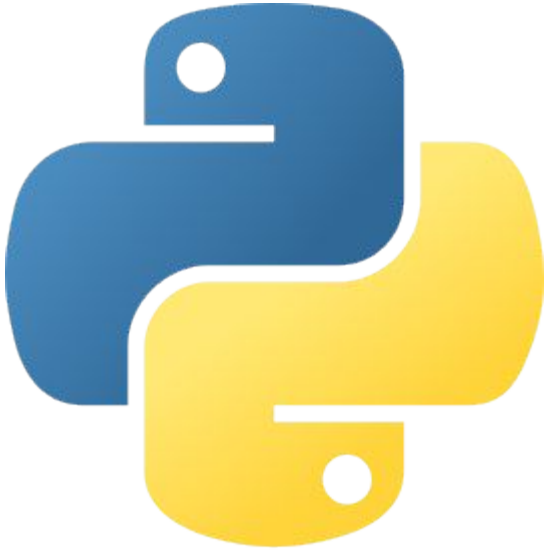
EPIGENE LABS

InMoose
Bridging the reproducibility gap between R and Python

Maximilien Colange, Abdelkader Behdenna

R vs Python

VS



R vs Python

historical standard

- focused on **stats**
- **bioinformatics** ecosystem
- **visualization** features

the new kid on the block

- **general-purpose** language
- standard for **AI/ML**
- **multi-disciplinary** ecosystem

Epigene Labs

EPIGENE LABS

🌱 Founded in 2019

💶 Raised €10M+

🇫🇷 R&D HQ in France

🇺🇸 Business HQ in US

👩‍💻👨‍🔬 20+ Scientists & Engineers

Unlocking insights for faster precision oncology breakthroughs

Potential of omic data remains vastly untapped

Petabytes of genomic data are generated each year[4] . . .

Due to inconsistencies and lack of a standardized data collection process, only a fraction of the data is often used.

Epigene Labs aims to unlock the full potential of genomic data to create opportunities in extracting depth of clinical and molecular insights.

(4) Banks. Nature Medicine. 2020.

mCUBE: a multi-cancer, -omic, -source platform

Data Source

Mapped Data*

- **microarrays**
  15K datasets
  940K samples
- **bulkRNA-seq**
  5K datasets
  170K samples
- **scRNA-seq**
  2K datasets
  190K samples

+++

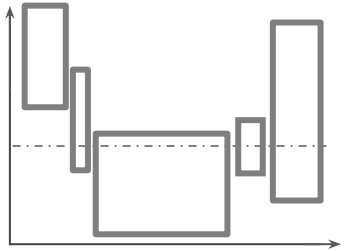
Partner

Harmonization & Normalization

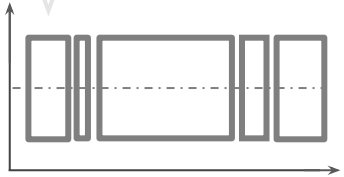
Clinical Data

AI-powered

Molecular Data

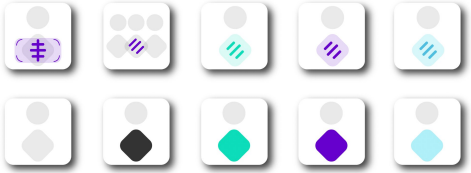


Next-generation bioinformatics powered by InMoose

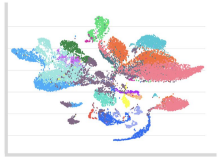


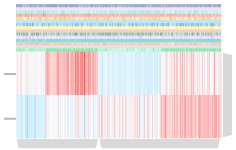
Visualization & Analysis

Disease Atlas



Insight Generation





**mapped databases include primarily GEO in addition to TCGA, MET1000, CPTAC, HPA, GTEx, CCLE*

Motivations for InMoose

Epigene Labs activity spans multiple fields

Python: one language to rule them all

- AI and ML
- data science
- bioinformatics
- website

↳ **simplifies our technical stack**

From R to Python: handling the legacy

- find **Python equivalent** to R tools
- check for **regressions**
- maintain **compatibility** before/after migration
  - not realistic to migrate overnight
  - not realistic to rerun all previous analysis

