RUCova: Removal of Unwanted Covariance in mass cytometry data

Rosario Astaburuaga-García[1,2], Thomas Sell[1,2], Samet Mutlu[3,4,5], Anja Sieber[1,2], Kirsten Lauber[4,6,7], and Nils Blüthgen[1,2,8,✉]

[1]Institute of Pathology, Charité-Universitätsmedizin Berlin, Berlin, Germany.
[2]Institute of Biology, Humboldt University of Berlin, Berlin, Germany.
[3]Department of Radiation Oncology, University Hospital, LMU München, Munich, Germany.
[4]German Cancer Consortium (DKTK), Partner site Munich, Munich, Germany.
[5]German Cancer Research Center (DKFZ), Heidelberg, Germany.
[6]Department of Radiation Oncology, University Hospital, LMU München, Munich, Germany.
[7]Clinical Cooperation Group 'Personalized Radiotherapy in Head and Neck Cancer' Helmholtz Center Munich, German Research Center for Environmental Health GmbH, Neuherberg, Germany.
[8]German Cancer Consortium (DKTK), Partner Site Berlin, Berlin, Germany.

High dimensional mass cytometry is confounded by unwanted covariance due to variation in cell size and staining, making analysis and interpretation challenging.

We present RUCova, a novel method designed to address confounding factors such as heterogeneous cell size and staining efficiency in mass cytometry data. RUCova removes unwanted covariance using multivariate linear regression on Surrogates of Unwanted Covariance (SUCs), and Principal Component Analysis (PCA). We exemplify the use of RUCova and show that it effectively removes unwanted covariance while preserving genuine biological signals. Our results demonstrate the efficacy of RUCova in elucidating complex data patterns, facilitating the identification of activated signaling pathways, and improving the classification of important cell populations. By providing a robust framework for data normalization and interpretation, RUCova enhances the accuracy and reliability of mass cytometry analyses, contributing to advancements in our understanding of cellular biology and disease mechanisms. R package is available on https://github.com/molsysbio/RUCova. Detailed documentation, data, and the code required to reproduce the results are available on https://doi.org/10.5281/zenodo.10913464. Supplementary material: Available at bioRxiv.

mass cytometry | covariance | cell size | R package

Correspondence: ***nils.bluethgen@charite.de***

Introduction

Mass cytometry allows the simultaneous quantification of numerous cellular markers in individual cells and across multiple samples. It is used widely in immunology research to quantify surface proteins and classify immune cells [Spitzer and Nolan, 2016; Bendall et al., 2011; Horowitz et al., 2013; Giesen et al., 2014; Georg et al., 2022]. Mass cytometry is also increasingly used to study intracellular signalling pathways by measuring phospho-protein abundance, providing insights into diverse cellular processes such as the differentiation pathways of colorectal cancer [Brandt et al., 2019; Sell et al., 2023], organoid heterogeneity [Sufi et al., 2021], acute myeloid leukaemia stem/progenitor cells [Han et al., 2015] and prediction of drug sensitivity in breast cancer [Tognetti et al., 2021]. While the distributions of surface proteins typically show a bimodal pattern, those of intracellular signalling markers show a unimodal distribution with rather small quantitative shifts in response to perturbations. These distributions are affected by both biological and technical variability. Biological variability arises from inherent differences between individual cells, including variations in cell state, type and size, whereas technical variability arises from experimental procedures and instrumentation, such as heterogeneous staining efficiency. While some biological variability is essential, unwanted variability, such as that caused by differences in cell size, carries the risk of confounding the data. This unwanted covariance, influenced by factors such as cell size, can obscure the detection of small cell populations and prevent accurate comparisons between different experimental conditions, cell lines and cell states.

In recent years, a class of methods called Remove Unwanted Variation (RUV) has been developed to address this problem, primarily variation coming from batch effects. These methods have been successfully applied to various high-throughput data types, including microarrays [Gagnon-Bartsch and Speed, 2011], RNA-seq (RUV by Risso et al. [2014]), Nanostring nCounter gene expression (RUV-III by Molania et al. [2019]), single-cell RNA-seq (scMerge by Lin et al. [2019]), and mass cytometry (CytofRUV by Trussart et al. [2020]). While differences between batches are a significant source of unwanted variability, single-cell mass cytometry datasets can exhibit considerable covariance within a single batch due to uncorrected heterogeneity in cell size and staining efficiency. Mass cytometry builds upon flow cytometry by increasing the number of measurable markers, as the mass spectrum is more specific than the fluorescence spectrum, which suffers from spectral overlap. In flow cytometry, marker abundance can be normalized by cell size using the Forward Scatter (FSC) parameter, which is proportional to the cell's relative size. However, mass cytometry lacks an intrinsic parameter that directly serves as a proxy for cell size. Conventional normalisation methods, such as those used in single-cell RNA sequencing, are impractical due to the absence of total protein content information in mass cytometry data. Previous attempts to normalize cell volume using Ruthenium isotopes [Rapsomaniki et al., 2018] faced chal-

lenges with complexity and assumptions about marker-cell volume relationships. To address these limitations, we introduce RUCova, a novel approach utilising linear modeling based on Surrogates of Unwanted Covariance (SUCs). By incorporating factors like mean DNA, mean barcoding isotopes, pan Akt, and total ERK, RUCova effectively removes technical artifacts and enhances the accuracy of mass cytometry analyses. Our study demonstrates the utility of RUCova in revealing complex data patterns, identifying activated signalling pathways, and improving the classification of important cell populations like apoptotic cells. In this paper, we showcase the unique advantages of RUCova in advancing our understanding of cellular biology and disease mechanisms.

The RUCova method

RUCova comprises two major steps. First, it fits a multivariate model for each measured marker (m) across cells (i) from samples (j_i) with respect to the surrogates of unwanted covariance (SUC) $\vec{x}_i$. Second, it eliminates such dependency by assigning the residuals ϵ of the model as the new modified expression of the marker. The fit can be expressed as:

$$y_i^m(\vec{x}_i, j_i) = \underbrace{O^m(j_i) + S^m(\vec{x}_i, j_i)}_{M(\vec{x}_i, j_i)} + \epsilon_i^m \qquad \textbf{(1)}$$

where $S^m(\vec{x}_i, j_i)$ describes the slope of the fit and $O^m(j_i)$ the intercept or offset. The predictors are the surrogates of unwanted covariance (SUC) $\vec{x}$, which can be either specific markers as proxies of the confounding factors or the principal components (PCs) derived from a Principal Component Analysis (PCA) performed on such markers.

RUCova offers 3 different uni- or multi-variate linear models to describe the relationship between marker expression and SUC: (1) $M_1(\vec{x}_i)$: simple, (2) $M_2(\vec{x}_i, j_i)$: offset, and (3) $M_3(\vec{x}_i, j_i)$: interaction (Fig. 1).

1. **Simple model (Fig. 1A,B):** Consists of one fit per measured marker y_i^m across the entire data set.

$$M_1(\vec{x}_i) = \underbrace{\beta^m}_{=O_1^m} + \underbrace{\sum_{p=1}^{N_{SUC}} \alpha_p^m \cdot x_{i,p}}_{=S_1^m(\vec{x}_i)} \qquad \textbf{(2)}$$

   where β_0^m is the intercept and α_p^m is the slope coefficient for each predictor or SUC p.

