



HAL
open science

New tools to optimise the analysis of a large RNA-Seq dataset from non model species: development of a hybrid assembly strategy and assessment of library complexity from raw sequencing output.

Jacques Lagnel, - Khalid Belkhir, - Tereza Manousaki, - Erick Desmarais, -
Anastasia Tsagkarakou, - Alban Mancheron

► **To cite this version:**

Jacques Lagnel, - Khalid Belkhir, - Tereza Manousaki, - Erick Desmarais, - Anastasia Tsagkarakou, et al.. New tools to optimise the analysis of a large RNA-Seq dataset from non model species: development of a hybrid assembly strategy and assessment of library complexity from raw sequencing output.. 2014. hal-02799163

HAL Id: hal-02799163

<https://hal.inrae.fr/hal-02799163>

Preprint submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

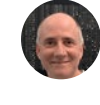
See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/319932729>

New tools to optimise the analysis of a large RNA-Seq dataset from non model species: development of a hybrid assembly strategy....


Conference Paper · September 2014
DOI: 10.13140/RG.2.2.17467.11945

CITATIONS
0

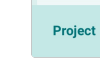
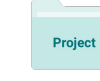
6 authors, including:

 **Jacques Lagnel**
INRA French National Institute for Agricultural Research
49 PUBLICATIONS 1,574 CITATIONS
[SEE PROFILE](#)

READS
17

 **Khalid Bekhir**
Université de Montpellier
70 PUBLICATIONS 3,911 CITATIONS
[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:

-  **Bioanalysis** [View project](#)
-  **Bioinformatics** [View project](#)

New tools to optimise the analysis of a large RNA-Seq dataset from non model species: development of a hybrid assembly strategy and assessment of library complexity from raw sequencing output

Jacques Lagnel¹, Khalid Belkhir², Tereza Manousaki¹, Erick Desmarais², Anastasia Tsagkarakou³ and Alban Mancheron⁴

¹ Institute of Marine Biology Biotechnology and Aquaculture, Hellenic Centre for Marine Research-HCMR Greece, Gournes Pediasos, P.O. Box 2214, Heraklion 71003, Crete, Greece

² UMR CNRS 5554, Institut des Sciences de l'Évolution de Montpellier, Université Montpellier 2 - cc63, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

³ Hellenic Agricultural Organisation-Demeter, NAGREF, Plant Protection Institute of Heraklion, P.O. Box 2228, 71003 Heraklion, Crete, Greece

⁴ Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM) UMR 5506, Université Montpellier 2

