCRB;IRS1;NDUFB11;NDUFB	33/149	3.9	101
Ex6b	up	Non-alcoholic fatty liver disease (NAFLD)	Energy	8.5	GSK3A;COX7B;NDUFA12;COX4I1;IRS2;COX7C;UQCRFS1	49/149	2.6	61
Ex3e	up	Oxidative phosphorylation	Energy	8.0	COX7B;NDUFB10;UQCRB;NDUFB6;NDUFA12;NDUFB11;	48/133	2.6	62
Ex3e	up	Thermogenesis	Energy	7.6	ATF2;COX7B;KDM1A;NDUFA12;COX4I1;NDUFA10;PRKA	68/231	2.2	47
In6b	up	Non-alcoholic fatty liver disease (NAFLD)	Energy	7.4	COX7B;NDUFB7;NDUFB10;UQCRB;IRS1;NDUFA12;NDUI	34/149	3.2	66
In6b	up	Synaptic vesicle cycle	Synapse	10.1	ATP6V1A;SNAP25;RAB3A;STXBP1;CLTC;CLTB;CLTA;AP2/	27/78	4.8	131
Ex6a	up	Synaptic vesicle cycle	Synapse	8.2	ATP6V1A;SNAP25;ATP6V1G1;ATP6V0B;RAB3A;STXBP1;(	22/78	5.0	113
In8	up	Long-term potentiation	Synapse	6.8	GRIA2;GRIN2A;CREBBP;PRKCB;CAMK4;EP300;ITPR2;CA	Nov-67	13.3	283
In6b	up	GABAergic synapse	Synapse	6.5	GABARAPL2;GABBR2;GABRA1;SLC38A1;GABARAPL1;GA	24/89	3.8	70
In6b	up	Dopaminergic synapse	Synapse	6.0	ATF2;ITPR1;CACNA1C;GNAI1;PPP1CC;GRIN2A;PPP3CB;(	29/131	3.1	53
Ex6b	up	Endocytosis	Endo	7.2	WASHC4;VPS29;ARF1;WIPF2;WIPF3;CLTC;ARPC5L;CLTB	65/244	2.1	43
Ex6b	up	Protein processing in endoplasmic reticulum	Endo	6.9	ERO1B;HSP90AB1;UBE2D3;UBE2D1;HERPUD1;CAPN1;S	49/165	2.4	46
In6b	up	Vibrio cholerae infection	Immune	6.8	ATP6V1A;ATP6V1G1;ATP6V0B;ATP6AP1;ACTB;ACTG1;S	18/50	5.0	97
Microglia	down	Staphylococcus aureus infection	Immune	6.3	C1QB;ITGAM;FPR1;FPR2;ITGAL;ICAM1;C3;FCGR3A;FCG	16/68	6.6	133
Ex1	up	Vibrio cholerae infection	Immune	6.0	ATP6V1A;ARF1;ATP6V1G1;ATP6V0B;ATP6AP1;ADCY3;A	20/50	3.9	67
Ex1	up	Mitophagy	Autophagy	6.5	GABARAPL2;BNIP3L;FIS1;GABARAPL1;JUN;BNIP3;AMBR	24/65	3.6	67
Ex6b	up	Proteasome	Autophagy	6.2	PSMD12;PSMD11;POMP;PSMA7;PSMD8;PSMB6;PSMA	21/45	3.7	66
Oligo	down	Ubiquitin mediated proteolysis	Autophagy	6.0	DET1;UBE2D4;UBE3C;UBE2D2;KEAP1;UBE2D1;UBE3A;C	88/137	1.6	32
Astro	up	Mineral absorption	Other	8.2	MT1M;MT1X;ATP1B3;SLC30A1;ATP1A2;ATP1A1;ATP1B	13/51	12.5	307
In6b	up	Cardiac muscle contraction	Other	6.4	COX7B;UQCRB;COX4I1;ATP2A2;CACNA2D2;ATP1B3;CO	22/78	3.9	71
Ex1	up	Retrograde endocannabinoid signaling	Signaling	7.1	GABRB3;NDUFB7;NDUFB10;NDUFA12;NDUFB11;NDUFI	41/148	2.7	54
In6b	up	Retrograde endocannabinoid signaling	Signaling	5.9	NDUFB7;NDUFB10;NDUFA12;NDUFB11;NDUFB5;NDUFI	31/148	2.9	49

FTD: Regulation, NeuroD, Energy, Synapse

Pineda, *et al*, In Submission

Huntington's: Single-cell changes in striatum

- 6 human post-mortem striata: caudate and putamen (12 samples)
- 10 mouse samples: HTT repeat length allelic series: 20,50,111,170,175

- iSPNs (indirect pathway spiny projection neurons) merge with dSPNs
- Human and mouse changes confirm each other

Lee, Fenster, *et al*, Neuron 2020

Cellular state characterization of microglia + AD-associated changes

Microglia subtypes: Inflammatory (AD) vs. Synaptic (Sz)

AD DEGs: Homeostatis, age-related, inflammatory, DAMs

AD: Expansion of DAM (MG4), Immuno-metabolism (MG5)

Up: matrix adh, migration, metab, transp., Aβ clearance, lipid transport
Down: chaperones, vesicle transport, cell communication, cytokine resp.

