Enhancing Reproducibility Through Bioimage Analysis: The Significance of Effect Sizes and Controls

David J. Barry[1*], Stefania Marcotti[2], Lina Gerontogianni[3] and Gavin Kelly[3]

[1] Crick Advanced Light Microscopy Science Technology Platform, The Francis Crick Institute, London, UK

[2] Randall Centre for Cell and Molecular Biophysics, King's College London, London, UK

[3] Bioinformatics and Biostatistics Science Technology Platform, The Francis Crick Institute, London, UK

* david.barry@crick.ac.uk

Abstract

Bioimage analysis is a powerful tool for investigating complex biological processes, but its robustness depends on both technical precision and rigorous experimental design. In particular, the use of appropriate controls is critical for drawing meaningful biological conclusions. However, controls are often inadequately applied or overlooked in favour of achieving statistical significance, frequently obtained through the misuse or misinterpretation of statistical tests. In this study, we reanalyse a publicly available image dataset to highlight the crucial role of controls in interpreting experimental outcomes. Our findings underscore the importance of focusing on effect sizes and biological relevance over arbitrary statistical thresholds. We also discuss the diminishing returns of increased data collection once statistical stability has been achieved, advocating for more efficient experimental designs. By refining control usage and emphasising effect sizes, this work aims to enhance the reproducibility and robustness of research findings. We provide open-access code to allow researchers to engage with the dataset, promoting better practices in experimental design and data interpretation.

Introduction

Bioimage analysis has revolutionised our ability to quantify and interpret complex biological phenomena. However, the robustness of these analyses hinges not only on technical advancements but also on the rigour of experimental design and the correct interpretation of results. Unfortunately, there are many examples in the literature where neither of these criteria are met: experiments are poorly designed with inadequate controls and a tendency to falsely equate biological and statistical significance. Both of these problems stem from a lack of understanding of the variability and measurement noise that is inherent in biological experimentation - a measured difference may appear significant, but this difference could result from day-to-day, sample-to-sample or even cell-to-cell variability, rather than any real biological effect.

Despite widespread awareness of their importance, controls are often inadequately implemented or poorly chosen, leading to misinterpretations of experimental results. Effective controls are crucial for distinguishing true biological effects from artifacts introduced by the experimental procedure. However, their role is frequently overshadowed by the pursuit of statistical significance, which is widely acknowledged as a substantial contributor to reproducibility issues (Cobey et al., 2024).

Additionally, despite recognising the importance of repeating experiments to demonstrate reproducible results, researchers often treat individual measurements from within the same sample as "replicates". Often termed "pseudo-replication" by statisticians, this issue is commonly encountered in bioimage analysis when researchers present measurements from individual cells, drawn from the same sample, as independent data points. These data points are then grouped according to treatment and subjected to statistical tests in pursuit of a "significant" p-value. Applying tests such as ANOVA to these data is inappropriate, as cells from the same sample are necessarily subjected to the same treatment and cannot generally be considered independent. These issues are often exacerbated when non-independent data points from multiple experiments are pooled, artificially inflating sample size and leading to misleadingly low p-values (Lord et al., 2020).

This pursuit of statistical significance can obscure the importance of technically sound experimental design, as small or irrelevant effects may appear significant in the absence of rigorous comparative frameworks. Furthermore, the use of inappropriate or recycled controls undermines the biological

validity of findings, leading to conclusions that lack reproducibility. As bioimage data often involve complex and variable biological systems, interpreting results without the proper contextual grounding provided by controls can lead to misleading conclusions.

Contrary to popular belief, obtaining a very large number of data points from a sample to achieve a statistically sound result is often unnecessary and can lead to an unnecessary expenditure of resources. The distribution of values within a large population can often be well-represented by a relatively small sample size. While increasing this sample size will result in greater confidence in the obtained statistical descriptors, the law of diminishing returns eventually kicks in. The point at which one should cease collecting data ultimately depends on the level of confidence needed in the statistical description of a particular population, which in turn depends on the size of the effect one is attempting to detect. However, it is important to acknowledge that, even with a well-designed experiment and well-chosen controls, the results can still be ambiguous, and researchers need to be more accepting of this uncertainty in their results.

In this study, we focus on the critical role of experimental controls in bioimage analysis, using a publicly available dataset to demonstrate how conclusions can differ significantly based on control selection. Our goal is to emphasise the importance of control conditions that are not just technically adequate, but biologically meaningful. By refining control usage and focusing on effect sizes over arbitrary statistical thresholds, we aim to guide researchers toward more robust and reproducible outcomes.

While there are many guides on statistics for biologists, such as the excellent 'Points of Significance' series in Nature (Krzywinski and Altman, 2013) and the recently published work by (Zweifach, 2024), they can be perceived as somewhat general, abstract, and difficult to apply to actual experimental results. In contrast, we present an analysis of previously published image data, emphasizing that (a) beyond a certain threshold, effect sizes are relatively invariant to sample size, (b) interpreting biological effects is challenging without appropriate controls, and (c) when interpreting effect sizes, there is a diminishing return with continued data collection. We have made the code used for generating the figures in this manuscript openly accessible online, encouraging readers to engage with the data themselves and reproduce the figures using their own datasets. This approach, augmented by actual microscopy data,

seeks to make these statistical concepts tangible and directly applicable for cell biologists. It is our hope that this helps to "demystify" statistical analysis, in the concept of bioimage analysis, and encourage greater transparency and reproducibility and encourage researchers to embrace ambiguity and uncertainty to a much greater extent. To quote from Fisher and Prance (1974): "*No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur ... In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure.*"

Results & Discussion

Overview

We reanalysed a subset of the data previously published by Lawson et al. (2022) (Fig. 1A). We chose this particular dataset as it met the following desirable criteria: (1) the images are publicly available via the Image Data Resource (https://idr.openmicroscopy.org), (2) the dataset contains appropriate controls, (3) the data consists of 2D images of cultured cells, labelled with nuclear and actin markers, and are thus relatively easy to analyse, (4) the metric used in the published screen (proportion of Fascin in cell nucleus) is intuitive and should be easily grasped by life scientists.

Lawson et al. (2022) examined the effects on nuclear localisation of Fascin following treatment with a library of pharmacological compounds. The compounds are part of AstraZeneca's annotated screening library and, as such, their function is not known (https://openinnovation.astrazeneca.com/preclinical-research/target-identification). Here we selected one of these compounds, SN0212398523, for reanalysis; however, the purpose of this work is not to attempt to elucidate any of its biological or clinical roles. The aim here is to illustrate that, given an unknown treatment, measuring the purported effect of that treatment can be achieved with high confidence from a relatively small sample size. But, drawing conclusions about biological effects is difficult in the absence of appropriate controls.

