Single Cell Plus, Banff International Research Station, 05 July 2023

Universität Zürich^{UZH}

URPP Evolution in Action

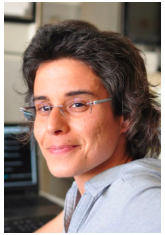
SIB Swiss Institute of Bioinformatics

Benchmarking computational methods for single cell and spatial transcriptomics data: anecdotes, questions and OMNIBENCHMARK

Mark D. Robinson
Statistical Bioinformatics Group, DMLS@UZH+SIB

@markrobinsonca

https://robinsonlabuzh.github.io/



Izaskun Mallona

Charlotte Soneson

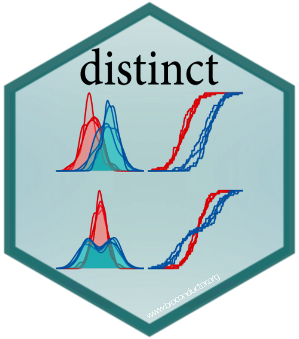
Almut Luetge

Anthony Sonrel

Single cell data tool spectrum

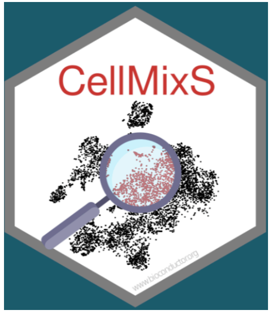
diffcyt = Differential discovery in high-dim cytometry (via high-resolution clustering)

treeclimbR = pinpoint the data-dependent resolution on hierarchical hypotheses.



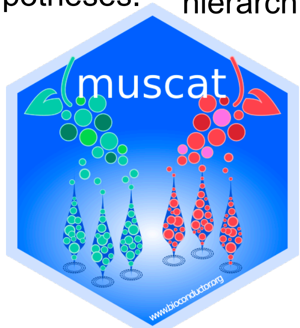
distinct = differential distribution analysis via hierarchical permutation tests

CATALYST = Cytometry dATa anALYSis Tools



CellMixS = Evaluate cell-specific mixing (batch correction)

scDblFinder = doublet detection for scRNA-seq data



muscat = multi-sample multi-group scRNA-seq data analysis tools

pipeComp = comparison of pipelines involving various steps and parameters

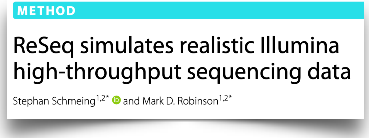
SampleQC = robust multivariate, multi-celltype, multi-sample quality control for single cell data

censcyt = diff. abundance analysis with a right-censored covariates in high-dim cytometry

Theme: infrastructure + benchmarking

iSEE = interactive (Shiny-based) SummarizedExperiment explorer

iCOBRA = interactive comparative evaluation of binary classification and ranking methods

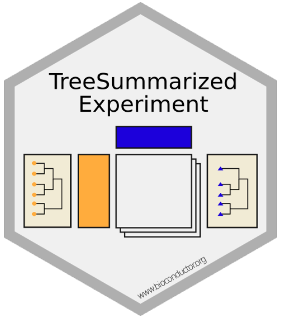

METHOD

ReSeq simulates realistic Illumina high-throughput sequencing data

Stephan Schmeing[1,2*] and Mark D. Robinson[1,2*]

ReSeq = authentic synthetic sequencing data

SpatialExperiment = data structure for Spatially Resolved Transcriptomics Data



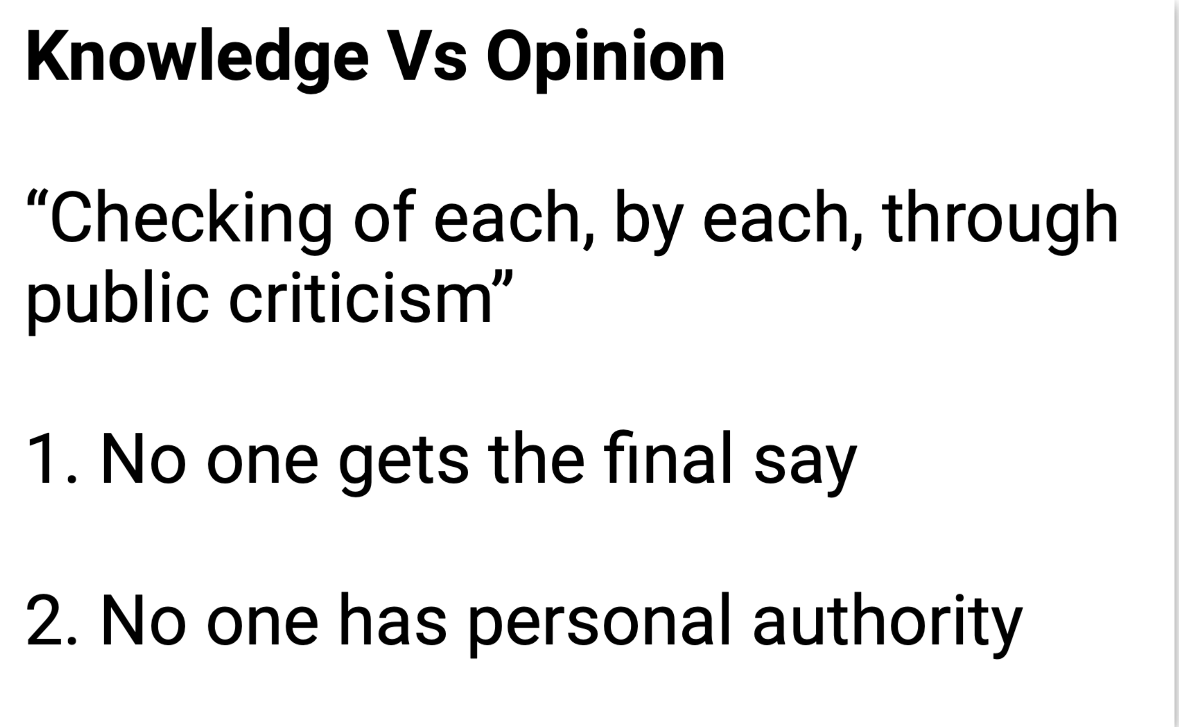
TreeSummarized-Experiment = data structure for Data with Tree Structures

OMB = OMNIBENCHMARK framework for general benchmarking

What I think about when I see (talks/ papers that include) benchmarks

- What are the metrics for success?
- Are the simulations reasonable?
- Could I reproduce this benchmark result?
- To what data (and for how long) are these benchmark results valid?

The philosophy of benchmarking?



Knowledge Vs Opinion

“Checking of each, by each, through public criticism”

