Cell Systems

Transparent Peer Review Record

Reproducibility metrics for context-specific CRISPR screens

Maximilian Billmann, Henry N. Ward, Michael Aregger, Michael Costanzo, Brenda J. Andrews, Charles Boone, Jason Moffat, Chad L. Myers

Summary

Initial Submission:	Received February 21, 2022 Scientific editor: Ernesto Andrianantoandro, Ph.D.
First round of review:	Number of reviewers: Three *Three confidential, Zero signed* Revision invited April 26, 2022 *Major changes anticipated* Revision received August 17, 2022
Second round of review:	Number of reviewers: Three *Three original, Zero new* *Three confidential, Zero signed* Revision invited September 12, 2022 *Minor changes anticipated* Revision received December 20, 2022

Third round of review:

Number of reviewers: One

One original, Zero new
One confidential, Zero signed

Accepted April 07, 2023

This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Billmann,

I'm enclosing the comments that reviewers made on your paper, which I hope you will find useful and constructive. As you'll see, they express interest in the study, but they also have a number of criticisms and suggestions. Based on these comments, it seems premature to proceed with the paper in its current form; however, if it's possible to address the concerns raised with additional experiments and/or analysis, we'd be interested in considering a revised version of the manuscript. ***We appreciate that the COVID-19 pandemic challenges and limits what you and your lab can do, so to make sure we're absolutely on the same page about the feasibility of revisions, let's schedule a Zoom call at our earliest mutual convenience.***

As a matter of principle, I usually only invite a revision when I'm reasonably certain that the authors' work will align with the reviewers' concerns and produce a publishable manuscript. In the case of this manuscript, the reviewers and I have make-or-break concerns that can be addressed through:

1. Better articulating the conceptual advance and distinguishing the use of WBC from your previously published analyses and competing analyses.
2. Solidifying the demonstration of utility of WBC.

To move beyond being a commentary on previous analyses, the paper needs to convincingly address reviewer concerns about lack of an advance. To help support this, the utility of the WBC metric needs to be better substantiated though clarifying its role in the DepMap analysis and/or applying it other datasets. In addition to the concerns I've detailed above, I've highlighted portions of the reviews that strike me as particularly critical I'd also like to be explicitly clear about an almost philosophical stance that we take at Cell Systems…

We believe that understanding how approaches fail is fundamentally interesting: it provides critical insight into understanding how they work. We also believe that all approaches do fail and that it's unreasonable, even misleading, to expect otherwise. Accordingly, when papers are transparent and forthright about the limitations and crucial contingencies of their approaches, we consider that to be a great strength, not a weakness. Please keep this in mind when addressing concerns about the novelty of the approach.

As you address these concerns, it's important that you and I stay on the same page. I'm always happy to talk, either over email or by Zoom, if you'd like feedback about whether your efforts are moving the manuscript in a productive direction. Do note that we generally consider papers through only one major round of revision, so the revised manuscript would be either accepted or rejected based on the next round of comments we receive from the reviewers. If you have any questions or concerns, please let me know. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: In this manuscript Billmann et al. present a succinct discussion of reproducibility metrics for context specific CRISPR screens. The primary advancement here is the development of the Within vs between-context replicate correlation (WBC). The goal of this metric is to quantify the reproducibility of a context-dependent signal as opposed to the traditional Pearson correlation which simply addresses technical noise within the experimental replicates. The authors present several panels of data (real and simulated) which highlight the utility of this metric. Their primary example involves analyzing a FASN genetic interaction screen, and showing that the WBC metric better captures the reproducibility of the identified genetic interactions. They also extend this analysis to the large scale CRISPR screening resource DepMap.

Primary strengths:

One of the primary strengths of this manuscript is its conciseness. The manuscript doesn't stretch longer than it needs to, and makes its point quickly and directly. I felt as though panel 1b-c, and 1g-h were the most convincing in regards to showing the utility of the WBC metric. Specifically 1h was very useful, showing visually how the low WBC in the MEL202 cell line accurately captured its low dLFC reproducibility.

Primary critiques:

The title I think could be rephrased to better highlight the focus of this work. A change as simple as "Reproducibility metrics for context specific CRISPR screens" might better alert a reader to what's being discussed. I think this edit is worth making because the authors readily acknowledge from the opening paragraph that the core set of "essential" genes has largely been defined, and the primary utility here is identifying reproducibility metrics for context specific screens. This edit is only one of many possible rephrasings, with the main goal being to better align the title with the second sentence of the abstract. I think this is important to distinguish what type of reproducibility is being discussed, because normally people think of reproducibility in terms of "how much technical noise was there between biological replicates".

Figure 1d is somewhat confusing insofar as it doesn't use the WBC metric. Only the GI score captures the 'true" hit LUR1, but their GI score method was already presented in Aregger et al., 2020. I suppose the idea is that only the GI score captures the true hit, therefore the GI score should have a high "correlation metric". Therefore, the WBC is a good metric, because it is high for the GI score? It all makes some sense, but the presentation is a little hard to follow, perhaps moving panel 1d earlier would make the logic a little more clear.

I find the comparisons made in 1e-f very unconvincing. The text says "the WBC score indicates that the dLFC metric reflects context-specific signal with much higher quality…", which is true for the comparison between LFC and dLFC. But this isn't the type of comparison that this metric would have most utility for. For example, I would imagine a big use for this metric would be to compare different CRISPR screens to identify which ones have high signal (as in 1g-h). In that sense, I think a better comparison would be to compare the WBC scores and the Pearson correlations for dLFCs across multiple cell lines. From looking at the definition of WBC, my intuition is that they would be highly correlated. On that same note, I'm not really sure how much novel information the WBC score tells us about the data of interest. Doesn't the Pearson correlation between the dLFC capture at least some of the same information? Insofar as for panel 1h SKBR3, and HCC1187 both have high WBC scores, as well as good gene level dLFC reproducibility between replicates. This point could be somewhat clarified by more text explaining what "dLFC reprod." means on the panel annotation. The legend simply says reproducibility is the "product" between dLFC replicates.

