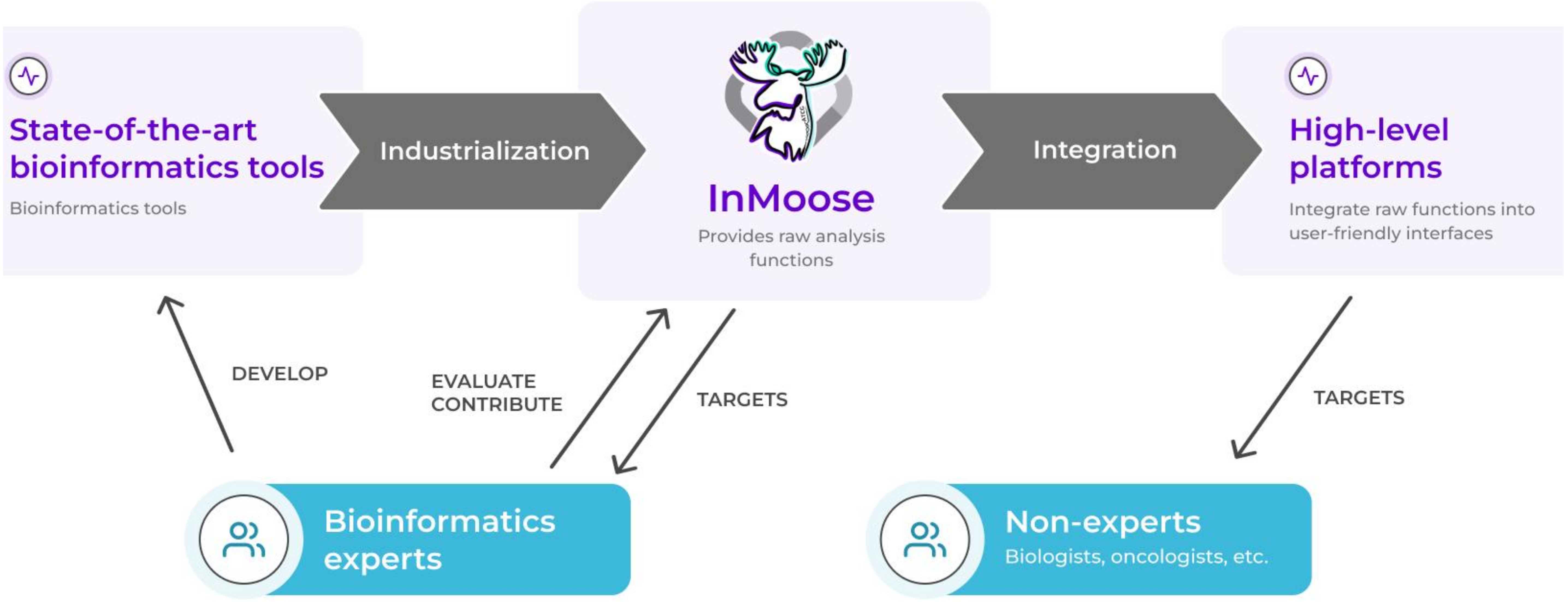
# Introducing InMoose

## An Integrated Multi-omic Open Source Package for Python Analyses

M. Colange[1], G. Appé[1], A. Nordor[1], A. Behdenna[1]
[1]Epigene Labs, Paris, France

American Association for Cancer Research



**State-of-the-art bioinformatics tools**
Bioinformatics tools

Industrialization

**InMoose**
Provides raw analysis functions

Integration

**High-level platforms**
Integrate raw functions into user-friendly interfaces

DEVELOP

EVALUATE CONTRIBUTE

TARGETS

TARGETS

**Bioinformatics experts**

**Non-experts**
Biologists, oncologists, etc.

## Introduction

Exponentially increasing amount of omic data exacerbates the need for high-quality, high-performance and highly interoperable computational tools for cancer research.
Our goal is to industrialize tried-and-tested, high-quality tools, by making them more efficient and more interoperable.

## Open Source

- InMoose is open source (GPL3 License).
- Open source is our way to **bind with the community**.
- Open source is a great way to achieve **quality and performance**:
  - anyone can audit the code (evaluation)
  - anyone can contribute to the code (improvement)
- Open source fosters interoperability:
  - easier to **interface with other tools**
- Open source is a prerequisite for **open science**.
- pyComBat GitHub repository:
  10+ external issues, 42 stars, 18 forks
- pyComBat preprint: 16 citations

## Approach

Our method is to port existing tools to Python, and integrate them in a single package.
As a general-purpose, mainstream language, Python offers several perks:

- easy integration into large-scale frameworks (*e.g.* web platform)
  ⇒ **versatility**, user-friendliness, wide target audience
- widespread language
  ⇒ **accessibility** (*e.g.* cross-disciplinary collaboration)
- trendy in bioinformatics
  ⇒ harmonizing ecosystem, momentum
- porting = opportunity window for improvement
  ⇒ functionality, **performance** (*e.g.* ComBat-Seq)

First tools ported:

- batch effect correction (ComBat, ComBat-Seq)
- differential expression analysis (DESeq2)

