InterTADs: Integration of Multi-Omics Data on Topological Associated Domains

Maria Tsagiopoulou
Centre for Research and Technology-Hellas https://orcid.org/0000-0002-1653-0327

Nikolaos Pechlivanis
Centre for Research and Technology-Hellas

Fotis Psomopoulos (✉ fpsom@certh.gr)
Centre for Research and Technology-Hellas https://orcid.org/0000-0002-0222-4273

Methodology

Keywords: TAD, multi-omics, integration, chromatin, epigenetics, genetics, transcriptomics

Posted Date: August 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-54194/v1

InterTADs: Integration of multi-omics data on topological associated domains

Maria Tsagiopoulou[1*], Nikolaos Pechlivanis[1*] and Fotis Psomopoulos[1,2]

[1] Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece

[2] Dept of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

Corresponding author: fpsom@certh.gr

* MT and NP contributed equally to this work.

Abstract

Background: The integration of multi-omics data can greatly facilitate the advancement of research in Life Sciences by providing new insights on how biological systems interact. However, there is currently no widespread procedure for a robust, efficient and meaningful multi-omics data integration; the approach presented here is a first attempt towards increasing the reliability of data discovery power compared to the processing of individual biodata sets.

Results: Here, we proposed a high-speed framework, called InterTADs, for integrating multi-omics data from the same physical source (e.g. patient) taking into account the chromatin configuration of the genome, i.e. the topologically associating domains (TADs). The main concept of the proposed methodology is to create a single matrix with all different events (e.g. DNA methylation, expression, mutation) combined with their genome coordinates and the respective quantitative metrics after application of the appropriate scaling. The events are divided into their related TADs according to the chromosomal location and each TAD is evaluated for statistically significant differences between the groups of interest (e.g. normal cells vs cancer cells). Finally, several visualization approaches are available, including the mapping of the events on the chromosomal location of the TAD as well as the distribution of the counts within a given TAD across the different study groups.

Conclusions: InterTADs provides a general framework for integrating multi omics data and relating them with the TADs. This could lead to the extraction of new biological insight of the examined case study. InterTADs is an open-source tool implemented in R and licensed under the MIT License. The source code is freely available from https://github.com/nikopech/InterTADs.

Keywords: TAD, multi-omics, integration, chromatin, epigenetics, genetics, transcriptomics

Background

The study of the molecular mechanisms that may lead to cancer was revolutionized by the advent of Next Generation Sequencing (NGS)(1, 2). NGS include studies of whole genomes (whole-genome sequencing), smaller regions of the genome (exome sequencing), of the transcriptome (RNA-seq), the methylome (Bisulfite-seq) and protein-DNA binding sites (ChIP-seq)(3). Using NGS to sequence the entire human genome can produce more than 100 GB of raw data(4), thus leading to a whole cadre of challenges towards their analysis. The raw NGS-data are consequently analyzed by established and widely accepted bioinformatics tools (e.g. *bwa*, *TrimGalore*, *HISAT2*, *MACS2*, *R*)(5). The process of analyzing omics data usually leads to a high dimensional matrix, with the different cases listed as columns and the locations on the genome in which the examined event happened (e.g. mutation, gene expression etc.) as rows. The integration of several types of data that originate from the same physical source (e.g. patient) but focus on different mechanisms (gene expression, DNA Methylation, histone modifications etc.) is remaining a promising field since there is no widely accepted approach for it. The most common processes to compare different omics data are by (i) comparing the gene lists produced at the end of each individual analysis, with the assumption that overlapping genes were influenced by different mechanisms(6, 7) and (ii) checking the correlation of two events that are associated with the same gene, using statistical methods such as spearman or pearson correlation test(8, 9). However, as interactions in biological systems are generally not linear, methods such as PCA, Bayesian or non-Bayesian network-based were applied as extended data integration approaches(10). Although these methods are promising, they show instability and tend to over-fit to a given dataset. Moreover, there are several existing tools that integrate different kinds of omics data but they perform the analysis at the gene level (e.g. *CNAmet*, *iGC*, *PLRS*, *Oncodrive-CIS*) or they focus on sample classification based on the driving clinical perspective (e.g. *iClusterPlus* and *mixOmics*)(10, 11). Moreover, many existing tools consider pathway databases for further evaluation of the biological meaning(12, 13)

Going to a level of organization further than the simple chromosomal position, the introduction of NGS methods, like Hi-C, provides insight into chromatin organization such as the topologically associated domains (TADs). TADs represent segments of chromatin domains characterized by frequent interactions within themselves, and are conserved in mammals(14, 15). Since the human genome is organized in three dimensions and the chromosomes fold locally driving gene regulation, our tool is integrating the multi-omics data in relation to TADs. Moreover, recent studies have been shown that integrating multi-omics data that included TAD information revealed novel insights into the mechanism of the regulation of genes causing tumor development(16, 17).

We developed an R-based framework, called InterTADs, that integrates the tabular output of multiple experiments of different NGS types (e.g. tables with expression values, mutation and DNA methylation values) into a single file. The tool then combines the joined representation of the multiple experiments, with the 3D organization of the genome, the TADs. Statistical analysis is performed according to predefined groups of interest (e.g. normal cells vs cancer cells) and the events related to multi-omics data (CpG site – CpGs, transcript, mutation, histone marker, etc.) which are divided into the associated TADs based on the overlap of the chromosomal locations. Finally, visualization options are available for the statistically significant results.

Our approach was tested on different omics data sets from 9 patients with Chronic Lymphocytic Leukemia (CLL)(18). CLL is the most common adult leukemia in Western countries(19) and is characterized by clinical and biological heterogeneity at both the genetic and epigenetic levels(20, 21). Due to its great heterogeneity, CLL provides a paradigmatic case to decode complex associations of the events within the same TADs.

Results

Implementation

The proposed method was evaluated on data from CLL stereotyped subsets #6 and #8. These are two distinct subgroups of CLL, with subset #8 showing more aggressive disease than subset #6. In our previous study(18), we explored the DNA methylation values (450K) and the expression values as independent datasets. Here, the RNA-seq raw data were analyzed for variants using the GATK pipeline for Haplotype detection(22).

