PLOS ONE

RESEARCH ARTICLE

OpenStats: A robust and scalable software package for reproducible analysis of high-throughput phenotypic data

Hamed Haselimashhadi[1]*, Jeremy C. Mason[1], Ann-Marie Mallon[2], Damian Smedley[3], Terrence F. Meehan[1], Helen Parkinson[1]

1 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, **2** MRC Harwell Institute, Harwell Campus, Oxfordshire, United kingdom, **3** Queen Mary University of London, London, United Kingdom

* hamedhm@ebi.ac.uk

Citation: Haselimashhadi H, Mason JC, Mallon A-M, Smedley D, Meehan TF, Parkinson H (2020) OpenStats: A robust and scalable software package for reproducible analysis of high-throughput phenotypic data. PLoS ONE 15(12): e0242933. https://doi.org/10.1371/journal.pone.0242933

Editor: Giorgio F. Gilestro, Imperial College London, UNITED KINGDOM

Received: May 30, 2020

Accepted: November 12, 2020

Published: December 30, 2020

Data Availability Statement: The data that support the findings of this study are publicly available in the International Mouse Phenotyping Consortium (IMPC) portal at https://www.mousephenotype.org. The following link is a direct link to the data access page: https://www.mousephenotype.org/help/non-programmatic-data-access/.

Funding: This work was supported by the NIH Common Fund UM1-HG006370. We further confirm that the research reported in this publication was supported by the European

Abstract

Reproducibility in the statistical analyses of data from high-throughput phenotyping screens requires a robust and reliable analysis foundation that allows modelling of different possible statistical scenarios. Regular challenges are scalability and extensibility of the analysis software. In this manuscript, we describe OpenStats, a freely available software package that addresses these challenges. We show the performance of the software in a high-throughput phenomic pipeline in the International Mouse Phenotyping Consortium (IMPC) and compare the agreement of the results with the most similar implementation in the literature. OpenStats has significant improvements in speed and scalability compared to existing software packages including a 13-fold improvement in computational time to the current production analysis pipeline in the IMPC. Reduced complexity also promotes FAIR data analysis by providing transparency and benefiting other groups in reproducing and re-usability of the statistical methods and results. OpenStats is freely available under a Creative Commons license at www.bioconductor.org/packages/OpenStats.

Introduction

Statistics is the main inferential tool used in science and medicine to extract information from data. It provides a set of proven steps for drawing conclusions and making decisions in spite of the uncertainty inherent in any data, which are unavoidable due to biological variation as well as the constraints of cost, time, and measurement precision. The inference made from the data is subject to reproducibility in the analysis requiring precise, transparent, comprehensive and well-documented rules to prevent unreliable, costly and even invalid results [1]. Reviewing the literature shows that reproducibility in the data and analysis is the subject of an extensive range of publications in different areas of science, e.g., life science, bioscience, medical and pharmaceutical science and translational science [2–6]. Studies have shown irreproducibility of results is often due to poor documentation of the statistical method [7–9]. This is especially critical for the high-throughput phenomic screening when tens of thousands of data points are generated and analysed.

Molecular Biology Laboratory core funding and the National Human Genome Research Institute of the National Institutes of Health under Award Number UM1HG006370. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

The International Mouse Phenotyping Consortium (IMPC www.mousephenotype.com) is a G-7 recognised global research infrastructure dedicated to generating and characterising a knockout mouse line for every protein-coding gene [10–13]. Currently, in its 8th year, the IMPC has phenotyped over 188K+ knockout and 69K+ control mice across 13 research centres from 9 countries in 3 continents (data release 11, February 2020, www.mousephenotype.org/data/release). These centres adhere to a set of standardised phenotype assays defined in the International Mouse Phenotyping Resource of Standardised Screens (IMPReSS www.mousephenotype.org/impress) that represents over 496 (116 IMPC specific) procedures and 10,266 (3,054 IMPC specific) parameters measured on mice. As part of the operating design protocol, critical factors that can impact data collection such as reagent type or equipment are captured as mandatory metadata. Phenotype data is then centrally collected and quality controlled by trained professionals before being released for statistical analyses.

The data is then processed by PhenStat [14], a freely-available R [15] package that provides a set of statistical methods for the identification of genotype to phenotype associations by comparing mutants to controls [16]. PhenStat imposes the same statistical model on the entire continuous data regardless of the nature of the original measurements. That is, the continuous measurements are analysed using a Linear Mixed Model (LMM) [17] under the following setting for the fixed effects,

$$\text{Response} = \text{Genotype} + \text{Sex} + \text{Genotype} \times \text{Sex} + \text{BodyWeight} \tag{1}$$

and Batch (defined as the date of measurement) in the random effect. For the cases where LMM fails, PhenStat proceeds to an alternative method called Reference Range Plus (RR+) [16]. The RR+ method relies on an initial setting of a quantile, default 95% in PhenStat, to form the initial classes (Low/Normal/High) that discretises a continuous response based on the control, wild type (WT), mice population. Mutants are then stratified into the (Low-Normal versus High) and similarly to (Low versus Normal-High) classes. One can present the RR + model as below,

$$\begin{cases} \text{Response} \times \text{Genotype}|\text{Distribution of control mice} & , \quad \text{Main effect} \\ \text{Response} \times \text{Genotype}|\text{Distribution of sex specific control mice} & , \quad \text{Sex specific effects} \end{cases} \tag{2}$$

where ('|') and ('×') represent the *conditional operation* and *interaction* respectively. Fisher's Exact Test is applied to each contingency table to test the hypothesis of the independence of rows (discretised response) and columns (genotype). The categorical data in the IMPC are analysed using Fisher's Exact test for combined sexes as well as the individual sexes under the model below,

$$\begin{cases} \text{Response} \times \text{Genotype} & , \quad \text{Main effect} \\ \text{Response} \times \text{Genotype}|\text{Sex}, & \quad \text{Sex specific effects} \end{cases}. \tag{3}$$