We considered three channels in the images, highlighting cell nuclei (Hoescht), Fascin and actin; we used the actin channel as a proxy for the cell body (Fig. 1B). We composed a simple and openly

available pipeline using CellProfiler (Stirling et al., 2021) to quantify the proportion of Fascin-mScarlet that localised to the nuclei (L_F) of Fascin knockdown (KD) HeLa cells under different experimental conditions. This was achieved by dividing the mean nuclear Fascin signal intensity (L_{F_N}) by the sum of the mean nuclear and cytoplasmic Fascin signal intensity (L_{F_C}):

$$L_F = \frac{I_{F_N}}{I_{F_N} + I_{F_C}} \tag{1}$$

The results of analysing all cells in each sample using the above metric are presented in Table 1.

Table 1. Statistical properties of the data produced by our analysis pipeline across four different treatments. The metric being compared, nuclear Fascin localisation, is described in Equation 1

Treatment	Mean	Lower 95% CI of mean	Upper 95% CI of mean	Median	25th Percentile	75th Percentile	Standard Deviation	n_{cells}
Untreated	0.4649	0.4579	0.4718	0.5108	0.4177	0.5465	0.1306	1362
DMSO	0.4757	0.4685	0.4828	0.5241	0.4409	0.5544	0.1308	1299
SN0212398523	0.4882	0.4788	0.4977	0.5328	0.4479	0.5578	0.1398	844
Leptomycin b	0.5791	0.5704	0.5877	0.5693	0.5431	0.6262	0.1082	606

Treating Fascin KD HeLa cells with SN0212398523 causes an increase in localisation of Fascin to the cell nucleus

Let us begin by estimating the effect of treating the cells with SN0212398523, relative to untreated cells (negative control). If we were to compare the mean values for these two samples, it would appear that SN0212398523 causes an increase in Fascin localisation in the nucleus (Table 1). The size of the effect appears small (0.4882 versus 0.4649 - an increase of approximately 2.3%), but significant, given that the 95% confidence intervals (0.479 - 0.498 compared to 0.458 - 0.472) do not overlap. However, we should consider comparisons with appropriate controls before reaching any conclusions.

Treatment with DMSO also results in an increase in nuclear localisation of Fascin

Using untreated cells as a control when the drug is dissolved in DMSO can introduce bias, as, while unlikely, DMSO itself may influence cellular behaviour. A more appropriate comparison would be between drug-treated cells (dissolved in DMSO) and cells treated with DMSO alone. This ensures that

any observed effects are due to the drug itself and not the DMSO vehicle, providing a fair and accurate control.

Our results show that treatment with DMSO (vehicle control) leads to an apparent increase in nuclear Fascin localization relative to untreated cells (0.4757 versus 0.4649; Table 1), although the confidence intervals overlap (0.469 - 0.483 versus 0.458 - 0.472). As expected, DMSO as the vehicle control does not significantly affect Fascin localization to the cell nucleus, supported by the overlapping confidence intervals. However, quantified Fascin localization in SN0212398523-treated cells (0.479 - 0.498) was higher than in DMSO-treated cells (0.469 - 0.483; Table 1). DMSO may introduce a systematic technical variation that could influence the results for SN0212398523-treated cells. The observed effect in SN0212398523-treated cells is likely a combination of this technical variation, a biological component due to the active agent, and random variation. Therefore, additional controls are required to conclusively determine any biological significance.

However, while the observed effect of DMSO may not be biologically relevant, it is important to note that technical variation can still lead to detectable differences. In this scenario, DMSO-treated cells should not significantly differ from untreated cells, but unwanted differences (artifacts) may occur. Researchers must be aware that apparently significant differences can stem from random variation rather than biological effects. Indeed, these differences can be statistically significant, but this does not guarantee biological significance.

Treatment with Leptomycin B results in a far larger effect size

Leptomycin B inhibits nuclear export (Lawson et al., 2022), which should result in Fascin accumulating in the nucleus (positive control). Quantifying the proportion of Fascin localised in the nucleus in Leptomycin B-treated cells, we find that this effect is much larger than all other treatments (Table 1). This helps to put the other measured differences in context. For example, the difference in measured Fascin localisation between Leptomycin b-treated cells and SN0212398523 is 0.073 - 0.109, which is several multiples of the difference between SN0212398523-treated and DMSO-treated cells (0.0 - 0.029; Table 1).

However, it is important not to over-interpret this result, as it is based on a single experimental

replicate. There is a possibility that any of the sampled wells may be outliers and not truly representative of the treatment group due to natural variation or experimental error. Without additional replicates to validate the findings, the results may not be generalizable. In other words, the result might not be reproducible, and conclusions drawn from this single experiment may not hold true for other samples or under slightly different conditions. It is crucial to acknowledge and be aware of this limitation when interpreting results from single experiments.

Increasing sample size provides more confidence in statistical descriptors

It is possible to ascertain at what point data collection should cease by examining statistical properties as data is being collected. In Figure 2A, we show that the uncertainty in a particular statistical descriptor (inter-quartile range) decays exponentially with increasing sample size. In other words, as more data is collected, the error associated with a particular statistical parameter declines in a highly predictable manner.

This is further illustrated in Figure 2B, which shows how four commonly-used statistical parameters describing the untreated cells data (Table 1) vary as a random sample of the data is increased in size. We observe in our example a threshold effect around a sample size of approximately 250 - 300 cells, beyond which additional data yields diminishing returns in terms of statistical insight. This can be seen graphically in Figure2C-H - the distribution of the data for sample sizes between 200 and 500 looks very similar.

This evidence suggests a pragmatic approach to data collection: researchers should aim to gather enough data to reach statistical stability, where the key metrics of interest no longer change significantly with additional sampling. Once this level of statistical confidence is achieved — often visible through a plateau in the variability of statistical parameters — further data collection becomes less informative and may unnecessarily expend resources. This approach not only ensures efficient use of research resources but also guards against so-called p-value “hacking” (Amrhein et al., 2019; Miola and Miot, 2021).

It is important once again at this point to emphasize the difference between increasing the number of data points obtained from a single experiment and increasing the number of independent experiments. In the case of the former, as we discuss above, drawing additional samples (here, cells) from a single

experiment will have diminishing returns beyond a certain point - once we have a stable, representative estimate of the mean, for example, additional data may offer little new insight. This differs from collecting independent samples from multiple units, which enhances the generalisability and reliability of the results (Lord et al., 2020).

Increasing sample size has little influence on effect size

The number of cells in each sample in Table 1 is relatively large and the difference in means is relatively small. However, the data presented in Figure 2 suggests that we should have high confidence in the result provided by a random sample of approximately 200 - 300 cells from each treatment group.

We therefore explored the influence of randomly sub-sampling from these larger samples and examining the impact this has on the results (Figure 3A - F). For each of these random samples, the distribution shapes might differ, particularly for the smaller sample size of 50 cells (Figure 3A - C). It is also notable that, as we would expect, the 95% confidence intervals are narrower for larger sample sizes (Figure 3D - F). However, the overall trend in the result is virtually identical and the differences in measured effect are similar.