Finally, we used the matrix of DNA methylation values (based on Illumina 450K BeadChip Arrays), the matrix with expression values (based on RNA-seq), and the variants (based on RNA-Seq) as the input multi-omics file for our tool. Also, TADs from the cell line GM12878(23, 24) were used. The DNA methylation and RNA-seq expression data have been deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under the accession numbers E-MTAB-6955 and E-MTAB-6962, respectively. The three omics data contained 448,372 rows with DNA methylation events (CpGs), 24,192 rows with gene expression events (transcripts) and 1,736,345 rows with variant events. For whole analysis, only events on chromosomes 1 to 22 were included.

Multi-omics data from patients with CLL

The InterTADs was applied on previous published omics data from two distinct and aggressive subgroups of CLL as described in implementation section. The tool includes three main phases: (i) data integration of multi omics data, (ii) TADiff: association with the 3D organization of the genome through the TADs, and (iii) visualization of the statistically significant TADs (**Figure 1A**). The integrated table contains 2,208,909 events i.e. CpG sites, transcripts or variants. The 2,066,262 out of 2,208,909 events were found on 3,029 of the 3,036 TADs in the GM12878 cell line. The analysis was performed according to the predefined metadata file in which the samples are characterized based on the group of interest i.e. stereotyped subset #6 (*ss6*) and stereotyped subset #8 (*ss8*).

First, we calculated the FDR between the two groups on the DNA methylation matrix, on gene expression matrix, on variants matrix and based on our approach regarding the TADs. To demonstrate the ability of the tool in revealing the most significant result compared to the individual dataset, we checked for the distribution of the FDR in each file and the FDR on our approach (**Figure 1B**). Statistically significant results were found only in our approach, thus clearly showing the value of the tools as a data discovery approach.

Regarding the differential analyses, we applied a threshold of 0.05 on FDR. Additionally, the cut off of the differences between the groups regarding the TADs was automatically selected on 16.24, which represent the 3rd quartile and considered as the most significant. We found 459 statistically significant TADs between the two categories, #6 and #8. The most significant TAD based on FDR was TAD2130 (*diff* = 19.8, FDR < 0.0001), which included 4,299 events (CpGs, transcripts and variants), and based on the *diff* was TAD854 (*diff* = 27.7, FDR < 0.0001), which included only 43 events (CpGs and variants).

We continued on the visualization phase focusing on TAD2130, which showed the lower FDR and includes high number of associated events (n = 4,299). The first option of the visualization phase is to provide a dotplot with connecting lines based on the mean of each event is generated as a second option combining with a violin plot of the distribution of the mean values (**Figure 2A**). Also, a dot plots with the cases values of the associated events in the TAD of interest for the two groups separately (**Figure 2B**) is provided. The black line highlights the median in each group. The third option generates scatter plots regarding the chromosomal location of each TAD on x axis and the value of each associated event on y axis. The most significant TAD based on FDR, the TAD2130, was plotted for stereotyped subset #6 and stereotyped subset #8 separately (**Figure 2C, D**).

Benchmarking

The tool was evaluated on its computational time by generating artificial inputs from the original multi-omics data, and comparing it to the quick sort algorithm. For this purpose, a set of experiments were used as benchmarking. During these experiments, fourteen artificial datasets were created, by randomly sampling different number of rows from the original dataset. The datasets consisted of 10,000, 100,000, 150,000, 200,000, 300,000, 500,000, 1,000,000, 1,500,000, 2,000,000, 2,500,000, 3,000,000, 3,500,000, 4,000,000 and 4,500,000 rows (e.g. events; CpG, expression, mutation etc.). All experiments were executed on an SSD drive computer with 16 GB RAM at 1.80 GHz and a 64-bit operating system. **Figure 3A** shows the compute times for the Data integration phase of the algorithm, and **Figure 3B** shows the compute times for the TADiff phase. For smaller dataset's sizes the Data Integration needs more time to produce the required results, while on larger datasets the TADiff part takes longer to terminate. **Figure 3C** depicts the elapsed time for the Data Integration and the TADiff part, given a particular size of dataset produced (400KB), compared to Singleton (1969)'s implementation of Hoare's Quicksort method(25). In general, both of the InterTADs' functions (Data Integration and TADiff) perform relatively good in comparison with the Quicksort algorithm, with TADiff giving better compute times than Data Integration. Particularly, the Quiqcsort method needs almost 0.001 seconds to terminate, the Data Integration script takes ~10 seconds to perform and TADiff almost a second.

Discussion

NGS technologies have impacted massively on the life sciences, especially in cancer research. Through global scientific communities and consortia, such as The Cancer Genome Atlas (TCGA)(26), the International Cancer Genome Consortium (ICGC)(27), BLUEPRINT(28) etc., high-quality data and corresponding metadata of over 20,000 tumor genomes are available worldwide.

Despite the increasing amount of data though, there is no single approach to efficiently integrate multi-omics data that originate from the same source (e.g. patient). Here we propose a novel tool, namely InterTADs, which provides a complete framework to analyze multi-omics data, either available in-house or through public repositories. InterTADs implementation supports very fast execution in order to (i) generate a single file from multi omics inputs, (ii) find significant differences on the TADs between predefined groups of interest, and (iii) visualize the TADs of interest. Our approach clearly supports efficient data discovery in multi-omics data by increasing the statistically significant differences across higher level organizational units (i.e. TADs), as compared to the individual data sets.

In relationship to other existing tools(10), InterTADs is rather on a genome-wide approach by separating the genome into TADs. Applying this approach, we omit the gene level analysis and the coincidence windows analysis, which generates sliding windows within the chromosome by taking into account the chromatin configuration and the high level of interactions within the TADs.

InterTADs is an open-source R package, easily applicable to any type of omics data. The tool is in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable) since it is freely available on GitHub and it is also accompanied by detailed documentation and examples, highlighting the reproducibility of the source code.

Altogether, there are currently several publicly available resources for multi-omics data but also several community-supported bioinformatics tools and databases that can manipulate this kind of data. However, the downstream analysis and especially the integration of different technologies of NGS data is a very promising area since it is still an ongoing process, with no single widely accepted approach. Our method gives a new perspective towards analyzing multi-omics data, by offering short execution time, and meaningful representation of information structure, and clear visualization options.

Conclusions

InterTADs offers a novel software framework for integrating multi omics datasets considering the chromatin configuration, i.e. the interactions within the TADs. It takes advantage of the new aspect of analysis and directly benefits by adhering to the open science principles, in order to produce high quality and reproducible scientific results based on already published data.

