bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A field-wide assessment of differential high throughput sequencing reveals widespread bias

A Preprint

Taavi Päll
Department of Microbiology
Institute of Biomedicine and Translational Medicine, University of Tartu, Estonia
Ravila 19, 50411, Tartu, Estonia
taavi.pall@ut.ee

Hannes Luidalepp
Quretec (https://www.quretec.com)
Ülikooli 6a, 51003, Tartu, Estonia
luidale@gmail.com

Tanel Tenson
Institute of Technology, University of Tartu, Estonia
Nooruse 1, 50411, Tartu, Estonia
tanel.tenson@ut.ee

Ülo Maiväli *
Institute of Technology, University of Tartu, Estonia
Nooruse 1, 50411, Tartu, Estonia
ymaivali@gmail.com

May 14, 2021

Abstract

Here we assess reproducibility and inferential quality in the field of differential HT-seq, based on analysis of datasets submitted 2008-2019 to the NCBI GEO data repository. Analysis of GEO submission file structures places an overall 56% upper limit to reproducibility without querying other sources. We further show that only 23% of experiments resulted in theoretically expected p value histogram shapes, although both reproducibility and p value distributions show marked improvement over time. Uniform p value histogram shapes, indicative of <100 true effects, were extremely few. Our calculations of π_0, the fraction of true nulls, showed that 36% of experiments have π_0 <0.5, meaning that in over a third of experiments most RNA-s were estimated to change their expression level upon experimental treatment. Both the fraction of different p value histogram types and π_0 values are strongly associated with the software used for calculating these p values by the original authors, indicating widespread bias.

Introduction

Over the past decade a feeling that there is a crisis in experimental science has increasingly permeated the thinking of methodologists, captains of industry, working scientists, and even the lay public (Ioannidis 2005; Baker 2016; Begley and Ellis 2012; Prinz, Schlange, and Asadullah 2011; Harris 2017). This manifests in poor statistical power to find true effects (Button et al. 2013), in poor reproducibility (defined as getting identical results when reanalysing the original data by the original analytic workflow), and in poor replicability (defined as getting similar results after repeating the entire experiment) of the results (Goodman, Fanelli, and Ioannidis 2016). While reproducibility depends on the availability of data and on the quality of description of the data analytic workflow, replicability depends on the quality of experiments, on the quality of data analysis,

*corresponding author

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

including data pre-processing, and on the power of the experiment to detect true effects. The proposed reasons behind the crisis include sloppy experimentation, selective publishing, perverse incentives, difficult-to-run experimental systems, insufficient sample sizes, over-reliance on null hypothesis testing, and much-too-flexible analytic designs combined with hypothesis-free study of massively parallel measurements (Grimes, Bauch, and Ioannidis 2018; Maiväli 2015; Munafò et al. 2017; Szucs and Ioannidis 2017; Botvinik-Nezer et al. 2020; Leng and Leng 2020). Although there have been attempts at assessing experimental quality through replication of experiments, mostly in psychology, prohibitive cost and theoretical shortcomings in analysing concordance in experimental results have encumbered this approach in biomedicine (Hardwicke et al. 2020; Patil, Peng, and Leek 2016; Leek and Jager 2017).

Another way to assess the large-scale quality of a science is to employ a surrogate measure for quality that can be more easily obtained than full replication of a study. The most often used such measure is technical reproducibility, which involves checking for availability and/or running the original analysis code on the original data. Although the evidence-base for reproducibility is still sketchy, it seems to be well <50% in several fields of biomedicine (Leek and Jager 2017). However, as there are many reasons, why a successful reproduction might not indicate a good quality of the original study, or why an unsuccessful reproduction may not indicate a bad quality of the original study, the criterion of reproducibility is clearly insufficient. Yet another quality proxy can be found in published p values, especially the distribution of p values. In a pioneering work Jager and Leek extracted ca. 5000 statistically significant p values from abstracts of leading medical journals and pooled them to formally estimate, from the shape of the ensuing p value distribution, the "science-wide false discovery rate" or SWFDR as 14% (Jager and Leek 2014a). However, as this estimate rather implausibly presupposes that the original p values were calculated uniformly correctly, and that unbiased sets of significant p values were obtained from the abstracts, they subsequently revised their estimate of SWFDR upwards, as "likely not >50%"(Jager and Leek 2014b). For observational medical studies, by an independent method, a plausible estimate for field-wide FDR was found to be somewhere between 55% and 85%, depending on the study type (Schuemie et al. 2014).

While our work uses published p values as evidence for field wide quality and presupposes access to unbiased full sets of p values, it does not pool the p values across studies, nor does it assume that they were correctly calculated. In fact, we assume the opposite and do a study-by-study analysis of the quality of calculation of p values. This makes the quality of the p value a proxy for the quality of the experiment and/or of the scientific inferences based on these p values. We do not see our estimate of the fraction of poorly calculated p values as a formal quality metric. We merely hope that by this measure we can shed some light into the overall quality of a field. However, we chose the field whose quality to assess, so as to maximize the potential weight of p values on scientific inferences.

Here we assess *in silico* the reproducibility and replicability of high throughput differential expression studies by next-generation sequencing (HT-seq). We concentrate on the HT-seq field for two reasons. HT-seq has become the gold standard for whole transcriptome gene expression quantification, both in research and in clinical applications. And secondly, due to the massively parallel testing in individual studies of tens of thousands of features per experiment, we have access to study-wide unbiased lists of p values. From the shapes of histograms of p values we can find the experiments where p values were calculated apparently correctly, and from these studies we can determine the study-wise relative frequencies of true nulls (the π_0-s). Also, we believe that the very nature of the HT-seq field, where a single biological experiment entails comparing the expression levels of about 20,000 features (e.g. RNA-s) on average, predicates that the quality of data analysis, and specifically statistical inference based on p values (directly, or indirectly through FDR) must play a decisive part in scientific inference. Simply, one cannot analyse an HT-seq experiment intuitively, without resorting to formal statistical inference. Therefore, quality problems of statistical analysis would very likely directly and substantially impact the quality of science. Thus we use the quality of statistical analysis as a proxy for the quality of science, with the understanding that this proxy may work better for modern data-intensive fields, where scientist's intuition has a comparatively smaller role to play.

Results

Assessing reproducibility by NCBI GEO database supplementary files

We queried the NCBI GEO database for "expression profiling by high throughput sequencing" (for exact query string, please see Methods), retrieving 32,017 datasets (GEO series) from 2006, when first HT-seq dataset was submitted to GEO, to Dec-31, 2019. The number of yearly new HT-seq submissions increased from 1 in 2006 to 8770 by 2019, making up 27.4% of all GEO submissions in 2019. Most of the GEO HT-seq

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

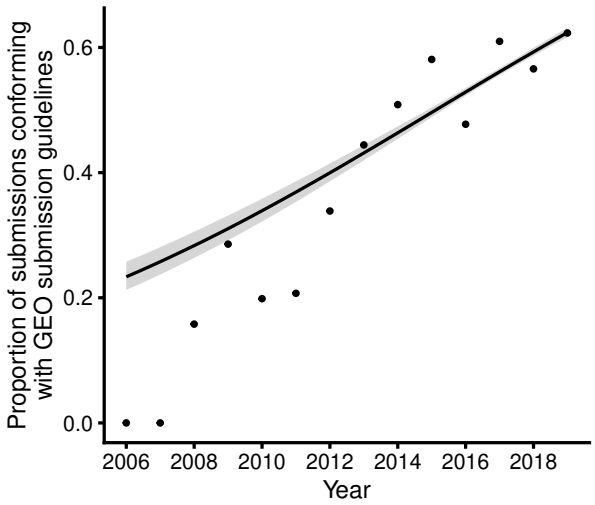


Figure 1: The increasing proportion of GEO submissions conforming with submission guidelines in regard of inclusion of processed data is estimated from a bernoulli logistic model. Line denotes linear model best fit using model formula *conforms* $\sim$ *year*, bernoulli likelihood. Download model object conforms_year.rds. Shaded area denotes 95% credible region. N = 32,017. Points denote yearly proportions of conforming GEO submissions.

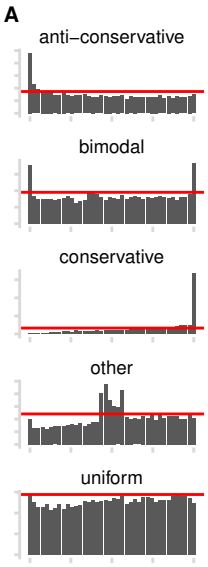
submissions are from human and mouse, and they increase in a much faster rate than submissions from other taxa (data not shown). We filtered the GEO series containing supplementary processed data files. NCBI GEO database submissions follow MINSEQE guidelines (Functional Genomics Data Society 2008, 2012). Processed data are a required part of GEO submissions, defined as the data on which the conclusions in the related manuscript are based. The format of processed data files submitted to GEO is not standardized, but in case of expression profiling such files include, but are not limited to, quantitative data for features of interest, e.g. mRNA, in tabular format. Sequence read alignment files and coordinates (SAM, BAM, and BED) are not considered as processed data by GEO. As reproducibility by definition entails arriving at the original conclusions by independently repeating original analysis (Peng 2011), we surmise that the fraction of GEO submission with processed data files gives an upper limit of reproducibility (upper limit, because processed data files *per se* do not guarantee reproducibility). According to our analysis 17,920 GEO series, containing 43,340 supplementary processed data files, conform with GEO submission guidelines. After further excluding the submissions with potentially non-tabular supplementary files, based on file extensions (see Methods for details), the number of GEO series was reduced to 15,520, containing 31,862 supplementary files, including tar.gz archives, which we downloaded from the GEO server. From those we programmatically imported 32,764 files as tables, resulting in 32,414 (99%) successfully imported files. For the purpose of reproducibility analysis, we considered all 17,920 GEO series with supplementary processed files out of 32,017 published GEO series in our time window as potentially reproducible. Therefore, the observed overall 56% retention rate of GEO submissions with processed data files should be seen as an upper bound for reproducibility without querying other sources. There is a substantial increase of the retention rate over time, from 16% (3/19) in 2008 to 62% (5,465/8,770) in 2019, indicating increasing reproducibility of the HT-seq field (Figure 1).

According to GEO submission requirements, the processed data files may contain raw counts of sequencing reads, and/or normalized abundance measurements. Therefore, a valid processed data submission may or may not contain lists of p values. We identified p values from 2,109 GEO series, from which we extracted 6,267 unique p value sets. While the mean number of p value lists, each list corresponding to a separate experiment, per 2,109 GEO submissions was 2.97 (max 66), 49% of submissions contained a single p value list and 78% contained 1-3 p value lists. For further analysis we randomly selected one p value list per GEO series.

P value histograms

We algorithmically classified the p value histograms into five classes (Breheny, Stromberg, and Lambert 2018) (See Methods for details and Figure 2A for representative examples). The "Uniform" class contains

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



B

Class	N	Fraction [95% CI]
anti−conservative	485	0.23 [0.21; 0.25]
bimodal	629	0.3 [0.28; 0.32]
conservative	242	0.11 [0.1; 0.13]
other	750	0.36 [0.33; 0.38]
uniform	3	0.0016 [0.00019; 0.0033]

Figure 2: Classes of p value histograms. A. Examples of p value histogram classes. Red line denotes QC threshold used for dividing p value histograms into discrete classes. B. Summary of p value histograms identified from GEO supplementary files. One p value set was randomly sampled from each GEO series where p values were identified. N = 2,109. 95% CI denotes credible intervals calculated from a binomial model with formula *Class* $\sim$ *1*. Download model object Class_1.rds.

flat p value histograms indicating no true effects (at the sample sizes used to calculate these p values). The "Anti-Conservative" class contains otherwise flat histograms that contain a spike near zero. The "Conservative" class contains histograms that have a distinct spike close to one. The "Bimodal" histograms have two peaks, one at either end. The class "Other" contains a panoply of malformed histogram shapes (humps in the middle, gradual increases towards one, spiky histograms, etc.). The "Uniform" and "Anti-Conservative" histograms are the theoretically expected shapes of p value histograms.

We found that overall, 23% of the histograms fall into anti-conservative class, 12% were conservative, 30% bimodal and 36% fell into class "other" (Figure 2B). Only 3 of the 2,109 histograms were classified as "uniform". Median number of features in our sample was 21,084; interestingly there is an apparent leftward shift in the peak of distribution of features in anti-conservative histograms, as compared to histograms with all other shapes, suggesting different data pre-processing for datasets resulting in anti-conservative histograms (Figure 2–figure supplement 1). Logistic regression reveals a clear trend for increasing proportion of anti-conservative histograms, starting from <10% in 2010 and topping 25% in 2018 (Figure 3–figure supplement 1). Hierarchical modelling indicates that all differential expression (DE) analysis tools and sequencing platforms exhibit similar temporal increases of anti-conservative p value histograms (Figure 3–figure supplement 2-3). Multinomial hierarchical logistic regression further demonstrated that the increase in the fraction of anti-conservative histograms is accomplished by decreases mostly in the class "other", irrespective of the DE analysis tool (Figure 3A).

This positive temporal trend in anti-conservative p value histograms suggests improving quality of the HT-seq field. Rather surprisingly, Figure 3A also indicates that different DE analysis tools are associated with very different proportions of p value histogram classes, suggesting that quality of p value calculation, and therefore, quality of scientific inferences based on these p values depends on DE analysis tool. We further tested this conjecture in a simplified model, restricting our analysis to 2018-2019, the final years in our dataset (Figure 3B). As no single DE analysis tool dominates the field – cuffdiff 23%, deseq 33%, edgeR 13%, limma 2%, unknown 29% (see Figure 3–figure supplement 4 for temporal trends) –, a state of affairs where proportions of different p value histogram classes do not substantially differ between analysis tools would indicate lack of DE analysis tool-specific bias to the results. However, we found by multinomial regression that all p value histogram classes, except "uniform", which is largely unpopulated, depend strongly on the DE analysis tool used to calculate the p values (Figure 3B). This is confirmed by modelling the frequency of the anti-conservative p value histograms in binomial logistic regression (Figure 3–figure supplement 5A).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

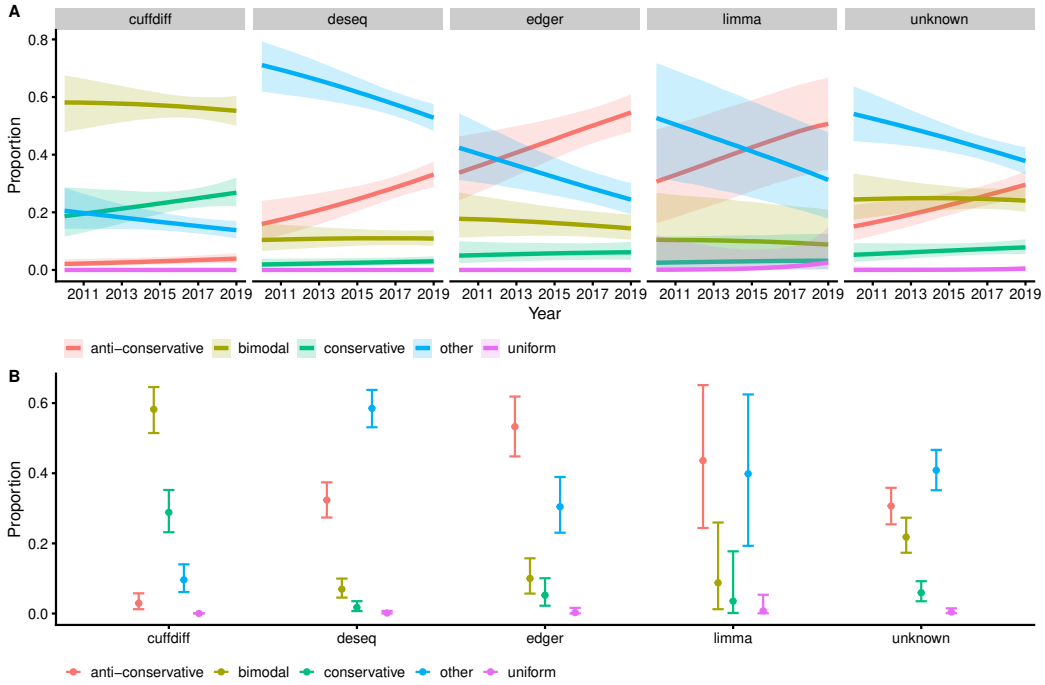


Figure 3: Association of the p value histogram class with differential expression analysis tool. A. The increase in the proportion of anti-conservative histograms is accompanied by decreases mostly in the class "other", irrespective of the DE analysis tool. Lines denote best fit of model *class* $\sim$ *year + (year | de_tool)*, categorical likelihood. Download model object Class_year___year_detool_year.rds. Shaded areas denote 95% credible regions. N = 2,109. B. Association of p value histogram type with DE analysis tool; data is restricted to 2018-2019 GEO submissions. Points denote linear model fit *class* $\sim$ *de_tool*, categorical likelihood. Download model object Class_detool_2018-19.rds. Error bars denote 95% credible intervals. N = 980.