1. No one gets the final say

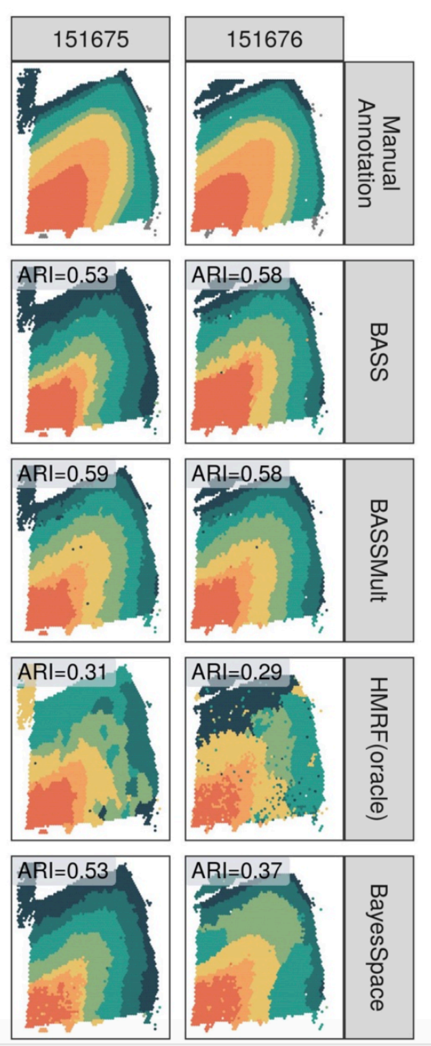
2. No one has personal authority

Talk by Marcel Salathé,
EPFL Open Science Day 2019

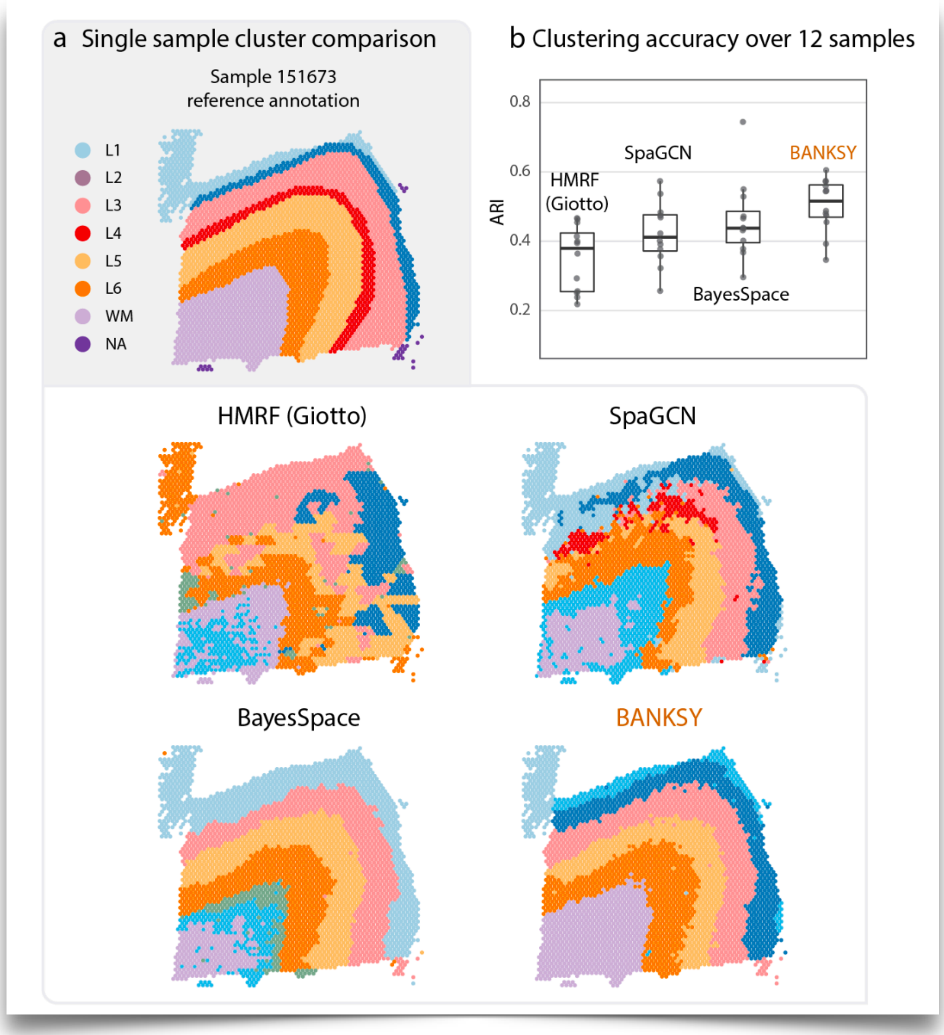
"Kindly Inquisitors: The New Attacks on Free Thought" by J. Rauch.

Benchmarking anecdote 1:
if multiple people compare a method, they’ll get roughly the same results, right?

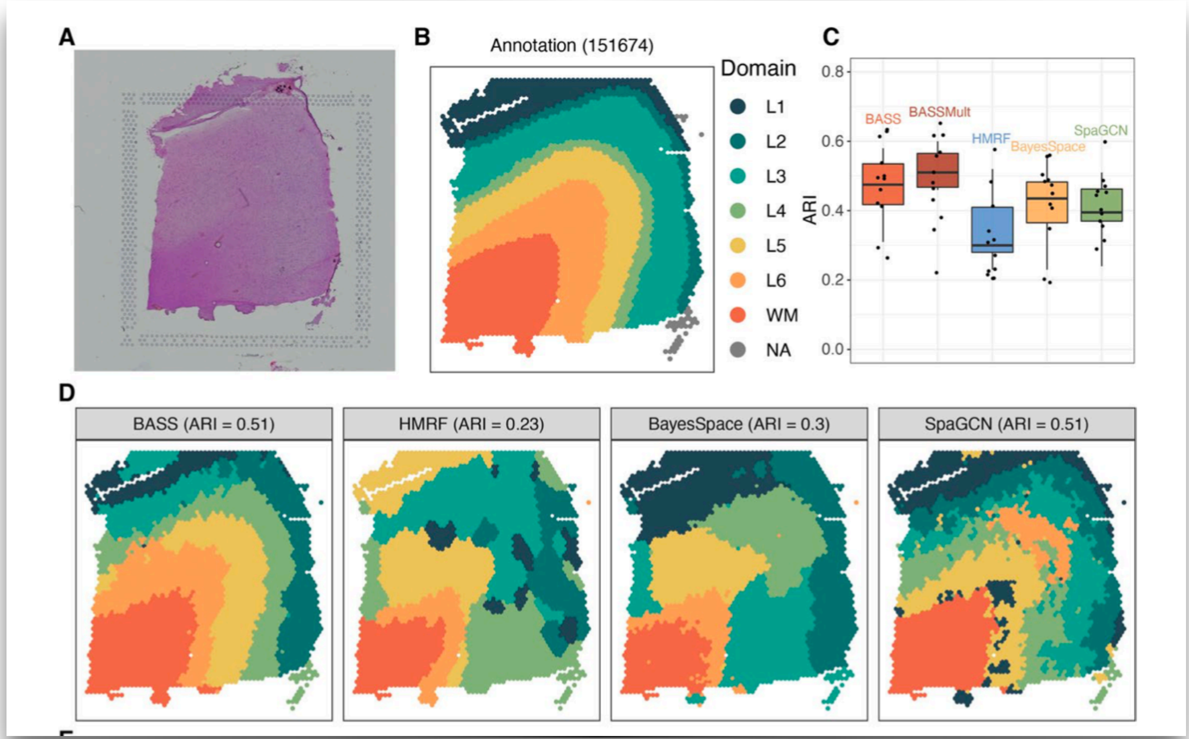
{same data, same method} in the hands of many.



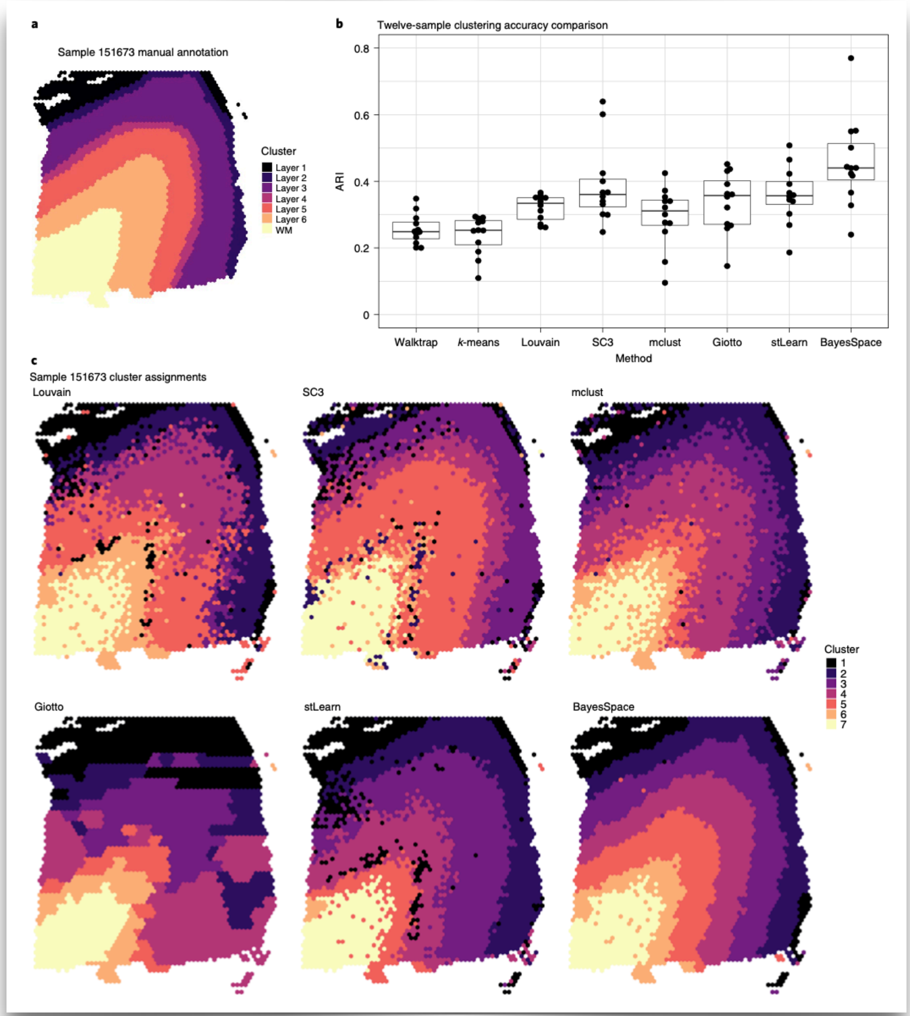
{same data, same method} in the hands of many.