Reviewer #2: Billmann et al. describe an approach for measuring reproducibility metrics in CRISPR screens that are designed to identify a context specific signal. The CRISPR functional genomics community has some established standards for data reproducibility (e.g. as cited: DepMap but also reviewed in PMID: 29199283 and 32284587) but there is certainly room for innovation and clarity. Reproducibility is key for taking functional genomics from the current perception of having high false positive towards the potential of not being thought of as a noisy screen but rather being a quantitative experiment that defines how genes impact a phenotype of interest with a low false positive rate. My concerns with this manuscript relate to the novelty of the analytic approach. The authors have previously published a portion of the analytic framework reiterated in this current manuscript and they cite/discuss this work. For example a key metric in this manuscript is the qGI score or dLFC for identification of context specific biology. However, in the current manuscript they state the "a fully normalized dLFC (expressed as a qGI, see Aregger et al. 2020 for details)". So what about this manuscript is novel other than the elaboration that their previous approach is particularly useful for identification of context specific specific effects? The DepMap Chronos or CERES approach has a framework for identifying genes categorized as strongly selective vs common essential. The idea of using the dLFC to identify context specific effects has been established for both RNAi and CRISPR functional genomics platforms (e.g. PMID: 23739767 as an early example). The Within vs Between (WBC) score is to my knowledge a novel metric although not conceptually dissimilar from other approaches that leverage negative control pseudogene distributions constructed from negative control sgRNAs (e.g. PMID 27661255- in that both WBC and other approaches use a distribution of screen results to establish a distribution of non-scoring or non-hit data and then evaluate hits relative to this) and WBC is mostly used for analysis of the genetic interaction data as it seems when they extend their analysis to DepMap data the authors return to using LFC and dLFC analysis. In summary, highlighting and clarifying the novelty of the analytic approach in Billmann et al. relative to the authors previous efforts and relative to other similar approaches in the field would be very useful.

Reviewer #3: In this brief report the authors introduce a new metric for assessing the reproducibility of CRISPR screens. Their starting observation is that current correlation based metrics are unable to account for genomic context-specific sgRNA depletion fold-changes (computed between final versus initial time point post library selection, and usually considered as proxy measures of the targeted gene's essentiality), or gene depletion fold-changes (obtained by aggregating the sgRNA depletion fold-changes on a targeted gene basis). Going further, the authors demonstrate that a correlation based metric is not able to distinguish replicates of experiments within the same genomic context from those of experiments performed across genomic contexts. To tackle this problem the authors propose to use an alternative metric which estimates the reproducibility of a screen computing a ratio between the correlation of that screen's replicates over its expectation, i.e. the average correlation scores obtained when comparing replicates of different screens.
The author show that this metric is indeed able to discriminate same screen's replicates (yielding higher values) from different screens' replicates (yielding lower values) using data from a previous publication, as well as using data from the cancer dependency map. In particular in this last case the author show that their metric indeed yields higher values when computed between cell line specific profiles of gene fold changes (computed based on their distance from the consensus essentiality profile across screened cell lines).

As the authors rightly point out, most current uses of CRISPR-Cas9 screens aim at identifying context-specific phenotypes, for example cancer vulnerabilities associated to a given genomic aberration. This makes this study timely and important.

Although, I found the authors' conclusion that informative reproducibility measures are inconsistently reported because they tend to be relatively low and most of the published CRISPR screens are of poor quality very questionable. Furthermore, both topic and solution proposed in this brief report are not original or novel.

For example in Behan et al. Nature 2019, the authors reported that a genome wide correlation metric computed to assess individual screens' reproducibility was generally unable to distinguish replicates of the same screen from replicates of different screens (similarly as the authors do here). This was because the existence of a relatively small (compared to a genome-wide library) set of core-fitness genes that were consistently depleted in every screen, with large effects, thus inflating all correlation scores, regardless of the screens (this should be mentioned by the authors).

However Behan et al. solved this issue by first identifying a set of highly informative sgRNAs which were of sufficiently high on-target efficiency, as well as had a quite variable depletion signal across screens (thus reflective of their targeted genes being not core-fitness or never-essential, thus context-specifically essential). Second, they computed Pearson's correlation scores across replicates (of the same or different screens) on the domain of these sgRNAs only. In this way they were not only able to neatly distinguish the distributions of values yielded by within/between screen replicates' correlation respectively but also to define a robust quality threshold based on which some experiments were discarded, as deemed of scarce quality.
Particularly this is summarised in the Extended figure 1DEF of Behan et al. As a conclusion, at least Behan et al. included in their follow up analysis only high quality and robustly assessed data reporting all reproducibility statistics. This clearly contradicts the authors' conclusions.

The authors should mention the study by Behan et al, explicitly discussing conceptual commonalities and differences with respect to their approach as at this stage these are not clear to me.

For example a limitation of the approach proposed in Behan et al, which could favour the metric proposed in this short report, is that the template of sgRNA used to compute correlation scores was library specific thus unusable for screen performed using other libraries. In addition, deriving such template for a different library would require the availability of a large number of screens which would allow estimating sgRNA on target efficiency and signal variability. For example this could be done for the AVANA library (used in the other Cancer dependency map dataset) but it is generally impractical for other screens/library.

Finally, the two following minor points should be addressed:

- Although the author refer to their previous publication, a brief definition of qGI score and GI score should be included.

- a fold-change (log [ending measurement / initial measurement]) is by definition already a logarithmic measurement (i.e. the fold is the base of the considered logarithm). Are the authors actually considering the log2 of a gene (or sgRNA) depletion fold-change or it is just a redundancy ? in the latter case this should be corrected and the authors should use just fold-changes (FCs) instead of log fold-changes (LFCs).

