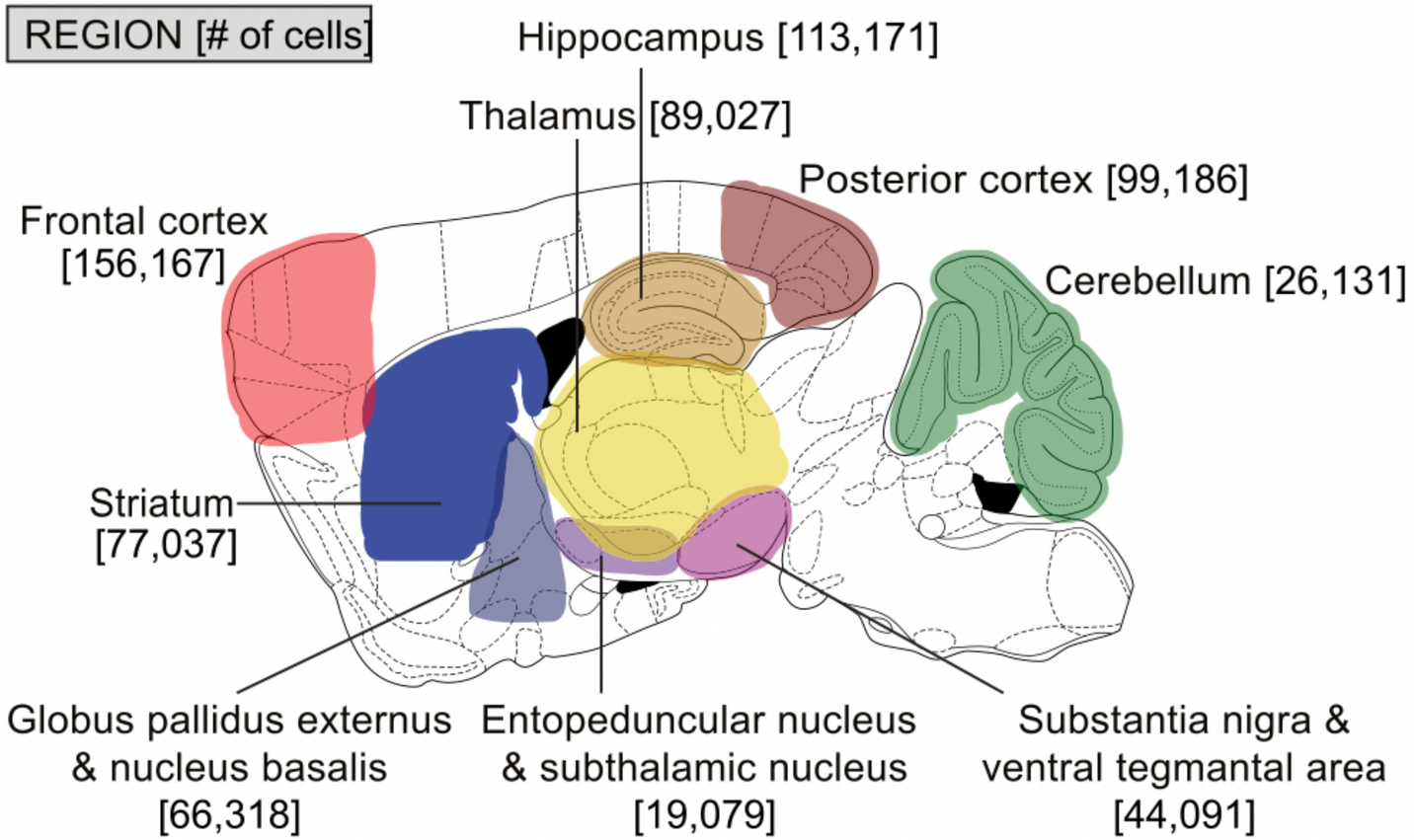
Computational Methods for Phenotyping a Perturbation in Single Cell Data

Stanley Center Primer
11/27/2022

Aleksandrina Goeva

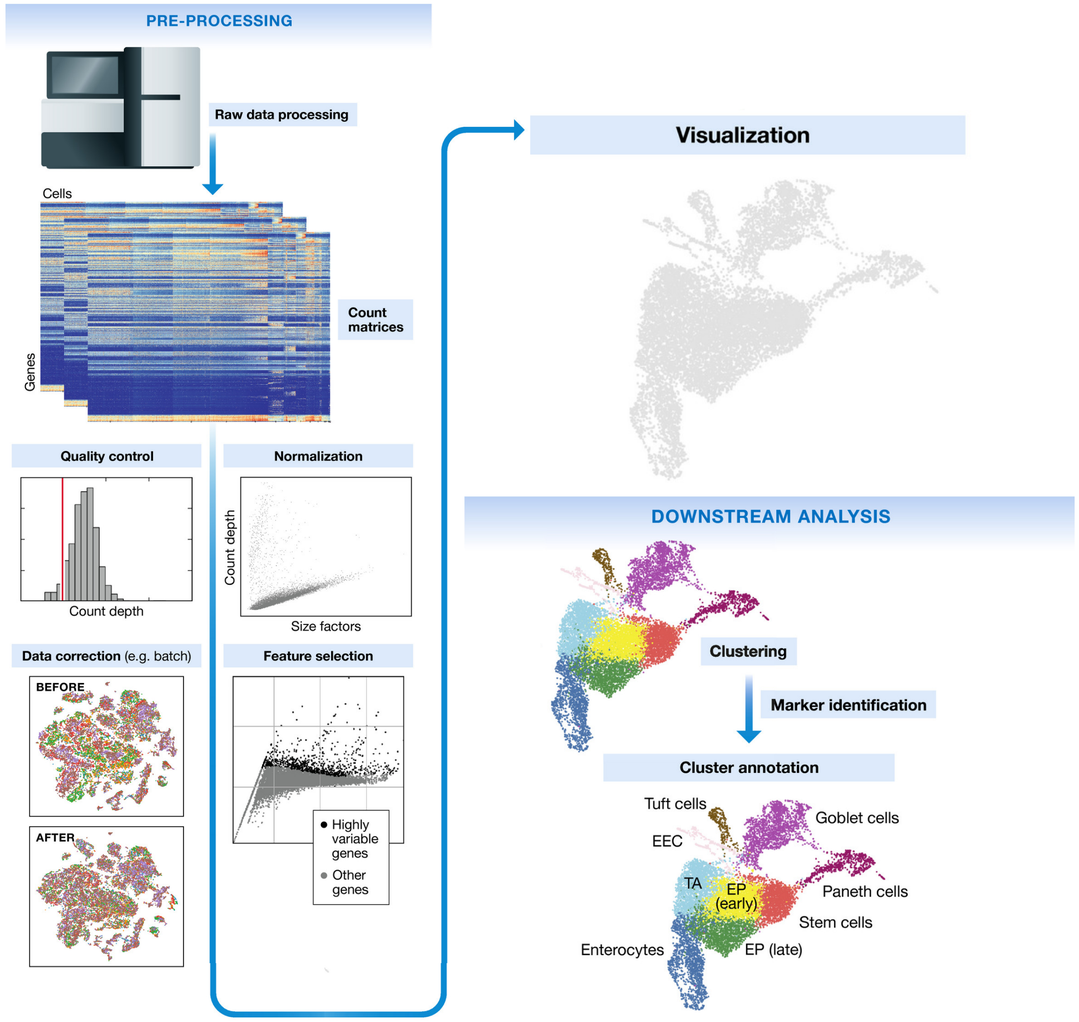
Macosko Lab

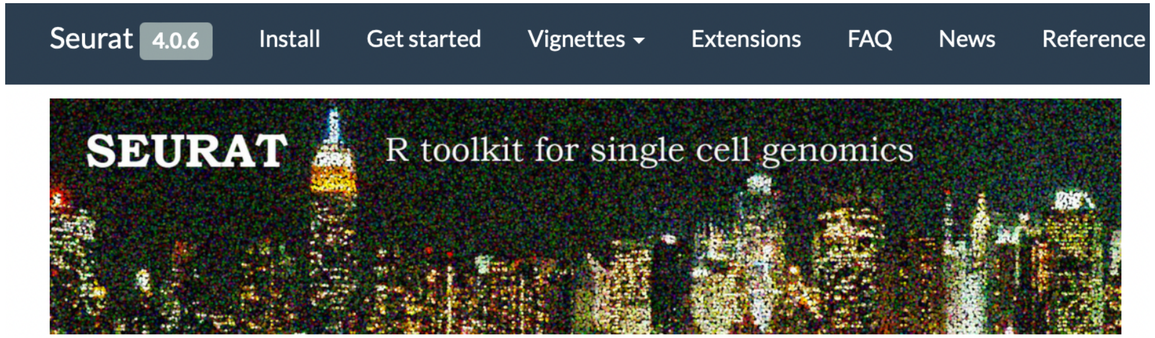
Single cell atlases create comprehensive maps of transcriptional heterogeneity in complex tissues



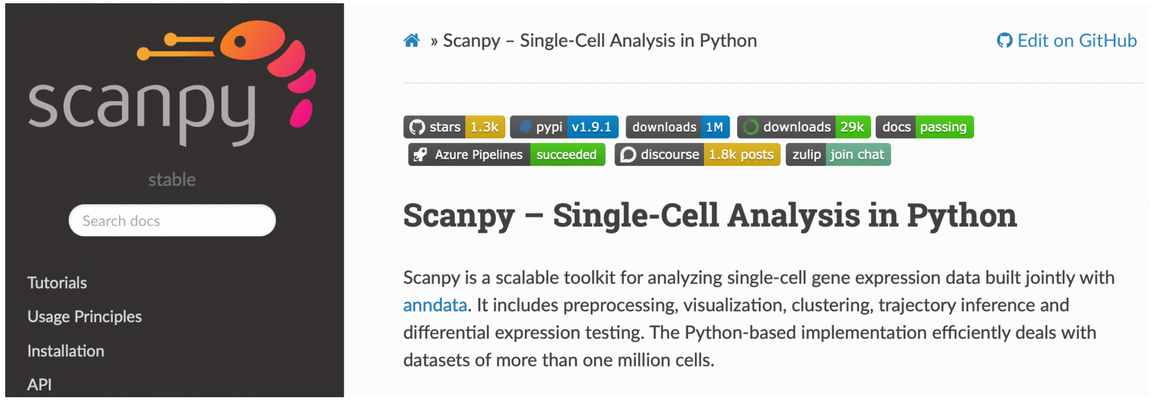
Saunders, A., Macosko, E.Z., et al. Cell (2018)

