Systematic benchmarking of omics computational tools

Serghei Mangul, Ph.D
Department of Clinical Pharmacy
USC School of Pharmacy

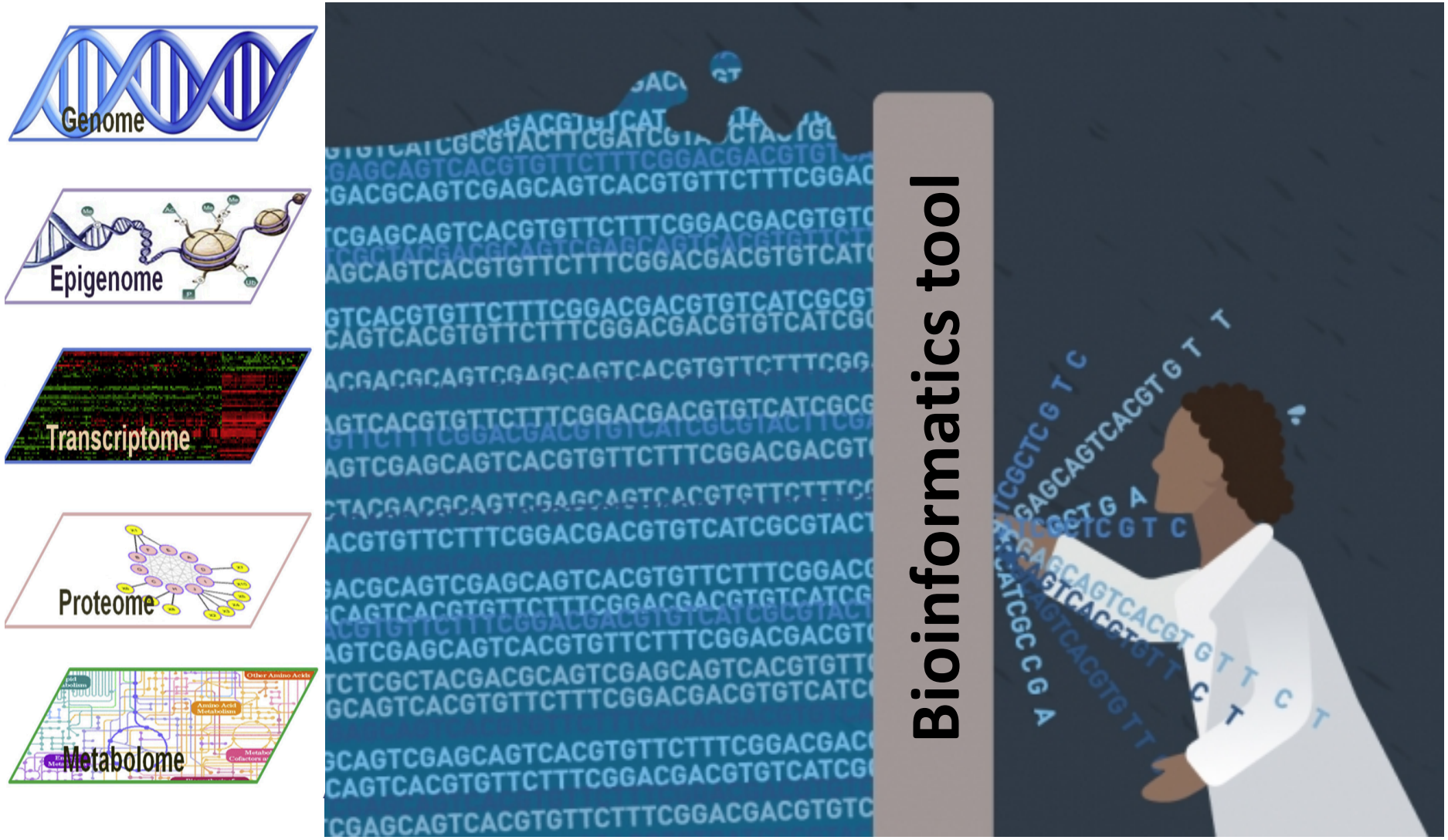
Review Article | OPEN | Published: 27 March 2019

Systematic benchmarking of omics computational tools

Serghei Mangul ✉, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin & Jonathan Flint

Nature Communications **10**, Article number: 1393 (2019) | Download Citation ↓

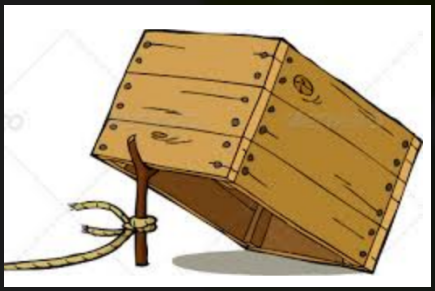
Flood of genomic data



S. Mangul "Interpreting and integrating big data in the life sciences." *Emerging Topics in Life Sciences* (2019)

How to choose the best tool?