2. **Offset model (Fig. 1C,D):** Consists of one fit per measured marker y_i^m and sample j_i. The fits for the samples share the same slope, while differing in the intercept (offset term $O_2^m(j_i)$).

$$M_2(\vec{x}_i, j_i) = \underbrace{\beta_{j_i}^m}_{=O_2^m(j_i)} + \underbrace{\sum_{p=1}^{N_{SUC}} \alpha_p^m \cdot x_{i,p}}_{=S_2^m(\vec{x}_i)} . \qquad \textbf{(3)}$$

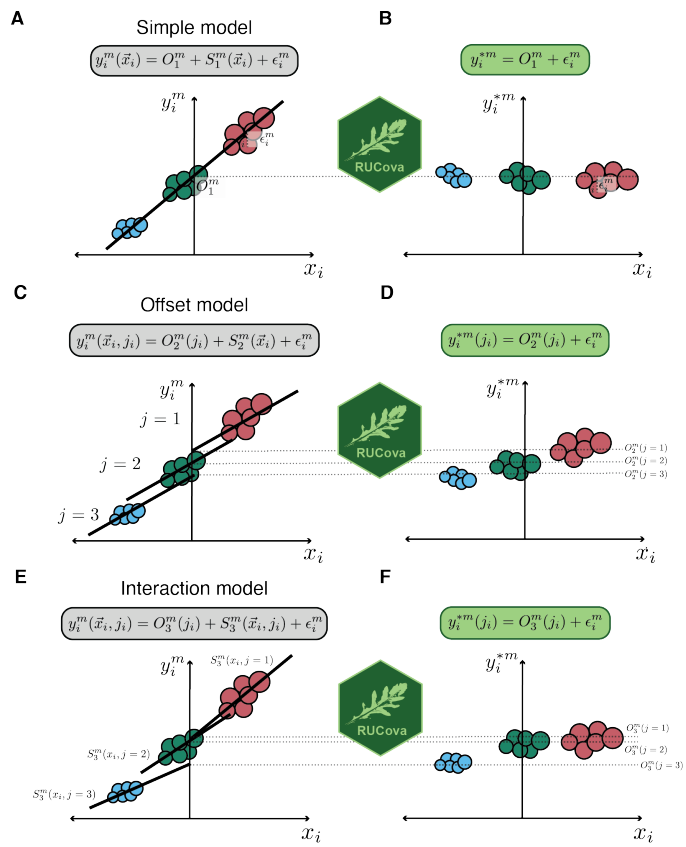


Fig. 1. Illustration of the RUCova method and its three different models. A, C, E) Original expression y_i^m of a marker m before RUCova as a function of a centered expression of a SUC or PC. Illustrative regression line and equation corresponding to each model. **B, D, F)** Modified expression y_i^{*m} of a marker m after RUCova. **D, F)** Keeping the offset $O^m(j_i)$ between samples j. **A, B)** Simple model: one fit across the input data set with intercept O_1^m and residuals ϵ_i. **C, D)** Offset model: one slope $S_2^m(\vec{x}_i)$ for the whole input data set and different intercepts $O_2^m(j_i)$ between samples. **E, F)** Interaction model: one fit per sample j with intercepts $O_3^m(j_i)$ and slope $S_3^m(\vec{x}_i, j_i)$.

3. **Interaction model (Fig. 1E,F):** Consists of one fit per measured marker y_i^m and sample j_i. The fits for the samples can have different slopes (interaction term $S_3^m(\vec{x}_i, j_i)$) and intercepts (offset term $O_3^m(j_i)$).

$$M_3(\vec{x}_i, j_i) = \underbrace{\beta_{j_i}^m}_{=O_3^m(j_i)} + \underbrace{\sum_{p=1}^{N_{SUC}} \alpha_{j_i,p}^m \cdot x_{i,p}}_{=S_3^m(\vec{x}_i, j_i)} . \qquad \textbf{(4)}$$

Samples j_i can be either different cell lines, perturbations, conditions, *metacells* (clusters), or even batches. By taking the zero-centered distributions of the SUCs ($x_i^c = x_i - \frac{1}{N_i}\sum_i^{N_i} x_i$), the mean values of the markers m across all cells are kept after applying RUCova (Fig. 1). If a more conservative approach is desired where the fold changes between samples should be kept, each SUC should be centered per sample (Fig. S1). Similarly, when using PCs as the predictive variables, SUCs can be z-score normalised by sample before performing PCA.

The RUCova method eliminates the dependency of each measured marker on the SUCs by computing the residuals of the model, in addition to the intercept as the revised expression for each marker (y_i^{m*}). The offset - intercept - term O^m for different samples can either be wanted or unwanted. For the first case, the new, modified abundance of the marker is independent of $\vec{x}_i$ and can be expressed as:

$$y_i^{*m}(j_i) = O^m(j_i) + \epsilon_i^m \quad \textbf{(5)}$$

More information about the RUCova model, cell cultures and mass cytometry measurements can be found in the Supplementary Material.

Results

1. Mass cytometry data is confounded by multiple factors. Mass cytometry enables the quantification of protein and phospho-protein abundance in single-cells using antibodies conjugated with metal isotopes, facilitating the investigation of intracellular signals that determine the state and response to treatments. However, challenges such as heterogeneous cell volume and labeling efficiency confound the data, leading to spurious correlations between markers and hindering comparisons between cell lines, perturbations, and cell states. To address this, we developed RUCova, an R package designed to remove unwanted covariances in mass cytometry data.

To illustrate the need and benefit of using the RUCova method, we chose a mass cytometry data set with 8 different Head-and-Neck Squamous Cell Carcinoma (HNSCC) cell lines in control (0 Gy) and irradiated (48 h after 10 Gy) condition (Fig. 2A, S4).

Since cell volume and labeling efficiency cannot be directly measured with mass cytometry, we use four Surrogates of Unwanted Covariance (SUCs, Eq. S1): (1) Mean DNA: Mean value of normalised iridium channels (Eq. S2), (2) Mean BC: Mean value of the highest (used) barcoding isotopes per cell (Eq. S3), (3) pan Akt, and (4) total ERK (Fig. 2B).

Iridium is a common DNA stain in mass cytometry as it binds to DNA. Interestingly, we noted that Ruthenium, previously proposed by Rapsomaniki et al. [2018] to correct for the cell volume, strongly correlates with DNA staining (Fig. S3A).

Mass cytometry is often performed with multiplexed samples which are stained with specific combination of isotopes, *e.g.:* palladiums or telluriums [Zunder et al., 2015; Willis et al., 2018], acting as a barcode. These barcording reagents bind unspecifically to surface proteins and, when stained while cells are fixed, also intracellular proteins [Zunder et al., 2015]. Thus barcode signals may be used as surrogate of cell volume.

Total ERK and pan Akt are commonly used as loading controls for normalization in *e.g.:* Western blotting experiments, as they are typically abundant proteins that are relatively stable under different experimental conditions.

These four SUCs strongly correlate between each other and to the majority of the markers in all studied HNSCC cell lines (Fig. S5, S6). The correlations for the Cal27 cell line are depicted in Fig. 2C as an example. Some correlations are authentic and expected, such as between proliferation markers (p-Rb and Ki-67), members of the MAPK pathway (p-MEK1/2 and p-ERK1/2), regulators of cell cycle progression and protein synthesis (p-Rb and p-4EBP1), proteins in the DNA damage-response pathway (p-p53 and p-γH2AX), and proteins involved apoptotic cell death (p-Chk2, cCasp3, cPARP, NF-κB and IκBα). However, other correlations are suspicious and most likely driven by unwanted covariance, specially between p-Rb, p-38, p-CDC25c, p-Smads, YAP and the majority of the measured markers (Fig. 2C). This distribution of correlation coefficients is observed in all the studied HNSCC lines (Fig. 2D, S5, S6)

We quantified the cell area via microscopy images for each HNSCC cell line and condition (Fig. 2E, Fig. S4A,B). We observed variations in the median cell area, both across different cell lines within the same condition and within each cell line across radiation conditions. The median area of the cells increased after irradiation with 10 Gy, which agrees with previously observed cellular enlargement of irradiated cells [Rene and Nardone, 1968; Ronny Sham et al., 2020; Ren et al., 2023]. The UPCISCC131 cell line displayed the most increase in median cell area following irradiation.

PCA calculated on the four SUCs shows that more than half of the variance across SUCs is explained by PC1 (Fig. 2F,G), which strongly correlated with the quantified median cell area ($\rho = 0.87$, 2H), in contrast with subsequent PCs (Fig. S6D).

In an uncorrected mass cytometry data set, these differences in cell area (and therefore marker abundance) and other factors like staining efficiency will confound the comparisons between cell lines and perturbations, masking meaningful biological information.

2. RUCova enables improved classification of apoptotic cells by uncovering previously obscured cell populations. To remove the unwanted covariance coming from heterogeneous cell area, but also potentially from heterogeneous staining efficiency, we applied the RUCova method using the interaction model $M_3(\vec{x}_i, j_i)$ with the cell lines as samples (j).