Vasculature/BBB characterization, zonation, vc cells, DEGs

Brain vascular atlas: *ex-vivo* enrichmenetand + *in silico* sorting

- **Validate: human-specific fibroblast subtype**
- **Validate: vasculature-coupled vcNeurons, vcAstro**

Zonation gradients for endothelial, pericyte, smooth muscle cells

HD-associated changes in vasculature

Na Sun
Francisco Garcia
Myriam Heiman

Vasculature-dysregulated AD genes in AD GWAS + sub-threshold loci

Vasculature composition differs between brains regions

• Vasculature DEGs help explain 70 GWAS loci

AD DEGs: cEndo down: cytokine stimulus, zinc ion
Upstream regulators: FOXD1, KLF4, HIF1A, CEBFD

Increased interactions between capEndo, Pericytes and ExcN, diverse cell types

Na Sun
Li-Huei Tsai

RORA (lipid sensing receptor, infl. reg.)
- **Down** in capEndo, vEndo, peric, fibr.

ABCA1 (cholesterol efflux, Aβ depositn/clearnce)
- **Up** in pericytes: cholesterol efflux
- **Down** in fibroblast: Aβ dep. & clearance

SCARB1 (HDL sign., SMC-Aβ interaction)
- **Up** in capEndo and pericytes

IL6R (inflammatory response)
- **Up** in capEndo

MYRIP (insulin exocytosis)
- **Down** in capEndo

PFDN1 (prefoldin complex subunit)
- **Down** in SMCs

- **Link non-coding variants to their target genes**
- **Corr.** (EpiMap), **eQTLs** (GTEx), **sc-eQTLs** (ROSMAP), **Hi-C**

scATAC-scRNA Epigenome-Transcriptome linking & integration

scATAC cell-type-specific marker accessibility

scATAC subtypes of ExcN and InhN

Micr/ExcN/Astr-specific peaks

Late AD loss of ExcN+InhN (esp. male)

scRNA-scATAC coordinated changes in AD

scRNA-scATAC alignment

AD differential peaks enriched for AD, tauopathies, immune functions, and other neuronal disorders

TFs upstream of AD-assoc. peaks reveal candidate master regulators

Xushen Xiong
Carles Boix
Shahin Mohammadi
Sebastian Pineda
Kiki Galani
Li-Lun Ho

Cell-type-specific GWAS enrichments for scATAC-seq peaks

scATAC-seq: cell-type-specific GWAS enrichments

SPI motifs repeatedly disrupted by AD GWAS SNPs in microglial scATAC peaks
➔ Infer master regulators for each trait in each cell type

Xushen Xiong
Carles Boix
Kiki Galani
Li-Lun Ho
Li-Huei Tsai

Mosaicism: Somatic mutation burden in disease from scRNA

- Variant calling + tx: single-cell genotype-phenotype
- Somatic burden: senescent cells (betw ExN+Olig)

- Mutational burden⇔Cognitive decline
- Incr. missense + high-impact variants

- Mutational burden concentrates in specific genes+pathways
- Energy production, signaling, lipid metabolism, transport
- ExcN and Olig show highest burden (specifically in clusters near senescent cells)

- AlzD: concentrated mutational burden
- Including known AD GWAS+Famil. genes

Maria Kousi
Carles Boix

scATAC evidence of epigenome erosion in late AD across cell types

- Early AD: active states are 'open', repressed are 'closed'
- Late AD: reversal, closing of promoters/enhancers, opening of heterochromatin/lamina/repressed regions

- scATAC: senescent/de-identified cluster
- Global regulatory shutdown, epigenome erosion
- Cells come specifically from <u>late AD</u> individuals

Xushen Xiong
Carles Boix

- Across all cell types
- Specific to late AD

- Eroded cells found in both non-AD, early-AD, and late-AD
- But dramatic shift in the distribution, especially for late-AD
- <u>de-ID cells, and de-ID oligos/neurons particularly pronounced</u>

- Broad impact on epigenome, spanning Mbs
- Some cells show clear reversal, others just erosion

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

4. Stepping back: Basic Tasks of Single-cell Analysis

single-cell transcriptome analysis

github.com/theislab/scanpy

pypi v0.1.8 docs passing build passing

scanpy

Wolf et al, *Genome Biology* 2018

Luecken & T, *Mol Sys Bio* 2019

Differential expression analysis

Association analysis plays an important role in scRNA-seq data analysis

1. Generating residuals for clustering & downstream analysis

2. Identification of cell type marker genes

3. Identification of disease-associated genes

4. co-expression analysis

Genes over-expressed in AD Early: Modifications specific to the cell type Late: Global changes

Genes under-expressed in AD: Cell-type specific signatures in early and late AD

Multi-level structure is ubiquitous in scRNA-seq data

- Example of large-scale multi-subject droplet-based snRNA-seq data sets in the brain from ROSMAP involving multiple subjects
  - ~100k 24-subject snRNA-seq data in the frontal cortex (Habib, Avraham-Davidi, Basu et al. 2017)
  - ~80k 48-subject snRNA-seq data in the BA10 frontal cortex (Mathys et al. 2019)
  - ~66k 32-subject snRNA-seq data in the prefrontal cortex (Zhou et al. 2020)
- An upcoming snRNA-seq dataset in PFC including >400 subjects (Park, ASHG 2021)

(Mathys et al., 2019)

Existing methods for single-cell data do not account for hierarchical structure

- Previous statistical regression models
  - 1. Directly modeling the count.
    - Generalized Linear Models: Poisson regression, Negative binomial regression, etc.
    - Zero-inflated models: Zero-inflated Poisson, Zero-inflated negative binomial, etc.
    - Advantage: Accurately capturing the data generating process, overdispersions, and excessive zeros.
    - Disadvantage: computationally more expensive compared to a linear model
  - 2. Modeling a transformed variable
    - Convert the count data into a continuous variable using e.g., logarithm. (limma, voom, MAST)
    - Advantage: Fast.
    - Disadvantage: Highly depending on how to transform the data and how to deal with zeros, less statistically powerful, not modeling the data generating process
- However, neither accommodates hierarchical structure.

