Technical note

Multi-omics integrations and Machine Learning analysis of datasets deposited at TriTrypDB database

Michele Tinti and Michael A. J. Ferguson

The Wellcome Centre for Anti-Infectives Research, School of Life Sciences, University of Dundee, Dundee, UK

Affiliation

Highlights

Abstract

The number of Proteomics and Genomics experiments is growing. This large amount of data poses new challenges to the scientific community. In particular, the amount of available data is not matched by the capacity of common bioinformatics tools for retrieving and extracting useful insights from the publicly accessible datasets. Machine learning can mitigate this issue by providing powerful analysis tools. However, data needs to be prepared and formatted for machine learning analysis. In this paper, we share with the trypanosome research community a series of Jupyter notebooks to demonstrate the utility of machine learning applied to the analysis of our own OMIC experiments and those deposited at the TryTripDB database.

Abbreviations

ML, machine learning; PCF, procyclic form; BSF bloodstrem form; LS, long slender; SS short stumpy; ATF, adipose tissue forms; PRIDE, proteomics identifications database; GEO, Gene Express Omnibus; ENA, European Nucleotide Archive; FAIR, findability, accessibility, interoperability, and reusability; SILAC, stable isotope labeling by/with amino acids in cell culture; RNA, ribonucleic acid; RNA-seq, RNA-sequencing; RNAi, RNA interference; SEC, size-exclusion chromatography; MW, molecular weight; CSV, comma-separated values; SNP, single nucleotide polymorphism; GFF, general feature format; CV, cross-validation; ROC, operating characteristic curve; AUC, area under the curve; SHAP Shapley additive explanations; EF, Experimental Feature; AF, amino acid feature; ELM, eukaryotic linear motif; HT, High-throughput.

Keywords

Trypanosoma; Machine Learning; OMIC dataset; Prediction; Data Analysis; Data Visualization

