countfitteR: efficient selection of count distributions to assess DNA damage

Jarosław Chilimoniuk[1,3]^, Alicja Gosiewska[2]^, Jadwiga Słowik[2]^, Romano Weiss[3]^, P. Markus Deckert[4]^, Stefan Rödiger[3,5]^, Michał Burdukiewicz[3,6]

[1]Department of Bioinformatics and Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland; [2]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland; [3]Faculty of Natural Sciences, Brandenburg University of Technology Cottbus-Senftenberg, Senftenberg, Germany; [4]Faculty of Medicine and Psychology, Brandenburg Medical School Theodor Fontane, and Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Brandenburg, Germany; [5]Faculty of Health Sciences Brandenburg, Brandenburg University of Technology Cottbus-Senftenberg, Senftenberg, Germany; [6]Medical University of Białystok, Białystok, Poland

Contributions: (I) Conception and design: PM Deckert, S Rödiger, M Burdukiewicz; (II) Administrative support: S Rödiger, M Burdukiewicz; (III) Provision of study materials or patients: J Chilimoniuk, J Słowik, R Weiss, M Burdukiewicz; (IV) Collection and assembly of data: J Chilimoniuk, S Rödiger, R Weiss, A Gosiewska; (V) Data analysis and interpretation: J Chilimoniuk, A Gosiewska, J Słowik, S Rödiger, M Burdukiewicz; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Stefan Rödiger. Faculty of Natural Sciences, Brandenburg University of Technology Cottbus-Senftenberg, 01968 Senftenberg, Germany; Faculty of Health Sciences Brandenburg, Brandenburg University of Technology Cottbus-Senftenberg, 01968 Senftenberg, Germany. Email: stefan.roediger@b-tu.de; Michał Burdukiewicz. Faculty of Natural Sciences, Brandenburg University of Technology Cottbus-Senftenberg, 01968 Senftenberg, Germany; Medical University of Białystok, 15-089 Białystok, Poland. Email: michalburdukiewicz@gmail.com.

Background: DNA double-strand breaks can be counted as discrete foci by imaging techniques. In personalized medicine and pharmacology, the analysis of counting data is relevant for numerous applications, e.g., for cancer and aging research and the evaluation of drug efficacy. By default, it is assumed to follow the Poisson distribution. This assumption, however, may lead to biased results and faulty conclusions in datasets with excess zero values (zero-inflation), a variance larger than the mean (overdispersion), or both. In such cases, the assumption of a Poisson distribution would skew the estimation of mean and variance, and other models like the negative binomial (NB), zero-inflated Poisson or zero-inflated NB distributions should be employed. The model chosen has an influence on the parameter estimation (mean value and confidence interval). Yet the choice of the suitable distribution model is not trivial.

Methods: To support, simplify and objectify this process, we have developed the countfitteR software as an R package. We used a Bayesian approach for distribution model selection and the shiny web application framework for interactive data analysis.

Results: We show the application of our software based on examples of DNA double-strand break count data from phenotypic imaging by multiplex fluorescence microscopy. In analyzing numerous datasets of molecular pharmacological markers (phosphorylated histone H2AX and p53 binding protein), countfitteR demonstrated an equal or superior statistical performance compared to the usually employed two-step procedure, with an overall power of up to 98%. In addition, it still gave information in cases with no result at all from the two-step procedure. In our data sample we found that the NB distribution was the most frequent, with the Poisson distribution taking second place.

Conclusions: countfitteR can perform an automated distribution model selection and thus support the data analysis and lead to objective statistically verifiable estimated values. Originally designed for the analysis of foci in biomedical image data, countfitteR can be used in a variety of areas where non-Poisson distributed

^ ORCID: Jarosław Chilimoniuk, 0000-0001-5467-018X; Alicja Gosiewska, 0000-0001-6563-5742; Jadwiga Słowik, 0000-0003-3466-8933; Romano Weiss, 0000-0001-9569-5607; P. Markus Deckert, 0000-0001-9569-5607; Stefan Rödiger, 0000-0002-1441-6512.

counting data is prevalent.

Keywords: DNA damage; count data; zero-inflation; overdispersion

Submitted Sep 16, 2020. Accepted for publication Jan 17, 2021.
doi: 10.21037/atm-20-6363
View this article at: http://dx.doi.org/10.21037/atm-20-6363

Introduction

DNA double-strand breaks (DSBs) are complementary breaks in the phosphodiester backbone of both strands of a DNA molecule, leading to its complete segregation at the break point. They are considered a common, but very severe form of DNA damage as they can promote genomic instability and increase the risk of cancer (1). This is reflected in elevated levels of DSBs in various types of cancer, and prognostic relevance of this finding has been shown, e.g., in breast cancer [for a review, see (2)].

Typical inducers of DSBs are ionizing radiation and certain cytotoxic agents such as etoposide, rapamycin, doxorubicin and others, all of which are used in cancer therapy (2). This opens a wide field for diagnostic and prognostic use of DSBs in precision medicine. For example, DSBs occurring due to pathological cellular processes have been described as prognostic markers in radiological and pharmacological cancer management (3,4).

Physiologically, once a DSB has occurred, a cell will either perform an immediate repair or undergo apoptosis (becoming a cancer cell is the pathological alternative). Achieving the first requires a complex repair process to compensate for the lack of a complementary DNA template. Thus, cells can react to DSBs with several pathways. Among the very first events is the accumulation of histone H2AX at the break site, presumably to stabilize the disrupted molecule. H2AX becomes phosphorylated at serine 139 either by ATM or other phosphatidylinositol 3-kinase-related kinase (PIKK) kinases (5), then called γH2AX, as a sensing mechanism (6,7). This event recruits further DNA repair proteins like p53 binding protein (53BP1), MDC1, BRCA1/2 and RAD51 (2,5,8). An alternative sensing mechanism is provided by the Ku70/80 pathway (9,10).

Similarly, different pathways are involved in the repair process itself, whose selection mainly depends on cell cycle state: While cells in late S and G2/M phase can perform homology-directed repair (HDR; activated through BRCA1/2 pathway) (9,11) due to the close vicinity of sister alleles and thus achieve a highly accurate repair, cells in G0/G1 and G2/M phase use non-homologous end joining (NHEJ) activated through the Ku70/80 pathway to maintain genomic integrity (11,12). As DNA damage promotes tumorigenesis, key players of the DNA damage repair system are often found mutated (12-14) neoplastic disease, leading to ineffective or absent damage response and thus, genomic instability without programmed cell death (11). In contrast, cancers with high resistance against ionizing radiation or other DNA-damaging agents have shown increased DNA damage response (DDR), thus avoiding potentially lethal damage (5,15). In some of these, susceptibility to DNA damaging agents could be restored by targeted inhibition of DRR proteins such as ATM or DNA-PKcs or via “baiting” of the DDR system with small interfering DNA (siDNA) (10-15).

Obviously, the critical process of DNA-damage recognition and repair interacts with numerous other processes, of which only few are briefly mentioned here. For a more thorough review, see (2,4,11).

Poly(ADP-ribose) polymerases (PARP) catalyze post-transcriptional modification of proteins and play a crucial role, among other processes, in cell death, DNA repair and DNA modification. Previously only described in the repair of single-strand DNA nicks, newer research connects them to DSBs and γH2AX formation, too. This is of clinical importance, as PARP inhibitors are a new class of drugs in cancer therapy (16-18). Challenges lie in understanding the polypharmacology of current PARP inhibitors. While PARPs can be qualitatively associated with various processes, their quantitative levels vary not only between patients but also between replicates from the same patient (19), hindering their employment for diagnostic purposes.

A positive association was shown between nuclear γH2AX levels and PD-L1 expression in squamous cell lung carcinoma (20). Whether this finding points to a clinically exploitable cellular pathomechanism, e.g., via the cGAS-STING or the PI3K-Akt pathways, or is just an epiphenomenon of tumour mutational burden (TMB) remains to be elucidated. TMB is being discussed as a

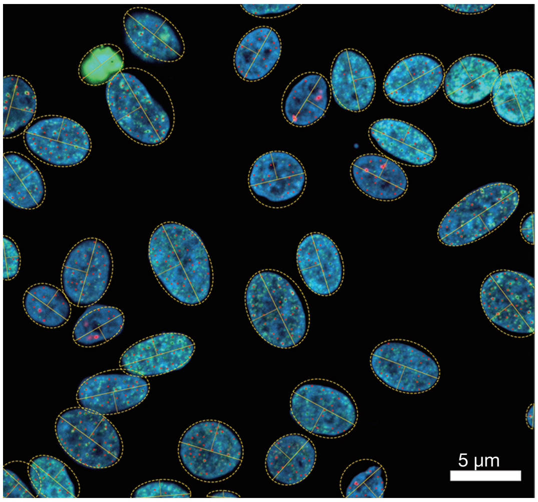


Figure 1 Foci detection of DNA double-strand breaks (DSBs) and cell nuclei by phenotypic imaging as shown in the software NucDetect. In a cascade of phosphorylation events, thousands of H2AX molecules may surround a single DSB. Depending on the damage induced, few to a hundred foci per cell can occur. Cell nuclei of the human HEp-2 line were labelled with the DNA-binding dye DAPI (blue) and DSBs were detected by an H2AX-specific monoclonal antibody with a FITC-conjugated secondary antibody (green). The number of foci (red dot) was determined via digital image processing. Example conditions: 24 h incubation with 5 µM Etoposide.

predictive marker for immunotherapy, as it correlates with the emergence of tumor-associated neoepitopes forming immunological targets (21). Measuring of DSBs from patient samples allows to quantify the extent of DNA damage, be it intended or adverse (22,23), it may directly predict treatment response as well as toxicity. Also, it may indirectly indicate DSB repair capacity as a proxy for drug resistance. This may be of particular interest in adjuvant therapy, where patients potentially cured by surgery receive additional therapy to reduce their risk of relapse (24-26). In observational studies, the predictive potential of DDR foci has been demonstrated by correlating foci data with clinical endpoints such as tumour control probability (27,28). As the association between γH2AX levels and PD-L1 expression suggests, it may also predict response to checkpoint-inhibitor immunotherapy.