Classify variables into the cell level or the subject level

Subject level

- Age, sex, disease/treatment, genotype, batch, etc
- Both pseudo-bulk methods and hierarchical models work
- Hierarchical models are more accurate in the sense that it basically gives different weights to cells

Cell level

- Subpopulation, cell cycle, gene expression, mitochondrial gene abundance, etc
- The interaction term between subject-level and cell-level should a be treated as a cell- level variable

$$y_{ij} = x_{ij}\beta + z_i\gamma + x_{ij}z_i\delta + \varepsilon_{ij}$$

Cell-level variable

Subject-level variable

Interaction term

Importance of adjusting for the hierarchical structure using a mixed model

- In testing associations with a cell-level variable x_{ij}, ignoring the subject-level random effects φ_i would lead to a biased estimated β, which captures the subject-level confounder effects of φ_i
- The subject-level random effects φ_i also need to be included in the model when testing a subject-level variable z_i. Otherwise, it would lead to inflated p-values of γ. See (Landeghem et al., 2005; Moerbeek, 2004)

$$y_{ij} = x_{ij}\beta + z_i\gamma + \boxed{\varphi_i} + \varepsilon_{ij}$$

$$i \in \{1, \dots, m\}, number\ of\ subjects$$
$$j \in \{1, \dots, n_i\}, number\ of\ cells\ in\ subject\ i$$

Cell-level variable

Subject-level variable

Subject-level random effects capture all unknow variables (Cannot be ignored)

Computational challenges for differential analysis using large-scale multi-subject single-cell data

- Current methods designed for bulk data are not sufficient to deal with scRNA-seq data because of
  - Very large sample size
    - Requires an efficient method
  - High-sparsity
    - Poses challenges for data transformation
  - Hierarchical structure
    - Requires a hierarchical model such as mixed models

Difference between Bulk and scRNA-seq		
	Bulk	scRNA-seq
Sample size	Small	Large
Count matrix	Dense	Sparse
Hierarchical	No	Yes
Overdispersion	Subject	Subject and cell
Model	Negative binomial model	Negative binomial mixed model

Computational challenge in fitting NBMMs

- More challenging than the mixed models used for genome-wide association analysis
- Current methods for fitting an NBMM are prohibitively slow (A challenge for decades)
  - Frequentist
    - Laplace approximation
    - Adaptive Gaussian quadrature
    - Penalized quasi-likelihood
    - H-likelihood
  - Bayesian
    - INLA
    - MCMC
    - Variational Bayes

$$y_{ij} \sim NB\left(\mu_{ij} = \pi_{ij} \exp\left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\omega_i)\right), \phi\right)$$

The challenge lies in the difficulty in integrating out ω_i in the following marginal likelihood. The integral is not in closed form, which requires a two-layer optimization procedure which converges slowly.

$$L_i(\boldsymbol{\beta}, \sigma^2, \phi) \propto \int_0^{+\infty} \prod_{\boldsymbol{j}} \Big(\Gamma(y_{ij} + \phi) \exp\left(y_{ij}\left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\pi_{ij}) + \log(\omega_i)\right)\right) \left(\phi + \exp\left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\pi_{ij}) + \log(\omega_i)\right)\right)^{-(y_{ij}+\phi)} \Big) f(\omega_i) d\omega_i$$

- Can we find a fast method for this model?

NEBULA: a fast NBMM using a large-sample approximation

- NEBULA: **NE**gative **B**inomial mixed model **U**sing a **L**arge-sample **A**pproximation
- NEBULA is based on a negative binomial gamma mixed model (NBGMM)

$$y_{ij} \sim NB\left(\mu_{ij} = \pi_{ij} \exp\left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\omega_i)\right), \phi\right) \quad \omega_i \sim \text{Gamma}(\alpha, \lambda)$$

- Using an approximated marginal likelihood by analytically integrating out the random effects instead of Laplace approximation by leveraging the large number of cells per subject (NEBULA-LN)
- Using an h-likelihood method in situations the above large-sample approximation is not sufficiently accurate (NEBULA-HL)
- Using higher-order Laplace approximation for low-expression genes

Account for the hierarchical structure using Negative Binomial Mixed Models (NBMMs)

- We want to test or detect which genes in $Y \in \mathbb{Z}_{+}^{n \times m}$ are associated with the variable $X \in \mathbb{R}^{n \times 1}$ of interest (n: number of cells, m: number of genes).
- NBMMs Decompose the total overdispersion into the cell-level and the subject-level overdispersions

Raw count — Scaling factor — Subject-level random effects — Cell-level overdispersion

$$y_{ij} \sim NB\left(\mu_{ij} = \pi_{ij} \exp\left(\boldsymbol{x}_{ij}{}^{T}\boldsymbol{\beta} + \log(\omega_i)\right), \phi\right)$$

$$\omega_i \sim \text{lognormal}(0, \sigma^2)$$ ← Subject-level overdispersion

Negative binomial lognormal mixed model *(NBLMM)*

or

$$\omega_i \sim \text{Gamma}(\alpha, \lambda) \qquad \alpha = \lambda = \sigma^2$$

Negative binomial gamma mixed model *(NBGMM)*

$$\alpha = \frac{1}{\exp(\sigma^2) - 1} \qquad \lambda = \frac{1}{\exp(\sigma^2/2)\,(\exp(\sigma^2) - 1)}$$

The NEBULA R package

- The R package can be installed from Github or R- forge
- The vignettes include detailed usage and examples
- Any feedback is welcome and please send them to liang.he@duke.edu
- Next update is scheduled in December (Output Pearson residuals)

https://github.com/lhe17/nebula

https://github.com/lhe17/nebula

README.md

- Installation
  - Most recent version
- Functions
- Basic usage
  - Example
- Specifying scaling factors
  - Example
- Selection between NEBULA-LN and NEBULA-HL
- Filtering low-expressed genes
- Checking convergence for the summary statistics
- Using other mixed models
  - Example
- Testing contrasts

NEBULA v1.1.8

Overview

The R package, *nebula*, provides fast algorithms for fitting negative binomial and Poisson mixed models for analyzing large-scale multi-subject single-cell data. The package *nebula* accounts for the hierarchical structure of the data by decomposing the total overdispersion into between-subject and within-subject components using a negative binomial mixed model (NBMM). The package nebula can be used for e.g., identifying marker genes, testing treatment effects, detecting genes with differentail expression, and performing cell-level co-expression analysis.