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Billmann,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication. Reviewr #2 still has some concerns, but I believe they can be straight-forwardly addressed with additional discussion.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager.

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

Transparent Peer Review: Thank you for electing to make your manuscript's peer review process transparent. As part of our approach to Transparent Peer Review, we ask that you add the following sentence to the end of your abstract: "A record of this paper's Transparent Peer Review process is included in the Supplemental Information." Note that this ***doesn't*** count towards your 150 word total! Also, if you've deposited your work on a preprint server, that's great! Please drop me a quick email with your preprint's DOI and I'll make sure it's properly credited within your Transparent Peer Review record.

Manuscript text:
The supplemental PDF is intended to contain only figure and their legends, as well as tables of length 3 pages or less. Please move the methods to the main text and convert to our STAR Methods format (see below).

Figures and Legends:
Please look over your figures keeping the following in mind:

- Please ensure that every time you have used a graph, you have defined "n's" specifically and listed statistical tests within your figure legend.

STAR Methods:
Please convert your methods section to our STAR Methods format. See the STAR Methods guidelines for additional information.

Thank you!

Reviewer comments:

Reviewer #1: The authors have adequately addressed this reviewer's comments.

Reviewer #2: The authors have responded to my concerns. The revised manuscript is substantially clearer and provides context which I feel is quite helpful.
It would be helpful (assuming the authors think it statistically appropriate) to comment on how the WBC is powered for detection of context specific weak, medium and strong hit genes within data paradigms or experimental plans of a variable N= for experimental and control conditions. For example with the WBC metric in mind-- how should scientists design their experiments to identify context specific signal of a given magnitude? In the current FASN example the authors are comparing experimental conditions (N=3) to control (N=5). I am asking for this because many genome-scale functional genomics experiments are carried out as N=2 or 3 replicates for control and experimental conditions and controls are rarely present as N=5 for genome scale screens and thus many individual labs are potentially not able to robustly utilize WBC. The WBC uses the mean and standard deviation of its correlation with screens performed in a different context. Is there another way to do this if N= is limiting e.g. for primary cell screens etc.? The concept of power analysis is common in other areas of biological inquiry but not used in functional

genomics.

Reviewer #3: The authors have put a lot of effort in addressing mine and other reviewers' comments. Particularly they have clarified conceptual advances and advantages of their metric over previously published one and properly discussed previous literature.
I do believe this manuscript will make a good contribution to the field and will be of interest to computational biologists working with CRISPR screens.

Authors' response to the reviewer's comments

Attached.

Editorial decision letter with reviewers' comments

Dear Dr. Billmann,

I'm very pleased to let you know that your manuscript is now "accepted in principle," that is, provisionally accepted pending our receipt of final files that meet the journal's formatting requirements and the *Final Editorial Changes* below. Congratulations!

Final Editorial Changes: Please move the Methods to the main text and convert these to our STAR Methods format (including having a Key Resources Table) - see the STAR Methods guidelines for additional information.

Please review the information below along with the detailed formatting requirements listed in the Final Files Checklist. We've also put together this FAQ (click the Final Formatting Checks tab) for your convenience. Please ask any questions you may have, make any necessary changes to your manuscript files, and then upload your final files into Editorial Manager. Once we receive your formatted files, we will go through our formatting checks and let you know if further changes are needed.

We hope to receive your formatted files within 5 business days. Please email me directly if this timing is a problem or you're facing extenuating circumstances. Alternatively, if this manuscript needs to be officially accepted by a particular date because of grant deadlines, applications, or because it will help your trainees, please let me know.

Introducing new referencing style

To standardize the referencing style across Cell Press journals, starting from October 2022, we ask that all in-text citations be formatted as superscripted numbers (e.g. “Multiple reports support this observation.[1,2]”). Moving away from the Harvard referencing style (e.g. Smith *et al*., 2020) will improve author and reader experiences. All manuscripts accepted from now on must use **the superscript numbered Cell Press referencing style**. Make sure to use this numbered referencing style for all new and revised submissions as well. Switching is easy. Just use the updated CSL and EndNote referencing styles for Cell Press articles.

Important update: Cell Press recently created a mechanism to include a formal Inclusion and Diversity statement within published papers that authors may use if they wish. It was introduced in a *Cell* editorial that you can read here, and more practical information can be found in our Final Files Checklist. *Cell Systems* appreciates that there are many valid perspectives on how to address the well-documented and long-standing systemic inequalities in science. We are eager to know your view of this Cell Press initiative, and if you would like to talk to us about it, please reach out.

Below my signature, you'll find specific information about what to expect next regarding formatting checks and working with our Production Department after acceptance. It’s been a pleasure working with you, please feel free to contact our journal team with questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewer comments:

Reviewer #2: the authors have addressed all my concerns and should be commended on a very nice manuscript which will be quite useful for the field if functional genomics.

Reviewers' comments:

Reviewer #1: In this manuscript Billmann et al. present a succinct discussion of reproducibility metrics for context specific CRISPR screens. The primary advancement here is the development of the Within vs between-context replicate correlation (WBC). The goal of this metric is to quantify the reproducibility of a context-dependent signal as opposed to the traditional Pearson correlation which simply addresses technical noise within the experimental replicates. The authors present several panels of data (real and simulated) which highlight the utility of this metric. Their primary example involves analyzing a FASN genetic interaction screen, and showing that the WBC metric better captures the reproducibility of the identified genetic interactions. They also extend this analysis to the large scale CRISPR screening resource DepMap.

Primary strengths:

One of the primary strengths of this manuscript is its conciseness. The manuscript doesn't stretch longer than it needs to, and makes its point quickly and directly. I felt as though panel 1b-c, and 1g-h were the most convincing in regards to showing the utility of the WBC metric. Specifically 1h was very useful, showing visually how the low WBC in the MEL202 cell line accurately captured its low dLFC reproducibility.