Such information could help to personalize treatment and improve patient outcomes by adapting drug or radiation doses or choice of immunotherapy to cellular effects. As opposed to clinical effects observed weeks or even years later, DSBs could be measured between therapy sessions and thus help to individually adapt treatment in real-time. Similarly, drug development may be facilitated with regard to effectiveness as well as safety (29). To understand why this apparently attractive field has so far been cultivated only scarcely requires a closer look at its methodology.

The most widely used method for quantification of DSBs relies on intranuclear detection of fluorescence-labeled antibodies γH2AX by fluorescence microscopy (3,25,30). Visually, DSBs appear as intranuclear fluorescence foci varying in number and intensity. Manual counting of these foci in conventional microscopy is time-consuming and error-prone as it depends on the attention and experience of the examiner (31). Automated immunofluorescence microscopy has been developed to obtain more objective and reproducible results and to allow for high throughput in diagnostics (25). Phenotypic imaging is also an approach for testing pharmacodiagnostic and drug combination strategies and is used, among others, to predict drug resistance (32).

Image processing algorithms implemented in software like CellProfiler (33) or AutoFoci (34) allow both counting and sizing of the foci (*Figure 1*). In this method the algorithm calculating the result is of paramount importance. In a previous study we have examined image analysis software for counting focus data (35). Most of the programs used different algorithms and optimizations regarding the recognition of foci and cell nuclei. Consequently, the reported counts were different.

If DSBs are to be used as a key factor in clinical decision-making, a precise and unbiased method to estimate the mean foci count per cell is indispensable. A Poisson distribution is typically assumed to model count data. DNA DSBs are prerequisite for the formation of chromosome aberrations (36). For radiation-induced chromosome aberrations it was shown that the Poisson distribution is not optimal to model the data (37). Our studies of DSB data support the observation that Poisson, zero-inflated Poisson (ZIP), negative binomial (NB) and zero-inflated negative binomial (ZINB) distributions are appropriate to describe foci counts. Mathematically, assuming a Poisson distribution can be incorrect for two reasons: overdispersion and zero-inflation (31).

Overdispersion and zero-inflation

Count data represents the number of occurrences of a signal as non-negative integer values. Since this is one of the most common types of data, the modeling of counting data is of primary interest in many areas, including public health, medicine and epidemiology. It is generally assumed that such data can best be fitted to the Poisson distribution, but other distribution models have been described (38).

The Poisson distribution relies on the assumption that the mean (λ) and the variance are equal (equidispersion). If this is not true, the Poisson distribution cannot appropriately model the data. The case of the variance exceeding the mean is termed overdispersion, a variance smaller than the mean, underdispersion (39,40). Distribution models exist for both cases. Others have shown that deviation from the Poisson model may also depend on the scoring method (41). Overdispersion in Poisson distributions can be tested by a test proposed by Cameron and Trivedi (42). Within the biomedical scope of this article, we will focus on overdispersion as by far the most frequent case (43,44).

The NB distribution best describes data with a high variance due to numerous extremely high or low counts (45). As an example, this kind of overdispersion can arise from partial body exposure with radiation. Even though the NB distribution is used primarily for counting the number of failures before a predetermined number of successes occurs, it can be alternatively parameterized to describe count data with non-equal variance and mean. One of the critical properties of the NB distribution is that the maximum likelihood (ML) estimator of its mean ($\hat{\mu}$) (e.g., the mean number of occurrences) is equal to the arithmetic mean of counts in a data set. The additional variability provided by the parameter θ is known as size. When θ is approaching infinity, the NB distribution becomes a Poisson distribution.

Another cause of overdispersion is zero-inflation, i.e., an excessive number of zeros in a data set (46). To describe this phenomenon, we can use the ZIP and the ZINB distributions. The former depicts Poisson-distributed data with excessive zeros. The latter describes data where overdispersion arises from both increased variability of counts and zero-inflation. It is important to note that in the case of zero-inflated distributions, the mean number of counts λ is not equal to the average number of occurrences (μ). To describe their relationship, we need to introduce another parameter, r, which is equal to the fraction of counts faulty turned to zeros. Using the notation: $\lambda = \frac{\mu}{1-r}$. Henceforth, if we do not correctly identify zero-inflation, we underestimate the real number of occurrences. Further information on overdispersion can be found in the Supplement document in section 5 (available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf).

Usually, the NB distribution is parameterized using μ and θ, but to make comparison clearer, we use λ instead of μ. Poisson and NB distributions have the same expected value (table S1 available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf). In the case of ZIP and ZINB, the expected value is smaller than the real average number of foci per cell. Depending on the value of *r*, the variance of ZIP and ZINB may be smaller or bigger than the variance of the Poisson distribution. In the case of the NB distribution, the variance is always bigger than for the Poisson distribution, although the difference becomes negligible when the θ is much bigger than λ^2.

Selection of the most appropriate distribution

As the selection of the appropriate distribution is the key to successful modeling, the wrong choice leads to biased results of analysis (47,48). There are plenty of methods to test the equidispersion of data (49) or find out if the data does not contain exceeding amounts of zeros (50,51). However, even though these tests are statistically rigorous, they do not point to a specific distribution, but rather detect over- or underdispersion.

The next class of solutions are decision-making procedures designed to help in choosing the most appropriate count distribution (46,52). Here, the results of statistical tests are used to find out which distribution is underlying the data. Nevertheless, these procedures are often limited to a very specific set of distributions. Moreover, their power (here, the ability to select the most appropriate distribution) is reduced in the case of zero-inflated models (53).

Considering the situation described above, we implemented countfitteR as a framework for the selection of the underlying count distribution. Our software fits count data to four distributions that describe foci counts: Poisson, NB, ZIP and ZINB. The countfitteR framework selects the most appropriate model using the Bayesian information criterion (BIC). Additionally, countfitteR also estimates parameters of the distribution of choice and their confidence intervals. As our goal was to enable experimentalists to work with their own data, countfitteR is available not only as the R package but also on a web server.

We present the following article in accordance with the MDAR reporting checklist (available at http://dx.doi.

org/10.21037/atm-20-6363).

Methods

Model selection

In our parameterization, NB and ZINB are treated as a mixture of Poisson and Gamma (Γ) distributions (table S2, available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf). Moreover, Poisson, ZIP and NB distributions can be seen as special cases of ZINB distribution. From the modeling point of view, all of them belong to the family of General Linear Models (54). Poisson, ZIP and NB models are nested in ZINB, as ZINB contains all terms necessary to describe the other three distributions. Moreover, Poisson is nested in both ZINB and ZIP. However, ZIP is not a nested model to NB. Therefore, to compare fits with all four distributions, we must use a measure that is suitable for comparing fits with both nested and non-nested models.

Therefore, we have decided to use BIC, as this model selection criterion is appropriate for comparison of both nested and non-nested models (55). We decided to choose BIC instead of similar criteria for model complexity (e.g., Akaike's Information Criterion) based on the assumption that the distributions underlying the data we have examined here could be either Poisson, ZIP, NB, ZINB. In this case, when the sample size approaches infinity, the probability that BIC will select the correct distribution reaches certainty, which does not hold for AIC.

Thus, for samples with almost only zeros, BIC would choose the model with the smallest number of parameters, i.e., Poisson. It is the desired outcome for our framework, as we expect from the decision-support tool that it will be able to provide a conclusive answer about the underlying distribution.

The study was approved by the ethics committee of the Brandenburg University of Technology (BTU) Cottbus-Senftenberg (Ethikkommissionssatzung BTU, document number EK2018-3).

Statistical analysis

Statistical analysis to fit above mentioned models was performed by countfitteR, which uses R (36) packages pscl (37) and MASS (56). The data is presented as BIC values. The differences between models are interpreted according to the guidelines published elsewhere (55). The confidence level by default is set as 0.95. The assessment of the empirical power of the countfitteR framework was performed using the likelihood ratio test and the Vuong test (52).

Empirical power analysis

The empirical power of the countfitteR framework was compared with the two-step distribution selection (52). We have performed statistical simulations with a range of distributional assumptions based on the scheme published elsewhere (46). We have extended the proposed scheme by considering more values of parameters defining the distributions.