More details can be found in the manuscript "NEBULA: a fast negative binomial mixed model for differential expression and co-expression analyses of large-scale multi-subject single-cell data" (https://www.nature.com/articles/s42003-021-02146-6).

Installation

Most recent version

To install the lastest version from github:

Trajectories through cell space

Cell-cycle phase prediction

Scialdone *et al.*, 2015

Trajectory inference

Bargaje *et al*, PNAS 2017

- Identify key branching points in development/disease
- Regulatory circuits that drive these transitions

Trajectory inference methods

TSCAN pseudotime reconstruction with monocle

Ji *et al*, NAR 2016

- Start with dimension reduction
- Build a graph among cells/inferred cell types
  - Typically underlying structure is based on minimum spanning tree (**MST**) or k-nearest neighborhood (**kNN**) graph.
- Either infer a linear (pseudo-time) ordering, or identify branching points

Overview of trajectory identification methods

Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	–	+	–	–
Scalability w.r.t. cells	–	–	±	±	–	±	+	+	–	±
Scalability w.r.t. genes	+	+	+	+	–	+	±	±	±	+
Code and documentation	–	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	–	±	–	+	+	+

First Author	Marco	Bendall	Setty	Welch	Matsumoto	Shin	Chen	Ji	Trapnell	Marco
Last Author	GC Yuan	Dana Pe'er	Dana Pe'er	Hartemink, Prins	Kiryu	Hongjun Song	Poidinger	Ji	Rinn	GC Yuan
Journal	PNAS	Cell	Nature Biotechnology	Genome Biology	BMC Bioinformatics	Cell Stem Cell	Nature Communications	NAR	Nature Biotechnology	PNAS
Year	2014	2014	2016	2016	2016	2015	2016	2016	2014	2014

Trajectory identification: meta-method view

hemberg-lab.github.io/scRNA.seq.course

Dataset completion & missing data imputation

Spatial reconstruction of single-cell gene expression

Satija Nature Biotech 2015

Missing value imputation with MAGIC (Markov Affinity-based Graph Imputation of Cells)

Random walk on cell-cell similarity graph

- uses neighborhood-based Markov-affinity matrix
- shares weight information across cells
- generate an imputed count matrix

Imputation reveals gene-gene correlation patterns

A: model-based imputation

bayNorm	[47]	Binomial model, empirical Bayes prior
BISCUIT	[48]	Gaussian model of log counts, cell- and cluster-specific parameters
CIDR	[49]	Decreasing logistic model (DO), non-linear least-squares regression (imp)
SAVER	[50]	NB model, Poisson LASSO regression prior
ScImpute	[51]	Mixture model (DO), non-negative least squares regression (imp)
scRecover	[52]	ZINB model (DO identification only)
VIPER	[53]	Sparse non-negative regression model

B: data smoothing

DrImpute	[54]	*k*-means clustering of PCs of correlation matrix
knn-smooth	[55]	*k*-nearest neighbor smoothing
LSImpute	[56]	Locality sensitive imputation
MAGIC	[57]	Diffusion across nearest neighbor graph
netSmooth	[58]	Diffusion across PPI network

C: data reconstruction, matrix factorizatio

ALRA	[59]	SVD with adaptive thresholding
ENHANCE	[60]	Denoising PCA with aggregation step
scRMD	[61]	Robust matrix decomposition
consensus NMF	[62]	Meta-analysis approach to NMF
f-scLVM	[63]	Sparse Bayesian latent variable model
GPLVM	[64]	Gaussian process latent variable model
pCMF	[65]	Probab. count matrix factorization with Poisson model
scCoGAPS	[66]	Extension of NMF
SDA	[67]	Sparse decomposition of arrays (Bayesian)
ZIFA	[68]	ZI factor analysis
ZINB-WaVE	[69]	ZINB factor model

C: data reconstruction, machine learning

AutoImpute	[70]	AE, no error back-propagation for zero counts
BERMUDA	[71]	AE for cluster batch correction (MMD and MSE loss function)
DeepImpute	[72]	AE, parallelized on gene subsets
DCA	[73]	Deep count AE (ZINB / NB model)
DUSC / DAWN	[74]	Denoising AE (PCA determines hidden layer size)
EnImpute	[75]	Ensemble learning consensus of other tools
Expression Saliency	[76]	AE (Poisson negative log-likelihood loss function)
LATE	[77]	Non-zero value AE (MSE loss function)
Lin_DAE	[78]	Denoising AE (imputation across *k*-nearest neighbor genes)
SAUCIE	[79]	AE (MMD loss function)
scScope	[80]	Iterative AE
scVAE	[81]	Gaussian-mixture VAE (NB / ZINB / ZIP model)
scVI	[82]	VAE (ZINB model)
scvis	[83]	VAE (objective function based on latent variable model and t-SNE)
VASC	[84]	VAE (denoising layer; ZI layer, double-exponential and Gumbel distribution)
Zhang_VAE	[85]	VAE (MMD loss function)

T: using external information

ADImpute	[86]	Gene regulatory network information
netSmooth	[58]	PPI network information
SAVER-X	[87]	Transfer learning with atlas-type resources
SCRABBLE	[88]	Matched bulk RNA-seq data
TRANSLATE	[77]	Transfer learning with atlas-type resources
URSM	[89]	Matched bulk RNA-seq data

47. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, et al. bayNorm: Bayesian gene expression recovery, imputation and normalisation for single cell RNA-sequencing data. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/384586v2.abstract.
48. Azizi E, Prabhakaran S, Carr A, Pe'er D. Bayesian inference for single-cell clustering and imputing. Genomics Comput Biol. 2017;3(1):46. https://doi.org/10.18547/gcb.2017.vol3.iss1.e46. Accessed 27 Mar 2019.
49. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18(1):59. https://doi.org/10.1186/s13059-017-1188-0. Accessed 27 Mar 2019.
50. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15(7):539. https://doi.org/10.1038/s41592-018-0033-z. Accessed 27 Mar 2019.
51. Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell RNA-seq data. Nat Commun. 2018;9(1):997.
52. Miao Z, Li J, Zhang X. scRecover: discriminating true and false zeros in single-cell RNA-seq data for imputation. bioRxiv. 2019665323. https://doi.org/10.1101/665323. Accessed 15 Oct 2019.
53. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. Genome Biol. 2018;19(1):196.
54. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018;19(1):220. https://doi.org/10.1186/s12859-018-2226-y. Accessed 27 Mar 2019.
55. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. bioRxiv. 2018217737. https://doi.org/10.1101/217737. Accessed 15 Oct 2019.
56. Moussa M, Măndoiu II. Locality sensitive imputation for single cell RNA-Seq data. J Comput Biol. 2019. https://doi.org/10.1089/cmb.2018.0236. Accessed 27 July 2019.
57. Dijk DV, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bierie B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–72927. https://doi.org/10.1016/j.cell.2018.05.061. Accessed 27 Mar 2019.
58. Jonathan Ronen AA. netsmooth: network-smoothing based imputation for single cell RNA-seq. F1000Res. 2018;7:. https://github.com/BIMSBbioinfo/netSmooth.
59. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/397588v1.abstract.
60. Wagner F, Barkley D, Yanai I. Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis. bioRxiv. 2019655365. URL https://doi.org/10.1101/655365. Accessed 15 Nov 2019.
61. Chen C, Wu C, Wu L, Wang Y, Deng M, Xi R. scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. bioRxiv. 2018459404. https://doi.org/10.1101/459404. Accessed 15 Oct 2019.
62. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, Sabeti PC. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife. 2019;8:43803.
63. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. 2017;18(1):212. https://doi.org/10.1186/s13059-017-1334-8.
64. Verma A, Engelhardt BE. A robust nonlinear low-dimensional manifold for single cell RNA-seq data. bioRxiv. 2018443044. https://doi.org/10.1101/443044. Accessed 15 Nov 2019.
65. Durif G, Modolo L, Mold JE, Lambert-Lacroix S, Picard F. Probabilistic count matrix factorization for single cell expression data analysis. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz177.
66. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, Liu S, Qian J, Colantuoni C, Blackshaw S, Goff LA, Fertig EJ. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. Cell Syst. 2019;8(5):395–4118. https://doi.org/10.1016/j.cels.2019.04.004.
67. Jung M, Wells D, Rusch J, Ahmad S, Marchini J, Myers SR, Conrad DF. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. eLife. 2019;8:. URL https://doi.org/10.7554/eLife.43966.
68. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1):241. https://doi.org/10.1186/s13059-015-0805-z. Accessed 27 Mar 2019.
69. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Commun. 2018;9(1):284. https://doi.org/10.1038/s41467-017-02554-5.
70. Talwar D, Mongia A, Sengupta D, Majumdar A. AutoImpute: autoencoder based imputation of single-cell RNA-seq data. Sci Rep. 2018;8(1):16329. https://doi.org/10.1038/s41598-018-34688-x.
71. Wang T, Johnson TS, Shao W, Lu Z, Helm BR, Zhang J, Huang K. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol. 2019;20(1):165. https://doi.org/10.1186/s13059-019-1764-6. Accessed 15 Nov 2019.
72. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire L. DeepImpute: an accurate, fast and scalable deep neural network method to impute single-cell RNA-Seq data. bioRxiv. 2018. https://www.biorxiv.org/content/10.1101/353607v1.abstract.
73. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390. https://doi.org/10.1038/s41467-018-07931-2. Accessed 27 Mar 2019.
74. Srinivasan S, Johnson NT, Korkin D. A hybrid deep clustering approach for robust cell type profiling using single-cell RNA-seq data. bioRxiv. 2019. https://www.biorxiv.org/content/10.1101/511626v1.abstract.
75. Zhang X-F, Ou-Yang L, Yang S, Zhao X-M, Hu X, Yan H. EnImpute: imputing dropout events in single cell RNA sequencing data via ensemble learning. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz435.
76. Kinalis S, Nielsen FC, Winther O, Bagger FO. Deconvolution of autoencoders to learn biological regulatory modules from single cell mRNA sequencing data. BMC Bioinformatics. 2019;20(1):379. https://doi.org/10.1186/s12859-019-2952-9.
77. Badsha MB, Li R, Liu B, Li YI, Xian M, Banovich NE, Fu AQ. Imputation of single-cell gene expression with an autoencoder neural network. bioRxiv. 2018504977. https://doi.org/10.1101/504977. Accessed 15 Oct 2019.
78. Lin C, Jain S, Kim H, Bar-Joseph Z. Using neural networks for reducing the dimensions of single-cell RNA-Seq data. Nucleic Acids Res. 2017;45(17):156.
79. Amodio M, Dijk DV, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, Desai A, Ravi V, Kumar P, Montgomery R, Wolf G, Krishnaswamy S. Exploring single-cell data with deep multitasking neural networks. bioRxiv. 2019237065. https://doi.org/10.1101/237065. Accessed 15 Oct 2019.
80. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat Methods. 2019. https://doi.org/10.1038/s41592-019-0353-7.
81. Grønbech CH, Vording MF, Timshel P, Sønderby CK, Pers TH, Winther O. scVAE: Variational auto-encoders for single-cell gene expression data. bioRxiv. 2019318295. https://doi.org/10.1101/318295. Accessed 15 Oct 2019.
82. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2.
83. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018;9(1):2002.
84. Wang D, Gu J. VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. Genomics Proteomics Bioinforma. 2018;16(5):320–31.
85. Zhang C. Single-cell data analysis using mmd variational autoencoder for a more informative latent representation. bioRxiv. 2019613414. https://doi.org/10.1101/613414. Accessed 15 Oct 2019.
86. Leote AC, Wu X, Beyer A. Network-based imputation of dropouts in single-cell RNA sequencing data. bioRxiv. 2019611517. URL https://doi.org/10.1101/611517. Accessed 23 Apr 2019.
87. Wang J, Agarwal D, Huang M, Hu G, Zhou Z, Ye C, Zhang NR. Data denoising with transfer learning in single-cell transcriptomics. Nat Methods. 2019;16(9):875–8. https://doi.org/10.1038/s41592-019-0537-1. Accessed 15 Oct 2019.
88. Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. Genome Biol. 2019;20(1):88. https://doi.org/10.1186/s13059-019-1681-8.
89. Zhu L, Lei J, Devlin B, Roeder K. A unified statistical framework for single cell and bulk RNA sequencing data. Ann Appl Stat. 2018;12(1):609–32. https://doi.org/10.1214/17-AOAS1110. Accessed 15 Nov 2019.