After removing the covariance with all SUCs (*i.e*, using all four PCs as predictive variables) spurious correlations were removed and authentic correlations were kept (Fig. 3A-C), enabling differentiation between activated (apoptotic and MAPK pathway), and non-activated signalling pathways (JAK/STAT and TGFβ). The covariance between marker abundances was substantially decreased in all HNSCC cell lines, specially after regressing-out the correlation with PC1 and PC2 (Fig. 3D, S5, S6). Consistently, we observed a general decrease in the standard deviation of the signals for each condition after applying RUCova (Fig. 3E). However, for some markers we observed an unexpected increase in the standard deviation relative to the original distributions (arrows in Fig. 3E). For IdU in the UPCISCC154 line, p-γH2AX in UPCISCC131 line, and p-p53 in the Cal33 line, the higher standard deviation observed after RUCova was attributed to

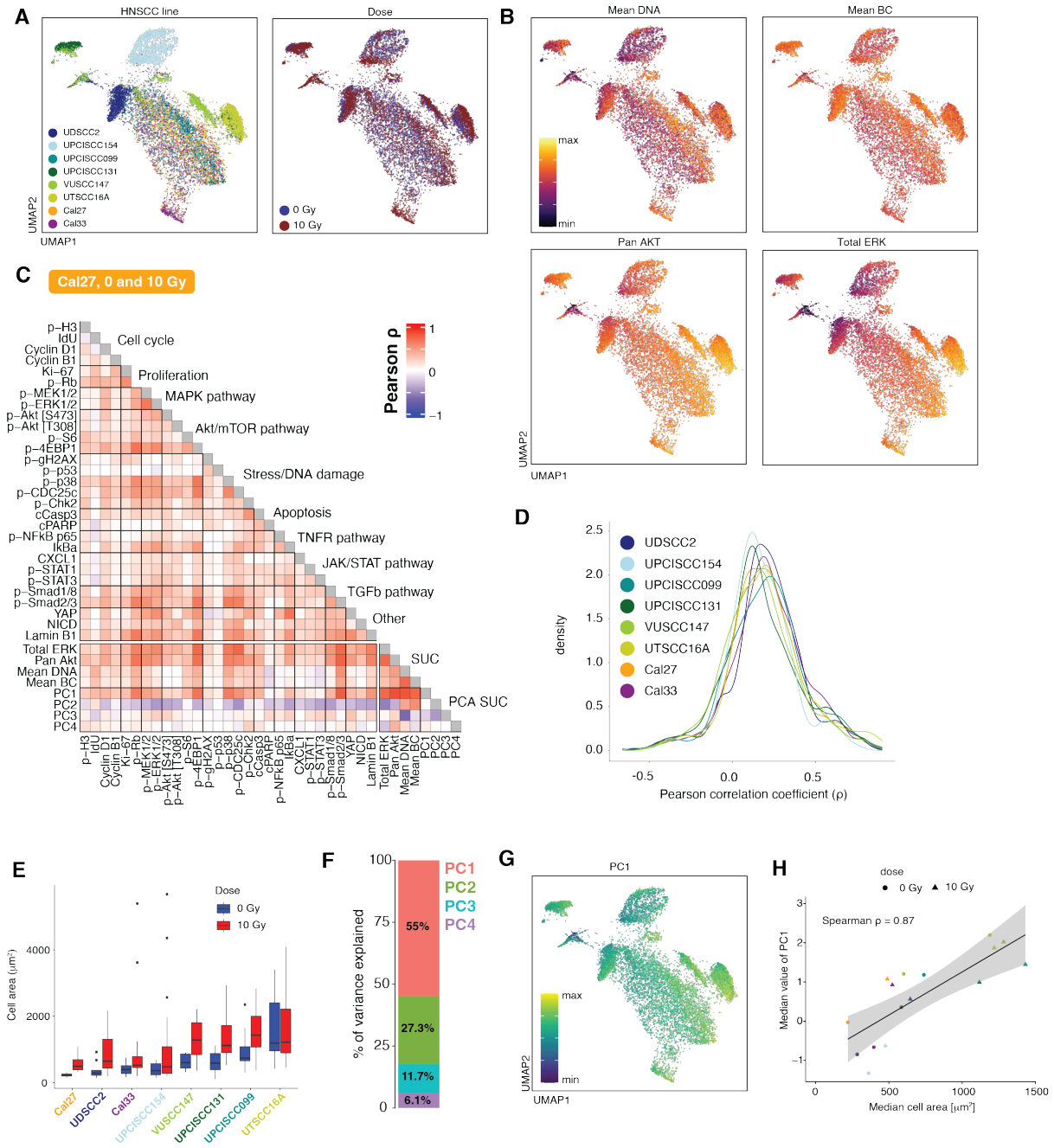


Fig. 2. Unwanted covariance in a mass cytometry data set coming from heterogeneous cell area and other factors. A). UMAP coloured by cell line (left) and dose (right) calculated by excluding cell cycle and proliferation markers (n = 800 cells per line and dose). **B)**. UMAP coloured by the expression values of the 4 SUCs: mean DNA (top left), mean BC (top right), pan Akt (bottom left), and total ERK (bottom right). Expression values were asinh-transformed and min-max normalised. **C)**. Pearson correlation coefficients between z-scored and asinh-transformed expression values of markers in the Cal27 cell line across cells from 0 and 10 Gy. **D)**. Distribution of Pearson correlation coefficients between markers in all cell lines across cells from 0 and 10 Gy. **E)** Measurements of the cell area via microscopy images before fixation (n = 16.5 $\pm$ 3.7 cells per cell line and condition, Fig. S4B). **F)**. Percentage of variance explained by each PC of a PCA based on the SUCs. PCA was calculated based on the asinh-transformed and z-scored expression values of the 4 SUCs in B). **G)**. UMAP coloured by PC1 of a PCA based on the 4 SUCs. PC1 values were min-max normalised. **H)** Spearman correlation coefficients between the median cell area (μm^2) from E) and median value of PC1 from G) per cell line and dose.

the assignment of non-zero values to the artificial zero values typically present in a mass cytometry measurement (Fig. 3F). The rise in the standard deviation (σ) of the distribution of the apoptotic marker cleaved Caspase-3 (cCasp3) in the UPCISCC131 line under irradiated condition can be attributed to two factors: the assignment of non-zero values and an increased dissimilarity in cCasp3 signal between the non-apoptotic and apoptotic populations (Fig. 3F,G). We categorized the apoptotic cells in the irradiated UPCISCC131 cell line by analyzing their cCasp3 signals and establishing decisions thresholds before and after RUCova (based on PC1 to PC4) (dashed vertical lines in Fig. 3G). The classification based on the data after RUCova allowed to increase the identification capabilities of apoptotic cells by 47.5 % (Fig. 3H). Apoptotic cells showed lower original cCasp3 signals (before RUCova), making them less distinguishable