Methods

The tool is implemented as an R script. The input multi-omics files are BED-formatted containing the coordinates of each event (mutation, CpG site, transcript etc), and the corresponding score values. In more detail, the input files have to include in the 1st column a unique identifier (e.g. cg00000029, XLOC_032721, mut_1, etc.) for each event, in the next 2nd to 4th columns the BED format information (i.e. chromosome, start, end), and in the rest of the columns the values for each patient. These files are produced by tools performing the analysis of the raw data such as *HISAT2*(29), *featureCounts*, *MACS2*(30), *minfi*(31), *GATK*(22), etc. On our study case, the format of the omics data were transformed to a BED-format adding the scores of each patient using the library *IlluminaHumanMethylation450kanno.ilmn12.hg19* in R for the CpG events and using the GTF file from *StringTie* of the HISAT2 pipeline. In order for the algorithm to run properly, all files are placed into two folders, named *freq* and *counts*, based on the type of information they are carrying (frequency score values or count values). Along with these files a *meta-data* file is created containing information about the mapping between the files' columns.

Workflow

Briefly, the data integration module contains functions for loading, reformatting and scaling of the input files and generates a single table. Subsequently, each event of the integrated table is characterized according to the related gene and the genomic features (exon, intron etc.). Regarding the 3D organization, all events are grouped into corresponding TADs based on the overlap of the chromosomal regions. A statistical analysis is then performed, which includes the evaluation of the differences of the TADs between the predefined groups of interest (e.g. normal cells VS cancer cells), retrieved by a user-provided *meta-data* file. Additionally, visualization scripts produce plots of the events on the chromosomal location of a TAD and dot plots based on the values of the events on a TAD, considering the predefined groups for both options. Our approach can be applied to any kind of NGS or array-based experiment, and any cohort size and integrates with the TAD boundaries using either publicly available Hi-C data or custom resources.

We split the InterTADs workflow into three main phases: (i) automation of the multi-omics data integration, (ii) introduction of the biological knowledge regarding the 3D organization of the genome through the TADs and (iii) visualization of the statistically significant results (**Figure 2A**):

- **Data integration**

The first phase includes the automated process of reading and formatting all inputs into a single file. The input multi-omics files are BED-formatted containing the coordinates of each event (CpG, transcript, mutation etc.), and the corresponding score values. These files are produced by tools performing the analysis of the raw data such as HISAT2(29), MACS2(30), minfi(31), GATK(22) etc. Along with these files, a *meta-data* file is created containing information about the mapping between the files' columns. The Data integration phase consists of four steps (**Figure 1**):

- **Loading**: First, all inputs are read and loaded regardless of the format of each individual file.
- **Reformatting**: Next, each file is transformed into a data table based on the given *meta-data* file. This transformation ensures that same index columns from different tables, point to the same physical source (chromosome information, patient ID etc.).
- **Scaling**: In order for any further analysis to be possible, all tables are transformed so that they correspond to the same scale. A range between $0 - 100$ has been chosen for convenience purposes. Hence, numeric data, which contain frequency score values in the above range, are slightly (DNA methylation) or not at all changed (mutation data). On the other hand, a function is applied to count values so that they correspond to the desired range. The transformation process is as follows; supposing that E corresponds to expression counts, then a logarithmic scale is applied:

  $$E_{log} = \ln(E)$$

  Later on, a vector with all maximum values of the columns of E_{log} is created:

  $$\boldsymbol{E_{max}} = \max_{j} E_{log}, \text{ where } j \text{ refers to column index}$$

  and a new matrix is generated by calculating the ratio between the maximum values and the desired range:

  $$E_{new} = E_{log} \cdot 100/\boldsymbol{E_{max}}$$

- **Gene names / Location:** Finally, for every event on the new integrated matrix, the gene names and locations (exon, intron, cds etc.) are retrieved based on the chromosomal location of the event. This module includes options for either hg19 or hg38 annotation according to the reference genome of the multi omics data.

- **TADiff**

The output file of the Data Integration phase is the main input of the TADiff phase together with a BED file containing information of the TADs. The TADs are conserved sites for a specific cell type and there are several publicly available files on UCSC, on ENCODE project and also Hi-C experiments on GEO DataSets. The chromosomal coordinates of each event of the data integration output are tested for the overlap with the chromosomal coordinates with the TADs. The overlap of the ranges between the events and the TADs is tested using the R package *GenomicRanges*(32). Next, the tool splits the samples in two subgroups, according to a predefined metadata file that includes a list of sample IDs and the corresponding group e.g. normal/tumor. Then, the statistical analysis of the two subgroups includes the calculation of the Benjamini-Hochberg False Discovery Rate (FDR) choosing automatically the parametric or nonparametric test according to the cohort of the metadata file (i.e. >30 samples, parametric method). Also, there is an option for paired data e.g. diagnosis/progression. Additionally, the tool calculates the differences on the values between the two subgroups for each TAD, as follows:

$$mean\begin{pmatrix} mean(subgroup_2)_{event1} - mean(subgroup_1)_{event1} \\ mean(subgroup_2)_{event2} - mean(subgroup_1)_{event2} \\ \vdots \\ mean(subgroup_2)_{eventN} - mean(subgroup_1)_{eventN} \end{pmatrix}_{TAD}$$

where $N = total\ events/TAD$, $subgroup_1$ refers to the data related to the first group and $subgroup_2$ refers to the data related to the second group. Events with no difference between the two subgroups are excluded from the downstream analysis. Finally, the user can select the cutoff of the FDR significance; however, the cutoff of the differences (*diff*) is automatically set to be equal to the 3rd Quartile, based on the distribution of the *diff* value.

- **Visualization**

The visualization phase includes three options: (i) dot plots for the two subgroups based on the values of the cases of the TAD's events, (ii) dot plots for the two subgroups based on the mean values of the cases of each event and (iii) chromosomal representation of TADs.

In more detail, the dot plot takes into account the associated events of a TAD of interest and plot the values of the cases between the two subgroups. The second option is a dot plot based on the mean of each event in each subgroup accompanied by a connecting line. Also, as violin plot is generated on the same plot showing the distribution of the mean values.

The third option takes as input the integrated matrix and a desired chromosomal location, and produces plots showing the chromosomal location of the TAD of interest on the x axis and the associated events combined with their values on the y axis. The plots are generated based on each case separately or on each group. Also, a single plot with the differences of the events between the groups is produced. All plotting functions were generated using *ggplot2*, *gghalves* and *karyoploteR*(33).