Cell Projected Phenotypes

Beyond Clinical Labels: Comparing cell transcriptional states to uncover key disease genes and pathways

- Not every cell in an AD brain presents disease phenotypes
- Many aged non-AD patients have AD-like cells
- New methods to capture this heterogeneity

Cell Neighborhood Associations from KNN graphs

- Construct KNN graph in scRNA transcriptional embedding space
- Cells derived from multiple donors, some with AD, some with no AD
- Define ‘neighborhoods’ as a set of ‘index cells’ and their neighbors
- Statistical enrichment testing for phenotype enrichments in neighborhoods

Dann et. al, Nature Biotechnology 2022

Cell Projected Phenotypes: pairing cells with clinical labels

Bouland et al, in preparation

Cell Projected Phenotypes in Alzheimer's Disease

3.4M cells from 599 donors, postmortem PFC scRNA profiles

Donor heterogeneity - cell type specific molecular pathologies

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

5. Deep Learning methods for scRNA-seq

MMD-ResNet: Autoencoder for batch correction

Autoencoder ResNet arch.

MMD (Gretton *et al.*, 2006, 2012) is a measure for distance between two probability distributions p, q. It is defined with respect to a function class $\mathcal{F}$ by

$$\mathrm{MMD}(\mathcal{F}, p, q) \equiv \sup_{f \in \mathcal{F}} (\mathbb{E}_{x\sim p} f(x) - \mathbb{E}_{x\sim q} f(x)).$$

When $\mathcal{F}$ is a reproducing kernel Hilbert space with kernel k, the MMD can be written as the distance between the mean embeddings of p and q

$$\mathrm{MMD}^2(\mathcal{F}, p, q) = ||\mu_p - \mu_q||^2_{\mathcal{F}}, \quad (1)$$

where $\mu_p(t) = \mathbb{E}_{x\sim p} k(x, t)$. Equation (1) can be written as

$$\mathrm{MMD}^2(\mathcal{F}, p, q) = \mathbb{E}_{x,x'\sim p} k(x, x') - 2\mathbb{E}_{x\sim p, y\sim q} k(x, y) + \mathbb{E}_{y,y'\sim q} k(y, y'), \quad (2)$$

where x and x' are independent, and so are y and y'. Importantly, if k is a universal kernel, then $\mathrm{MMD}(\mathcal{F}, p, q) = 0$ iff $p = q$. In practice, the distributions p, q are unknown, and instead we are given observations $X = \{x_1, \ldots x_n\}, Y = \{y_1, \ldots y_m\}$, so that the (biased) sample version of (2) becomes

$$\mathrm{MMD}^2(\mathcal{F}, X, Y) = \frac{1}{n^2} \sum_{x_i, x_j \in X} k(x_i, x_j) - \frac{2}{nm} \sum_{x_i \in X, y_j \in Y} k(x_i, y_j) + \frac{1}{m^2} \sum_{y_i, y_j \in Y} k(y_i, y_j).$$

Maximum Mean Discrepancy (MMD) loss function

Fig. 5. Histograms of the 25 *P*-values of Kolmogorov-Smirnov tests, comparing the distributions of the calibrated data with the target distribution of each of the 25 markers

MMD-ResNet outperforms PCA, Combat, and

$$L(w) = \sqrt{\mathrm{MMD}^2(\{\widehat{\psi}(x_1), \ldots \widehat{\psi}(x_n)\}, \{y_1, \ldots y_m\})},$$

Train ResNet with loss MMD score function

t-SNE plots before (left) and after (right) calibration

Removal of batch effects using distribution-matching residual networks

Uri Shaham[1,†], Kelly P. Stanton[2,3,†], Jun Zhao[3], Huamin Li[4], Khadir Raddassi[5], Ruth Montgomery[6] and Yuval Kluger[2,3,4,*]

Shaham et al., Bioinf, 2017

DESC: Deep embedding for cell-type-specific batch correction

DESC (Deep Embedding for single-cell clustering):