Text

The Wellcome Centre for Anti-Infectives Research in Dundee UK is active in fundamental and translational research, including drug discovery and the development of diagnostics, for neglected tropical diseases (NTDs). These NTDs include human African trypanosomiasis (or sleeping sickness) caused by tsetse-transmitted *Trypanosoma brucei* parasites [[1](#_ENREF_1)]. The parasite multiplies as procyclic form (PCF) in the tsetse midgut and migrates to the salivary glands where it differentiates via an epimastigote form to a non-dividing metacyclic trypomastigote form that is adapted for transmission to a mammalian host during a bloodmeal. Once in the bloodstream, the parasite differentiates to the actively replicating long slender (LS) bloodstream form (BSF). Some of these differentiate into non-dividing stumpy forms (SS) that are adapted for differentiation back to PCF trypanosomes once taken up by a tsetse fly in a subsequent bloodmeal. Over the years, scientists have performed several OMICs (genomic, transcriptomic and proteomic) experiments on *T. brucei* to get better insights into the biology of this parasite. The OMICs datasets produced on *T. brucei* are generally uploaded by researchers in open repositories such as the PRoteomics IDEntifications database (PRIDE) for proteomics datasets or the Gene Express Omnibus (GEO) and European Nucleotide Archive (ENA) databases among the others [[2-6](#_ENREF_2)] for transcriptomics and genomics datasets. Since its establishment in early 2009, the TriTrypDB database has gone to a great deal of effort to retrieve data from open repositories and to annotate and share back these data with the trypanosome community all these OMICs datasets in one integrated platform [[7](#_ENREF_7)]. At the time of writing, the TriTrypDB database (version 45) contains more than sixty OMICs experiments performed in *T. brucei*. The data within TriTrypDB can be interrogated with search strategies allowing the construction of complex queries [[7](#_ENREF_7)]. Conveniently, TriTrypDB provides analysis tools to perform GO term enrichment, metabolic pathway enrichment or word enrichment analyses from a user input series of gene identifiers (gene set). These enrichment strategies depend on either manual or computer-aided (generally inference from the closest annotated orthologue) annotation of gene function. Nevertheless, it can be challenging to combine and fully exploit the potential of every OMICs dataset. For example, several OMICSs datasets deposited at TriTrypDB cannot immediately be used to annotate a gene function, even if the dataset itself carries useful information. For example, our laboratory and others performed proteomics experiments to quantify the amount of protein expressed in the BSF and PCF life stages of *T. brucei* [[8-11](#_ENREF_8)]. These experiments can be used to create a continuous response variable or biological descriptor (feature) reporting the relative fold change in expression between the BSF and PCF life stages. However, the decision boundaries between what is preferentially expressed in PCF or BSF life stages is arbitrary, hampering the creation of a gene set based on such property. On the other hand, Machine Learning (ML) algorithms are potentially able to identify associations between gene sets and biological features by automatically selecting such decision boundaries [[12](#_ENREF_12)]. For this reason, we decided to extract biological features from the OMIC datasets deposited at TriTrypDB to create a ML dataset for *T. brucei*. To achieve this goal, we used all the gene identifiers of *T. brucei* clone TREU927 to query the TriTrypDB database version 45. The search strategy used for this step is available at theTriTrypDB database with the accession identifier d7c5277bb3cfbf2d (<https://tritrypdb.org/tritrypdb/im.do?s=d7c5277bb3cfbf2d>). From the TriTrypDB interface, we downloaded all the available information for this search strategy into a comma-separated values (CSV) file. Broadly speaking, this step assembles a collection of proteomics and transcriptomics (RNA-seq / gene chip) datasets along with several protein sequence descriptors, such as the presence of a signal peptide, the counts of transmembrane domains or the genomic location of the protein gene. Afterwards, we proceeded to assemble an ML dataset from the downloaded CSV file by engineering experimental features (EF) from the dataset deposited at TriTrypDB. The datasets mentioned above studying the differentially expressed proteins between the PCF and BSF life stage of *T. brucei* are present in TritrypDB as six different experiments, reporting the log2 fold changes of the PCF versus the BSF SILAC intensities [[8-11](#_ENREF_8)]. While only one experiment used the Short Stumpy (SS) BSF life stage, the other five experiments compared the Long Slender (LS) BSF life stage to the PCF life stage [[11](#_ENREF_11)]. These five datasets using the LS BSF life stage showed good experimental reproducibility (Pearson coefficients above 0.8). Unsurprisingly, the experiment using the SS BSF life stage showed a lower reproducibility with the other five LS BSF experiments; for this reason, this experiment was removed and used to create a different biological descriptor. We then averaged the experiments using the LS BSF life stage in one descriptor, after normalization using the z-scores, to avoid redundancy in the ML dataset. We also used any available data point for the average to decrease the amount of missing values. Further, we decided to create an absolute protein abundance feature in BSF and PCF taking advantage of a recent experiment published by our laboratory [[13](#_ENREF_13)]. This dataset (not yet uploaded in TryTripDB), is available from the public repository Zenodo [[14](#_ENREF_14)] and uses the SILAC technique to study the BSF and PCF protein half-lives. In this work, fully labelled BSF and PCF parasites grown in medium SILAC culture media (M) were placed in light SILAC culture media (L) to follow the synthesis and degradation rate of the proteome in a time-course experiment. The experiment consisted of seven and nine time points for the BSF and PCF, respectively, with three biological replicates each. The samples of the time course experiment were also mixed 1:1 with fully labelled parasites in heavy SILAC culture media (H) to provide an internal standard for normalization. The protein intensity values of those H labelled samples were used as a proxy for protein abundance, after averaging the H intensity values across the biological replicates and the time points. Another experiment from our