Besides the advantage of providing a reliable analysis pipeline by PhenStat, a number of limitations arise with the scalability of the input data flow and the diversity of scenarios that can be handled by the statistical software in high-throughput screening pipelines. For instance, the internal optimisation of PhenStat for continuous measurement relies on the repetitive use of Likelihood Ratio Tests (LRT) for model selection that requires a predefined threshold (default 0.05 in PhenStat). This leads to a reduction in transparency of the statistical analysis in addition to an increase in the computational complexity of the analysis for large scale screening projects such as IMPC. There are also many instances of the data in IMPC with repetition in the measurement values that lead to a misleading inference from the current

implementation of the RR+ in PhenStat. These issues, coupled with the ever-growing screens in the IMPC, especially ageing (e.g., longitudinal data), require scalable and more versatile statistical pipeline with user-defined models for different possible scenarios.

In this paper, we address the issues of scalability, extensibility, versatility, and efficiency in the current IMPC statistical pipeline implemented using the R package PhenStat by introducing a new package that we call ***OpenStats*** in the same development environment, R. The new software offers versatile modelling of high-throughput phenotypic data, such as modelling time dependency in data, for example, the longitudinal data in the IMPC ageing pipeline, with a focus on simplicity and efficiency. We assess the performance of OpenStats on the IMPC data including more than 2.5M datasets and analyses and compare the results with the current implementation of the IMPC stats pipeline. OpenStats is available from the Bioconductor repository (www.bioconductor.org/packages/OpenStats) that can be installed using the standard R workflow.

Methods

The IMPC data are collected from 13 research centres around the world and consists of 45M + data points (Data release 11, February 2020) from different measurements on the control and mutant mice. The Genotype effect is measured on a set of knockout mice [18], typically 14 (7 males and 7 females), and a large group of controls, normally several hundred mice, spread over a moderately long period of time from months to years. Fig 1 shows the increase in the number of phenotyped mice lines (left) as well as the data points (right) that are collected along with the IMPC data major releases starting from the first release in 2012 to the current release in 2020. This figure shows that on average the total number of the data points and the phenotyped lines between major data releases are increased by a factor of 21% and 23% respectively. Along with the scalability of the data, one challenge is the long term variability such as seasonal effects, changes in personnel and unknown time-dependent environmental factors in the IMPC data that is addressed by SoftWindowing method [19]. However, an accurate solution to the estimation of the long term variation in SoftWindowing requires a precisely formed *initial model* that is fitted to the data. This motivates a reimplementation of the current

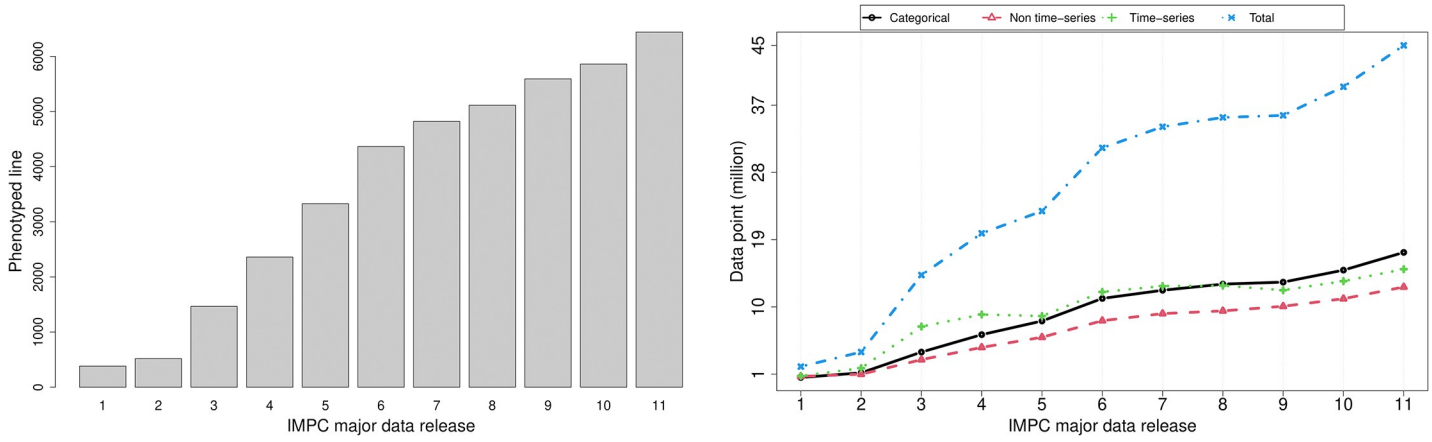

Fig 1. The increase in the total number of the IMPC mouse lines/data points along with the IMPC major data releases from the first release in 2012 to the current release in 2020. There is always a chance of more than one release or some minor releases per year. Here the y-axis shows the mouse lines/data points for the specific major release or the average from the minor releases. (Left) The total number of IMPC phenotyped lines corresponding to the IMPC data releases. (Right) The overall increasing trend in the data points divided by the type of the data, non-time series (red), the time series (green), categorical (black) and total (blue) corresponding to the IMPC data releases. These plots show that on average the total number of data points and phenotyped mouse lines increase by a factor of 20% between IMPC major data releases.

https://doi.org/10.1371/journal.pone.0242933.g001

methods with more versatility in the configuration and robustness in the implementation for the initial model.

Building block of the software

Fig 2 shows the four-layer structure of the software package namely, input data and model specification, data preparation, statistical analysis and reports.

The first layer, input data and model specification, data and the statistical model are mandatory, however, if the model is not specified, the default model is set to the same standard model [20] as PhenStat in Eqs (1), (2) and (3) corresponding to the type of input data.