We have an established pipeline for interrogating the heterogeneity in a sample



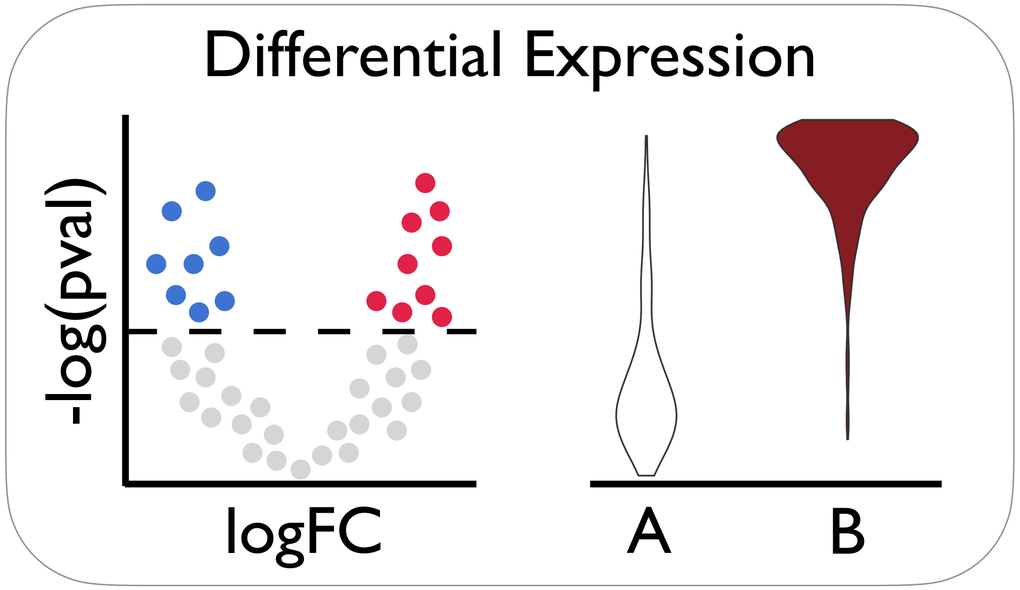


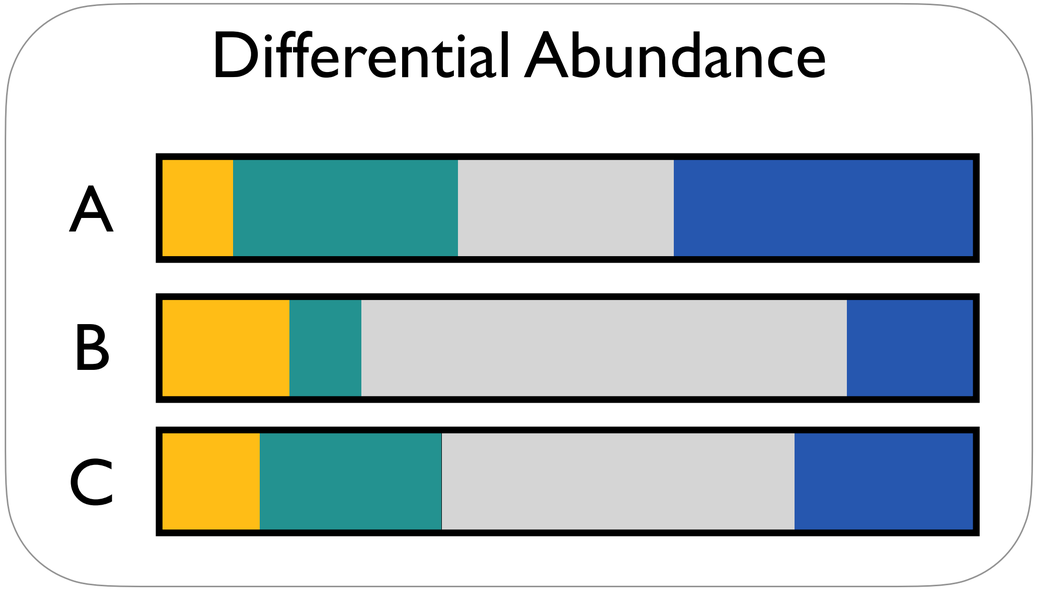
Official release of Seurat 4.0

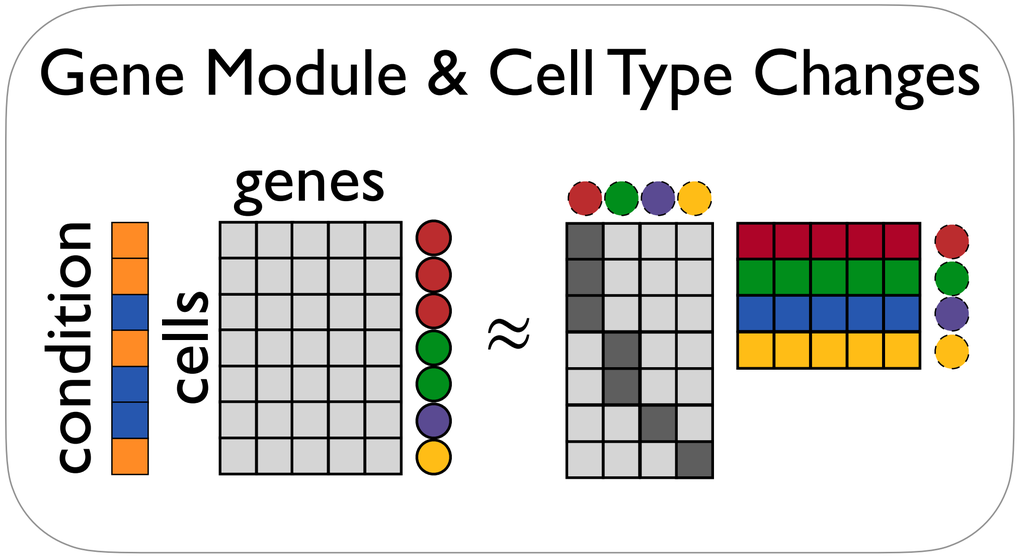


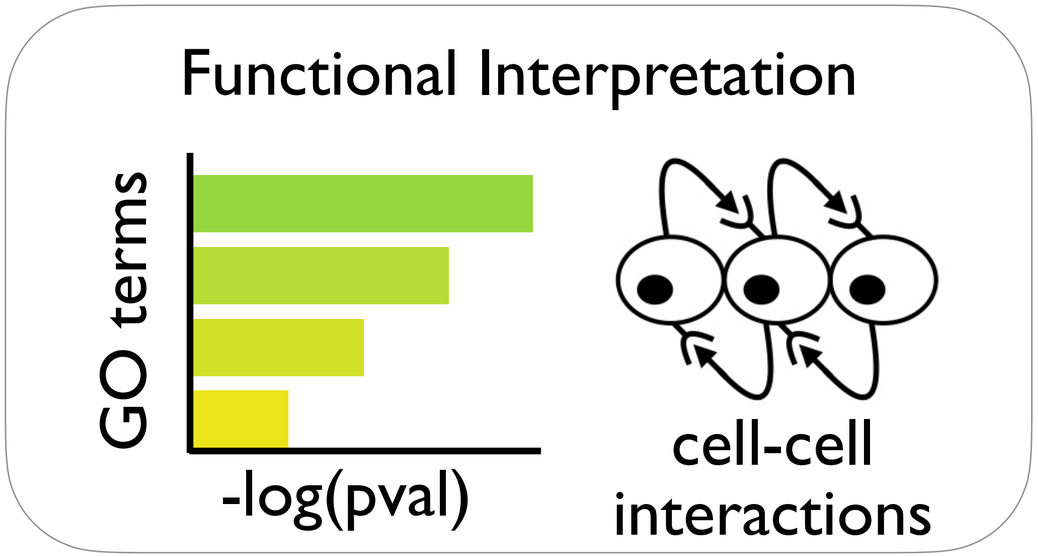
Luecken, M.D., Theis, F.J. Molecular Systems Biology (2019)

Today we will focus on methods for finding differences across conditions



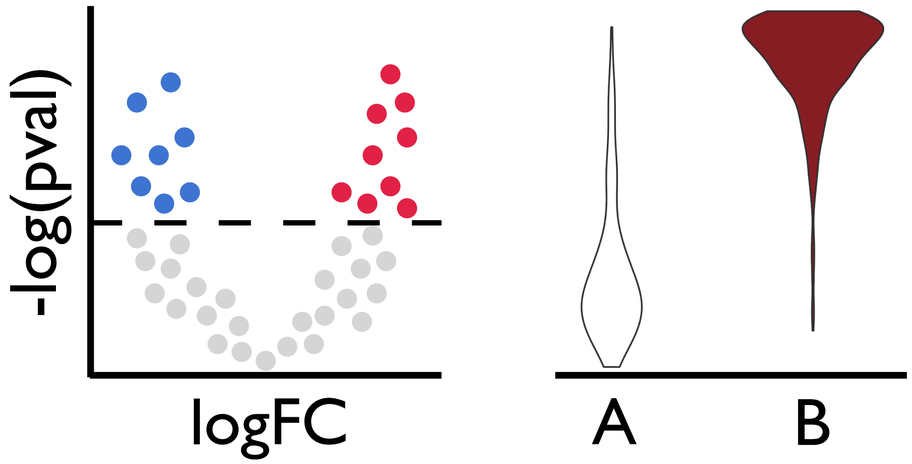






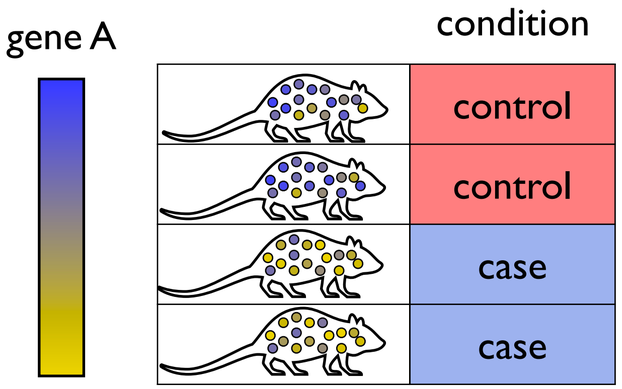
Differential Expression

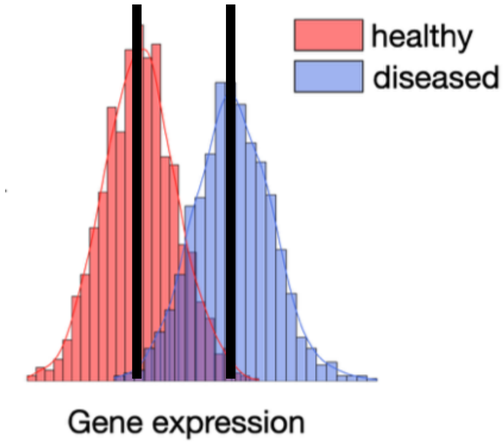
Identification of genes associated with a perturbation



Condition-level aggregation

The default method





Seurat - Guided Clustering Tutorial

Compiled: January 11, 2022

Finding differentially expressed features (cluster biomarkers)

```
# find all markers of cluster 2
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2, min.pct = 0.25)
head(cluster2.markers, n = 5)
```

Gene expression markers of identity classes

Finds markers (differentially expressed genes) for identity classes

```
FindMarkers(object, ...)

# S3 method for default
FindMarkers(
  object,
  slot = "data",
  counts = numeric(),
  cells.1 = NULL,
  cells.2 = NULL,
  features = NULL,
  logfc.threshold = 0.25,
  test.use = "wilcox",
```

scanpy.tl.rank_genes_groups

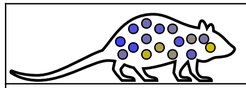
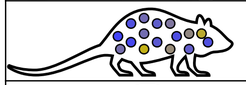
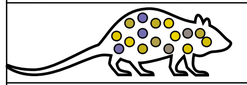
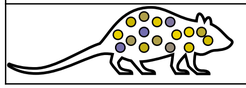
method : `Optional` [`Literal` ['logreg', 't-test', 'wilcoxon', 't-test_overestim_var']] (default: `None`)

The default method is `'t-test'` , `'t-test_overestim_var'` overestimates variance of each group, `'wilcoxon'` uses Wilcoxon rank-sum, `'logreg'` uses logistic regression. See [Ntranos18], here and here, for why this is meaningful.

Sample-level aggregation

Allows for accounting for covariates

gene A

	condition	sex	weight	animal ID
	control	F	53	1
	control	M	61	2
	case	F	55	3
	case	M	58	4

$$= NB(\beta_0 + \beta_1 + \beta_2 + \beta_2, \sigma^2)$$

Bioinformatics. 2010 Jan 1; 26(1): 139–140.
Published online 2009 Nov 11. doi: 10.1093/bioinformatics/btp616
PMCID: PMC2796818
PMID: 19910308

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson,[1,2,*†] Davis J. McCarthy,[2,†] and Gordon K. Smyth[2]

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber & Simon Anders

Genome Biology **15**, Article number: 550 (2014) | Cite this article
395k Accesses | **28144** Citations | **110** Altmetric | Metrics

Nucleic Acids Res. 2015 Apr 20; 43(7): e47.
Published online 2015 Jan 20. doi: 10.1093/nar/gkv007
PMCID: PMC4402510
PMID: 25605792

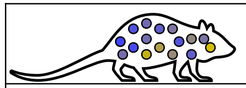
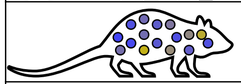
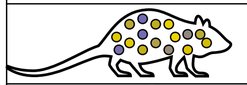
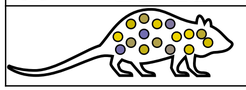
limma powers differential expression analyses for RNA-sequencing and microarray studies

Matthew E. Ritchie,[1,2] Belinda Phipson,[3] Di Wu,[4] Yifang Hu,[5] Charity W. Law,[6] Wei Shi,[5,7] and Gordon K. Smyth[2,5,*]

Bespoke single-cell methods

Can account for cell-level covariates

gene A

	condition	sex	weight	animal ID
	control	F	53	1
	control	M	61	2
	case	F	55	3
	case	M	58	4

$$= \mathrm{NB}(\beta_0 + \beta_1 + \beta_2 + \beta_2 + \beta_3, \sigma^2)$$

glmGamPoi: fitting Gamma–Poisson generalized linear models on single cell count data

Constantin Ahlmann-Eltze, Wolfgang Huber

Bioinformatics, Volume 36, Issue 24, 15 December 2020, Pages 5701–5702,
https://doi.org/10.1093/bioinformatics/btaa1009

MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

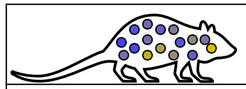
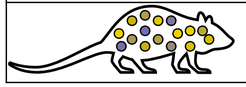
Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley & Raphael Gottardo

Genome Biology **16**, Article number: 278 (2015) | Cite this article

Differential Expression

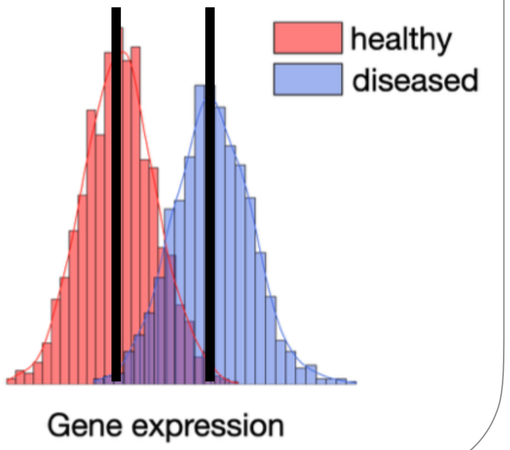
Summary

gene A

	condition	sex	weight	animal ID
	control	F	53	1
	control	M	61	2
	case	F	55	3
	case	M	58	4

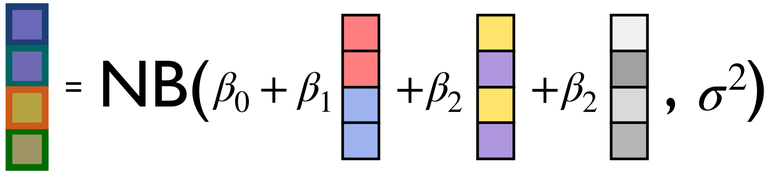
t-test, Wilcoxon

- one gene at a time
- average per condition



edgeR, DEseq2, limma

- regress one gene at a time
- count model
- typically used in pseudo-bulk mode
- can be used per cell but are slow
- can account for covariates

