

Cell Reports Medicine, Volume 2

Supplemental information

**C-MORE: A high-content single-cell morphology
recognition methodology for liquid biopsies
toward personalized cardiovascular medicine**

Jennifer Furkel, Maximilian Knoll, Shabana Din, Nicolai V. Bogert, Timon Seeger, Norbert Frey, Amir Abdollahi, Hugo A. Katus, and Mathias H. Konstandin

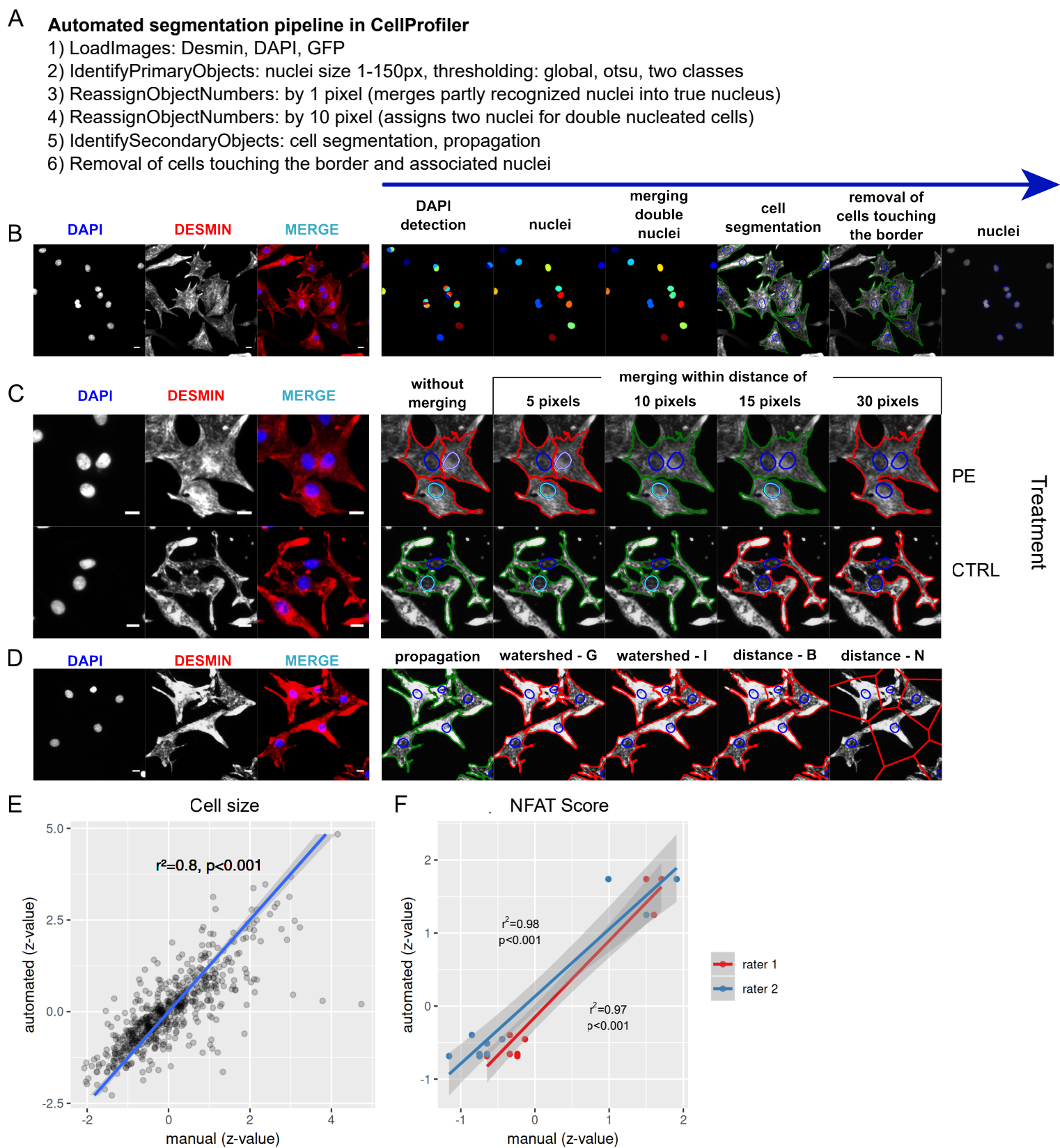


Figure S1. CellProfiler pipeline functionality and validity assessment of automated measurements. Related to Figure 1 and STAR methods. (A) Overview of the CellProfiler analysis steps and important settings. (B) Overview of the CellProfiler analysis step with an image example. (C) shows the detection of double nucleated cells by distance of two nuclei. The distance of 10 pixels was validated manually to be optimal to identify true multi-nucleated cardiomyocytes, while minimizing misidentification of nearby single nucleated cells. Unstimulated and hypertrophic cardiomyocytes were used as ground truth. We show representative recognition on images of unstimulated (CTRL) and phenylephrine (PE) stimulated cardiomyocytes. (D) shows a comparison of cell segmentations performed by different algorithms available in CellProfiler. For our pipeline we chose the propagation algorithm. (E) Correlation between manual and automatically determined cell sizes (z-transformation, Spearman correlation, $n=612$ cells, p -value: likelihood ratio test between linear models). (F) Numbers of nuclear GFP positive cells as determined automatically (density thresholding) and manually by two raters. Metrics are calculated as in E. Images were cropped for better visualization.