We have generated univariate data from Poisson (λ), ZIP (λ, r), NB (λ, θ), and ZINB (λ, θ, r) distributions. Each data set consists of 1,000 replications of samples with one of three possible sample sizes (n=50, 100, 200), and one of three possible means (λ =2, 5, 10). For zero-inflated distributions, the r ranges from 0.1 to 0.9 In the case of NB and ZINB distributions, similarly to Perumean-Chaney *et al.* (46), we have parameterized the dispersion parameter θ as $\frac{\lambda}{2}$. Additionally, we have considered θ equal to λ and 2λ.

Here, we have defined empirical power as the ability to select the correct distribution among Poisson, ZIP, NB, and ZINB. However, depending on the distribution, the two-step procedure does not work correctly and provides no answer (labeled as "uncomputable"). Although this behavior stems from the numerical assumptions of the two-step procedure, we have decided to highlight it in results to keep the comparison with the countfitteR framework fair.

The code necessary to reproduce the simulation study is available in the supplementary repository: https://github.com/BioGenies/countfitteR-simulations.

Acquisition of data for case study

The protocol for foci quantification was adopted from previous studies (30,57). Hereby, the AKLIDES® Nuk Human Lymphocyte Complete Combi (4268) Kit (MEDIPAN GmbH, Germany) was used for the immunofluorescence staining of DNA repair foci. In detail, 10^4 HEp-2 (ATCC® CCL-23™) cells were seeded in Dulbecco's modified Eagle's medium (DMEM; 10% fetal calf serum (FCS), 2 mM L-glutamine, 100 U/mL penicillin/streptomycin) on 10 well slides and incubated for 24 h (37 ℃, 5% CO_2). After incubation, the medium was discarded and the cells were incubated for 24 h with 5 µM etoposide in DMEM (10% FCS, 2 mM L-glutamine, 100 U/mL penicillin/streptomycin) to introduce DSBs. The medium was again discarded and the cells were fixated with 2% formaldehyde for 15 min at room temperature. The wells were rinsed three times with phosphate buffered saline (PBS; 140 mM NaCl, 2.7 mM KCl, 1 mM $Na_2HPO_4·2H_2O$, 2 mM KH_2PO_4, pH 7.4) and blocked/permeabilized with 50 µL/well blocking/permeabilization buffer (5% bovine serum albumin, 0.3% Triton X-100 in PBS) for 30 min at room temperature. After removing the blocking buffer, the primary antibody (anti-γH2AX/53BP1) was added (1:500 dilution in blocking/permeabilization buffer) and the slides were incubated for 60 min at room temperature. Upon rinsing the slides three times with PBS, 25 µL/well of secondary antibody solution [1:500 dilution of anti-mouse/rabbit antibody, 5 µg/mL, 4,6-diamidino-2-phenylindole (DAPI)] was added. The slides were incubated for 1 h in the dark at room temperature and then rinsed three times with PBS. A drop of mounting medium was added to each well and slides were sealed with a cover slip. Images were taken via immunofluorescence microscopy and analyzed via bioimage informatics (35). Foci numbers per cell were determined using the CellProfiler (33) (v. 3.1.9) with standard speckle counting pipeline and NucDetect (v. 0.11.15.dev2) (58) software.

Results

Modi operandi of countfitteR

countfitteR (v. 1.4) uses R as a basis because this statistical computing language is also used in clinical research (59). It can be used in different modes. These are the use of the R console for the programmatic use of specific functions and the use of the software as an interactive application with a graphical user interface. countfitteR can be installed in the R console using the command "install.packages('countfitteR')". Once installed successfully, the functions and sample data sets from the package are available with the command "library('countfitteR')" and can be used in dedicated R graphical user interfaces like RStudio or RKWard (60) (*Figure 2*). A detailed example is given in the Supplement document in section 3.

The countfitteR web server is an implementation of our framework as a graphical user interface running in the majority of modern browsers. The online version limits the user to datasets smaller than 5 MB. The local version can be used to analyze larger datasets.

The main functionalities of the web server are in two panels: Fitted models and Compare distributions. The first one presents results of the countfitteR framework: count data fitted to the distribution with the lower BIC. This panel contains the information, both in graphical and tabular format, about the fitted parameters as λ and their confidence intervals (*Figure 3*).

The second panel, compare distributions, allows in-depth exploration of the fitted models. The user can investigate all fits and their BIC values. countfitteR is a decision-support system, but the actual decision remains in the user's responsibility. This panel allows the researcher to determine whether the distribution suggested by countfitteR is plausible.

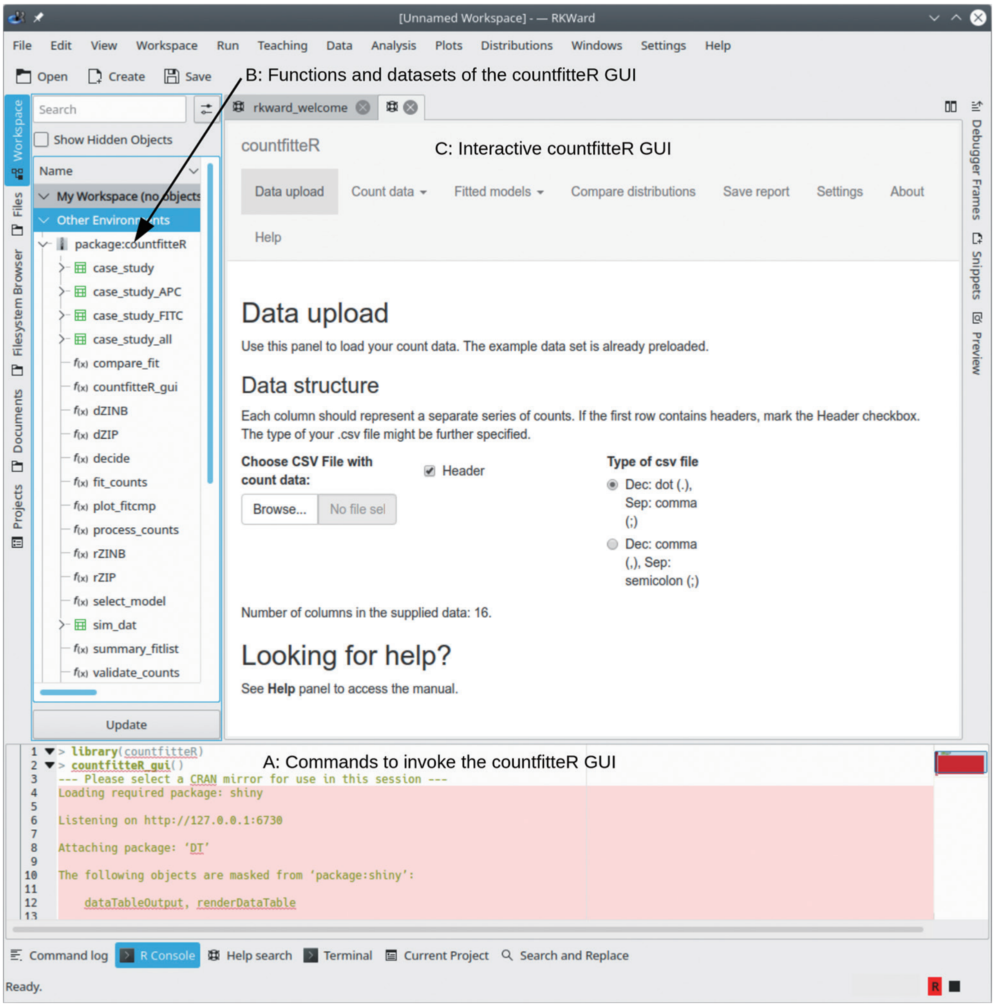
The important part of the web server is the report generation. All inputs changed by the user, such as confidence levels, are included in the report to ensure the reproducibility of the analysis. The report also includes the version of R and all R packages required to perform the analysis. This is necessary because the version of R and the package versions on which countfitteR depends may change due to updates. The report is also enriched with the md5 checksum of the input file and information about it to check if the file has been modified with the internal spreadsheet tool (*Figure 4*).

Empirical power summary

To validate the countfitteR framework, we have analyzed its empirical power with a simulation study and compared it with an existing two-step procedure (52). The empirical power is defined as the fraction of correctly identified distributions. The detailed description of simulations is available in the Methods section.

The countfitteR framework outperformed the two-step procedure for ZIP distribution [*Figure 5*; table S3 (available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf)] for all considered values of λ and n (number of counts in a sample). This is reflected by the overall empirical power of 0.94 for countfitteR, compared to 0.22 for the two-step procedure (*Table 1*). Moreover, countfitteR detects the ZIP distribution even for very high values of *r*, with the exception for the lowest considered λ=2. The two-step test performed adequately only for the highest value of λ and low-to-medium *r*. For other cases, the two-step test was wrongly pointing to NB distribution.