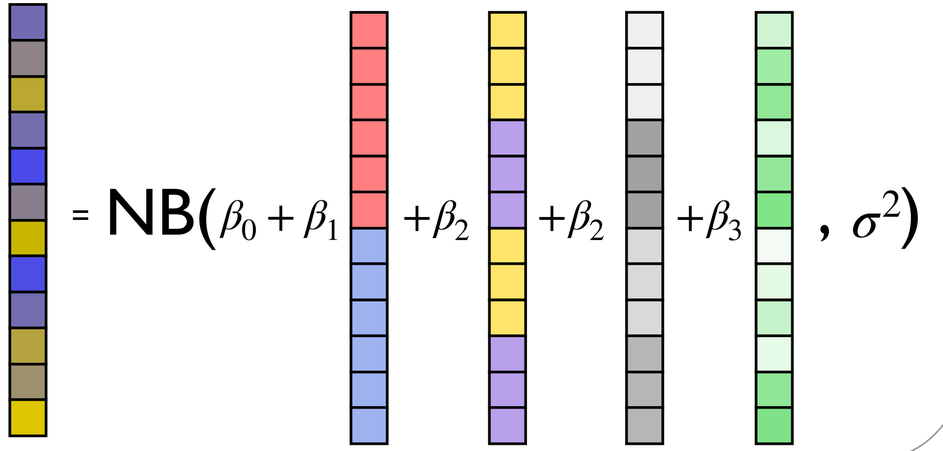


glmGamPoi

- fast alternative to edgeR
- ran per cell
- fixed effects only

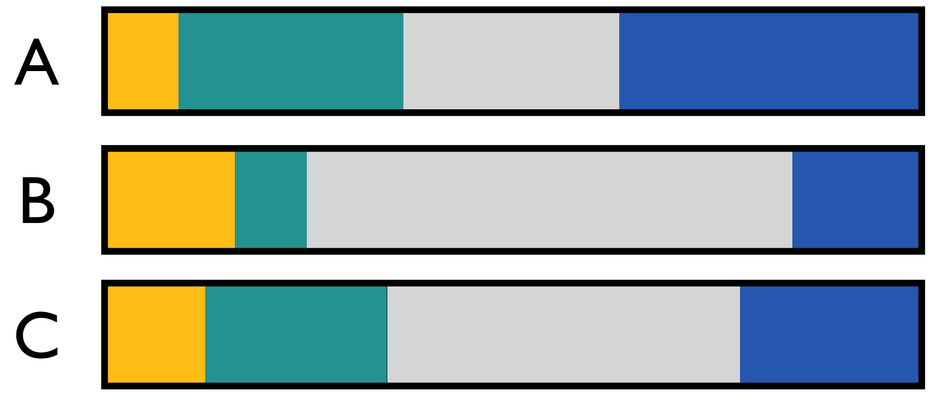
MAST

- ran per cell
- can include random effects



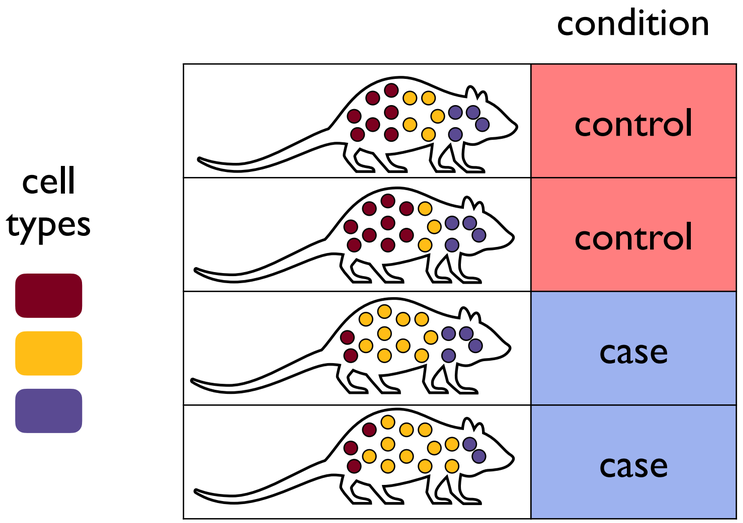
Differential Abundance

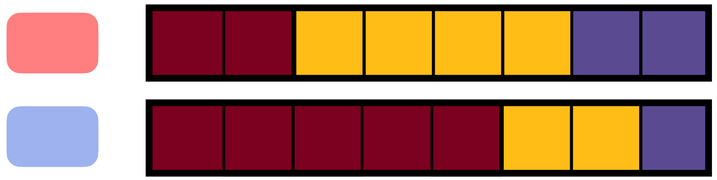
Identification of cell type proportion changes across conditions



Univariate tests

Reflects how enriched each cell type is across conditions, but does not account for how changes in other cell types might affect that





vs rest

		rest	
	2	6	8
	5	3	8
	7	9	16

vs rest

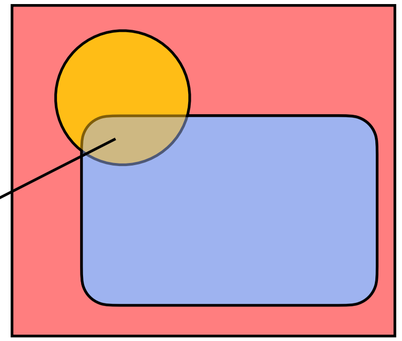
		rest	
	4	4	8
	2	6	8
	6	10	16

vs rest

		rest	
	2	6	8
	1	7	8
	3	13	16

$$\frac{\binom{\text{total case}}{\text{yellow case}}\binom{\text{total control}}{\text{yellow control}}}{\binom{\text{all cells}}{\text{total yellow}}}$$

overlap

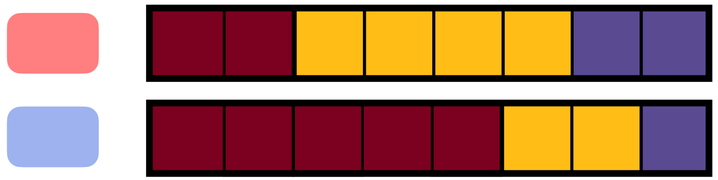


Fisher's exact test

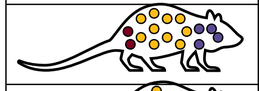
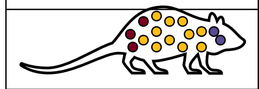
- test of independence of row and column variable in the 2x2 table
- assumes fixed row and col totals
- equivalent to hypergeometric
- relax fixed marginals assumptions
- elevated FDR

Cell type proportions are not independent of each other

Since all proportions sum to 1, an increase in the proportion of one cell subset will necessarily lead to a decrease in the proportions of other cell subsets



cell types

	condition	sex	weight	animal ID
	control	F	53	1
	control	M	61	2
	case	F	55	3
	case	M	58	4

DirichletReg: Dirichlet Regression for Compositional Data in R

Marco Maier

Institute for Statistics and Mathematics

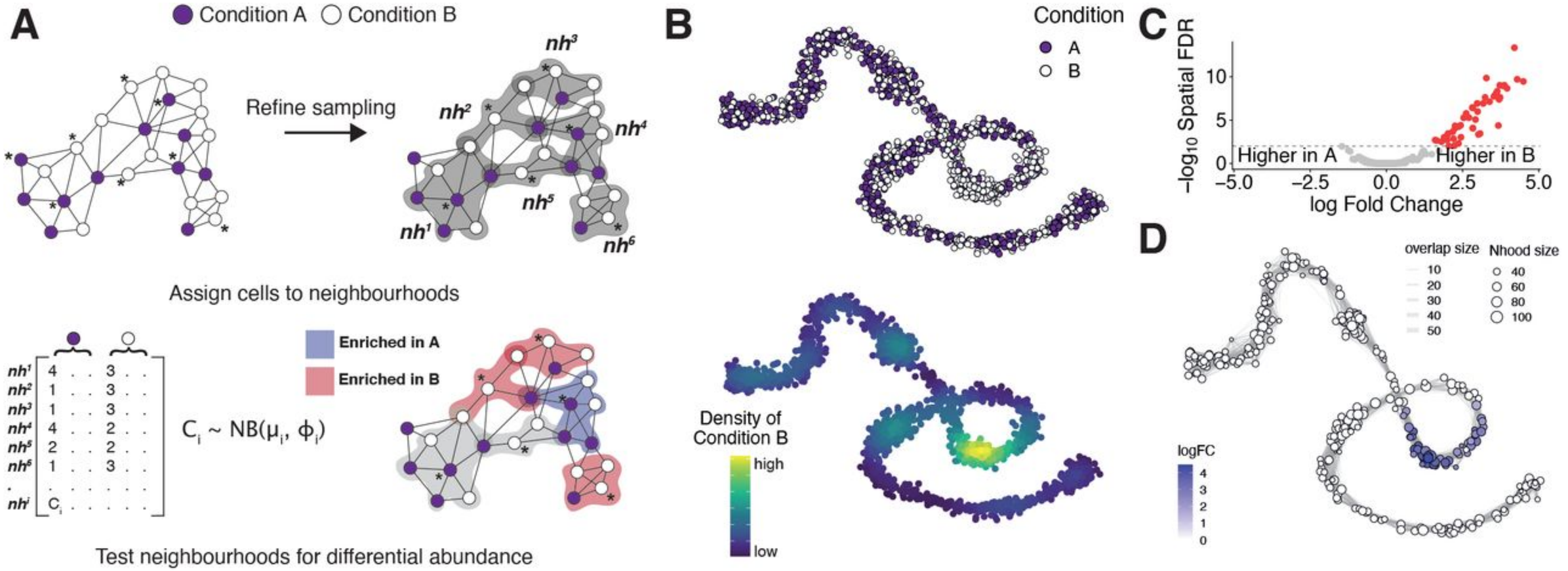
nature > nature communications > articles > article

Article | Open Access | Published: 25 November 2021

scCODA is a Bayesian model for compositional single-cell data analysis

M. Büttner, J. Ostner, C. L. Müller ✉, F. J. Theis & B. Schubert ✉

Compositional changes when cells lie on a continuum

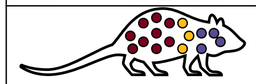
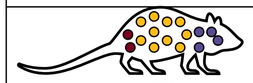
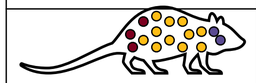


Dann, E., et al. Nature Biotechnology (2021)

Differential Abundance

Summary

cell types

	condition	sex	weight	animal ID
	control	F	53	1
	control	M	61	2
	case	F	55	3
	case	M	58	4

Univariate tests

- one vs all
- Fisher's exact test
- elevated FDR

Dirichlet-multinomial regression

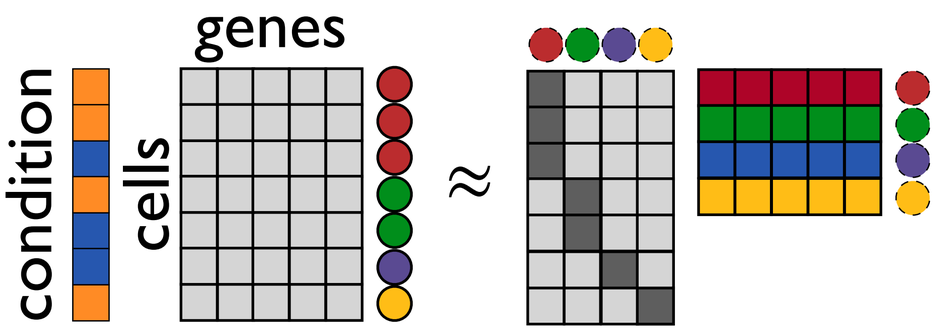
- accounts for proportions of all cell types
- can account for covariates

Milo

- differential abundance testing when cells are on a continuum rather than belonging to discrete clusters