The second layer, data preparation, the data and model terms are checked for missing values where OpenStats removes the variables with more than 50% missings; or alternatively allows basic substitution of the missings with the user specified values. Further checks are essential terms (such as genotype effect), redundancies (e.g. repeated variables in the model or variables that have the absolute 1 Pearson correlation), mismatching between model terms and the input data, and normalising the different terminologies in the data e.g. sex, Sex, gender to a unified semantic (e.g. "Sex"). Furthermore, it allows basic standard operations such as missing specification, visualisations and summary via standard plot and summary functions.

The third layer, statistical analysis, is managed by the OpenStatsAnalysis function. This works as a hub for different statistical methods that can be selected using the "method" argument. Regardless of the chosen method, the function checks for the concordance of the input model and the underlying data and whether the type of data fits the chosen statistical model.

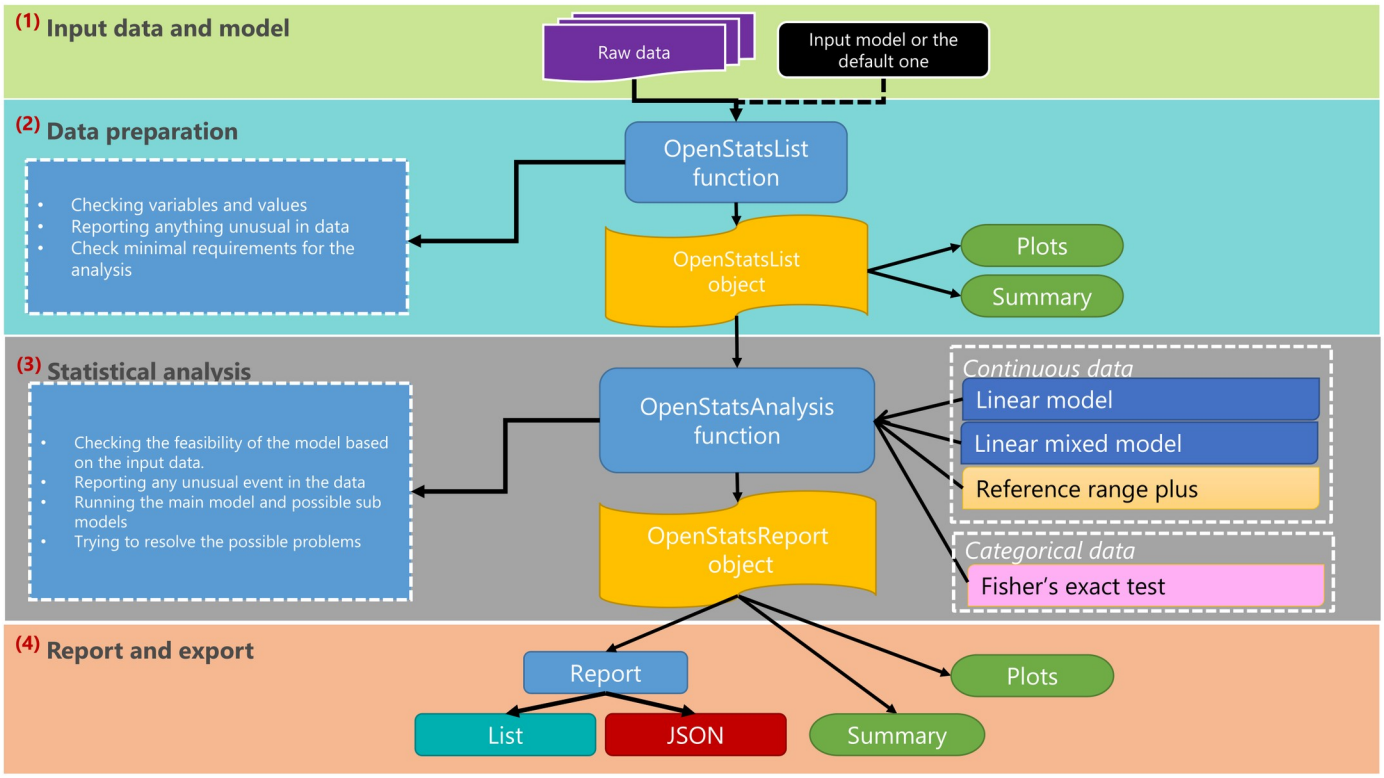

Fig 2. The schematic illustration of the OpenStats workflow. The OpenStats software is designed with a four-layer structure namely Input data and model specification, dataset processing and preparation, statistical analysis, and reporting/exporting the results.

https://doi.org/10.1371/journal.pone.0242933.g002

This further checks for other input arguments to the function and reports any potential errors. For the implemented statistical frameworks, the statistical significance is assessed and several measures such as standardised effect sizes, confidence intervals, sex-specific effects (provided sex is included in the input data), summary statistics of the input data and several other measures are reported. Moreover, the plot and summary are available for all methods.

The final layer in the workflow, report and export, is managed by the function OpenStatsReport and allows the key elements of the analysis to be extracted in the form of either *list* or *JavaScript Object Notation* (JSON www.json.org). The output of the OpenStatsReport function has a schema that makes it versatile to be populated into databases or used by other software in a pipeline.

Available statistical frameworks

Selection of the statistical method in the high-throughput screening pipelines that best fits the input data and the goal of the project is crucial otherwise can lead to misleading or weak evidence or increase the chance of producing Type I or Type II error in hypothesis testing [21]. However, efficiency and simplicity are essential when hundreds of thousands of analyses are performed in the high-throughput screening pipelines such as in the IMPC. Below we describe the three main analysis frameworks that are implemented in the OpenStats software package.

