



A Permutation-Based Correction for Multiple Testing in High-Content Imaging Screens

Marcus Ingram^{1*}, Kevin Grigoryan², Mateo Afolabi³

¹Department of Life Sciences, Periyar School of Life Sciences, Kenya

²Department of Life Sciences, Yangtze Center for Life Sciences, Thailand

³Department of Life Sciences, Kaveri School of Life Sciences, Romania

Article Info

Volume 1, Issue 1, 2026

[doi: 10.62715/ijbb.2026.v1.i1.pp15-20](https://doi.org/10.62715/ijbb.2026.v1.i1.pp15-20)

Received: 30 December 2025

Accepted: 23 February 2026

Published: 09 May 2026

Abstract

A high-content imaging screen compares many treatments with negative controls on hundreds or thousands of morphological features, so the number of hypothesis tests is the product of two large numbers. The features are strongly correlated, the replicate wells are few, and the measurements carry plate and position effects. Standard corrections are poorly matched to these conditions: Bonferroni-type procedures ignore the correlation and discard power, while false discovery rate procedures rest on dependence assumptions and on p-values that are unreliable when computed from a handful of wells. This article proposes a permutation-based correction designed around the layout of a screen. Whole wells are relabelled within plates, which respects the plate as a block and keeps each well's feature vector intact, and the maximum studentised statistic across all tests is recorded for each relabelling. The resulting reference distribution yields p-values adjusted for the family-wise error rate that reflect the actual dependence among features. The procedure is described in seven steps with the assumption behind each, together with a variant that draws the null from control wells alone and a built-in check on held-out controls. It is a proposal assembled from established resampling theory and has not been validated on screening data here.

Keywords: High-content screening, Multiple testing, Permutation test, Family-wise error rate, Image-based profiling, Max-statistic

© 2026 Marcus Ingram et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Automated microscopy has turned the cell-based screen from a single-readout assay into a high-dimensional measurement. Software such as CellProfiler segments every cell in every image and returns measurements of size, shape, intensity and texture for each stained compartment (Carpenter *et al.*, 2006), and multiplexed staining protocols such as Cell Painting are designed to make that description as broad as possible (Bray *et al.*, 2016). A screen of a compound library or a gene perturbation set therefore produces, for each treatment, not one number to be compared with control but a vector of several hundred to a few thousand.

* Corresponding author: Marcus Ingram, Department of Life Sciences, Periyar School of Life Sciences, Kenya. E-mail: marcus.ingram@periyar.edu

© 2026 Marcus Ingram et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Statistical practice has not kept pace with this change. Hit selection in screening grew up around single readouts and plate-quality measures such as the Z'-factor (Zhang et al., 1999), and although the case for formal inference in screening was made some time ago (Malo et al., 2006), the question of how to control error rates across treatments and features together is often handled by an arbitrary threshold or by a correction chosen for convenience. This article proposes a correction built on permutation of well labels. It describes the procedure and its rationale. It does not report an application to data, and its performance remains to be demonstrated.

2. Background

The design of the procedure follows from three observations about imaging screens and about the corrections usually applied to them.

2.1. The structure of an imaging screen

Cells are grown in multi-well plates. Each treatment occupies a small number of wells, commonly spread over replicate plates, and each plate carries a larger number of negative-control wells and some positive controls. Per-cell measurements are summarised to one value per feature per well, typically a median, and the well is the unit that received the treatment. The many cells within it are subsamples, not replicates (Lazic et al., 2018). Wells on one plate share reagents, incubation and imaging conditions, and systematic gradients across rows and columns are common, so normalisation against on-plate controls is a routine first step (Birmingham et al., 2009; Caicedo et al., 2017).

2.2. Why the usual corrections fit poorly

Testing every feature for every treatment creates a very large family of hypotheses. The Bonferroni correction and its step-down refinement by Holm (1979) control the probability of any false positive under arbitrary dependence, but they treat the tests as if each carried independent risk. Morphological features are far from independent: many are near-duplicates computed at different scales or in different channels. The effective number of tests is much smaller than the nominal number, and the correction is correspondingly too severe.

The false discovery rate procedure of Benjamini and Hochberg (1995) is less conservative and is the common alternative. Its guarantee was proved for independent tests and extended to a particular form of positive dependence, with a more stringent variant required for arbitrary dependence (Benjamini and Yekutieli, 2001). Feature sets contain both positive and negative correlations, so which case applies is unclear. Both families of correction also take the unadjusted p-values as given. With two to four wells per treatment and feature distributions that are often skewed or heavy-tailed, p-values from the t-distribution are unreliable in exactly the far tail where a stringent threshold operates.