We further investigated this relationship between sample size (number of cells), random sampling, and effect size. We randomly sampled (with replacement) from the different datasets in Table 1 and once again measured the proportion of Fascin localised in the nucleus across all treatment conditions, using sample sizes ranging from 10 up to 500. We repeated this 100 times for each sample size to produce distributions of mean values (Figure 3G - I).

Let us first examine the influence of sample size and random sampling on the difference between Fascin localisation in untreated cells and cells treated with SN0212398523 (Figure 3G). For small sample sizes (< 50), there is a large amount of uncertainty in measured effect size (the difference in nuclear Fascin localisation between the two samples). But, above a sample size of approximately 50, the effect of random sampling is minimal and the result changes little. We can reasonably conclude that, in this particular experiment, for any sample of approximately 50 or more cells, the nuclear Fascin localisation will be higher in SN0212398523-treated cells than in untreated cells. Adding more cells to the samples will not alter this conclusion. However, it is not clear how biologically significant this

difference is.

We therefore repeated this analysis, but this time included the vehicle control, DMSO-treated cells (Figure 3H). Again, we can see that, above a certain threshold (approximately 250 - 300 cells per sample) there is little overlap between the uncertainty bands and we can reasonably conclude that nuclear Fascin localisation is higher in SN0212398523-treated cells than it is in DMSO-treated cells. But the fact that Figure 3H also shows that nuclear Fascin localisation is higher in DMSO-treated cells than it is in untreated cells should cause us to question the biological significance of any measured effect of SN0212398523, because we know that DMSO almost certainly does not promote nuclear localisation of Fascin. At this point, it is difficult to distinguish between technical variability (which likely accounts for the difference between untreated and DMSO-treated cells) and potential biological effects of SN0212398523. In other words, the higher level of nuclear Fascin we have measured in the presence of SN0212398523 might be a real biological effect, or it might just be a technical artifact, caused by random well-to-well variability. Therefore, additional controls are required to aid our interpreation.

Repeating the same analysis one more time with a positive control puts these differences into context (Figure 3I). There are two important points to highlight from these results. The first is that, while there is a measured increase in nuclear Fascin localisation in DMSO- and SN0212398523-treated cells, it is not biologically significant, relative to the increase in localisation measured in Leptomycin B-treated cells. The second is that, even for small sample sizes (< 50 cells), this result is apparent. Adding more cells into the samples only serves to refine the measured differences (that is, the uncertainty in the estimates of the means is reduced) - it does not alter the overall conclusion.

Effect size is relatively insensitive to random sampling when comparing multiple experiments

Up to this point, we have only considered data from one single “experiment”, or one replicate. To test the influence of random sampling and sample size when comparing multiple experiments, we analysed additional data from Lawson et al. (2022) (Figure 4), including more wells and two additional compounds. We found that increasing the number of cells analysed and averaged from 50 to 200 had a small influence on the result (Figure 4A & B). The distribution of data points changes slightly, but the

overall conclusion does not change - Leptomycin B has a far larger effect on nuclear Fascin localisation than any of the treatments being tested. Consistent with other data presented up to this point, we found that increasing the number of cells in each sample beyond 200 had little effect - we obtained an almost identical result when 500 cells were randomly sampled (Figure 4C). One could employ an ANOVA test at this point to add additional weight to the result. But given that all box and whiskers plots overlap, except for those associated with Leptomycin B, no statistical test should be necessary to conclude that Leptomycin B has a far more significant biological effect than any other treatment.

There is an important point to be made at this juncture regarding the misuse or misinterpretation of p-values. It is not unusual for data similar to that presented in Table 1 to be presented as a swarm plot, annotated with p-values, where each cell is treated as an independent data point (Supplemental Figure 1). For examples, see Cai et al. (2024), Figure 1 D & E or Buglak et al. (2024), Figure 1 G - I. Without getting into a detailed discussion of what constitutes a technical or biological replicate, it is not unreasonable to state that different cells within the same dish/sample all subjected to the same treatment are *not* replicates (Zweifach, 2024) - they constitute one single experiment. As such, subjecting such data to statistical tests makes little sense. Regardless, blindly relying on the p-values presented in Supplemental Figure 1 might lead us to conclude that SN0212398523 has a highly significant influence on nuclear Fascin localisation. However, all of the data we have presented up to this point suggests that this is not the case.

We know that significant p-values can occur by chance and we know that the probability of a significant p-value emerging increases as the number of data points increases (so-called “p-hacking”; Lord et al. (2020)). Given the large number of datapoints presented in Supplementary Figure 1 (606 - 1362 per condition), it is not surprising that each distribution is statistically significantly different from all others. But we also know that, biologically, most of these differences are insignificant. As such, one must be cautious about equating statistical significance with biological significance.

Recommendations

Our analysis has highlighted the limitations of over-interpreting small treatment effects, particularly in the absence of appropriate controls. Biological significance can only be conclusively inferred when

effect sizes are measured against controls that deliver biologically meaningful and measurable effects. Relying instead on other metrics, such as p-values, to distinguish between significant and insignificant results can lead to misconceptions about the biological importance of findings. We propose several strategies to help researchers develop a more nuanced understanding of statistical results and to encourage practices that prioritise biological relevance over arbitrary statistical thresholds (Box 1).

In practice, experiments often default to the simplest form of control: comparing treated samples against untreated ones. However, this binary approach can overlook nuanced biological responses and fail to account for variables such as the effects of the treatment process itself or batch effects. Moreover, the practice of reusing control samples across different experimental repeats—a shortcut that compromises the integrity of the data—highlights the need for a more rigorous approach to control selection.

Conclusion

Our findings emphasise the importance of using appropriate controls and focusing on effect sizes rather than, for example, p-values. By considering biological relevance over arbitrary thresholds of significance, researchers can derive more robust and reproducible conclusions. These recommendations aim to guide future bioimage analysis toward more meaningful interpretations and efficient experimental designs.

We caution against solely relying on statistical tests for interpreting bioimage data. Collaborating with statisticians for data analysis is highly recommended. More importantly, we stress the use of appropriate controls to discern biological meaningfulness. Although this might introduce greater uncertainty in research outcomes, it is a necessary step towards reducing irreproducible “highly significant” results, thereby advancing the credibility and robustness of scientific research.

Box 1: Recommendations.

Consult a statistician prior to collecting any data
The frequent misuse of p-values in life sciences, highlighted by numerous statisticians (Amrhein et al., 2019; Nuzzo, 2014; Wasserstein et al., 2016), underscores the need for a foundational shift in how research is approached. Consulting a statistician from the initial stages of research design is not merely a formality but a crucial step in ensuring the robustness and reliability of scientific findings. This early collaboration can aid in selecting appropriate statistical tests, planning effective experiments, and avoiding common pitfalls associated with p-value interpretation.