Nominal data. The majority of the nominal data in the IMPC measure the occurrence of a rare event such as *abnormal behaviour* or *absent/present of the tail* in the mice. To comply with the goal of the analysis, OpenStats applies Fisher's Exact Test [22–25] with p-values computed by Monte Carlo simulation for larger than *2 by 2* contingency tables. Depending on the specification of the model, sub-tables such as male, female, lifestage (defined as Early/Late adult mice), male/female × LifeStage interactions etc. are also formed and tested. The confidence interval for the odds ratios of 2×2 tables and effect sizes, defined as the maximum percentage change from the corresponding contingency table, are estimated for all tables.

Continuous data. *Linear mixed model*. The majority of the phenotypic data from the IMPC are continuous and analysed by performing the linear mixed model [17, 26] with Genotype, Sex and Bodyweight in the covariates and Batch (as the date of measurement) in the random effect [27, 28]. OpenStats allows an open structure for the covariates, random effect and the further structures on the within/between-group variation. This, in contrast to the PhenStat allows modelling of complex data in the IMPC such as repeated measures by including custom covariates/random effects. To cope with the low N, typically 4–7 mutant animals per sex in high-throughput pipelines, OpenStats applies an optional forward/backward/stepwise [29] optimisation to all model terms on the basis of comparing mutual AICc, a version of Akaike information criterion (AIC) that has a correction for small sample sizes [30]. Further to the initial model that is specified by the user, the sub-models are also fitted to the data for special purposes such as the detection of the sex specific (sexual dimorphism) effects. The confidence intervals, standard effect sizes [31] and standardised coefficients are also estimated for each possible sub-model. OpenStats further allows diagnosing the fitted model by providing visualisation tools and also reporting Shapiro-Wilk or Kolmogorov-Smirnov normality test statistics [32, 33] for assessing the normality of the model residuals.

Reference Range Plus method. The Reference Range Plus (RR+) method is an intuitive, simple and conservative method that is introduced in [16, 34, 35] and is based on the concept that a significant phenotype can be called when the majority of the mutant mice lie outside the natural variation seen in the controls/WT. More extensive implementation of the RR+ compared to PhenStat is performed in the OpenStats that includes an open structure to make the comparison amongst all specified covariates as well as sub-levels. To overcome the misleading

results from the data with repeated values, OpenStats reports empirical quantiles and adjust the threshold to the first distinct quantile.

Comparison between analysis from OpenStats vs PhenStat

The data analysis pipeline in the IMPC requires several steps that are shown briefly in Fig 3. This consists of importing data to the operational environment and QC'ing before applying the statistical methods. The working datasets are formed carefully by splitting data based on predefined metadata to ensure all relevant information are packaged into a single dataset prior to applying the statistical analysis. The analysis engine is controlled by the statistical analysis software namely PhenStat or OpenStats. Ultimately, the statistical results are quality controlled for failures, errors and the agreement with the other available/manual implementations and if approved, passed to the downstream processes.

The comparison between the statistical analyses from OpenStats versus PhenStat is shown in Table 1 where two instances of the IMPC statistical pipeline equipped with OpenStats and PhenStat simultaneously ran on a cluster computing machine. The results consist of processing 226 distinct procedures (S1 Table) including 41 IMPC specific procedures such as IMPC calorimetry (IMPC_CAL), clinical blood chemistry (IMPC_CBC), haematology (IMPC_HEM), acoustic startle, pre-pulse inhibition (IMPC_ACS), insulin blood level (IMPC_INS) and body composition (IMPC_DXA), and over 3.8K parameters (1K+ IMPC parameters)