Considering all the steps, the InterTADs generates 6 different output files:

- *integrated_table.csv*: A table contains all the events of the input omics data included the ID of the event (e.g. cg02913364, chr1 100503564:T:C etc.), the chromosomal location (e.g. chromosome, start, end), gene names, gene locations (transcript, exon, threeUTR)
- *integrated_table_with_tads.csv*: It is the *integrated_table* adding the information about the TAD in which each event belonging.
- *tad_statistics.csv*: A table with the TAD IDs and the statistical measurements (count, mean, IQR, ttest, Wilcoxon, FDR)
- *integrated_table_with_sign_tads.csv*: A filtered table of *integrated_table_with_tads by the user's cut-off*
- *sign_tad_statistics.csv*: A filtered table of *tad_statistics by the user's cut-off*
- *genes_found.txt*: A list with the genes that were found on the statistically significant TADs

Finally, the size of the input and output files range from 4.19 to 130 MB and 3.28 to 239 MB, respectively (**Table 1**)

Abbreviations

NGS: Next Generation Sequecning; TAD: Topologically Associated Domains; CLL: Chronic lymphocytic leukemia; FDR: False Discovery Rate; TCGA: The Cancer Genome Atlas; ICGC: International Cancer Genome Consortium

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and materials

The datasets analyzed during the current study are available in the in the ArrayExpress database at EMBL-EBI (https://www.ebi.ac.uk/arrayexpress/) under the accession numbers E-MTAB-6955 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6955/) and E-MTAB-6962 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6962/). InterTADs is an open-source tool implemented in R and licensed under the MIT License. The source code and is freely available from https://github.com/nikopech/InterTADs. Detailed instructions could be found on https://github.com/nikopech/InterTADs/wiki.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been produced with the support of the RDA Europe Ambassador programme. It has received funding from the European Union's Horizon 2020 (H2020) research and innovation programme via the RDA EU 4.0 project (ref. GA no. 777388). Additionally, this work was funded by ELIXIR, the research infrastructure for life-science data.

Author contributions

M.T designed the research, performed data and statistical analysis, assisted in data interpretation and wrote the study, N.P. performed data analysis, interpretation and designed the tool; F.P. designed and supervised the research and wrote the study.

Acknowledgements

The authors would like to thank Kostas Stamatopoulos (Institute of Applied Biosciences, Centre for Research and Technology Hellas) for providing the data and for his productive comments on the manuscript, and Jose I. Martin-Subero (Institut d'Investigacions Biomèdiques August Pi i Sunyer, Departamento de Fundamentos Clínicos, Universitat de Barcelona) for his contribution on the TADs knowledge.

Authors' information

Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece

Maria Tsagiopoulou, Nikolaos Pechlivanis, Fotis Psomopoulos

Dept of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden

Fotis Psomopoulos

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304-51.
3. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56(2):61-4, 6, 8, passim.
4. He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. Int J Mol Sci. 2017;18(2).
5. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinform Biol Insights. 2015;9(Suppl 1):29-46.
6. He W, Ju D, Jie Z, Zhang A, Xing X, Yang Q. Aberrant CpG-methylation affects genes expression predicting survival in lung adenocarcinoma. Cancer Med. 2018;7(11):5716-26.
7. Del Real A, Perez-Campo FM, Fernandez AF, Sanudo C, Ibarbia CG, Perez-Nunez MI, et al. Differential analysis of genome-wide methylation and gene expression in mesenchymal stem cells of patients with fractures and osteoarthritis. Epigenetics. 2017;12(2):113-22.
8. Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. 2012;44(11):1236-42.
9. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15(2):R37.
10. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated Omics: Tools, Advances, and Future Approaches. J Mol Endocrinol. 2018.
11. Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Brief Bioinform. 2019.
12. Kim SY, Jeong HH, Kim J, Moon JH, Sohn KA. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. Biol Direct. 2019;14(1):8.
13. Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun. 2020;11(1):735.
14. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376-80.
15. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10(8):1297-309.
16. Speedy HE, Beekman R, Chapaprieta V, Orlando G, Law PJ, Martin-Garcia D, et al. Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics. Nat Commun. 2019;10(1):3615.
17. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stutz AM, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet. 2017;49(1):65-74.
18. Papakonstantinou N, Ntoufa S, Tsagiopoulou M, Moysiadis T, Bhoi S, Malousi A, et al. Integrated epigenomic and transcriptomic analysis reveals TP63 as a novel player in clinically aggressive chronic lymphocytic leukemia. Int J Cancer. 2019;144(11):2695-706.
19. Kipps TJ, Stevenson FK, Wu CJ, Croce CM, Packham G, Wierda WG, et al. Chronic lymphocytic leukaemia. Nat Rev Dis Primers. 2017;3:17008.
20. Guieze R, Wu CJ. Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. Blood. 2015;126(4):445-53.

21. Tsagiopoulou M, Papakonstantinou N, Moysiadis T, Mansouri L, Ljungstrom V, Duran-Ferrer M, et al. DNA methylation profiles in chronic lymphocytic leukemia patients treated with chemoimmunotherapy. Clin Epigenetics. 2019;11(1):177.

22. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8.

23. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.

24. Beekman R, Chapaprieta V, Russinol N, Vilarrasa-Blasi R, Verdaguer-Dot N, Martens JHA, et al. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat Med. 2018;24(6):868-80.

25. Singleton RC. An Efficient Algorithm for Sorting with Minimal Storage. Commun Acm. 1969;12(3):185-+.

26. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77.

27. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464(7291):993-8.

28. Martens JH, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica. 2013;98(10):1487-9.

29. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357-60.

30. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

31. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-9.

32. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.

33. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017;33(19):3088-90.