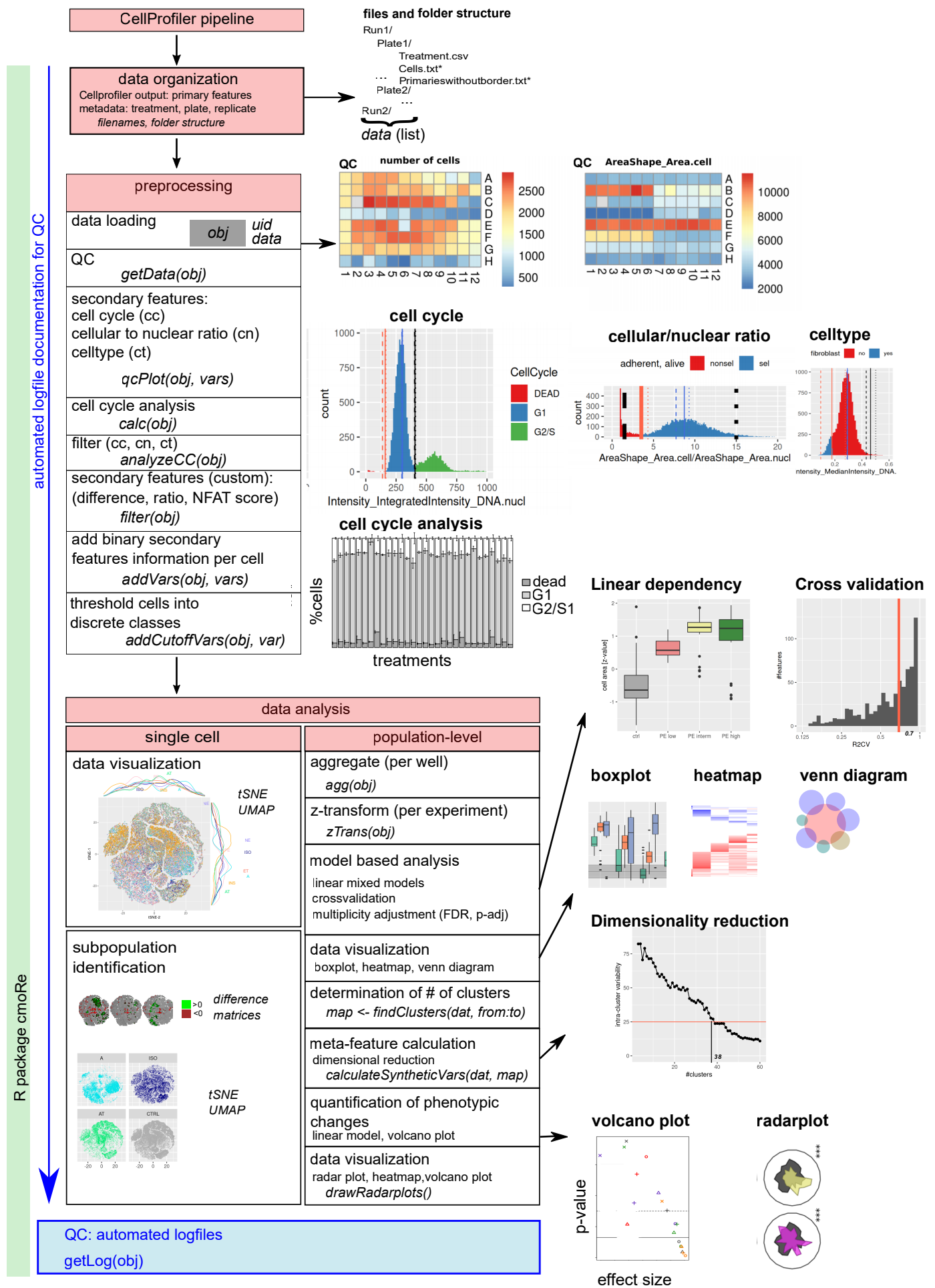


Figure S2. Detailed overview of the R package cmoRe and the proposed workflow. Related to Figure 2 and STAR methods. The analysis workflow comprises four main parts: 1) Loading data matrices calculated by CellProfiler using a convenient file-folder structure, 2) Preprocessing includes quality control using median feature distribution over the plate (e.g. cell number or cell area). Secondary features are calculated using the cmoRe thresholding function. Specific secondary features - cell cycle (cc), cellular to nuclear ratio for detection of non-attached cells (cn) and the celltype (ct) - are used to filter for vital and properly attached cardiomyocytes for downstream analyses. Custom secondary features, e.g. when custom reporters as the NFAT-GFP reporter are used can be calculated additionally on the curated data set. For data analysis cmoRe offers 3) a single cell level phenotyping with functions to quantify patterns for subpopulation identification. And 4) a population-level phenotyping aggregating single cell data per well. cmoRe functions to be used for the respective analysis step are indicated in italic font. In-depth documentation, a handbook of all package functions and an example are provided in Methods S1 and the package vignette online.