We thank the reviewer for the positive summary of our work and provide a point-by-point response below.

Primary critiques:

The title I think could be rephrased to better highlight the focus of this work. A change as simple as "Reproducibility metrics for context specific CRISPR screens" might better alert a reader to what's being discussed. I think this edit is worth making because the authors readily acknowledge from the opening paragraph that the core set of "essential" genes has largely been defined, and the primary utility here is identifying reproducibility metrics for context specific screens. This edit is only one of many possible rephrasings, with the main goal being to better align the title with the second sentence of the abstract. I think this is important to distinguish what type of reproducibility is being discussed, because normally people think of reproducibility in terms of "how much technical noise was there between biological replicates".

We thank the reviewer for this suggestion. We agree and have updated the title of the revised manuscript to: "Reproducibility metrics for context-specific CRISPR screens".

Figure 1d is somewhat confusing insofar as it doesn't use the WBC metric. Only the GI score captures the 'true" hit LUR1, but their GI score method was already presented in Aregger et al., 2020. I suppose the idea is that only the GI score captures the true hit, therefore the GI score should have a high "correlation metric". Therefore, the WBC is a good metric, because it is high for the GI score? It all makes some sense, but the presentation is a little hard to follow, perhaps moving panel 1d earlier would make the logic a little more clear.

We thank the reviewer for the helpful comment regarding the logical flow. Yes, the logic as you demonstrated is close. We would like to stress that we do not necessarily expect the PCC to be high, but we expect a QC metric to be in a range that suggests good quality, particularly for the data processing level that matters most for the experiment, which is not true for the PCC. We agree this point is somewhat complex but important, and that our previous logic was hard to follow. We have expanded the section of the text that refers to Figure 1d to make this more clear. Specifically, we expanded this previous *sentence:: "Moreover, only dLFC and qGI scores identified the recently discovered link between FASN and LUR1 (Figure 1d) (Aregger et al., 2020)." to this*

"We note that only the context-specific scores (dLFC and qGI) capture the biologically relevant signal in this case, which are genetic interactions with the FASN query mutation. For example, only dLFC and qGI scores are able to identify the gene LUR1 as a top interacting partner (Figure 1d), which was recently characterized as playing a functional role in lipid metabolism with FASN (Aregger et al., 2020). Using simple replicate PCC as a measure of reproducibility, one would conclude that these context-specific scores are of lower quality than the less biologically relevant scores from earlier stages of data processing, but a context-specific reproducibility score such as the WBC score suggests the opposite. Both the metric one chooses to quantify reproducibility and the stage of data processing at which this measurement is taken are important for making accurate conclusions about data quality."

To better support the logical flow, we also streamlined the header of panels 1a-b.

I find the comparisons made in 1e-f very unconvincing. The text says "the WBC score indicates that the dLFC metric reflects context-specific signal with much higher quality…", which is true for the comparison between LFC and dLFC. But this isn't the type of comparison that this metric would have most utility for. For example, I would imagine a big use for this metric would be to compare different CRISPR screens to identify which ones have high signal (as in 1g-h). In that sense, I think a better comparison would be to compare the WBC scores and the Pearson correlations for dLFCs across multiple cell lines. From looking at the definition of WBC, my intuition is that they would be highly correlated.

We thank the reviewer for this suggestion. We had originally only compared the traditional metric, PCC on LFC, with our new suggestion, the WBC on dLFC in panel 1g. We agree with the reviewer that it is also important to show the direct comparison between WBC and PCC with both applied on dLFC data, which we have now added as a new panel to Figure 1 (new 1h). To cover the dLFC PCC high and dLFC WBC low area in 1h, we chose the cell line A2780, which is also shown in the new panel i now. WBC and PCC applied to dLFC for the same set of cell lines are moderately but not highly correlated (PCC = 0.53, SCC=0.52)). Thus, if one were to rank screens on the basis of either metric, there would be substantial differences in this ranking.

To clarify the coherence of the panels Figure 1e-h (new 1e-i), we would like to highlight that panel 1e shows that the concept we established in 1a-d is generalizable: the more biologically relevant data processing level tends to exhibit poorer performance based on a standard PCC metric. Specifically, panels 1e and 1f demonstrate that PCC measures are substantially higher on LFC as compared to dLFC (panel e), but that this trend is reversed for the WBC score (panel f). Panels 1h-i (new) add to the trends illustrated in 1e-f by visualizing the non-trivial relation between the traditional metric (PCC) and the WBC.

Again, our general concern when evaluating differential (e.g. cell line-specific) CRISPR screening results is that the dominant reproducibility metric used, a simple (mostly Pearson) correlation coefficient, provides erroneous intuition. Specifically, it tends to become smaller (a negative interpretation) once the data of interest becomes more sparse (sparsity occurs at the transition from general to context-specific fitness). We realize that our message is somewhat complex– we aim to raise awareness on not only which metrics to use but also on which data processing level such metrics should be applied. We have therefore included the following description in our manuscript and hope the intent of the demonstrated analysis becomes more clear: " We made the assumption that screens performed in the same cell line, here replicates, contain context (cell line)-specific effects that distinguish a given cell line from other cell lines, and that those effects are quantified by dLFC rather than LFC values. We tested how the PCC and WBC quantify screen replication and how those metrics change when we focus on the cell line-specific (dLFC) signal."

We believe that this additional description now better clarifies why we conclude that the WBC is an improved metric for QC cell line-specific CRISPR screens relative to a standard PCC of replicates.

On that same note, I'm not really sure how much novel information the WBC score tells us about the data of interest. Doesn't the Pearson correlation between the dLFC capture at least some of the same information?