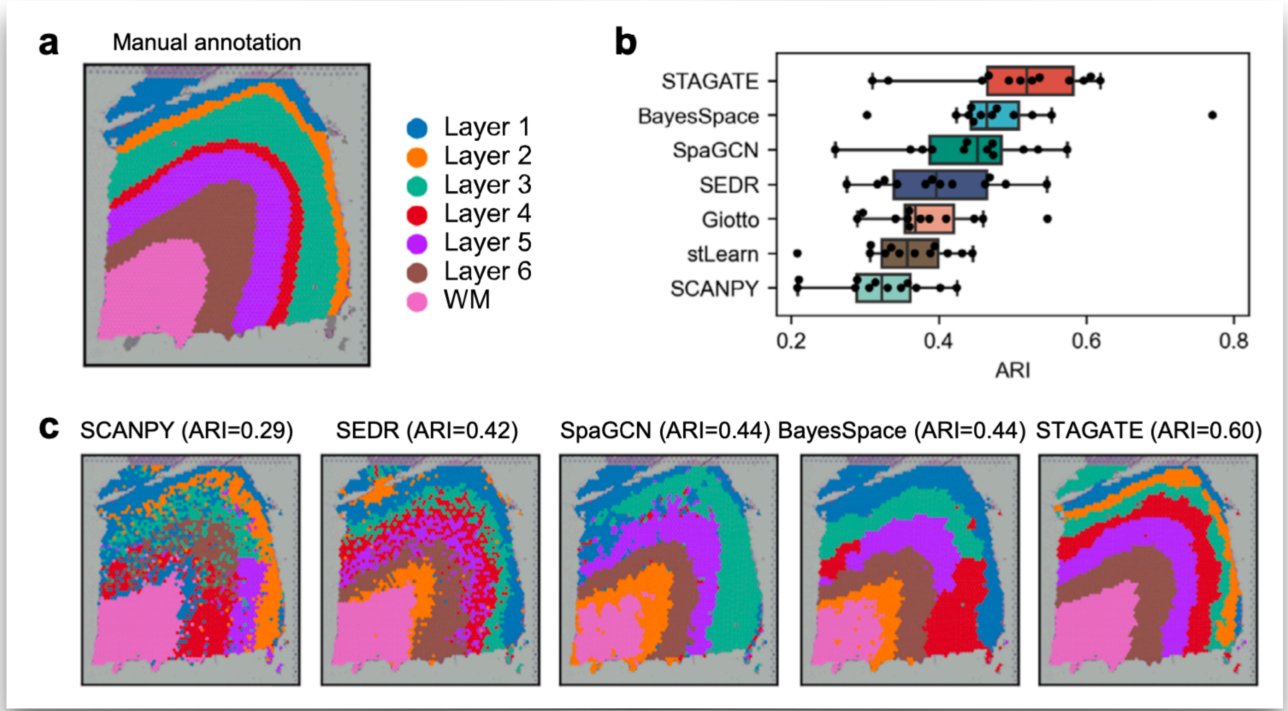
{same data, same method} in the hands of many.



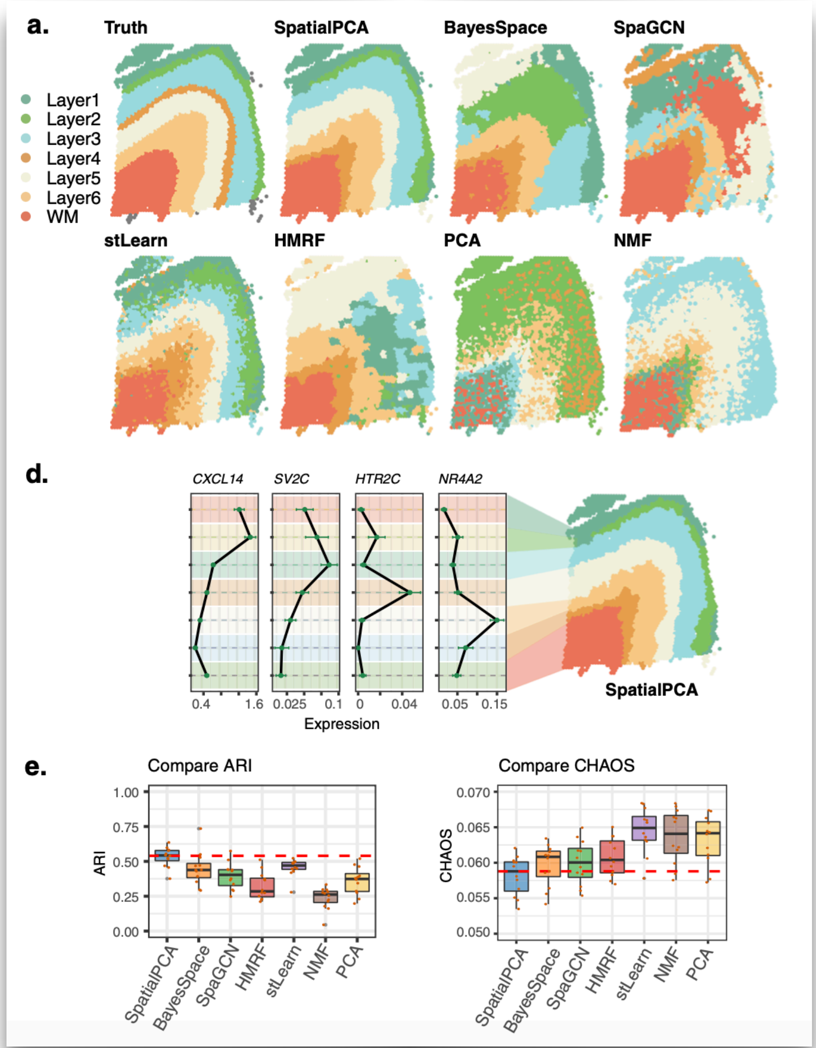
{same data, same method} in the hands of many.



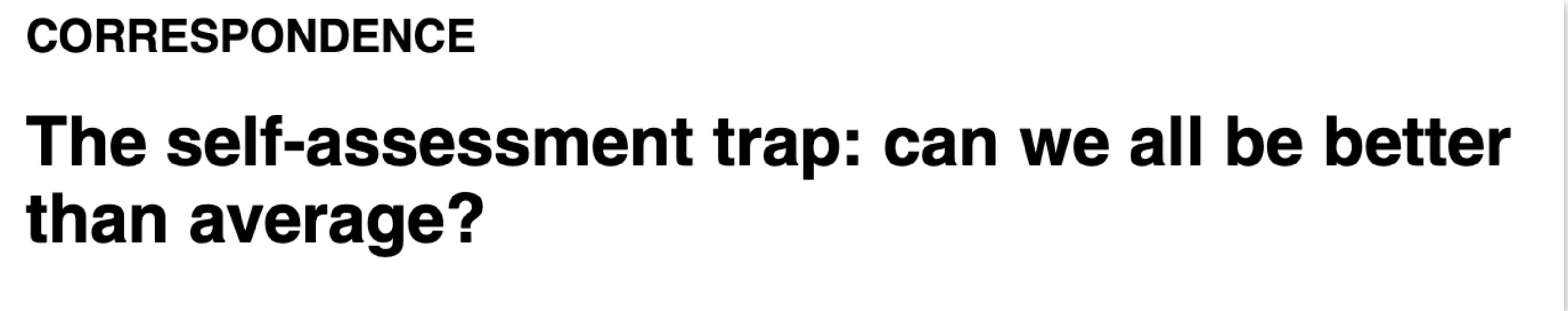
{same data,
same method}
in the hands
of many.