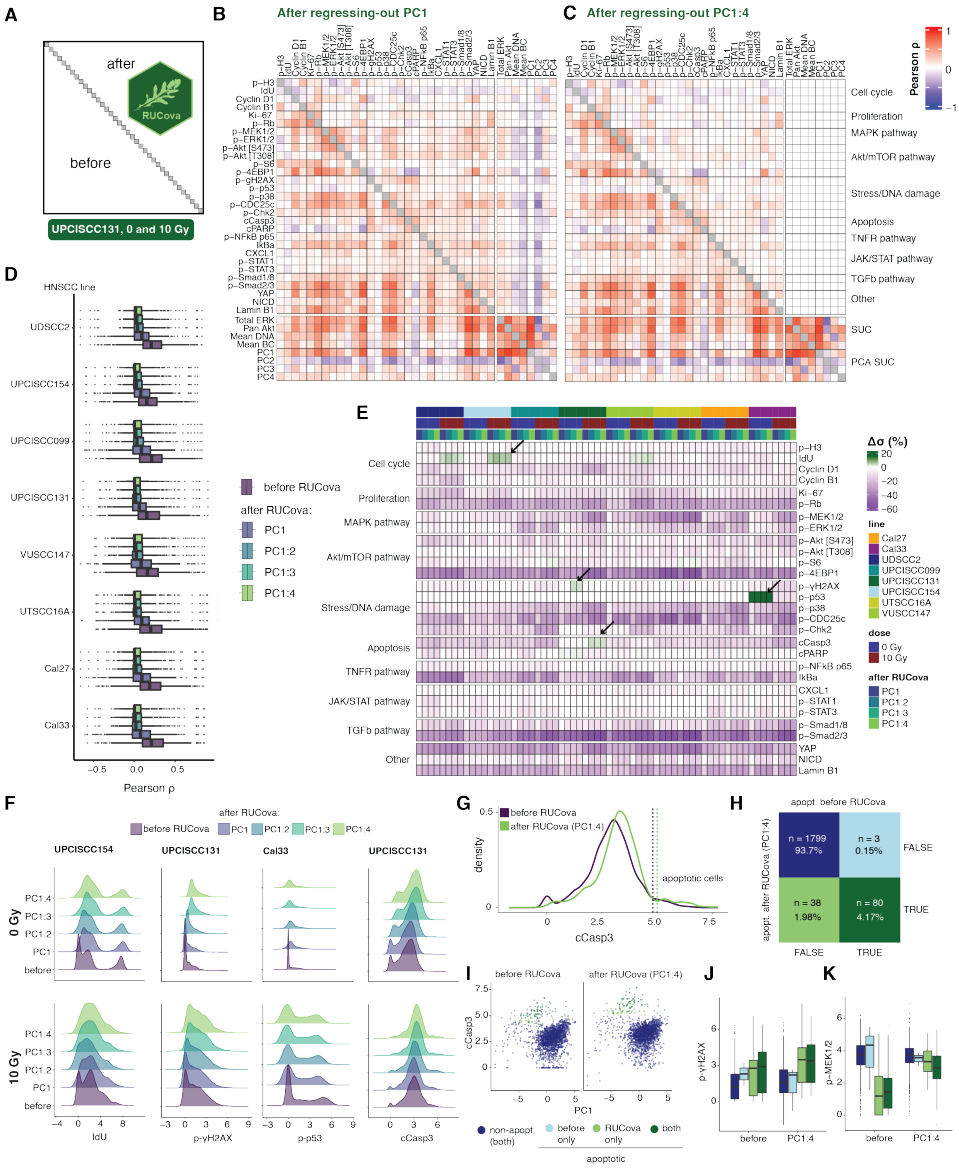


Fig. 3. RUCova removes unwanted covariance and allows improved classification of apoptotic cells. A) Scheme of the correlation heatmap, where the lower and upper triangles show the Pearson correlation coefficients between marker values before and after RUCova (using the interaction model $M_3(\vec{x}_i, j_i)$ per cell line), respectively. Diagonal unity values are depicted in grey. Pearson correlation coefficients were calculated between z-scored and asinh-transformed expression values of markers. **B,C)** Correlation heatmap with the upper triangle showing the Pearson correlation coefficients between marker values in the UPCISCC131 cell line across cells from 0 and 10 Gy after RUCova based on **B)** PC1, **C)** PC1 to PC4. **D)** Boxplots for Pearson correlation coefficients between markers per cell line across cells from 0 and 10 Gy, based on the expression values before and after RUCova. **E)** Heatmap of percentual difference in the standard deviation σ of each marker's distribution after applying RUCova relative to the distributions before RUCova ($\Delta\sigma = \frac{\sigma_{\text{RUCova}} - \sigma_{\text{before}}}{\sigma_{\text{before}}} \cdot 100$) per cell line and dose levels. Arrows indicate specific cases where $\Delta\sigma > 0$. **F)** Density plots for asinh-transformed expression values of IdU, p-$\gamma H2AX$, p-p53, and cleaved Casp3 before and after RUCova (y-axis) in cell lines where $\Delta\sigma < 0$ (arrows in panel E) for 0 Gy (first row) and 10 Gy (second row). **G)** Density plot for asinh-transformed expression of cleaved Casp3 in the irradiated (10 Gy) UPCISCC131 cell line, before (purple) and after RUCova based on all four PCs (green). Dashed vertical lines indicate the decision thresholds for apoptotic (higher values of cleaved Casp3) and non-apoptotic populations according to the cleaved Casp3 distribution before and after RUCova. **H)** Confusion matrix for classification of apoptotic cells in the irradiated UPCISCC131 cell line before and after RUCova (based on all four PCs) according to the decision thresholds in G). **I)** Scatter plots of PC1 vs. asinh-transformed values of cleaved Caspase-3 before (left) and after RUCova (right) for irradiated cells in the UPCISCC131 line. Cells are coloured by apoptotic status. **J,K)** Boxplots of asinh-transformed expression of **J)** p-γH2AX and **K)** p-MEK1/2 before and after RUCova in irradiated UPCISCC131 cells according to their apoptotic status.

from non-apoptotic cells. This lower cCasp3 signal corresponded to lower PC1 values (Fig. 3I). After RUCova, the expected differences in markers like the DNA-damage marker γH2AX between apoptotic and non-apoptotic cells became discernible (Fig. 3J, S7), while potential artifacts, like lower p-MEK1/2 signals in apoptotic cells, were reduced (Fig. 3K, S7).

The use of RUCova in this data set enabled the preservation of authentic correlations while removing spurious ones, facilitating the differentiation between activated and non-activated signaling pathways after irradiation in different HNSCC lines. The application of RUCova allowed a clearer understanding of apoptotic marker distribution and its relation to cell size reduction, highlighting its efficacy in elucidating complex data patterns in mass cytometry analysis.