As described above, the correlation between PCC applied to dLFC and the WBC applied to dLFC on all DepMap cell lines is 0.53 (Pearson), with a Spearman correlation of 0.52. Thus, the WBC captures different information (i.e. the screens that you would rank highest by PCC and WBC are substantially different).

The second important difference between WBC and PCC is the interpretability of the range of scores. The relevant data processing steps result in relatively low PCCs (mean of about 0.5) between replicates, which in our experience is mostly not accepted as indicative of "good quality" data. Importantly, "acceptable" PCC ranges are somewhat subjective. In contrast, the WBC score provides a clear cut-off with a statistical interpretation, since it is a specific version of a z-score. For example, most researchers would agree a z-score of much greater than 3 reflects reasonable quality data while a z-score of < 1 (i.e. replicate similarity is less than 1 standard deviation from the mean of randomly paired screens) is of questionable quality.

Insofar as for panel 1h SKBR3, and HCC1187 both have high WBC scores, as well as good gene level dLFC reproducibility between replicates. This point could be somewhat clarified by more text explaining what "dLFC reprod." means on the panel annotation. The legend simply says reproducibility is the "product" between dLFC replicates.

We thank the reviewers for pointing this out, and we agree that this should be clarified. Those values to scale dots are the per-gene contributions to the PCC of the pair of screen replicates at the dLFC level. In the new version of this figure (new Figure 1i, old 1h), we added this explanation to the figure legend: "Circle size indicates each gene's dLFC reproducibility and corresponds to the per-gene dLFC product between replicate screens.". Additionally, we added for each panel in new Figure 1i a dLFC replicate comparison in Figure S1d, which uses the same scaling factor and shows that this scaling factor increases when per-gene dLFC values become larger AND agree in their direction.

Reviewer #2: Billmann et al. describe an approach for measuring reproducibility metrics in CRISPR screens that are designed to identify a context specific signal. The CRISPR functional genomics community has some established standards for data reproducibility (e.g. as cited: DepMap but also reviewed in PMID: 29199283 and 32284587) but there is certainly room for innovation and clarity. Reproducibility is key for taking functional genomics from the current perception of having high false positive towards the potential of not being thought of as a noisy screen but rather being a quantitative experiment that defines how genes impact a phenotype of interest with a low false positive rate. My concerns with this manuscript relate to the novelty of the analytic approach.

We thank the reviewer for placing our work into the context of recent literature. We agree that PMID: 29199283 and 32284587 provide some of the most comprehensive guides to analysis of data from CRISPR screens. Based on this and Reviewer #3's comment, we have added a paragraph that better puts our work in context with previous discussion of data reproducibility in the CRISPR community, including those two publications (see our 2nd response to reviewer #3 below). We address the comment about the novelty of our contributions in the response that immediately follows this.

The authors have previously published a portion of the analytic framework reiterated in this current manuscript and they cite/discuss this work. For example a key metric in this manuscript is the qGI score or dLFC for identification of context specific biology. However, in the current manuscript they state the "a fully normalized dLFC (expressed as a qGI, see Aregger et al. 2020 for details)". So what about this manuscript is novel other than the elaboration that their previous approach is particularly useful for identification of context specific specific effects?

The reviewer is correct that we previously published the CRISPR screen data we use for illustration here and also, in that paper, we used the qGI score to identify genetic interactions from CRISPR screens (Aregger et al. 2020). The current manuscript is not focused on the qGI score, the dLFC measure, or in general, computational methods designed to identify genetic interactions. Instead, this manuscript is focused on reproducibility metrics best suited for context-specific CRISPR screens. As discussed above, the main contributions of the current manuscript are to: (1) raise the point that the level of data processing at which reproducibility is measured for CRISPR screens can substantially affect the result and the extent to which this actually reflects the quality of the context-specific signal, (2) show that the commonly used reproducibility metric (the PCC) is sensitive to the sparsity of signal in context-specific CRISPR screen analysis, and (3) provide an alternative metric (the WBC score), which is more informative for measuring reproducibility in this context. None of these points were discussed in our previous publication (Aregger et al. 2020), nor have they been discussed in previous literature on guidelines for CRISPR screen analysis (e.g. PMID 29199283 or 32284587). In our opinion, these are important additions to the discussion of reproducibility in the CRISPR functional genomics community.

Why do we reuse our previously published data (the FASN KO screen) here? We decompose this data, where different dominant signals (gRNA library representation standard deviation, general gene essentiality, etc.) are successively excluded. This helps us to illustrate what a correlation coefficient quantifies and serves as a practical guide for anyone testing reproducibility of their CRISPR screening data. We complement this data with orthogonal data sets and a simple simulation. We hope that this generates a helpful precedent for the community to put into perspective what correlation coefficients on CRISPR screening data might indicate and some caveats on their interpretation.

The DepMap Chronos or CERES approach has a framework for identifying genes categorized as strongly selective vs common essential. The idea of using the dLFC to identify context specific effects has been established for both RNAi and CRISPR functional genomics platforms (e.g. PMID: 23739767 as an early example). The Within vs Between (WBC) score is to my knowledge a novel metric although not conceptually dissimilar from other approaches that leverage negative control pseudogene distributions constructed from negative control sgRNAs (e.g. PMID 27661255- in that both WBC and other approaches use a distribution of screen results to establish a distribution of non-scoring or non-hit data and then evaluate hits relative to this)...

We thank the reviewer for pointing us to the important work described in PMID 27661255, which uses control distributions to identify individual gene effects in CRISPR screens. This does not address the issue we tackle with the WBC score, which is to quantify reproducibility of CRISPR screens. It is correct that leveraging control distributions has been widely used, but these approaches do not measure the degree to which replicated screens agree with each other. Rather, they compare the measured effects of guides targeting regions of interest with control regions to establish statistical measures within a single screen. We now place our work more directly into the context of previous literature such as Behan et al., 2019 to discuss and clarify the novelty and potential utility of the WBC score (see our 2nd response to reviewer #3 below).