Using the whole dataset of 6,267 p value histograms – as a check for robustness of results – or adjusting the analysis for GEO publication year, of taxon (human, mouse, and pooled other), of the RNA source or sequencing platform – as a check for possible confounding – does not change this conclusion (Figure 3–figure supplement 5B-E). The lack of confounding in our results allow a causal interpretation, indicating that DE analysis tools bias the analysis of HT-seq experiments (Pearl, Glymour, and Jewell 2016).

Proportion of true nulls

To further enquire into DE analysis tool-driven bias we estimated from user-submitted p values the fraction of true null effects (the π_0) for each HT-seq experiment. As non-anti-conservative sets of p values (excepting the "uniform") indicate problems during the respective experiments and/or data analyses, we only calculated the π_0 for datasets with anti-conservative and uniform p value distributions (n = 488). Nevertheless, the π_0-s show an extremely wide distribution, ranging from 0.999 to 0.06. Remarkably, 36% of the π_0 values are smaller than 0.5, meaning that in those experiments over half of the features (e.g. mRNA-s) are estimated to change their expression levels upon experimental treatment (Figure 4A). Conversely, only 21% of π_0-s exceed 0.8, and 8.5% exceed 0.9. Intriguingly, the peak of the π_0 distribution is not near 1, as might be expected from experimental design considerations, but there is a wide peak between 0.5 and 0.8 (median and mean π_0-s are both at 0.59). The median π_0-s range over 20 percentage points, from 0.5 to 0.7, depending on the DE analysis tool (Figure 4B). Using the whole dataset qualitatively confirms the robustness of this analysis, also producing narrower credible intervals, due to larger sample (N = 1,567) (Figure 4–figure supplement 1A).

In addition, mean π_0 tend to rise in time, similarly to fraction of anti-conservative p value histograms, and this increase is also common to all DE analysis tools (Figure 4C). Controlling for time, taxon or sequencing platform, did not substantially change the association of DE analysis tools with the π_0-s, except for the multilevel models, which resulted in substantially larger estimation uncertainty (Figure 4–figure supplement 1B-E). Recalculating the π_0-s with a different algorithm (Storey 2002) and reanalysing the data did not

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

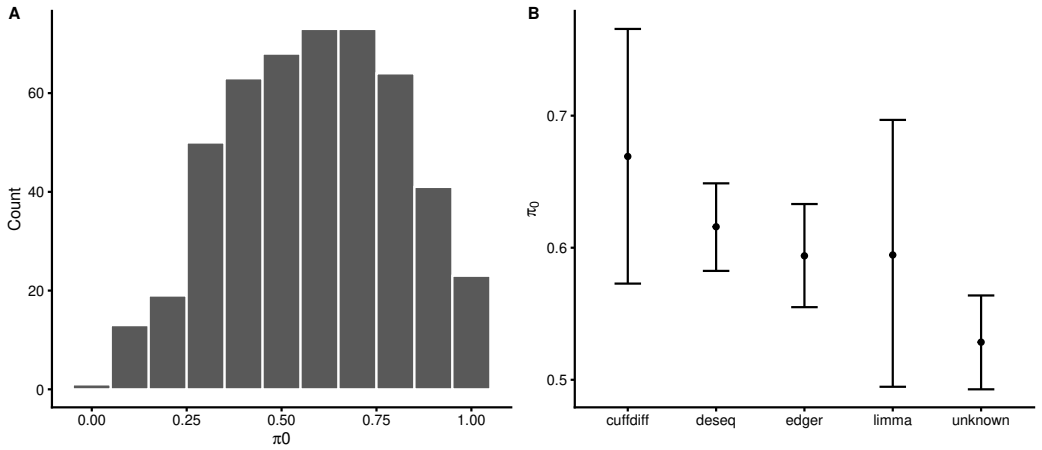


Figure 4: Association of the proportion of true null effects (π_0) with DE analysis tool. A. Histogram of π_0 values estimated from anti-conservative and uniform p value sets. N = 488. B. Robust linear model (*pi0* ∼ *de_tool*, student's t likelihood) indicates association of π_0 with DE analysis tool. Download model object pi0_detool_sample.rds. Points denote best estimate for the mean π_0 and error bars denote 95% credible intervals. N = 488.

change these conclusions (data not shown). As there is a strong association between both proportion of anti-conservative p value histograms and π_0 with DE analysis tool, we further checked for, and failed to see, similar associations with variables from raw sequence metadata, such as the sequencing platform, library preparation strategies, library sequencing strategies, library selection strategies, and library layout (single or paired) (Figure 3–figure supplement 6-9, Figure 4–figure supplement 2-5). These negative results support the conjecture of specificity of the associations with DE analysis tools.

Curing of p value histograms by removing low-count features

Removal of low-expressed genes before model fitting in DE analysis is a recommended step in the considered DE analysis tools, like edgeR, DESeq2, limma-voom. Threshold for filtration is arbitrary and should be decided by researcher based on data on hand. Removal of low-expressed genes from a dataset before modeling is suggested as their levels might not be biologically relevant, their removal benefits computation, and increases the sensitivity of finding true effects (Dialsingh, Austin, and Altman 2015). Moreover, we observed a small leftward shift in the p value length distribution (number of p values in a set) of anti-conservative histograms compared to a length distribution of all other p value set shapes, suggesting that p value sets with anti-conservative shape are more likely to be pre-filtered (Figure 2–figure supplement 1). Accordingly, the "Conservative" and "Other" histograms represent unsuccessful attempts at calculating p values. we speculated that we could "rescue" some of the untoward p value histograms by converting them into anti-conservative or uniform types, simply by filtering out features with low counts. Our goal here was not to provide and optimal interventions for individual datasets, which would require tailoring the filtering algorithm for the compositional properties of each dataset, but merely to provide proof of principle evidence for or against the general hypothesis that by a simple filtering approach we could increase the proportion of anti-conservative p value sets and/or reduce the dependence of results on the analysis platform. Therefore we applied arbitrary conservative thresholds to 1,720 p value sets where we were able to identify gene expression values (see Methods for details). We found that overall we could increase the proportion of anti-conservative p value histograms by 2.6-fold, from 368 (21.4%) to 955 (55.5%), and the number of uniform histograms from 2 (0.1%) to 9 (0.5%) (Figure 5A).

The proportion of rescued p value histograms differ considerably between analysis platforms. The platform with the lowest pre-rescue proportion of anti-conservative histograms (2.6%), cuffdiff, increased to 37% (14.2-fold; Figure 5B). In contrast, DESeq/DESeq2 increased from 28% to 71% (2.5-fold; Figure 5C), edgeR increased from 51% to 68% (1.34-fold; Figure 5D), limma increased from 58% to 76% (1.34-fold; Figure 5E), and the class "other" increased from 25% to 58% (2.33-fold; Figure 5F). For all platforms, the vast majority of rescued p value distributions came from classes "bimodal" and "other", while almost no rescue was detected from conservative histograms. To see, whether our intervention would reduce or abolish the dependence

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

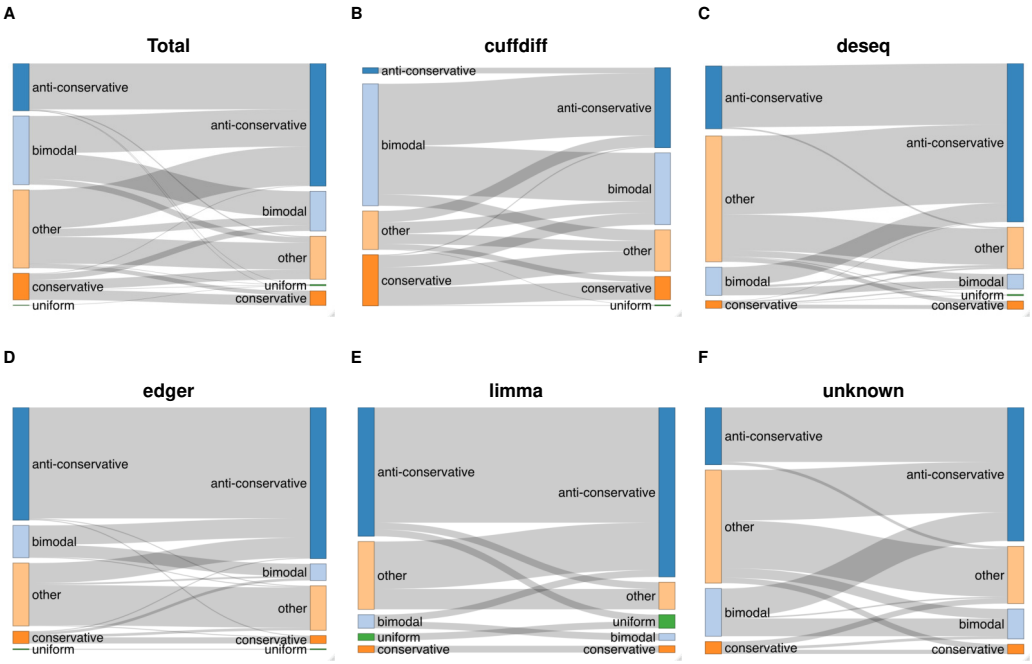


Figure 5: Schematic representation of the influence on p value histogram class of removing low-count features. Line thickness is linearly proportional to the number of p value sets that change their distributional class. Only the 1,720 experiments that could be subjected to this treatment are depicted. A. Full data, N=1,720. B. The subset where the p values were calculated with cuffdiff, N=693. C. The subset where the p values were calculated with DESeq/DESeq2, N=572. D. The subset where the p values were calculated with edgeR, N=253. E. The subset where the p values were calculated with limma, N=33. F. The subset where the p values were calculated with unassigned analysis platform, N=169.

of p value distribution type and π_0 on the analysis platform, we modeled the dependence on data analytic platform of the type of post-intervention p value sets and of π_0-s, as calculated from rescued anti-conservative p value sets (Figure 5–figure supplement 1A and 1B, respectively). The results did not substantially differ from those obtained from the full pre-rescue p value sets, indicating that applying low-counts filtering does not change the dependency on analysis platform (compare Figure 5–figure supplement 1A to Figure 3–figure supplement 5B and Figure 5–figure supplement 1B to Figure 4–figure supplement 1A).

Publication impact

We investigated whether the experiments, whose statistical analysis resulted on anti-conservative p value distributions, were published, on average, in higher performing journals, and whether such papers collect more citations. For journal performance we used the Elsevier CiteScore, which employs a relatively long three-year window and all document types published in a given journal, making it a more stable and robust metric than journal IF (Zijlstra and McCullough 2016; Okagbue and Silva 2020). Our analysis suggests a negative association, where, in comparisons with the lowest ranked journals, the experiments with anti-conservative p value distributions are underrepresented by 0.26–17.7 percentage points (95% CI) in journals with the highest cite scores (Figure 6A). This is qualitatively similar to an observed negative association of reproducibility of qPCR experiments with a journal quality metric (Bustin et al. 2013). However, it must be noted that in our analysis the weight of evidence for the negative association is moderate at best (the posterior probability of a negative association is 0.98, and the probability that the full effect size (CiteScore of 0.1 vs. 60) is in an arbitrarily meaningful range, above 5 percentage points, is 0.74).

In our analysis of citations we failed to find a statistically supported association of experiments containing anti-conservative p value histograms with citations per resulting paper (Figure 6B). However, the wide confidence band indicates that this result cannot be construed as a claim of no effect, as reasonably big effects in both directions are consistent with the evidence.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

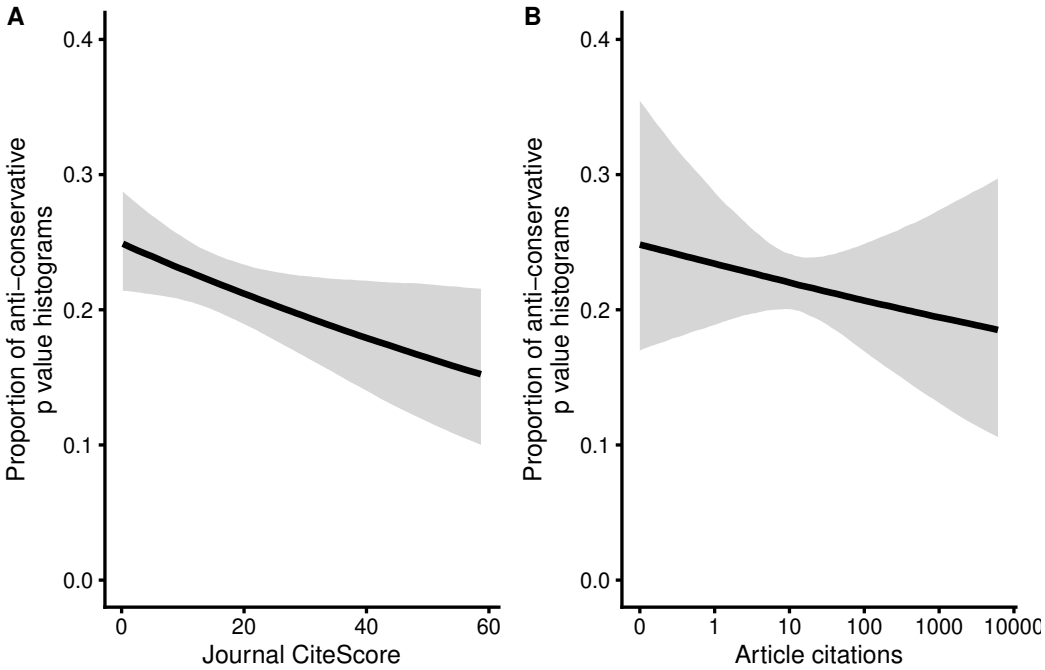


Figure 6: Association of anti-conservative p value distributions with publication impact. A. Increased journal CiteScore is related to decreasing proportion of anti-conservative p value histograms. Line denotes best fit of linear model *anticons* $\sim$ *CiteScore + year*, bernoulli likelihood. Download model object anticons___CiteScore_year.rds. Shaded area denotes 95% credible region. N=1544. B. Relationship of article citations with the proportion of anti-conservative p value histograms. Line denotes best fit of linear model *anticons* $\sim$ *citations + year*, bernoulli likelihood. Download model object anticons___log_citations_year.rds. Shaded area denotes 95% credible region. N=1544.

Discussion

In this work we have calculated five indicators of health of the HT-seq field: (i) the proportion of GEO submissions with published analysis endpoints in a tabular format, (ii) the relative fractions of classes of p value distribution shapes, (iii) the association of p value distribution shapes with DE analysis tool, (iv) the estimated proportion of true null effects (the π_0) for each experiment, and (v) the association of π_0 values with DE analysis tool. We believe that (i) indicates reproducibility, (ii) the quality of p value calculations (use of correct distributional models, etc.), (iv) the experimental design and data pre-processing choices, and that (iii) and (v) indicate bias.