3. RUCova enhances the reliability of perturbation comparisons by eliminating cell size artifacts. To un-

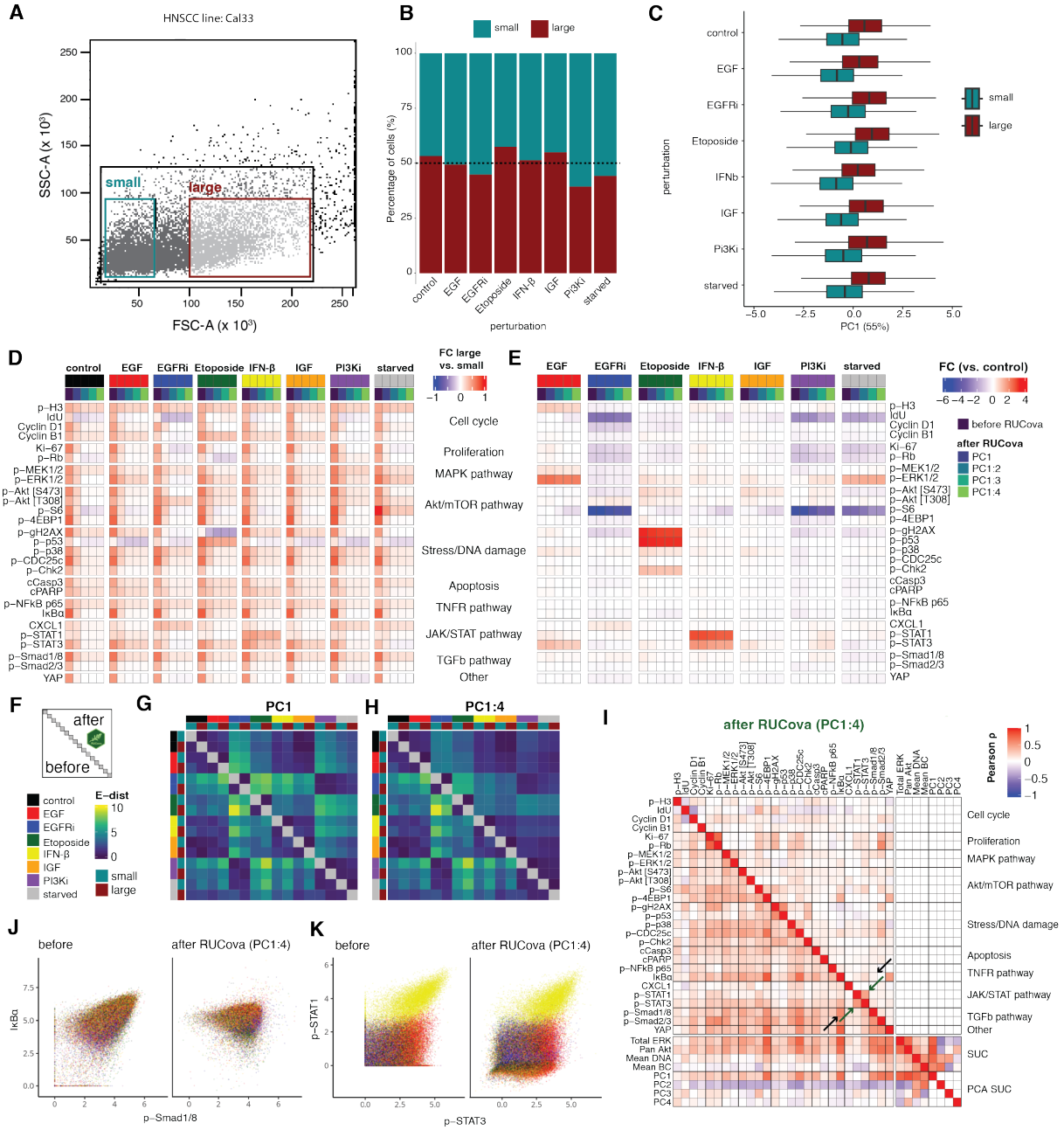


Fig. 4. RUCova successfully removes cell size artifacts enabling accurate perturbation analyses. A) Scatter plot of the FACS parameters SSC-A vs FSC-A. Gates defining the size-sorted groups of small (blue) and large (red) cells in the Cal33 cell line before mass cytometry measurement. B) Percentage of small and large cells per perturbation condition in the Cal33 cell line. **C)** Boxplots of PC1 (of a PCA based on the four SUCs) per perturbation and size group. **D)** Fold Change (FC) of asinh-transformed values between large and small cells all conditions, before and after applying the simple RUCova model ($M_1(\vec{x}_i)$) using different number of PCs. **E)** Fold Change (FC) of asinh-transformed values between perturbation and control condition, before and after applying the simple RUCova model ($M_1(\vec{x}_i)$ using different number of PCs. **F)** Scheme of the E-distance heatmap, where the lower and upper triangles show the E-distance between conditions before and after RUCova (using the simple model $M_1(\vec{x}_i)$), respectively. **G,H)** E-distance heatmap between 1000 cells per condition and sorted population for data after RUCova using **G)** only PC1, **H)** PC1 to PC4. **I)** Correlation heatmap with the upper triangle showing the Pearson correlation coefficients between marker values across all pertubations and sorted pupulations after RUCova based on PC1 to PC4. Black arrows indicate an artifactual correlation which is removed after RUCova, and green arrows indicate a real correlation which is kept. **J, K)** Scatter plots of Cal33 cells coloured by perturbation before (left) and after (right) RUCova based on all four PCs. **J)** asinh-transformed values of p-Smad1/8 vs. IκBα (artifactual correlation). **K)** asinh-transformed values of p-Stat1 and p-Stat3 (real correlation driven by IFN-β-stimulated cells).

derstand how cell size and other factors can confound mass cytometry data and especially analyses of cellular responses, we perturbed cells from the HNSCC Cal33 cell line using EGF stimulation (30 min), EGFR inhibition (Gefitinib, 24 h), Etoposide treatment (2 h), IFN-β stimulation (30 min), IGF stimulation (30 min), PI3K inhibition (24 h), and starvation alone (24 h). We then sorted the cells into two groups based on size (small and large cells) using fluorescence-activated cell sorting (FACS) (Fig. 4A).

Overall, most of the perturbations resulted in similar proportions of sorted cells, as shown in Fig. 4B. However, in the case of PI3K inhibition the majority of sorted cells were smaller in size. The PC1 of a PCA based on the SUCs was substantially higher for large compared to small cells

(Fig. 4C). Correspondingly, average marker values were consistently higher in large cells compared to small cells across all perturbations (Fig. 4D), illustrating once again how cell size confounds mass cytometry data. Upon applying RUCova using the simple model $M_1(\vec{x}_i)$, these differences in mean values were notably reduced, particularly after removing correlations with all four PCs. For certain markers and perturbations (e.g., p-p53 in Etoposide treatment and p-STAT1 in IFN-β stimulation) considerable fold-changes between large and small cells persisted after RUCova, indicating the method's ability to preserve genuine biological signals while eliminating artificial ones. Across large and small cells, the fold-changes between perturbation and control condition were usually maintained after applying RUCova (Fig. 4E). Nevertheless, the removal of unwanted covariances led to modified fold-changes in some cases: p-ERK1/2 was increased after PI3K inhibition relative to control, which might indicate a compensatory signaling mechanism as a means of maintaining cell survival and proliferation. These reductions or increments in the fold-change after RUCova could also be due to unbalanced proportion of small and large cells in a perturbation compared to control condition. By assessing the pairwise E-distance (Peidli et al. [2024]) between perturbation conditions and sorted populations both before and after applying RUCova, we observed that differences between small and large cells were effectively eliminated, along with dissimilarities arising from variations in cell size distribution, such as those seen in the PI3K condition compared to the others. We also observed small increments in the E-distance after implementing RUCova, specially between EGFR inhibition and starvation (Fig. 4F,G,H, Fig. S8A).

The Pearson correlation coefficients between markers across the entire data set after RUCova provides reliable information. An example for a regressed-out artifactual correlation is p-Smad1/8 and IκBα (black arrow Fig. 4I,J), which was mainly driven by the size of the cells (Fig. S8C). An example of a real correlation that was kept after RUCova is between p-Stat1 and p-Stat2 mainly driven by IFN-β stimulation (green arrow Fig. 4I,K, S8D).

This suggests that RUCova successfully reduced the dissimilarities between conditions due to cell size and other potential sources of unwanted covariance, providing a clearer understanding of the underlying biological responses to different stimuli.

Discussion

Mass cytometry data is contaminated by variance that is induced by variations in cell size, staining and due to other technical artifacts, leading to spurious correlations between markers. Here we describe RUCova, a method to regress out such unwanted co-variation. The method consists on fitting a model for each marker based on Surrogates of Unwanted Covariance (SUCs), such as mean DNA, mean barcode signal, and total protein marker such as total ERK and AKT. Previous approaches used fixed relation between marker abundance and cell size stains [Rapsomaniki et al., 2018] yielding suboptimal results, as the extend of correlation between protein abundance and cell size varies and depends *e.g.* on protein localisation [Lanz et al., 2023].

Cell size can exhibit variability across distinct cell lines, cell types (e.g., various immune cells), and even within different tissue microenvironments [Liu et al., 2022]. If the association between cell size and protein abundance varies, RUCova can incorporate this using the interaction model ($M_3(\vec{x}_i, j)$), where slopes of the regression fits can account for differences among samples (cell lines, cell types, or cell clusters).

In Fig. 2, S5, S6 we illustrated that the Pearson correlation coefficients between markers display consistently high values, posing challenges in distinguishing between activated and non-activated signaling pathways. The majority of these markers exhibited strong correlations with the SUCs, specially with the first principal component (PC1) across SUCs which explained 55 % of the variance and highly correlated with cell area, thereby confirming the artifactual nature of these elevated correlations.

In Fig. 3 we demonstrated how spurious correlations and resulting variance in marker distributions were effectively mitigated with RUCova. We noticed that PC1 of the SUC represents cell area, and that regressing out PC1 and PC2 removes spurious correlation effectively. We recommend to use PCA components on the SUC as predictors in the model. By adjusting the number of PCs used, the user of RUCova can adjust the extent of unwanted correlation to be removed. PC1 may also serve as a proxy of cell size in later analysis. Mass cytometry data often contains an inflation of zero values that are technically induced by sensitivity thresholds of the instruments. While RUCova widely reduces variance within condition, we observed that variance of markers with a large number of zeros increased as those will be non-zero values after regression. Some approaches developed for mass cytometry data address these zero values by imputing them with estimated values [Li et al., 2017; Minoura et al., 2020] or by extrapolating measurements from other panels using k-nearest neighbor methods [Abdelaal et al., 2019]. However, applying these imputation methods to uncorrected data may introduce bias, as the resulting non-zero values could be influenced by existing covariance in the dataset. In contrast, RUCova addresses this issue by assigning non-zero values while simultaneously removing such covariance.

An important application of RUCova is to identify populations of cells that may be obscured due to cell-size. In this article we showed that RUCova allowed to identify about 50% more apoptotic cells in a data set. These previously hidden cells exhibited the lowest PC1 values, suggesting they were smaller in size, a known phenomenon in apoptotic cells which undergo cell shrinkage during the early stages of apoptosis [Kerr et al., 1972; Jänicke et al., 1998; Albeck et al., 2008]. Importantly, RUCova also corrected marker expression for apoptosis related signals, including higher levels of the DNA damage marker p-$\gamma H2AX$ compared to non-apoptotic cells.