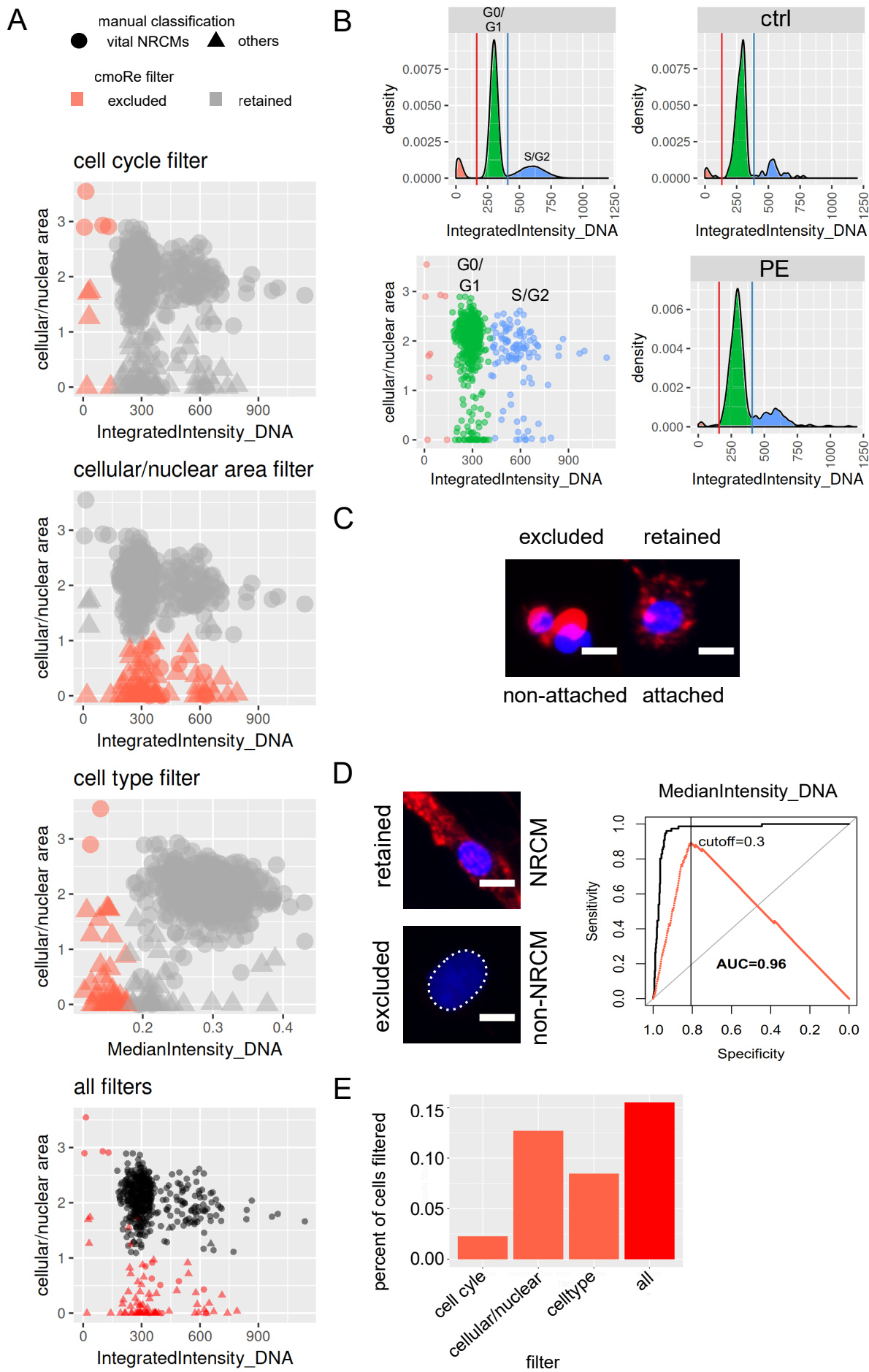


Figure S3. Filters implemented in the cmoRe workflow validated against a manually curated data set. Related to Figure 2 and STAR methods. (A) Scatter plots of the respective secondary features used for filtering via the cmoRe automated thresholding function. (B) Representative plots of nuclear integrated DNA intensity and cmoRe calculated thresholds for automated cell cycle analysis analogously to usage e.g. in flow cytometry (FACS). (C) Filter for non-attached cells: A representative non-attached cell is shown which is characterized by a very small cellular to nuclear ratio and excluded by the filter. In contrast, we show a very small, but properly attached cell which is correctly retained by the filter. (D) Left: Filter for non-cardiomyocytes: Non-cardiomyocytes are characterized by their low intense nucleus. A representative retained NRCM cell, as well as an excluded non-NRCM nucleus (dotted line) is shown. Right: The feature was chosen based on its high sensitivity and specificity in separating NRCMs and non-NRCMs as depicted on the right (receiver operating characteristic (ROC); data: manually annotated dataset, black: ROC curve, red: Youden index). (E) Fraction of cells excluded with the respective filter step and fraction of overall excluded cells.