… and WBC is mostly used for analysis of the genetic interaction data as it seems when they extend their analysis to DepMap data the authors return to using LFC and dLFC analysis.

We thank the reviewer for allowing us to clarify this point. The WBC score helps to interpret reproducibility of context-specific *biological signal in CRISPR screens, and genetic interaction screens are an ideal example of such data. Another example is cell line-specific CRISPR screens. Much of the focus surrounding the analysis of the DepMap dataset centers on dependencies that are specific to small subsets of cell lines rather than shared dependencies as those reflect core essential genes. Thus, context-specific signal is often the focus. Since cell lines differ in more than just one genetic mutation, the dLFC is a more appropriate processing level of the DepMap data as opposed to the qGI score. The emphasis is not on the distinction between the qGI or dLFC scores– they both capture context-specific signal on their respective datasets. We also include the simple fitness measure (LFC) to highlight shortcomings of a simple correlation coefficient in Figure 1e-h.*

To clarify this point, we added the text now saying: "We made the assumption that screens performed in the same cell line, here replicates, contain context (cell line)-specific effects that distinguish a given cell line from other cell lines, and that those effects are quantified by dLFC rather than LFC values. We tested how the PCC and WBC quantify screen replication and how those metrics change when we focus on the cell line-specific (dLFC) signal.".

In summary, highlighting and clarifying the novelty of the analytic approach in Billmann et al. relative to the authors previous efforts and relative to other similar approaches in the field would be very useful.

We thank the reviewer for the helpful comments and agree that clarifying how our work fits with the state-of-the-art literature and which aspects of our work are conceptually complementary made our manuscript more relevant. We now cite and discuss additional papers that are central to the community's discussion about CRISPR screen data quality and reproducibility. We also further clarify the novelty and utility of our proposed metric.

We added this new paragraph to the text:

We note that there have been other complementary efforts to establish best practices for conducting CRISPR screens and analyzing the resulting data (Behan et al., 2019; Doench 2018; Hanna & Doench 2020). In particular, Behan et al. recognized the challenges of computing correlation between replicate screens based on whole dependency profiles. Specifically, they noted that including the core essential genes in this calculation inflates the correlation such that replicates of the same screen are generally less *distinguishable from replicates of different screens. Second, they noted that including guides targeting genes that never showed phenotypes led to pessimistic estimates of reproducibility due to the sparsity of signal across the dependency profile. Behan et al. addressed these issues by pre-processing the data to find the most variable signal (excluding both core essential genes and genes with no phenotypes) and to compute correlation on that subset of the data, which provides a more informative report of the data reproducibility. We address related issues here, but rather than pre-filtering of profiles, which may depend on the specific gRNA library used or a large collection of screens, we instead suggest that reproducibility analysis should be performed on scores that capture context-specific signal (e.g. dLFC). Furthermore, we propose a new metric, the WBC score, that is more directly interpretable than a correlation coefficient when applied to a sparse profile. Our suggested approach can be applied to a variety of CRISPR screening contexts.*

Reviewer #3: In this brief report the authors introduce a new metric for assessing the reproducibility of CRISPR screens. Their starting observation is that current correlation based metrics are unable to account for genomic context-specific sgRNA depletion fold-changes (computed between final versus initial time point post library selection, and usually considered as proxy measures of the targeted gene's essentiality), or gene depletion fold-changes (obtained by aggregating the sgRNA depletion fold-changes on a targeted gene basis). Going further, the authors demonstrate that a correlation based metric is not able to distinguish replicates of experiments within the same genomic context from those of experiments performed across genomic contexts. To tackle this problem the authors propose to use an alternative metric which estimates the reproducibility of a screen computing a ratio between the correlation of that screen's replicates over its expectation, i.e. the average correlation scores obtained when comparing replicates of different screens.

The author show that this metric is indeed able to discriminate same screen's replicates (yielding higher values) from different screens' replicates (yielding lower values) using data from a previous publication, as well as using data from the cancer dependency map. In particular in this last case the author show that their metric indeed yields higher values when computed between cell line specific profiles of gene fold changes (computed based on their distance from the consensus essentiality profile across screened cell lines).

As the authors rightly point out, most current uses of CRISPR-Cas9 screens aim at identifying context-specific phenotypes, for example cancer vulnerabilities associated to a given genomic aberration. This makes this study timely and important.

We thank the reviewer for these comments and the summary of the main contributions of our work. We have included our response to the specific concerns raised inline below.

Although, I found the authors' conclusion that informative reproducibility measures are inconsistently reported because they tend to be relatively low and most of the published CRISPR screens are of poor quality very questionable. Furthermore, both topic and solution proposed in this brief report are not original or novel. For example in Behan et al. Nature 2019, the authors reported that a genome wide correlation metric computed to assess individual screens' reproducibility was generally unable to distinguish replicates of the same screen from replicates of different screens (similarly as the authors do here). This was because the existence of a relatively small (compared to a genome-wide library) set of core-fitness genes that were consistently depleted in every screen, with large effects, thus inflating all correlation scores, regardless of the screens (this should be mentioned by the authors).

We would like to thank the reviewer for placing our work in context of what has been investigated by Behan et al. Nature 2019, and for pointing out that our conclusions should be made more precise. ***We do not intend to say, nor do we believe, that most CRISPR screens are of poor quality****. Indeed, we are frequently impressed by the quality of most published data sets and their utility for systematic meta-analyses. To be more precise and place our work into the context of previous work, we now discuss and cite Behan et al. Nature 2019 and additional previous literature related to CRISPR screen analysis (see the new text we added in our 3rd response below).*

To further clarify our intent, based on numerous impressions during scientific communications over the past years, we speculated that correlation coefficients smaller than about 0.7 (acceptable thresholds vary, typically between 0.6 and 0.9) are generally considered a sign of “poor quality”. In fact, we are convinced that most such interpretations are overly pessimistic and speculate that the issue with this prevalent conception is that a correlation coefficient does not quantify reproducibility when used on a sparse dataset. To demonstrate our point in the manuscript, we simulated how amplitude and sparsity of signal in CRISPR screening data translate into correlation coefficient scores (Figure 1j).