{same data, same method} in the hands of many.



I think the main tension is ..



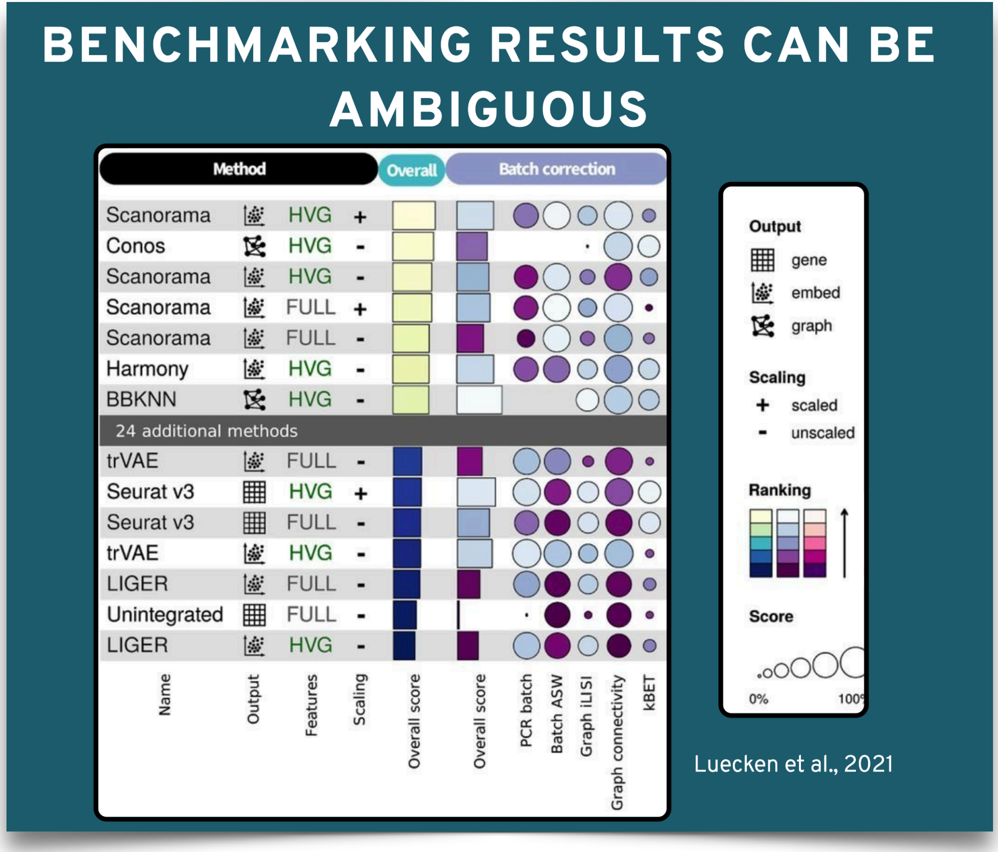
CORRESPONDENCE

The self-assessment trap: can we all be better than average?

"*researchers wishing to publish their analytical methods are required by referees to compare the performance of their own algorithms against other methodologies, thus being forced to be judge, jury and executioner. The result is that the authors' method tends to be the best ..*" (from 2011!)

Benchmarking anecdote 2: do we know how to quantify performance?

Do multiple metrics agree? (e.g., batch correction)

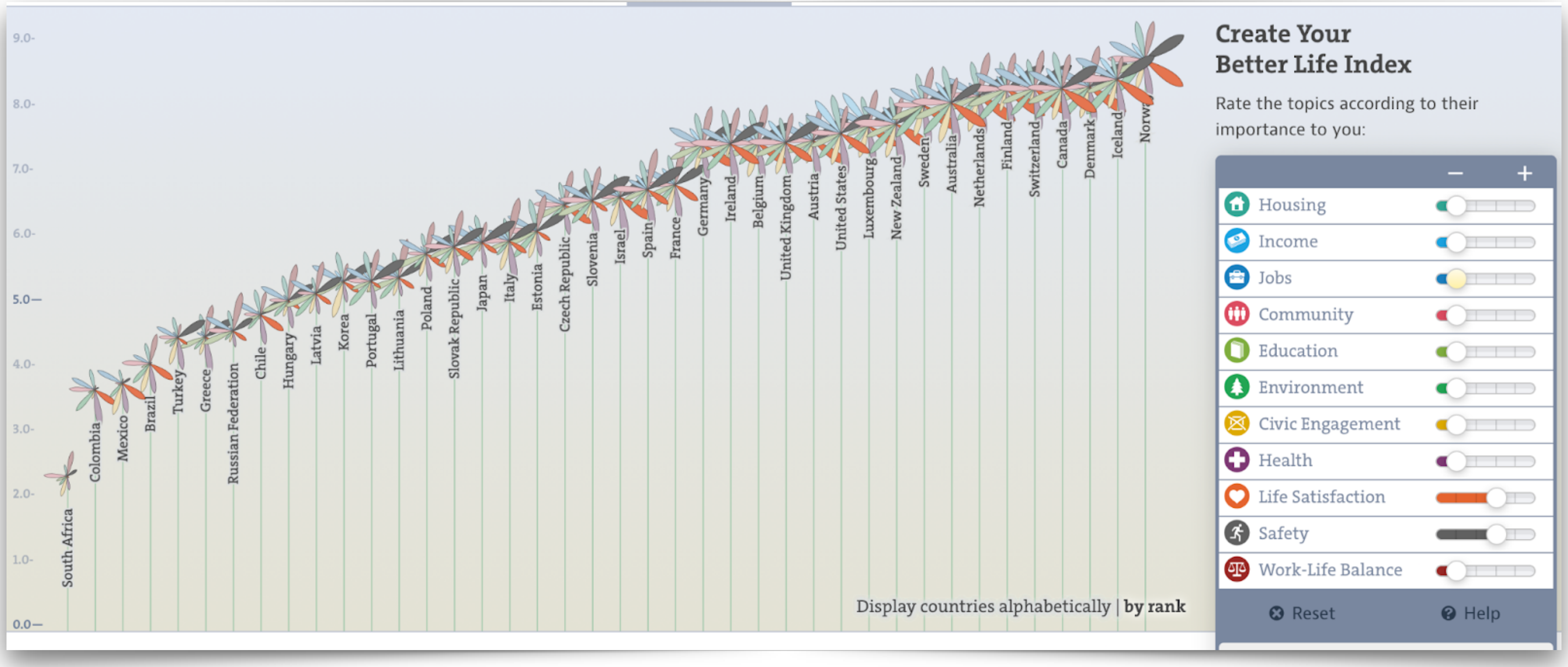


BENCHMARKING RESULTS CAN BE AMBIGUOUS

Name	Output	Features	Scaling
Scanorama	⁘	HVG	+
Conos	✢	HVG	-
Scanorama	⁘	HVG	-
Scanorama	⁘	FULL	+
Scanorama	⁘	FULL	-
Harmony	⁘	HVG	-
BBKNN	✢	HVG	-
24 additional methods			
trVAE	⁘	FULL	-
Seurat v3	▦	HVG	+
Seurat v3	▦	FULL	-
trVAE	⁘	HVG	-
LIGER	⁘	FULL	-
Unintegrated	▦	FULL	-
LIGER	⁘	HVG	-

Luecken et al., 2021

“Better tool index” for computational biology?