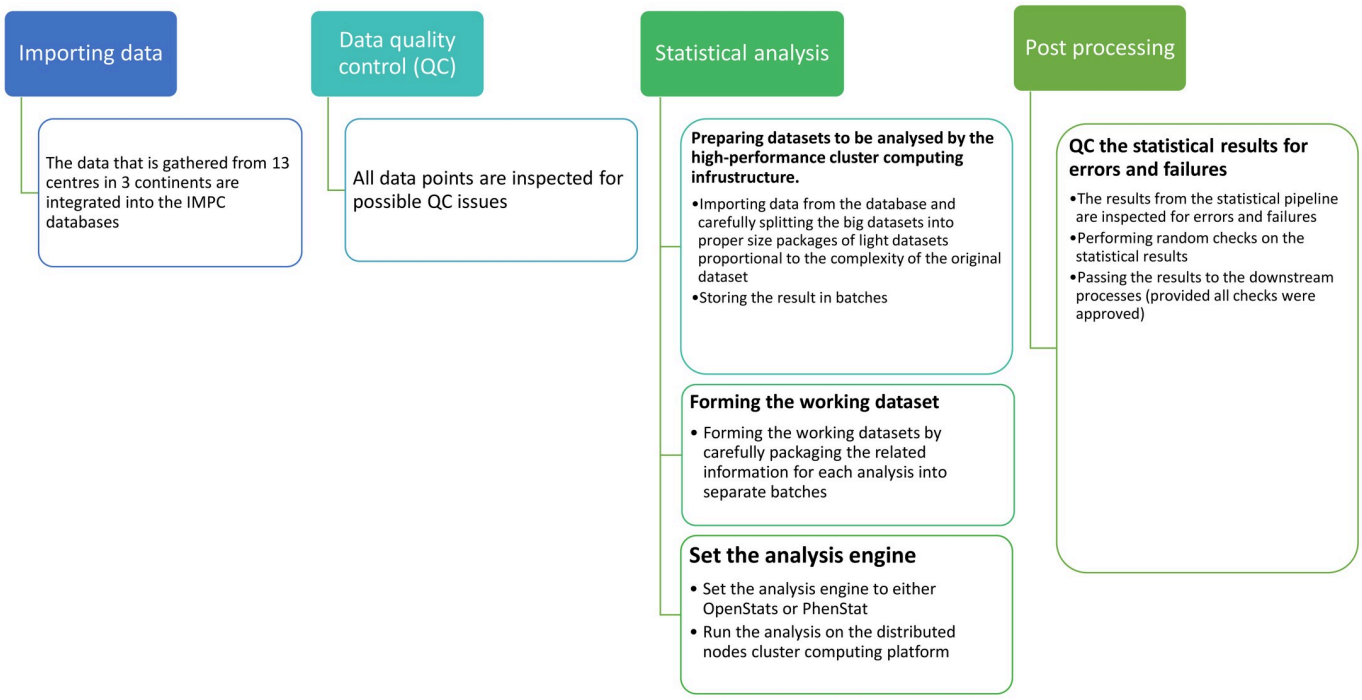

Fig 3. Schematic view of the IMPC statistical pipeline. The measurement of several parameters per specimen are collected from 13 centres all over the world, inspected for possible QC issues, carefully filtered to form individual working datasets, pre-optimised for being processed by the cluster computing platform and ultimately passed to the statistical analysis engine either PhenStat or OpenStats for the statistical analysis. The analysis engine is in charge of applying a proper statistical method to each working dataset and stores the analysis results in a format that enhances the downstream processes. All outputs from the statistical engine are inspected for the failures, errors and must pass a random QC check prior to being released to the downstream processes.

https://doi.org/10.1371/journal.pone.0242933.g003

Table 1. The comparison between OpenStats and PhenStat for analysing the IMPC continuous and categorical data.

	Total procedures	Total parameters	Total processed datasets	Total failures (individual analysis)	Total average time in hours (months)
OpenStats	226 (41)	5190 (1533)	2,551,026	135	1497 (2|1.5 IMPC)
PhenStat	226 (41)	5190 (1533)	2,550,141	1020	16168 (22|20 IMPC)

Each procedure contains several parameters that need to be analysed separately. Time elapse is estimated by adding up all spent times by carefully generated multi-processing jobs that equally fed into the OpenStats and PhenStat simultaneously. In all cases, the same procedure, parameter and data and analysis method are processed by OpenStats and PhenStat.

https://doi.org/10.1371/journal.pone.0242933.t001

(S2 Table). The complete list of IMPC procedures and parameters are available from www.mousephenotype.org/impress/pipelines. Because the analyses are performed on a farm of machines controlled by the cluster computing software, the direct measurement of time elapsed for each procedure/parameter is not straightforward. To alleviate this issue, the statistical pipeline reports the time spend per analysis/machine. Then, we report the average time for analysing the datasets in each procedure/parameter, which is normally the average of several hundred analyses. This ensures an unbiased estimation of the performance of the software.

The analysis of the entire IMPC data using OpenStats and PhenStat on a hypothetical single-core machine takes a total of 24 months (~21.5 months for the IMPC specific procedures), including the categorical and continuous data. From that 24 months, 22 (92%) months (20 months for IMPC procedures) are accounted for by PhenStat and 2 (8%) months (1.5 months IMPC procedures) by OpenStats. Fig 4 (top row) shows the distribution of average time saved (in minutes) by utilising OpenStats software across the specific IMPC procedures (top ten) (left) and parameters (top 30 to save space) (right) whereas the bottom plots show the cases where PhenStat performs more efficient in time than OpenStats. These plots show a significant reduction in computational time from the OpenStats versus PhenStat software for processing the entire IMPC data. The full table of comparisons over all procedures is available in supplementary materials S3 Table.

We further performed a confirmatory analysis to compare the agreement between the statistical results, Genotype effect p-value in particular, from PhenStat and OpenStats. Our results show an overall agreement of 99% between the two software, 99.9%, 99.9%, and 98.9% for Fisher's exact test, Reference Range plus, and Linear Mixed model frameworks respectively. The main cause of the disagreement between the two software is using Monte Carlo simulation-based method for the Fisher's exact test and Reference Range plus as well as new model selection strategy based on Akaike information criterion for small samples (AICc) for the internal optimisation of the software and the natural diversity in the data that sometimes violates the assumptions of the applied model for the Linear Mixed model frameworks. For instance, https://bit.ly/2WxI2Io is an example dataset from IMPC Acoustic Startle and Prepulse Inhibition (PPI) procedure. This data is right-skewed and should be considered carefully. In this example, the internal optimisation of PhenStat removes the effect of the bodyweight (at the level of 0.05) as a covariate in the linear mixed model under the settings in Eq. (1) whereas OpenStats preserves it in the model. The consequence is a 2-fold decrease in the magnitude of the genotype effect p-value, 0.49 in PhenStat and 0.18 in OpenStats. We should stress that there is no right answer in this example as there is a violation of the model assumptions/residuals. The next example in https://bit.ly/2YB4int represents a dataset from the IMPC Haematology procedure and depicts the inconsistency in the statistical results that is a product of the missing values in body weight (14% in total) and the outliers in response (5% based on Tukey's criteria with k = 3). In this case, PhenStat omits the bodyweight effect from the linear mixed model in Eq (1) (Genotype effect p-value 0.0004) whereas OpenStats keeps the body weight in