Our analysis puts an upper bound of about 60% to differential HT-seq reproducibility in recent years, based on the presence of processed data files in the GEO submissions. To elucidate the meaning of this estimate, we must look into the relationship between the GEO submission data structures and the HT-seq workflows. The GEO repository requires three components for data submission: a metadata spreadsheet providing experimental design details, tabular processed data files, defined as the data on which the conclusions in the related manuscript are based (containing normalized abundance measurements or raw counts of sequencing reads), and raw data files (reads and quality scores as generated by the sequencing instrument). Even in the presence of raw data files, processed data files are practically necessary for assessing and reproducing the evidence behind the conclusions of a study, because of the large number of choices available to a discerning analyst. To put this in another way: scientific inference in the functional genomics field is based on null hypothesis testing and p values, and there isn't a single correct method to calculate them (nor is there one for FDR). Therefore, there is no guarantee that two independent analyses of the same data will obtain the same p values, and thus the same conclusions (Maiväli 2015). This means that access to original analytic choices (analysis code), or at least to p values, is needed for reproducibility.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The major points of divergence in the analytic pathway include aligning the sequences to genomic DNA, counts normalization, and differential expression testing (Nookaew et al. 2012; Sun and Zhu 2012). The analysis tools that are used for differential expression testing allow for a plethora of choices, including different distributional models, data transformations and basic analytic strategies, which can lead to different results through different trade-offs (Everaert et al. 2017; Nookaew et al. 2012). A similar state of affairs exists in the field of fMRI experiments, where reproducibility is questionable, and where 70 teams, testing the same hypotheses on the same data, used 70 different workflows, getting variable results (Carp 2012a, 2012b; Botvinik-Nezer et al. 2020).

Taken together, this leads us to conclude that access to correctly annotated read counts is a minimum requirement for assessing reproducibility of read pre-processing and alignment. More desirably, also raw p values are necessary to judge the evidence behind conclusions of a paper. For full reproducibility, starting from raw sequencing data, complete analysis instructions and modelling choices should be provided, which is currently not a part of the GEO submission protocol. Thus our reproducibility estimate does not refer to full reproducibility, but to mere potential of independently recreating conclusions of a paper. Although full reproducibility of the HT-seq field still seems elusive, the robust temporal trend of improvement, documented by us, gives reason for optimism.

While there is also a positive trend for the increasing fraction of anti-conservative p value sets, a strong majority of them yet fall into shapes that are considered problematic for successful analysis, including FDR/q value calculations. In fact, the most common class of p value histograms, "other", encompasses a diverse mixture of unruly shapes least likely to lead to good downstream analysis and interpretation of these p values.

There are very few uniform p value distributions, suggesting relatively few true effects. This surprising result was confirmed by visual re-examination of p value histograms. As a technical comment, it should be noted, that the assigned class of the p value histogram depends on arbitrarily set bin size. Our use of 40 bins leads to histograms, where an experiment with even around 100 true effects could reasonably lead to uniform histogram shape, because of swamping of the lowermost bin with p values emanating from true null effects (see Figure 2–figure supplement 2).

Importantly, the proportions of different classes of p value distributions differ greatly between DE analysis tools, indicating analysis tool-specific bias. We see similar tool-specific bias (and similar temporal increase) in the value of estimated fractions of true null effects (π_0). The insensitivity of this result to adding a time covariate into the model is consistent with our implicit assumption that the composition of deposited experiments didn't change over time, allowing a causal interpretation. As π_0 -s were calculated from anti-conservative p value sets only, this this analysis tool specific bias could mean (i) that the anti-conservativeness of a p value set is not a sufficient predictor of its quality in terms of further analysis, and/or (ii) that the algorithms implemented in differential expression analysis tools differ to a degree that introduces such bias downstream of the p value calculation.

To infer bias means attributing a causal interpretation to regression analysis, which is fundamentally acausal. We do so for three main reasons. Firstly, examination of manuals of the DE-analysis tools indicates that the causal interpretation is reasonable. Namely, different tools apply different amount of automation to data pre-processing and analysis – cuffdiff tries to automate everything and assures user that all biases are taken into account (Trapnell et al. 2012) DESeq2 workflow suggest some pre-processing steps to speed-up computations but uses automation to remove biases from input data (Love, Huber, and Anders 2014), whereas edgeR (McCarthy, Chen, and Smyth 2012a) and limma-voom require more interactive execution of separate pre-processing and analysis steps (Ritchie, Phipson, Wu, et al. 2015b; Law et al. 2014). We assume that popularity of cuffdiff and DESeq2 partly lies in their automation, as the user is largely relieved from decision making and can expect that the more experienced creators of these functions direct the analysis with robust choices. However, we found that cuffdiff is associated with the smallest proportion of anti-conservative p value histograms, whereas limma and edger, with their more hands-on approach, is associated with the biggest proportions of anti-conservative histograms.

Secondly, we performed a quasi-experimental intervention, where we removed low-expressed features from p value sets (Figure 5). Thus we were able to shift many of the unruly p value sets into the anti-conservative class. Interestingly, cuffdiff, which originally has the lowest fraction of anti-conservative p value distributions, as well as the lowest potential user input in the analysis, exhibits the highest rescue efficiency, consistently with causal effect of the platform. Also, edgeR and limma manuals recommend the use of the same data pre-processing tool from edgeR, which seems to result in identical rescue efficiencies, also indicating causal effect.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thirdly, and most importantly, we did several adjusted analyses (see figure supplements for Figure 3, Figure 4 and Figure 5), which one-by-one examine the effects on the statistical association of potential confounders. The apparent lack of such effects also suggests feasibility of a causal interpretation. This causal interpretation would be put into doubt, if it was the case that we have factor(s) that causally influence both which analysis platform researchers choose for their experiments and, independently of the analysis platform, the class of p value histogram/π_0 value of the resulting analysis (Pearl, Glymour, and Jewell 2016). For example, if some experimental scheme would necessitate using of a particular analysis program and also, and independently, would lead to a particular p value distribution. A possible candidate for such an experiment would be single-cell RNA-seq. However, single-cell experiments make up 8% (2,667) of GEO series in our dataset, yielding 2% (42) of p value sets, and they are not over-represented in any of the analysis tools (data not shown). More generally, as a possible caveat it cannot be excluded that some informal groups of labs use the same analysis platforms, effectively forming subcultures, and at the same time prefer to use some experimental designs that result in suboptimal/different p value distributions and π_0-s. However far-fetched by the criterion of Occam's razor, such a scenario, if true, would certainly modify our conclusions, turning the observed bias into a more local phenomenon. The DE analysis tool-specific medians of π_0 values range from 0.43 (tool "unknown") to 0.68 (tool "cuffdiff"), and the total median π_0 is 0.52, showing that by this criterion in an average experiment about half of the cellular RNA-s are expected to change their expression levels (Figure 4). In a biased situation the measured effects are a mixture of effects of the intended experimental treatment and of the undesirable effects of a more-or-less accidental choice of DE analysis tool, as no analytic workflow has been shown to systematically outperform the others (Everaert et al. 2017; Nookaew et al. 2012). Determining the relative weight of DE analysis tool in this mixture requires careful case-by-case study, as this depends both on the performed experiment (its variations, effect sizes, and actual laboratory implementation), and on the particular analytic choices reflected in the shapes of p value distributions and in π_0 values. However, from the evidence at hand, notably from the strong associations with DE analysis tool, in combination with the preponderance of extreme p value distributions and low π_0-s, we can conclude that such bias must be substantial, as it is widespread.

Another limitation of our study, also due to its large-scale nature, is our inability to pinpoint the sources of DE analysis tool-specific bias. However, a recent close study of 35 datasets shows data normalization to be a potentially important source of bias, which cannot be corrected by many of the currently widely used data normalization methods (Mandelboum et al. 2019; Quinn et al. 2019). Indeed, the widely used DE analysis tools use different RNA-seq data pre-processing strategies, all vulnerable to the situation where a large fraction of features change expression (McGee et al. 2019). In the light of our findings, the sources of bias clearly merit further case-by-case study of individual GEO submissions. To see whether we could shift p value distributions into desired shapes (anti-conservative and uniform), and possibly to abolish the analysis platform driven bias, by a simple measure, we did an intervention whereby we excluded low-count features from the p value sets from which distributional shapes were determined (Figure 5). Indeed, by this arbitrary and simple method we could increase the overall fraction of anti-conservative p value distribution by more than two-fold: from 21% to 56%. Interestingly, the rescue efficiencies differed greatly between experiments analyzed by different platforms, from >14-fold in the case of cuffdiff to 1.34-fold for edgeR and limma. These differences could be caused by differences in both the actual workflows and/or in the recommendations given to users in platform manuals. For example, the increase in proportion of anti-conservative p value distributions is very similar, at 1.34-fold, for edgeR and limma. While the edgeR "User's Guide" gives a biology-based explanation, why such filtering is beneficial, the limma "User's Guide" implies that filtering is just one of the required steps, and directs users to edger package function "filterByExpr" to carry out pre-filtering. In contrast, DESeq2, whose rescue efficiency is 2.5-fold, uses in-function heuristics to filter low-count genes, and their vignette accordingly recommends filtering only to increase computational efficiency. Also cuffdiff, which on its own generates very few anti-conservative p value sets, but provides highest efficiency rescue, tries to handle all biases automatically: "Cufflinks and Cuffdiff can automatically model and subtract a large fraction of the bias in RNA-seq read distribution across each transcript, thereby improving abundance estimates" (Trapnell et al. 2012). We speculate that this could provide users with a false sense of security.

Interestingly, the identical rescue efficiencies of edgeR and limma, which likely reflect their users employing the same filtering function, do not lead to reduction in the dependency of proportion of anti-conservative p value sets, or of π_0-s, on respective analysis platforms (Figure 5–figure supplement 1 and 2, respectively). Taken together with our general inability to abolish analysis platform driven bias by our simple and robust intervention, this suggests that bias comes from other aspect(s) of data processing and modeling. Nevertheless, this conjecture should be tempered with caution, as our robust filtering approach could hide biases of its own. On the other hand, the very inflexibility of our filtering, which does not take into account the idiosyncrasies of individual datasets, makes it highly probable that individual researchers can do better by individualizing the

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

filtering according to sequencing depth, number of features and scientific questions. As the goal of filtering is to get an anti-conservative p value distribution, one could simply opt for the least possible amount of filtering that leads to this outcome in their particular experiment. Could the relatively high frequency of low π_0-s reflect reality in the sense that in many experiments most RNA-s actually change expression levels? Although there is a small number of well-supported examples of this (Lin et al. 2012; Nie et al. 2012; Hu et al. 2014), it has been argued that the vast majority of genome wide differential gene expression studies ever conducted, including by HT-seq, have used experimental designs that would make it impossible to uncover such global effects, at least in a qualitatively accurate way (Lovén et al. 2012; Chen et al. 2016). The issue lies in the use of internal standards in normalizing the counts for different RNA-s (commonly normalizing for total read counts in each sample), which leads to completely wrong interpretations, if most genes change expression in one direction. To overcome this problem, one could use spike-in RNA standards like the External RNA Controls Consortium (ERCC) set (Ambion) or the Spike-in RNA Variants (SIRV) set (Lexogen) (Lun et al. 2017) and compositional normalization (McGee et al. 2019). However, even spike-in normalization requires great care to properly work in such extreme cases (Risso et al. 2014; Quinn et al. 2019), and outside single-cell RNA-seq it is used infrequently (McGee et al. 2019). In the absence of spike-in normalization, it seems likely that many or most of the low π_0 experiments represent technical failures, most likely during data normalization (McGee et al. 2019).

While the aim of this paper was to provide a birds-eye view on the HT-seq field, the dataset created to support this work provides added value by allowing access at individual GEO submission/experiment level. The dataset that accompanies this study allows to get a first estimate for the reproducibility and quality of conclusions of more than 30,000 HT-seq studies deposited in NCBI GEO. The dataset contains for every GEO submission the date of submission, association with publications, organisms studied, association with tabular files; and for every submitted tabular file it contains the name of file, the number of p value columns, the number of features, the p value histogram type and depiction of the histogram, π_0.
Finally, we note that our methodology can be adapted to study any type of experiment that results in at least several hundreds of parallel measurements/ p values, such as quantitative protein mass spectroscopy and metabolomics.

Strengths of the study:

- Our study is based on a large unbiased dataset, which allows for reliable quantification.
- All steps of the analysis are transparent and reproducible, as we provide full workflow and code for data mining, p value histogram classification, π_0 calculations, modelling, and figures.
- To the best of our knowledge this is the first large-scale study to offer quantitative insight into general quality of experimentation/data analysis/reproducibility of a fairly large field of biomedical science.
- The accompanying full dataset includes, for individual GEO submissions, useful quality control measures, like presence of tabular files describing analysis end-points, the shapes of the p value distributions and the estimated proportions of true null effects, providing evidence on the quality of inference of almost any HT-seq work in the literature.

Limitations of the study:

- It is a general study, whose measure of bias (statistical association inferred from many experiments) cannot be directly extended to single experiments. To do so requires (at least) incorporating additional information about the effect sizes encountered in the study of interest.
- From our dataset we cannot determine, which features of which differential expression analysis tools are responsible for the inferred bias. Neither can we say, which (if any) of the tools are better than the others. However, these questions have been addressed in the literature, and we hope that by stressing the seriousness of the problem, our study will inspire further work on the subject.
- Our estimate for reproducibility only provides an upper limit (presence of tabular files, which have the mere potential of containing enough pertinent information). Obtaining the true level of reproducibility requires actual reproduction of individual analyses, which we feel, cannot be automated. Although presence of tabular supplementary files is not a good measure of reproducibility of an individual study, their absence give a strong indication for irreproducibility of a study.

Methods

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

NCBI GEO database query and supplementary files

NCBI GEO database queries were performed using Bio.Entrez Python package and by sending requests to NCBI Entrez public API. The exact query string to retrieve GEO HT-seq datasets was 'expression profiling by high throughput sequencing[DataSet Type] AND ("2000-01-01"[PDAT] : "2019-12-31"[PDAT])'. FTP links from GEO datasets document summaries were used to download supplementary file names. Supplementary file names were filtered for downloading, based on file extensions, to keep file names with "tab", "xlsx", "diff", "tsv", "xls", "csv", "txt", "rtf", and "tar" file extensions. We dropped the file names where we did not expect to find p values using regular expression "filelist.txt|raw.tar$|readme|csfasta|(big)?wig|bed(graph)?|(broad_)?lincs".

NCBI supplementary file processing

Downloaded files were imported using Python pandas package, and searched for unadjusted p value sets. Unadjusted p value sets and summarized expression level of associated genomic features were identified using column names. P value columns from imported tables were identified by regular expression "p[ˆa-zA-Z]{0,4}val", from these, adjusted p value sets were identified using regular expression "adj|fdr|corr|thresh" and omitted from further analysis. Columns with expression levels of genomic features were identified by using following regular expressions: "basemean", "value", "fpkm", "logcpm", "rpkm", "aveexpr". Where expression level data were present, raw p values were further filtered to remove low-expression features using following thresholds: basemean=10, logcpm=1, rpkm=1, fpkm=1, aveexpr=3.32. Basemean is a mean of library-size normalized counts of all samples, logcpm is a mean log2 counts per million, rpkm/fpkm is reads/fragments per kilobase of transcript length per million reads, aveexpr is an average expression across all samples, in log2 CPM units, whereas CPM is counts per million. Row means were calculated when there were multiple expression level columns (e.g for each contrast or sample) in table. Filtered p value sets were stored and analysed separately.

Classification of p value histograms

Raw p value sets were classified based on their histogram shape. Histogram shape was determined based on the presence and location of peaks. P value histogram peaks (bins) were detected using a quality control threshold described in (Breheny, Stromberg, and Lambert 2018), a Bonferroni-corrected alpha-level quantile of the cumulative function of the binomial distribution with size m and probability p. Histograms, where none of the bins were over QC-threshold, were classified as "uniform". Histograms, where bins over QC-threshold started either from left or right boundary and did not exceeded 1/3 of the 0 to 1 range, were classified as "anti-conservative" or "conservative", respectively. Histograms with peaks or bumps in the middle or with non-continuous left- or right-side peaks were classified as "other". Histograms with peaks on both left- and right-side were classified as "bimodal".