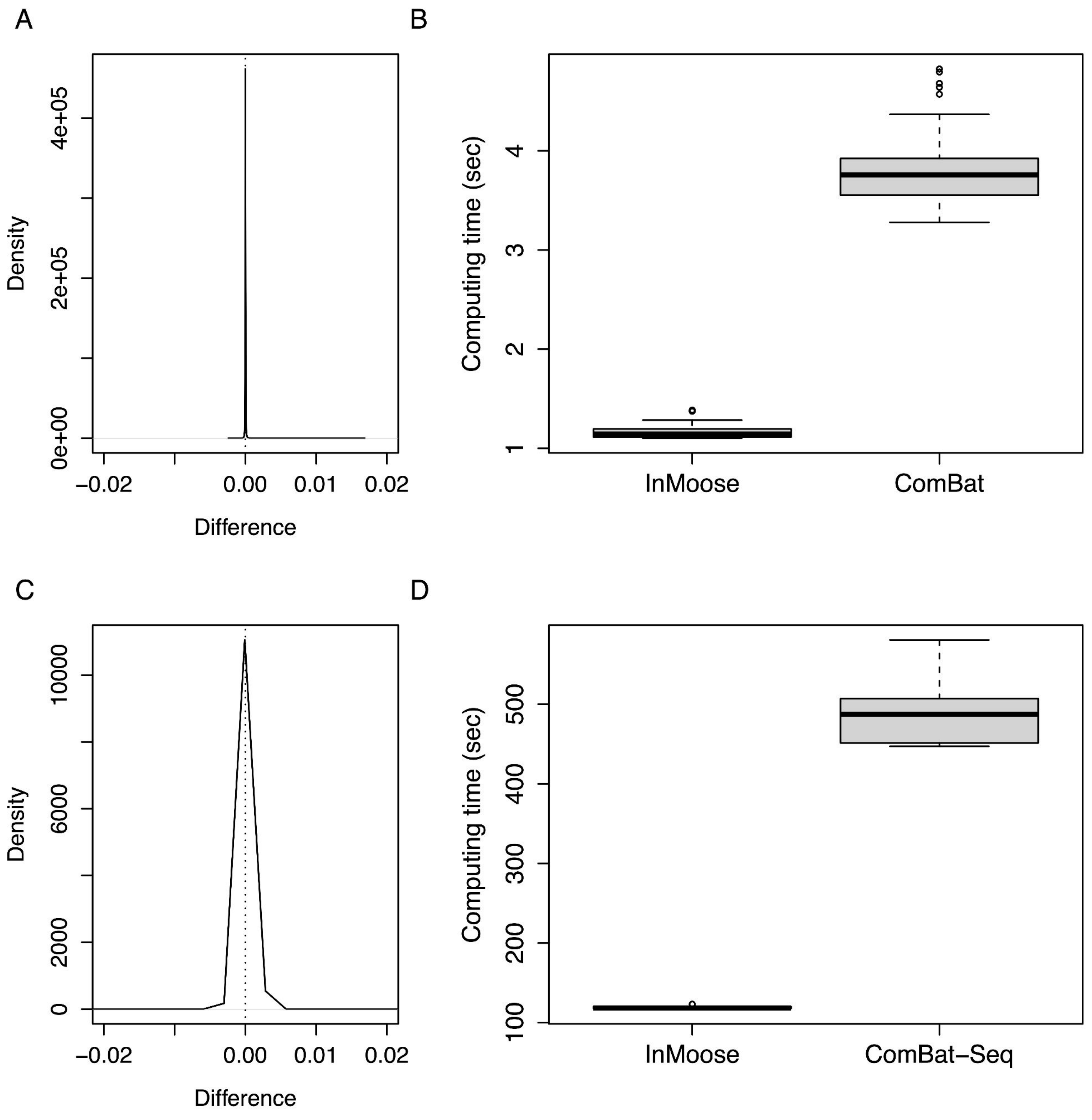
**Our approach aims to foster a larger collaborative effort to build and grow a consistent state-of-the-art Python ecosystem for cancer bioinformatics.**

## Results

**Our approach is to faithfully port the code from R/C++ to Python/C++, to preserve the quality of the ported tools.**

We assess the quality of our ports by replicating results of the original tools.

- Microarray batch effect correction
  - Results are reproduced with a mean **relative difference of $2.5e^{-7}$%** (CI95%: [$3.4e^{-11}$,$1.6e^{-6}$]).
  - InMoose runs almost **4x faster** than the original R implementation.
- RNAseq batch effect correction
  - Results are reproduced with a mean **relative difference of $5.4e^{-8}$%** (first non-zero quantile 0.9995).
  - InMoose runs more than **4x faster** than the original R implementation.



Performance of InMoose *vs*. Combat and ComBat-Seq.
A - Distribution of the relative differences between the expression matrices corrected for batch effects, respectively by ComBat and InMoose (parametric version). The vertical dotted line corresponds to zero.
B - Computation time in seconds for InMoose and ComBat for the parametric method.
C - Distribution of the relative differences between the expression matrices corrected for batch effects, respectively by ComBat-Seq and InMoose. The vertical dotted line corresponds to zero.
D - Computation time in seconds for InMoose and ComBat-Seq.

## References

1. Johnson, W. E., Li, C., & Rabinovic, A. (2007). *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 8(1), 118–127.
2. Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). *ComBat-Seq: batch effect adjustment for RNASeq count data.* NAR Genomics and Bioinformatics, 2(3), lqaa078. a
3. Behdenna, A., Haziza, J., Azencott, C.-A., & Nordor, A. (2020). *pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods*. bioRxiv.
4. Love, M. I., Huber, W., & Anders, S. (2014). *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 15(12), 550.
5. Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Zhang, Y., Storey, J. D., & Torres L. C. (2022). sva: Surrogate Variable Analysis. R package version 3.46.0.

## Contact Information

- Akpéli Nordor, PharmD, PhD
- Email: akpeli@epigenelabs.com

## Check it out!

- QR code to PyPI page
- Call to contribution and feedback

European Innovation Council
Co-funded by the European Union