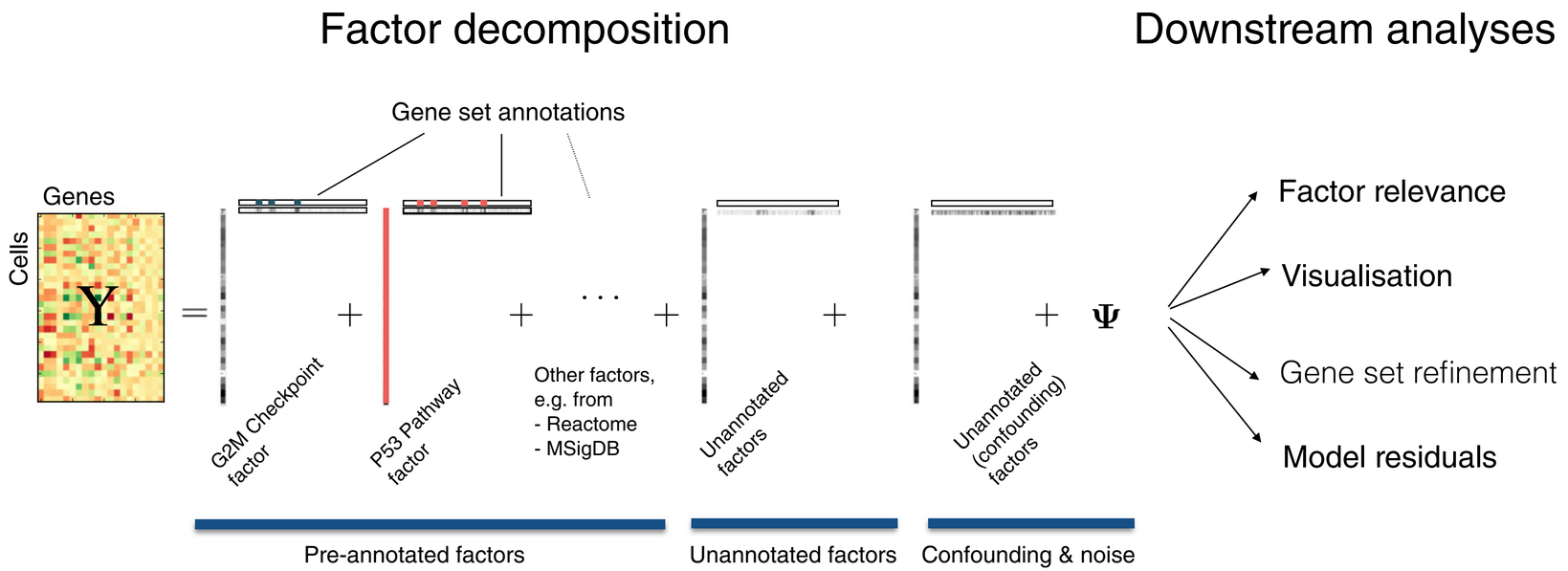
Gene Module and Cell Type Changes

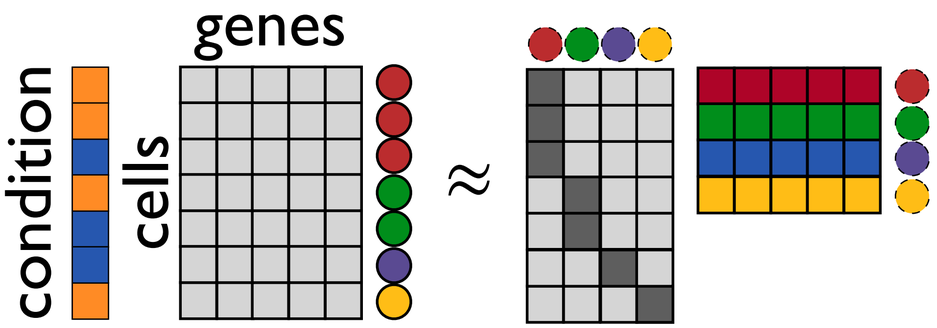
Detection of groups of genes and groups of cells correlated with the perturbation

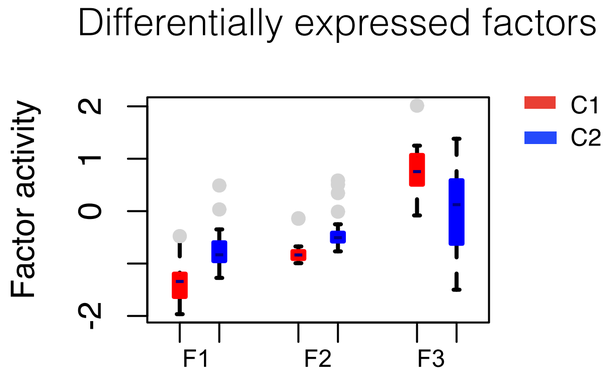


Factor analysis approach

Correlating groups of genes with the perturbation effect



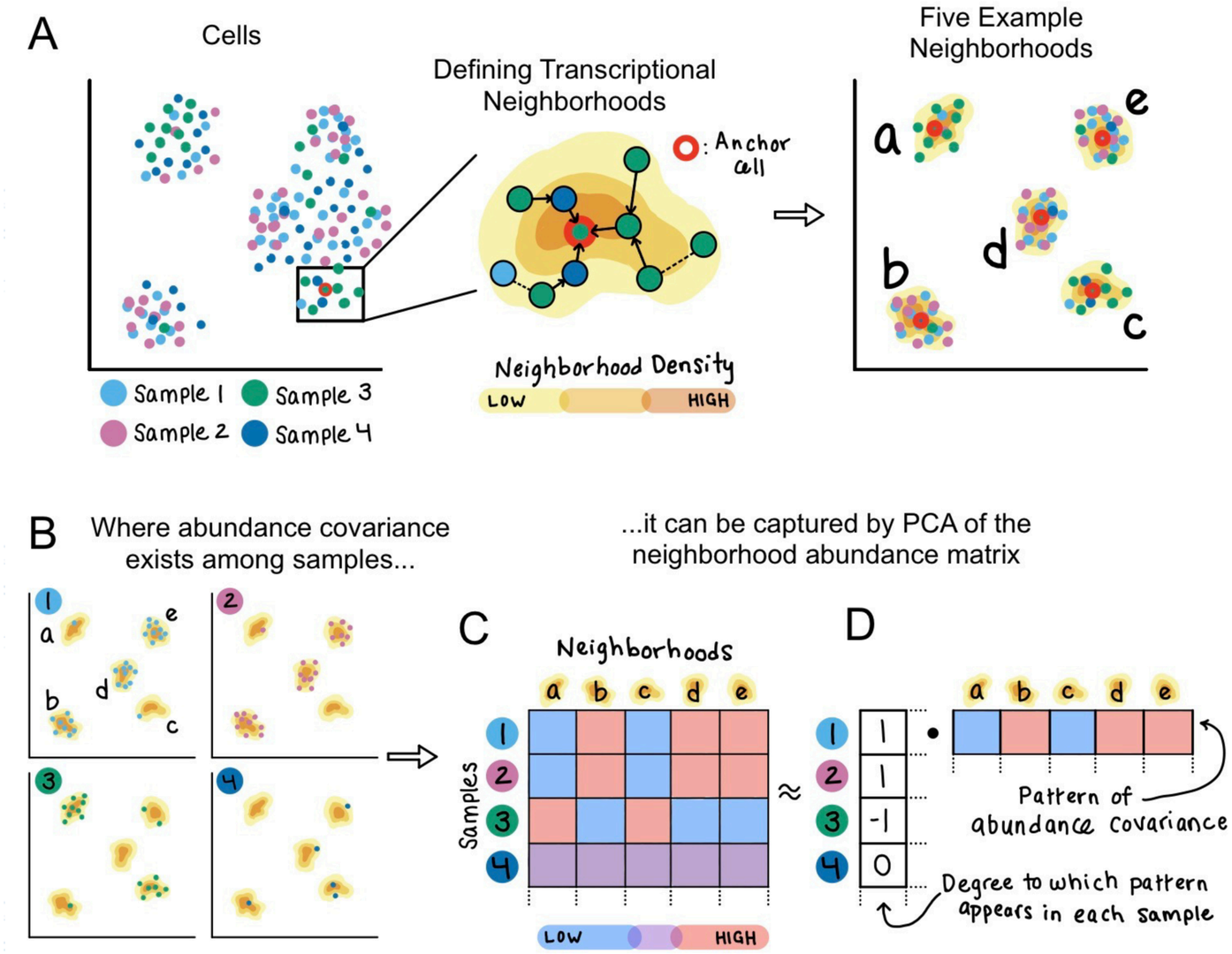




F1: Transmission across chem. synapses
F2: Signalling by Egfr
F3: Peptide chain elongation

Buettner, F., et al. Genome Biology (2017)

Cluster-free approach for identifying cell states associated with a perturbation

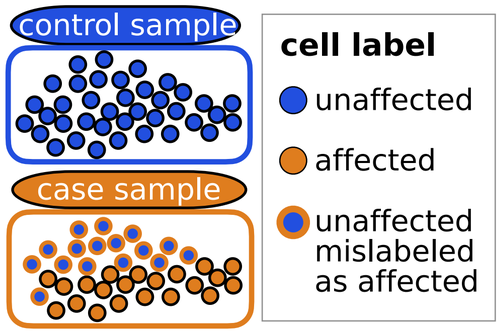


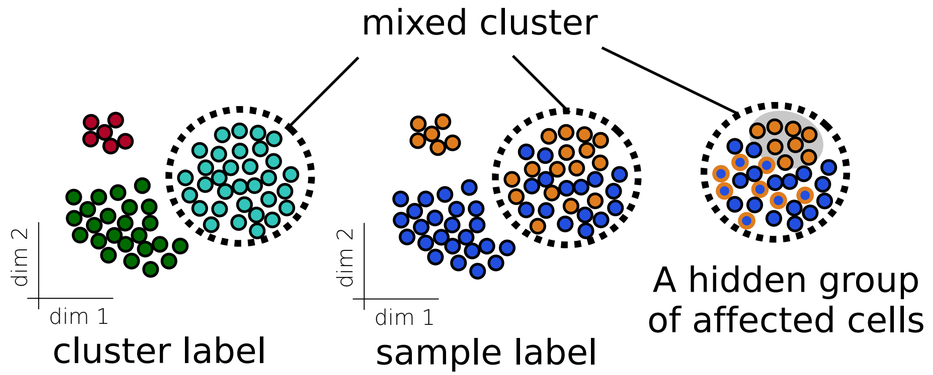
Reshef, Y.A., et al. Nature Biotechnology (2022)

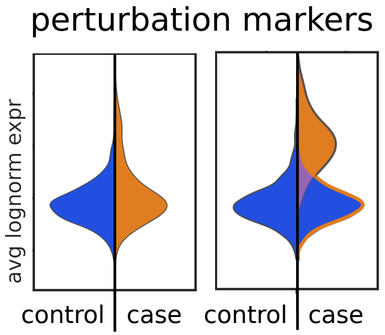
Key assumptions of all prior methods

- The latent space is a complete and accurate representation of variation in the data
- Unsupervised analyses are sufficient to identify groups of cells affected by the perturbation
- Perturbation labels correctly represent the presence or absence of an effect in individual cells

HiDDEN: A machine learning label refinement method for detection of disease-relevant populations in case-control single-cell transcriptomics








Mike Dolan


Judy Luu


Eric Garcia

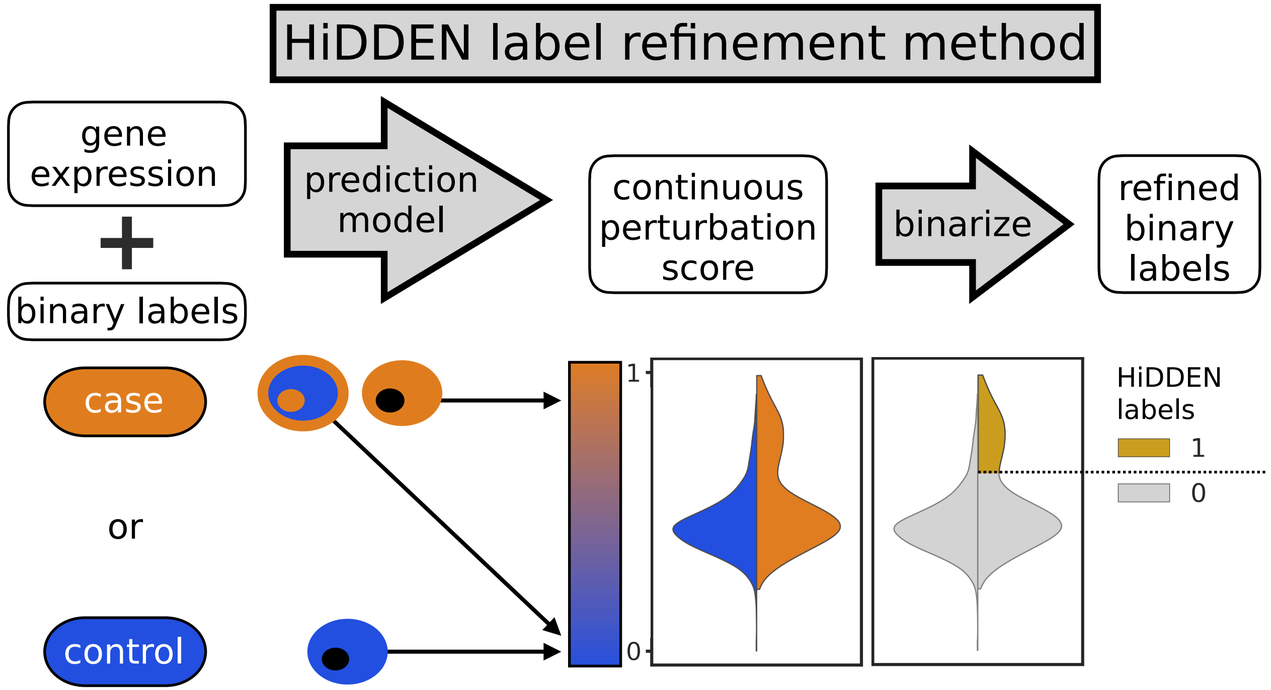

Rebecca Boiarsky


Rajat Gupta


Evan Macosko

Goeva, A., et al. *[to appear on bioRxiv]* (2022)

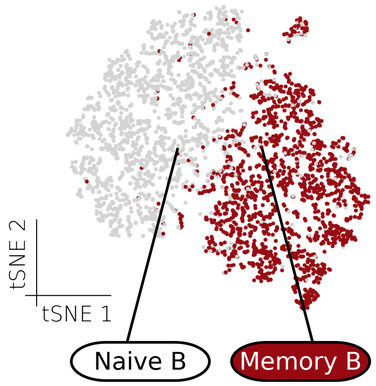
Method overview

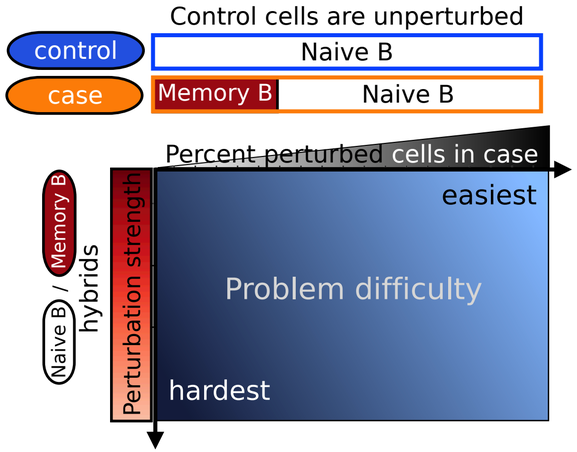


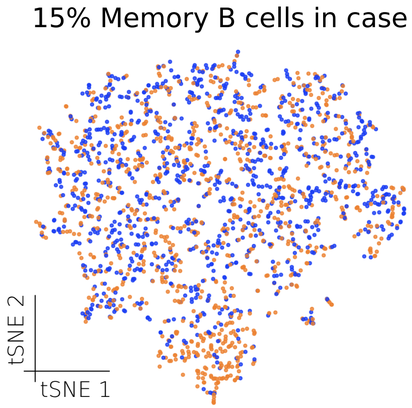
Goeva, A., et al. *[to appear on bioRxiv]* (2022)