Create Your Better Life Index

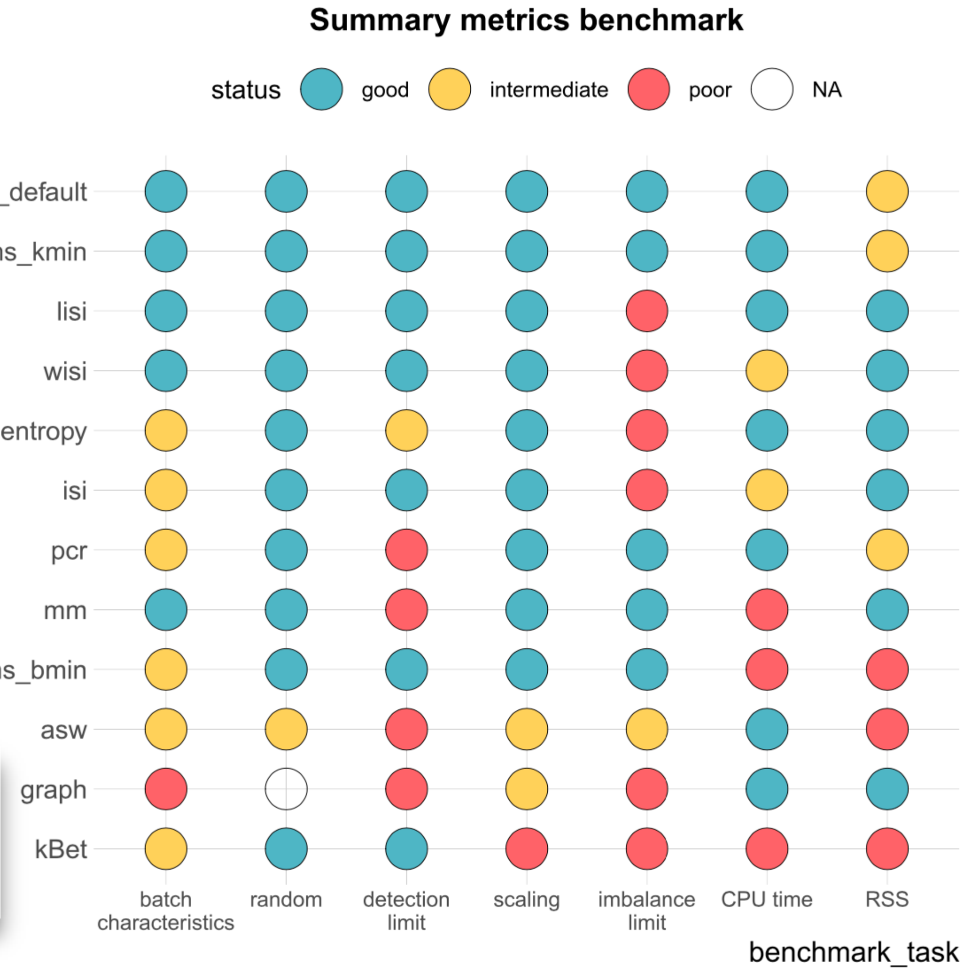
Rate the topics according to their importance to you:

- Housing
- Income
- Jobs
- Community
- Education
- Environment
- Civic Engagement
- Health
- Life Satisfaction
- Safety
- Work-Life Balance

https://www.oecdbetterlifeindex.org/

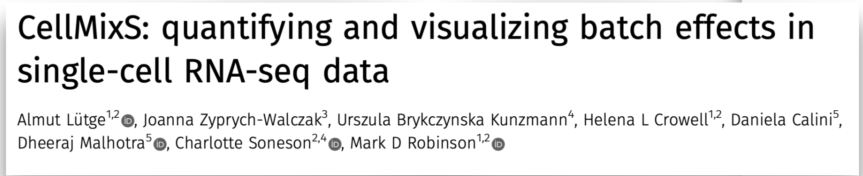
Some metrics are better than others

Almut



Summary metrics benchmark

metric	batch characteristics	random	detection limit	scaling	imbalance limit	CPU time	RSS
cms_default	good	good	good	good	good	good	intermediate
cms_kmin	good	good	good	good	good	good	intermediate
lisi	good	good	good	good	poor	good	good
wisi	good	good	good	good	poor	intermediate	good
entropy	intermediate	good	intermediate	good	poor	good	good
isi	intermediate	good	good	good	poor	intermediate	good
pcr	intermediate	good	poor	good	good	good	intermediate
mm	good	good	poor	good	good	poor	good
cms_bmin	intermediate	good	good	good	good	poor	poor
asw	intermediate	intermediate	poor	intermediate	intermediate	good	poor
graph	poor	NA	poor	intermediate	poor	good	good
kBet	intermediate	good	good	poor	poor	poor	poor

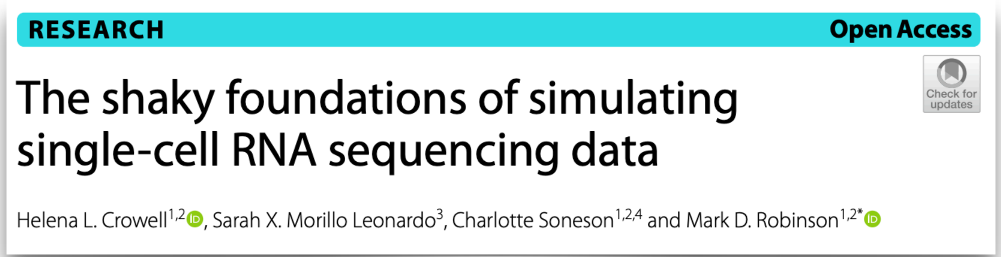


CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq data

Almut Lütge[1,2], Joanna Zyprych-Walczak[3], Urszula Brykczynska Kunzmann[4], Helena L Crowell[1,2], Daniela Calini[5], Dheeraj Malhotra[5], Charlotte Soneson[2,4], Mark D Robinson[1,2]

Benchmarking anecdote 3: are simulations good?

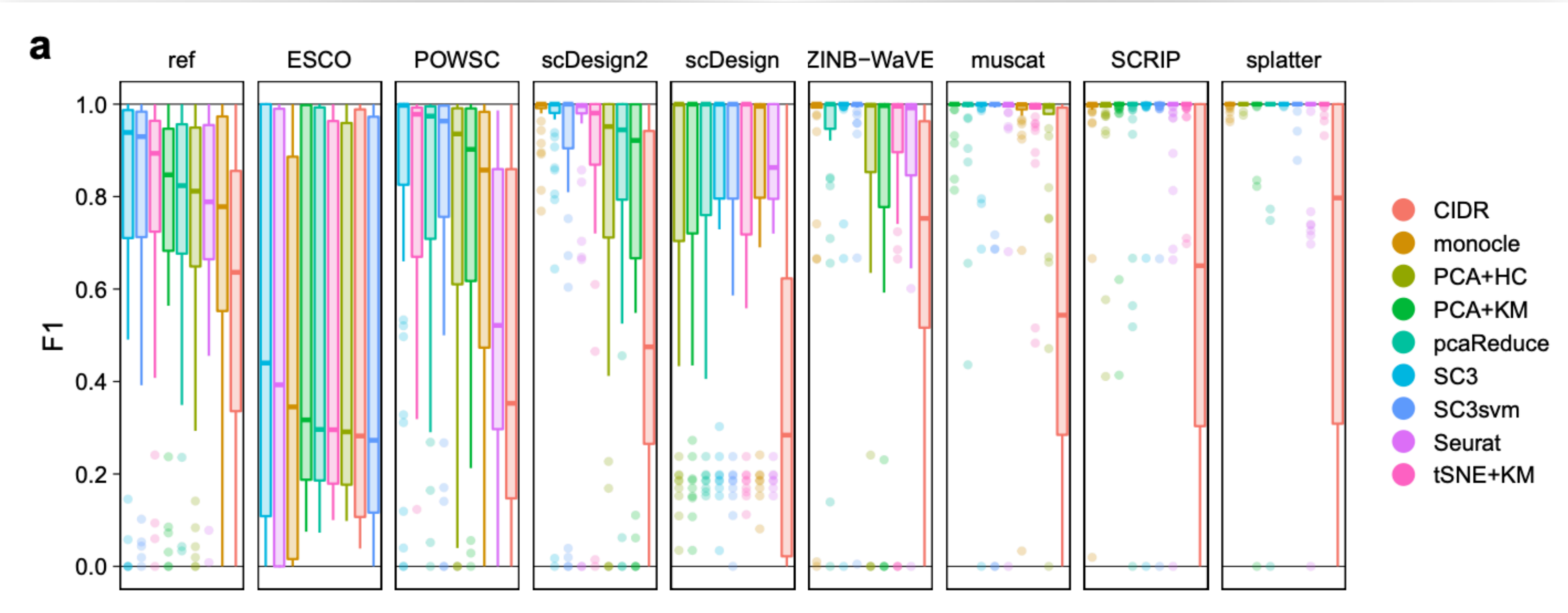
Do results on simulated data reflect results from real datasets?