Results for the ZINB distribution were less conclusive


[Unnamed Workspace] - — RKWard
File Edit View Workspace Run Teaching Data Analysis Plots Distributions Windows Settings Help
Open Create Save
Workspace Files Filesystem Browser Documents Projects
Search
Show Hidden Objects
Name
My Workspace (no objects
Other Environments
package:countfitteR
case_study
case_study_APC
case_study_FITC
case_study_all
compare_fit
countfitteR_gui
dZINB
dZIP
decide
fit_counts
plot_fitcmp
process_counts
rZINB
rZIP
select_model
sim_dat
summary_fitlist
validate_counts
Update
rkward_welcome
Debugger Frames Snippets Preview
countfitteR
Data upload Count data Fitted models Compare distributions Save report Settings About
Help
Browse... No file sel
Command log R Console Help search Terminal Current Project Search and Replace
Ready.

B: Functions and datasets of the countfitteR GUI

C: Interactive countfitteR GUI

A: Commands to invoke the countfitteR GUI

Data upload

Use this panel to load your count data. The example data set is already preloaded.

Data structure

Each column should represent a separate series of counts. If the first row contains headers, mark the Header checkbox. The type of your .csv file might be further specified.

Choose CSV File with count data:

☑ Header

Type of csv file

● Dec: dot (.), Sep: comma (;)

○ Dec: comma (,), Sep: semicolon (;)

Number of columns in the supplied data: 16.

Looking for help?

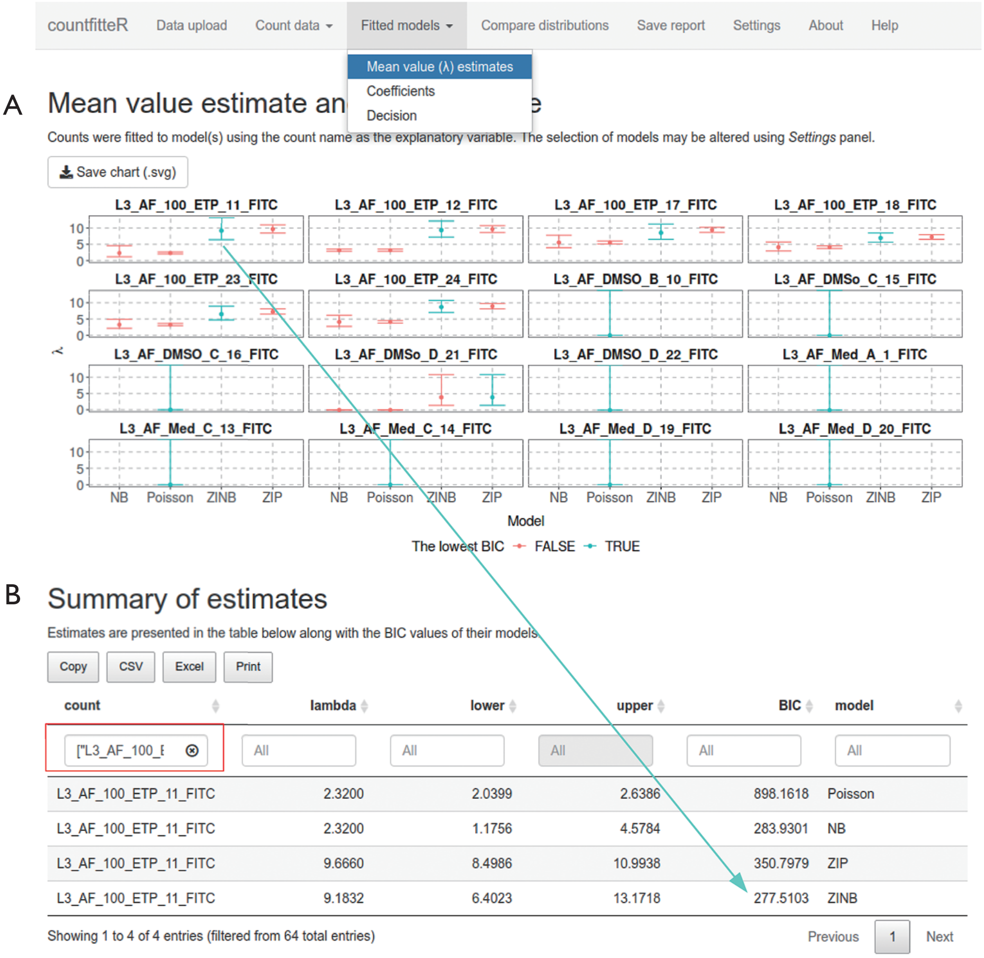
See **Help** panel to access the manual.

```
> library(countfitteR)
> countfitteR_gui()
--- Please select a CRAN mirror for use in this session ---
Loading required package: shiny

Listening on http://127.0.0.1:6730

Attaching package: ‘DT’

The following objects are masked from ‘package:shiny’:

    dataTableOutput, renderDataTable

```

Figure 2 Modi operandi of the countfitteR software. The countfitteR software was started in an R environment. Within this example RKWard (60) (0.7.1z+0.7.2+devel4) was used. A: On all operating systems the commands "library('countfitteR')", to load the countfitteR package, and "countfitter_gui()", to start the interactive graphical user interface (GUI), are identical. B: A list of functions and data sets of the countfitteR software is shown as an example. Details to the functions can be taken from the documentation. C: By executing the command "countfitter_gui()" the interactive GUI of the countfitteR software will be started automatically as soon as an environment which supports ECMAScript and HTML5 is detected [e.g., modern browser, RStudio (https://rstudio.com/), RKWard]. In this interface, data can be uploaded ("Data upload") and the steps for analysis and report generation can be defined.

as for some combinations of λ and θ. Both algorithms performed equally poorly [*Figure 6*; table S4 (available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf)]. For low lambda and high variance (θ= λ×0.5 and λ=2), the two-step procedure performed slightly better than the countfitteR framework, while countfitteR outperformed the two-step procedure for larger values of lambda, regardless of the θ. Due to this, the overall performance of countfitteR for the ZINB distribution was 0.43 compared to 0.25 for the two-step procedure. In the case of wrong decisions, the two-step procedure pointed towards NB and was unable to recognize even very high values of zero-inflation (r=0.5 and higher). The countfitteR software did not show such preference and reported a mixture of NB and ZIP distributions.

By intention, one of the most critical aspects of our tool is its practical applicability for experimental researchers. The two-step procedure, although statistically sound, often does not yield any results due to numerical limitations. In the case of only (or almost only) zeros in a small sample, it



Counts were fitted to model(s) using the count name as the explanatory variable. The selection of models may be altered using *Settings* panel.

Estimates are presented in the table below along with the BIC values of their models.

count	lambda	lower	upper	BIC	model
["L3_AF_100_E	All	All	All	All	All
L3_AF_100_ETP_11_FITC	2.3200	2.0399	2.6386	898.1618	Poisson
L3_AF_100_ETP_11_FITC	2.3200	1.1756	4.5784	283.9301	NB
L3_AF_100_ETP_11_FITC	9.6660	8.4986	10.9938	350.7979	ZIP
L3_AF_100_ETP_11_FITC	9.1832	6.4023	13.1718	277.5103	ZINB

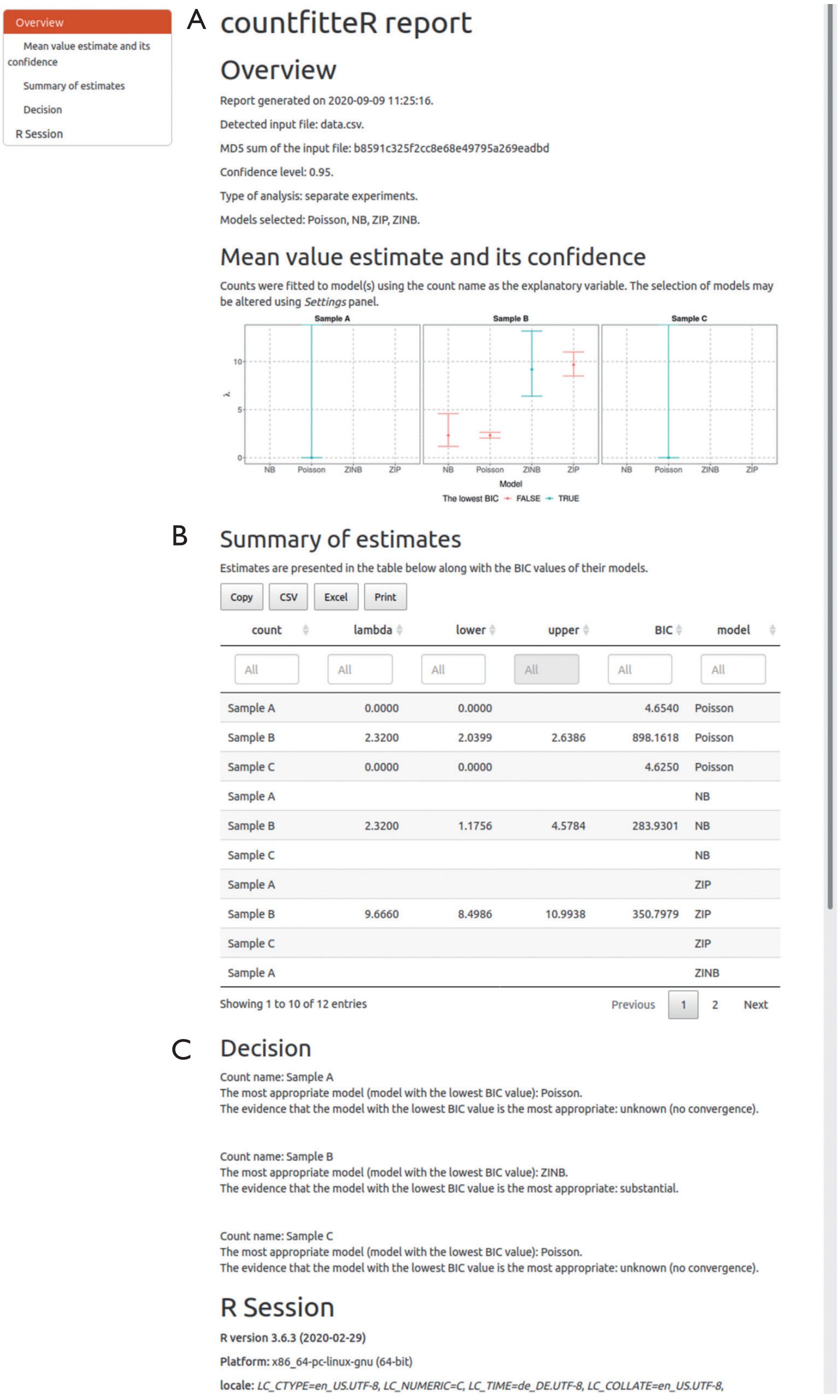
Showing 1 to 4 of 4 entries (filtered from 64 total entries)