- Stacked auto-encoder learns cluster-specific gene expression representation and cluster assignments for scRNA-seq data clustering
- Initialize clustering obtained from autoencoder
- Learn non-linear mapping from original space to a low-dimensional space
- iteratively optimize clustering objective function
  - Move each cell to nearest cluster
  - balance biological and technical differences between clusters
  - reduce influence of batch effect
- Enables soft clustering by assigning cluster-specific probabilities to each cell
- Facilitates clustering of cells with high confidence

Deep learning enables accurate clustering and batch effect removal in single-cell RNA-seq analysis

Xiangjie Li, Yafei Lyu, Jihwan Park, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Gang Hu, Mingyao Li

doi: https://doi.org/10.1101/530378

DESC avoids cluster-specific batch effects found in other methods

Iterative approach progressively removes cluster-specific batch effects

Rand Index (RI) = measure of the similarity between two data clusterings

ARI = Adjusted Rand Index, adjusted for the chance grouping of elements

AutoImpute: Overcomplete autoencoder for filling in zeros

Autoimpute captures more non-zero values for highly-expressed genes

Iterative approach progressively removes cluster-specific batch effects

AutoImpute

Filter raw gene expression data for bad genes
(normalize by library size, prune by gene-selection, log transform)
Feed processed matrix to AutoImpute model

- learn expression data representation
- reconstruct imputed matrix

Use overcomplete autoencoders to capture distribution of sparse gene expression data, and regenerate complete version of it

- Feeding sparse gene expression matrix as input to autoencoder
- train it to learn the encoder and decoder functions that best regenerate imputed expression with no dropouts
- back-propagating errors only for non-zero counts in sparse matrix

Training and Hyper-parameter Selection. The autoencoder network consists of a fully-connected multi-layer perceptron (MLP), with three layers: input, hidden and the output layer. It is trained using gradient descent with gradients computed by back-propagation to reach the minimum of the cost function (equation 8). RMSProp Optimizer was used to adjust the learning rate, such that, we avoid getting stuck at local minima and reach the minimum of the cost function faster. Both E - encoder matrix and D - decoder matrix were initialized from a random normal distribution.

The hyper-parameter selection was done after doing an extensive grid search on the following hyper-parameters:

- λ - the regularization coefficient, to control the contribution of the regularization term in the loss or cost function.
- Size of the hidden layer or latent space dimensionality.
- Initial value of learning rate.
- Threshold value - We stop the gradient descent after the change in loss function value in consecutive iterations is less than the threshold value, implying convergence.

The best results were observed on the hyper-parameter choices shown in Table 2.

AutoImpute: Autoencoder based imputation of single-cell RNA-seq data

Divyanshu Talwar[1], Aanchal Mongia[1], Debarka Sengupta[1,3] & Angshul Majumdar[2]

scVI: Use NN to estimate params in variational inference

scVI retains biological signal in diverse datasets

scVI enables differential expression analysis

scVI: Learn non-linear embedding of cells for multiple analysis tasks
NN=Neural networks used to compute embedding and expr. distribution
f_w,f_h: functional representations NN5,6 to capture parameters of Gaussians

Modeled observed expression x_{ng} (gene g, cell n) as sample
Drawn from zero-inflated negative binomial (ZINB) distribution
Conditioned on the batch annotation s_n of each cell (if available)
And on two additional, unobserved random variables:

- ρ_g^n **nuisance variation**, 1-D Gaussian, model differences in capture efficiency & sequencing depth, cell-specific scaling factor
- z_n, **remaining variation**, 10-D Gaussian, model biological differences between cells.

Represent each cell as point in low-dimensional latent space (for visualization and clustering).
Neural network maps the latent variables to ZINB distribution parameters (Fig. 1a, neural networks 5 and 6).
This mapping goes through intermediate variables:

- batch-corrected, normalized estimate of the percentage of transcripts in each cell n that originate from each gene g

Use these estimates for differential expression analysis
Use scaled version (multiplying by estimated library size) for imputation.
Derived approximation for posterior distribution of latent variables q by training another neural network using variational inference and a scalable stochastic optimization procedure (NN1-NN4).

CycleGAN for scRNA-scATAC translation

Unsupervised distribution matching [Zhu2017]

Learn latent representation for scATAC and for scRNA, and their correspondence

Deep Learning transformers across diverse domains: scRNA, scATAC, Imaging, Phenotypes, Clinical data

Today: Single-cell Genomics

1. Why single cells, traditional approaches, scRNA-seq
2. Scaling up single-cell technologies: evolution of scRNA-seq
3. Case studies: Single-cell Dissection of Brain Disorders
4. Computational challenges in scRNA-seq data analysis

   Clustering similar cells; Clustering similar genes; Dimensionality Reduction; Cell Type Annotation; Multi-Resolution analysis; Cell variability; Differential Expression Analysis; Module Analysis; Spatial Transcriptomics; Trajectories through cell space; Missing Data Imputation; MultiDataset Integration
5. Deep learning methods for scRNA-seq data analysis
6. Beyond scRNA-seq: scATAC-seq, multi-omics