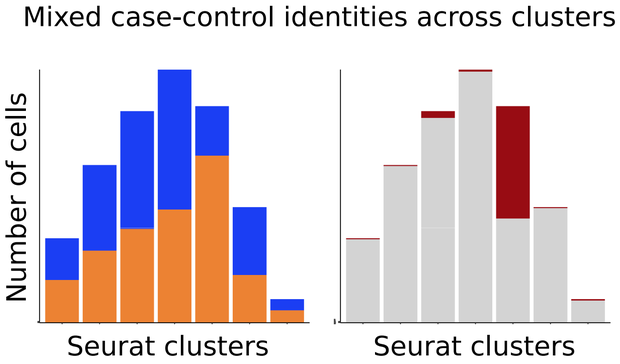
The standard analysis workflow can miss the biological signal

in simulated ground truth datasets of cell type mixtures



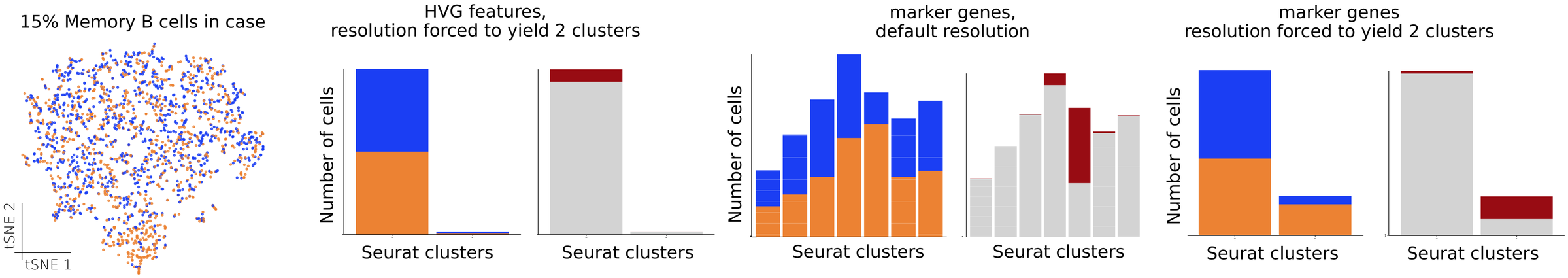






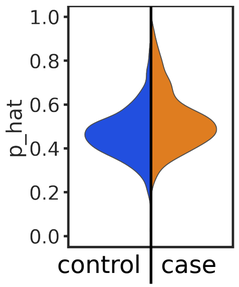
Goeva, A., et al. *[to appear on bioRxiv]* (2022)

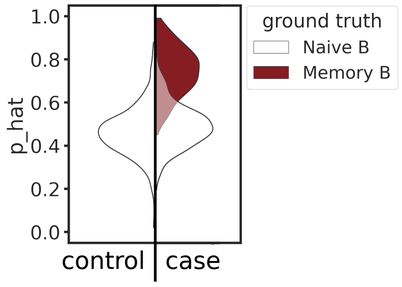
The recovery of the Memory B cluster could not be improved by gene selection or adjusting the resolution parameter alone

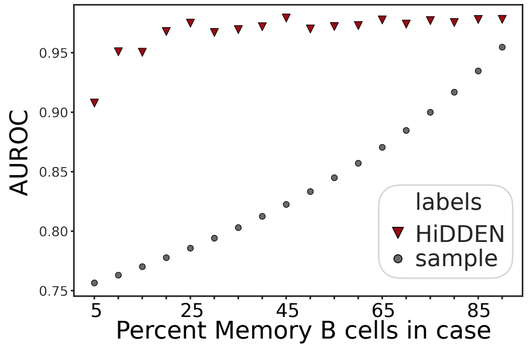


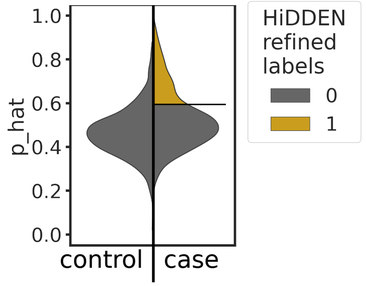
Goeva, A., et al. *[to appear on bioRxiv]* (2022)

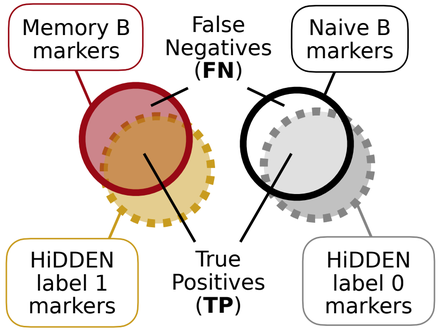
HiDDEN detects biological signal missed by the standard analysis workflow

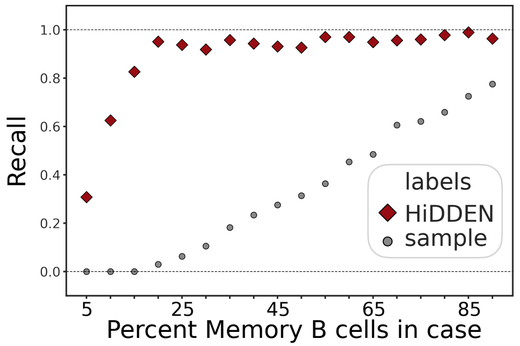








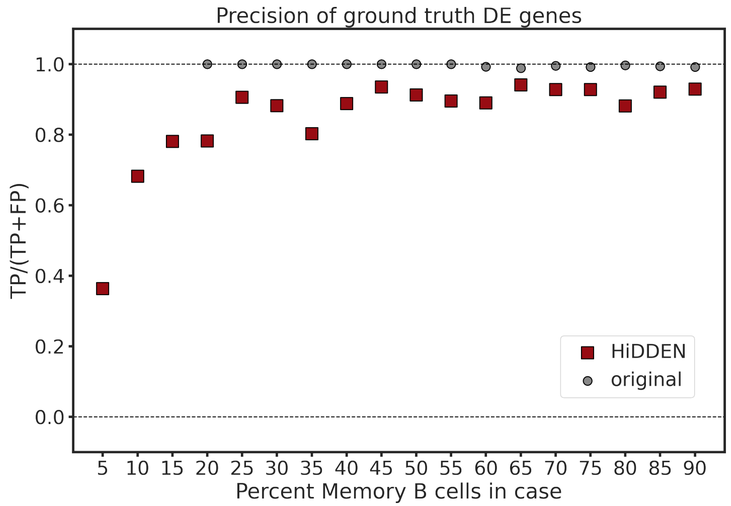


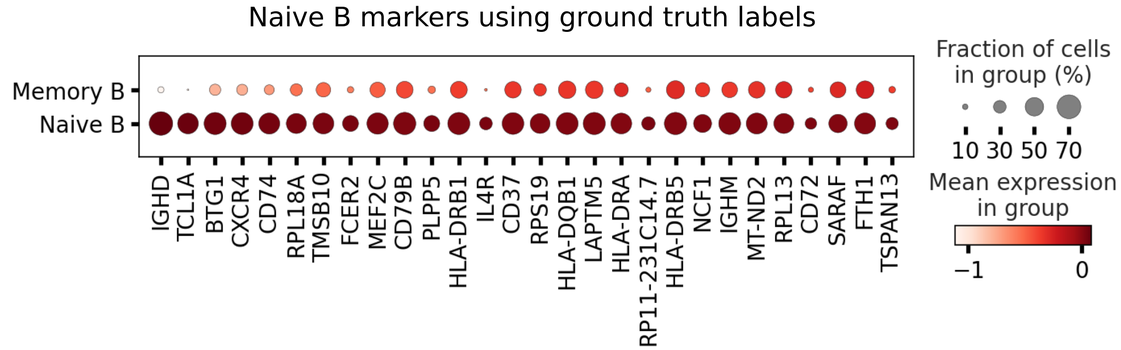


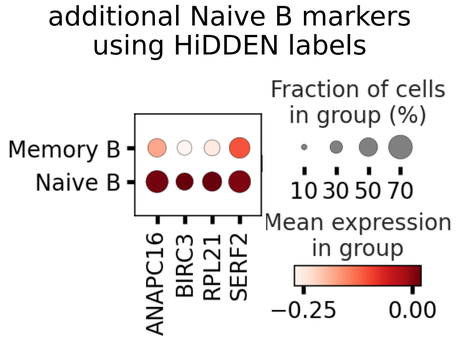
Goeva, A., et al. *[to appear on bioRxiv]* (2022)

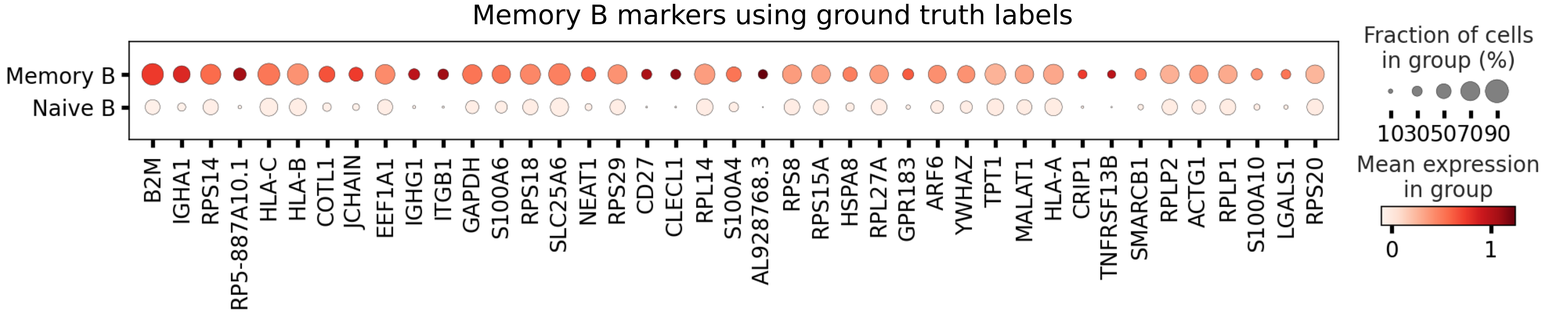
HiDDEN identifies additional marker genes

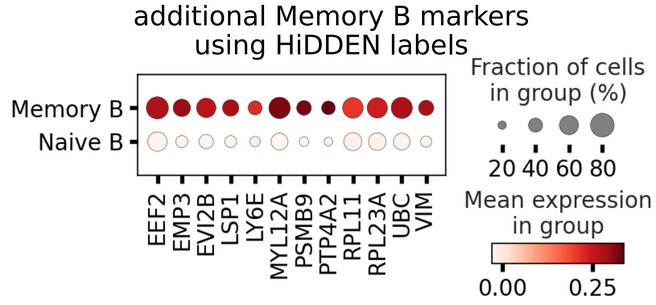
indicating a slight amount of misclassification might have occurred in the original annotation







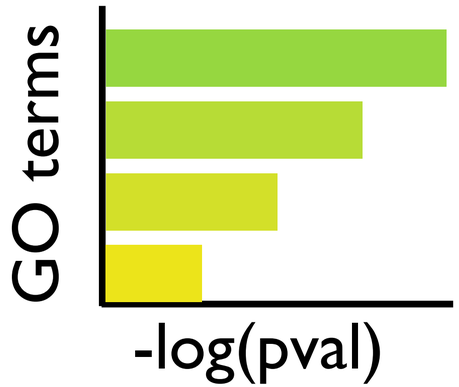




Goeva, A., et al. *[to appear on bioRxiv]* (2022)