Use appropriate controls
The cornerstone of any robust experimental design is the use of well-chosen control conditions. This principle might appear fundamental, yet the frequent encounters with inadequately controlled experiments in bioimage analysis underscore its critical importance. Effective controls are essential not only for validating the experimental setup but also for ensuring that the observed effects are attributable to the experimental variables rather than external factors or inherent variability.

Cease data collection once statistical properties of the sample are known
But what of cases where access to a statistician is not available and no controls for a particular experimental scenario are known? In Figure 2A, we show that the uncertainty in a particular statistical descriptor (inter-quartile range) decays exponentially with increasing sample size. In other words, as more data is collected, the error associated with a particular statistical parameter declines in a highly predictable manner. It is therefore recommended that, so far is possible, the statistical properties of the data are interrogated as the data is being collected. Once those statistical properties are known with a reasonable degree of confidence, data collection should cease.

Materials & Methods

Image Data

The image data used in this study, first published by Lawson et al. (2022), is freely available on the Image Data Resource (https://idr.openmicroscopy.org) under accession number IDR0139. We purposefully selected four wells (J05, O02, E22, L08) from plate number 1093711385 for use in this study. Well O02 (DMSO) was chosen (on the basis of our analysis outputs) as the measure of nuclear Fascin in this well was slightly higher than in untreated cells (J05). Similarly, well E22 (SN0212398523) was chosen as nuclear Fascin was slightly higher in this well than in DMSO-treated cells (O02).

To simulate experimental repeats (Figure 4), we randomly selected additional wells in the same plate for untreated, DMSO-treated and Leptomycin b-treated cells. Because there was only one well with SN0212398523-treated cells, we chose different treatments that yielded similar effects on nuclear Fascin localisation for comparison: SN1054616339 and SN1066932540.

Image Analysis

All image analysis was performed with a custom image analysis pipeline implemented in CellProfiler (Stirling et al., 2021), which is freely available on Github, along with all other code associated with this study (https://github.com/djpbarry/Enhancing-Reproducibility). Briefly, the Hoechst and actin channel of each image were smoothed using Gaussian filtering. Nuclei were then identified in the smoothed Hoechst channel and cell bodies in the smoothed actin channel using minimum cross-entropy thresholding. Measurements of the Fascin signal were then made within these segmented objects. To obtain the percentage of nuclear Fascin used as quantitative output, the Fascin signal intensity in the nuclear region was divided by the total cell area (nucleus + cell body).

Data Analysis

All data analysis was performed in Python using the numpy (Harris et al., 2020), pandas (McKinney, 2010), scikit-posthocs (Terpilowski, 2019) and scipy (Virtanen et al., 2020) packages. Plots were generated using matplotlib (Hunter, 2007) and seaborn (Waskom, 2021). A companion Jupyter notebook, inclusive of instructions on how to set it up locally using Python environments, is available on GitHub (https://github.com/djpbarry/Enhancing-Reproducibility).

Individual data points in swarm plots represent individual cells, unless otherwise stated. Box-and-whiskers plots show median (central line), inter-quartile range (box edges) and 1.5 times the inter-quartile range (whiskers). Red error bars (where present) show the 95% confidence interval of the mean. Effect size (d) is calculated according to Cohen (2013) as:

$$d = \frac{\bar{x}_t - \bar{x}_u}{s_u} \tag{2}$$

where $\bar{x}_t$ and $\bar{x}_u$ represent the mean of the treated and untreated cell samples and s_u is the standard deviation of the untreated cell sample. A notebook that reproduces all the plots in this paper is freely accessible online (https://github.com/djpbarry/Enhancing-Reproducibility).

Funding

This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001157, FC001999), the UK Medical Research Council (FC001157, FC001999) and the Wellcome Trust (FC001157, FC001999). SM has received funding from the Biotechnology and Biological Sciences Research Council (BBSRC, grant number BB/V006169/1).

Code Availability

All code used to generate the figures in this manuscript is available online (https://github.com/djpbarry/Enhancing-Reproducibility).

Acknowledgements

The authors would like to thank Maddy Parsons for constructive feedback on the manuscript.

References

Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., and Moher, D. (2024) Biomedical researchers' perspectives on the reproducibility of research. *PLOS Biology 22*, 1–15.

Lord, S. J., Velle, K. B., Mullins, R. D., and Fritz-Laylin, L. K. (2020) SuperPlots: Communicating reproducibility and variability in cell biology. *Journal of Cell Biology 219*, e202001064.

Krzywinski, M., and Altman, N. (2013) Importance of being uncertain. *Nature Methods 10*, 809–810.

Zweifach, A. (2024) Determining how many cells to average for statistical testing of microscopy experiments. *Journal of Cell Biology 223*, e202401074.

Fisher, R., and Prance, G. *The Design of Experiments*; Hafner Press, 1974.

Lawson, C. D., Peel, S., Jayo, A., Corrigan, A., Iyer, P., Dalrymple, M. B., Marsh, R. J., Cox, S., Audenhove, I. V., Gettemans, J., and Parsons, M. (2022) Nuclear fascin regulates cancer cell survival. *eLife 11*, e79283.

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., and Goodman, A. (2021) CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics 22*, 433.

Amrhein, V., Greenland, S., and McShane, B. (2019) Scientists rise up against statistical significance. *Nature 567*, 305–307.

Miola, A. C., and Miot, H. A. (2021) P-value and effect-size in clinical and experimental studies. *Jornal Vascular Brasileiro 20*.

Cai, R., Bai, P., Quan, M., Ding, Y., Wei, W., Liu, C., Yang, A., Xiong, Z., Li, G., Li, B., Deng, Y., Tian, R., Zhao, Y. G., Wu, C., and Sun, Y. (2024) Migfilin promotes autophagic flux through direct interaction with SNAP29 and Vamp8. *Journal of Cell Biology 223*, e202312119.

Buglak, D. B., Holmes, K. H. M., Galletta, B. J., and Rusan, N. M. (2024) The proximal centriole-like structure maintains nucleus–centriole architecture in sperm. *Journal of Cell Science 137*, jcs262311.

Nuzzo, R. (2014) Scientific method: Statistical errors. *Nature 506*, 150–152.

Wasserstein, R. L., Lazar, N. A., Wasserstein, R. L., Lazar, N. A., and Wasserstein, R. L. (2016) The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician 70*, 129–133.

Harris, C. R. et al. (2020) Array programming with NumPy. *Nature 585*, 357–362.

McKinney, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010; pp 56 – 61.

Terpilowski, M. A. (2019) scikit-posthocs: Pairwise multiple comparison tests in Python. *Journal of Open Source Software 4*, 1169.

Virtanen, P. et al. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17*, 261–272.

Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering 9*, 90–95.

Waskom, M. L. (2021) seaborn: statistical data visualization. *Journal of Open Source Software 6*, 3021.

Cohen, J. *Statistical power analysis for the behavioral sciences*; Routledge, 2013.