RESEARCH Open Access

The shaky foundations of simulating single-cell RNA sequencing data

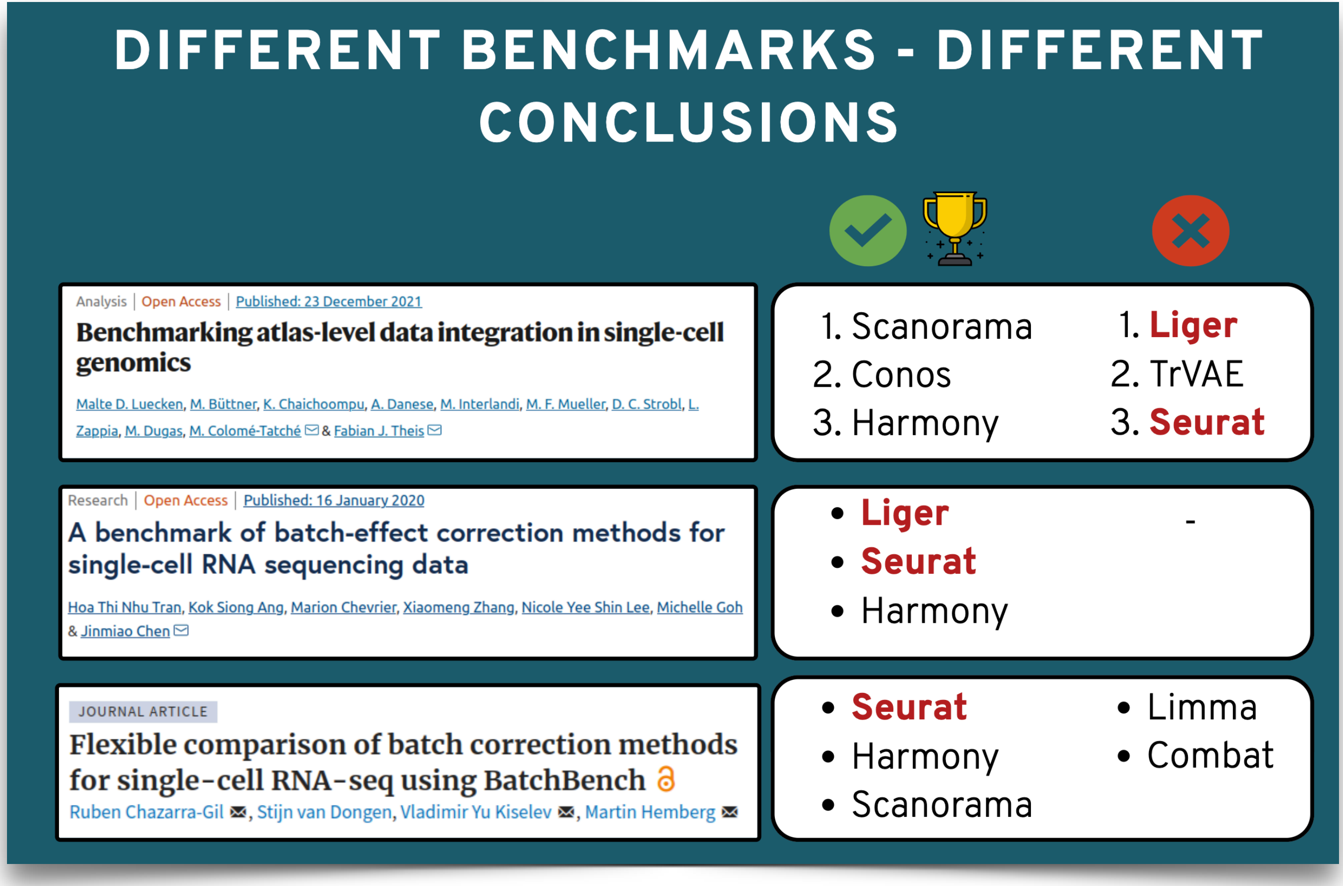
Helena L. Crowell[1,2], Sarah X. Morillo Leonardo[3], Charlotte Soneson[1,2,4] and Mark D. Robinson[1,2*]



Helena

Benchmarking anecdote 4: do multiple benchmarks agree?

Do multiple benchmarks agree? (e.g., batch correction)

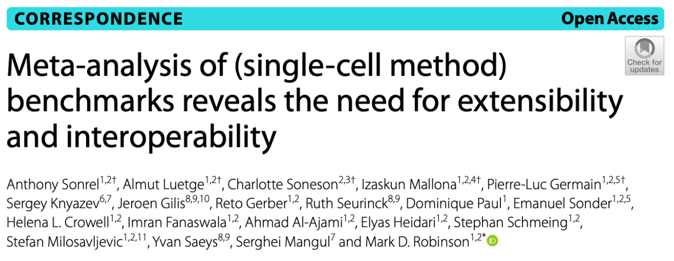


DIFFERENT BENCHMARKS - DIFFERENT CONCLUSIONS

	✓ 🏆	✗
Analysis | Open Access | Published: 23 December 2021 **Benchmarking atlas-level data integration in single-cell genomics** Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché & Fabian J. Theis	1. Scanorama 2. Conos 3. Harmony	1. **Liger** 2. TrVAE 3. **Seurat**
Research | Open Access | Published: 16 January 2020 **A benchmark of batch-effect correction methods for single-cell RNA sequencing data** Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh & Jinmiao Chen	• **Liger** • **Seurat** • Harmony	-
JOURNAL ARTICLE **Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench** Ruben Chazarra-Gil, Stijn van Dongen, Vladimir Yu Kiselev, Martin Hemberg	• **Seurat** • Harmony • Scanorama	• Limma • Combat

Should we benchmark the benchmarks?

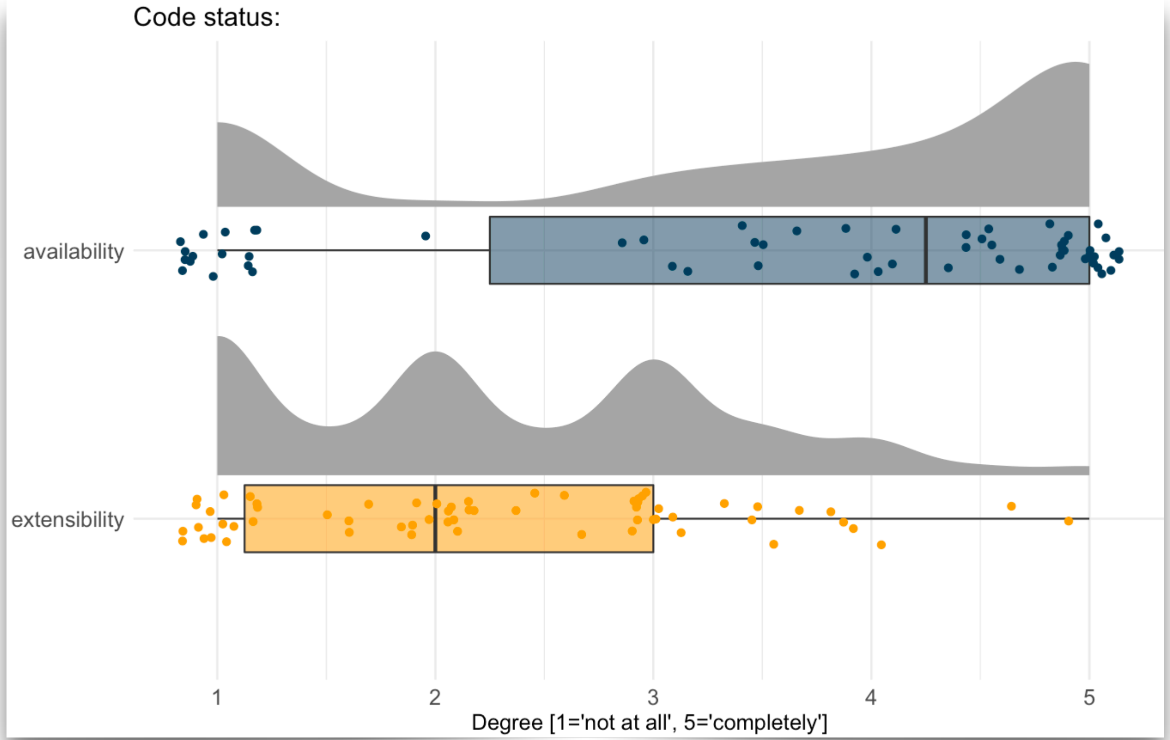
Code availability: good
Code extensibility: not good