Figures and Tables

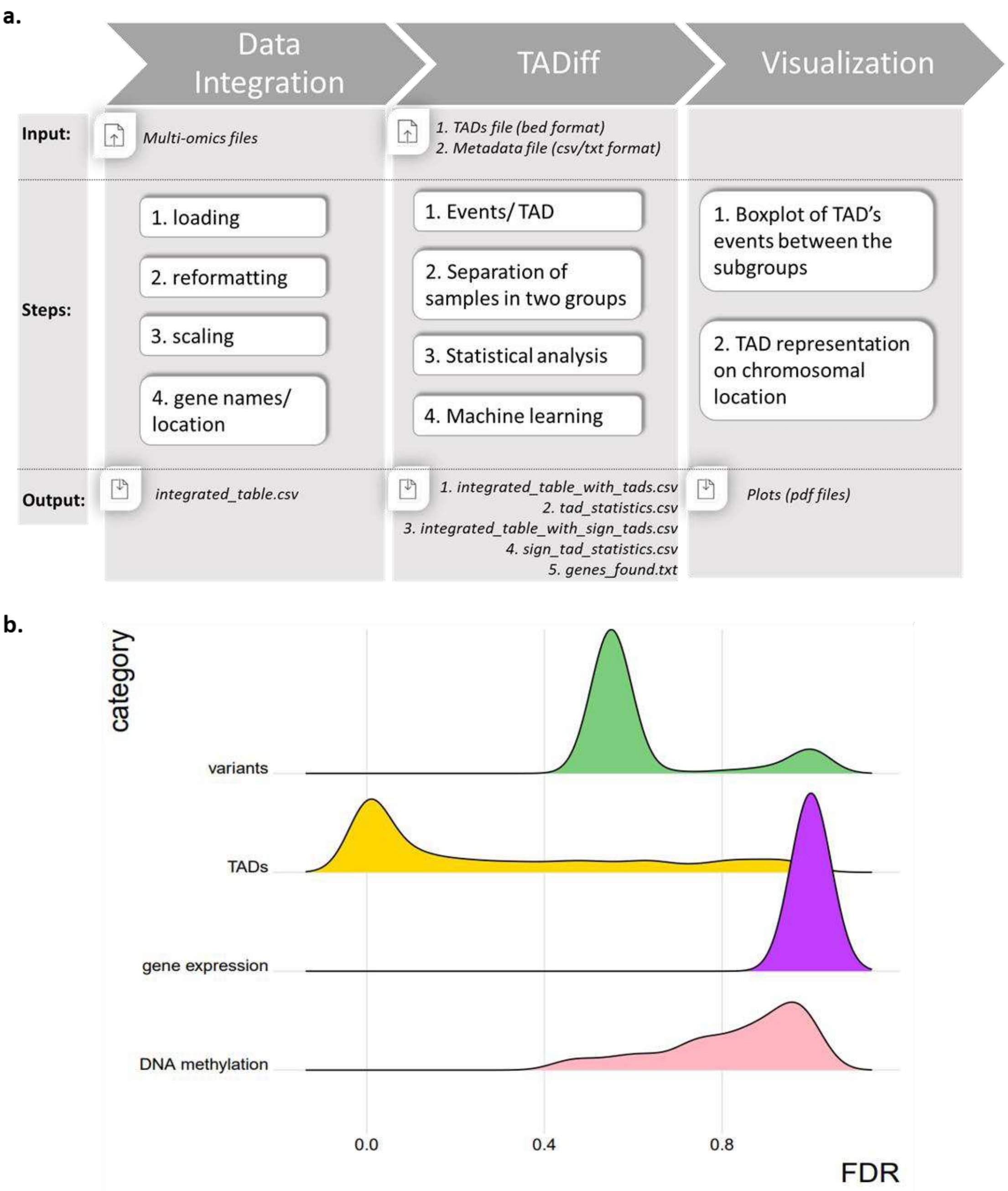

Figure 1. A. Schematic diagram of InterTADs. Input multi-omics data are loading, reformatting, scaling and annotating to genome in order to generate a single integrated file. The events of the integrated file were mapped on the TADs and statistics are provided. Finally, plotting functions are available. and summary statistics supplement the data preprocessing module. **B.** Boxplots showing the different omics-data (DNA methylation, gene expression, variants) and also our proposed approach (TADs) (y-axis) and the statistically significant results comparing the two subgroups (FDR) (x-axis)