Calculation of π_0 statistic

Raw p value sets with anti-conservative shape were used to calculate the π_0 statistic. The π_0 statistic was calculated using local FDR method implemented in limma::PropTrueNullByLocalFDR (Ritchie, Phipson, Wu, et al. 2015b) and, independently, Storey's global FDR smoother method (Storey 2002) as implemented in gdsctools (Cokelaer et al. 2017) Python package. Differential expression analysis tools were inferred from column names pattern for cuffdiff (column name = "fpkm" and "p_value") (Trapnell et al. 2013), DESeq/DESeq2 (column name = "basemean") (Love, Huber, and Anders 2014), EdgeR (column name = "logcpm") (McCarthy, Chen, and Smyth 2012b), and limma (column name = "aveexpr") (Ritchie, Phipson, Wu, et al. 2015a), all other unidentified sets were binned as "unknown".

Publication data

Publication data were downloaded from NCBI PubMed database using PubMedId-s from GEO document summaries. Article citation data was downloaded from Elsevier Scopus database using PubMedId-s. Journal CiteScore data from years 2011-2019 (October 2020) was downloaded from Elsevier. Yearly journal CiteScore data was merged with GEO series journal publications using journal ISSN/ESSN and articles' publication year. Sequence read library metadata were downloaded from NCBI SRA database using GEO accessions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Modelling

Bayesian modelling was done using R libraries rstan vers. 2.21.2 (Stan Development Team 2020) and brms vers. 2.13.3 (Bürkner 2018). Models were specified using extended R lme4 (Bates et al. 2015) formula syntax as implemented in R brms package. We used weak priors to fit models. We run minimally 2000 iterations and four chains to fit models. When suggested by brms, Stan NUTS control parameter adapt_delta was increased to 0.95–0.99 and max_treedepth to 12–15.

RNA-seq simulation

RNA-seq experiment simulation was done with polyester R package (Frazee et al. 2015) and differential expression was assessed using DESeq2 R package (Love, Huber, and Anders 2014) using default settings. Code and workflow used to run and analyze RNA-seq simulations is deposited in Zenodo with doi: 10.5281/zenodo.4463804 (`https://doi.org/10.5281/zenodo.4463804`). Processed data, raw data and workflow with input fasta file is deposited in Zenodo with doi: 10.5281/zenodo.4463803 (`http://doi.org/10.5281/zenodo.4463803`).

Code and raw data

The code to produce raw dataset is available as a snakemake workflow (Köster and Rahmann 2012) on rstats-tartu/geo-htseq Github repo (`https://github.com/rstats-tartu/geo-htseq`). Raw dataset produced by the workflow is deposited in Zenodo `https://zenodo.org` with doi: 10.5281/zenodo.4046422 (`http://doi.org/10.5281/zenodo.4046422`). The code to produce article's figures and fit models is deposited on rstats-tartu/geo-htseq-paper Github repo (`https://github.com/rstats-tartu/geo-htseq-paper`). Article's input data, code, software required to produce all models and figures in Linux, along with fitted model objects is deposited in Zenodo with doi: 10.5281/zenodo.4469911 (`https://doi.org/10.5281/zenodo.4469911`). Individual model objects are stored on GIN repo `https://gin.g-node.org/tpall/geo-htseq-paper` (doi:). ggplot2 vers. 3.3.1 (Wickham 2016) R library was used for graphics. Data wrangling was done using tools from tidyverse package (Wickham et al. 2019). Bayesian models were converted to tidy format and visualised using tidybayes R package (Kay 2020).

Acknowledgments

We are grateful for Toomas Mets (University of Tartu) for critically reading the manuscript, and Niilo Kaldalu (University of Tartu) and Margus Pihlak (Tallinn University of Technology) for useful discussions. The work was supported by the European Union from the European Regional Development Fund through the Centre of Excellence in Molecular Cell Engineering (2014-2020.4.01.15-0013) and by the grants from the Estonian Research Council (PRG335, PUT1580).

References

Baker, M. 2016. "1,500 scientists lift the lid on reproducibility." *Nature* 533 (7604): 452–54. `https://doi.org/10.1038/533452a`.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. `https://doi.org/10.18637/jss.v067.i01`.

Begley, C. G., and L. M. Ellis. 2012. "Drug development: Raise standards for preclinical cancer research." *Nature* 483 (7391): 531–33. `https://doi.org/10.1038/483531a`.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature*, 1–7. `https://doi.org/10.1038/s41586-020-2314-9`.

Breheny, Patrick, Arnold Stromberg, and Joshua Lambert. 2018. "P-Value Histograms: Inference and Diagnostics." *High-Throughput* 7 (3): 23. `https://doi.org/10.3390/ht7030023`.

Bürkner, Paul-Christian. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. `https://doi.org/10.32614/RJ-2018-017`.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bustin, Stephen A, Vladimir Benes, Jeremy Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Reinhold Mueller, et al. 2013. "The Need for Transparency and Good Practices in the qPCR Literature." *Nature Methods* 10 (11): 1063–7.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Carp, Joshua. 2012a. "On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of Fmri Experiments." *Frontiers in Neuroscience* 6: 149. https://doi.org/10.3389/fnins.2012.00149.

———. 2012b. "The Secret Lives of Experiments: Methods Reporting in the fMRI Literature." *Neuroimage* 63 (1): 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004.

Chen, Kaifu, Zheng Hu, Zheng Xia, Dongyu Zhao, Wei Li, and Jessica K Tyler. 2016. "The Overlooked Fact: Fundamental Need for Spike-in Control for Virtually All Genome-Wide Analyses." *Molecular and Cellular Biology* 36 (5): 662–67. https://doi.org/10.1128/MCB.00970-14.

Cokelaer, Thomas, Elisabeth Chen, Francesco Iorio, Michael P Menden, Howard Lightfoot, Julio Saez-Rodriguez, and Mathew J Garnett. 2017. "GDSCTools for mining pharmacogenomic interactions in cancer." *Bioinformatics* 34 (7): 1226–8. https://doi.org/10.1093/bioinformatics/btx744.

Dialsingh, Isaac, Stefanie R Austin, and Naomi S Altman. 2015. "Estimating the Proportion of True Null Hypotheses When the Statistics Are Discrete." *Bioinformatics* 31 (14): 2303–9.

Everaert, Celine, Manuel Luypaert, Jesper LV Maag, Quek Xiu Cheng, Marcel E Dinger, Jan Hellemans, and Pieter Mestdagh. 2017. "Benchmarking of Rna-Sequencing Analysis Workflows Using Whole-Transcriptome Rt-qPCR Expression Data." *Scientific Reports* 7 (1): 1–11. https://doi.org/10.1038/s41598-017-01617-3.

Frazee, Alyssa C., Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2015. "Polyester: simulating RNA-seq datasets with differential transcript expression." *Bioinformatics* 31 (17): 2778–84. https://doi.org/10.1093/bioinformatics/btv272.

Functional Genomics Data Society. 2008. "A draft proposal for the required Minimum Information about a high- throughput Nucleotide SeQuencing Experiment – MINSEQE." http://fged.org/site_media/pdf/MINSEQE_draft_2008.pdf.

———. 2012.

Goodman, S. N., D. Fanelli, and J. P. Ioannidis. 2016. "What does research reproducibility mean?" *Sci Transl Med* 8 (341): 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027.

Grimes, D. R., C. T. Bauch, and J. P. A. Ioannidis. 2018. "Modelling science trustworthiness under publish or perish pressure." *R Soc Open Sci* 5 (1): 171511. https://doi.org/10.1098/rsos.171511.

Hardwicke, Tom E, Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N Goodman, and John PA Ioannidis. 2020. "Calibrating the Scientific Ecosystem Through Meta-Research." *Annual Review of Statistics and Its Application* 7: 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104.

Harris, Richard. 2017. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*. Basic Books.

Hu, Zheng, Kaifu Chen, Zheng Xia, Myrriah Chavez, Sangita Pal, Ja-Hwan Seol, Chin-Chuan Chen, Wei Li, and Jessica K Tyler. 2014. "Nucleosome Loss Leads to Global Transcriptional up-Regulation and Genomic Instability During Yeast Aging." *Genes & Development* 28 (4): 396–408. https://doi.org/10.1101/gad.233221.113.

Ioannidis, John PA. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Jager, Leah R, and Jeffrey T Leek. 2014a. "An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature." *Biostatistics* 15 (1): 1–12.

———. 2014b. "Rejoinder: An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature." *Biostatistics* 15 (1): 39–45.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake—a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28 (19): 2520–2. https://doi.org/10.1093/bioinformatics/bts480.

Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for Rna-Seq Read Counts." *Genome Biology* 15 (2): 1–17.

Leek, Jeffrey T, and Leah R Jager. 2017. "Is Most Published Research Really False?" *Annual Review of Statistics and Its Application* 4: 109–22.

Leng, Gareth, and Rhodri Ivor Leng. 2020. *The Matter of Facts: Skepticism, Persuasion, and Evidence in Science*. Mit Press.

Lin, Charles Y, Jakob Lovén, Peter B Rahl, Ronald M Paranal, Christopher B Burge, James E Bradner, Tong Ihn Lee, and Richard A Young. 2012. "Transcriptional Amplification in Tumor Cells with Elevated c-Myc." *Cell* 151 (1): 56–67. https://doi.org/10.1016/j.cell.2012.08.026.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lovén, Jakob, David A Orlando, Alla A Sigova, Charles Y Lin, Peter B Rahl, Christopher B Burge, David L Levens, Tong Ihn Lee, and Richard A Young. 2012. "Revisiting Global Gene Expression Analysis." *Cell* 151 (3): 476–82. https://doi.org/10.1016/j.cell.2012.10.012.

Lun, Aaron TL, Fernando J Calero-Nieto, Liora Haim-Vilmovsky, Berthold Göttgens, and John C Marioni. 2017. "Assessing the Reliability of Spike-in Normalization for Analyses of Single-Cell Rna Sequencing Data." *Genome Research* 27 (11): 1795–1806. https://doi.org/10.1101/gr.222877.117.

Maiväli, Ülo. 2015. *Interpreting Biomedical Science: Experiment, Evidence, and Belief*. Academic Press.

Mandelboum, Shir, Zohar Manber, Orna Elroy-Stein, and Ran Elkon. 2019. "Recurrent Functional Misinterpretation of Rna-Seq Data Caused by Sample-Specific Gene Length Bias." *PLoS Biology* 17 (11): e3000481. https://doi.org/10.1371/journal.pbio.3000481.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012a. "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97. https://doi.org/10.1093/nar/gks042.

———. 2012b. "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97. https://doi.org/10.1093/nar/gks042.

McGee, Warren A, Harold Pimentel, Lior Pachter, and Jane Y Wu. 2019. "Compositional Data Analysis Is Necessary for Simulating and Analyzing Rna-Seq Data." *bioRxiv*. https://doi.org/10.1101/564955.

Munafò, Marcus R, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1): 1–9. https://doi.org/10.1038/s41562-016-0021.

Nie, Zuqin, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, et al. 2012. "C-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells." *Cell* 151 (1): 68–79. https://doi.org/10.1016/j.cell.2012.08.033.

Nookaew, Intawat, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. 2012. "A Comprehensive Comparison of Rna-Seq-Based Transcriptome Analysis from Reads to Differential Gene Expression and Cross-Comparison with Microarrays: A Case Study in Saccharomyces Cerevisiae." *Nucleic Acids Research* 40 (20): 10084–97. https://doi.org/10.1093/nar/gks804.

Okagbue, Hilary I, and Jaime A Teixeira da Silva. 2020. "Correlation Between the Citescore and Journal Impact Factor of Top-Ranked Library and Information Science Journals." *Scientometrics* 124 (1): 797–801.

Patil, Prasad, Roger D Peng, and Jeffrey T Leek. 2016. "What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science." *Perspectives on Psychological Science* 11 (4): 539–44.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science* 334 (6060): 1226–7. https://doi.org/10.1126/science.1213847.

Prinz, Florian, Thomas Schlange, and Khusru Asadullah. 2011. "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?" *Nature Reviews Drug Discovery* 10 (9): 712–12. https://doi.org/10.1038/nrd3439-c1.

Quinn, Thomas P, Ionas Erb, Greg Gloor, Cedric Notredame, Mark F Richardson, and Tamsyn M Crowley. 2019. "A Field Guide for the Compositional Analysis of Any-Omics Data." *GigaScience* 8 (9): giz107. https://doi.org/10.1093/gigascience/giz107.

Risso, Davide, John Ngai, Terence P Speed, and Sandrine Dudoit. 2014. "The Role of Spike-in Standards in the Normalization of Rna-Seq." In *Statistical Analysis of Next Generation Sequencing Data*, 169–90. Springer.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015a. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Ritchie, Matthew E, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015b. "Limma Powers Differential Expression Analyses for Rna-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47–e47. https://doi.org/10.1093/nar/gkv007.

Schuemie, Martijn J, Patrick B Ryan, William DuMouchel, Marc A Suchard, and David Madigan. 2014. "Interpreting Observational Studies: Why Empirical Calibration Is Needed to Correct P-Values." *Statistics in Medicine* 33 (2): 209–18.

Stan Development Team. 2020. "RStan: The R Interface to Stan." http://mc-stan.org/.

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3): 479–98. https://doi.org/10.1111/1467-9868.00346.

Sun, Zhaonan, and Yu Zhu. 2012. "Systematic Comparison of Rna-Seq Normalization Methods Using Measurement Error Models." *Bioinformatics* 28 (20): 2584–91. https://doi.org/10.1093/bioinformatics/bts497.

Szucs, Denes, and John Ioannidis. 2017. "When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment." *Frontiers in Human Neuroscience* 11: 390. https://doi.org/10.3389/fnhum.2017.00390.

Trapnell, C., D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. 2013. "Differential analysis of gene regulation at transcript resolution with RNA-seq." *Nat Biotechnol* 31 (1): 46–53. https://doi.org/10.1038/nbt.2450.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of Rna-Seq Experiments with Tophat and Cufflinks." *Nature Protocols* 7 (3): 562.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zijlstra, Hans, and Rachel McCullough. 2016. "CiteScore: A New Metric to Help You Track Journal Performance and Make Decisions." *Elsevier*, 8.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure supplements

Figure 2 figure supplement 1

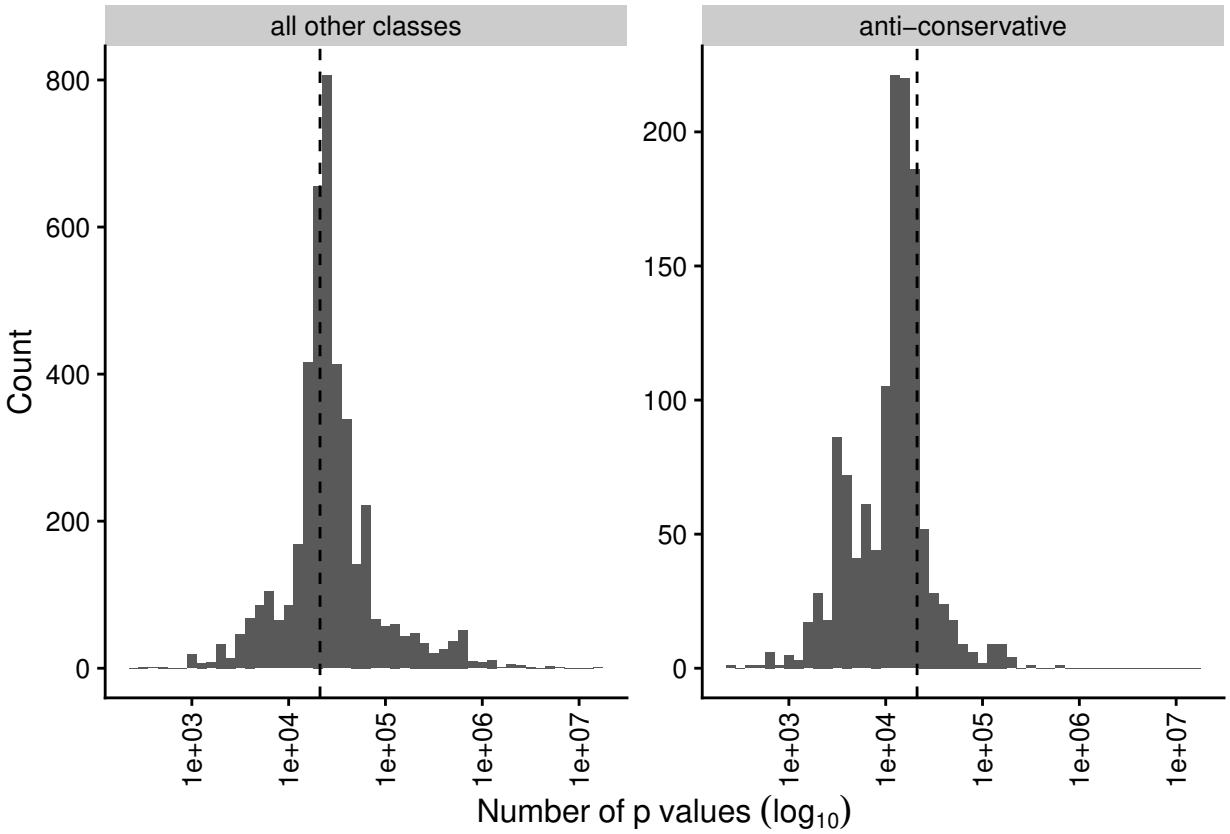


Figure 2–figure supplement 1. P value set size distribution. Dashed line denotes the median (21084) number of features. From each GEO series only one of each unique length was considered, N=5469 p value sets.