Figure 3 The countfitteR graphical user interface (GUI). The interface shows the top bar with drop-down selection menus. (A) This allows the user to create plots, tables and statistical analyses. In the example, counting data from the "case_study_FITC" data set were used [Supplement section 4: available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-Supplement.pdf]. Under "Fitted models" → "Man value (λ) estimate" plots and the estimated parameter values can be displayed. (B) Using the interactive table, the data set "L3_AF_100_ETP_11_FITC" was filtered out (red box). Consequently, the measured values of the other samples are not displayed. It can be seen that the λs (mean value estimate), with values from 2.3 to 9.66 foci per cell, show marked differences between the distributions. The ZINB distribution has the lowest BIC (plot with turquoise arrow) and is therefore the most likely distribution model for the data of the measurement "L3_AF_100_ETP_11_FITC". Its estimated values from λ and the confidence intervals ("lower" & "upper") should be used for further analysis. ZINB, zero-inflated negative binomial; BIC, Bayesian information criterion.

is statistically impossible to distinguish between the four distribution models. In such cases, the two-step procedure returned no information to the user, while countfitteR still worked according to the implemented methodology.

This difference is clearly visible in the case of a Poisson distribution [*Table 1*; table S3 and figure S1 (available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf)]. Overall statistical performance of countfitteR (0.98) is higher than the overall performance of two-step procedure (0.89). The two-step procedure worked properly only for the lowest considered λ=2. For higher values of λ, the two-step procedure was unable to return any answer most of the time. This problem seemed to be less prevalent for larger samples sizes (n=100 and higher). By contrast, countfitteR was able to assess these cases and therefore seems to offer a broader practical use.

Similarly, in the case of an NB distribution, countfitteR



countfitteR report

Overview

Report generated on 2020-09-09 11:25:16.

Detected input file: data.csv.

MD5 sum of the input file: b8591c325f2cc8e68e49795a269eadbd

Confidence level: 0.95.

Type of analysis: separate experiments.

Models selected: Poisson, NB, ZIP, ZINB.

Mean value estimate and its confidence

Counts were fitted to model(s) using the count name as the explanatory variable. The selection of models may be altered using *Settings* panel.

Summary of estimates

Estimates are presented in the table below along with the BIC values of their models.

count	lambda	lower	upper	BIC	model
All	All	All	All	All	All
Sample A	0.0000	0.0000		4.6540	Poisson
Sample B	2.3200	2.0399	2.6386	898.1618	Poisson
Sample C	0.0000	0.0000		4.6250	Poisson
Sample A					NB
Sample B	2.3200	1.1756	4.5784	283.9301	NB
Sample C					NB
Sample A					ZIP
Sample B	9.6660	8.4986	10.9938	350.7979	ZIP
Sample C					ZIP
Sample A					ZINB

Showing 1 to 10 of 12 entries

Decision

Count name: Sample A
The most appropriate model (model with the lowest BIC value): Poisson.
The evidence that the model with the lowest BIC value is the most appropriate: unknown (no convergence).

Count name: Sample B
The most appropriate model (model with the lowest BIC value): ZINB.
The evidence that the model with the lowest BIC value is the most appropriate: substantial.

Count name: Sample C
The most appropriate model (model with the lowest BIC value): Poisson.
The evidence that the model with the lowest BIC value is the most appropriate: unknown (no convergence).

R Session

R version 3.6.3 (2020-02-29)

Platform: x86_64-pc-linux-gnu (64-bit)

locale: *LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=de_DE.UTF-8, LC_COLLATE=en_US.UTF-8,*

Figure 4 The countfitteR report. All information that led to the result is contained in the countfitteR software report. (A) This includes information about the data used, the md5-sum of the data (unique assignment) and the parameters used for statistical analysis (upper area). (B) The report is saved as interactive HTML (e.g., middle section: tables with sorting and filter functions) and can be read or edited independent of platform and device. (C) In the decisions section, the strength of trust is displayed using Bayesian criteria. The report also contains information about the R packages and software environment used in the analysis (end of the report).

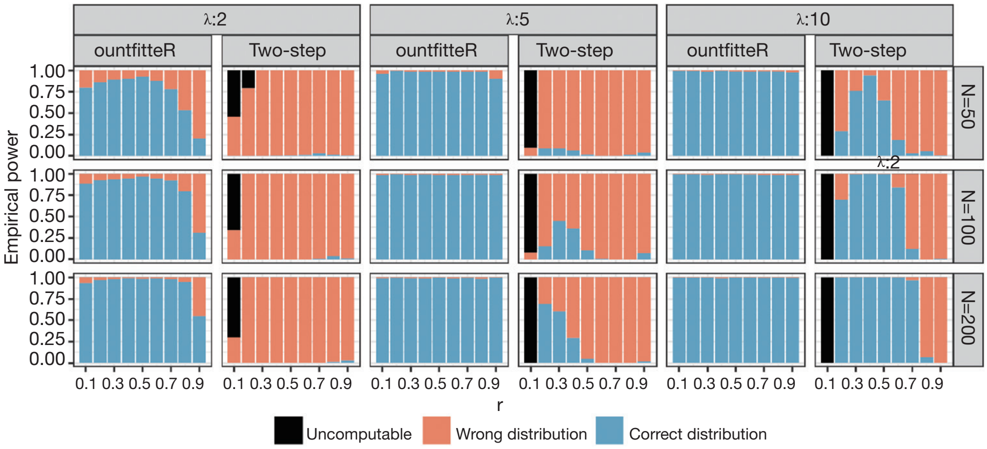


Figure 5 Empirical power of countfitteR and two-step test for ZIP distribution. n: number of counts in the sample. λ: Poisson parameter (number of occurrences, e.g., average number of foci per cell). r: zero inflation (fraction of occurrences treated as number of counts in the sample. λ: Poisson parameter (number of occurrences, e.g., average number of foci per cell). r: zero inflation (fraction of occurrences treated as zeros, e.g., fraction of cells treated by system as having no foci regardless of their real state). ZINB, zero-inflated negative binomial.

Table 1 Mean empirical power of countfitteR and two-step test for Poisson, ZIP, NB and ZINB distributions supplemented with percentages of solved cases

Distribution	Method	countfitteR	Two-step
Poisson	Mean	0.98	0.89
	Solved	100%	50.11%
ZIP	Mean	0.94	0.22
	Solved	100%	90.19%
NB	Mean	0.86	0.92
	Solved	100%	69.03%
ZINB	Mean	0.43	0.25
	Solved	100%	99.95%

ZIP, zero-inflated Poisson; NB, negative binomial; ZINB, zero-inflated negative binomial.

was able to make a decision in all simulations [*Table 1*; table S5 and figure S2 (available online: https://cdn.amegroups.cn/static/public/ATM-20-6363-1.pdf)]. The two-step procedure yielded a result in only 69.03% of cases, mostly for the lowest values of λ and for small sample sizes. Although the overall empirical power for the two-step procedure was higher (0.92), countfitteR still had a reasonable performance (0.86) while being able to analyze all cases. It is important to point out that this was irrelevant in both types of zero-inflated distribution (ZIP and ZINB). Here, the two-step procedure returned a decision for almost all analyzed samples.

Case study

We analyzed 2,253 images (as described in the Methods section) and used the countfitteR framework to find out which distribution is best describing the foci counts (*Figure 7*). To highlight the impact of the foci-counting software, we have compared counts obtained from CellProfiler and NucDetect.