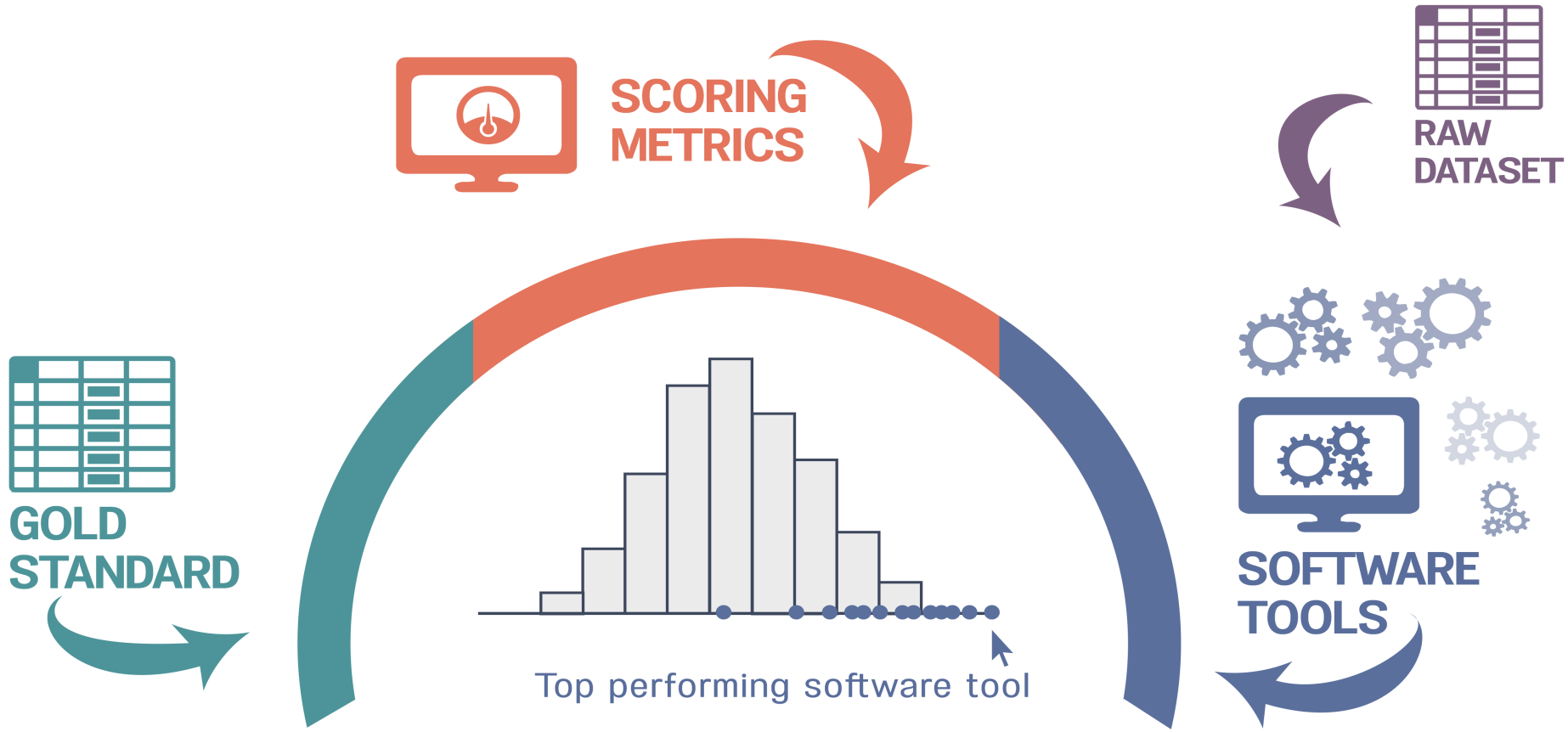
- Assessment of a newly-published algorithm is typically performed by the researchers who develop the tool
- Can each published tools be the best performing tool?