Figure Legends

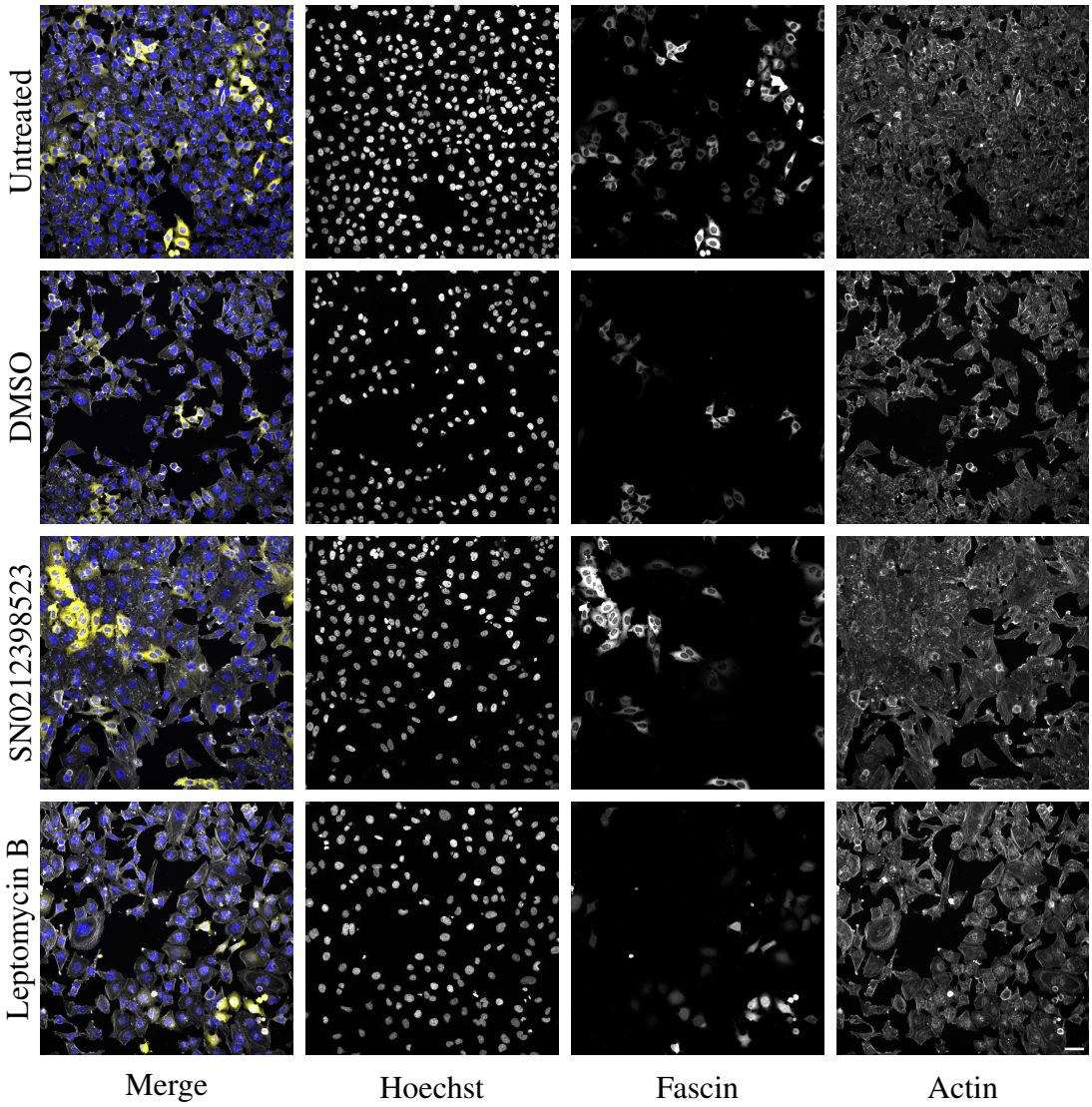
Fig 1. Overview of data and analysis workflow. A. Images of Fascin knockdown (KD) HeLa cells downloaded from the Image Data Resource (accession number IDR0139), originally published by (Lawson et al., 2022). The selected dataset includes a treated sample (pharmacological compound SN0212398523) and three controls (untreated or negative control, DMSO or vehicle control, and Leptomycin B or positive control). The images contain three channels, Hoechst (nuclei), Fascin and actin. Scale bar 50 μm. B. Schematic illustrating the calculation of Fascin intensity. Fascin localisation is calculated as the mean intensity of Fascin in the nucleus divided by the mean intensity of Fascin in the whole cell (nucleus plus cytoplasm). The CellProfiler (Stirling et al., 2021) pipeline used to perform these calculations is openly available in the GitHub repo accompanying this paper (https://github.com/djpbarry/Enhancing-Reproducibility).

Fig 2. Variation in statistical descriptors stabilises above a certain sample size. A. The difference between the maximum and minimum estimate of the inter-quartile range follows a predictable exponential decay as sample size increases. Results are based on 100 random samples at each sample size. B. Statistical properties of cumulative random samples drawn from the population of untreated cells (Table 1). For sample sizes below approximately 200, there is substantial variability in the measured statistical properties. However, above a sample size of approximately 250 - 300, the statistical properties are relatively stable. C - H. Histograms of the random samples described in 'B'. Solid vertical green and blue lines depict the width of the inter-quartile range (grey shaded area). The dashed vertical red line depicts the median.

Fig 3. Effect sizes are relatively insensitive to random sampling and sample size. A - F. Three different random samples of 50 (A - C) or 200 (D - F) cells for each treatment - each data point represents a single cell. In all cases, the measured Fascin localisation is similar. G. The influence of random sampling and sample size on the difference in Fascin localisation between treated and untreated cells. For each sample size, random sampling was performed 100 times. The solid line represents the median of those 100 random samples and shaded bands represent the width of the inter-quartile range. Above a sample size of approximately 250, the effect of random sampling on the result is minimal. We can confidently state that nuclear Fascin localisation in treated cells is approximately 0.02 higher than in untreated cells, but in the absence of controls, interpreting the significance of this value is difficult. H. A repeat of the bootstrapping analysis in 'G', but this time including the effect size of measured nuclear Fascin in DMSO-treated cells as a vehicle control when compared to untreated cells. I. One more repeat of the bootstrapping analysis in 'G' and 'H', this time including the positive control, Leptomycin B. Effect sizes are calculated according to Equation 2.

Fig 4. Effect size is relatively insensitive to random sampling when comparing multiple experiments. A - C. Nuclear Fascin localisation averaged across samples of cells subjected to the indicated treatments. Each large dot represents a technical replicate and is the average of 50 (A), 200 (B) or 500 (C) randomly selected cells within a well. Each small dot represents a single cell. Colours correspond to a particular experimental replicate.

A.