From R to Python: handling the legacy

- find **Python equivalent** to R tools
- check for **regressions**
- maintain **compatibility** before/after migration
  - not realistic to migrate overnight
  - not realistic to rerun all previous analysis

R ↔ Python reproducibility is critical

InMoose: INtegrated MultiOmic Open-Source Environment

Bridge the reproducibility gap between R and Python

- provide **drop-in replacements** for state-of-the-art R tools
  - ensuring results **comparability**
- **harmonize** formats
- **open-source**
  - our way to give back to the community

Implementation

Tools ported

focus on bulk transcriptomic data

- **batch effect correction**
  empirical Bayes methods
  - combat (microarray)
  - combat-seq (RNA-Seq)

- **differential gene expression**
  empirical Bayes methods
  - limma (microarray)
  - edgeR (RNA-Seq)
  - DESeq2 (RNA-Seq)

- **data simulation**
  - splatter (RNA-Seq, scRNA-Seq)

- **data clustering**
  - consensus clustering

Test-Driven Development to Ensure Reproducibility

- 🔎 identify **features** to port
- 🧪 create an **extensive test suite**
  - use R tool as ground truth => **reproducibility**
- ↺ develop, debug, improve... **until all tests pass**

Challenges

- **difference in API**
  - *e.g.* different parameterizations of statistical distributions
  - **whiteboard maths**

- **C++**
  - replace R/C++ framework with Python/C++ framework
  - most C++ code replaced by **Cython**

- **numerical stability**
  - *e.g.* overflows, underflows, NaN...
  - hard to detect, often occurs in corner cases
  - **whiteboard maths, extend the test suite**

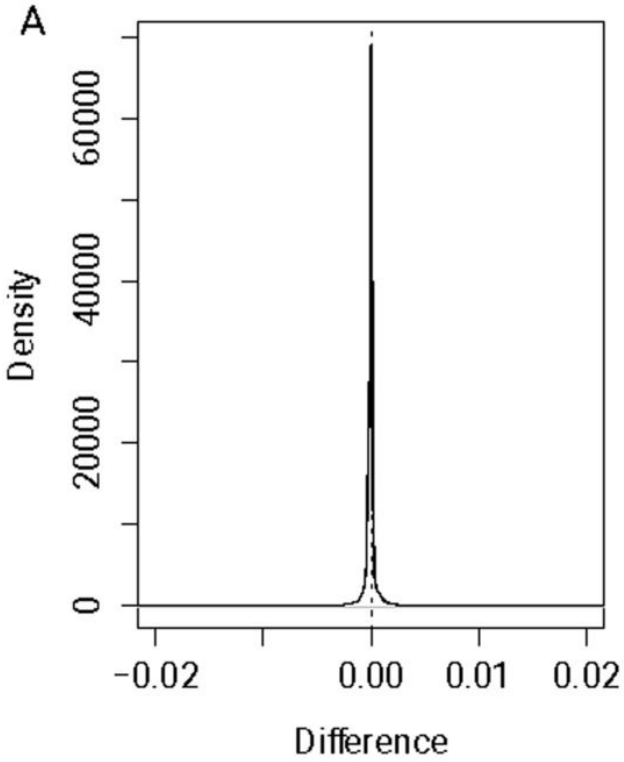
Challenges

- **adjusting tolerance in tests**
  - *e.g.* reproduce within 1e-6? within 1e-9?
  - question the notion of reproducibility
  - **trial-and-error, critical thinking**

- **licenses**
  - half a dozen licenses to harmonize
  - legal counseling
  - **GNU General Public License v3**