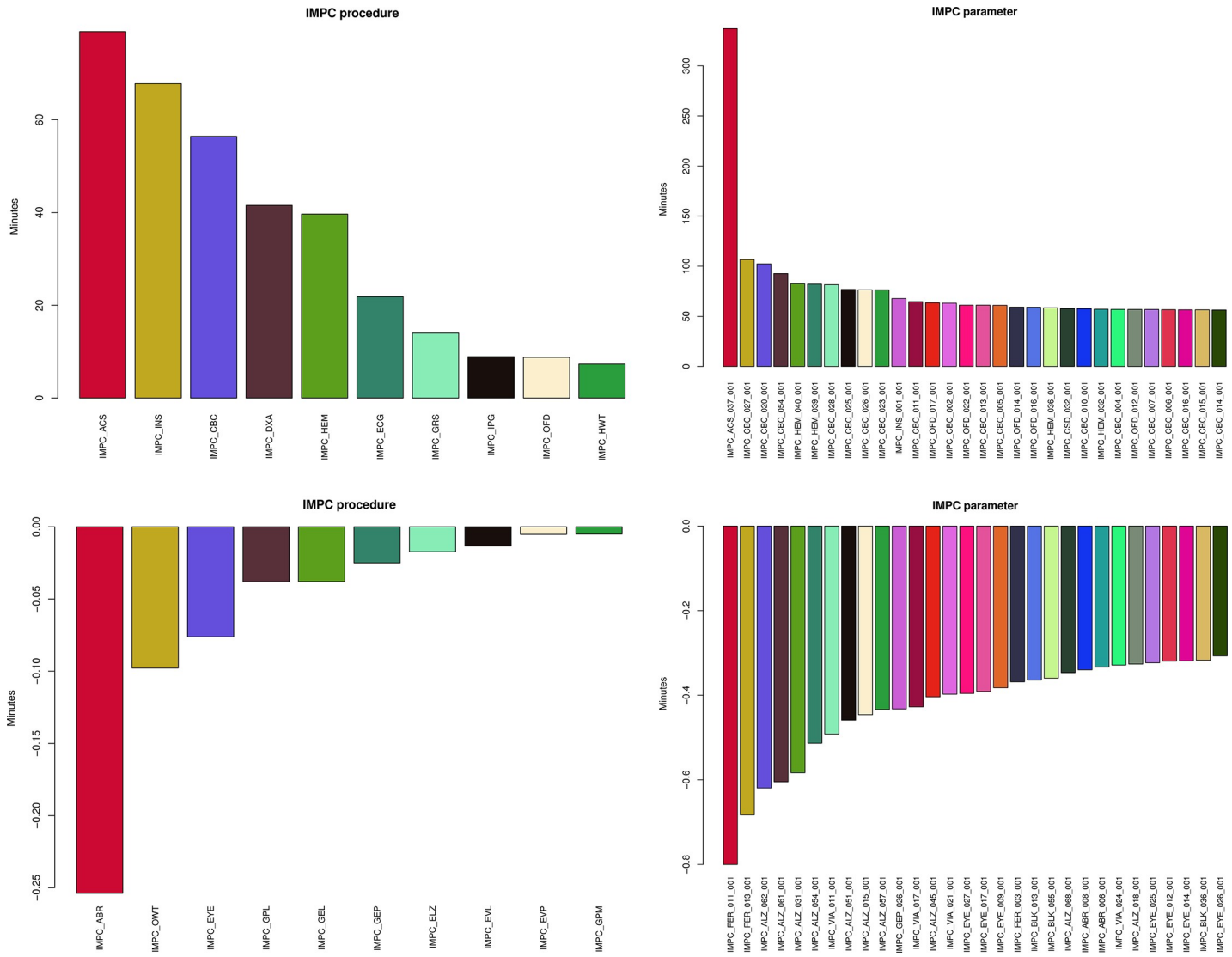

Fig 4. The comparison of the IMPC statistical pipeline analysed by OpenStats and PhenStat with respect to the time efficiency. (Top row) The left and right charts show the top (average) saving time in minutes by using OpenStats versus PhenStat over the IMPC procedures and parameters. The bottom row shows the top best (average) loses in minutes where PhenStat performs faster than OpenStats. These plots show that OpenStats improves the efficiency of the IMPC statistical pipeline.

https://doi.org/10.1371/journal.pone.0242933.g004

the model but excludes the missing values from the analysis (OpenStats Genotype effect p-value 0.97). We should stress that regardless of the statistical software, unusual cases should be interpreted carefully.

Discussion

Establishing precise, robust, reliable, and reproducible statistical pipelines for high-throughput phenotyping screens is challenging and is the subject of recent research topics [16, 19, 27, 28, 36–38]. With ever-growing phenotypic data, more consideration is required for scalability and versatility of the statistical pipeline to model different possible scenarios as efficiently as possible. One challenge the International Mouse Phenotyping Consortium (IMPC) faces is the diversity in data that demands more complex statistical methods with minimal latency. Here we introduced OpenStats, a freely available R package that allows systematically analysing different

statistical scenarios in the high-throughput phenotypic data. The R package allows a *fully customised* analysis plan for the implemented methods namely: linear mixed model, Fisher's exact test and Reference Range plus, as well as a comprehensive workflow with a focus on simplicity, efficiency, scalability and completeness that offers more than the raw statistical results and more than the counterparts in the literature. The performance of the new software compared to the current implementation of the statistical pipeline in IMPC is assessed on more than 45M data points and 4M+ analyses. Our comparisons show on average 90% reduction in time spent by adopting the new software while 99% of the results remain similar between OpenStats and the closest counterpart, PhenStat, for the IMPC data. The speed efficiency of OpenStats lies in the fact that the software utilises a "start-update" strategy instead of "start-terminate". That is, each model is formed by updating the previous one in contrast to fitting a brand-new model at each step. Besides the advantages of the software, there are a number of limitations to OpenStats including dependency to the statistical pipelines built on R and difficulties with exchanging the statistical methods with other statistical analysis software such as Python.

OpenStats addresses other challenges beyond the need to scale/versatile analysis to large datasets. One example is irreproducibility in results from animal experiments that is cited as a major contributor to explain many drugs failure in the development pipeline [39]. To address this concern, the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines have recently been updated to emphasis reporting on statistical methods including clearly stating the statistical method that is used and whether the data meet the assumptions of the statistical approach [40]. As part of its operation, OpenStats assesses whether data fits the requirements for a statistical test, such as not having enough data for performing interaction tests in the linear mixed model or assessing the normality of input data/model residuals. OpenStats also provides the visualisation tools required to diagnose the fitted model. The resulting analyses are clearly defined by which method was used, promoting reproducibility and repeatability of the results and the statistical models.