Figure 2 figure supplement 2

Frazee, Alyssa C., Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2015. "Polyester: simulating RNA-seq datasets with differential transcript expression." *Bioinformatics* 31 (17): 2778–84. https://doi.org/10.1093/bioinformatics/btv272.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

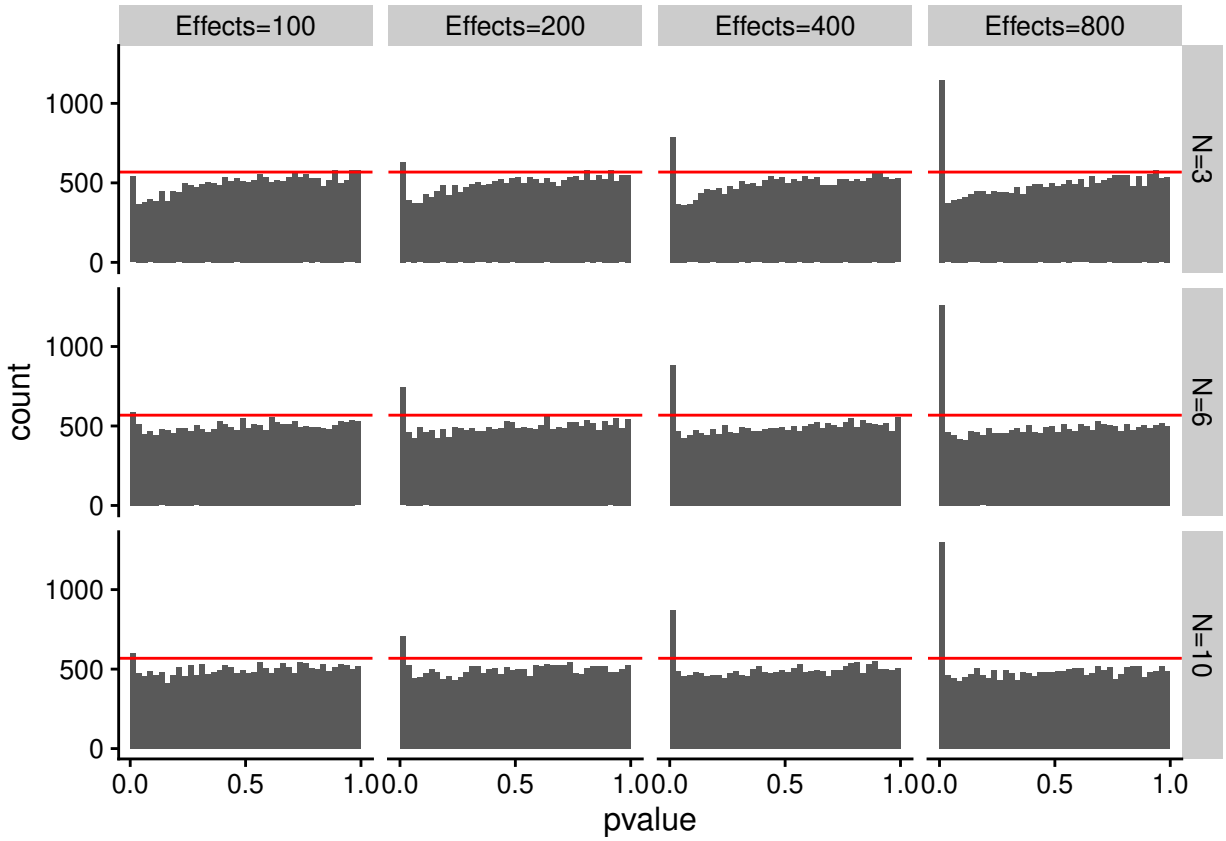


Figure 2–figure supplement 2. Simulated RNA-seq data shows that histograms from p value sets with around one hundred true effects out of 20,000 features can be classified as "uniform". RNA-seq data was simulated with polyester R package (Frazee et al. 2015) on 20,000 transcripts from human transcriptome using grid of 3, 6, and 10 replicates and 100, 200, 400, and 800 effects for two groups. Fold changes were set to 0.5 and 2. Differential expression was assessed using DESeq2 R package (Love, Huber, and Anders 2014) using default settings and group 1 versus group 2 contrast. Effects denotes in facet labels the number of true effects and N denotes number of replicates. Red line denotes QC threshold used for dividing p histograms into discrete classes. Code and workflow used to run these simulations is available on Github: https://github.com/rstats-tartu/simulate-rnaseq. Raw data of the figure is available on Zenodo https://zenodo.org with doi: 10.5281/zenodo.4463803.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

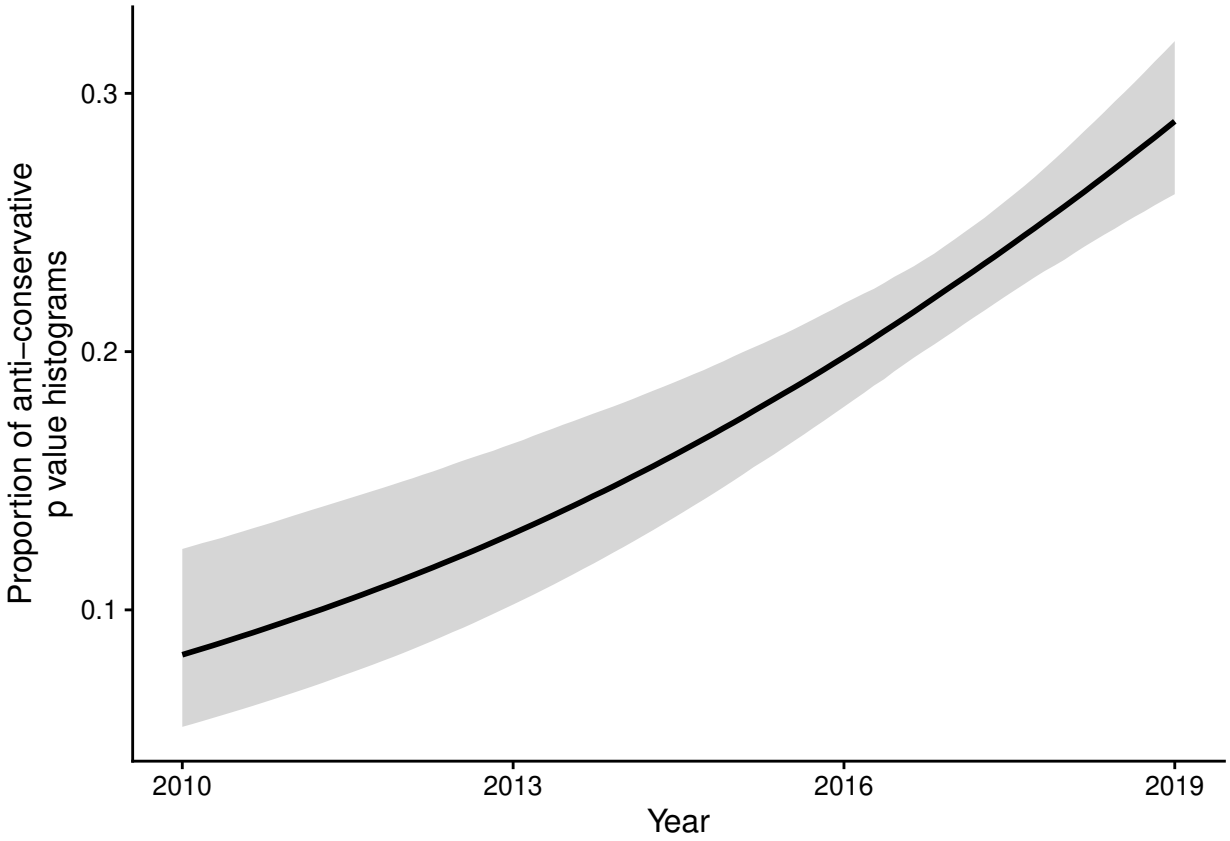


Figure 3–figure supplement 1. The increasing proportion of anti-conservative histograms. Binomial logistic model: *anticons* ~ *year*, N = 2,109. Download model object anticons_year.rds. Lines denote best fit of linear model. Shaded area denotes 95% credible region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

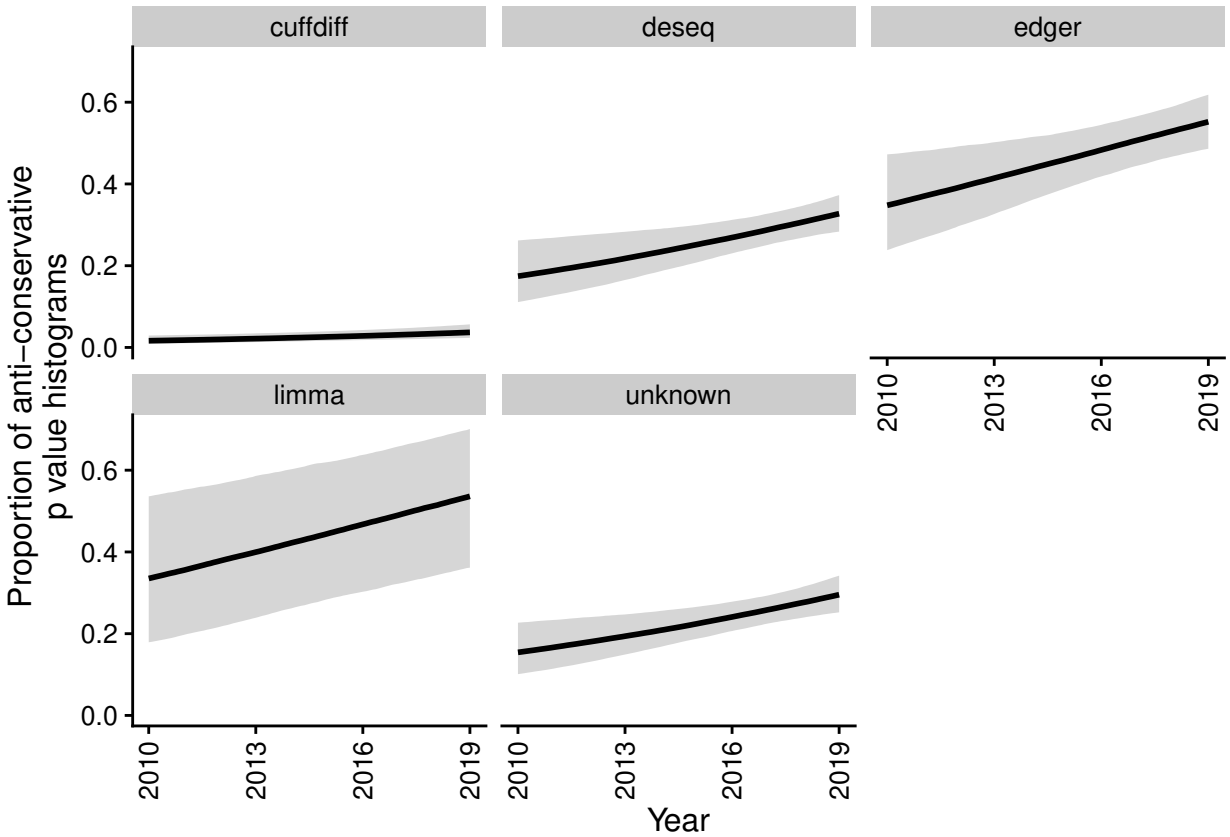


Figure 3–figure supplement 2. A 2-level binomial logistic model *anticons ~ year + (year | de_tool)* reveals that all differential expression analysis tools are associated with temporally increasing anti-conservative p value histograms, N = 2,109. Download model object anticons_year___year_detool.rds. Lines denote best fit of linear model. Shaded area denotes 95% credible region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

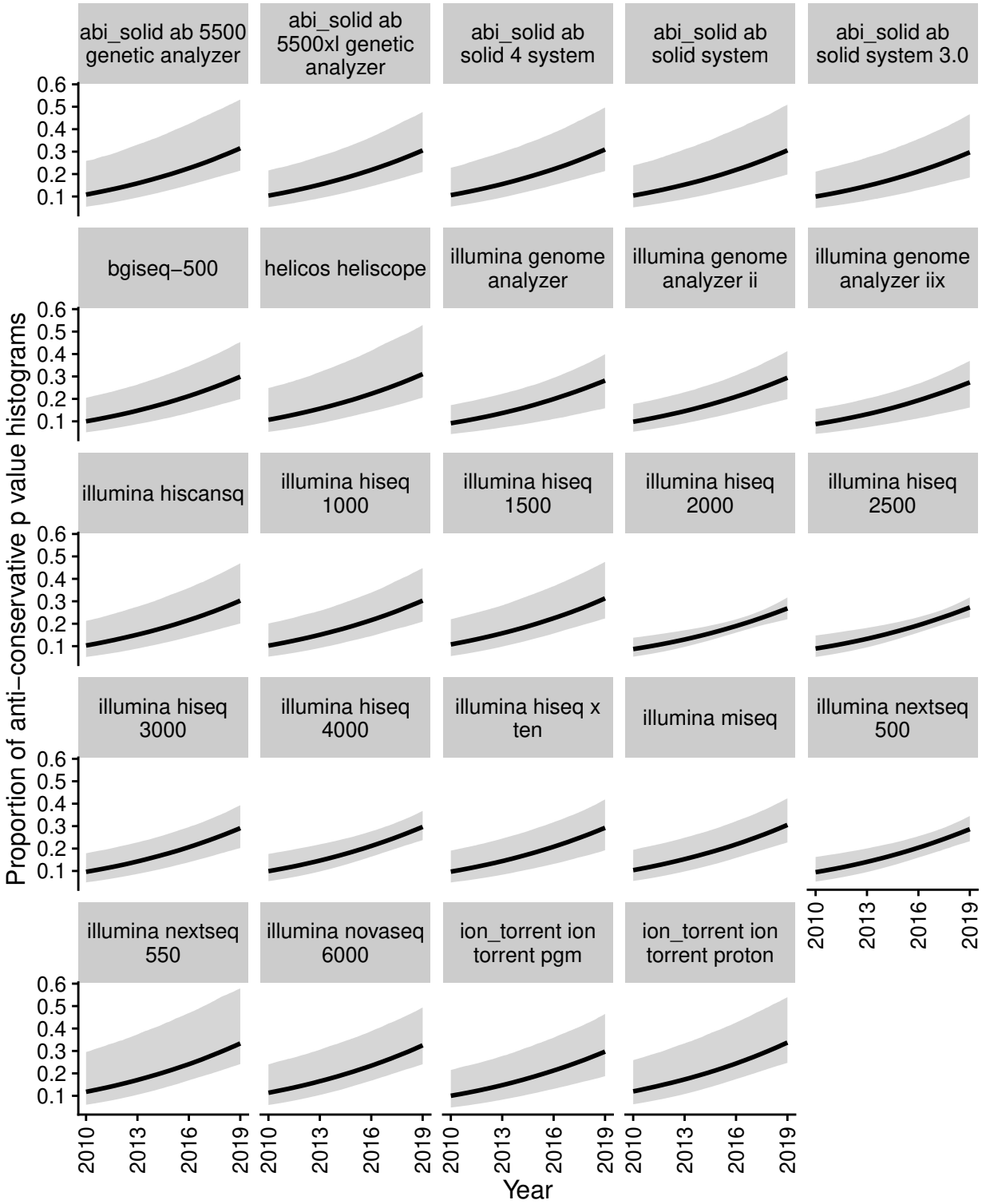


Figure 3–figure supplement 3. A 2-level binomial logistic model *anticons ~ year + (year | model)* reveals that all sequencing instrument models are associated with temporally increasing anti-conservative p value histograms, N = 1,718. Download model object anticons_year___year_model.rds. Only GEO submissions utilizing single sequencing platform were used for model fitting. Lines denote best fit of linear model. Shaded area denotes 95% credible region.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