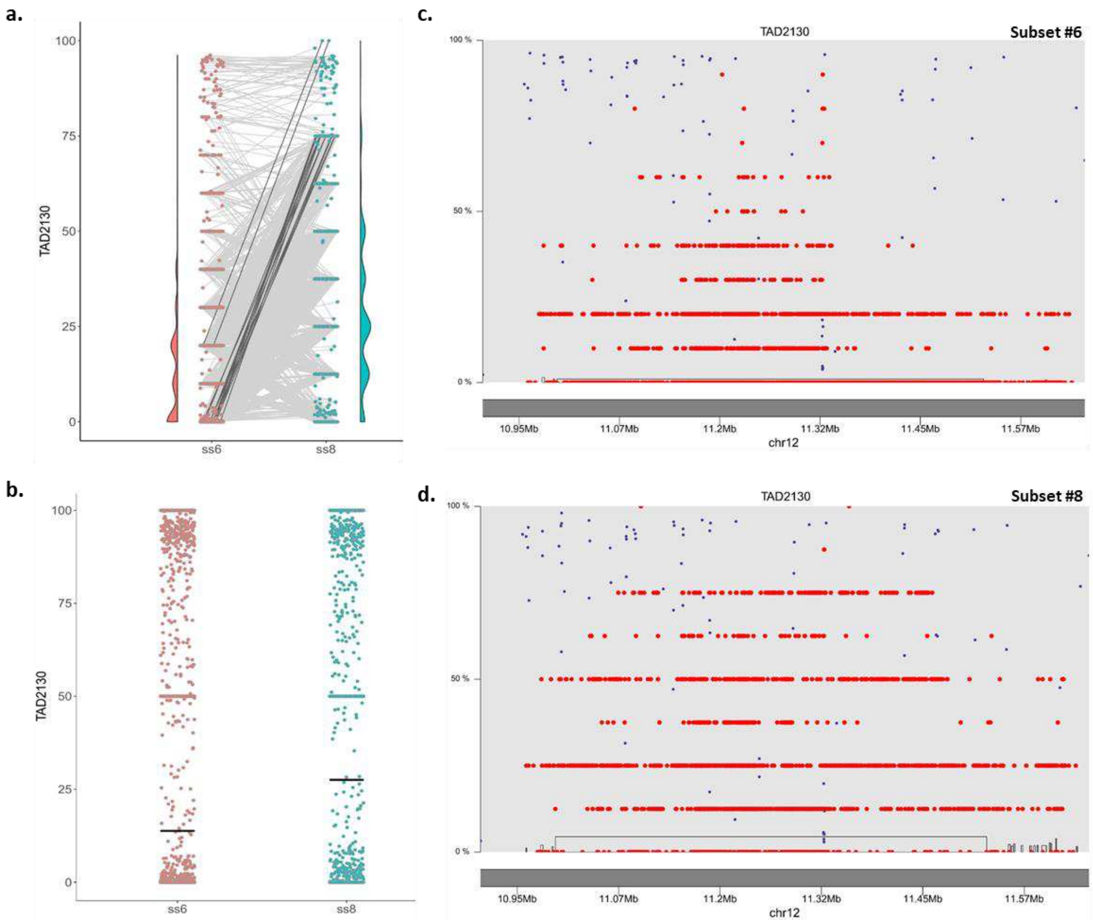

Figure 2. A. Dot plot of the mean values per event, with lines connecting related events between stereotyped subset #6 (red) and stereotyped subset #8 (blue). The plot visualizes the most significant TAD based on FDR (TAD2130)**.** The bold lines represent the top 30 events which showed the greatest difference between the two subgroups. The violin plots on the side of each scatterplot show the distribution of the mean values on the two subgroups **B. Dot plot with the mean of all events** belonging in stereotyped subset #6 (red) and in stereotyped subset #8 (blue). The plot visualizes the most significant TAD based on FDR (TAD2130). The black line corresponds to the mean of the all values in the particular subset (#6 or #8). **C-D. Scatter plot showing the chromosomal region of the TAD2130 (x-axis) and the values of the events (y-axis)** on **C.** stereotyped subsets #6 and **D.** on stereotyped subset #8. The blue boxes state the presence of a transcript, the red dot represents a variant and the black dot a CpG site.

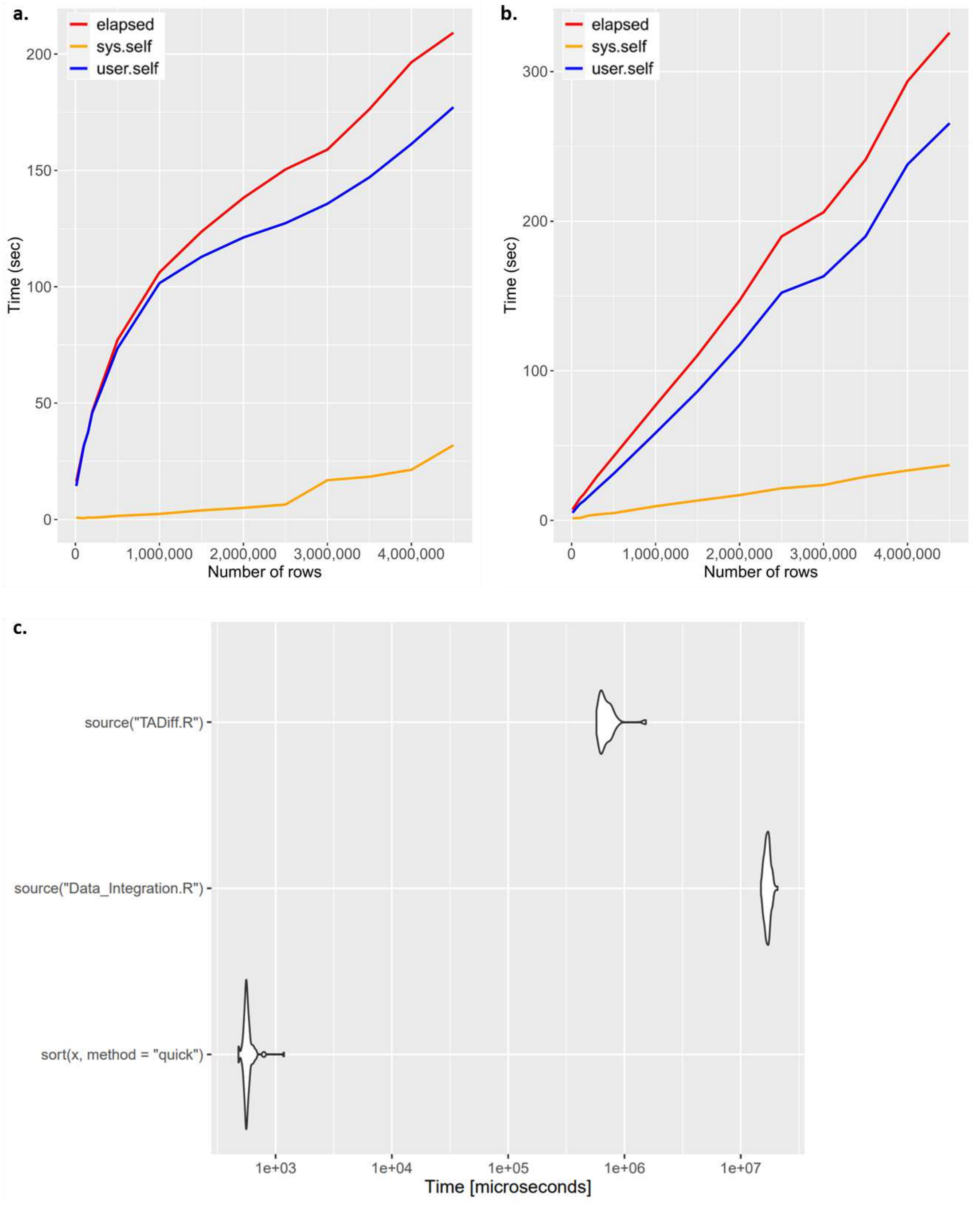

Figure 3.A-B. Line plots showing the time (sec) on y axis and number of rows on the input matrix for **A.** Data Integration phase and **B.** TADiff phase**.** The red line represents the elapsed time, the blue the system time and the yellow the user time. **C. The plot depicts the elapsed time for the Data Integration and the TADiff part compared to Singleton (1969)'s implementation of Hoare's Quicksort method.**

Table 1. A list of the input and output files combined with the size of the samples

Input files	
Methylation	84.4 MB
Variants	130 MB
Expression	4.19 MB
Output files	
integrated_table.csv	193 MB
integrated_table_with_tads.csv	239 MB
integrated_table_with_sign_tads.csv	38.7 MB
tads_statistics.csv	300 KB
sign_tad_statistics.csv	16.0 KB
genes_found.txt	3.28 MB