To directly compare how RUCova corrects the data between small and large cells, we employed fluorescence-activated

cell sorting (FACS) to separate large and small cells. RUCova removed differences in signals between small and large cells with only very few biologically plausible exceptions. These include p-Stat1 that showed differences between large and small cells following IFN-β. IFN-β induces inflammation which leads to larger cell sizes [Han et al., 2022], thus higher p-Stat1 signal may be a genuine signal indicative of inflammation-induced cell enlargement.

RUCova relies on the selection of SUCs to model and remove unwanted covariances in the mass cytometry data. However, the effectiveness of RUCova may be influenced by the choice and availability of suitable SUCs. In some cases, identifying appropriate SUCs that adequately capture all sources of unwanted covariance may be challenging. RUCova operates under the assumption of linearity in the relationship between marker expression and SUCs. While this assumption may hold true for many cases, there may be instances where the relationship between markers and SUCs is non-linear, leading to potential limitations in the accuracy of the correction performed by RUCova. The performance of RUCova may be sensitive to the specific modeling approach chosen (simple, offset or interaction model), the selection of the number of principal components or the decision to use uni- or multivariate modeling. Suboptimal model selection may lead to incomplete correction of unwanted covariances or overfitting of the data, potentially introducing biases or artifacts into the analysis. We recommend employing and comparing multiple modeling approaches. While RUCova aims to improve the interpretability and reliability of mass cytometry data by removing unwanted covariances, it is essential to consider the potential impact of data correction on biological interpretation. Over-correction or removal of genuine biological signals alongside technical artifacts may obscure meaningful biological insights or introduce biases into downstream analyses. For optimal utilization of RUCova and accurate interpretation of its results, it's recommended that mass cytometry users visually inspect cell morphology and estimate cell size. Additionally, verifying any correlation between cell size estimates and covariance levels in the dataset is essential. This approach ensures a comprehensive understanding of the data and boosts confidence in the regression process by identifying potential sources of unwanted variability.

Conclusion

In conclusion, our study introduces RUCova as a powerful tool for removing unwanted covariance in mass cytometry data, thereby enhancing the accuracy and reliability of downstream analyses. By effectively addressing technical artifacts associated with heterogeneous cell size and staining efficiency, RUCova facilitates the uncovering of genuine biological signals and contributes to a deeper understanding of cellular processes. Our findings demonstrate the utility of RUCova in elucidating complex data patterns, facilitating the identification of activated signaling pathways, and improving the classification of apoptotic cells. Furthermore, we emphasize the importance of thoughtful model selection and validation, as well as the critical interpretation of results in the context of biological insights. Moving forward, continued refinement and validation of RUCova and related methodologies will further enhance their utility in advancing our understanding of cellular biology and disease mechanisms.

References

T. Abdelaal et al. Cytofmerge: integrating mass cytometry data across multiple panels. *Bioinformatics*, 35(20): 4063â4071, Mar. 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz180. URL http://dx.doi.org/10.1093/bioinformatics/btz180.

J. G. Albeck et al. Quantitative analysis of pathways controlling extrinsic apoptosis in single cells. *Molecular Cell*, 30(1):11â25, Apr. 2008. ISSN 1097-2765. doi: 10.1016/j.molcel.2008.02.012. URL http://dx.doi.org/10.1016/j.molcel.2008.02.012.

S. C. Bendall et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687â696, May 2011. ISSN 1095-9203. doi: 10.1126/science.1198704. URL http://dx.doi.org/10.1126/science.1198704.

R. Brandt et al. Cell type-dependent differential activation of erk by oncogenic kras in colon cancer and intestinal epithelium. *Nature Communications*, 10 (1), July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10954-y. URL http://dx.doi.org/10.1038/s41467-019-10954-y.

J. A. Gagnon-Bartsch and T. P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539â552, Nov. 2011. ISSN 1468-4357. doi: 10.1093/biostatistics/kxr034. URL http://dx.doi.org/10.1093/biostatistics/kxr034.

P. Georg et al. Complement activation induces excessive t cell cytotoxicity in severe covid-19. *Cell*, 185(3):493–512.e25, Feb. 2022. ISSN 0092-8674. doi: 10.1016/j.cell.2021.12.040. URL http://dx.doi.org/10.1016/j.cell.2021.12.040.

C. Giesen et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417â422, Mar. 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2869. URL http://dx.doi.org/10.1038/nmeth.2869.

L. Han et al. Single-cell mass cytometry reveals intracellular survival/proliferative signaling in flt3âitdâmutated aml stem/progenitor cells. *Cytometry Part A*, 87(4): 346â356, Jan. 2015. ISSN 1552-4930. doi: 10.1002/cyto.a.22628. URL http://dx.doi.org/10.1002/cyto.a.22628.

X.-X. Han et al. Interferon-beta inhibits human glioma stem cell growth by modulating immune response and cell cycle related signaling pathways. *Cell Regeneration*, 11(1), July 2022. ISSN 2045-9769. doi: 10.1186/

s13619-022-00123-w. URL http://dx.doi.org/10.1186/s13619-022-00123-w.

A. Horowitz et al. Genetic and environmental determinants of human nk cell diversity revealed by mass cytometry. *Science Translational Medicine*, 5(208), Oct. 2013. ISSN 1946-6242. doi: 10.1126/scitranslmed.3006702. URL http://dx.doi.org/10.1126/scitranslmed.3006702.

R. U. Jänicke, M. L. Sprengart, M. R. Wati, and A. G. Porter. Caspase-3 is required for dna fragmentation and morphological changes associated with apoptosis*. *Journal of Biological Chemistry*, 273(16):9357–9360, 1998. ISSN 0021-9258. doi: https://doi.org/10.1074/jbc.273.16.9357. URL https://www.sciencedirect.com/science/article/pii/S002192581846139X.

J. F. R. Kerr, A. H. Wyllie, and A. R. Currie. Apoptosis: A basic biological phenomenon with wideranging implications in tissue kinetics. *British Journal of Cancer*, 26(4):239â257, Aug. 1972. ISSN 1532-1827. doi: 10.1038/bjc.1972.33. URL http://dx.doi.org/10.1038/bjc.1972.33.

M. C. Lanz, L. Fuentes Valenzuela, J. E. Elias, and J. M. Skotheim. Cell size contributes to single-cell proteome variation. *Journal of Proteome Research*, 22(12):3773â3779, Nov. 2023. ISSN 1535-3907. doi: 10.1021/acs.jproteome.3c00441. URL http://dx.doi.org/10.1021/acs.jproteome.3c00441.

H. Li et al. Gating mass cytometry data by deep learning. *Bioinformatics*, 33(21):3423â3430, July 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx448. URL http://dx.doi.org/10.1093/bioinformatics/btx448.

Y. Lin et al. scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets. *Proceedings of the National Academy of Sciences*, 116(20):9775â9784, Apr. 2019. ISSN 1091-6490. doi: 10.1073/pnas.1820006116. URL http://dx.doi.org/10.1073/pnas.1820006116.

S. Liu et al. What programs the size of animal cells? *Frontiers in Cell and Developmental Biology*, 10, Nov. 2022. ISSN 2296-634X. doi: 10.3389/fcell.2022.949382. URL http://dx.doi.org/10.3389/fcell.2022.949382.

K. Minoura et al. Cybertrack2.0: zero-inflated model-based cell clustering and population tracking method for longitudinal mass cytometry data. *Bioinformatics*, 37(11):1632â1634, Oct. 2020. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa873. URL http://dx.doi.org/10.1093/bioinformatics/btaa873.

R. Molania, J. A. Gagnon-Bartsch, A. Dobrovic, and T. P. Speed. A new normalization for nanostring ncounter gene expression data. *Nucleic Acids Research*, 47(12):6073â6083, May 2019. ISSN 1362-4962. doi: 10.1093/nar/gkz433. URL http://dx.doi.org/10.1093/nar/gkz433.

S. Peidli et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531â540, Jan. 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02144-y. URL http://dx.doi.org/10.1038/s41592-023-02144-y.

M. A. Rapsomaniki et al. Cellcycletracer accounts for cell cycle and volume in mass cytometry data. *Nat Commun*, 9(1):632, Feb 2018. doi: 10.1038/s41467-018-03005-5.

Y. Ren et al. Ionizing radiation triggers mitophagy to enhance dna damage in cancer cells. *Cell Death Discovery*, 9(1), July 2023. ISSN 2058-7716. doi: 10.1038/s41420-023-01573-0. URL http://dx.doi.org/10.1038/s41420-023-01573-0.