We will respond in-line below to clarify why our work contributes original and novel aspects to guide CRISPR screen reproducibility. We also added new panels to the main figure (1h-j) to illustrate more technical aspects of the WBC that contribute novel concepts to the field.

However Behan et al. solved this issue by first identifying a set of highly informative sgRNAs which were of sufficiently high on-target efficiency, as well as had a quite variable depletion signal across screens (thus reflective of their targeted genes being not core-fitness or never-essential, thus context-specifically essential). Second, they computed Pearson's correlation scores across replicates (of the same or different screens) on the domain of these sgRNAs only. In this way they were not only able to neatly distinguish the distributions of values yielded by within/between screen replicates' correlation respectively but also to define a robust quality threshold based on which some experiments were discarded, as deemed of scarce quality.

Particularly this is summarised in the Extended figure 1DEF of Behan et al. As a conclusion, at least Behan et al. included in their follow up analysis only high quality and robustly assessed data reporting all reproducibility statistics. This clearly contradicts the authors' conclusions.

To clarify again, we do not intend to say, nor do we believe, that the data presented in Behan et al., 2019 and similar studies like it (e.g. Meyers et al., 2017) are of poor quality, or that the quality has not been assessed properly. We specifically discuss the WBC score in comparison to the approach described in Behan et al. 2019 below.

The authors should mention the study by Behan et al, explicitly discussing conceptual commonalities and differences with respect to their approach as at this stage these are not clear to me.

We thank the reviewer for this suggestion and agree that a specific comparison with Behan et al. will strengthen the manuscript. Behan et al. recognized the challenges of computing correlation between replicate screens based on whole dependency profiles. Specifically, they noted that including the core essential genes in this calculation inflates the correlation such that replicates of the same screen are generally less distinguishable from replicates of different screens. Second, they noted that including guides targeting genes that never showed phenotypes led to pessimistic estimates of reproducibility due to the sparsity of signal across the dependency profile. Behan et al. addressed these issues by pre-processing the data to find the most variable signal (excluding both core essential genes and genes with no phenotypes) and to compute correlation on that subset of the data, which provides a more informative report of the data reproducibility. Behan et al. also precisely define thresholds based on a sophisticated evaluation, which helps guide a statistical interpretation of the measured correlation coefficient.

We believe this approach has the following limitations:

1. *Pre-selecting gRNAs changes the correlation coefficient for a given screen. Depending on how the gRNA signature is chosen, coefficients can have multiple values for the same dataset. This was not problematic in the context of the Behan et al. study, but would make the actual value of correlation coefficient difficult to interpret and would likely become a challenge when comparing completely different studies.*
2. *Selecting those gRNA to include/or exclude requires information that, in many cases, is not available in a more specialized screening context. As the reviewer noted, perhaps studies using the same library as Behan et al. could use the same selected set of gRNAs, but this does not necessarily generalize to other libraries or other screening contexts. As a field, ideally we could establish library-independent QC statistics that are applicable regardless of the biological focus of the screens.*

The proposed WBC score addresses both of these limitations as it is library- independent and doesn't require the preselection of individual guides. Furthermore, the z-score is directly interpretable regardless of the screen context (e.g. WBC scores < 1 would not be viewed as "high-quality" in any context). Beyond the WBC metric, our manuscript also raises the key point that the level of data processing at which reproducibility is measured for CRISPR screens can substantially affect the apparent reproducibility. If the goal of a set of screens is to identify context-specific signal, then a statistic that captures context-specificity should be the main focus of the analysis, including any measurements of reproducibility. Once a context-specific measure is used (e.g. dLFC or qGI), one no longer needs to filter out the generic signal that appears in every screen (e.g. core essential genes) and analyses can be applied to the whole profile. In general, we believe that our proposed WBC metric, along with the recommendation that this be applied at the appropriate data processing level, better generalizes to a variety of CRISPR screening contexts.

Finally, one more technical difference captured by the WBC statistic relative to the method described in Behan et al. is that the between-screen background expectation in Behan et al. is computed for the entire data set and therefore provides one global background distribution. Since a correlation coefficient is (i) highly dependent on the sparsity of the data and (ii) CRISPR screens have varying sparsity, which is also true for the DepMap (see our new Figure S3a), a sparser screen has a lower background correlation distribution (see new Figure S3b). The WBC computes a per-screen background distribution, which increases specificity for rich signal screens and sensitivity for sparse screens.

In general, we agree with the reviewer that putting our work in context of the Behan et al. work and properly citing their recognition of the same issues we address here is important. We have added a paragraph that discusses our contributions in the context of Behan et al. and other literature:

We note that there have been other complementary efforts to establish best practices for conducting CRISPR screens and analyzing the resulting data (Behan et al., 2019; Doench 2018; Hanna & Doench 2020). In particular, Behan et al. recognized the challenges of computing correlation between replicate screens based on whole dependency profiles. Specifically, they noted that including the core essential genes in this calculation inflates the correlation such that replicates of the same screen are generally less distinguishable from replicates of different screens. Second, they noted that including guides targeting genes that never showed phenotypes led to pessimistic estimates of reproducibility due to the sparsity of signal across the dependency profile. Behan et al. addressed these issues by pre-processing the data to find the most variable signal (excluding both core essential genes and genes with no phenotypes) and to compute correlation on that subset of the data, which provides a more informative report of the data reproducibility. We address related issues here, but rather than pre-filtering of profiles, which may depend on the specific gRNA library used or a large collection of screens, we instead suggest that reproducibility analysis should be performed on scores that capture context-specific signal (e.g. dLFC). Furthermore, we propose a new metric, the WBC score, that is more directly interpretable than a correlation coefficient when applied to a sparse profile. Our suggested approach can be applied to a variety of CRISPR screening contexts.