Figures

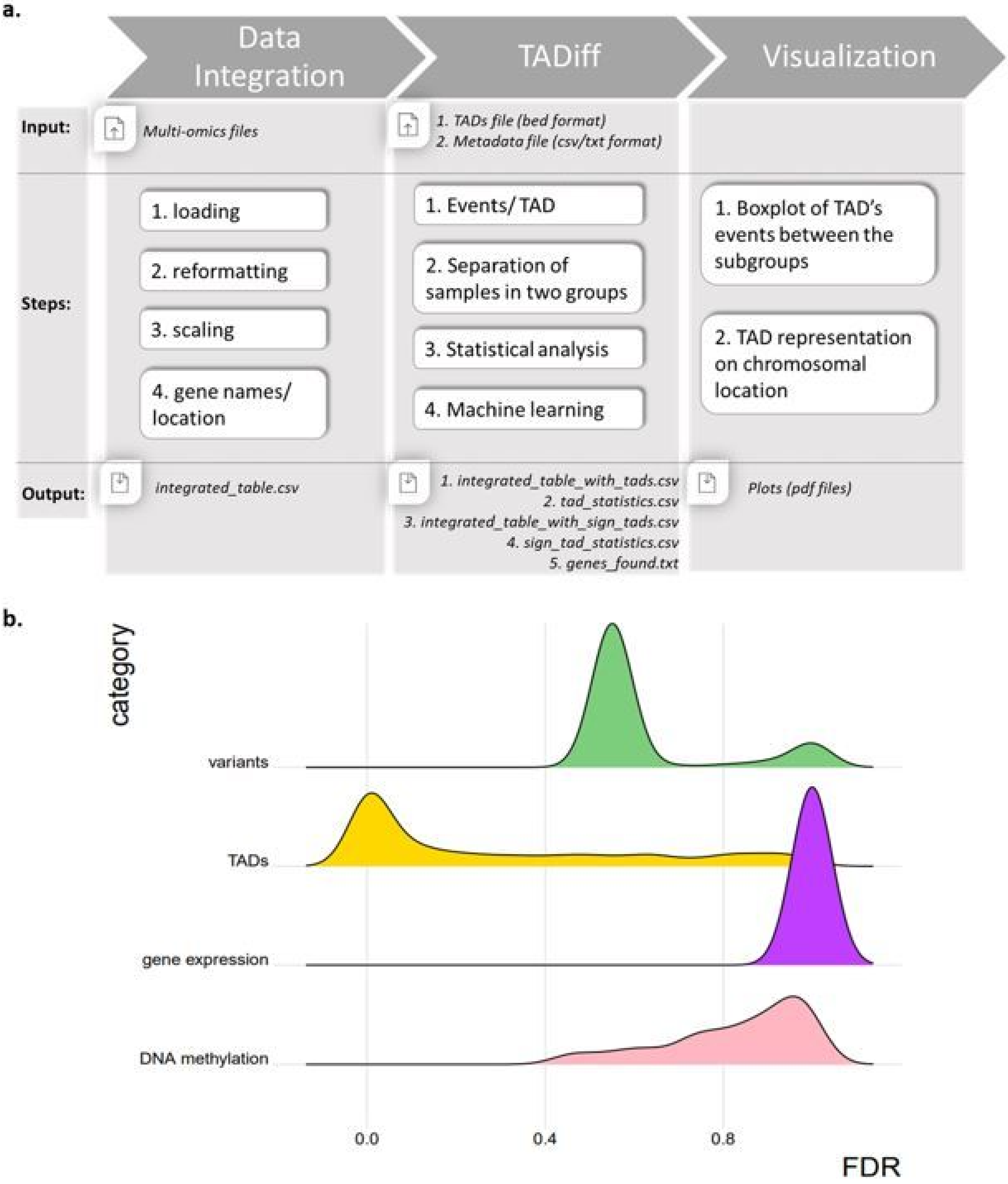

Figure 1

A. Schematic diagram of InterTADs. Input multi-omics data are loading, reformatting, scaling and annotating to genome in order to generate a single integrated file. The events of the integrated file were mapped on the TADs and statistics are provided. Finally, plotting functions are available. and summary statistics supplement the data preprocessing module. B. Boxplots showing the different omics-data (DNA methylation, gene expression, variants) and also our proposed approach (TADs) (y-axis) and the statistically significant results comparing the two subgroups (FDR) (x-axis)