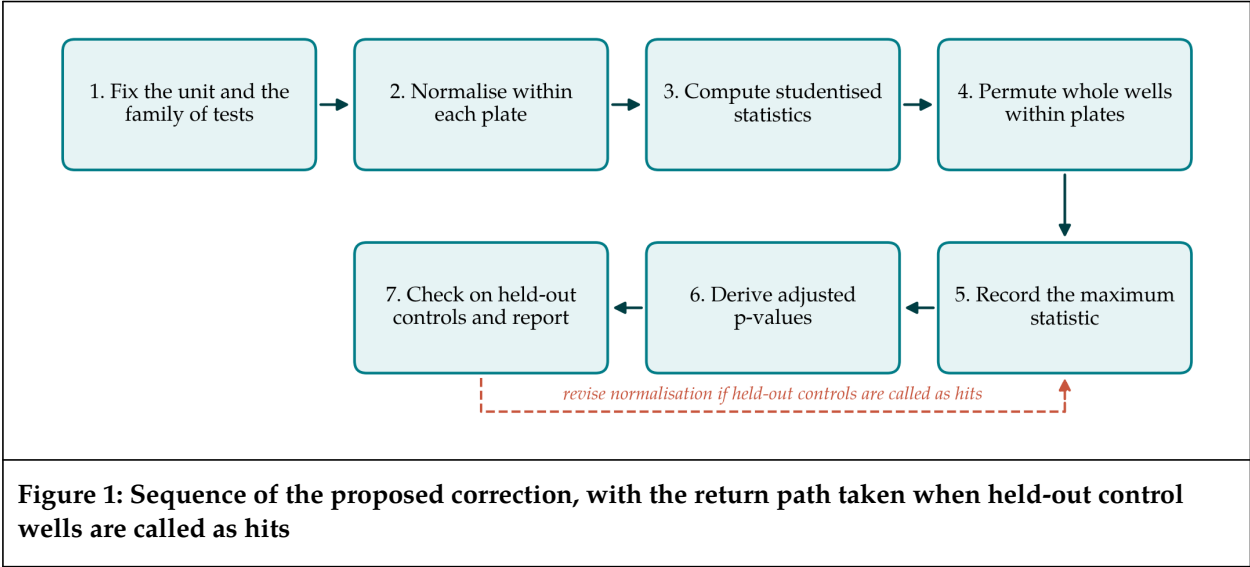
2.3. Resampling-based adjustment

Permutation tests obtain the null distribution of a statistic from the data by recomputing it under relabellings that would be equally likely if the null hypothesis held (Good, 2005). Westfall and Young (1993) showed how to extend the idea to many tests at once: if whole observation vectors are relabelled, the joint distribution of all the test statistics is reproduced, correlation included, and the distribution of the largest statistic gives adjusted p-values that control the family-wise error rate. The method became standard in microarray analysis (Ge et al., 2003) and in neuroimaging, where the maximum statistic over an image of correlated voxels is the usual device (Nichols and Holmes, 2002). The imaging screen presents the same problem in a different layout, and the proposal below adapts the same solution to it.

3. The Proposed Procedure

The procedure has seven steps. Table 1 lists them with the assumption or safeguard attached to each, and Figure 1 shows the sequence. It takes as input a table of well-level feature values with plate, well position and treatment labels.

Table 1: Steps of the proposed permutation-based correction and the assumption or safeguard attached to each		
Step	Action	Assumption or safeguard
1. Unit and family	Take the well as the unit; declare which treatment-feature tests form the family	Family fixed before the data are examined
2. Normalise	Scale each feature against negative controls on the same plate; remove positional trends	Wells on a plate comparable after correction
3. Statistic	Compute a studentised difference from control for each treatment and feature	Statistics on a common scale across features
4. Permute	Relabel whole wells at random within each plate, many times	Exchangeability within plates under the null
5. Maximum	For each relabelling, keep the largest absolute statistic in the family	Feature vectors kept intact, so dependence is preserved
6. Adjust	Compare each observed statistic with the distribution of maxima	Add-one estimator, so no p-value equals zero
7. Check and report	Run held-out controls as mock treatments; document all settings	Mock treatments should not be called as hits



3.1. Step 1: Fix the unit and the family

The well is the experimental unit and all resampling is done on wells. The family of tests must be declared in advance, since the correction is only meaningful relative to it. Two choices are natural. If the aim is to say which features change under which treatment, the family is every treatment-feature pair. If the aim is only to find active treatments, each treatment can be given a single multivariate statistic and the family is the set of treatments. Redundant features may be removed beforehand to ease computation, although the correction does not require it.