Untreated
DMSO
SN0212398523
Leptomycin B
Merge
Hoechst
Fascin
Actin

B.

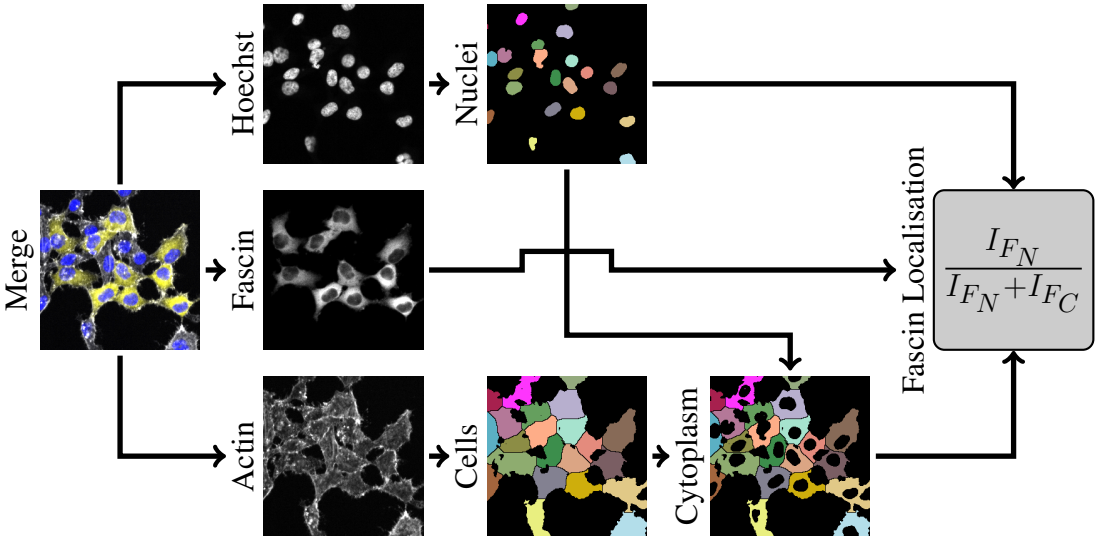

Merge
Hoechst
Fascin
Actin
Nuclei
Cells
Cytoplasm
Fascin Localisation
$\frac{I_{F_N}}{I_{F_N}+I_{F_C}}$

Figure 1

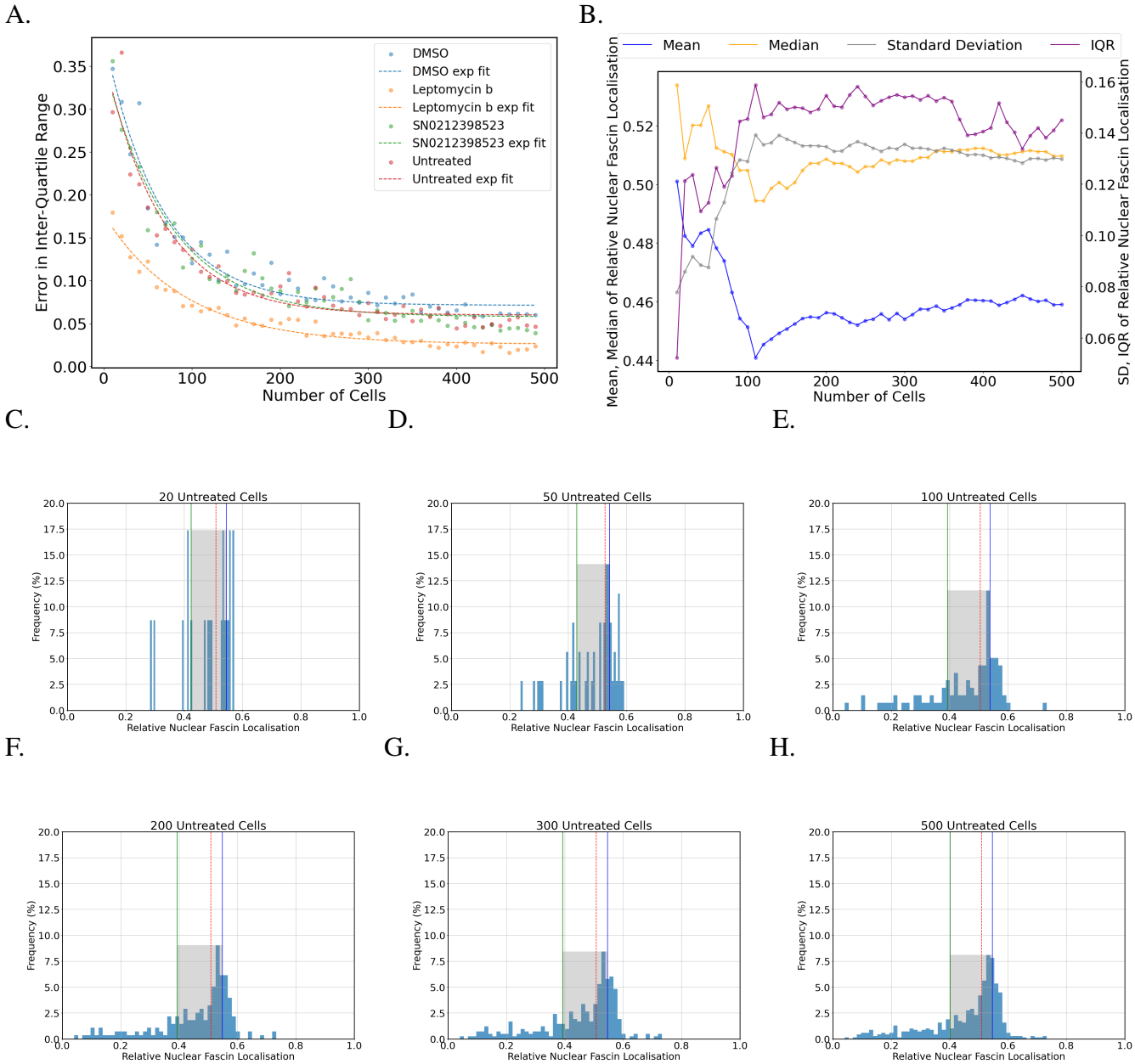

A.
B.
DMSO
DMSO exp fit
Leptomycin b
Leptomycin b exp fit
SN0212398523
SN0212398523 exp fit
Untreated
Untreated exp fit
Error in Inter-Quartile Range
Number of Cells
Mean
Median
Standard Deviation
IQR
Mean, Median of Relative Nuclear Fascin Localisation
SD, IQR of Relative Nuclear Fascin Localisation
Number of Cells
C.
D.
E.
20 Untreated Cells
50 Untreated Cells
100 Untreated Cells
Frequency (%)
Relative Nuclear Fascin Localisation
F.
G.
H.
200 Untreated Cells
300 Untreated Cells
500 Untreated Cells
Frequency (%)
Relative Nuclear Fascin Localisation