lagnel@hcmr.gr

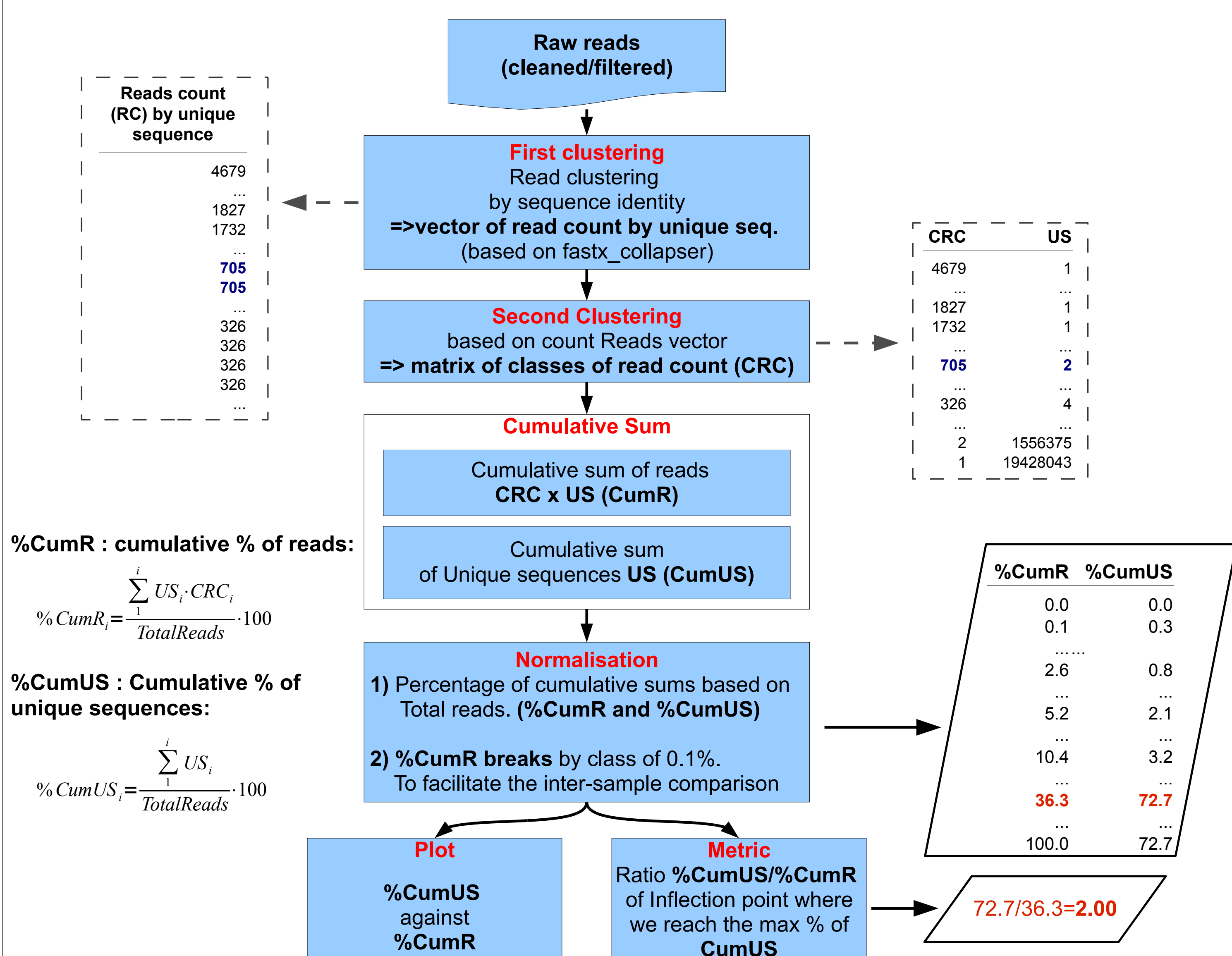
INTRODUCTION

Major challenges have emerged with the popularisation of the NGS technology, from library preparation up to data analysis. Here, we present two tools that respectively assess the complexity of the sequenced library and improve the quality and feasibility of the assembly process. The presented tools facilitate the analysis of RNA-Seq for non-model species improving the feasibility of the assembly and allowing for the post-sequencing evaluation of the library construction.