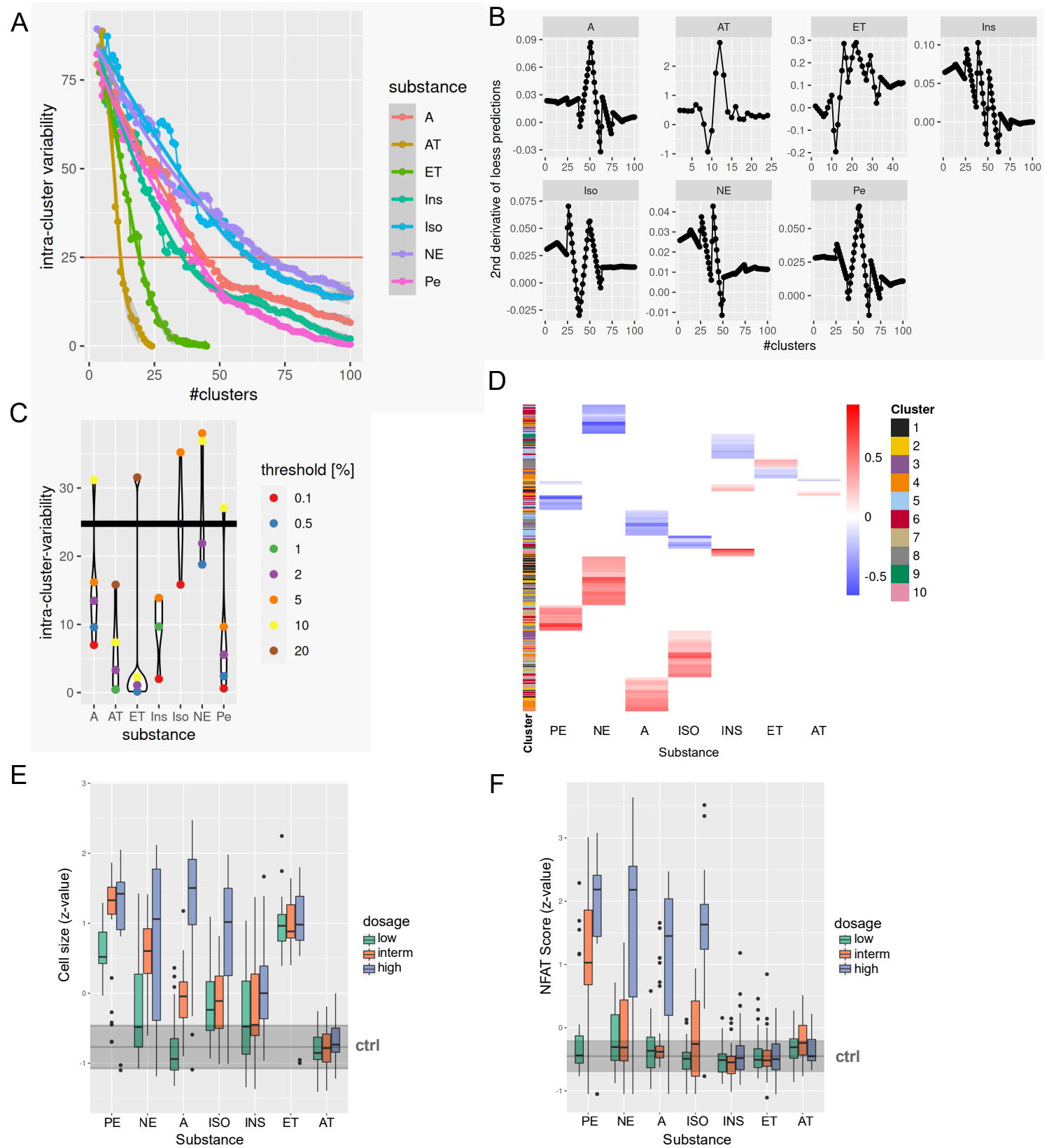


Figure S4. Selection of the intracluster variability threshold, stimulus specific feature changes and differences in cell sizes and nuclear NFAT positive cell fraction. Related to Figure 2,3, and STAR methods. (A) Intracluster variability for increasing number of clusters for all substances. (B) Second derivative of loess fit predictions trained on data from A. The cluster number corresponding to the value which surpasses its predecessor value for more than a fraction of the observed range for the respective substance is retained for multiple fractions (see C, thresholds). (C) Intracluster variability per identified threshold and median value of all thresholds (black horizontal line). (D) Heatmap of canonical hypertrophic stimulus specific feature alterations. Red indicates a dose-dependent increase, blue indicates a dose-dependent decrease. (E) Cell size for all stimuli and concentrations (z-values). (F) Nuclear GFP (NFAT) positive fraction of cells (z-score) for all stimuli and concentrations. For control condition the median and median absolute deviation is depicted in dark grey in E and F.