3.2. Step 2: Normalise within plates

Each feature in each well is centred on the median of the negative-control wells on its plate and divided by their median absolute deviation. Where the controls reveal row, column or edge trends, these are estimated and removed. This step matters more here than in a parametric analysis, because the permutation in Step 4 treats wells on a plate as interchangeable, and uncorrected positional effects would break that assumption.

3.3. Step 3: Compute studentised statistics

For each treatment and feature the mean of the normalised replicate values is divided by an estimate of its standard error. Studentising is essential when a maximum is to be taken, since otherwise the features with the greatest variance dominate the reference distribution and the rest lose power. Because variance estimates from two or three wells are unstable, the standard error is better taken from the pooled spread of the control wells for that feature than from the treatment wells alone.

3.4. Step 4: Permute whole wells within plates

The labels that say which treatment or control each well received are shuffled among the wells of a plate, separately for each plate, and the statistics of Step 3 are recomputed for all tests. Two restrictions define the scheme. Shuffling is confined to plates, so that the plate acts as an exchangeability block in the sense used for permutation inference in the general linear model (Winkler et al., 2014). And a well keeps its entire feature vector when its label moves, so that every relabelled data set has the same correlation among features as the real one. Positive-control wells are excluded from the shuffle.

3.5. Step 5: Record the maximum statistic

From each relabelled data set only one number is kept: the largest absolute statistic across the whole family. The collection of these maxima, one per relabelling, estimates the distribution of the most extreme result that chance alone produces in a screen of this size, layout and correlation structure. Highly redundant features add little to the maximum, which is how the method adapts to the effective number of tests without that number ever being estimated.

3.6. Step 6: Derive adjusted p-values

The adjusted p-value of an observed statistic is the proportion of permutation maxima that equal or exceed it, computed with one added to both numerator and denominator so that it can never be zero (Phipson and Smyth, 2010). Tests whose adjusted p-value falls below the chosen level are declared hits, with the family-wise error rate controlled at that level. A step-down version, in which the maximum is taken over successively smaller sets of hypotheses ordered by their observed statistics, rejects at least as many hypotheses under the same guarantee (Westfall and Young, 1993).

3.7. Step 7: Check on held-out controls and report

Before the analysis, a portion of the negative-control wells on each plate is set aside and labelled as mock treatments with the same replicate layout as real ones. They pass through every step. Since they are known to be null, calls among them at a rate above the nominal level indicate that normalisation has left structure the permutation does not account for, and the analyst returns to Step 2. The report should state the family, the unit, the permutation scheme, the number of relabellings, the random seed, the software version and the outcome of this check.

4. Practical Considerations

The number of relabellings sets the smallest adjusted p-value that can be obtained, which is the reciprocal of that number plus one. Family-wise control at conventional levels therefore needs only a few thousand relabellings, far fewer than would be required to push unadjusted permutation p-values below a Bonferroni threshold. Where finer resolution in the tail is wanted, the upper tail of the maxima can be approximated by a generalised Pareto distribution (Knijnenburg et al., 2009). Each relabelling involves only means over small groups of wells, so the computation is matrix arithmetic and parallelises over relabellings.

A variant addresses a known weakness. Shuffling all wells assumes that no treatment has any effect, and in a screen with many active treatments the strongly perturbed wells are mixed into the reference distribution and widen it, which makes the procedure conservative. If plates carry enough negative controls, the reference distribution can instead be built from controls alone, by repeatedly drawing sets of control wells to stand in for each treatment at its plate locations. The variant assumes that an inactive

treatment behaves like a control well, which is the working assumption of the screen itself. Investigators who prefer to control the false discovery rate can use the same relabellings to estimate it empirically, though the guarantee is then of a different kind.

5. Limitations and What Remains to Be Shown

The procedure is proposed and has not been validated in this article. Its guarantee rests on exchangeability of wells within a plate under the null hypothesis, and normalisation can only approximate that condition. Strong control of the family-wise error rate by the maximum-statistic method further requires that the joint distribution of statistics for the true null hypotheses does not depend on which other hypotheses are false, the condition Westfall and Young (1993) called subset pivotality. Compound effects on cell number, which alter the precision of every feature in a well, could violate it.