CORRESPONDENCE Open Access

Meta-analysis of (single-cell method) benchmarks reveals the need for extensibility and interoperability

Anthony Sonrel[1,2†], Almut Luetge[1,2†], Charlotte Soneson[2,3†], Izaskun Mallona[1,2,4†], Pierre-Luc Germain[1,2,5†], Sergey Knyazev[6,7], Jeroen Gilis[8,9,10], Reto Gerber[1,2], Ruth Seurinck[8,9], Dominique Paul[1], Emanuel Sonder[1,2,5], Helena L. Crowell[1,2], Imran Fanaswala[1,2], Ahmad Al-Ajami[1,2], Elyas Heidari[1,2], Stephan Schmeing[1,2], Stefan Milosavljevic[1,2,11], Yvan Saeys[8,9], Serghei Mangul[7] and Mark D. Robinson[1,2*]

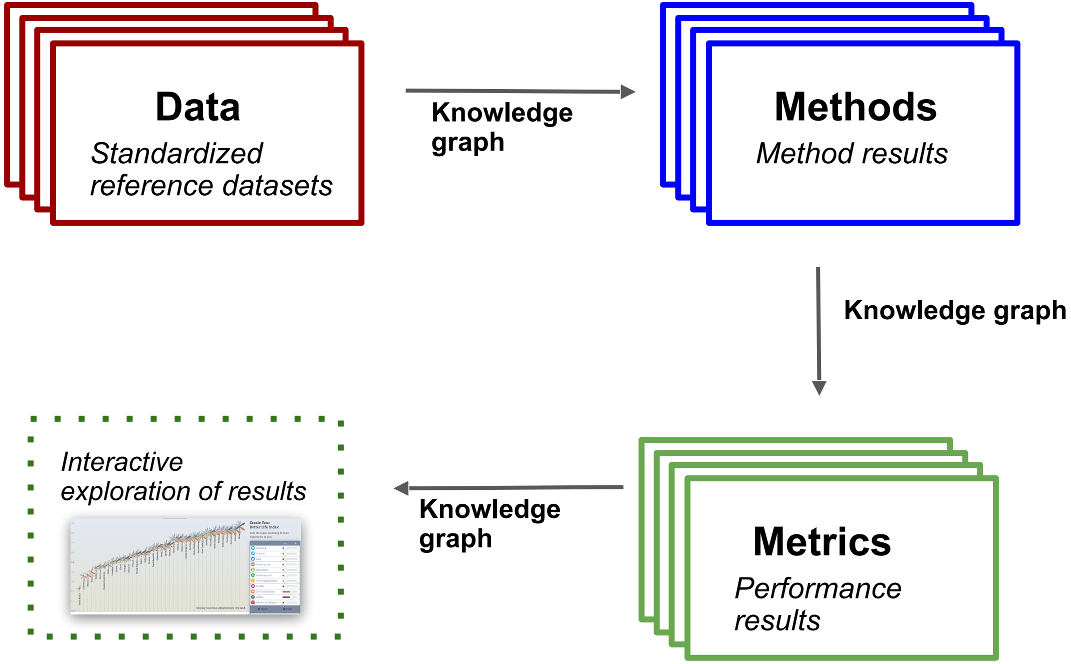


Each dot is a benchmark (62 were surveyed): reviewers' opinions on the extensibility and availability.

“It’s easy to be critical”

.. how about a rethink on benchmark design (open .. continuous .. can crowd-source ..)