laboratory (not yet uploaded in TriTrypDB) aimed at the identification of *T. brucei* protein complexes using size-exclusion chromatography (SEC) and mass spectrometry [[15](#_ENREF_15)]. Briefly, PCF cells were prepared for native protein complex analysis by sonication lysis. The resulting lysates were fractionated separating protein complexes based on their size and shape, and the fractions were analysed by mass spectrometry. From this dataset, we first focus on the experiment using the 30 nm pore column (SEC 300), that efficiently separates protein complexes in the range from 8 kDa to 1.2 MDa. Thanks to protein molecular weight (MW) standards, it is then possible to infer the apparent MW of the proteins in each of the SEC 300 elution fractions. Consequently, we assign to each protein an apparent MW corresponding to the fraction showing the maximum protein abundance. The log2 fold change between the apparent MW and the MW of the protein computed from the primary sequence was taken as a descriptor to describe the molecular state of protein, in high MW complexes (log2 fold change > 0), monomer (log2 fold change around 0) or degraded (log2 fold change < 0). Finally, we created a simple binary feature describing if a protein belongs (1) or does not belong (0) to a protein complex based on the analysis reported in Crozier et al. [[15](#_ENREF_15)].We then looked at the differences of the BSF and PCF transcriptomes by computing the log2 fold changes values of the RNA-seq read counts for total and ribosome-protected RNA using the Vasquez et al. study [[16](#_ENREF_16)]. The same study was used to create an RNA abundance descriptor by computing the log10 values of the total read counts for the transcriptome and ribosome-protected transcriptome of the *T. brucei* genes, in both BSF and PCF life stages. Further descriptors were engineered from the changes in individual protein group abundances within the PCF proteome during cell cycle progression [[17](#_ENREF_17)]. From this time-course experiment, we extracted two categorical features and one continuous variable. The two categorical features were the times of the maximum (MAX) and minimum (MIN) abundance for each protein in the time-course experiment. The continuous variable was the log2 fold change between the MAX and MIN protein abundance values across the time-course experiment. The same feature extraction strategy was further applied to several other time course experiments studying the changes in proteome and transcriptome abundance during *T. brucei* differentiation from the BSF to PCF life stage [[8](#_ENREF_8), [18-22](#_ENREF_18)]. We also assembled biological descriptors to capture the differential express genes between the BSF and ATF (adipose tissue forms) life stages [[23](#_ENREF_23)]. For this experiment, we computed the log2 fold changes between the ATF and BSF samples. We also considered the response of the parasites to glucose depletion by computing three features: The log2 fold changes between the RNA-seq read counts of 1) glucose-fed SS BSF versus glucose-starved SS BSF parasites, 2) glucose-fed LS BSF parasites versus glucose-fed SS BSF parasites and 3) glucose-fed LS BSF parasites versus glucose-starved SS BSF parasites [[24](#_ENREF_24)]. We further created a gene lethality descriptor by using the RNA interference (RNAi) high-throughput phenotyping experiment of Alsford et al. [[25](#_ENREF_25)]. From this study, we extracted the log2 fold change of the parasite RNA-seq read counts before and after three days of RNAi induction. The changes in RNA abundance after overexpressing the wild type RNA helicase DHH1, or after expressing a mutant form of this protein, have been investigating with a cDNA microarrays strategy [[26](#_ENREF_26)]. The dataset is reported in TriTrypDB as the log2 fold changes in signal intensity between the wild type versus the experimentally induced conditions. For this reason, the dataset was transferred over the ML dataset without further processing. A similar cDNA microarray strategy was used to study the *T. brucei* response to heat shock [[27](#_ENREF_27)] and the log2 fold changes between the wild type and treated condition were transferred over to the ML dataset. Similarly, the log2 fold change values of protein abundances (evaluated with label-free mass spectrometry) between the PCF nuclei and the whole-cell lysates [[28](#_ENREF_28)] were transferred over to the ML dataset. All the features described above were derived from experimental evidence, for this reason the names are preceded by an ‘EF’ (Experimental Feature) prefix for easier identification. We further complemented our ML dataset by using the ProFET package to extracts hundreds of features covering the elementary biophysical and sequence-derived attributes of the *T. brucei* proteome [[29](#_ENREF_29)]. Other sequence derived attributes were directly transferred from the CSV file downloaded from TriTrypDB without processing, such as the ortholog and paralog count for each gene. TriTrypDB reports the presence of a signal peptide in the proteome of *T. brucei* by using the SignalP3-NN and SignalP-HMM alghorithms implementend in SignalP 3.0. [[30](#_ENREF_30)]. The SignalP3-NN alghoritm produces two scores (NN Sum, NN D) while the HMM alghorithms output one score (HMM Prob). To create just one descriptor for the presence of a signal peptide, we scaled each score from 0 to 1 and computed the mean. Finally, from the ELM database [31], we downloaded the eukaryotic linear motifs and created a descriptor for each, reporting the number of motif instances in the primary sequence of the *T. brucei* proteome. The AF (amino acid feature) prefix was applied to the feature subsets that were derived from the analysis of amino acid sequences. We also analysed the single nucleotide polymorphism (SNP) information present in TryTripDB. The descriptor 'NonSyn/Syn SNP Ratio All Strains', reporting the ratio of non-synonymous versus synonymous SNPs was transferred to the ML dataset. The other SNP descriptors present in TryTripDB (‘Non-Coding SNPs All Strains', 'SNPs with Stop Codons All Strains', 'NonSynonymous SNPs All Strains', 'Synonymous SNPs All Strains', 'Total SNPs All Strains') were normalized by the gene length. The GFF file available in the TriTrypDB database for *T. brucei* TREU927 was downloaded and analysed to create a descriptor reporting the distance of a gene transcript from its polycystron start site [[31](#_ENREF_31)]. Polycystron boundaries were chosen based on changes in transcript orientation. With this final step, we ended up with a ML dataset of 1407 descriptors.