Figure 2

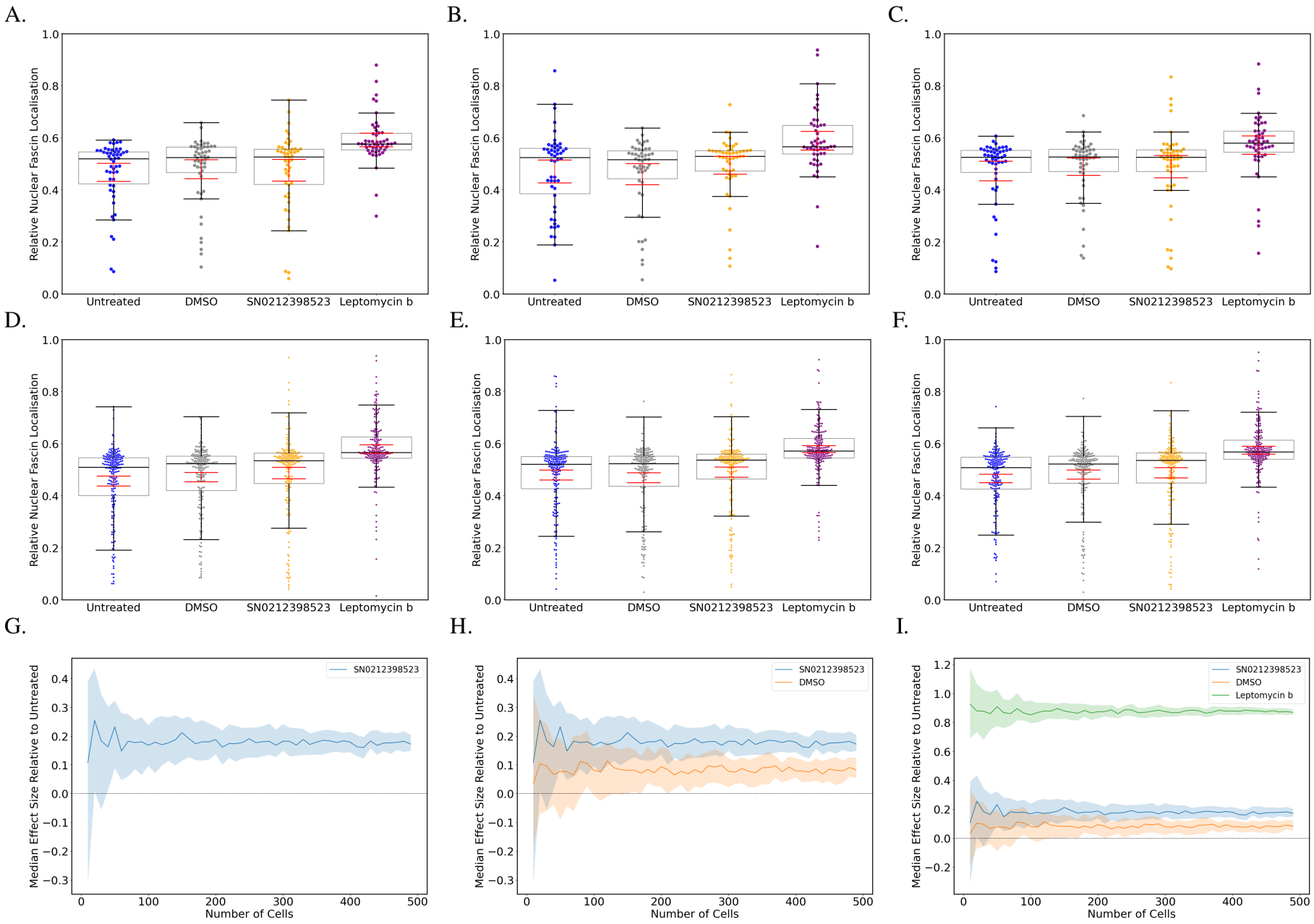
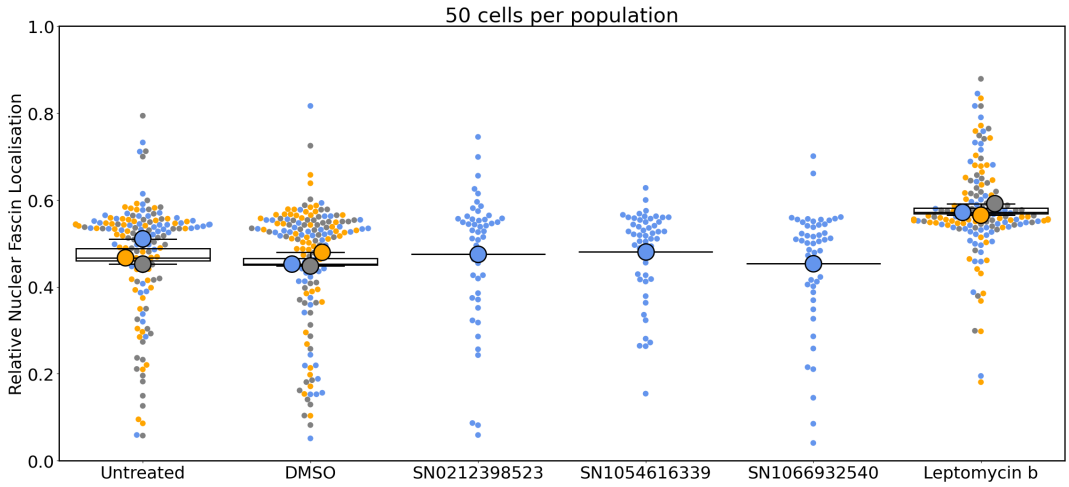
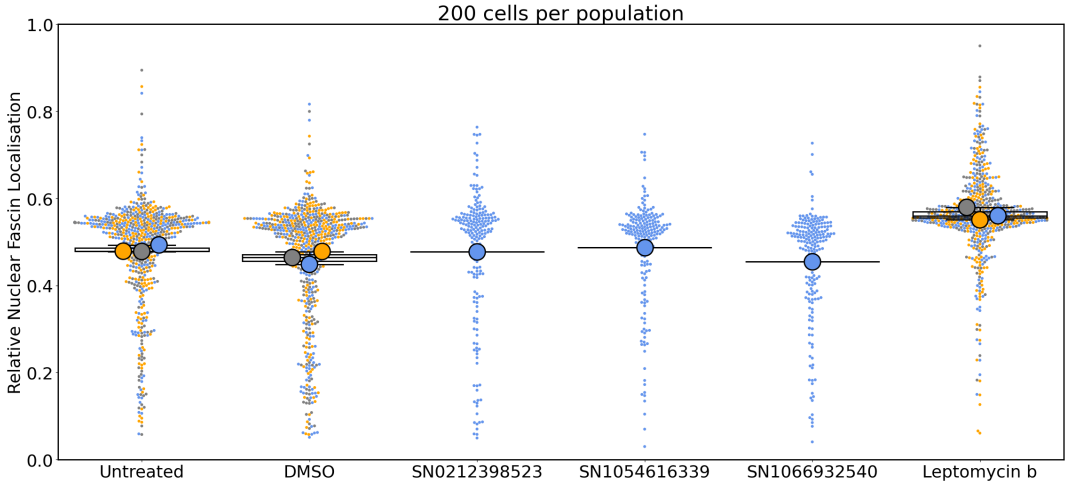

A.
B.
C.
D.
E.
F.
G.
H.
I.
Relative Nuclear Fascin Localisation
Untreated
DMSO
SN0212398523
Leptomycin b
Median Effect Size Relative to Untreated
Number of Cells

Figure 3

A.