self-assessment trap

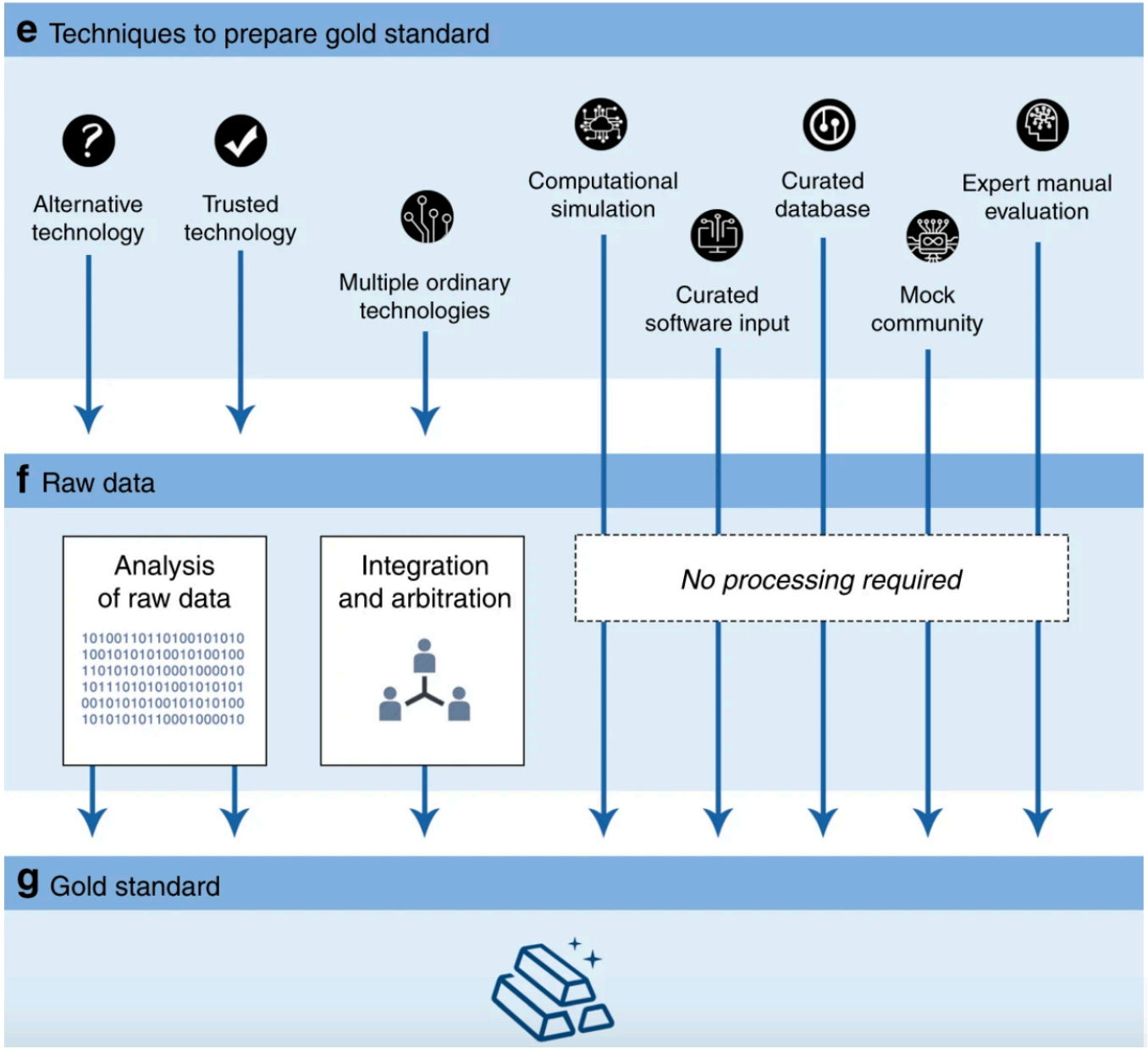
Norel, Raquel, et all. *Molecular systems biology* 7.1 (2011): 537.