A. Rene and R. Nardone. The effect of gamma radiation on cell enlargement and size distribution of strain l. *Biosystems*, 2(4):207–214, 1968. ISSN 0303-2647. doi: https://doi.org/10.1016/0303-2647(68)90005-1. URL https://www.sciencedirect.com/science/article/pii/0303264768900051.

D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896â902, Aug. 2014. ISSN 1546-1696. doi: 10.1038/nbt.2931. URL http://dx.doi.org/10.1038/nbt.2931.

N. F. Ronny Sham et al. Study of morphological changes and survival fraction in emt6 cell line post-gamma ray irradiation. *Journal of Physics: Conference Series*, 1497(1):012032, Mar. 2020. ISSN 1742-6596. doi: 10.1088/1742-6596/1497/1/012032. URL http://dx.doi.org/10.1088/1742-6596/1497/1/012032.

T. Sell et al. Oncogenic signaling is coupled to colorectal cancer cell differentiation state. *Journal of Cell Biology*, 222(6), Apr. 2023. ISSN 1540-8140. doi: 10.1083/jcb.202204001. URL http://dx.doi.org/10.1083/jcb.202204001.

M. H. Spitzer and G. P. Nolan. Mass cytometry: Single cells, many features. *Cell*, 165(4):780â791, May 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.04.019. URL http://dx.doi.org/10.1016/j.cell.2016.04.019.

J. Sufi et al. Multiplexed single-cell analysis of organoid signaling networks. *Nature Protocols*, 16(10):4897â4918, Sept. 2021. ISSN 1750-2799. doi: 10.1038/s41596-021-00603-4. URL http://dx.doi.org/10.1038/s41596-021-00603-4.

M. Tognetti et al. Deciphering the signaling network of breast cancer improves drug sensitivity prediction. *Cell Systems*, 12(5):401–418.e12, May 2021. ISSN 2405-4712. doi: 10.

1016/j.cels.2021.04.002. URL http://dx.doi.org/10.1016/j.cels.2021.04.002.

M. Trussart et al. Removing unwanted variation with cytofruv to integrate multiple cytof datasets. *eLife*, 9, Sept. 2020. ISSN 2050-084X. doi: 10.7554/elife.59630. URL http://dx.doi.org/10.7554/eLife.59630.

L. M. Willis et al. Tellurium based mass cytometry barcode for live and fixed cells. *Cytometry Part A*, 93 (7):685â694, July 2018. ISSN 1552-4930. doi: 10.1002/cyto.a.23495. URL http://dx.doi.org/10.1002/cyto.a.23495.

E. R. Zunder et al. Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nature Protocols*, 10(2):316â333, Jan. 2015. ISSN 1750-2799. doi: 10.1038/nprot.2015.020. URL http://dx.doi.org/10.1038/nprot.2015.020.

Supplementary data

Supplementary material is available online (PDF).

Availability and implementation

R package is available on https://github.com/molsysbio/RUCova. Detailed documentation, data, and the code required to reproduce the results are available on https://doi.org/10.5281/zenodo.10913464.

Competing interests

None declared.

Funding

The funding for this work was provided by the BMBF grants number 02NUK047C & 02NUK047E (ZiSS-Trans), 02NUK086A & 02NUK086E (Senirad), 16LW0239K (MSTARS-2), and DFG research training group RTG2424 (CompCancer).

Author contributions statement

R.A.G. and N.B. conceptualized the project. T.S., S.M., and A.S. conducted the experiments. R.A.G developed the RUCova method and R package in collaboration with T.S. and N.B. R.A.G. and N.B. wrote and reviewed the manuscript. N.B. and K.L. provided supervision.

Acknowledgments

We acknowledge the BIH Cytometry Core Facility for their help with cell sorting and mass cytometry data acquisition.