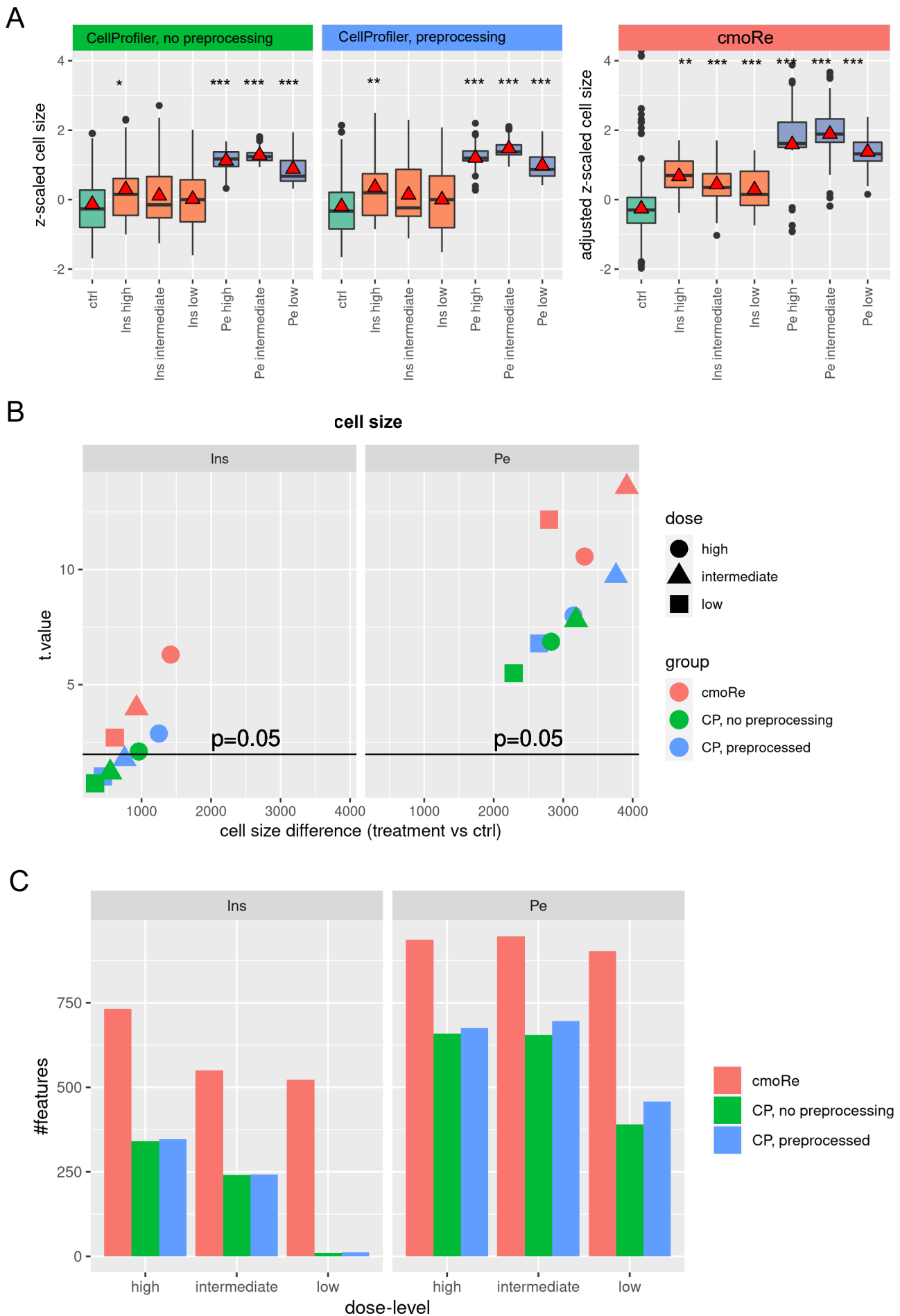


Figure S5. Benchmarking of cmoRe against CellProfiler raw data implemented with naïve linear models. Related to Figure 2. (A) Cell size as reported by CellProfiler (left) and adjusted with linear mixed models as implemented in cmoRe (right). Red triangles: mean, preprocessed: filtered data (fibroblasts, dead cells). z-transformed values. (B) Pairwise differences in cell size (ctrl vs treatment/dose) for compared analysis methods. (C) Number of significant features for different analysis methods with FDR < 0.05. P-values were calculated with linear mixed effect models; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$