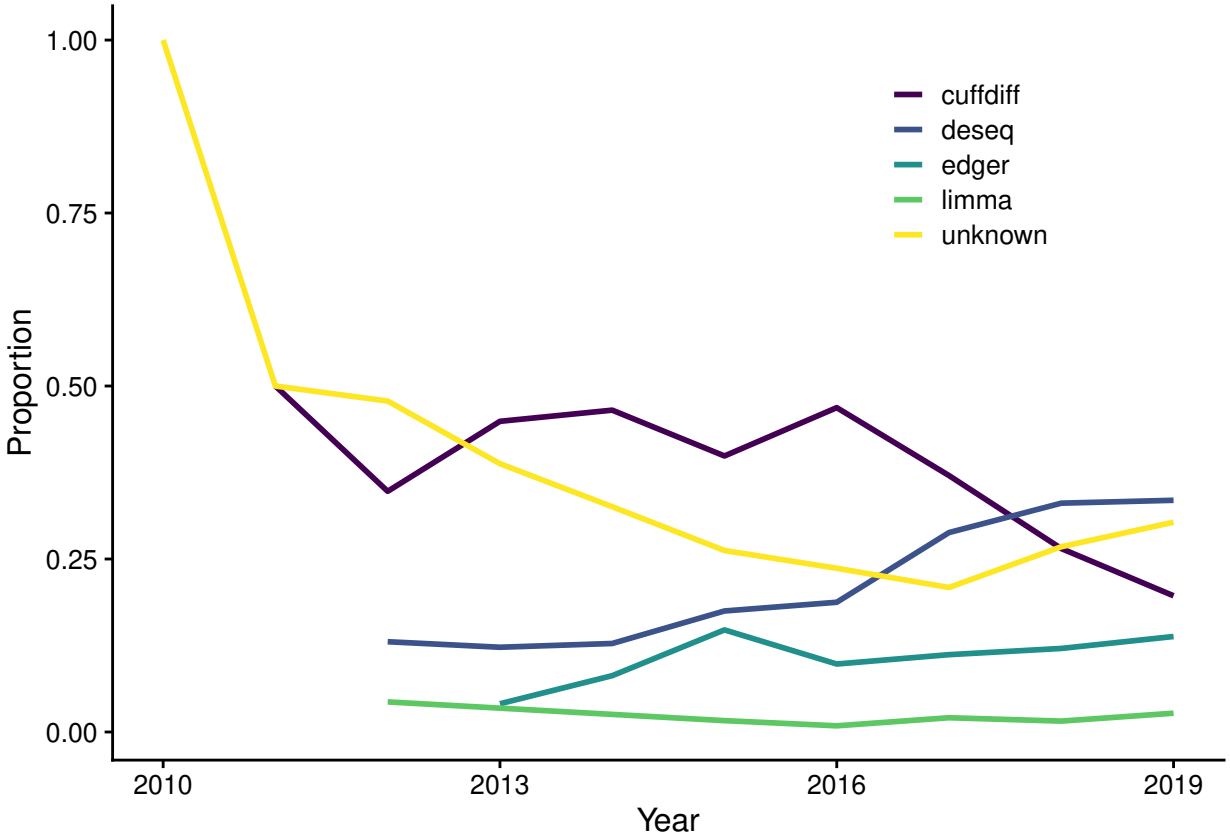


Figure 3–figure supplement 4. No single differential expression analysis tool dominates the field. Y-axis shows the proportion of analysis platforms, x-axis shows publication year of GEO submission, N = 1,733.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

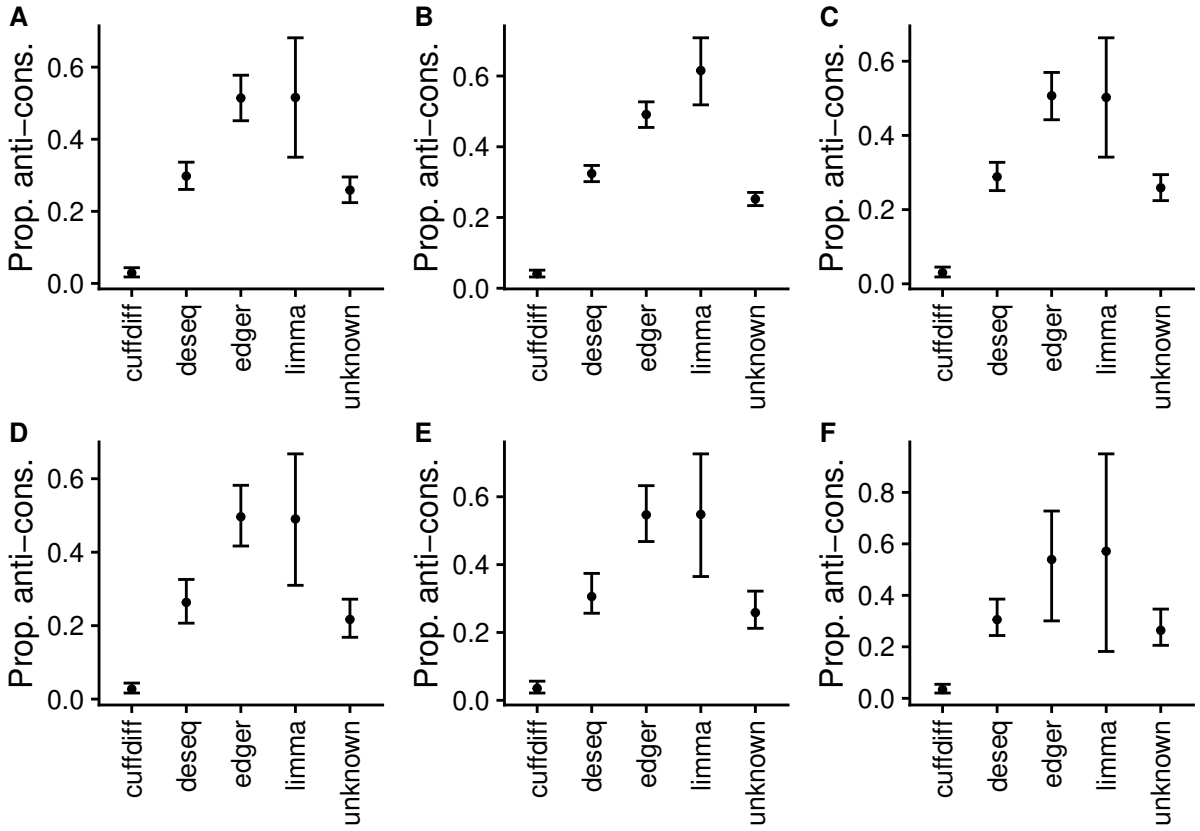


Figure 3–figure supplement 5. Binomial logistic models for proportion of anti-conservative p value histograms. A, simple model *anticons* ~ *de_tool*, N = 2,109. B, simple model *anticons* ~ *de_tool* fitted on complete data, N = 6,267. Download model object anticons_detool_all.rds. C, model conditioned on year of GEO submission: *anticons* ~ *year* + *de_tool*, N = 2,109. Download model object anticons_year_detool.rds. D, model conditioned on studied organism (human/mouse/other): *anticons* ~ *organism* + *de_tool*, N = 1,733. Download model object anticons_organism_detool.rds. E, varying intercept model *anticons* ~ *de_tool* + *(1 | model)* where "model" stands for sequencing instrument model, N = 1,718. Download model object anticons_detool___1_model.rds. F, varying intercept and slope model *anticons* ~ *de_tool* + *(de_tool | model)*, N = 1,718. Download model object anticons_detool___detool_model.rds. Points denote best fit of linear model. Error bars, 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

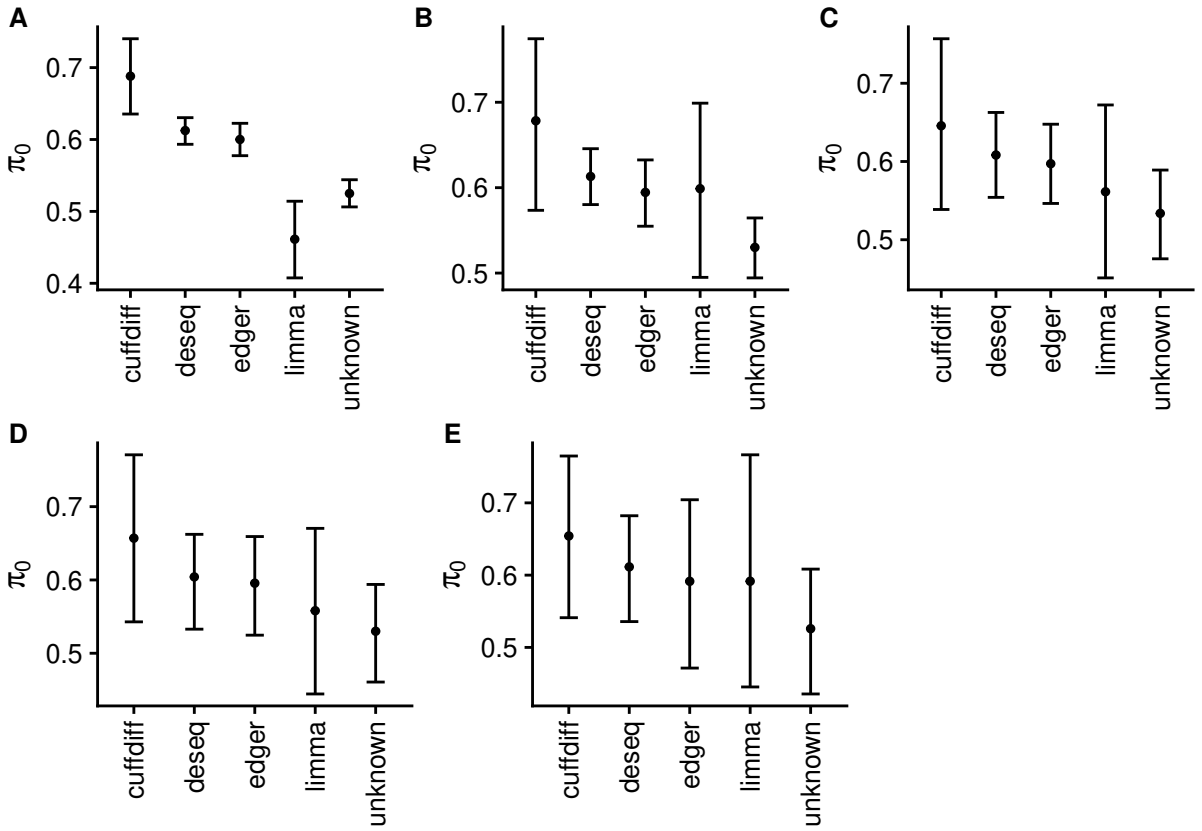


Figure 4–figure supplement 1. Robust (student's t likelihood) modeling of π_0. A, simple model *pi0 ~ de_tool* fitted on complete data, N = 1,567. Download model object pi0_detool_full_data.rds. B, model conditioned on year of GEO submission: *pi0 ~ year + de_tool*, N = 488. Download model object pi0_year_detool.rds. C, model conditioned on studied organism (human/mouse/other): *pi0 ~ organism + de_tool*, N = 400. Download model object pi0_organism_detool.rds. D, varying intercept model *pi0 ~ de_tool + (1 | model)* where 'model' stands for sequencing instrument model, N = 396. Download model object pi0_detool___1_model.rds. E, varying intercept/slope model *pi0 ~ de_tool + (de_tool | model)*, N = 396. Download model object pi0_detool___detool_model.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

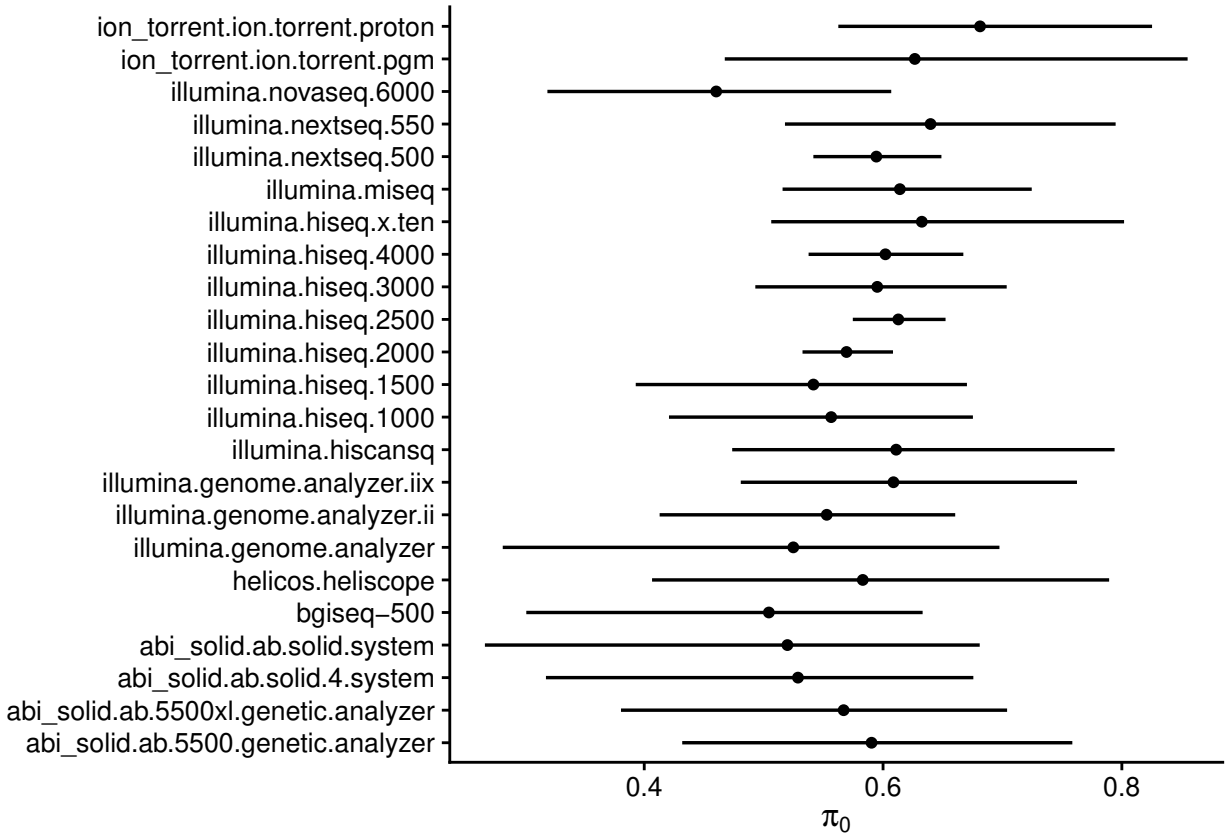


Figure 4–figure supplement 2. Modeling dependency of π_0 on sequencing instrument model: *pi0 ~ (1 | model)*, N = 396. Download model object pi0___1_model.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