Solution: systematic benchmarking

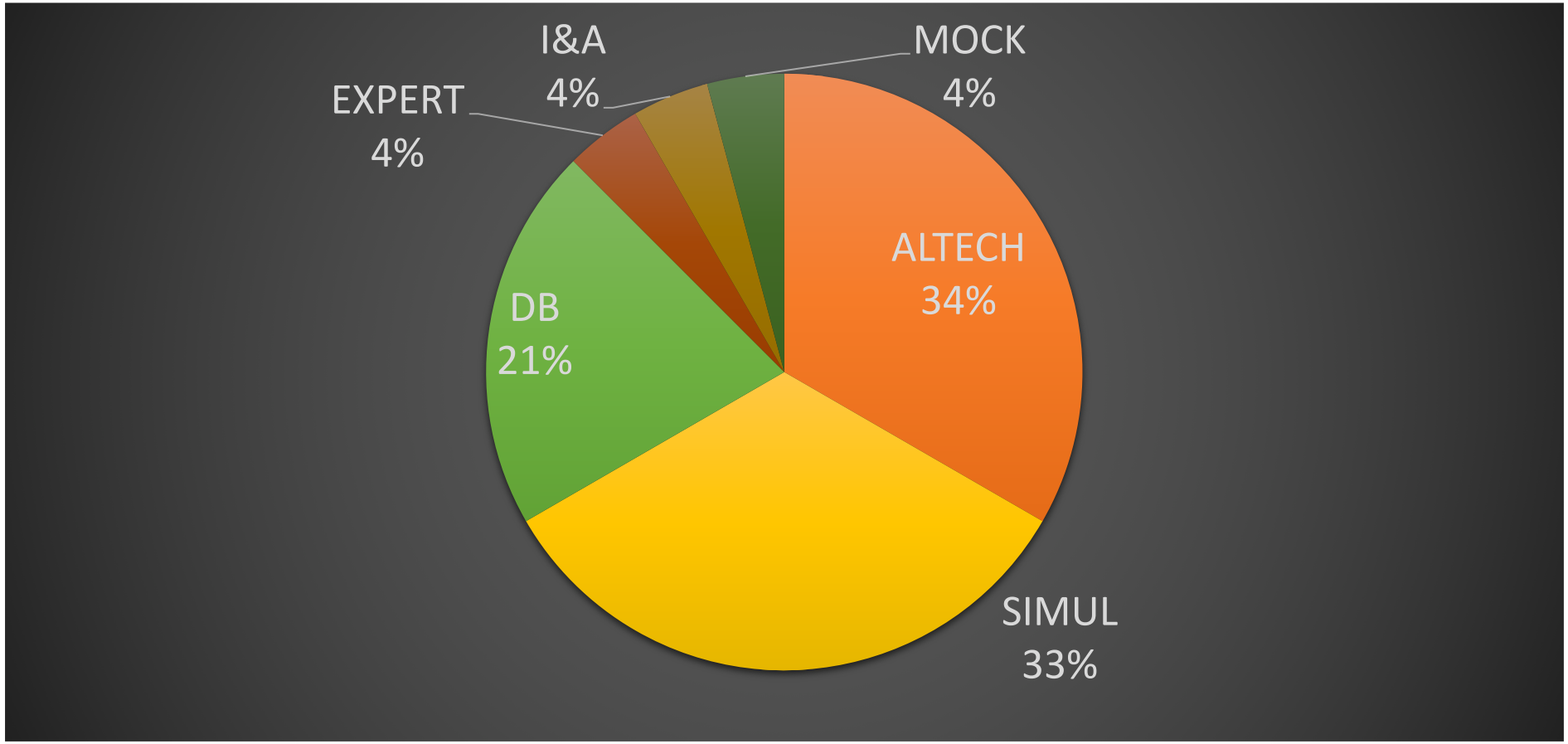


Benchmarking – a robust and comprehensive evaluation of the capabilities of existing algorithms to solve a particular computational biology problem

How to prepare gold standard?



Which techniques are the most popular?



*based on 25 benchmarking studies published across 10 relevant peer-reviewed journals from 2011 to 2017

TECHNIQUE	ADVANTAGES	LIMITATIONS
Trusted technology	High accuracy Direct, usually, no computational inference is required	Carries high cost Does not scale
Alternative technology	Direct, usually, no computational inference is required	**Not necessarily more accurate**
Multiple ordinary technologies	Using a consensus between the technologies allow reducing the number of false positives compared to each individual technology	Disagreement between used technologies results in the incompleteness of the gold standard
Mock community	Ground truth is fully known, because raw data is generated from prepared gold standard	The small number of items (e.g., microbial species) compared to reality The designed community is artificial
Curated database	Allows access to sensitivity, by comparing the number of elements in the sample and the database	Incompleteness of curated databases results in limited ability to define true positives and false negatives
Computational simulation	**Ground truth is fully known**, because raw data is generated from prepared gold standard **Cost-free** generation of multiple gold standards	Technology is simulated, and cannot capture true experimental variability and will always be less complex than real data Gold standard data is artificial