A METRIC TO EVALUATE LIBRARIES COMPLEXITY

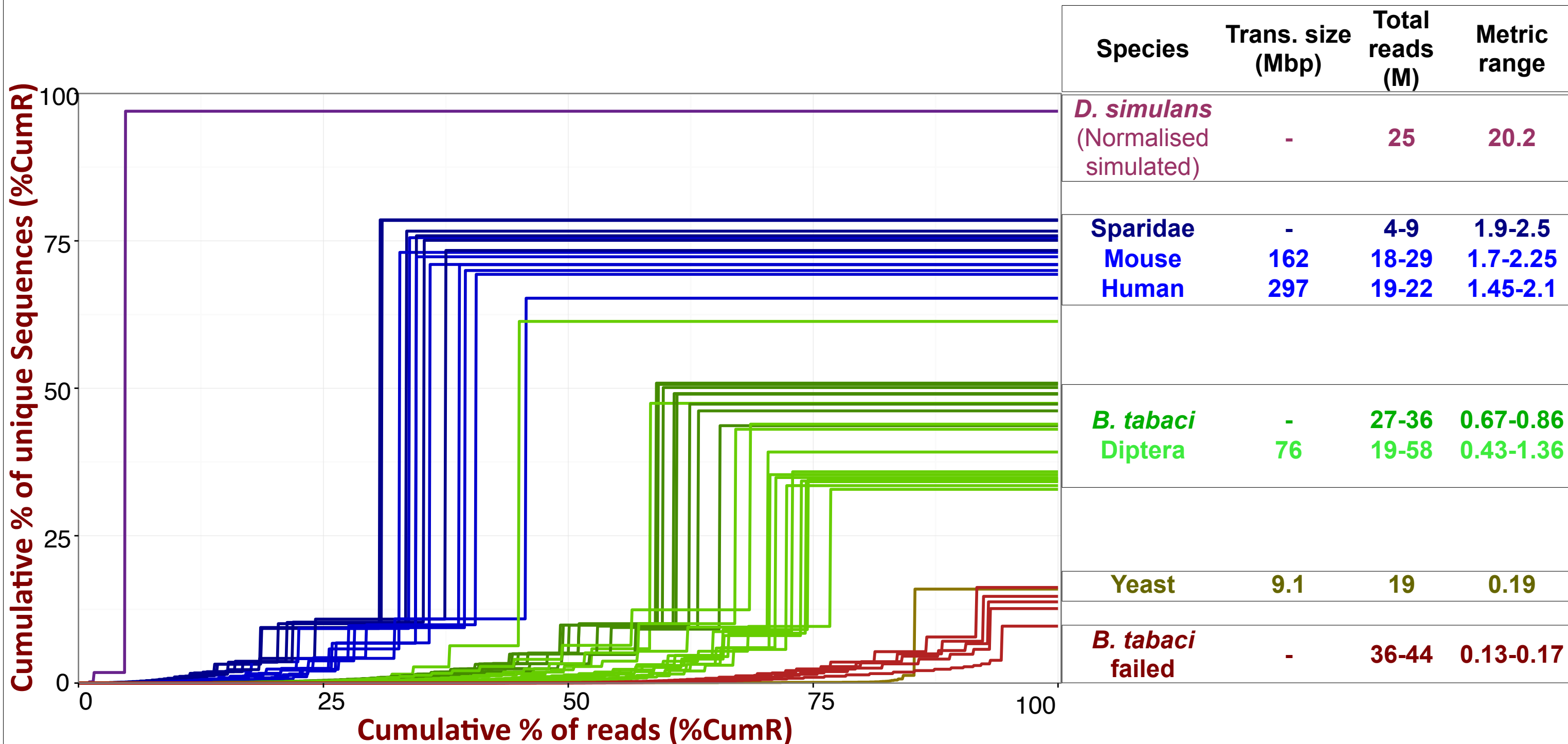
An important step prior to the analysis of RNA-Seq data is the quality assessment of both the library construction and the sequencing process. Indeed, technical failures in these steps can lead to biased RNA representation and over-sequencing of particular molecules (e.g., rRNA, PCR duplicates, contaminant RNA/DNA) that do not represent the starting biological sample. This may reduce the complexity of the library and, consequently, the coverage of the transcriptome. We developed a new metric that infers the diversity of unique sequences and assesses the complexity of a given library using two clustering steps of identical reads. A diversity index is calculated as the ratio of cumulative unique sequences over cumulative reads, representing the point at which the maximum percentage of unique sequences is reached. This metric can be used to compare the library construction success and evaluate multiple experiments while it can be used as a decision tool included in the workflow of omics analysis. The procedure is implemented in Perl and R scripts and results can be viewed as a plot.

METRIC WORKFLOW



RESULTS

Metric application to *Bemisia tabaci* (hemiptera) RNA-Seq experiment and various publicly available datasets (RNA-Seq on Illumina HiSeq2000 platform paired-end library, SRA 2013-2014).



Plot of the cumulative % of unique sequences (%CumUS) against cumulative % of reads (%CumR) from 42 quality controlled datasets. The transcriptome size, total reads and metric values are shown in the table.

The profile of the different libraries seems to follow a transcriptome size pattern with the exception of "failed" libraries that cluster at the bottom right side of the plot. The dataset that gave the highest metric value was the normalised simulated data from *Drosophila simulans*. Technical problems lead to sequence overrepresentation (e.g., rRNA, PCR duplicates, contaminant RNA/DNA) resulting to poor complexity libraries that tend to cluster at the bottom right part of the plot with a low metric value.