Functional Interpretation

Extracting the biological meaning of genes affected by the perturbation

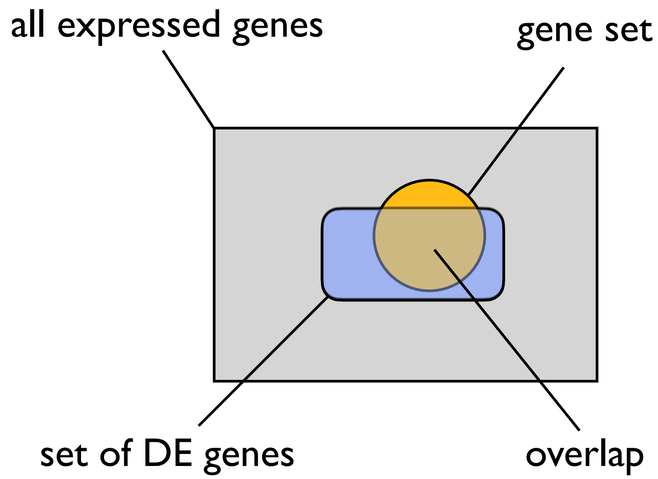


Differential expression analyses can return many genes

Endow genes affected by the perturbation with a biological interpretation

Overrepresentation test

set of DE genes



GSEA

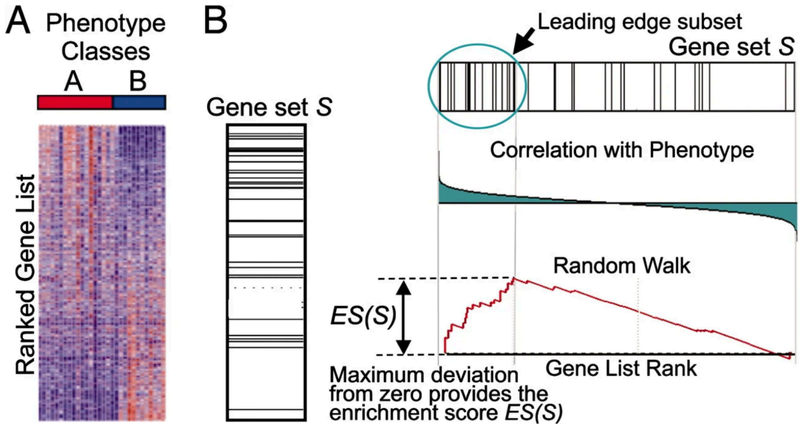
ranked list of all genes

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov -7 Authors Info & Affiliations

September 30, 2005 | 102 (43) 15545-15550 | https://doi.org/10.1073/pnas.0506580102



Subramanian, A., et al. PNAS (2005)

Gene set enrichment analyses often return many terms

Revigo can summarize them by removing redundant GO terms

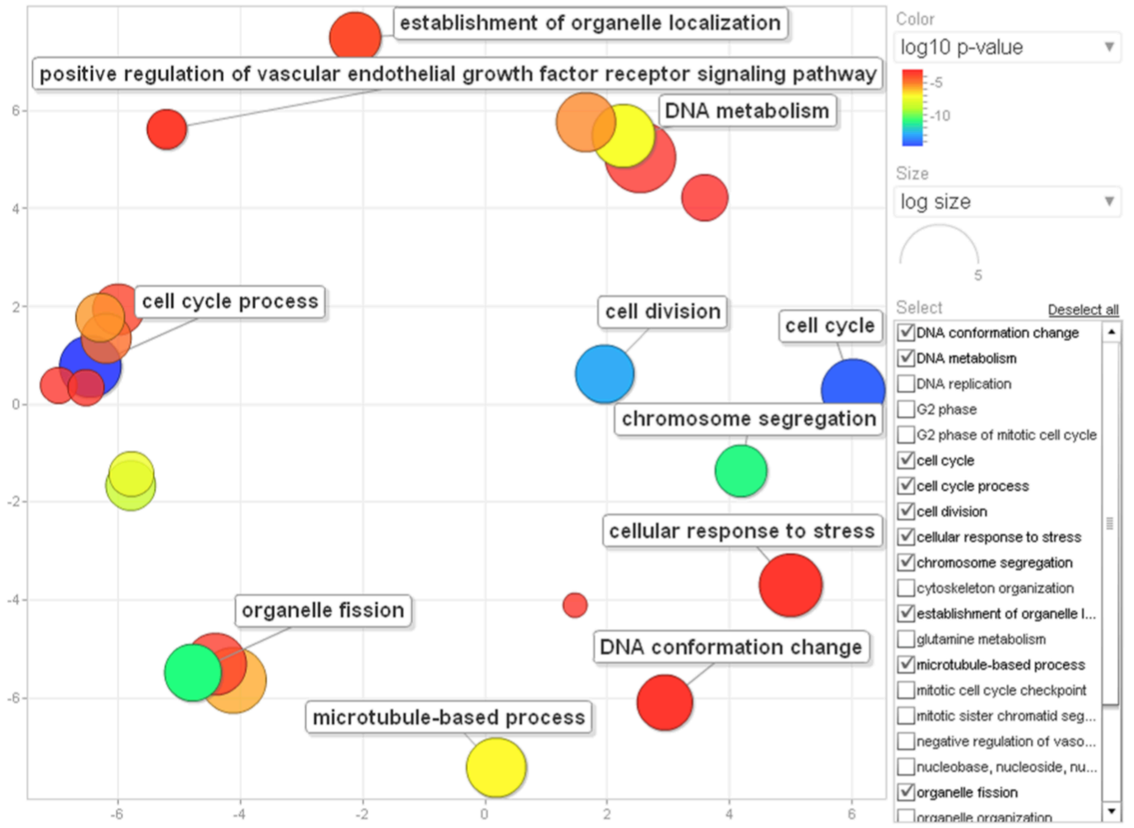
REVIGO

reduce + visualize Gene Ontology

REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms

Fran Supek, Matko Bošnjak, Nives Škunca, Tomislav Šmuc

Published: July 18, 2011 • https://doi.org/10.1371/journal.pone.0021800



Supek, F., et al. PLOS ONE (2011)

Gene set enrichment

Summary

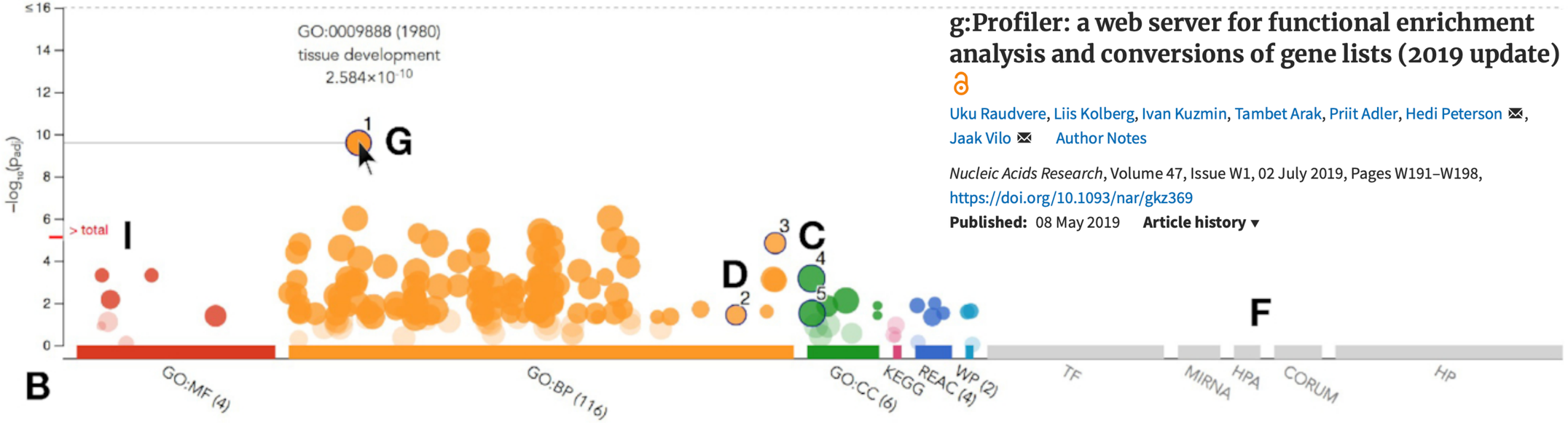
- The results can be more sensitive to the choice of gene set database rather than statistical methods

MSigDB Molecular Signatures Database

GENEONTOLOGY Unifying Biology

KEGG Kyoto Encyclopedia of Genes and Genomes

reactome



g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo Author Notes

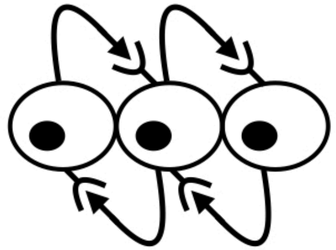
Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W191–W198, https://doi.org/10.1093/nar/gkz369

Published: 08 May 2019 **Article history**

Raudvere, U., et al. Nucleic Acids Res (2019)

Functional Interpretation

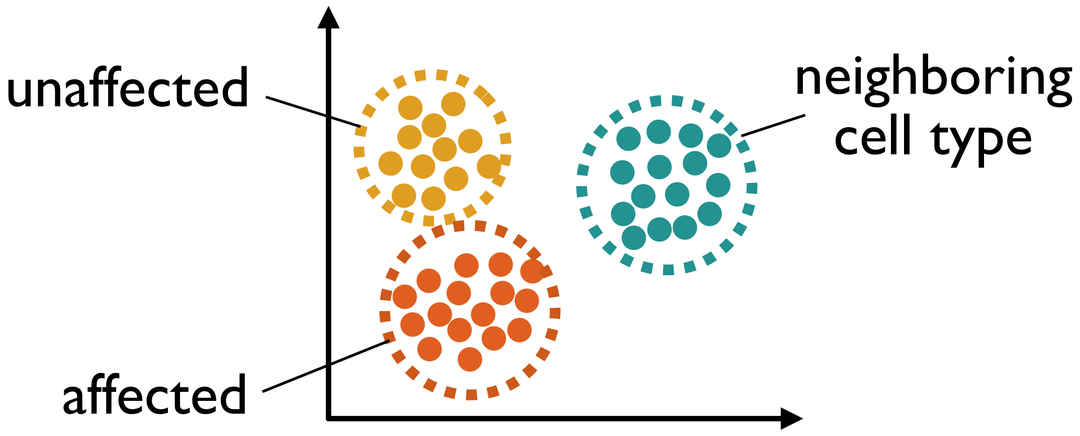
Testing for perturbation-associated cell-cell communication changes



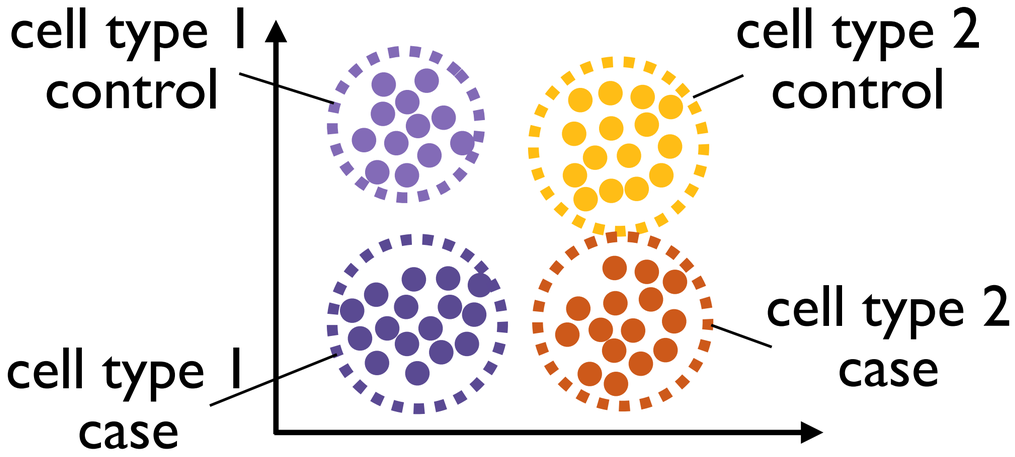
cell-cell
interactions

Contrast cell-cell interactions

of affected and unaffected cells
with a neighboring cell type

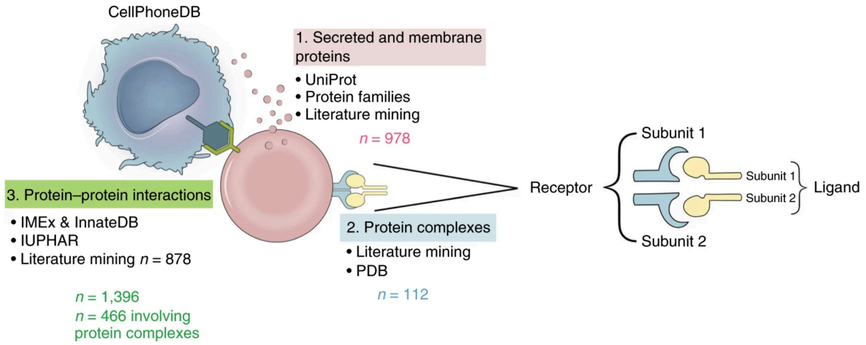


between two cell types
across conditions

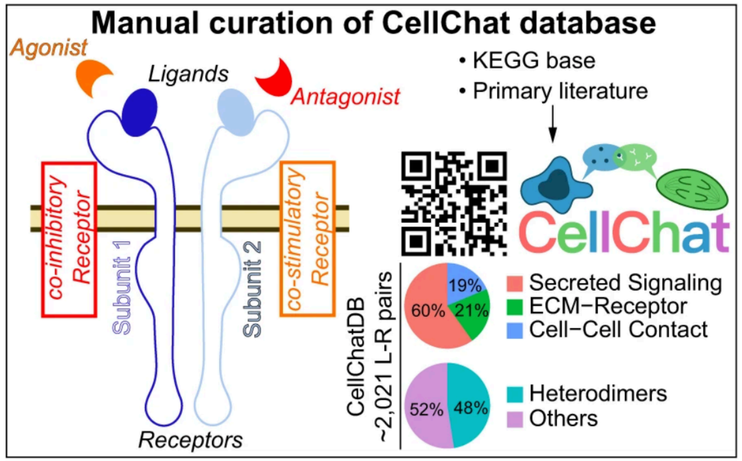


Ligand-receptor interaction repositories

CellphoneDB, CellChat
heteromeric complexes

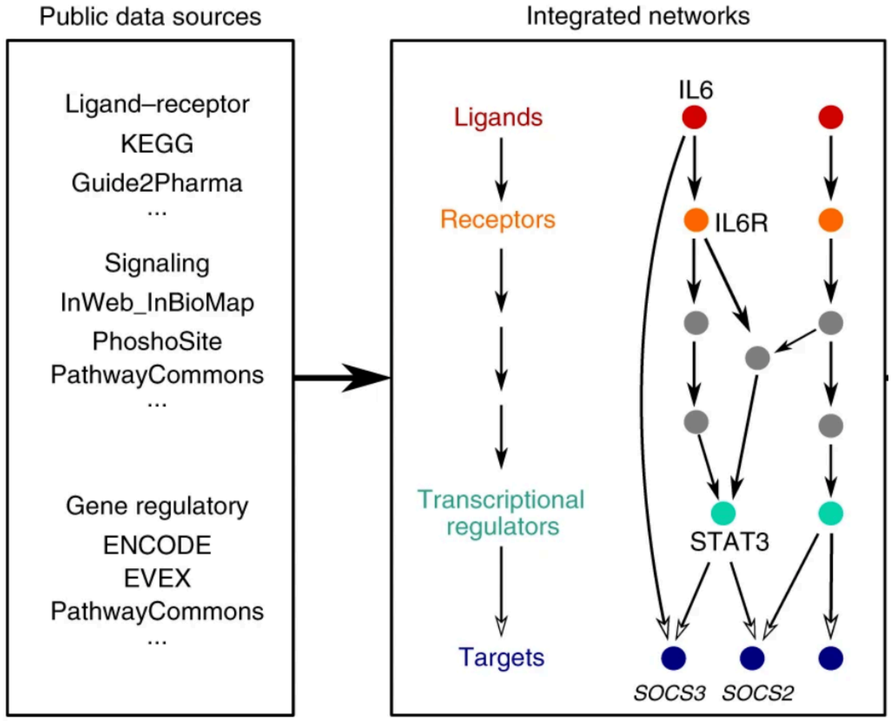


Efremova, M., et al. Nature Protocols (2020)