Increasingly researchers and stakeholders such as funders are demanding that biological research data and their analyses follow the FAIR principles—Findable, Accessible, Interoperable, and Reusable [41]. OpenStats contributes to FAIR data by assessing input data for completeness, redundancy, and any mismatch in variable format and/or labels. It also provides semantic normalisation such as sex, Sex and gender to a unified term "Sex" for common biological data variables in order to promote interoperability and reusability of data. Critically, OpenStats enables reusability of statistical methods by being freely available from the well-known BioConductor software project (www.bioconductor.org/packages/OpenStats) allowing any researcher to reproduce and reuse analyses from others' research while ensuring their own analysis is FAIR.

In summary, OpenStats promotes FAIR data and better reproducibility of biological research results while providing a means to scale to the larger and more complex datasets being generated by the research community.

Future study

Future work could be deriving a quality score that represents the quality of individual statistical analysis from the high-throughput genomic pipelines by testing the different aspects of the input data and the analysis results for the specified analysis framework.

Supporting information

S1 Table. The list of IMPC *procedures* that are involved in the IMPC statistical pipeline comparison between OpenStats and PhenStat.
(XLSX)

S2 Table. The list of IMPC *parameters* that are involved in the IMPC statistical pipeline comparison between OpenStats and PhenStat.
(XLSX)

S3 Table. The difference in the meantime spent by OpenStat and PhenStats for the IMPC procedures. The green cells show the improvement (in second) whereas the red cells show the cases where PhenStat outbeat OpenStats.
(XLSX)

Author Contributions

Conceptualization: Hamed Haselimashhadi.

Formal analysis: Hamed Haselimashhadi.

Funding acquisition: Helen Parkinson.

Methodology: Hamed Haselimashhadi, Jeremy C. Mason.

Writing – original draft: Hamed Haselimashhadi, Terrence F. Meehan.

Writing – review & editing: Jeremy C. Mason, Ann-Marie Mallon, Damian Smedley, Terrence F. Meehan, Helen Parkinson.

References

1. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011; 10: 712–712. https://doi.org/10.1038/nrd3439-c1 PMID: 21892149
2. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014; 505: 612–613. https://doi.org/10.1038/505612a PMID: 24482835
3. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The arrive guidelines for reporting animal research. Animals. 2013; 4: 35–44. https://doi.org/10.3390/ani4010035
4. Goktug AN, Ong SS, Chen T. GUItars: A GUI Tool for Analysis of High-Throughput RNA Interference Screening Data. PLoS One. 2012; 7. https://doi.org/10.1371/journal.pone.0049386 PMID: 23185323
5. Schulz JB, Cookson MR, Hausmann L. The impact of fraudulent and irreproducible data to the translational research crisis–solutions and implementation. J Neurochem. 2016; 139: 253–270. https://doi.org/10.1111/jnc.13844 PMID: 27797406
6. Holmes S. Statistical proof? The problem of irreproducibility. Bull Am Math Soc. 2018; 55: 31–55. https://doi.org/10.1090/bull/1597
7. Karp NA, Meehan TF, Morgan H, Mason JC, Blake A, Kurbatova N, et al. Applying the ARRIVE Guidelines to an In Vivo Database. 2015; 13: e1002151. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002151
8. Ozonoff DM, Grandjean P. What is useful research? The good, the bad, and the stable. Environ Heal A Glob Access Sci Source. 2020; 19. https://doi.org/10.1186/s12940-019-0556-5 PMID: 31910848
9. Hirsch C, Schildknecht S. In vitro research reproducibility: Keeping up high standards. Frontiers in Pharmacology. Frontiers Media S.A.; 2019. https://doi.org/10.3389/fphar.2019.01484 PMID: 31920667
10. Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res. 2014; 42: D802–D809. https://doi.org/10.1093/nar/gkt977 PMID: 24194600
11. Brown SDM, Moore MW. The International Mouse Phenotyping Consortium: Past and future perspectives on mouse phenotyping. Mamm Genome. 2012; 23: 632–640. https://doi.org/10.1007/s00335-012-9427-x PMID: 22940749
12. Bradley A, Anastassiadis K, Ayadi A, Battey JF, Bell C, Birling MC, et al. The mammalian gene function resource: The International Knockout Mouse Consortium. Mamm Genome. 2012; 23: 580–586. https://doi.org/10.1007/s00335-012-9422-2 PMID: 22968824

13. De Angelis MH, Nicholson G, Selloum M, White JK, Morgan H, Ramirez-Solis R, et al. Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. Nat Genet. 2015; 47: 969–978. https://doi.org/10.1038/ng.3360 PMID: 26214591

14. Kurbatova N, Karp N, Mason J, Haselimashhadi H. PhenStat: statistical analysis of phenotypic data. BiocIsmAcJp. 2016; 1–9. Available: http://bioc.ism.ac.jp/packages/devel/bioc/vignettes/PhenStat/inst/doc/PhenStatUsersGuide.pdf

15. R Team Core. R Foundation for Statistical Computing, Vienna, Austria. Vienna, Austria; 2019. p. 2019. Available: www.R-project.org/.