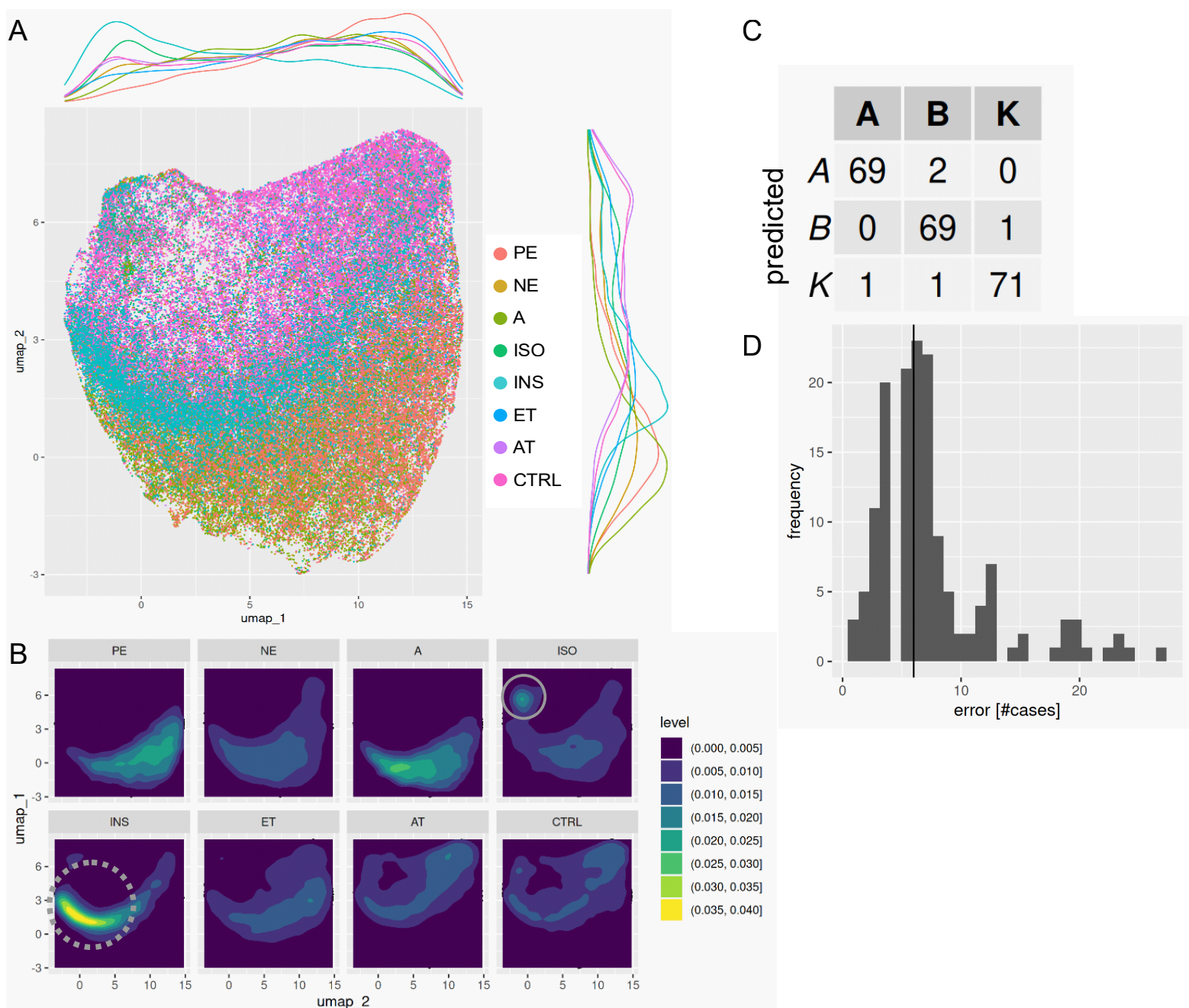


Figure S6. UMAP representation of single cell data of all canonical hypertrophic stimuli and random forest classification of aortic stenosis dataset. Related to Figure 3 and 5 and STAR methods. (A) UMAP of all canonical hypertrophic stimuli, with marginal distributions. (B) Cellular densities per stimulus color-coded on the UMAP plot. Identified substance specific subpopulations are circled for ISO, for INS circled with dotted line. (C) Accuracy of random forest classification of preTAVR (A), postTAVR (B) and healthy controls (K) per well: The table shows true vs. predicted classes. (D) Error for all permutations of random forest classification as shown in C, training and test set with fixed group sizes. Median number of misclassified wells are 5. For A and B: Single cell data of $n=1$ experiment is shown. For C and D: Training set consisted of patients 1,2,3, K1,K2,K3,K4; test set consisted of patients 4,5, K5,K6 for C. In D all permutations were used.

References

- Bakal C, Aach J, Church G, Perrimon N (2007) Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* 316: 1753-6
- Garvey CM, Spiller E, Lindsay D, Chiang C-T, Choi NC, Agus DB, Mallick P, Foo J, Mumenthaler SM (2016) A high-content image-based method for quantitatively studying context-dependent cell population dynamics. *Scientific reports* 6: 29752-29752
- Jentzsch C, Leierseder S, Loyer X, Floherschütz I, Sassi Y, Hartmann D, Thum T, Lagerbauer B, Engelhardt S (2012) A phenotypic screen to identify hypertrophy-modulating microRNAs in primary cardiomyocytes. *Journal of Molecular and Cellular Cardiology* 52: 13-20