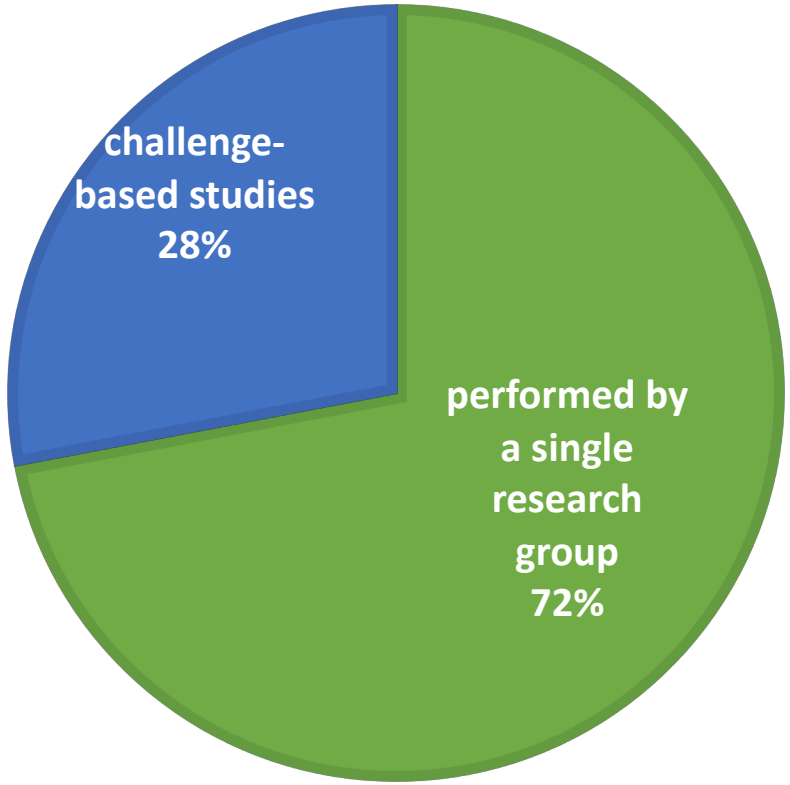
How to improve simulated data?

- incorporate real and simulated data in one comprehensive data set
- subsampling real datasets to generate new datasets with known properties

Individual vs. competition-based benchmarking

Individual. Single research groups conduct individual benchmarking studies of relevant computational problems

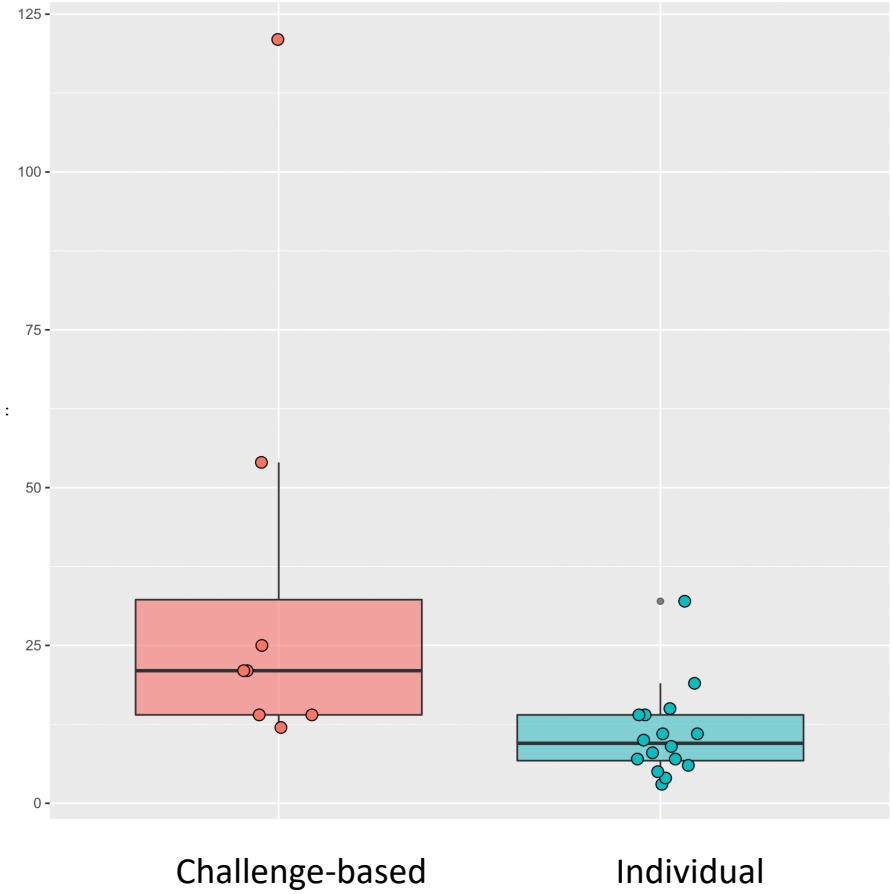
Challenge-based. An organized competition in which participants compete to solve the problems