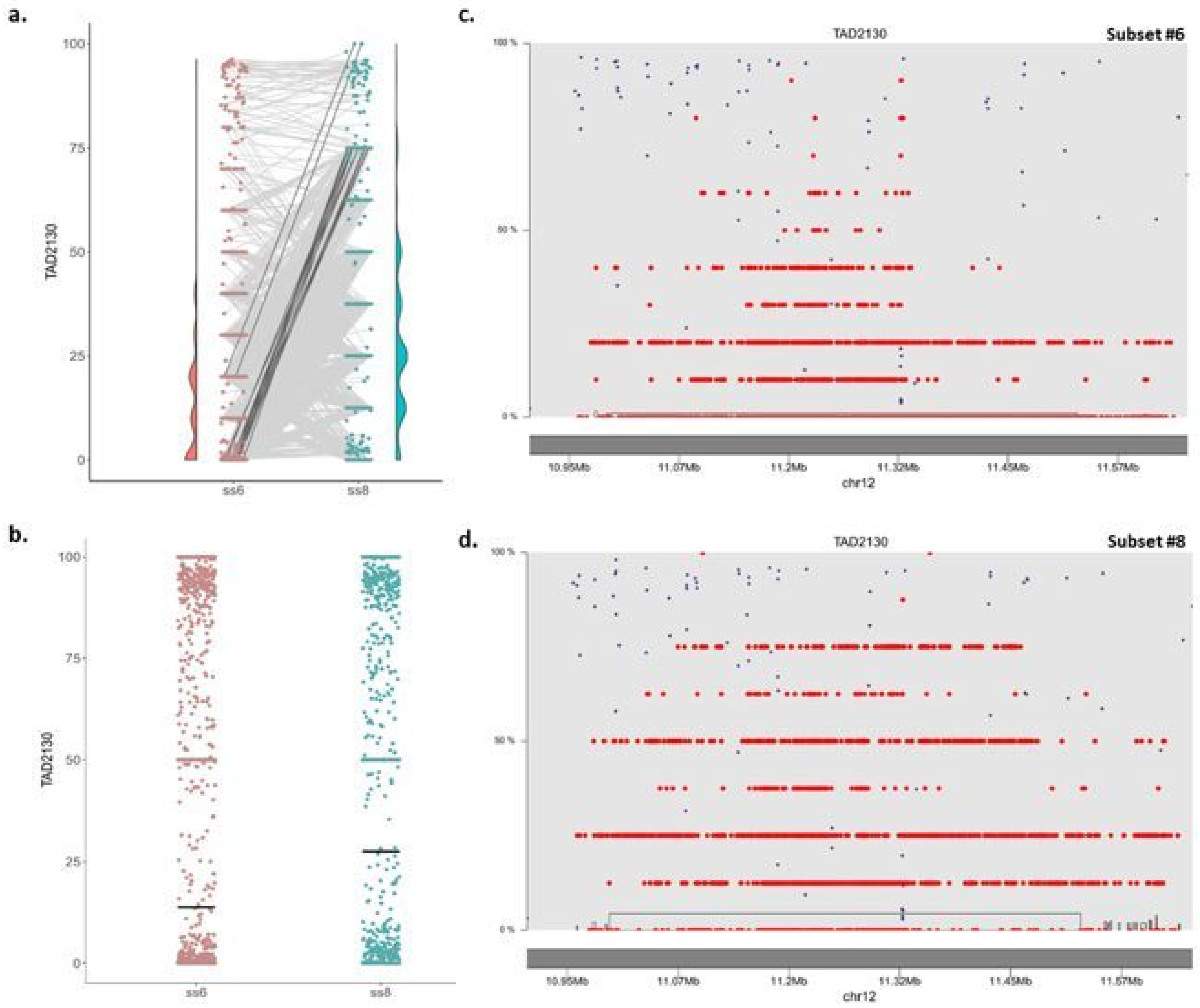

Figure 2

A. Dot plot of the mean values per event, with lines connecting related events between stereotyped subset #6 (red) and stereotyped subset #8 (blue). The plot visualizes the most significant TAD based on FDR (TAD2130). The bold lines represent the top 30 events which showed the greatest difference between the two subgroups. The violin plots on the side of each scatterplot show the distribution of the mean values on the two subgroups B. Dot plot with the mean of all events belonging in stereotyped subset #6 (red) and in stereotyped subset #8 (blue). The plot visualizes the most significant TAD based on FDR (TAD2130). The black line corresponds to the mean of the all values in the particular subset (#6 or #8). C-D. Scatter plot showing the chromosomal region of the TAD2130 (x-axis) and the values of the events (y-axis) on C. stereotyped subsets #6 and D. on stereotyped subset #8. The blue boxes state the presence of a transcript, the red dot represents a variant and the black dot a CpG site.

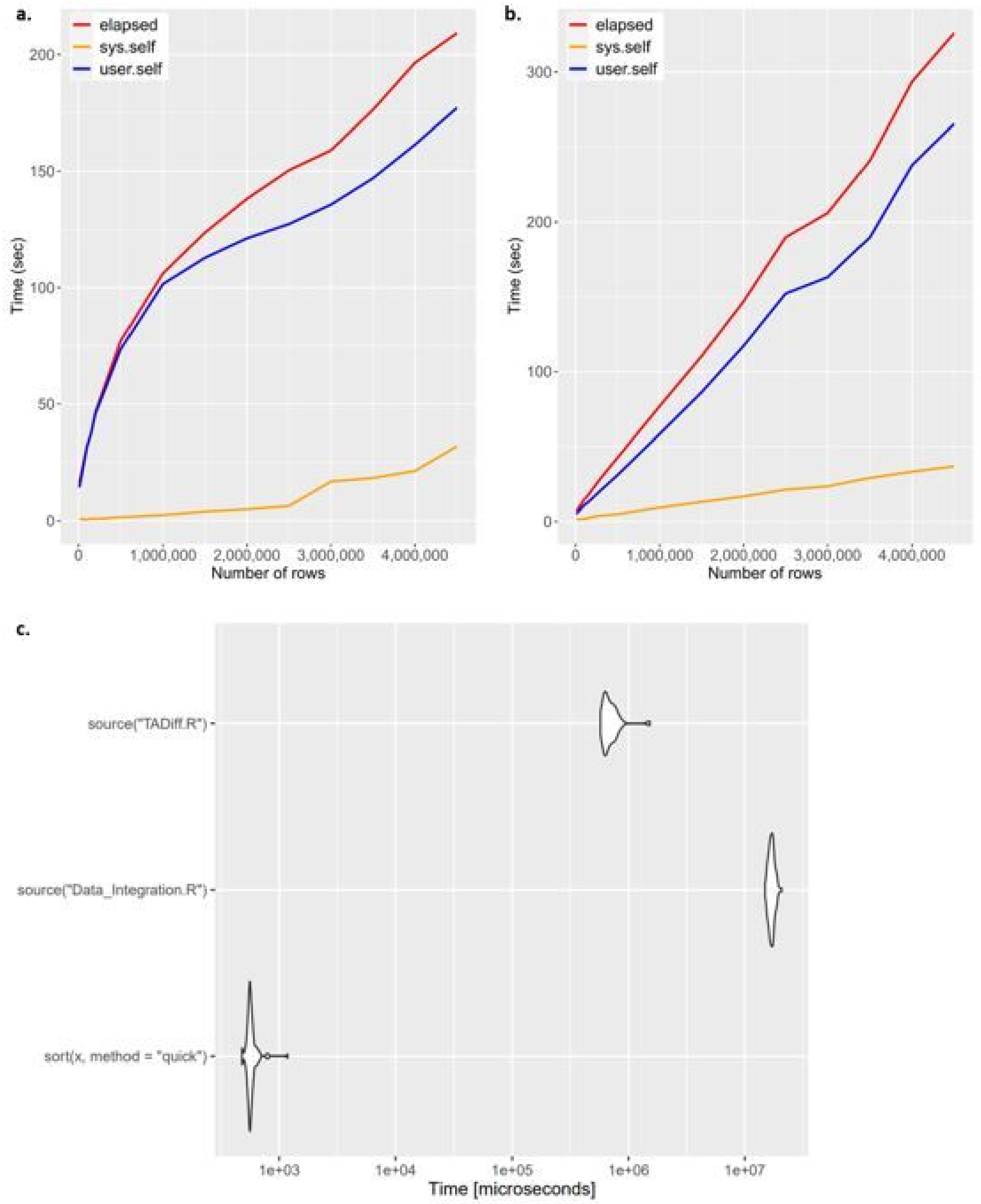

Figure 3

A-B. Line plots showing the time (sec) on y axis and number of rows on the input matrix for A. Data Integration phase and B. TADiff phase. The red line represents the elapsed time, the blue the system time and the yellow the user time. C. The plot depicts the elapsed time for the Data Integration and the TADiff part compared to Singleton (1969)'s implementation of Hoare's Quicksort method.