Validation

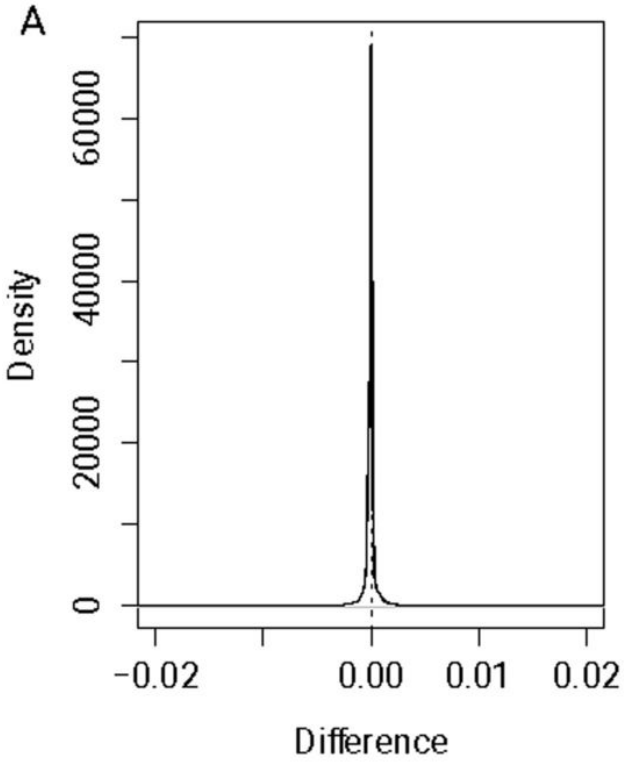
Batch Effect Correction



distribution of relative difference in output expression matrices between InMoose and ComBat

Behdenna, A., Colange, M., Haziza, J. *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics* **24**, 459 (2023). https://doi.org/10.1186/s12859-023-05578-5

Batch Effect Correction



distribution of relative difference in output expression matrices between InMoose and ComBat

- InMoose and ComBat-Seq have the **<u>same exact output</u>** on tested cohorts
- InMoose **4-5 times faster** than R tools

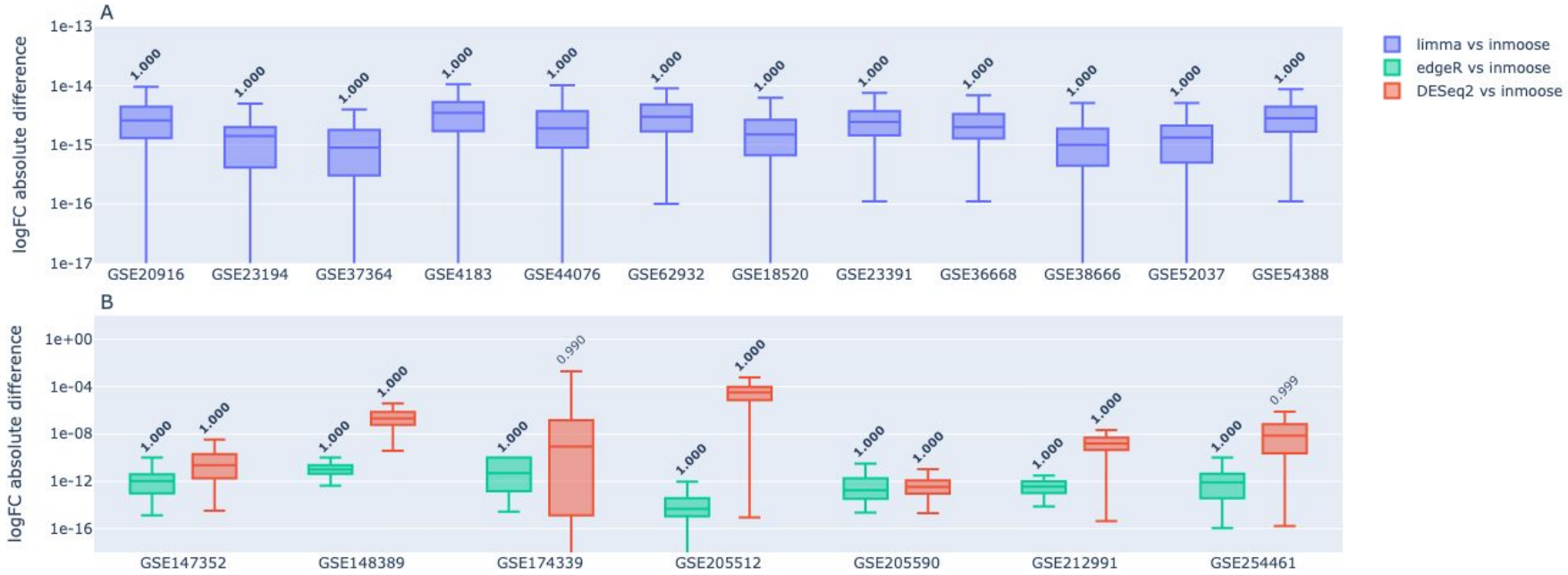
Behdenna, A., Colange, M., Haziza, J. *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics* **24**, 459 (2023). https://doi.org/10.1186/s12859-023-05578-5