6. Beyond RNA: scATAC-seq, Multi-Omics

Diverse technologies for sc profiling

SINGLE-CELL OMICS

Integrative single-cell analysis

Tim Stuart[1] *and Rahul Satija*[1,2]*

Data types	Method name	Feature throughput	Cell throughput	Refs
Unimodal				
mRNA	Drop-seq	Whole transcriptome	1,000- 10,000	4
	InDrop	Whole transcriptome	1,000- 10,000	5
	10XGenomics	Whole transcriptome	1,000- 10,000	6
	Smart-seq2	Whole transcriptome	100- 300	38
	MARS-seq	Whole transcriptome	100- 300	
	CEL-seq	Whole transcriptome	100- 300	
	SPLiT-seq	Whole transcriptome	2'50,000	8
	sci-RNA-seq	Whole transcriptome	2'50,000	7
Genome sequence	SNS	Whole genome	10- 100	9
	SCI-seq	Whole genome	10,000- 20,000	10
Chromatin accessibility	scATAC-seq	Whole genome	1,000- 2,000	13
	sciATAC-seq	Whole genome	10,000- 20,000	14
	scTHS-seq	Whole genome	10,000- 20,000	15
DNA methylation	scBS-seq	Whole genome	5- 20	17
	snmC-seq	Whole genome	1,000- 5,000	16
	sci-MET	Whole genome	1,000- 5,000	19
	scRRBS	Reduced representation genome	1- 10	18
Histone modifications	scChIP-seq	Whole genome + single modification	1,000- 10,000	24
Chromosome conformation	scHi-C-seq	Whole genome	1- 10	26
Multimodal				
Histone modifications+ spatial	NA	Single locus + single modification	10- 100	23
mRNA +lineage	scGESTALT	Whole transcriptome	1,000- 10,000	32

Integrative single-cell analysis

Tim Stuart e,[1] and Rahul Satija f>[1,2] *

Single-cell Epigenomics (scATAC-Seq)

Buenrostro *et al., 2015*

Trans-factors are associated with single-cell epigenomic variability

Integrate scATAC + scRNA using ChromVAR

Link single-cell epigenomics and single-cell transcriptomics

scATAC-scRNA Epigenome-Transcriptome linking & integration

scATAC cell-type-specific marker accessibility

scATAC subtypes of ExcN and InhN

Micr/ExcN/Astr-specific peaks

Late AD loss of ExcN+InhN (esp. male)

scRNA-scATAC coordinated changes in AD

scRNA-scATAC alignment

AD differential peaks enriched for AD, tauopathies, immune functions, and other neuronal disorders

TFs upstream of AD-assoc. peaks reveal candidate master regulators

Xushen Xiong
Carles Boix
Shahin Mohammadi
Sebastian Pineda
Kiki Galani
Li-Lun Ho

Cell-type-specific GWAS enrichments for scATAC-seq peaks

scATAC-seq: cell-type-specific GWAS enrichments

SPI motifs repeatedly disrupted by AD GWAS SNPs in microglial scATAC peaks
➔ Infer master regulators for each trait in each cell type

Xushen Xiong
Carles Boix
Kiki Galani
Li-Lun Ho
Li-Huei Tsai

scATAC evidence of epigenome erosion in late AD across cell types

- Early AD: active states are 'open', repressed are 'closed'
- Late AD: reversal, closing of promoters/enhancers, opening of heterochromatin/lamina/repressed regions

- scATAC: senescent/de-identified cluster
- Global regulatory shutdown, epigenome erosion
- Cells come specifically from <u>late AD</u> individuals

Xushen Xiong
Carles Boix

- Across all cell types
- Specific to late AD

- Eroded cells found in both non-AD, early-AD, and late-AD
- But dramatic shift in the distribution, especially for late-AD
- <u>de-ID cells, and de-ID oligos/neurons particularly pronounced</u>

- Broad impact on epigenome, spanning Mbs
- Some cells show clear reversal, others just erosion

scMulti-Omics: Multiple profiling of the same cell

Diverse approaches for sc multi-omics

- a | **Gathering cytometric single-cell measurements using multiparameter fluorescenceactivated cell sorting (FACS)** before single- cell RNA sequencing (scRNA- seq) canallow fluorescence- based measurements of protein levels to be later linked to cellular transcriptomes; hence, RNAand protein levels can be analysed jointly in the same cell.
- b | **A lyse-and-split strategy can allow parallel workflows to be performed on different cellular fractions**. For example, the cytosol can be physically separated from the nucleus to allow measurement of cytosolic mRNAs through scRNA- seq and measurements of the genomic DNA using whole- genome sequencing or bisulfite sequencing to gather complementary data on the cell genotype or methylome, respectively.
- c1 | Innovative **barcoding strategies** can enable standard scRNA- seq methods to capture important additional information to enhance the analysis of cell transcriptomes. **Cell surface protein abundance** can be capturedusing standard scRNA- seq methods by conjugating **polyadenylated antibody barcodes** to antibodies targeting cell surface proteins20,21 (left panel).
- c2 | These **antibody barcode sequences** can be captured alongside polyadenylated mRNAs and decoded to provide an estimate of protein levels for each cell. **Allelic information** can be encoded by the single- guide RNA (sgRNA) sequence used to guide Cas9 in pooled genetic screens, allowing **gene knockout information** to be associated with single- cell transcriptional profiles (middle panel).
- c3 | **Cell lineage can also be encoded in a polyadenylated barcode sequence** through the cumulative editing of a lineage array sequence by Cas9 (right panel). Over time, Cas9 will cut the lineage array , resulting in **mutations at different points in the array.** Cells sharing common mutations in the lineage array are likely to have originated from the same progenitor. By placing the **lineage array sequence** under the control of an RNA polymerase II promoter, these sequences can also be captured alongside endogenous mRNAs.
- d | Additional information can be extracted from scRNA- seq data beyond a typical analysis that provides only estimates of transcript counts in each cell. **Somatic mutations** can be identified from sequencing reads for each individual cell and can be used to reconstruct lineage relationships between cells. Retained introns can also be detected and can be used to give an estimate of the rate of change in transcript abundance (RNA velocity70). scBS-seq, single- cell bisulfite sequencing; scDNA- seq,single- cell DNA sequencing.

SINGLE-CELL OMICS

Integrative single-cell analysis

*Tim Stuart[1] and Rahul Satija[1,2]**