Three matters would need empirical study before the method could be recommended without reservation. The first is error control: whether, on plates containing only negative controls, the rate of false calls matches the nominal level across realistic plate layouts. The second is power relative to Holm and to false discovery rate procedures on data with known actives. The third is sensitivity to the normalisation method and to the number of control wells. Public image-based profiling data sets with annotated compounds would allow all three to be examined.

6. Conclusion

Multiple testing in imaging screens is a problem of dependence as much as of number. Relabelling whole wells within plates and referring each statistic to the distribution of the permutation maximum offers family-wise error control that uses the correlation among features instead of ignoring it, and that does not lean on distributional assumptions that a few replicate wells cannot support. The seven steps set out here make the assumptions explicit and include a check that can expose their failure. Whether the gain over standard corrections is large enough to justify the added computation is a question for empirical evaluation.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Benjamini, Y. and Hochberg, Y. (1995). [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Benjamini, Y. and Yekutieli, D. (2001). [The control of the false discovery rate in multiple testing under dependency](#). *The Annals of Statistics*, 29(4): 1165-1188. <https://doi.org/10.1214/aos/1013699998>
- Birmingham, A., Selfors, L.M., Forster, T., Wrobel, D., Kennedy, C.J., Shanks, E. et al. (2009). [Statistical methods for analysis of high-throughput RNA interference screens](#). *Nature Methods*, 6(8): 569-575. <https://doi.org/10.1038/nmeth.1351>
- Bray, M.-A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C. and Carpenter, A.E. (2016). [Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes](#). *Nature Protocols*, 11(9): 1757-1774. <https://doi.org/10.1038/nprot.2016.105>
- Caicedo, J.C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C. et al. (2017). [Data-analysis strategies for image-based cell profiling](#). *Nature Methods*, 14(9): 849-863. <https://doi.org/10.1038/nmeth.4397>

- Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P. and Sabatini, D.M. (2006). [CellProfiler: Image analysis software for identifying and quantifying cell phenotypes](#). *Genome Biology*, 7(10): R100. <https://doi.org/10.1186/gb-2006-7-10-r100>
- Ge, Y., Dudoit, S. and Speed, T.P. (2003). [Resampling-based multiple testing for microarray data analysis](#). *Test*, 12(1): 1-77.
- Good, P. (2005). [Permutation, Parametric and Bootstrap Tests of Hypotheses](#), 3rd edition. Springer, New York.
- Holm, S. (1979). [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2): 65-70.
- Knijnenburg, T.A., Wessels, L.F.A., Reinders, M.J.T. and Shmulevich, I. (2009). [Fewer permutations, more accurate P-values](#). *Bioinformatics*, 25(12): i161-i168. <https://doi.org/10.1093/bioinformatics/btp211>
- Lazic, S.E., Clarke-Williams, C.J. and Munafò, M.R. (2018). [What exactly is 'N' in cell culture and animal experiments?](#) *PLoS Biology*, 16(4): e2005282. <https://doi.org/10.1371/journal.pbio.2005282>
- Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J. and Nadon, R. (2006). [Statistical practice in high-throughput screening data analysis](#). *Nature Biotechnology*, 24(2): 167-175. <https://doi.org/10.1038/nbt1186>
- Nichols, T.E. and Holmes, A.P. (2002). [Nonparametric permutation tests for functional neuroimaging: A primer with examples](#). *Human Brain Mapping*, 15(1): 1-25. <https://doi.org/10.1002/hbm.1058>
- Phipson, B. and Smyth, G.K. (2010). [Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn](#). *Statistical Applications in Genetics and Molecular Biology*, 9(1): Article 39. <https://doi.org/10.2202/1544-6115.1585>
- Westfall, P.H. and Young, S.S. (1993). [Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment](#). Wiley, New York.
- Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M. and Nichols, T.E. (2014). [Permutation inference for the general linear model](#). *NeuroImage*, 92: 381-397. <https://doi.org/10.1016/j.neuroimage.2014.01.060>
- Zhang, J.-H., Chung, T.D.Y. and Oldenburg, K.R. (1999). [A simple statistical parameter for use in evaluation and validation of high throughput screening assays](#). *Journal of Biomolecular Screening*, 4(2): 67-73. <https://doi.org/10.1177/108705719900400206>