50 cells per population
Relative Nuclear Fascin Localisation
1.0
0.8
0.6
0.4
0.2
0.0
Untreated
DMSO
SN0212398523
SN1054616339
SN1066932540
Leptomycin b

B.


200 cells per population
Relative Nuclear Fascin Localisation
1.0
0.8
0.6
0.4
0.2
0.0
Untreated
DMSO
SN0212398523
SN1054616339
SN1066932540
Leptomycin b

C.

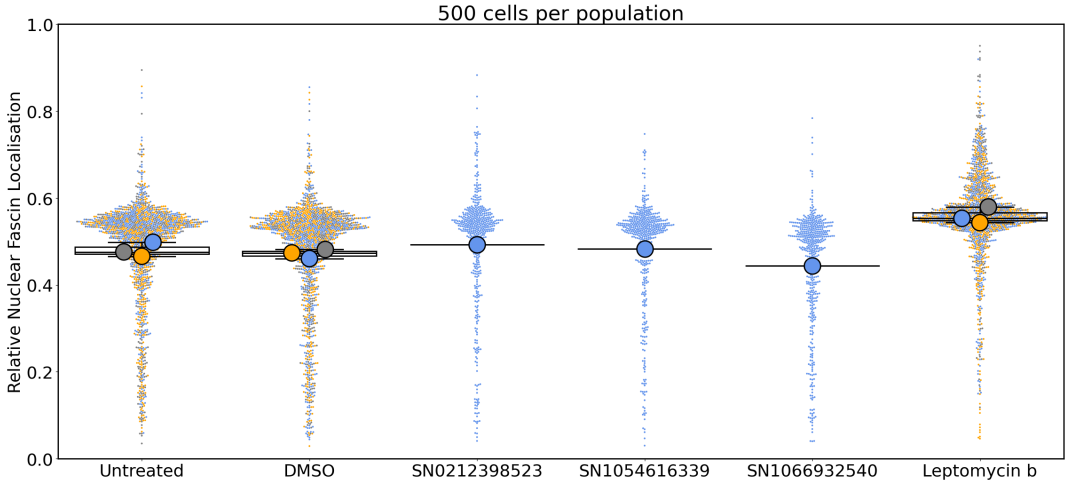
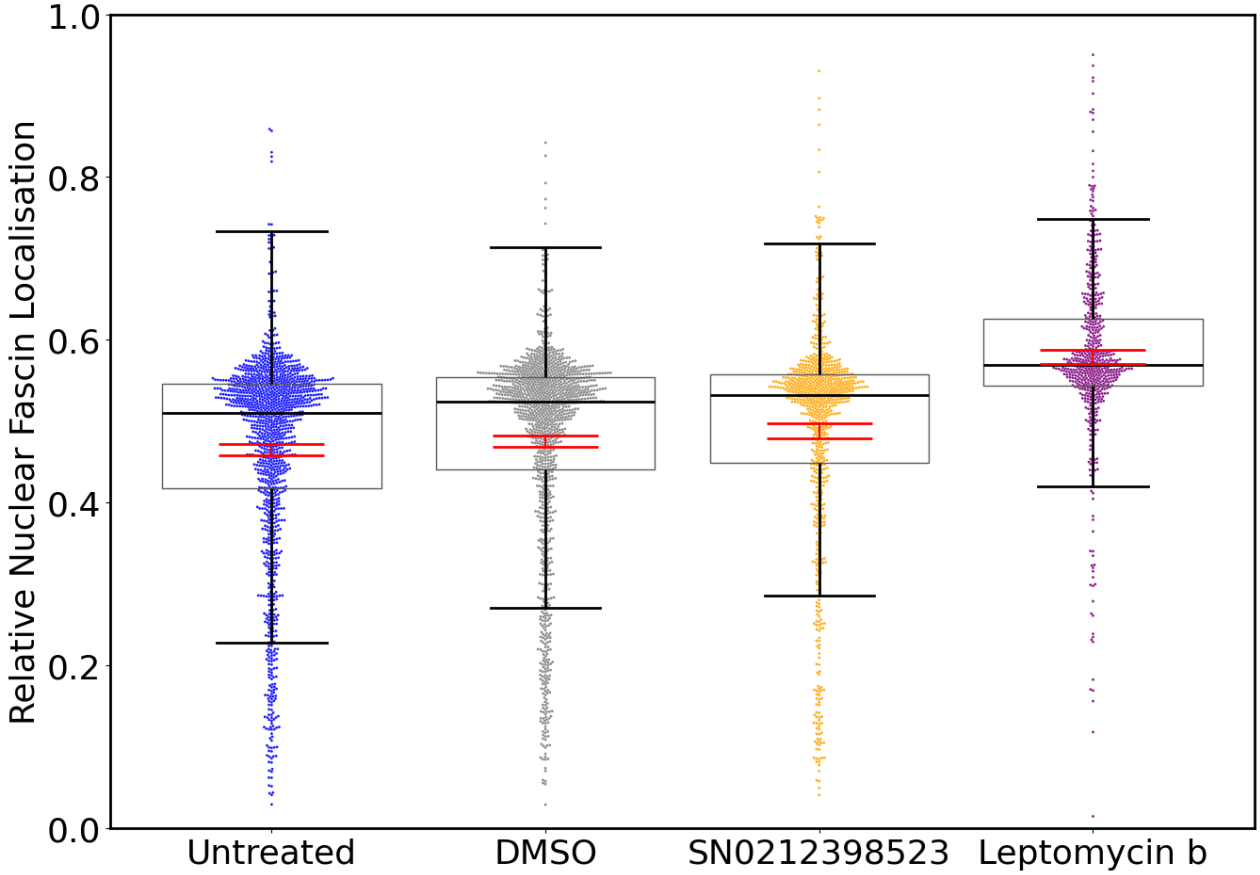

500 cells per population
Relative Nuclear Fascin Localisation
1.0
0.8
0.6
0.4
0.2
0.0
Untreated
DMSO
SN0212398523
SN1054616339
SN1066932540
Leptomycin b

Figure 4



Supplemental Figure 1