*based on 25 benchmarking efforts published across 10 relevant peer-reviewed journals from 2011 to 2017

Number of benchmarked methods

- many tools require a complicated installation process (~22%*)
- some tools are impossible to install and run in a reasonable amount of time (~28%*)
- many tools lack comprehensive documentation (~60% of installable tools*)

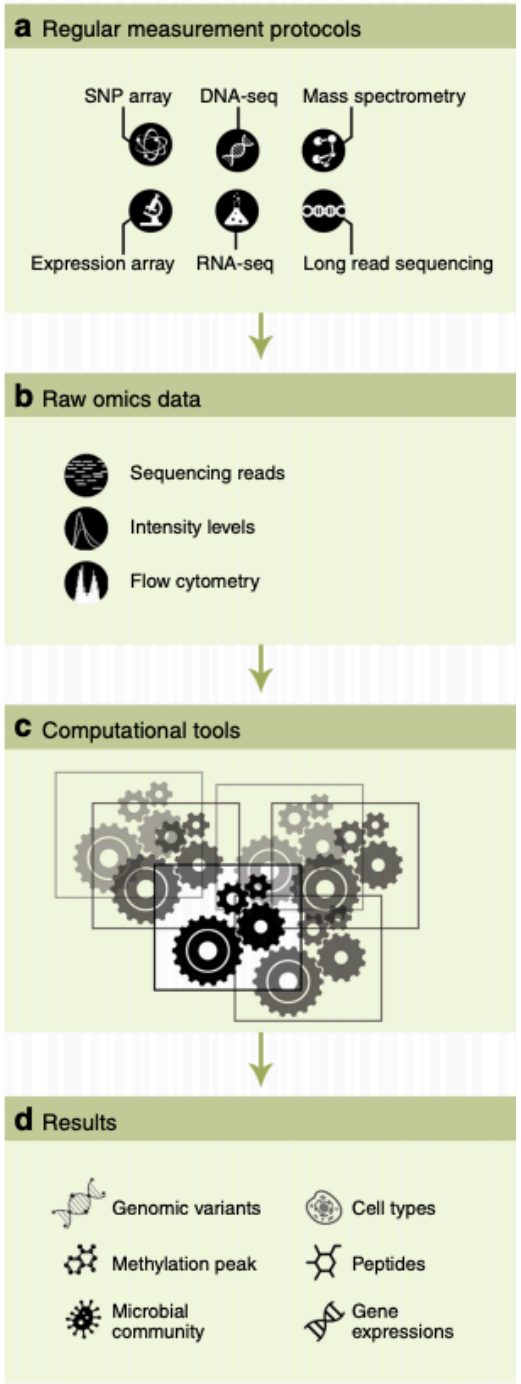


*based on 98 randomly selected tools across various domains of computational biology (Mangul et al. PLoS Biology 2019)

based on 25 benchmarking efforts published across 10 relevant peer-reviewed journals from 2011 to 2017

Parameter optimization

- Evaluating a software tool is non-binary decision
- Competition-based benchmarking studies rely on the expertise of the tool's developer to choose optimal parameters