The counts produced by CellProfiler mostly followed the NB distribution (63.2% for γH2AX and 59.4% for 53BP1). Still, the Poisson distribution was the second most common distribution (31.1% for γH2AX and 37.9% for 53BP1).

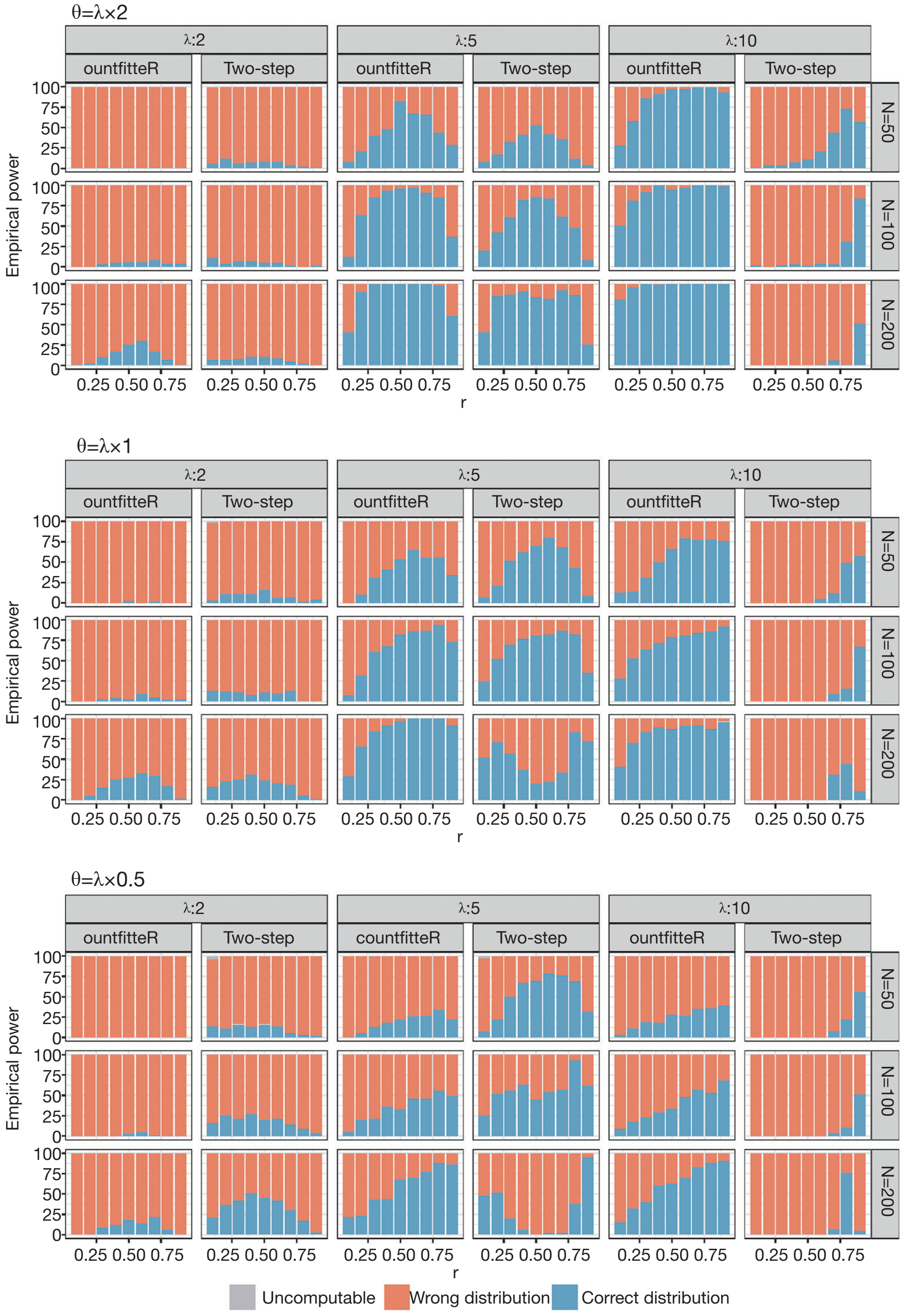


Figure 6 Empirical power of countfitteR and two-step test for ZINB distribution. λ: Poisson parameter (number of occurrences, e.g., average number of foci per cell). r: zero inflation (fraction of occurrences treated as zeros, e.g., fraction of cells treated by system as having no foci regardless of their real state). θ: dispersion parameter. ZINB, zero-inflated negative binomial.

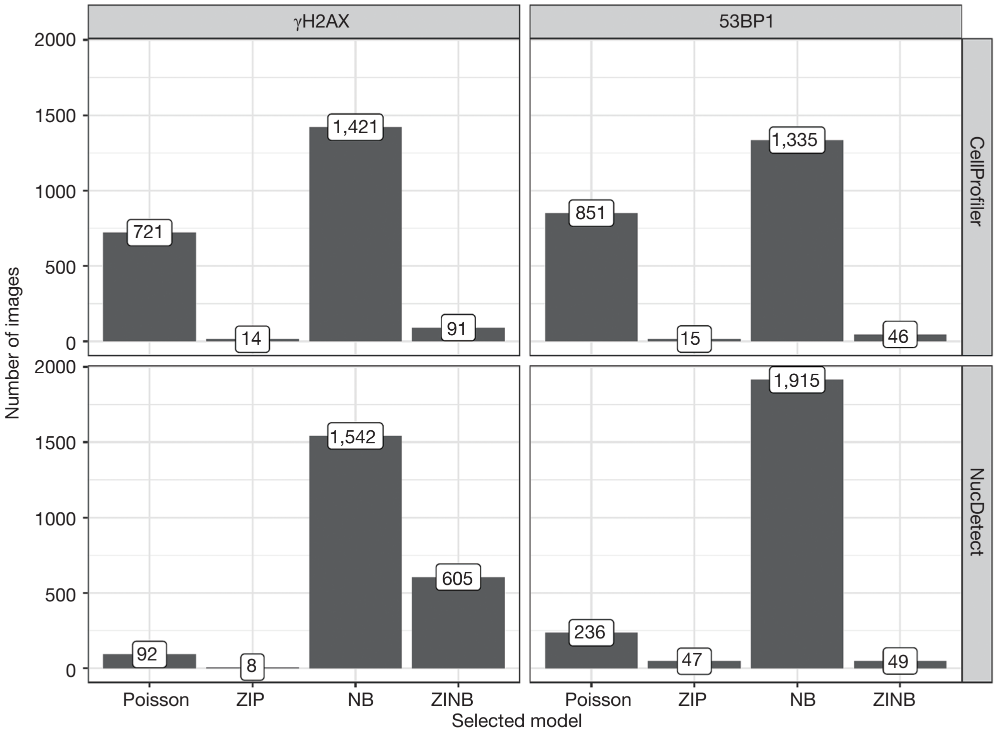


Figure 7 Distributions selected by countfitteR underlying data from the case study for γH2AX and 53BP1 biomarkers. The counts of 2,253 images were analyzed by CellProfiler (top row) and NucDetect (bottom row) software. ZIP, zero-inflated Poisson; NB, negative binomial; ZINB, zero-inflated negative binomial.

Both zero-inflated distributions were marginally present for both markers.

For NucDetect, the NB distribution dominated in both channels (70.3% for γH2AX and 87.2% for 53BP1) and was much more prevalent than in the case of foci counts produced by CellProfiler. However, the second most common distribution for the relevant γH2AX channel was ZINB (26.7%), and for 53BP1 it was Poisson (8.4%).

Nevertheless, as indicated elsewhere (61), the Poisson distribution was never the most frequent distribution. The equidispersion for γH2AX was observed in 4.1% of cases for NucDetect counts and 32.1% of cases for CellProfiler. Thus, the majority of the data expressed some kind of overdispersion, mostly caused by the presence of extremely high counts. The counts for the co-localized biomarker 53BP1 behaved similarly (10.5% of cases for NucDetect counts and 37.9% of cases for CellProfiler).

Discussion

Although both overdispersion and zero-inflation are problematic during the analysis of count data, it is very critical not to confuse these two phenomena. The increase in variance can result from different factors, such as the patients' resistance to DSB-inducing effects, and zero-inflation is usually linked to the measuring devices and foci counting algorithms. Both effects are unwelcome, and it is essential to be able to distinguish between, e.g., ZIP and NB. The two-step procedure seems to mix those two distributions easily. An important feature of countfitteR is the exact distinction between overdispersion due to increased variability and excessive zeros.

With our approach, we provide evidence that relying on the basic assumption of a Poisson distribution is not tenable. The consequences reach much deeper than to data post-processing alone, as the Poisson assumption could also affect foci counting software. For example, AutoFoci compares raw data to Poisson distribution. This gives the user the option to compare the theoretical and empirical distribution for this particular model. However, our data show that more models are needed for consideration. *Figure 3A* shows that the location (λ) and dispersion (95%

confidence interval) may vary considerably between models. We interpret this to mean that giving these values based on the evidence for a distribution (the lowest BIC among four likely models) is good practice.