$$y_i^m(\vec{x}_i) = O_1^m + S_1^m(\vec{x}_i) + \epsilon_i^m$$

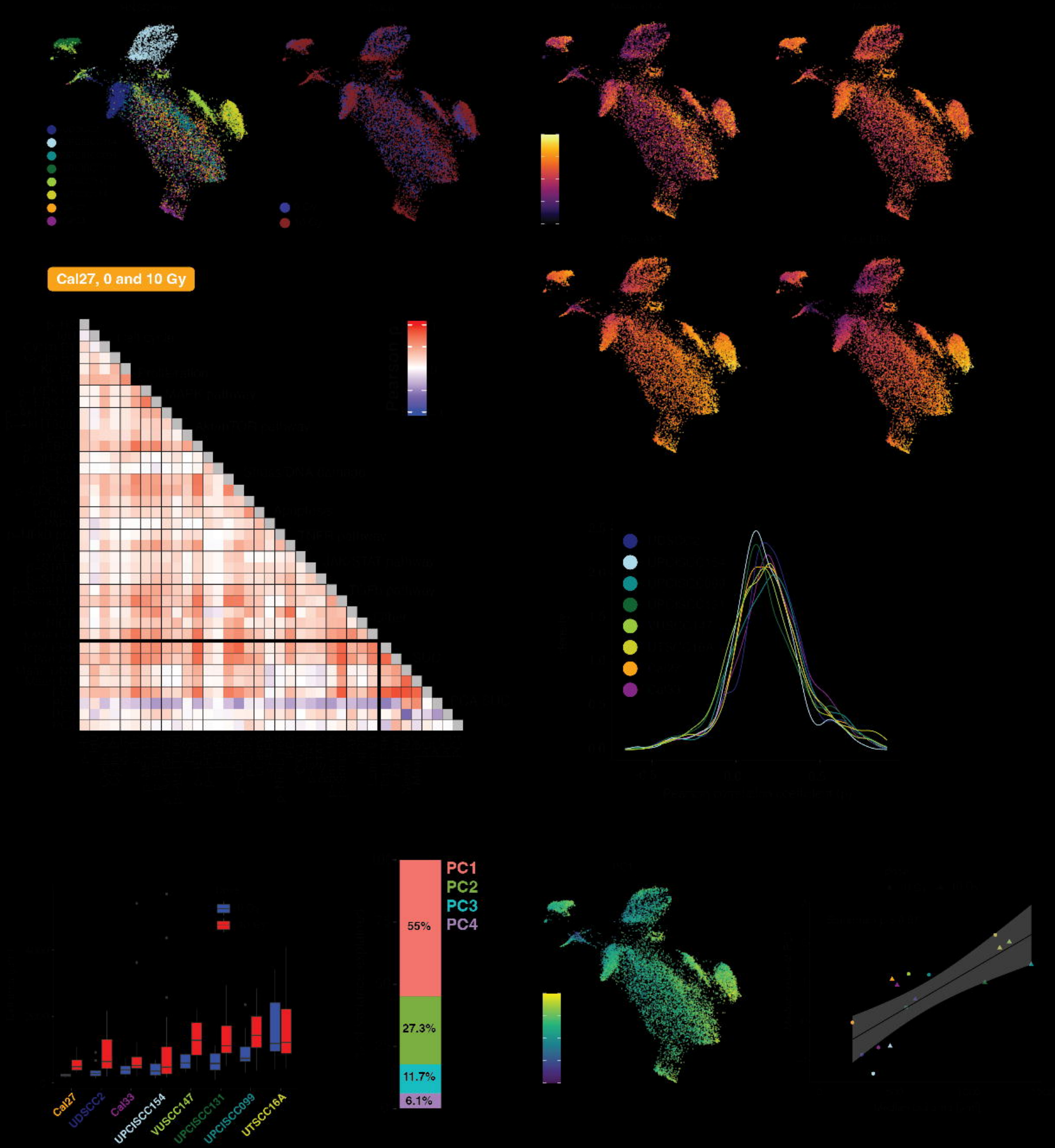
$$y_i^{*m} = O_1^m + \epsilon_i^m$$

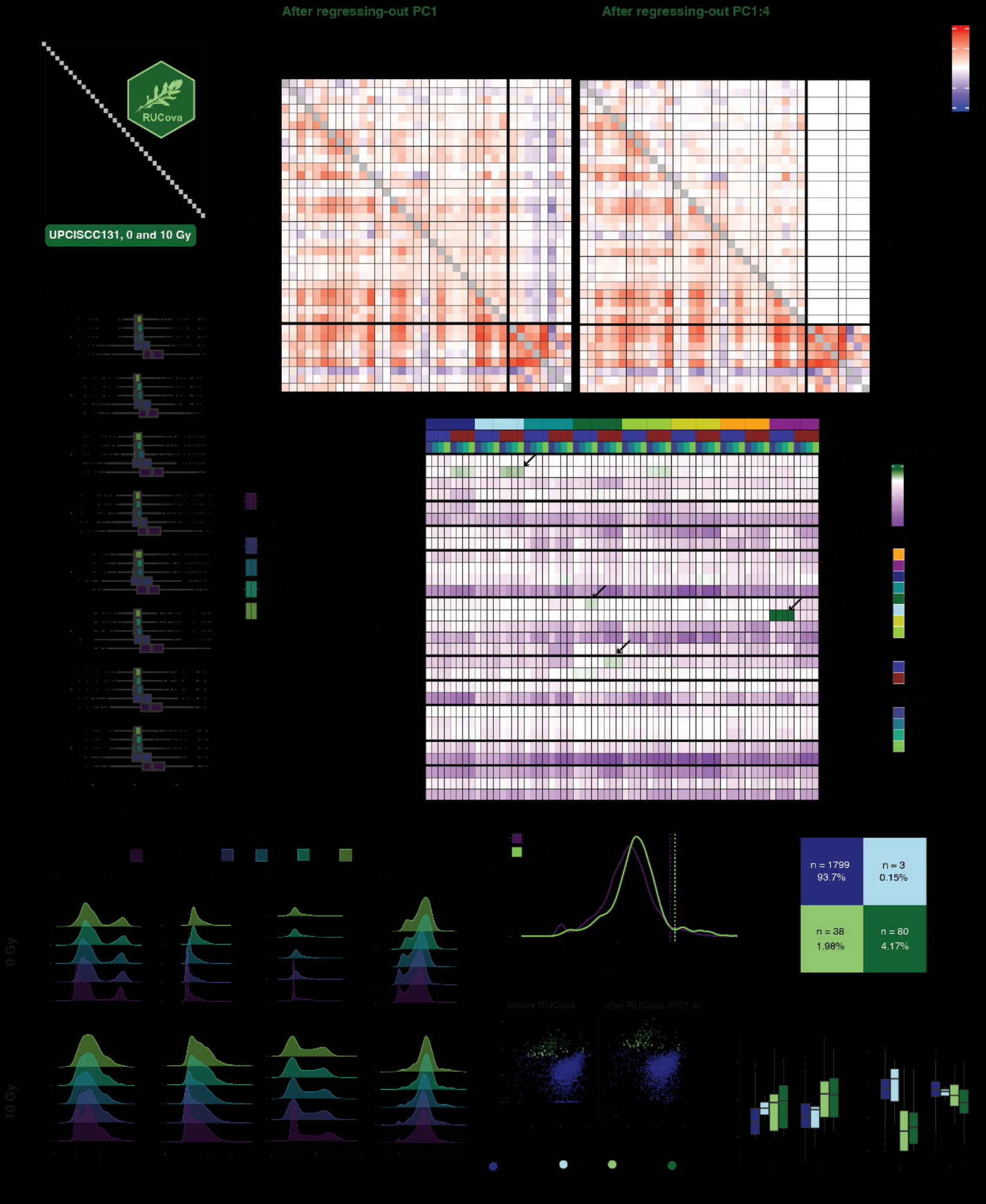
$$y_i^m(\vec{x}_i, j_i) = O_2^m(j_i) + S_2^m(\vec{x}_i) + \epsilon_i^m$$

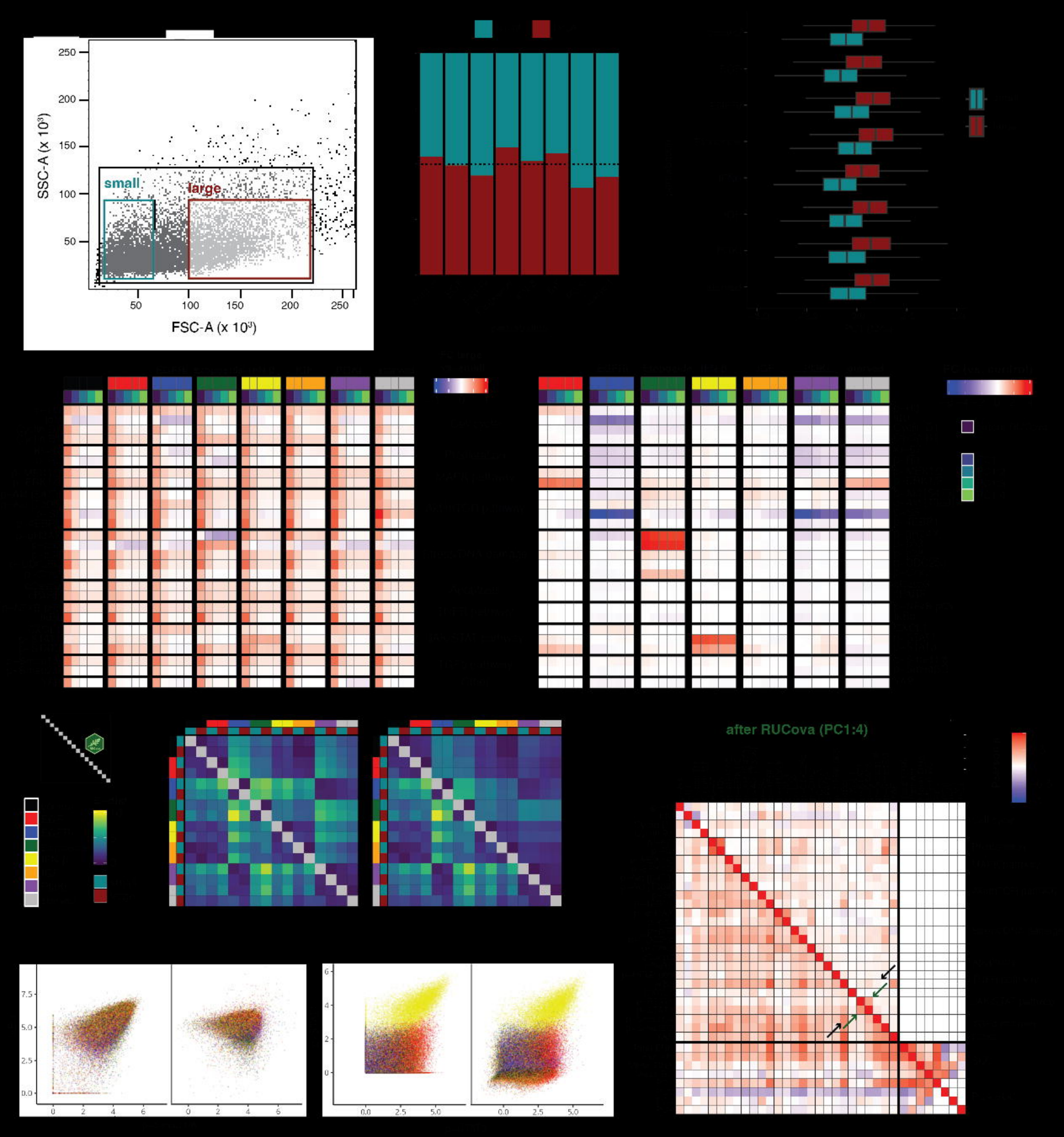
$$y_i^{*m}(j_i) = O_2^m(j_i) + \epsilon_i^m$$

$$y_i^m(\vec{x}_i, j_i) = O_3^m(j_i) + S_3^m(\vec{x}_i, j_i) + \epsilon_i^m$$

$$y_i^{*m}(j_i) = O_3^m(j_i) + \epsilon_i^m$$


Cal27, 0 and 10 Gy
PC1
PC2
PC3
PC4
55%
27.3%
11.7%
6.1%
Cal27
UDSCC2
Cal33
UPCISCC154
VUSCC147
UPCISCC131
UPCISCC099
UTSCC16A


After regressing-out PC1
After regressing-out PC1:4
RUCova
UPCISCC131, 0 and 10 Gy
n = 1799
93.7%
n = 3
0.15%
n = 38
1.98%
n = 80
4.17%
0 Gy
10 Gy
before RUCova
after RUCova (PC1:4)


SSC-A (x 10³)
FSC-A (x 10³)
small
large
after RUCova (PC1:4)