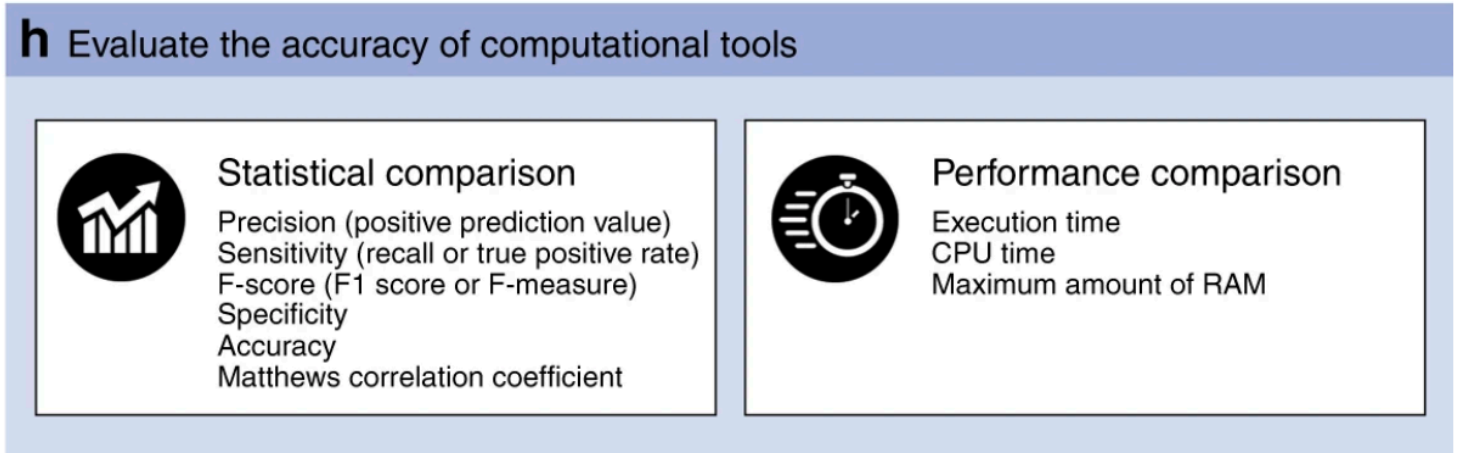


Benchmarking data and supporting documentation

- Data and code are often not shared (40%*)
  - an absence of journal policies requiring the public sharing of these resources
  - Infrastructural challenges to sharing large data generated by the benchmarking studies

*based on 25 benchmarking efforts published across 10 relevant peer-reviewed journals from 2011 to 2017

Evaluate performance of the tools



- defining statistical measures is an extremely complicated and ambiguous process
- sensitivity and PPV were the most popular measures
- performance comparison is often ignored (72%*)

*based on 25 benchmarking efforts published across 10 relevant peer-reviewed journals from 2011 to 2017