There is a plethora of software for count data analysis and very specific count distributions. Software for the (semi)automatic analysis of distributions [e.g., SPC (BPI Consulting, LLC, USA), Pelican (Vose software, Belgium), Weibull++ (reliasoft.com)] is closed source and tied to commercial platforms like EXCEL or specific operating systems. Most importantly they lack specific models for the analysis of DNA damage, as they were designed to be used for more general datasets. Other software like fitdistrplus (62) or fitter (63) are usually confined to command-line interfaces and lack objective and reproducible model selection with a report generation.

The distribution of digital data is important for diagnostic decisions and the generation of novel, data-driven hypotheses (64). We also see the countfitteR software as having biomedical significance for future diagnostic applications, which includes the recording and distribution of digital data. Since the software is based on R, it could be connected to existing technologies, such as the rEHR package (65), for working Electronic Health Record data and other medical data (66,67).

Our tool breaks ranks in both aspects, as countfitteR is designed to be as user-friendly and accessible to experimentalists as possible. We specifically aimed for the web server to make our tools available for users non-proficient in R. At this point it should be emphasized again that the countfitteR package contains the functions, exemplary data sets and the source code for calculation. Therefore, statistical bioinformaticians can develop pipelines for highly automated analyses.

As a downside, graphical user interfaces may be associated with a lower reproducibility than programmatic command lines, as the latter can easily be exactly recorded and repeated. Reproducibility means that a detailed description of the research workflow allows others to precisely replicate published results (68). Therefore, a leading design principle of the countfitteR web server was to enhance the reproducibility by providing advanced reporting functions. Every analysis comes with information about the version of used R packages and the md5 control sum of the input data.

countfitteR is a powerful and easily accessible tool for the selection of a proper count distribution among the four distribution models: Poisson, ZIP, NB and ZINB. Of course, its performance is limited to its area of competence defined by these models. In cases where countfitteR does not offer a plausible result, an even rarer distribution pattern has to be considered and to be tested for by other methods. For example, the zero-truncated Poisson distribution (value zero cannot occur) looks like a possible model for foci count data. However, in none of our experiments we ever had datasets where none zero values occurred. Therefore, we did not include it. In our experience and results by others regarding radiation-induced chromosome aberration (61), however, foci data appear to follow one of the four distributions implemented in the countfitteR software. The rationale for the selection of a specific model is documented for each analysis and thus supports reproducible research. We limited the software to Poisson, ZIP, NB and ZINB distribution, because there was not enough evidence in the literature for further models. However, since countfitteR is a cross-platform open-source software it can be extended by other models. The presented countfitteR software (v. 1.4) has a modular structure that contains 13 functions [as assessed by “lsf.str(‘package:countfitteR’)” in R] for statistical analysis and the graphical user interface. Since these functions are open source, the GitHub hosted software can be extended and forked. In the software we have implemented four distribution models (Poisson, ZIP, NB, ZINB distribution) based on the literature. If a new suitable model is described in the literature, it can be implemented in the software as a new function. The Bayesian selection algorithm adapts itself automatically. At this point we would like to mention that this can already be tested from the command line.

Conclusions

Other authors, and we expect that the tight integration of phenotypic imaging methods and automated data analysis will make a valuable scientific contribution. Our software can be integrated into your own bioinformatic analysis pipelines or users can use the software in a graphical user interface to get an objective assessment of the distribution. We recommend that the analysis of count data could be performed as follows and presented in studies:

(I) Documentation of data provenance [see (69) or further information];

(II) Data screening to identify possible errors in values and coding [see (70) for further information);

(III) Automated selection of the distribution model;

(IV) Determination of the location and dispersion parameters, at a defined confidence level, for the corresponding optimal distribution model;

(V) Preparation of a report containing information about the tested distribution models, the BIC for the optimal distribution model, the analyzed counts (location, dispersion and confidence level) along with the R/countfitteR version used.

The value of this approach was demonstrated by providing a tool for the automated statistical analysis of pharmacological responses to DNA damage. We tested our method on a specific pool of cells. This is important because we suspect that in other laboratories the distribution models may occur with a different frequency. Biological and technical factors are probably the primary reasons for this. In a review (35), we have listed other types of cancer cells that could be used for pharmacological studies and personalized medicine. These have the potential to advance personalized medicine and the development of novel therapeutic agents that include pharmacological compositions or polypharmacology in the disease context (71).

Acknowledgments

Funding: This work was supported by the "Gesundheitscampus Brandenburg - Konsequenzen der altersassoziierten Zell - und Organfunktionen" (GeCa: H228-05/002/008, 2017-2019) initiative of the Brandenburgian Ministry of Science, Research and Culture (MWFK) and the Deutscher Akademischer Austauschdienst/German Academic Exchange Service (DAAD) for scholarship (57507870 to JC).

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-6363

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-20-6363

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-6363). SR reports grants from Gesundheitscampus Brandenburg - Konsequenzen der altersassoziierten Zell - und Organfunktionen, grants from Initiative of the Brandenburgian Ministry of Science, Research and Culture (MWFK) during the conduct of the study; JC reports scholarship from Deutscher Akademischer Austauschdienst/German Academic Exchange Service (DAAD) during the conduct of part of the study. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Ethics committee of the Brandenburg University of Technology (BTU) Cottbus-Senftenberg (Ethikkommissionssatzung BTU, document number EK2018-3).

References

1. Van Der Schans GP. Gamma-Ray Induced Double-Strand Breaks in DNA Resulting from Randomly-Inflicted Single-Strand Breaks: Temporal Local Denaturation, a New Radiation Phenomenon? Int J Radiat Biol Relat Stud Phys Chem Med 1978;33:105-20.
2. Ruhe M, Rabe D, Jurischka C, et al. Molecular biomarkers of DNA damage in diffuse large-cell lymphoma—a review. J Lab Precis Med 2019;4:5.
3. Hernández L, Terradas M, Martín M, et al. Highly Sensitive Automated Method for DNA Damage Assessment: Gamma-H2AX Foci Counting and Cell Cycle Sorting. Int J Mol Sci 2013;14:15810-26.
4. Liontos M, Anastasiou I, Bamias A, et al. DNA damage, tumor mutational load and their impact on immune responses against cancer. Ann Transl Med 2016;4:264.
5. Krum SA, de la Rosa Dalugdugan E, Miranda-Carboni GA, et al. BRCA1 Forms a Functional Complex with γ-H2AX as a Late Response to Genotoxic Stress. J Nucleic Acids 2010;2010:801594.
6. Clingen PH, Wu JY, Miller J, et al. Histone H2AX phosphorylation as a molecular pharmacological marker for DNA interstrand crosslink cancer chemotherapy.

Biochem Pharmacol 2008;76:19-27.