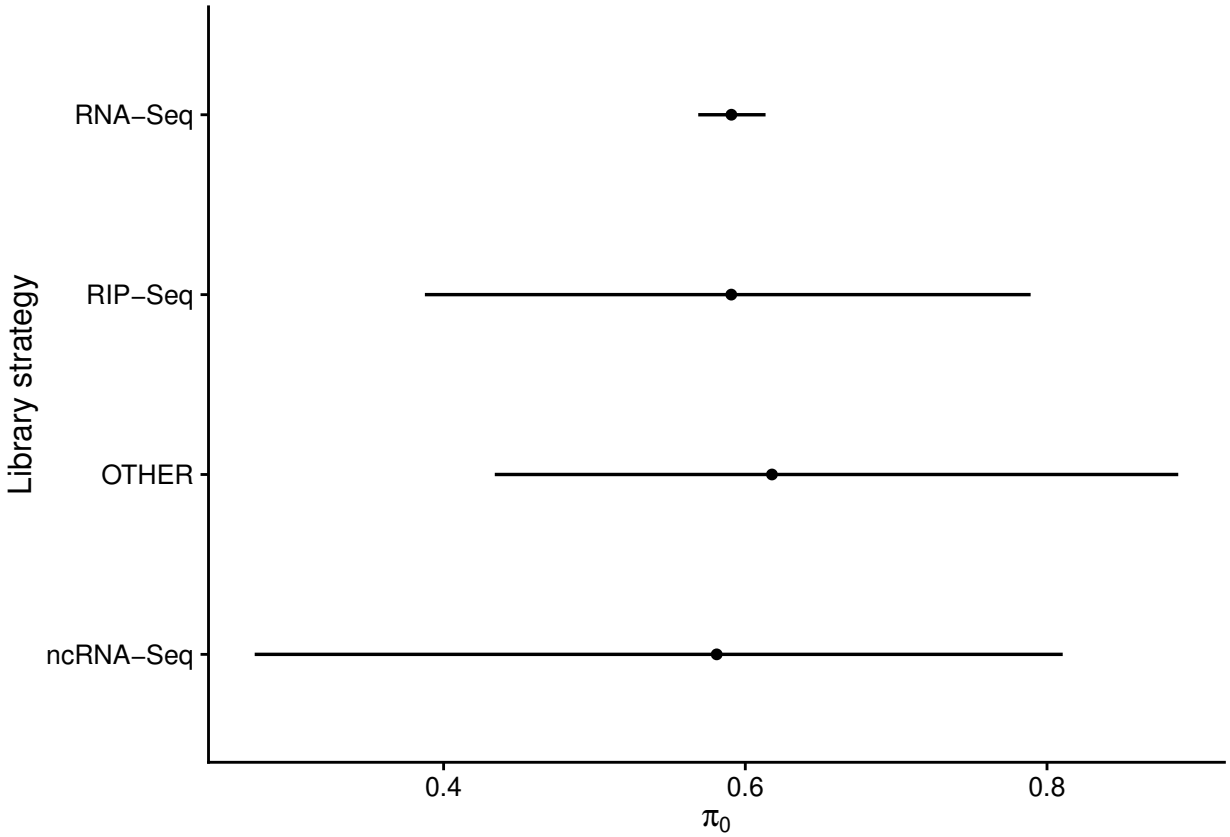


Figure 4–figure supplement 3. Modeling dependency of π_0 on library strategy: *pi0 ~ (1 | library_strategy)*, N = 396. Download model object pi0___1_librarystrategy.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

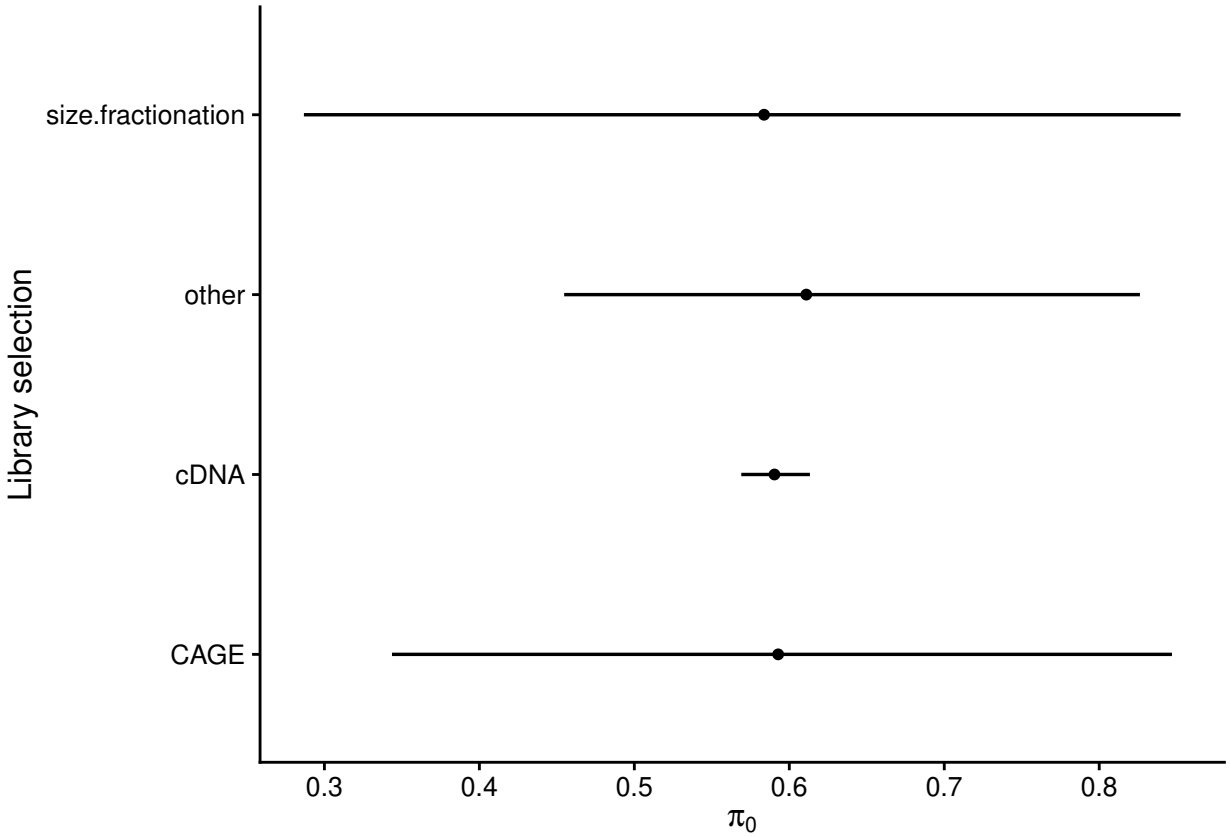


Figure 4–figure supplement 4. Modeling dependency of π_0 on library selection: *pi0 ~ (1 | library_selection)*, N = 396. Download model object pi0___1_libraryselection.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

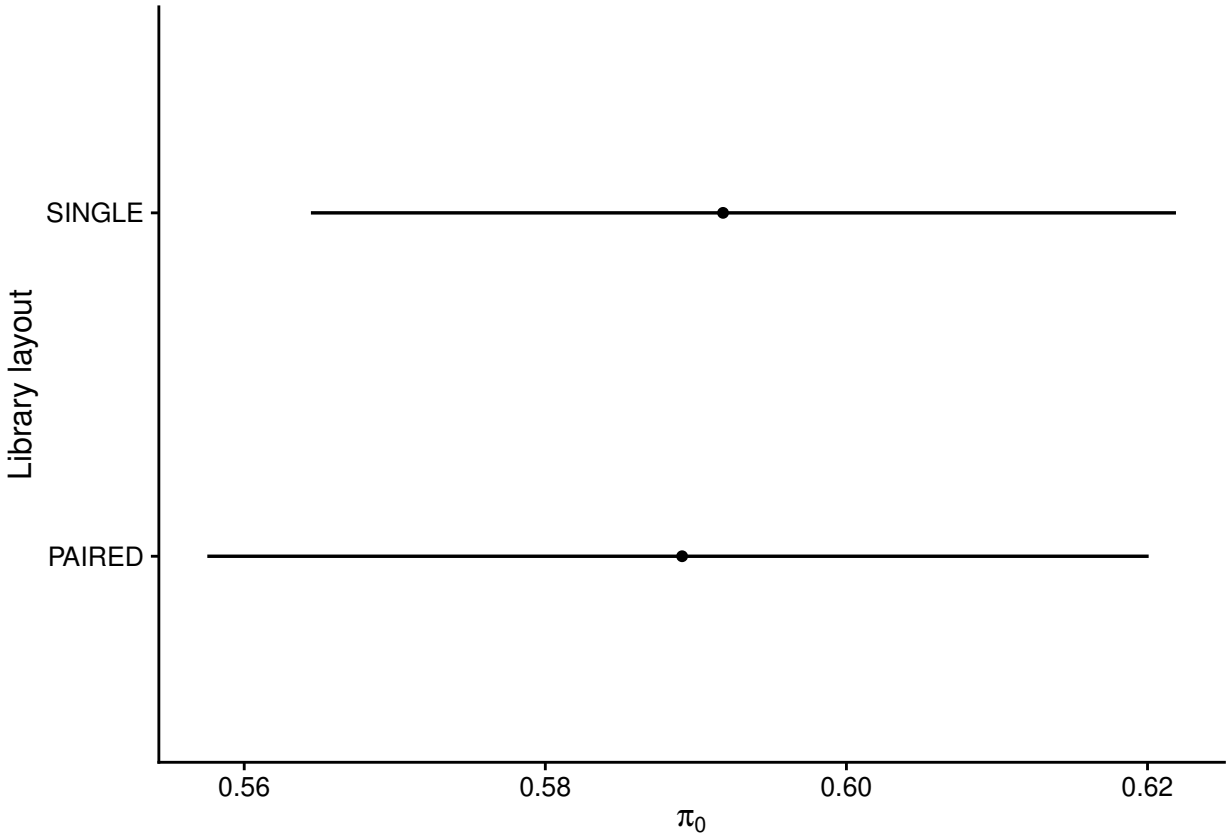


Figure 4–figure supplement 5. Modeling dependency of π_0 on library layout: *pi0 ~ (1 | library_layout)*, N = 396. Download model object pi0___1_librarylayout.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

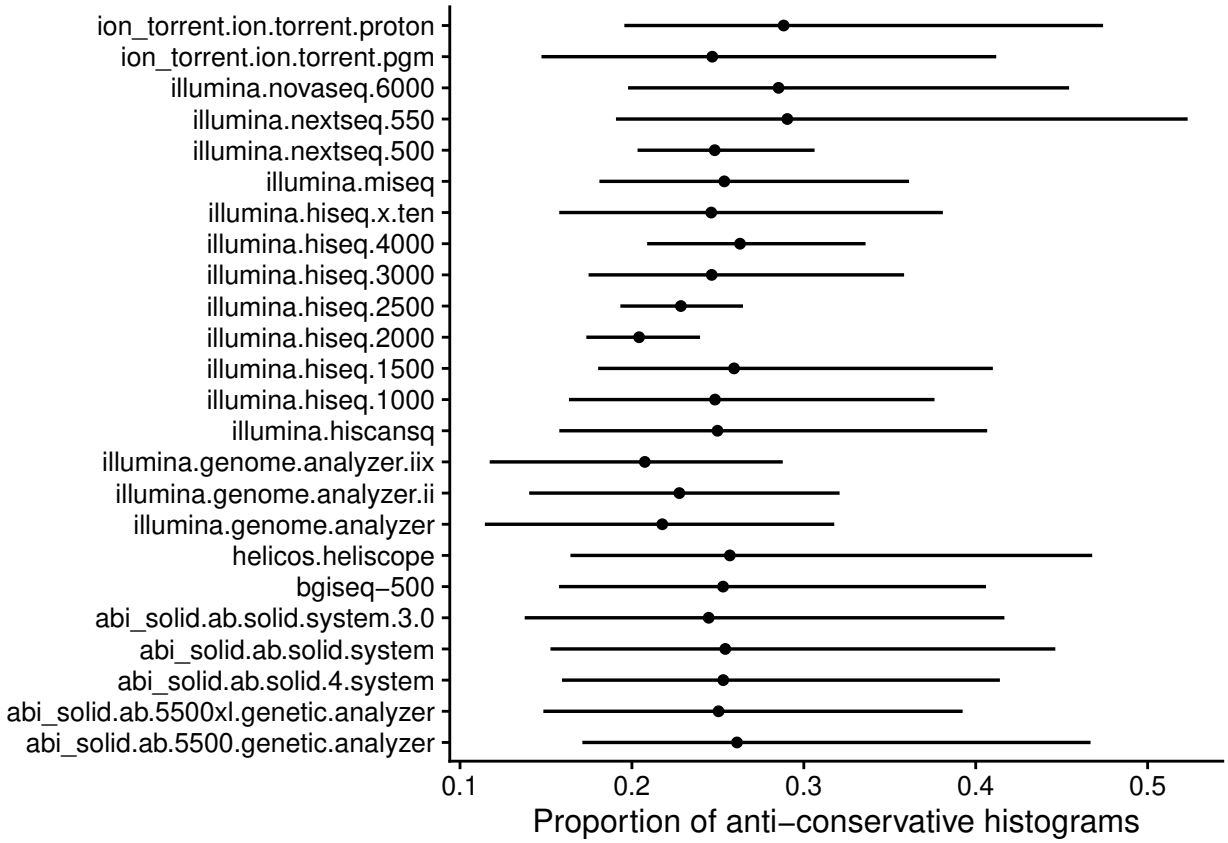


Figure 3–figure supplement 6. Modeling dependency of proportion of anti-conservative histograms on sequencing platform: *anticons ~ (1 | model)*, N = 1,718. Download model object anticons__1_model.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

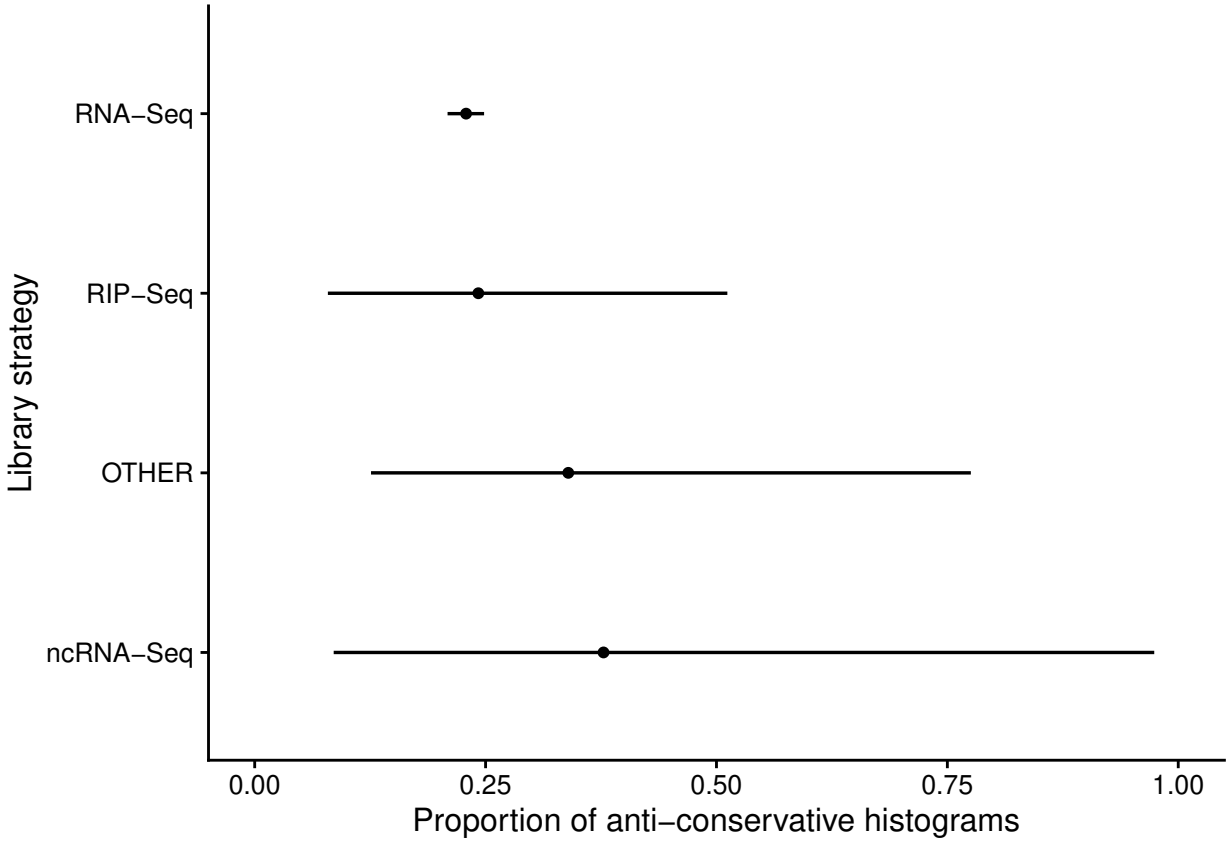


Figure 3–figure supplement 7. Modeling dependency of proportion of anti-conservative histograms on library strategy: *anticons ~ (1 | library_strategy)*, N = 1,718. Download model object anticons___1__librarystrategy.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