CONCLUSION

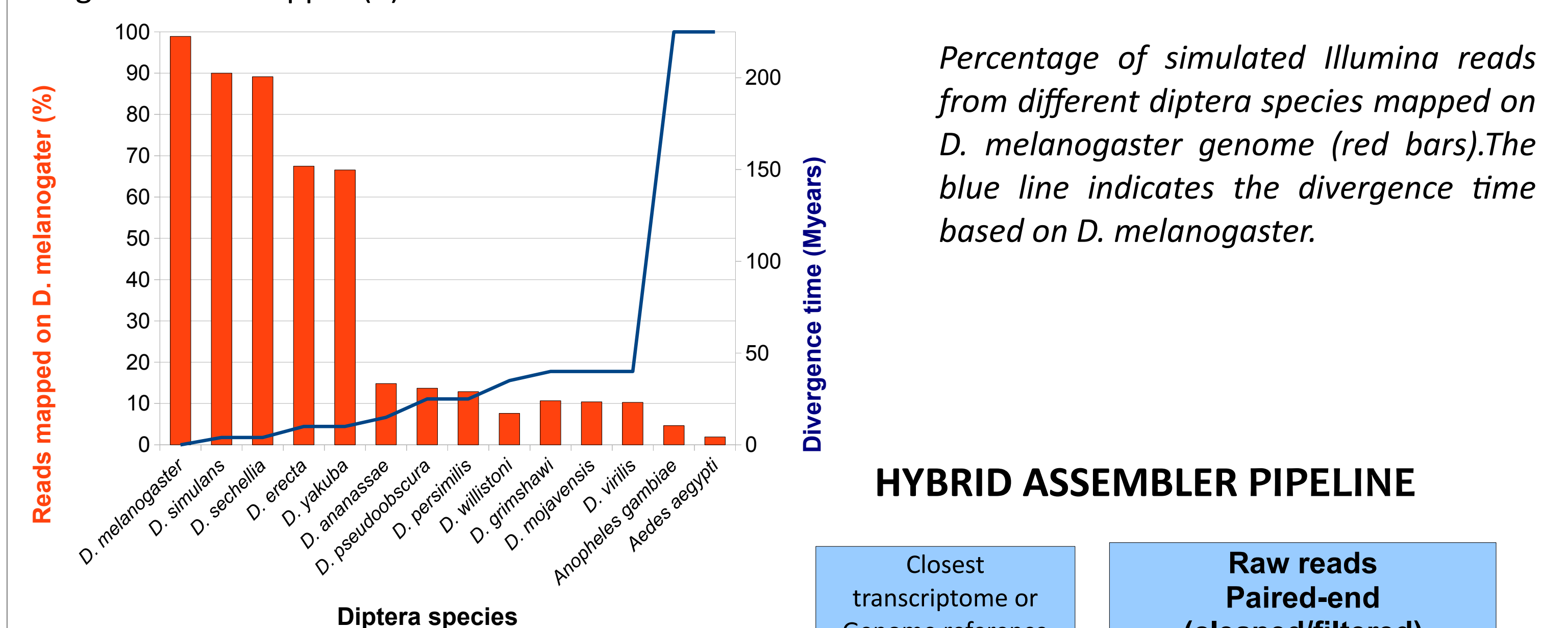
This metric can be used to compare the library construction success and evaluate multiple experiments while it can be used as a decision tool included in the workflow of omics analysis. The automatic procedure is implemented in Perl (with embedded R code for plot generation) script.