To provide a usage example for our ML dataset, we tried to identify descriptors able to discriminate between proteins with long or short half-lives. It is essential to remark that our main interest is not in creating a half-life predictor, as we know the ground truth of this variable for the majority of BSF and PCF proteome [[13](#_ENREF_13)]. Instead, we are interested in finding if any of the descriptors incorporated in the ML dataset can shed light onto the mechanisms underlying protein stability of the *T. brucei* proteome. This is analogous to, say, carrying out a GO term enrichment analysis using the short/long half-life gene sets except that in our case we simultaneously query 1407 features. As the first step, we imputed missing values in the categorical variables with a fake -1 class and replaced the missing values of continuous variables with mean values. In our aforementioned turnover paper [[13](#_ENREF_13)], we divided the BSF protein half-lives into decile bins, from the least stable to the most stable proteins. We then selected 837 proteins (fast half-life) from the first two half-life bins (first and second decile) and 837 protein (slow half-life) from the last two half-life bins (ninth and tenth) as the training set for ML analysis. When the number of descriptors (columns of the dataset) is similar or exceed the number of training instances (rows in the dataset), ML algorithms are prone to overfitting, i.e. learning associations with the response variable (fast or long half-lives) originating just by chance [[12](#_ENREF_12)]. Intending to reduce the number of descriptors, we eliminated 277 features that correlate to other features with a Pearson correlation coefficient of 0.75 or higher. We then trained a random forest algorithm implemented in LightGBM [[32](#_ENREF_32)] to discriminate between proteins with fast or slow half-lives. Before training, we removed at random 35% of the training instances (blind) to evaluate the training progression on a subset of instances never seen by the predictor. The LightGBM algorithm was trained with a three-fold cross-validation (CV) strategy. With this strategy, the input data is divided into 3 subsets (also known as folds). Afterwards, The ML algorithm is trained on all but one of the subsets and evaluated on the subset that was not used for training. The process repeats 3 times, with a different subset reserved for evaluation (and excluded from training) each time. In this way, all the data is used once for both training and evaluation. The LightGBM algorithm starts the training with just one tree in the random forest and evaluates the prediction performance using the evaluation subset. With an iterative process, the algorithm adds one tree to the random forest and evaluates the prediction performance again. This process continues until the addition of a tree to the random forest model does not improve the prediction performance (early stopping). At each step, The LightGBM algorithm evaluates the prediction performance using the receiver operating characteristic curve (ROC) area under the curve (AUC) score. At the end of the CV strategy, a new model is trained with all the data used for the CV step and evaluated on the blind dataset. The ROC AUC score computed on the blind dataset (0.69) suggested that the algorithm is separating proteins with fast and long half-lives better than random. The number of trees to evaluate the model on the blind dataset was chosen from the CV strategy. At each CV round, the algorithm records the number of trees used. By inspecting the training history, we selected 14 trees, equal to the maximum number of trees used in any of the three CV steps. This model was used to shrink further the number of relevant features. To this aim, we first applied the Boruta methods [[33](#_ENREF_33)] to identify all the features carrying information usable for prediction (82). Briefly, the Boruta algorithm randomises all descriptors in the dataset and evaluates the randomised features, looking for whether they are less or more important than the real features in separating fast and slow turnover protein. If the real features are better than the shuffled copies, they are marked as important. In a further feature selection, we applied a different permutation importance strategy implemented in the ELI5 python package [[34](#_ENREF_34)]. In this case, only one feature at a time is randomised, and the performance of the model evaluated. If the performance of the model decreases after randomisation (lower AUC score), the feature is annotated as important. The randomization process was performed three times with a different random seed and the average feature score difference stored. As the ELI5 strategy is more computationally expensive than the Boruta algorithm, the former was executed before the latter. From the ELI5 ranked list of features, we selected an