7. Nikolova T, Dvorak M, Jung F, et al. The γH2AX Assay for Genotoxic and Nongenotoxic Agents: Comparison of H2AX Phosphorylation with Cell Death Response. Toxicol Sci 2014;140:103-17.
8. Venkitaraman AR. Functions of BRCA1 and BRCA2 in the biological response to DNA damage. J Cell Sci 2001;114:3591-8.
9. Chen CC, Feng W, Lim PX, et al. Homology-Directed Repair and the Role of BRCA1, BRCA2, and Related Proteins in Genome Integrity and Cancer. Annu Rev Cancer Biol 2018;2:313-36.
10. Wu Z, Li S, Tang X, et al. Copy Number Amplification of DNA Damage Repair Pathways Potentiates Therapeutic Resistance in Cancer. Theranostics 2020;10:3939-51.
11. Huang RX, Zhou PK. DNA damage response signaling pathways and targets for radiotherapy sensitization in cancer. Signal Transduct Target Ther 2020;5:60.
12. Kelley MR, Logsdon D, Fishel ML. Targeting DNA repair pathways for cancer treatment: what's new? Future Oncol 2014;10:1215-37.
13. Biau J, Chautard E, Verrelle P, et al. Altering DNA Repair to Improve Radiation Therapy: Specific and Multiple Pathway Targeting. Front Oncol 2019;9:1009.
14. Gavande NS, VanderVere-Carozza PS, Hinshaw HD, et al. DNA repair targeted therapy: The past or future of cancer treatment? Pharmacol Ther 2016;160:65-83.
15. Das S, Camphausen K, Shankavaram U. Pan-Cancer Analysis of Potential Synthetic Lethal Drug Targets Specific to Alterations in DNA Damage Response. Front Oncol 2019;9:1136.
16. Liscio P, Camaioni E, Carotti A, et al. From Polypharmacology to Target Specificity: The Case of PARP Inhibitors. Curr Top Med Chem 2013;13:2939-54.
17. Redon CE, Nakamura AJ, Zhang YW, et al. Histone gammaH2AX and poly(ADP-ribose) as clinical pharmacodynamic biomarkers. Clin Cancer Res 2010;16:4532-42.
18. Kornberg Z, Chou J, Feng FY, et al. Prostate cancer in the era of "Omic" medicine: recognizing the importance of DNA damage repair pathways. Ann Transl Med 2018;6:161.
19. Mayeux R. Biomarkers: Potential Uses and Limitations. NeuroRx 2004;1:182-8.
20. Osoegawa A, Hiraishi H, Hashimoto T, et al. The Positive Relationship Between γH2AX and PD-L1 Expression in Lung Squamous Cell Carcinoma. In Vivo 2018;32:171-7.
21. Goodman AM, Kato S, Bazhenova L, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. Mol Cancer Ther 2017;16:2598-608.
22. Eberlein U, Peper M, Fernández M, et al. Calibration of the γ-H2AX DNA Double Strand Break Focus Assay for Internal Radiation Exposure of Blood Lymphocytes. PLoS One 2015;10:e0123174.
23. Ivashkevich A, Redon CE, Nakamura AJ, et al. Use of the γ-H2AX assay to monitor DNA damage and repair in translational cancer research. Cancer Lett 2012;327:123-33.
24. Lomax ME, Folkes LK, O'Neill P. Biological consequences of radiation-induced DNA damage: relevance to radiotherapy. Clin Oncol (R Coll Radiol) 2013;25:578-85.
25. Reddig A, Rübe CE, Rödiger S, et al. DNA damage assessment and potential applications in laboratory diagnostics and precision medicine. J Lab Precis Med 2018;3:31.
26. Weingeist DM, Ge J, Wood DK, et al. Single-cell microarray enables high-throughput evaluation of DNA double-strand breaks and DNA repair inhibitors. Cell Cycle 2013;12:907-15.
27. Willers H, Gheorghiu L, Liu Q, et al. DNA Damage Response Assessments in Human Tumor Samples Provide Functional Biomarkers of Radiosensitivity. Semin Radiat Oncol 2015;25:237-50.
28. Jeggo PA, Löbrich M. DNA double-strand breaks: their cellular and clinical impact? Oncogene 2007;26:7717-9.
29. Redon CE, Nakamura AJ, Martin OA, et al. Recent developments in the use of γ -H2AX as a quantitative DNA double-strand break biomarker. Aging 2011;3:168-74.
30. Rödiger S, Liefold M, Ruhe M, et al. Quantification of DNA double-strand breaks in peripheral blood mononuclear cells from healthy donors exposed to bendamustine by an automated γH2AX assay—an exploratory study. J Lab Precis Med 2018;3:47.
31. Rothkamm K, Barnard S, Ainsbury EA, et al. Manual versus automated γ-H2AX foci analysis across five European laboratories: Can this assay be used for rapid biodosimetry in a large scale radiation accident? Mutat Res 2013;756:170-3.
32. Carragher NO, Brunton VG, Frame MC. Combining imaging and pathway profiling: an alternative approach to cancer drug discovery. Drug Discov Today 2012;17:203-14.
33. Carpenter AE, Jones TR, Lamprecht MR, et al. CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. Genome Biol 2006;7:R100.
34. Lengert N, Mirsch J, Weimer RN, et al. AutoFoci, an automated high-throughput foci detection approach for

analyzing low-dose DNA double-strand break repair. Sci Rep 2018;8:17282.

35. Schneider J, Weiss R, Ruhe M, et al. Open source bioimage informatics tools for the analysis of DNA damage and associated biomarkers. J Lab Precis Med 2019;4:21.
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2020.
37. Zeileis A, Kleiber C, Jackman S. Regression Models for Count Data in R. J Stat Softw 2008. doi: 10.18637/jss.v027.i08.
38. Harris T, Hilbe JM, Hardin JW. Modeling Count Data with Generalized Distributions. Stata Journal 2014;14:562-79.
39. Cox DR. Some Remarks on Overdispersion. Biometrika 1983;70:269-74.
40. Vandersickel V, Beukes P, Van Bockstaele B, et al. Induction and disappearance of γH2AX foci and formation of micronuclei after exposure of human lymphocytes to 60 Co γ-rays and p(66)+ Be(40) neutrons. Int J Radiat Biol 2014;90:149-58.
41. Vinnikov VA, Ainsbury EA, Maznyk NA, et al. Limitations Associated with Analysis of Cytogenetic Data for Biological Dosimetry. Radiat Res 2010;174:403-14.
42. Cameron A, Trivedi P. Regression-based tests for overdispersion in the Poisson model. J Econom 1990;46:347-64.
43. Sellers KF, Morris DS. Underdispersion Models: Models That Are "under the Radar." Commun Stat Theory Methods 2017;46:12075-86.
44. Sellers KF, Swift AW, Weems KS. A flexible distribution class for count data. J Stat Distrib Appl. 2017;4:22.
45. Gardner W, Mulvey EP, Shaw EC. Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models. Psychol Bull 1995;118:392-404.
46. Perumean-Chaney SE, Morgan C, McDowall D, et al. Zero-Inflated and Overdispersed: What's One to Do? J Stat Comput Simul 2013;83:1671-83.
47. Guthrie KA, Gammill HS, Kamper-Jørgensen M, et al. Statistical Methods for Unusual Count Data: Examples from Studies of Microchimerism. Am J Epidemiol 2016;184:779-86.
48. Lee JH, Han G, Fulp WJ, et al. Analysis of Overdispersed Count Data: Application to the Human Papillomavirus Infection in Men (HIM) Study. Epidemiol Infect 2012;140:1087-94.
49. Dean CB. Testing for Overdispersion in Poisson and Binomial Regression Models. J Am Stat Assoc 1992;87:451-7.
50. Lim HK, Song J, Jung BC. Score Tests for Zero-Inflation and Overdispersion in Two-Level Count Data. Comput Stat Data Anal. 2013;61:67-82.
51. Thas O, Rayner JCW. Smooth Tests for the Zero-Inflated Poisson Distribution. Biometrics 2005;61:808-15.
52. Walters GD. Using Poisson Class Regression to Analyze Count Data in Correctional and Forensic Psychology: A Relatively Old Solution to a Relatively New Problem. Crim Justice Behav 2007;34:1659-74.
53. Yang Z, Hardin JW, Addy CL. Testing Overdispersion in the Zero-Inflated Poisson Model. J Stat Plan Inference 2009;139:3340-53.
54. McCullagh P, Nelder JA. Generalized Linear Models. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. 2nd edition. Chapman & Hall, 1989.
55. Raftery AE. Bayesian Model Selection in Social Research. Sociol Methodol 1995;25:111-63.
56. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th edition. New York: Springer, 2002.
57. Ruhe M, Dammermann W, Lüth S, et al. Effect of cryopreservation on the formation of DNA double strand breaks in human peripheral blood mononuclear cells. Journal of Cellular Biotechnology 2018;4:67-73.
58. Weiss R, Rödiger S. NucDetect - A python package for Detection and Quantification of DNA Doublestrand Breaks. 2020. Available online: https://pypi.org/project/NucDetect/01115.dev2/
59. Zhang Z. The role of big-data in clinical studies in laboratory medicine. J Lab Precis Med 2017;2:34.
60. Rödiger S, Friedrichsmeier T, Kapat P, et al. RKWard: a comprehensive graphical user interface and integrated development environment for statistical analysis with R. J Stat Softw 2012. doi: 10.18637/jss.v049.i09.
61. Oliveira M, Einbeck J, Higueras M, et al. Zero-Inflated Regression Models for Radiation-Induced Chromosome Aberration Data: A Comparative Study. Biom J 2016;58:259-79.
62. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. J Stat Softw 2015. doi: 10.18637/jss.v064.i04.
63. Cokelaer T, Brian, Stringari CE, et al. cokelaer/fitter: v1.2.3 synchronised on pypi. Zenodo, 2020.
64. Hulsen T, Jamuar SS, Moody AR, et al. From Big Data to Precision Medicine. Front Med (Lausanne) 2019;6:34.
65. Springate DA, Parisi R, Olier I, et al. rEHR: An R package

for manipulating and analysing Electronic Health Record data. PLoS One 2017;12:e0171784.
66. Choi G, Lee K, Seo D, et al. Analysis of Medical Data Using the Big Data and R. In: Park DS, Chao HC, Jeong YS, et al. editors. Advances in Computer Science and Ubiquitous Computing. Lecture Notes in Electrical Engineering; vol. 373. Singapore: Springer, 2015:867-73.
67. Zhou ZR, Wang WW, Li Y, et al. In-depth mining of clinical data: the construction of clinical prediction model with R. Ann Transl Med 2019;7:796.
68. Leeper TJ. Archiving Reproducible Research with R and Dataverse. R J 2014;6:151-8.
69. Schulz S, Stegwee R, Chronaki C. Standards in Healthcare Data. In: Kubben P, Dumontier M, Dekker A. editors. Fundamentals of Clinical Data Science. Cham: Springer International Publishing, 2019:19-36.
70. Petersen AH, Ekstrøm CT. dataMaid: Your Assistant for Documenting Supervised Data Quality Screening in R. J Stat Softw 2019. doi: 10.18637/jss.v090.i06.
71. Caie PD, Walls RE, Ingleston-Orme A, et al. High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells. Mol Cancer Ther 2010;9:1913-26.

Cite this article as: Chilimoniuk J, Gosiewska A, Słowik J, Weiss R, Deckert PM, Rödiger S, Burdukiewicz M. countfitteR: efficient selection of count distributions to assess DNA damage. Ann Transl Med 2021;9(7):528. doi: 10.21037/atm-20-6363