HYBRID ASSEMBLY STRATEGY

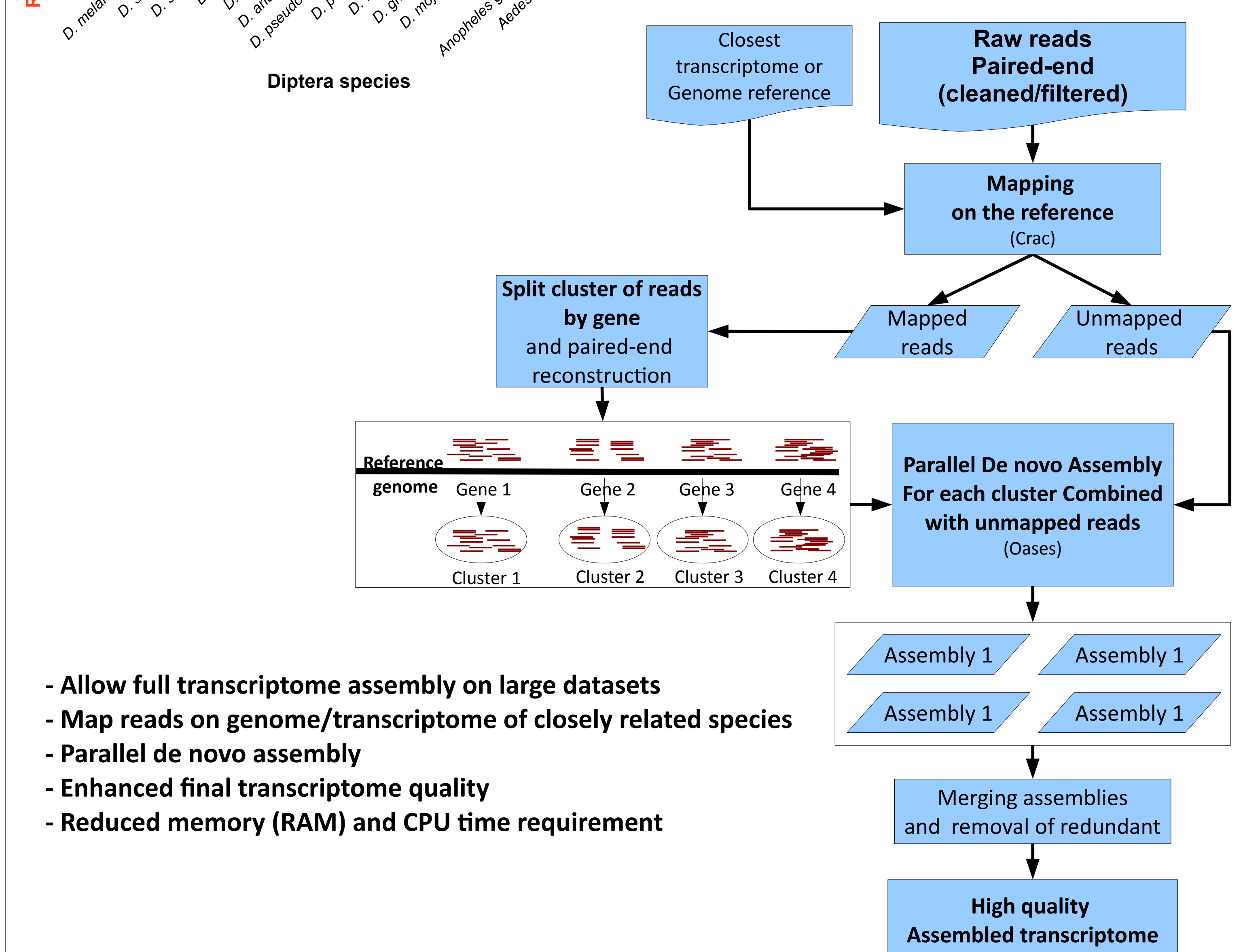
Depending on the availability of high-quality reference sequences, transcriptomes are either assembled following reference-based or *de novo* strategies. For non-model species where a high-quality reference transcriptome/genome is absent, a closely related species reference sequence can be used as a guide to improve the quality of the reconstructed transcriptome and decrease the computational requirements. By combining the two complementary assembly strategies, we can take advantage of the high sensitivity of reference-based assemblers, while leveraging the ability of *de novo* assemblers to detect novel transcripts. The *de novo* assembly requires important computing resources, particularly memory. The proposed strategy solves the problem of intensive memory requirements of a *de novo* assembly and greatly reduces the computational time by parallelising the process. Here, we test this strategy on a simulated *Drosophila* group dataset using various reference model species in a wide range of divergence times to assess the mapping success.

DIVERGENCE TIME VERSUS MAPPING EFFICIENCY: A CASE STUDY ON DIPTERA

in silico study evaluation of the mapping success of simulated paired-end reads from several diptera species on a reference genome with different divergence times (1). We generated 25M paired-end reads 2X100 bp from cDNA of each species. Each simulated data set was mapped on the *Drosophila melanogaster* genome using the CRAC mapper (2).



HYBRID ASSEMBLER PIPELINE



RESULTS: run test on D. simulans

	Hybrid assembly	classic Assembly
RAM (GB)	≤6*	50
runtime (minutes)	≤28*	154
Total nb of sequences	5262	2852
Total length	4383480	4075304
Average length	833	1428
Max size	24796	24544
N50	1561	1972

* by job (reads cluster mapped on one scaffold of *D. melanogaster*)

The hybrid assembler pipeline was tested with simulated *D. simulans* Illumina reads mapped on *D. melanogaster* and the parallel step was performed with reads mapped on each individual scaffold of *D. melanogaster* separately.

CONCLUSION

Our hybrid strategy pipeline significantly improves the standard assembly procedure, not only in quality but also in reducing the computer resources required. It facilitates the analysis of large datasets especially of non-model species. In order to decrease the proportion of unmapped reads we plan to include a second round of mapping with consensus reference sequences based on SNPs detection.

REFERENCES

- Granzotto, A. et al. (2009) The evolutionary dynamics of the Helena retrotransposon revealed by sequenced *Drosophila* genomes. *BMC Evol Biol*, 9, 174.
- Philippe, N. et al. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol*, 14, R30.
- Schulz, M.H. et al. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28, 1086–1092.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: THALES. This work has been partially supported by the LifeWatchGreece project, funded by the GSRT (structural funds).