Jin, S., et al. Nature Communications (2021)

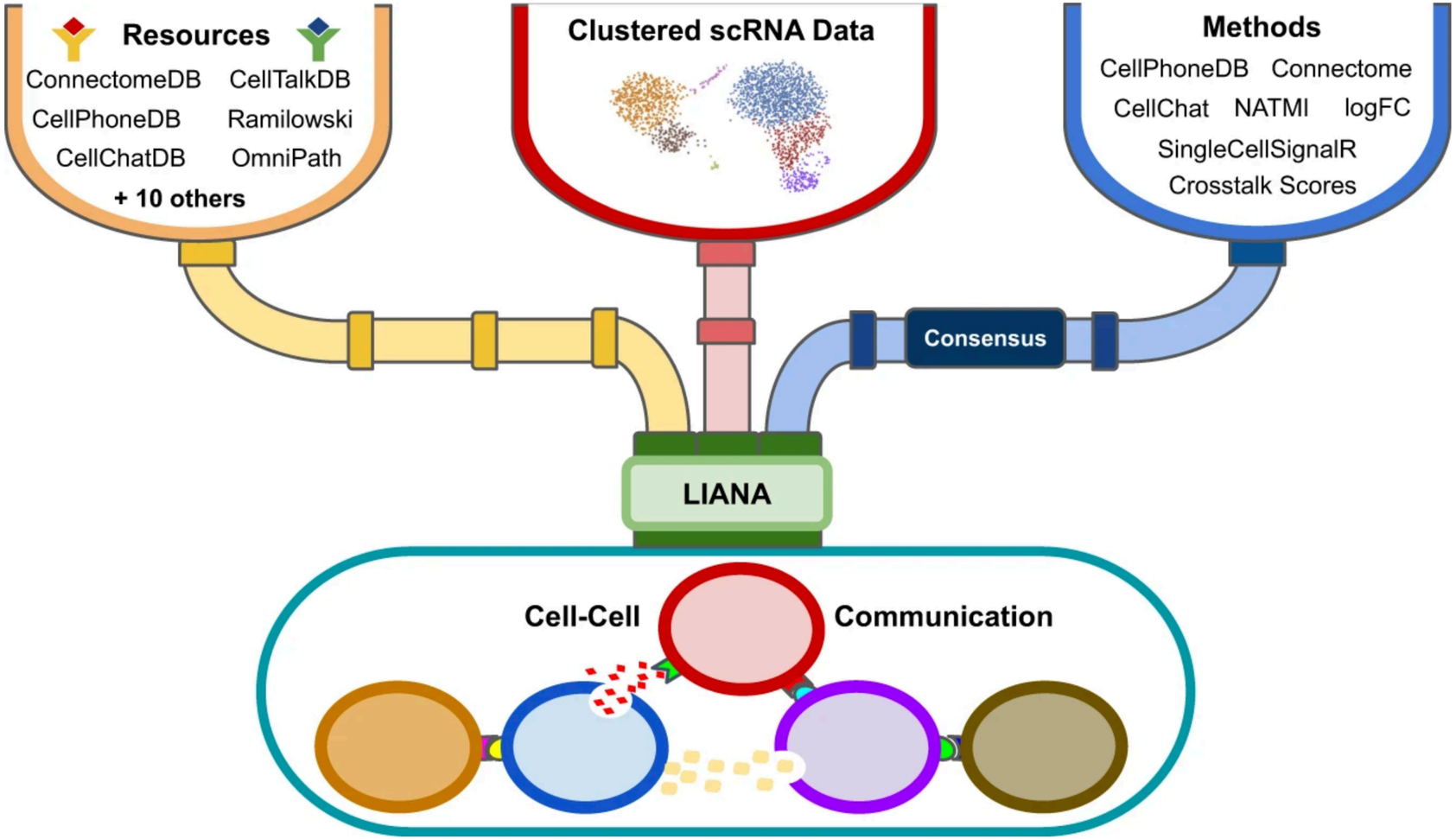
Nichenet
linking ligands to target genes



Browaeys, R., et al. Nature Methods (2020)

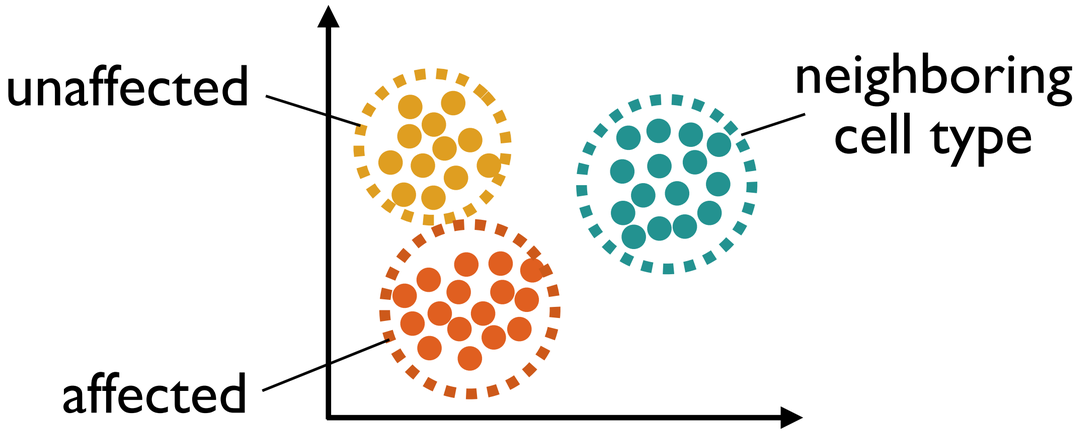
Lack of consensus between repositories and computational tools

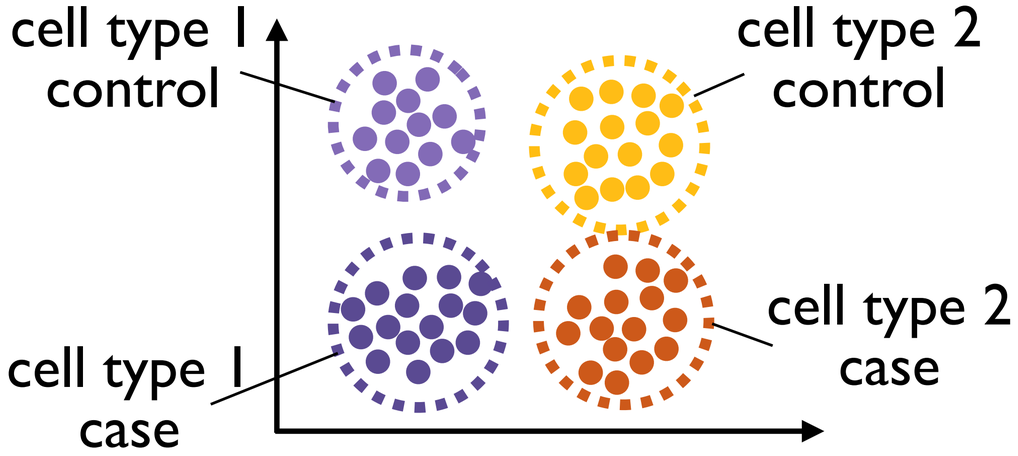
LIANA allows all combinations of resources and methods



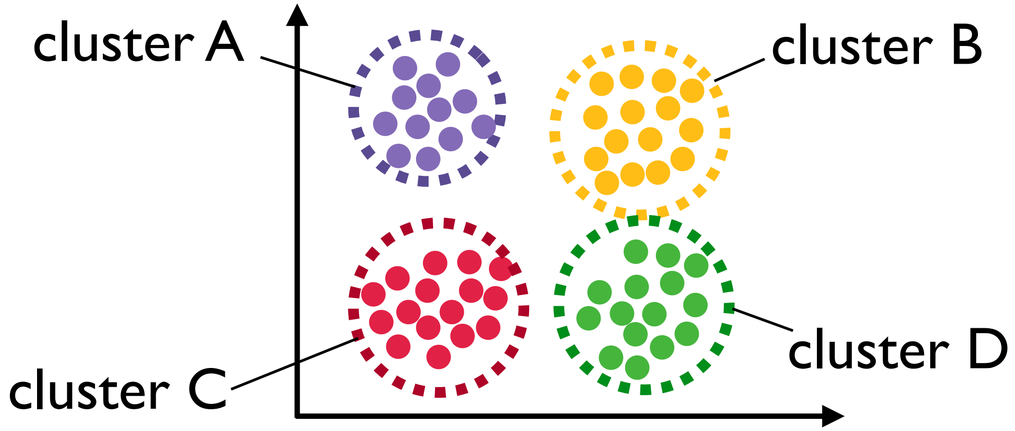
Dimitrov, D., et al. Nature Communications (2022)

Cell-cell communication methods are not built with our targeted contrasts in mind



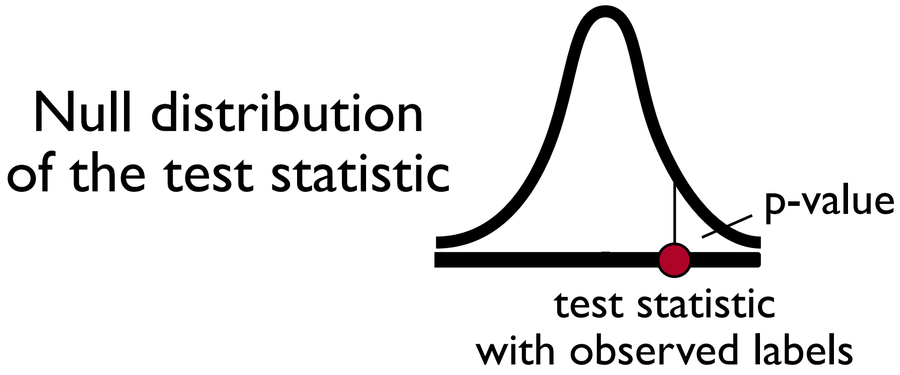


Take CellphoneDB as an example

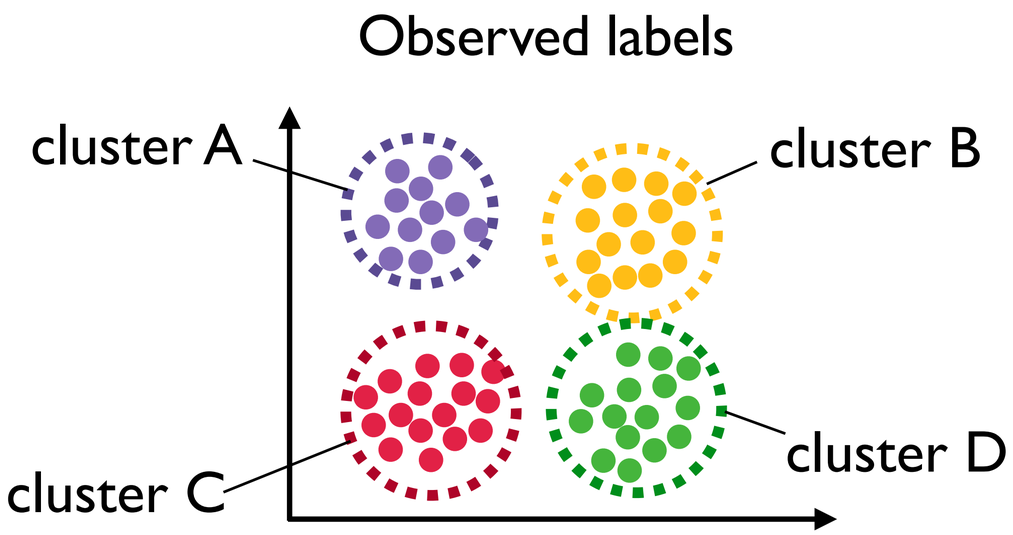


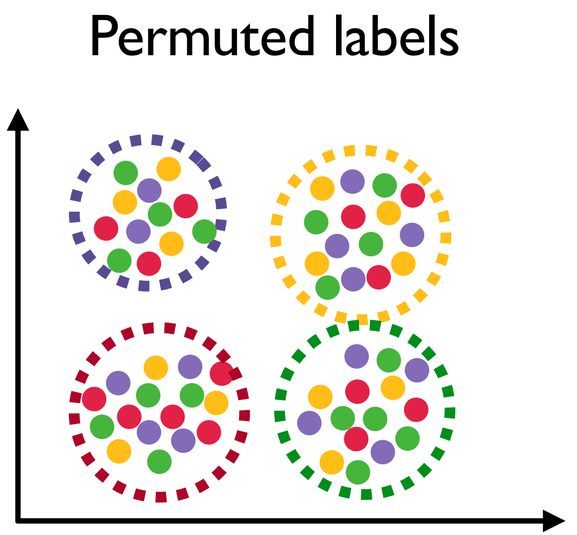
For any given ordered ligand-receptor pair (L, R) and an ordered pair of clusters (clA, clB):

$$\text{Test Statistic} = \text{mean}(\bar{L}_{clA}, \bar{R}_{clB})$$



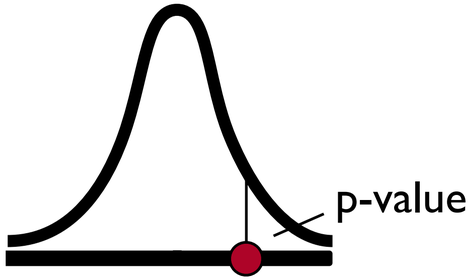
Permute labels to generate a null distribution