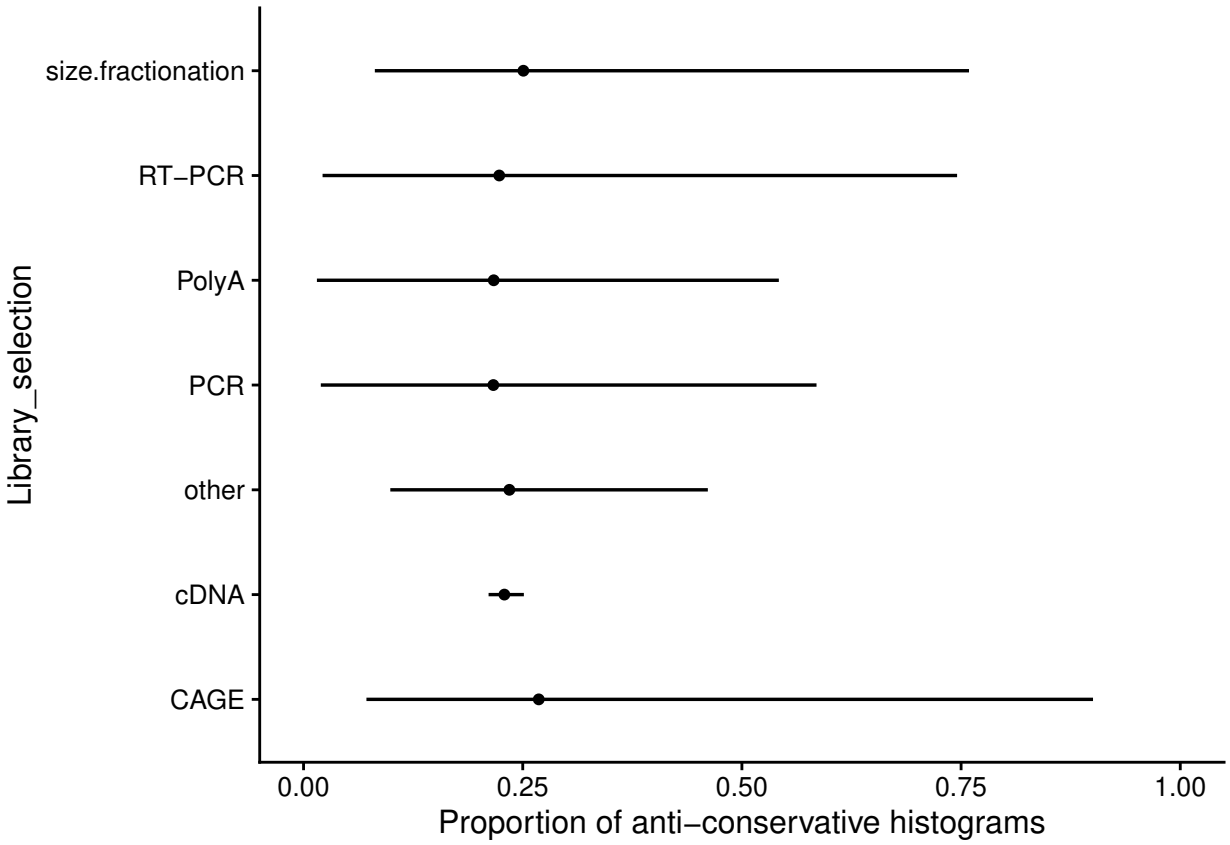


Figure 3–figure supplement 8. Modeling dependency of proportion of anti-conservative histograms on library selection: *anticons ~ (1 | library_selection)*, N = 1,718. Download model object anticons___1__libraryselection.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

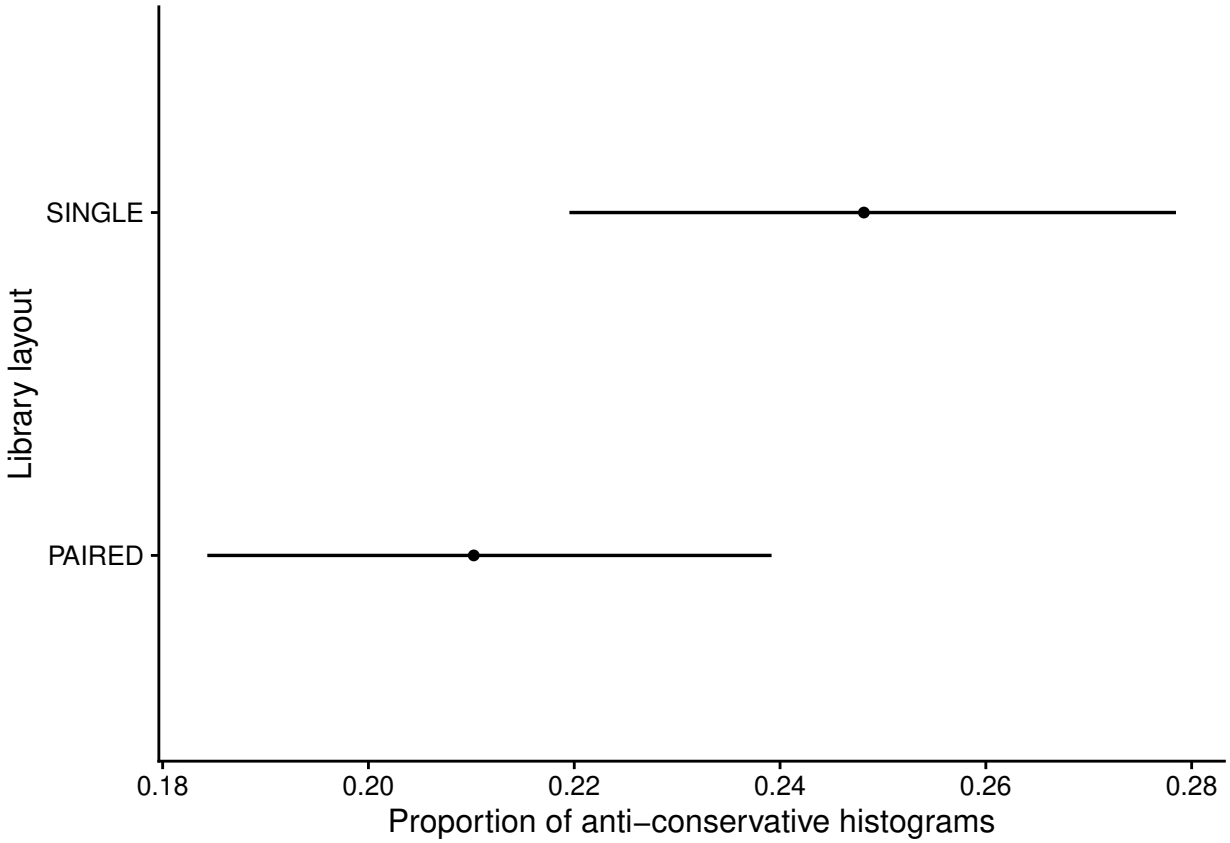


Figure 3–figure supplement 9. Modeling dependency of proportion of anti-conservative histograms on library layout: *anticons ~ (1 | library_layout)*, N = 1,718. Download model object anticons___1_librarylayout.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

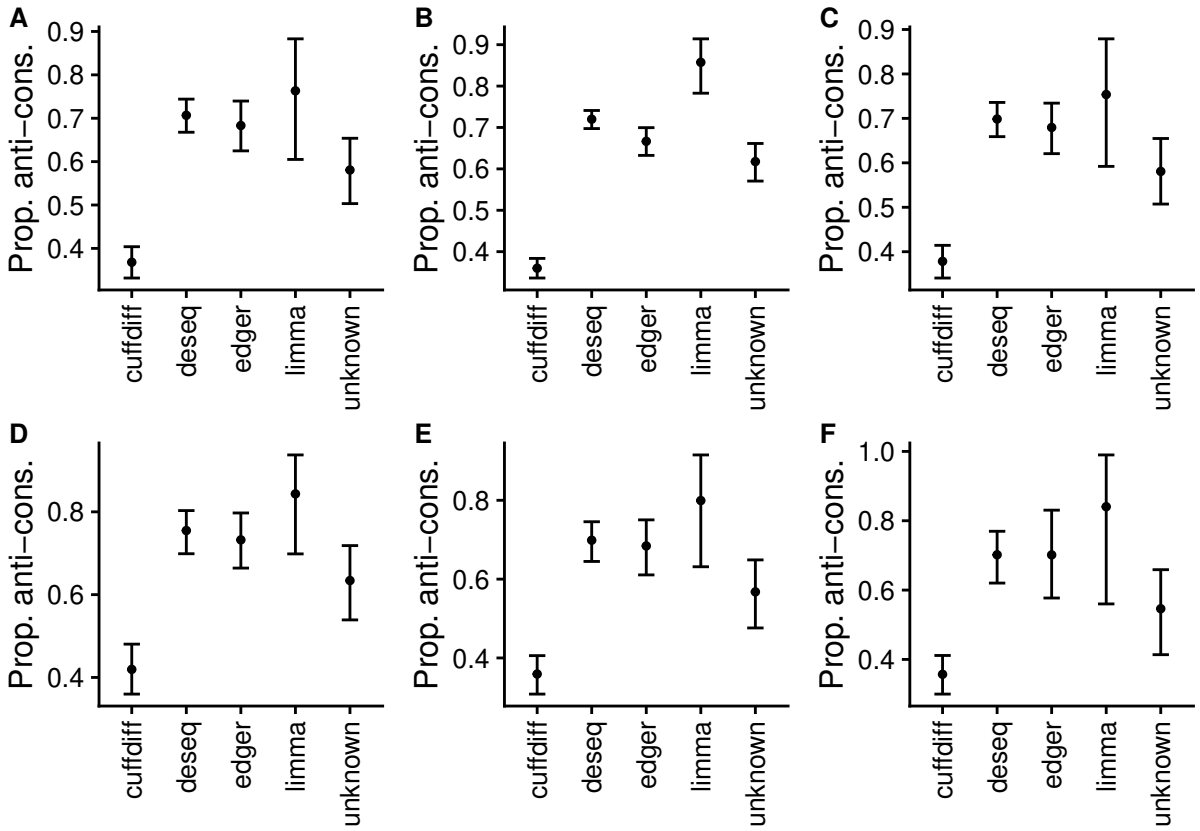


Figure 5–figure supplement 1. Binomial logistic models for proportion of anti-conservative p value histograms. A, simple model *anticons* ~ *de_tool*, N = 1,720. Download model object anticons_detool_filtered.rds. B, simple model *anticons* ~ *de_tool* fitted on complete data, N = 4,632. Download model object anticons_detool_all_filtered.rds. C, model conditioned on year of GEO submission: *anticons* ~ *year* + *de_tool*, N = 1,720. Download model object anticons_year_detool_filtered.rds. D, model conditioned on studied organism (human/mouse/other): *anticons* ~ *organism* + *de_tool*, N = 1,425. Download model object anticons_organism_detool_filtered.rds. E, varying intercept model *anticons* ~ *de_tool* + *(1 | model)* where "model" stands for sequencing instrument model, N = 1,418. Download model object anticons_detool___1_model_filtered.rds. F, varying intercept and slope model *anticons* ~ *de_tool* + *(de_tool | model)*, N = 1,418. Download model object anticons_detool___detool_model_filtered.rds. Points denote best fit of linear model. Error bars, 95% credible interval.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.04.424681; this version posted May 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

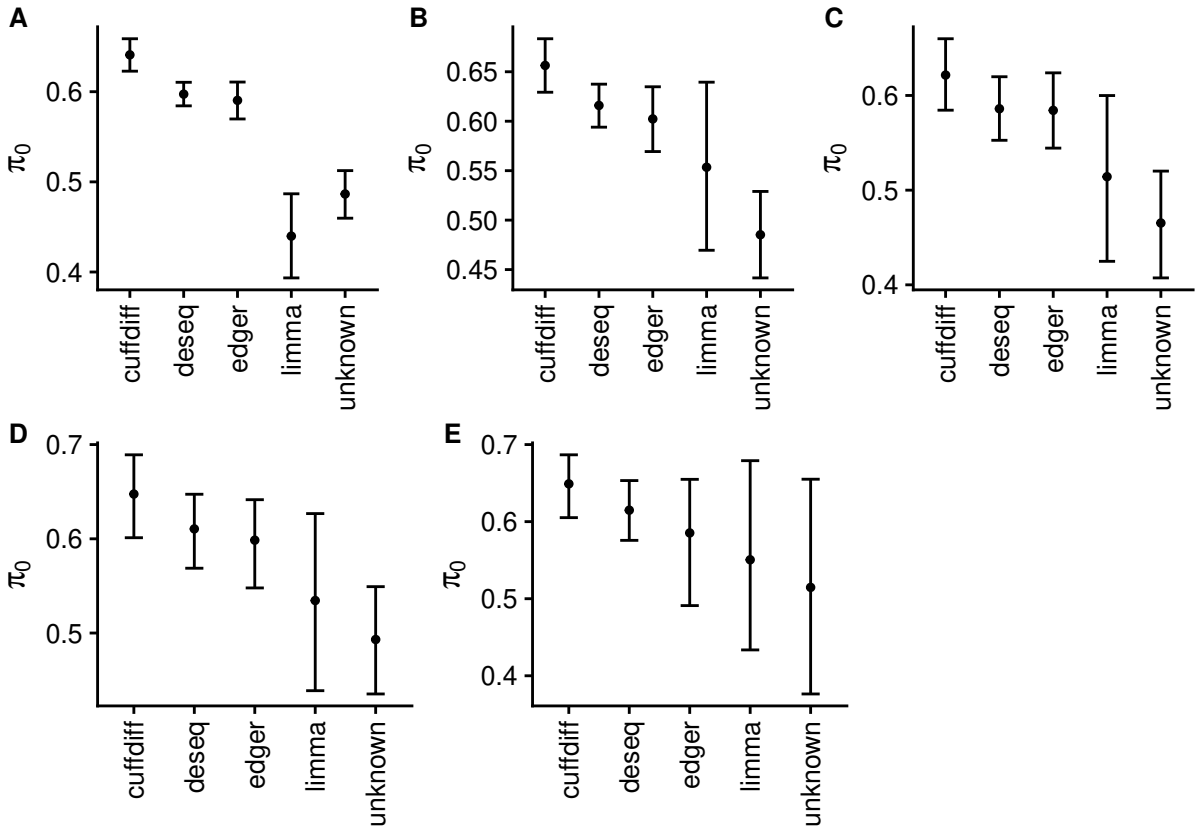


Figure 5–figure supplement 2. Robust (student's t likelihood) modeling of π_0. A, simple model *pi0 ~ de_tool* fitted on complete data, N = 2,682. Download model object pi0_detool_full_data_filtered.rds. B, model conditioned on year of GEO submission: *pi0 ~ year + de_tool*, N = 964. Download model object pi0_year_detool_filtered.rds. C, model conditioned on studied organism (human/mouse/other): *pi0 ~ organism + de_tool*, N = 791. Download model object pi0_organism_detool_filtered.rds. D, varying intercept model *pi0 ~ de_tool + (1 | model)* where 'model' stands for sequencing instrument model, N = 788. Download model object pi0_detool___1_model_filtered.rds. E, varying intercept/slope model *pi0 ~ de_tool + (de_tool | model)*, N = 788. Download model object pi0_detool___detool_model_filtered.rds. Points denote best fit of linear model. Error bars denote 95% credible interval.