16. Kurbatova N, Mason JC, Morgan H, Meehan TF, Karp NA. PhenStat a tool kit for standardized analysis of high throughput phenotypic data. Dalby AR, editor. PLoS One. 2015; 10: e0131274. https://doi.org/10.1371/journal.pone.0131274 PMID: 26147094

17. Gilbert GE. Linear Mixed Models: A Practical Guide Using Statistical Software. J Am Stat Assoc. 2009; 103: 427–428. https://doi.org/10.1198/jasa.2008.s216

18. De Angelis MH, Nicholson G, Selloum M, White JK, Morgan H, Ramirez-Solis R, et al. Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. Nat Genet. 2015; 47: 969–978. https://doi.org/10.1038/ng.3360 PMID: 26214591

19. Haselimashhadi H, Mason JC, Munoz-Fuentes V, López-Gómez F, Babalola K, Acar EF, et al. Soft Windowing Application to Improve Analysis of High-throughput Phenotyping Data. Bioinformatics. 2019 [cited 2 Jul 2019]. https://doi.org/10.1093/bioinformatics/btz744 PMID: 31591642

20. Kurbatova N, Mason JC, Morgan H, Meehan TF, Karp NA. PhenStat: A Tool Kit for Standardized Analysis of High Throughput Phenotypic Data. Dalby AR, editor. PLoS One. 2015; 10: e0131274. https://doi.org/10.1371/journal.pone.0131274 PMID: 26147094

21. Dennis B, Ponciano JM, Taper ML, Lele SR. Errors in Statistical Inference Under Model Misspecification: Evidence, Hypothesis Testing, and AIC. Front Ecol Evol. 2019; 7: 372. https://doi.org/10.3389/fevo.2019.00372

22. Patefield WM. Algorithm AS 159: An Efficient Method of Generating Random R × C Tables with Given Row and Column Totals. Appl Stat. 2006; 30: 91. https://doi.org/10.2307/2346669

23. Fisher RA. The Logic of Inductive Inference. J R Stat Soc. 2006; 98: 39. https://doi.org/10.2307/2342435

24. Clarkson DB, Fan Y, Joe H. A remark on algorithm 643: FEXACT: an algorithm for performing Fisher's exact test in r x c contingency tables. ACM Trans Math Softw. 1993; 19: 484–488. https://doi.org/10.1145/168173.168412

25. Agresti A. Categorical data analysis. 2nd ed. Wiley; 2003.

26. Pinheiro JC, Bates DM. Mixed-effects models in S and S-PLUS. Springer; 2000. Available: https://books.google.co.uk/books?id=y54QDUTmvDcC&printsec=frontcover&dq=Pinheiro,+J.C.,+and+Bates,+D.M.+(2000)+%22Mixed-Effects+Models+in+S+and+S-PLUS%22,+Springer.&hl=en&sa=X&ved=0ahUKEwiNsMj9oKXjAhWVqHEKHaYMAhIQ6AEIKjAA#v=onepage&q&f=false

27. Karp NA, Speak AO, White JK, Adams DJ, De Angelis MH, Hérault Y, et al. Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. Gkoutos G V., editor. PLoS One. 2014; 9: e111239. https://doi.org/10.1371/journal.pone.0111239 PMID: 25343444

28. Karp NA, Melvin D, Mott RF. Robust and Sensitive Analysis of Mouse Knockout Phenotypes. Dalby AR, editor. PLoS One. 2012; 7: e52410. https://doi.org/10.1371/journal.pone.0052410 PMID: 23300663

29. Suárez E, Pérez CM, Rivera R, Martínez MN. Applications of Regression Models in Epidemiology. Applications of Regression Models in Epidemiology. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2017. https://doi.org/10.1002/9781119212515

30. Burnham KP, Anderson DR, Losos JB. Model selection and multimodel inference. A practical information-theoretical approach. Ecology Letters. Springer; 2002. https://doi.org/10.2307/3802723

31. Cohen J. Statistical Power Analysis for the Behavioral Sciences. In: Statistical Power Analysis for the Behavioral Sciences [Internet]. 2013 [cited 22 Jan 2020]. https://doi.org/10.4324/9780203771587 PMID: 23888591

32. Royston JP. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. Appl Stat. 2006; 31: 115. https://doi.org/10.2307/2347973

33. Conover WJ, Conover WJ. Practical Nonparametric Statistics (Wiley Series in Probability and Statistics). 3rd edition, editor. John Wiley & Sons; 1980. Available: http://www.amazon.com/Practical-Nonparametric-Statistics-Series-Probability/dp/0471160687

34. White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, et al. XGenome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell. 2013; 154: 452. https://doi.org/10.1016/j.cell.2013.06.022 PMID: 23870131

35. Cook MN, Dunning JP, Wiley RG, Chesler EJ, Johnson DK, Miller DR, et al. Neurobehavioral mutants identified in an ENU-mutagenesis project. Mamm Genome. 2007; 18: 559–572. https://doi.org/10.1007/s00335-007-9035-3 PMID: 17629744

36. Willis R. Must try harder. Community Care. 2006; 483: 32–33. https://doi.org/10.1038/483509a PMID: 22460859

37. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483: 531–533. https://doi.org/10.1038/483531a PMID: 22460880

38. Baker D, Lidster K, Sottomayor A, Amor S. Two Years Later: Journals Are Not Yet Enforcing the ARRIVE Guidelines on Reporting Standards for Pre-Clinical Animal Studies. Eisen JA, editor. PLoS Biol. 2014; 12: e1001756. https://doi.org/10.1371/journal.pbio.1001756 PMID: 24409096

39. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLOS Biol. 2015; 13: e1002165. https://doi.org/10.1371/journal.pbio.1002165 PMID: 26057340

40. Sert NP du, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2019: updated guidelines for reporting animal research. bioRxiv. 2019; 703181. https://doi.org/10.1101/703181

41. Wilkinson MD, Dumontier M, Sansone SA, Bonino da Silva Santos LO, Prieto M, Batista D, et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci data. 2019; 6: 174. https://doi.org/10.1038/s41597-019-0184-5 PMID: 31541130