arbitrary threshold of 0.015 (increase in ROC AUC score ) to select 5 descriptors, namely: 1) 'EF\_Abundance\_BSF', the relative protein abundance in the BSF life stage; 2) 'EF\_CellCycle\_min', the time point showing the minimum protein abundance during the PCF cell cycle; 3) 'EF\_fc\_glucose\_1', the log 2 fold change in abundance between glucose-fed and glucose-starved SS BSF cells; 4) 'EF\_merged\_PCF\_BSF, the log 2 fold change of the PCF versus the BSF protein abundance, 5) 'EF\_CellCycle\_min\_max', the log2 fold change between the minimum and maximum protein abundance value during the PCF cell cycle. The importance of these selected features was visualised with the SHAP (SHapley Additive exPlanations) values (Figure 1). The SHAP values aggregate five different ranking feature methods into one, by using an algorithm first developed for game theory. This algorithm aims at explaining the prediction of an instance (protein) by computing the contribution of each feature to the prediction [[35-37](#_ENREF_35)]. In particular, for each feature this framework assigns a SHAP value to the proteins (positive for fast half-lives, negative for long half-lives), and the absolute sum of the SHAP values adds to the total importance of the feature. Consequently, the feature importance can be visualised along with the feature effect. As illustrated in Figure 1, the position of the feature on the y-axis of is determined by the feature importance, and the dots (protein/training instances) on the x-axis are arranged accordingly to the SHAP values. The dots are also jittered in the y-axis direction to assess the distribution. The colour code of the dots reflects the value of the feature from low (blue) to high (red). This visualization highlights the relationship between the value of a feature and the impact on the prediction. For example, training instances with low values for the protein abundance in the BSF life stage ('EF\_Abundance\_BSF') have the highest SHAP values, meaning that low protein abundance values are pushing the model towards a fast half-life prediction. The exact shape of this relationship is visualized in Figure 2 where we plotted on the x-axis the feature values and on the y-axis the SHAP values for each protein and feature in the training dataset. The relationship between protein turnover and protein abundance was already identified in our paper confirming the reliability of the described ML approach [[13](#_ENREF_13)]. In the same paper, we observed several cell-cycle regulated proteins with short half-lives. With approach described here, we could identify two descriptors, 'EF\_CellCycle\_min\_max' and 'EF\_CellCycle\_min', related to cell cycle regulation. The EF\_CellCycle\_min\_max reports the maximum fold change of the proteins during the cell cycle. For this descriptor, the higher is the protein abundance fold-change during the cell cycle, the higher the SHAP values, pushing the prediction towards short half-life (Fig 2B). Also, the 'EF\_CellCycle\_min' descriptor reports the time point where the protein shows the minimum abundance value. In Fig 2C it is possible to observe that proteins with a minimum abundance at the beginning of the cell cycle (G1, 0.5h) have higher SHAP values, directing the prediction towards short half-life (Fig 1 and 2C) predictions. While the relationships between protein abundance and cell cycle regulation with protein turnover were already identified in our paper [[13](#_ENREF_13)], our ML analysis suggests two new biological relationships, with glucose regulated proteins and with proteins differentially expressed between the PCF and BSF life stage. The 'EF\_merged\_PCF\_BSF’ descriptor reports the fold change in abundance of the *T. brucei* proteome between the PCF and BSF life stage. This descriptor show that both positive and negative high values of this descriptor are associated with high SHAP values, meaning that protein that are mostly expressed only in the PCF or BSF life stage are likely to have shorter half-lives Fig2A. This makes some sense as life-stage specific proteins need to be promptly removed during the progression from one life stage to the other, a task that would be impossible for proteins with long half-lives. Finally, the EF\_fc\_glucose\_1 descriptor reports the transcriptional changes of the SS BSF life stage after glucose starvation [[24](#_ENREF_24)]. Low values of this descriptor are associated with high SHAP values and consequently with short half-life predictions Fig2E. This descriptor is likely to capture genes under strong regulation when the SS form perceives a decrease in glucose level and starts the progression to the PCF form. Again, it is reasonable to expect short half-lives for this group of genes as they are likely to be promptly regulated upon the dramatic environmental changes associated with the transition from the human host to the vector host.