OMNIBENCHMARK technical design



Data
Standardized reference datasets

Knowledge graph

Methods
Method results

Knowledge graph

Metrics
Performance results

Knowledge graph

Interactive exploration of results

Repository

GitLab

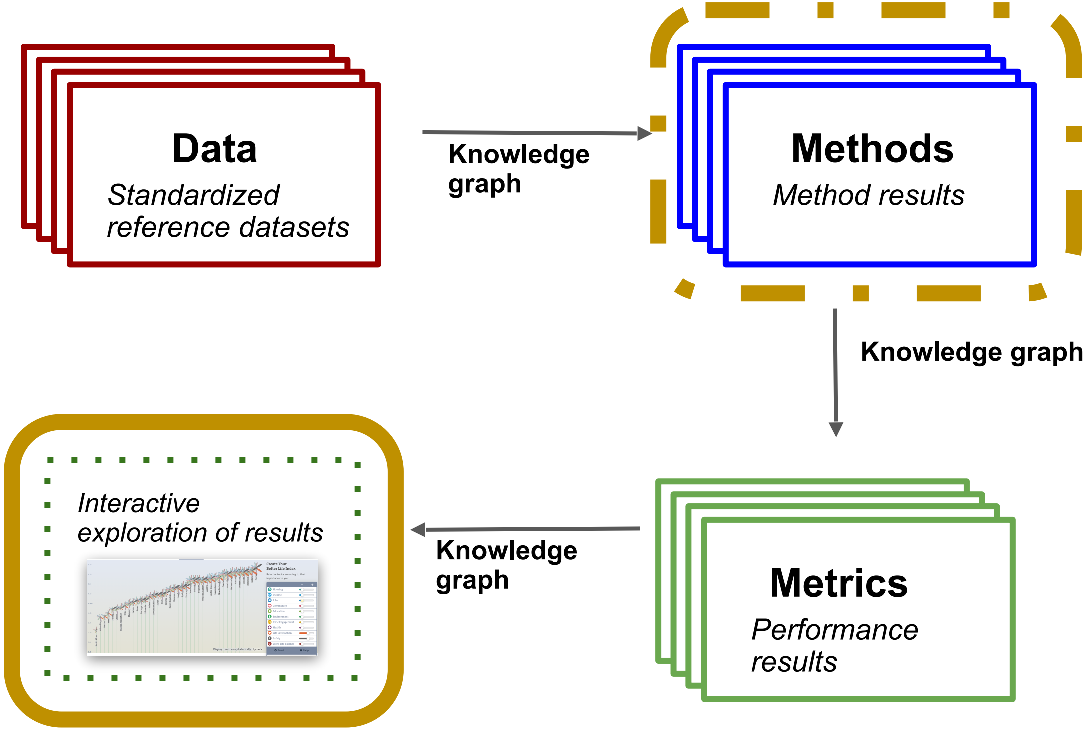
docker

CI/CD

Git LFS

renku

OMNIBENCHMARK users



Data

Standardized reference datasets

Knowledge graph

Methods

Method results

Knowledge graph

Metrics

Performance results

Knowledge graph

Interactive exploration of results

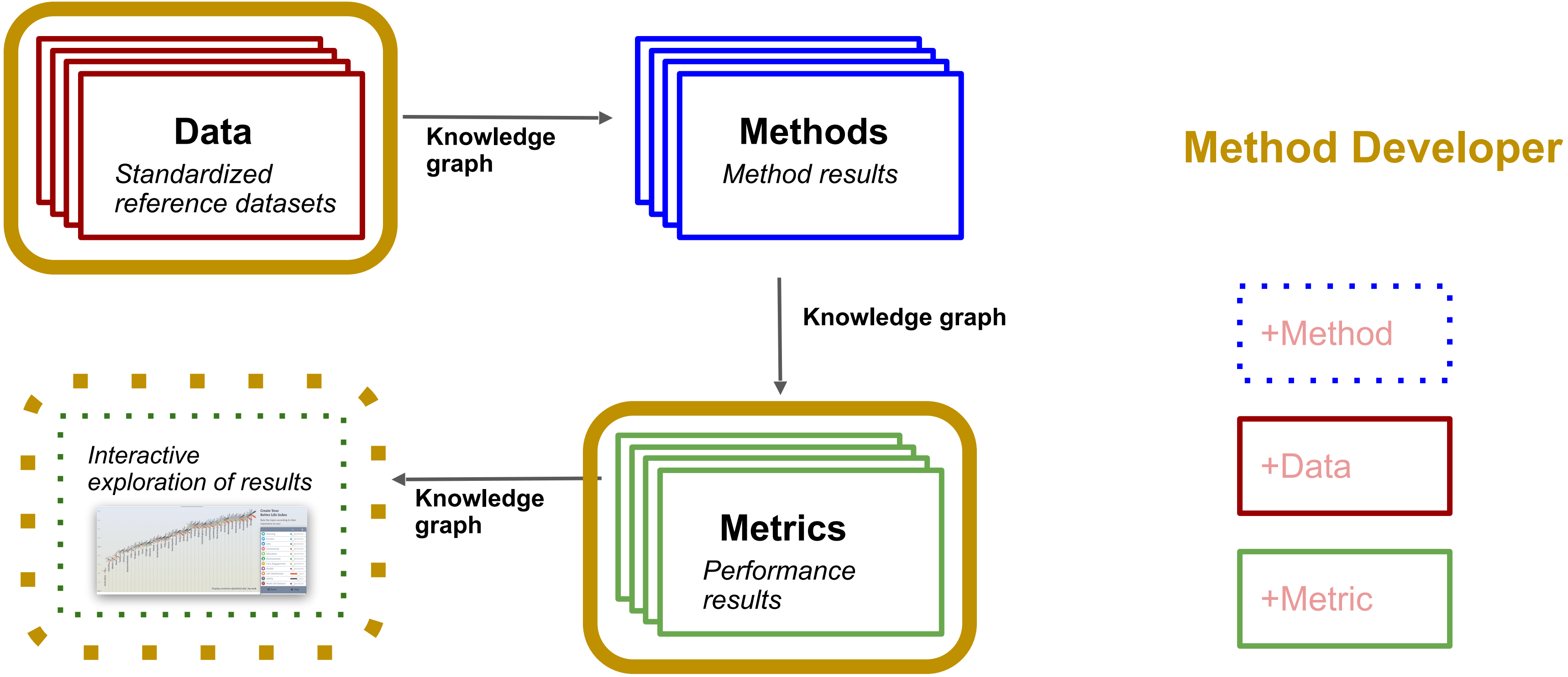
Biologist / Clinician / Data Analyst

+Method

+Data

+Metric

OMNIBENCHMARK users



Data

Standardized reference datasets

Knowledge graph

Methods

Method results

Knowledge graph

Metrics

Performance results

Knowledge graph

Interactive exploration of results

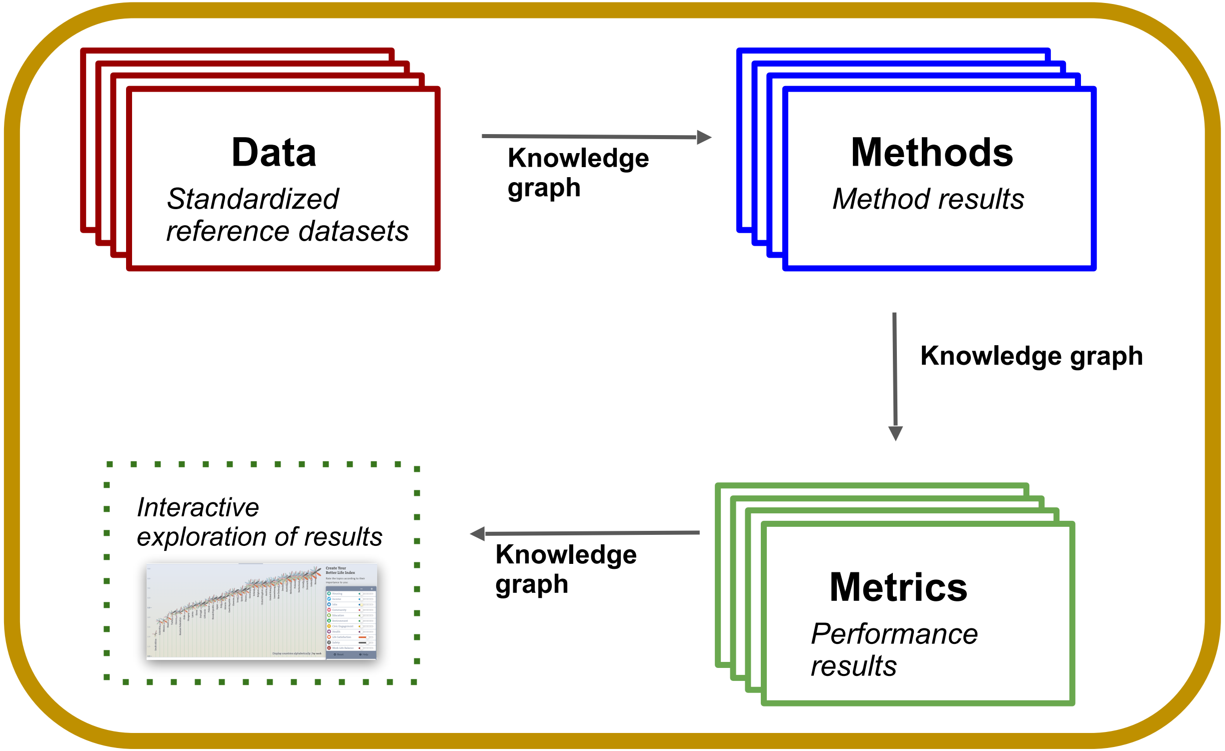
Method Developer

+Method

+Data

+Metric

OMNIBENCHMARK users



Data

Standardized reference datasets

Knowledge graph

Methods

Method results

Knowledge graph

Metrics

Performance results

Knowledge graph

Interactive exploration of results

Benchmarker

Benchmark commons

Discussion points

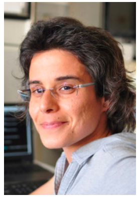
- Method explosion: gets more challenging every day
- Benchmarking is nuanced / difficult to do well; need to establish higher standards
- We don’t always know how to best evaluate methods: “test the tests”
- OMNIBENCHMARK gives a lot for free (transparency, systematization, reproducibility, flexible computing, provenance, efficiency), but steep learning curve
- Community engagement? Crowdsourcing?
- Publishing: continuous benchmark = database update
- Applications beyond computational biology

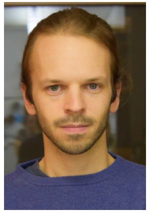
What OMNIBENCHMARK doesn’t do

- does not ensure **high quality** tests of methods (e.g., that simulations are representative), or high quality reference datasets (no standards are imposed, except technical)
- does not manage authority / gate-keeping (quality assurance, recognizing contributions)
- communities → ELIXIR, hackathons

Statistical Bioinformatics Group, DMLS, UZH

CURRENT MEMBERS


Izaskun

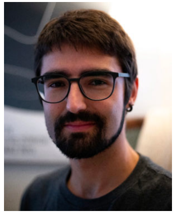

Pierre-Luc

<u>MSc / rotation / visitors:</u>
Frederik
Jiayi
Nidhi
Giulia
Ming
Sam

Helena

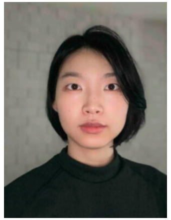
Almut

Anthony

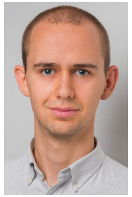

Reto

Martin

Samuel


Peiying

Emanuel


David

Vlad

Siyuan

Yin

Universität Zürich UZH

SIB Swiss Institute of Bioinformatics

University of Zurich UZH
URPP Evolution in Action: From Genomes to Ecosystems

Chan Zuckerberg Initiative

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Roche