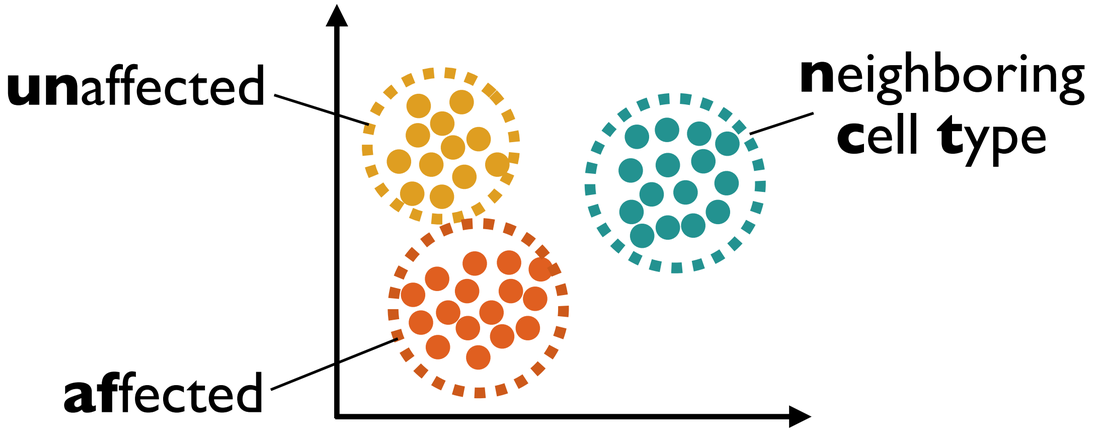
This tests whether this (L, R) interaction is significantly enriched in (clA, clB) relative to any other ordered pair of clusters.

Test Statistic = $\text{mean}(\bar{L}_{clA}, \bar{R}_{clB})$



Empirical distribution of the test statistic under permutation of all labels

Contrast cell-cell interactions of **af**fected and **un**affected cells with a **n**eighboring **c**ell **t**ype



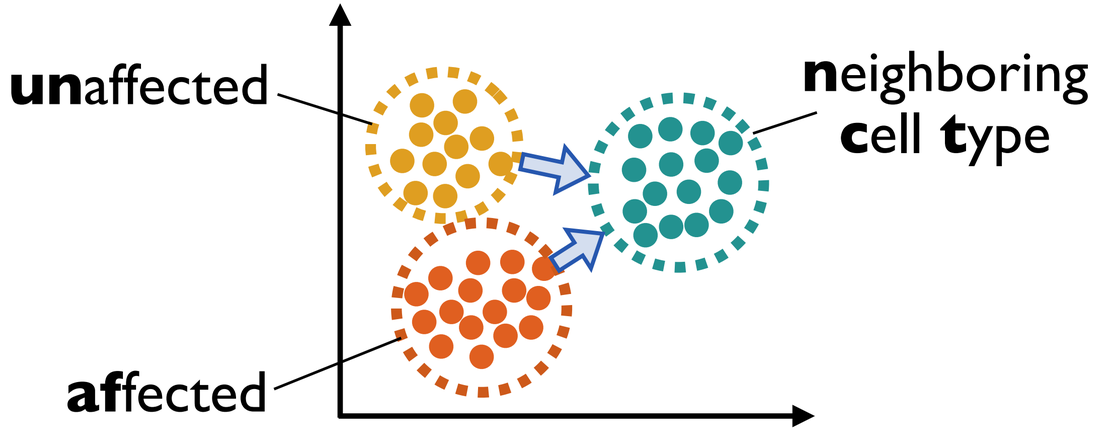
Test for each ordered (L, R) pair if
the interaction between
(**un** with **nct**) and (**af** with **nct**)
is significantly different.

$$H_o: \text{un} \rightarrow \text{nct} \equiv \text{af} \rightarrow \text{nct}$$

$$H_o: \text{nct} \rightarrow \text{un} \equiv \text{nct} \rightarrow \text{af}$$

CellphoneDB builds a general, not targeted, comparison

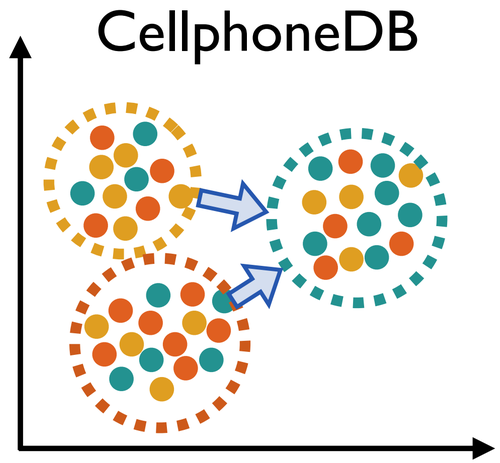
Testing for both directions and comparing to all pairwise cluster combinations

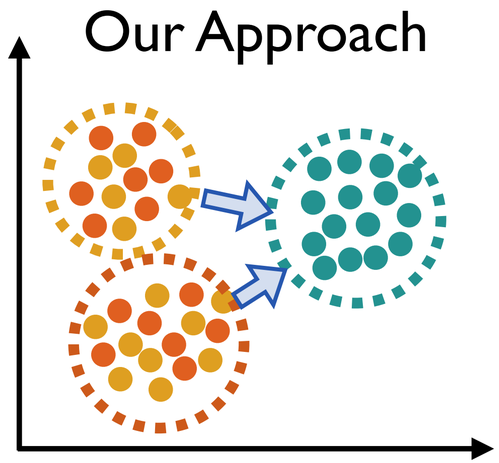


Test for each ordered (L, R) pair if the interaction between (**un** with **nct**) and (**af** with **nct**) is significantly different.

H_o: un → nct ≡ af → nct

H_o: nct → un ≡ nct → af





	un	*af*	*nct*
un			
af			
nct			

Test Statistic =

$$\text{mean}(\bar{L}_{un}, \bar{R}_{nct}) - \text{mean}(\bar{L}_{af}, \bar{R}_{nct})$$

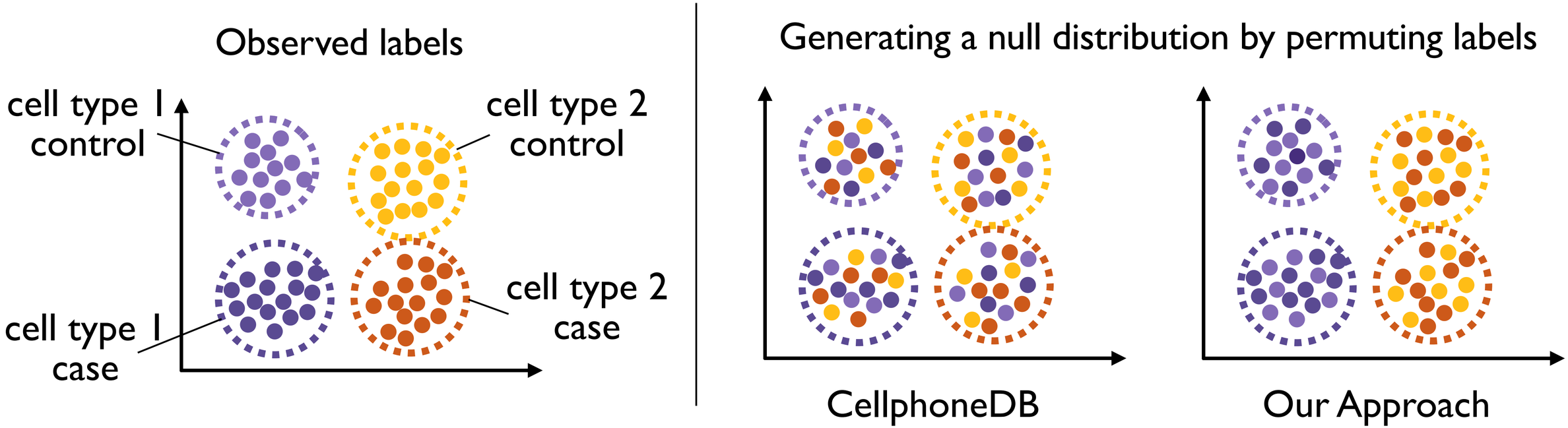
Goeva, A., et al. *[in preparation]*

Our approach generalizes to various cross-condition contrasts

For example, case-control comparisons

$$H_0: ct1 \rightarrow ct2 \equiv ct1 \rightarrow ct2$$

$$H_0: ct2 \rightarrow ct1 \equiv ct2 \rightarrow ct1$$



Goeva, A., et al. *[in preparation]*

Outlook

An inventory of perturbations

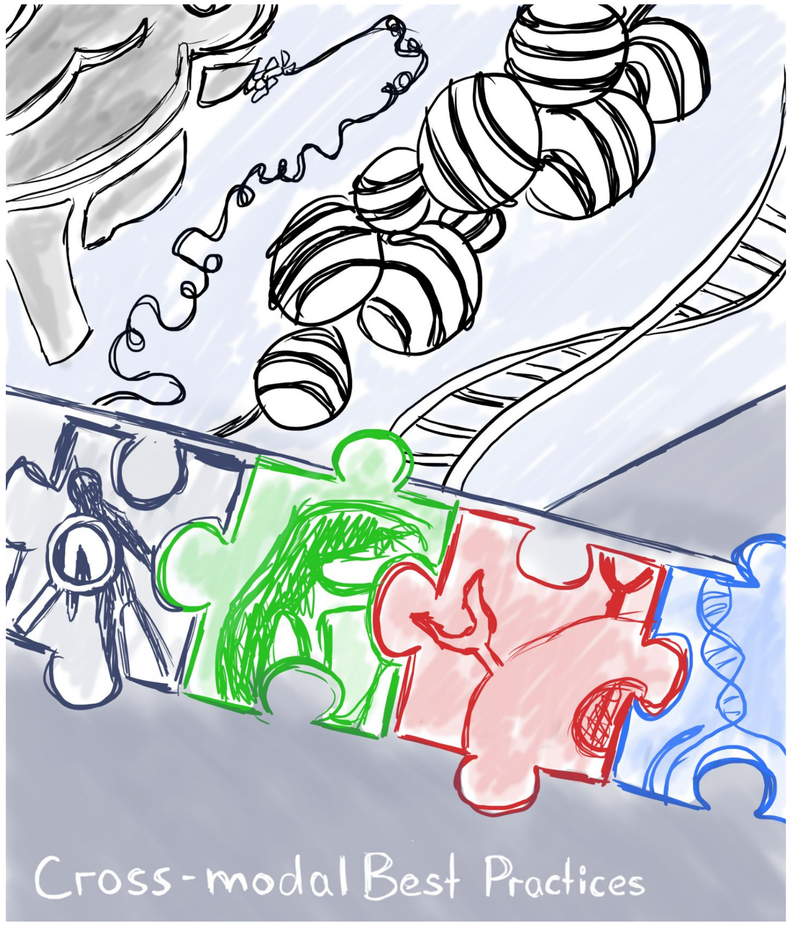
We need large-scale systematic projects that measure and characterize multi-modal cellular changes in response to a comprehensive collection of perturbations so we can relate our datasets to it.

Single cell best practices book

sc-best-practices.org


Lukas Heumos


Anna Schaar



> 40 contributors
> 50 chapters

- Online book with tutorials and best practices for single cell analysis steps
  - Unimodal and multimodal single cell analysis of RNA, ATAC, ADT, TCR/BCR and spatial omics
  - Preprocessing to advanced downstream analysis
  - Suggestions based on independent benchmarks
  - All chapters downloadable as Jupyter Notebooks
- Living resource for and with the community
  - Contribute content
  - Provide feedback
  - Work in progress…

github.com/theislab/single-cell-best-practices

lukas.heumos@helmholtz-munich.de

anna.schaar@helmholtz-munich.de