The computational strategy to assemble the ML dataset and the application example given here are shared as series of Jupyter notebooks in the GitHub https://github.com/mtinti/TriTrypML\_code to meet the FAIR standards of findability, accessibility, interoperability, and reusability, as already applied in other OMICs initiatives [[38](#_ENREF_38)]. With this work, we hope to kick start an interest in ML technique applied to the analysis of OMICS experiments in *T. brucei*. It should be fairly easy for user with some coding experience to accommodate new features and different analysis pipeline strategies using the notebooks as a starting template. In the future, we plan to offer a modified version of this analysis pipeline as a web service, if this work finds enough interest in the trypanosome community.

Funding

This work was supported by the Wellcome Trust Investigator Award to M.A.J.F. (101842).

Figure Legends

Figure 1. Feature importance.

The figure shows the SHAP values for each protein (x-axis) and for each feature (y-axis) used to classify fast and slow turnover proteins (dots). The relative position of the features on the y-axis is computed using the absolute sum of protein (dots) SHAP values. The colour code of the dots reflects the value of the feature from low (blue) to high (red). The dots are also jittered in the y-axis direction to assess the values distribution.

Figure 2. Feature values and SHAP values correlation.

The figure plots the SHAP values (y-axis) and the feature values (x-axis) for each feature (A to E) identified in the feature selection step and used to classify fast and slow turnover proteins (dots).

References

[1] Cox FE. History of sleeping sickness (African trypanosomiasis). Infectious disease clinics of North America. 2004;18:231-45.

[2] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30:207-10.

[3] Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic acids research. 2019;47:D442-D50.

[4] Silvester N, Alako B, Amid C, Cerdeno-Tarraga A, Clarke L, Cleland I, et al. The European Nucleotide Archive in 2017. Nucleic acids research. 2018;46:D36-D40.

[5] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic acids research. 2018;46:D41-D7.

[6] Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, et al. DNA Data Bank of Japan: 30th anniversary. Nucleic acids research. 2018;46:D30-D5.

[7] Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic acids research. 2010;38:D457-62.

[8] Dejung M, Subota I, Bucerius F, Dindar G, Freiwald A, Engstler M, et al. Quantitative Proteomics Uncovers Novel Factors Involved in Developmental Differentiation of Trypanosoma brucei. PLoS Pathog. 2016;12:e1005439.

[9] Urbaniak MD, Martin DM, Ferguson MA. Global quantitative SILAC phosphoproteomics reveals differential phosphorylation is widespread between the procyclic and bloodstream form lifecycle stages of Trypanosoma brucei. Journal of proteome research. 2013;12:2233-44.

[10] Urbaniak MD, Guther ML, Ferguson MA. Comparative SILAC proteomic analysis of Trypanosoma brucei bloodstream and procyclic lifecycle stages. PloS one. 2012;7:e36619.

[11] Gunasekera K, Wuthrich D, Braga-Lagache S, Heller M, Ochsenreiter T. Proteome remodelling during development from blood to insect-form Trypanosoma brucei quantified by SILAC and mass spectrometry. BMC genomics. 2012;13:556.

[12] Chicco D. Ten quick tips for machine learning in computational biology. BioData mining. 2017;10:35.

[13] Tinti M, Güther MLS, Crozier TWM, Lamond AI, Ferguson MAJ. Proteome turnover in the bloodstream and procyclic forms of Trypanosoma brucei measured by quantitative proteomics [version 1; peer review: awaiting peer review]. Wellcome Open Res. 2019;4:152.