Batch Effect Correction: Unique Features

- **batch of reference**
  - other batches are corrected relatively to the reference

- cohort **QC report**
  - assessment of residual batch effects
  - correlations with covariates

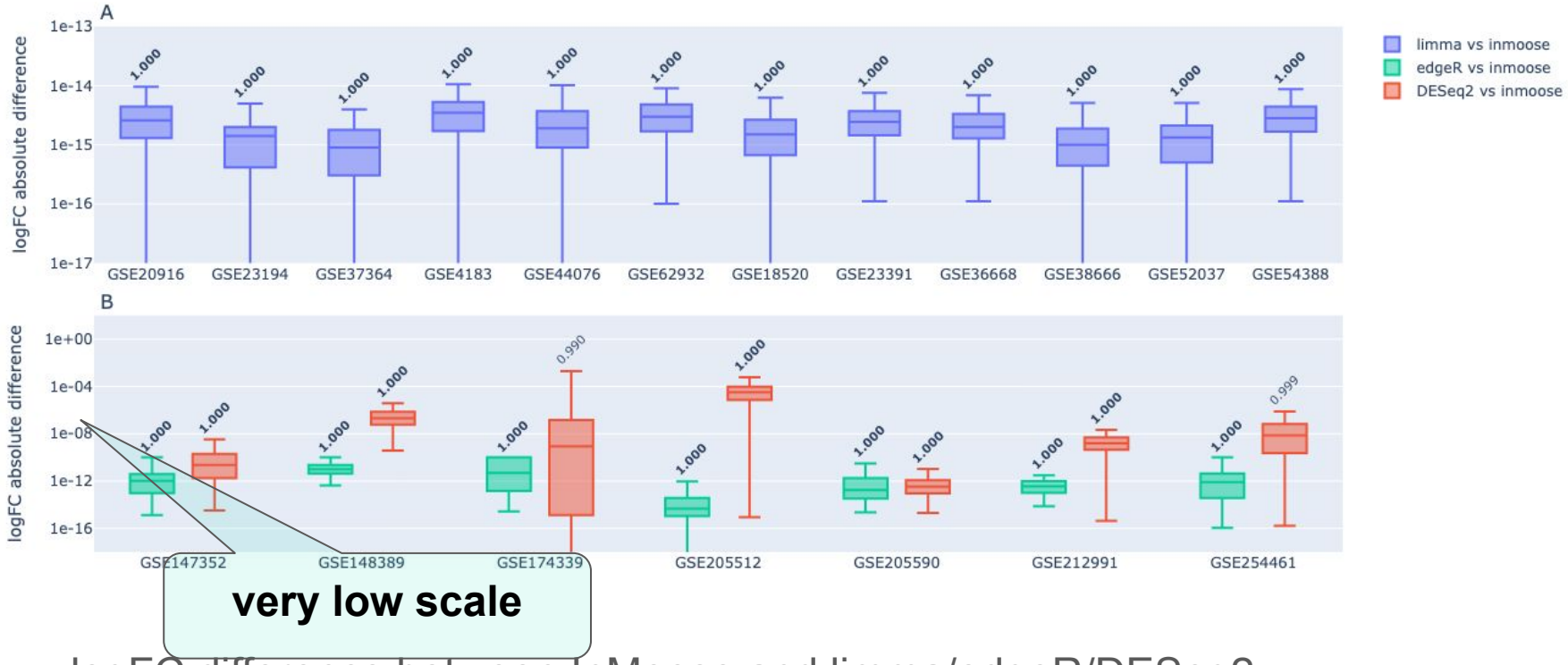
Differential Gene Expression Validation



logFC difference between InMoose and limma/edgeR/DESeq2

(mean, quartiles, 2.5%- and 97.5%-quantiles)