Other important aspect of benchmarking

- Continuous benchmarking
- Incentivizing community adoption

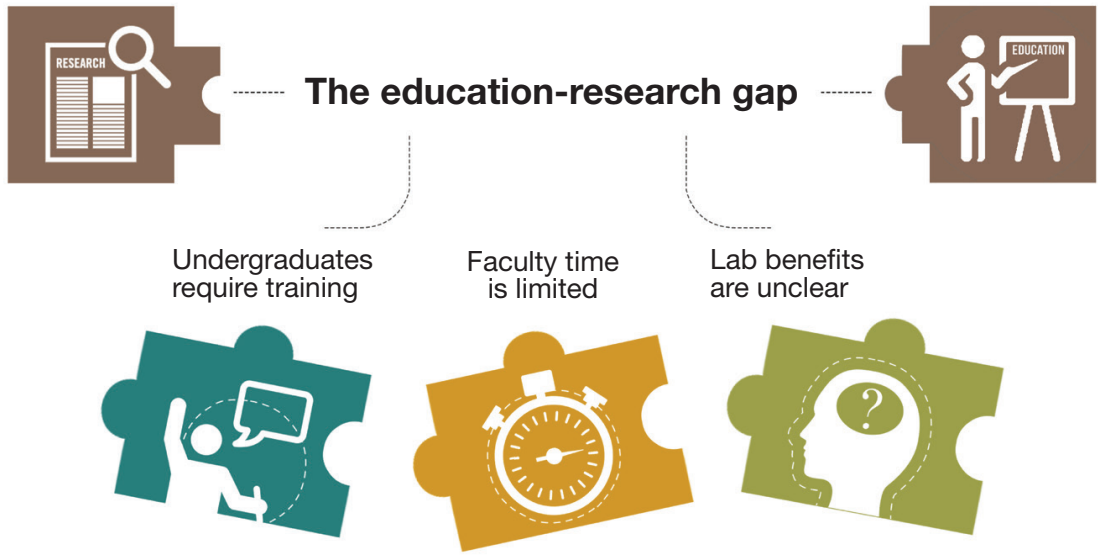
Principles for systematic benchmarking

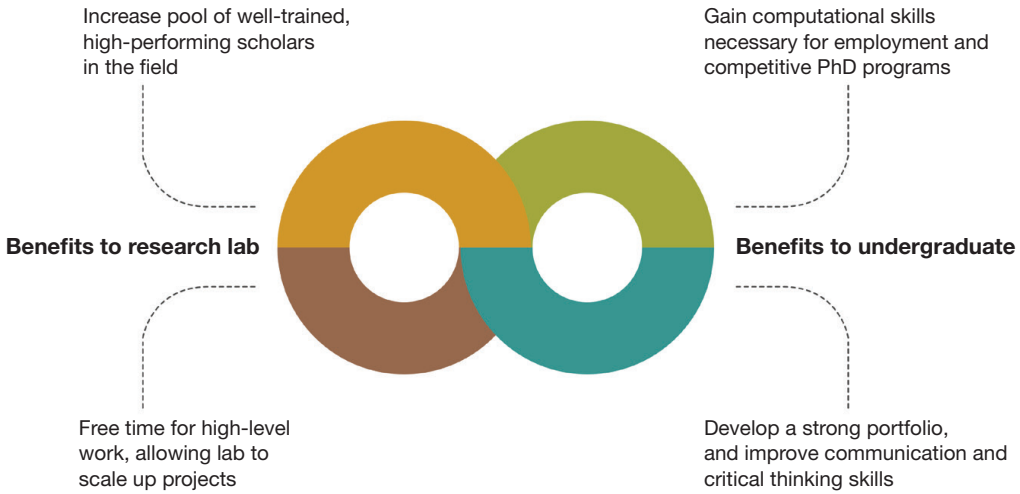
- Compile a comprehensive list of tools to be benchmarked
- Prepare and describe benchmarking data
- Carefully select evaluation metrics and pack in form of scripts (which the community can later use)
- Consider parameter optimization
- Summarize algorithm features and share commands for installing and running tools (preferably as virtual machine images or containers)
- Provide a flexible interface for downloading data

Open questions

- How can we encourage the research community to work on benchmarking?
- How can we address concern that benchmarking work may foster negatively competitive sentiments in the research community?
- How we can allocate funding for benchmarking research?

Involving undergraduates in genomics research to narrow the education–research gap





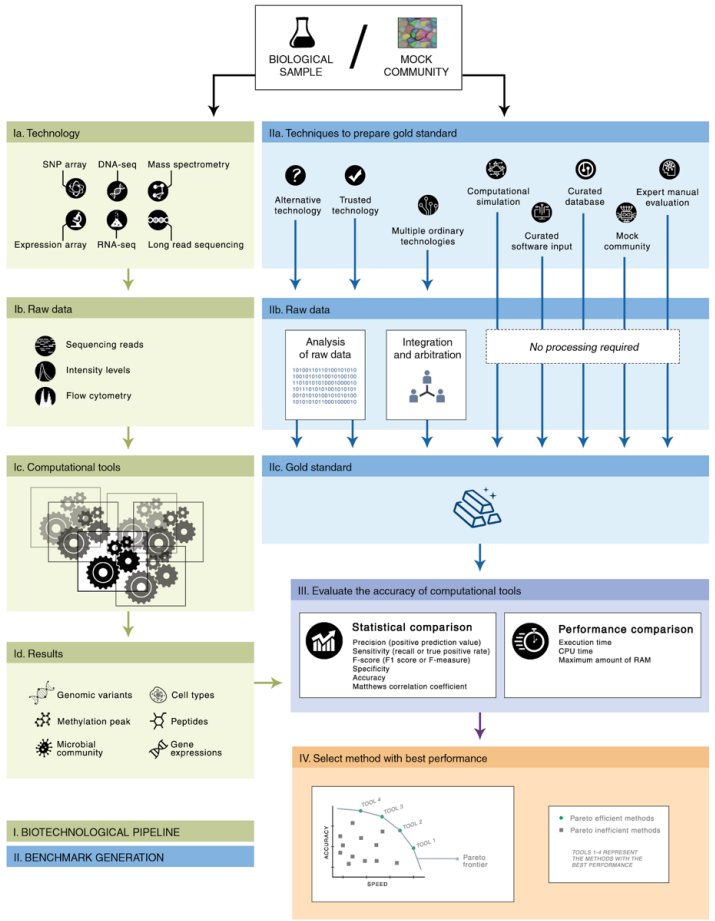
S. Mangul, L. S. Martin, and E. Eskin. "Involving undergraduates in genomics research to narrow the education–research gap." Nature Biotechnology

Review Article | OPEN | Published: 27 March 2019

Systematic benchmarking of omics computational tools

Serghei Mangul, Lana S. Martin, Brian L. Hill, Angela Ka-Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin & Jonathan Flint

Nature Communications **10**, Article number: 1393 (2019) | Download Citation



Acknowledgment

- Lana S. Martin
- Brian Hill
- Angela Ka-Mei Lam
- Margaret Distler
- Alex Zelikovsky
- Eleazar Eskin
- Jonathan Flint

The Mangul lab is looking for postdocs!

- Want to work at USC and live in Los Angeles?
- Want to develop cool cutting edge bioinformatics methods and apply them across the largest datasets?

USC School of Pharmacy