[14] Tinti M. https://doi.org/10.5281/zenodo.3417326.

[15] Crozier TWM, Tinti M, Larance M, Lamond AI, Ferguson MAJ. Prediction of Protein Complexes in Trypanosoma brucei by Protein Correlation Profiling Mass Spectrometry and Machine Learning. Molecular & cellular proteomics : MCP. 2017;16:2254-67.

[16] Vasquez JJ, Hon CC, Vanselow JT, Schlosser A, Siegel TN. Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. Nucleic acids research. 2014;42:3623-37.

[17] Crozier TWM, Tinti M, Wheeler RJ, Ly T, Ferguson MAJ, Lamond AI. Proteomic Analysis of the Cell Cycle of Procylic Form Trypanosoma brucei. Molecular & cellular proteomics : MCP. 2018;17:1184-95.

[18] Jensen BC, Sivam D, Kifer CT, Myler PJ, Parsons M. Widespread variation in transcript abundance within and across developmental stages of Trypanosoma brucei. BMC genomics. 2009;10:482.

[19] Queiroz R, Benz C, Fellenberg K, Hoheisel JD, Clayton C. Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. BMC genomics. 2009;10:495.

[20] Kabani S, Fenn K, Ross A, Ivens A, Smith TK, Ghazal P, et al. Genome-wide expression profiling of in vivo-derived bloodstream parasite stages and dynamic analysis of mRNA alterations during synchronous differentiation in Trypanosoma brucei. BMC genomics. 2009;10:427.

[21] Naguleswaran A, Doiron N, Roditi I. RNA-Seq analysis validates the use of culture-derived Trypanosoma brucei and provides new markers for mammalian and insect life-cycle stages. BMC genomics. 2018;19:227.

[22] Archer SK, Inchaustegui D, Queiroz R, Clayton C. The cell cycle regulated transcriptome of Trypanosoma brucei. PloS one. 2011;6:e18425.

[23] Trindade S, Rijo-Ferreira F, Carvalho T, Pinto-Neves D, Guegan F, Aresta-Branco F, et al. Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice. Cell host & microbe. 2016;19:837-48.

[24] Qiu Y, Milanes JE, Jones JA, Noorai RE, Shankar V, Morris JC. Glucose Signaling Is Important for Nutrient Adaptation during Differentiation of Pleomorphic African Trypanosomes. mSphere. 2018;3.

[25] Alsford S, Turner DJ, Obado SO, Sanchez-Flores A, Glover L, Berriman M, et al. High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. Genome research. 2011;21:915-24.

[26] Kramer S, Queiroz R, Ellis L, Hoheisel JD, Clayton C, Carrington M. The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes. Journal of cell science. 2010;123:699-711.

[27] Kramer S, Queiroz R, Ellis L, Webb H, Hoheisel JD, Clayton C, et al. Heat shock causes a decrease in polysomes and the appearance of stress granules in trypanosomes independently of eIF2(alpha) phosphorylation at Thr169. Journal of cell science. 2008;121:3002-14.

[28] Goos C, Dejung M, Janzen CJ, Butter F, Kramer S. The nuclear proteome of Trypanosoma brucei. PloS one. 2017;12:e0181884.

[29] Ofer D, Linial M. ProFET: Feature engineering captures high-level protein functions. Bioinformatics. 2015;31:3429-36.

[30] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. Journal of molecular biology. 2004;340:783-95.

[31] Clayton C. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. Open biology. 2019;9:190072.

[32] Guolin K, Meng Q, Finley T, Wang W, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. 2017;30:3149-57.

[33] Miron B, Kursa R, Witold R, Rudnicki R. Feature Selection with the Boruta Package. Journal of statistical software. 2010;36:1-13.

[34] ELI5. https://github.com/TeamHG-Memex/eli5.

[35] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 2017;30:4765-74.

[36] Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering. 2018;2:749-60.

[37] Molnar M. Interpretable Machine Learning https://christophm.github.io/interpretable-ml-book/shap.html.

[38] Mendez KM, Pritchard L, Reinke SN, Broadhurst DI. Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. Metabolomics : Official journal of the Metabolomic Society. 2019;15:125.