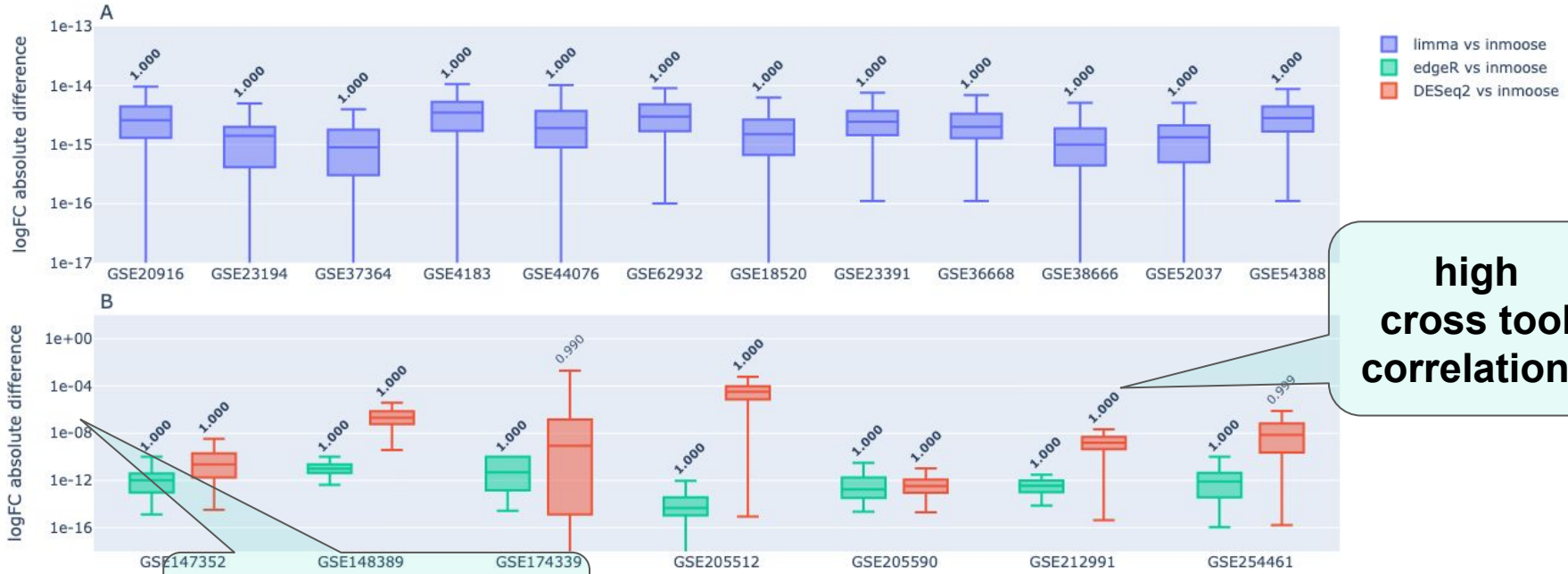
Differential Gene Expression Validation



logFC difference between InMoose and limma/edgeR/DESeq2

(mean, quartiles, 2.5%- and 97.5%-quantiles)