And to further clarify the technical details, we expanded the ***Methods*** *now stating: "To define the Within-vs-Between context replicate Correlation (WBC) score for a given screen, its biological replicate correlation is scaled to its expected background correlation distribution: the mean and standard deviation of its correlation with screens performed in another context (e.g. query mutation). This converts the correlation coefficients into a metric with an unambiguous statistical interpretation that can be interpreted as a z-score. Notably, each context (e.g. a set of screens done in a cell line within a larger set of screens covering multiple cell lines) creates its own background correlation distribution. This is important, because even at the same data processing level, signal sparsity substantially differs between contexts, and the* *background correlation is dependent on the abundance of the within-context signal. An example for this trend is illustrated for 693 distinct cell lines taken from the DepMap (Figure S3a, b).*

For the example shown in detail in this work, the FASN genetic interaction screens, …".

For example a limitation of the approach proposed in Behan et al, which could favour the metric proposed in this short report, is that the template of sgRNA used to compute correlation scores was library specific thus unusable for screen performed using other libraries. In addition, deriving such template for a different library would require the availability of a large number of screens which would allow estimating sgRNA on target efficiency and signal variability. For example this could be done for the AVANA library (used in the other Cancer dependency map dataset) but it is generally impractical for other screens/library.

We thank the reviewer for pointing out this possibility and limitation. We included this point in the discussion above (see our previous response).

Finally, the two following minor points should be addressed:

- Although the author refer to their previous publication, a brief definition of qGI score and GI score should be included.

We thank the reviewer for the interest in the qGI score. We have now added a short description of the qGI score to the relevant methods paragraph saying: "The quantitative genetic interaction (qGI) score represents the differential fitness effect between a wildtype control and query gene (here FASN) knockout screen after correcting query gene-unspecific screening artifacts. LFC and quantitative genetic interaction (qGI) scores were generated as described in Aregger et al. 2020."

The relevant information for the aspect of CRISPR screen reproducibility, which is the biological utility over a simple differential fitness effect, has been initially demonstrated in Aregger et al., 2020. A more detailed characterization and evaluation of the qGI beyond the information provided in Aregger et al., 2020 is beyond the scope of the current manuscript as our current focus is on measuring reproducibility of screens.

- a fold-change (log [ending measurement / initial measurement]) is by definition already a logarithmic measurement (i.e. the fold is the base of the considered logarithm). Are the authors actually considering the log2 of a gene (or sgRNA) depletion fold-change or it is just a redundancy ? in the latter case this should be corrected and the authors should use just fold-changes (FCs) instead of log fold-changes (LFCs).

By our understanding, "fold-change" simply refers to a ratio of two quantities (e.g. [ending measurement / initial measurement]). Here, when processing the readcounts, we follow the standard procedure to analyze CRISPR screening effects, which is to first build the ratio of the ending measurement of the initial measurement and then log2-transform this ratio. We refer to this combination as log fold-change (LFC), which we believe is standard.

Reviewer comments:

Reviewer #1: The authors have adequately addressed this reviewer's comments.

Reviewer #2: The authors have responded to my concerns. The revised manuscript is substantially clearer and provides context which I feel is quite helpful.

It would be helpful (assuming the authors think it statistically appropriate) to comment on how the WBC is powered for detection of context specific weak, medium and strong hit genes within data paradigms or experimental plans of a variable N= for experimental and control conditions. For example with the WBC metric in mind-- how should scientists design their experiments to identify context specific signal of a given magnitude? In the current FASN example the authors are comparing experimental conditions (N=3) to control (N=5). I am asking for this because many genome-scale functional genomics experiments are carried out as N=2 or 3 replicates for control and experimental conditions and controls are rarely present as N=5 for genome scale screens and thus many individual labs are potentially not able to robustly utilize WBC. The WBC uses the mean and standard deviation of its correlation with screens performed in a different context. Is there another way to do this if N= is limiting e.g. for primary cell screens etc.? The concept of power analysis is common in other areas of biological inquiry but not used in functional genomics.

We thank the reviewer for the practical question. While any statistical test benefits from as many replicates as possible, we agree that N=5 independent screens are rarely performed. To evaluate the stability of the WBC estimates, we sampled our 5 non-FASN KO screens down to subsets of 2, 3 and 4 screens and compared the 3 FASN KO screens against all the possible subsets. Furthermore, we sub-sampled 2 of the 3 FASN KO screens in addition. As the new Figure S4 shows, the WBC estimates become more variable at lower numbers of FASN and non-FASN-KO screens. However, even in these scenarios with fewer replicates, the WBC scores clearly distinguish the quality of the context-specific signal captured at various stages of data processing. For example, even for the case of only 2 FASN replicates and 2 non-FASN KO screens (the most limiting scenario), we observed a clear increase in the WBC of the qGI score vs. the LFC score (Fig. S4b, top left panel).

We added the following sentences to the method section describing the formula of the WBC: "While larger N and M provide more robust estimates of the WBC, we found that WBCs derived from any combination of 2, 3 or 4 of the LDLR, SREBF1, SREBF2, ACACA and C12orf49/LUR1 screens as well as only using 2 FASN KO screens provided stable measures of context-specific signal that distinguished scores derived from different stages of data processing (Figure S4a, b)."

Reviewer #3: The authors have put a lot of effort in addressing mine and other reviewers' comments. Particularly they have clarified conceptual advances and advantages of their metric over previously published one and properly discussed previous literature.

I do believe this manuscript will make a good contribution to the field and will be of interest to computational biologists working with CRISPR screens.