Differential Gene Expression Validation



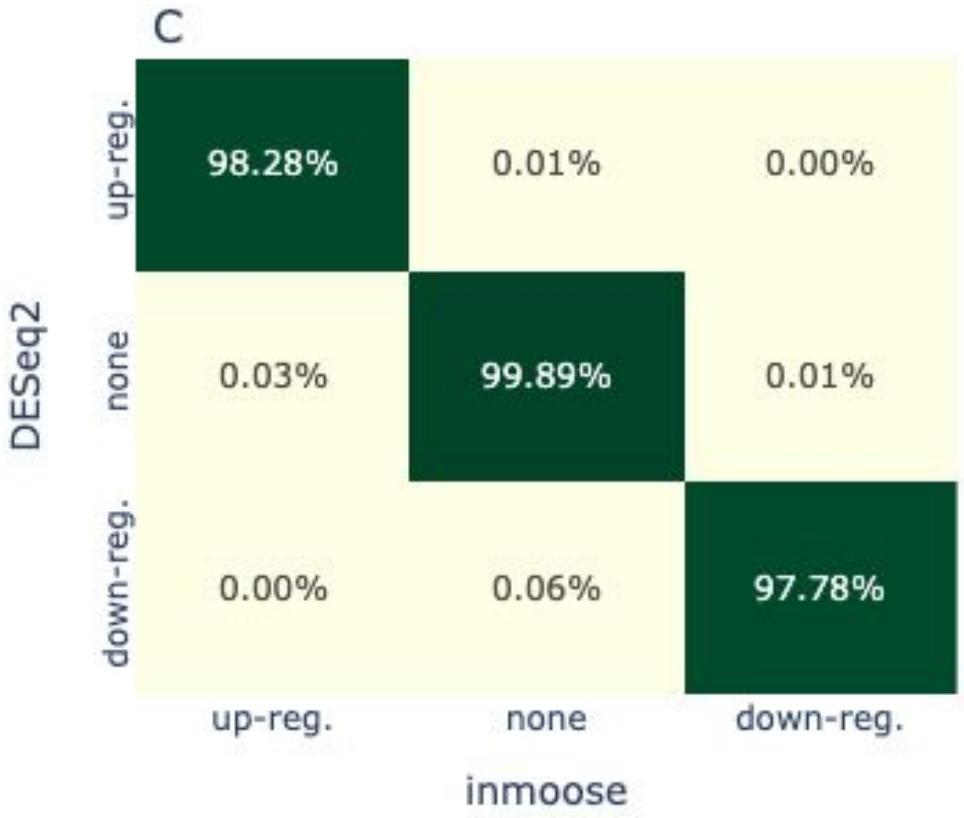
high cross tool correlations

very low scale

logFC difference between InMoose and limma/edgeR/DESeq2

(mean, quartiles, 2.5%- and 97.5%-quantiles)

Differential Gene Expression Validation



cross tool rank correlation of

- down-regulated
- up-regulated
- similarly expressed

genes

Colange, M., Appé, G., Meunier, L. *et al.* Differential Expression Analysis with InMoose, the Integrated Multi-Omic Open-Source Environment in Python. *Bioarxiv* **2024.11.14.623578** (2024). https://doi.org/10.1101/2024.11.14.623578

Differential Gene Expression: Unique Features

- **harmonized** output format
  - limma
  - edgeR
  - DESeq2

- **meta-analysis** module
  - allows to combine diff exp results
    - across datasets
    - across tools

Benefits for Epigene Labs

- **seamlessly integrate** our codebase
  - across teams, across fields of expertise
- **reduce** maintenance burden
- **streamline** work of bioinformaticians
  - inmoose is like an all-in-one toolbox
- **adjust** to specific needs
  - *e.g.* meta-analysis for differential gene expression

Check it out!

https://github.com/epigenelabs/inmoose

https://pypi.org/project/inmoose/

```
pip install inmoose
```


Behdenna, A., Colange, M., Haziza, J. *et al.* pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *BMC Bioinformatics* **24**, 459 (2023). https://doi.org/10.1186/s12859-023-05578-5

Colange, M., Appé, G., Meunier, L. *et al.* Bulk Transcriptomic Analysis with InMoose, the Integrated Multi-Omic Open-Source Environment in Python. *Bioarxiv* **2024.11.29.625982** (2024). https://doi.org/10.1101/2024.11.29.625982

Colange, M., Appé, G., Meunier, L. *et al.* Differential Expression Analysis with InMoose, the Integrated Multi-Omic Open-Source Environment in Python. *Bioarxiv* **2024.11.14.623578** (2024). https://doi.org/10.1101/2024.11.14.623578

QUESTIONS?