- Manzella G, Schreck LD, Breunis WB, Molenaar J, Merks H, Barr FG, Sun W, Römmele M, Zhang L, Tchinda J, Ngo QA, Bode P, Delattre O, Surdez D, Rekhi B, Niggli FK, Schäfer BW, Wachtel M (2020) Phenotypic profiling with a living biobank of primary rhabdomyosarcoma unravels disease heterogeneity and AKT sensitivity. *Nature Communications* 11: 4629
- Ryall KA, Bezzerides VJ, Rosenzweig A, Saucerman JJ (2014) Phenotypic screen quantifying differential regulation of cardiac myocyte hypertrophy identifies CITED4 regulation of myocyte elongation. *J Mol Cell Cardiol* 72: 74-84
- Sero JE, Sailem HZ, Ardy RC, Almuttaqi H, Zhang T, Bakal C (2015) Cell shape and the microenvironment regulate nuclear translocation of NF- κ B in breast epithelial and tumor cells. *Mol Syst Biol* 11: 790
- Slack MD, Martinez ED, Wu LF, Altschuler SJ (2008) Characterizing heterogeneous cellular responses to perturbations. 105: 19306-19311
- Snijder B, Vladimer GI, Krall N, Miura K, Schmolke A-S, Kornauth C, Lopez de la Fuente O, Choi H-S, van der Kouwe E, Gültekin S, Kazianka L, Bigenzahn JW, Hoermann G, Prutsch N, Merkel O, Ringler A, Sabler M, Jeryczynski G, Mayerhoefer ME, Simonitsch-Klupp I et al. (2017) Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. *The Lancet Haematology* 4: e595-e606
- Woo LA, Tkachenko S, Ding M, Plowright AT, Engkvist O, Andersson H, Drowley L, Barrett I, Firth M, Akerblad P, Wolf MJ, Bekiranov S, Brautigan DL, Wang QD, Saucerman JJ (2019) High-content phenotypic assay for proliferation of human iPSC-derived cardiomyocytes identifies L-type calcium channels as targets. *J Mol Cell Cardiol* 127: 204-214

Table S2. Stimulus and inhibitor information. Related to Figure 3 and 4.

hypertrophic stimulus	low concentration	medium concentration	high concentration	source	identifier
Adrenaline (Suprarenin/Epinephrin, 1mg/ml)	0,1 μ M	1 μ M	10 μ M	Sanofi	6053210
Angiotensin II ((R)-(-)-Phenylephrine hydrochloride)	0,1 μ M	1 μ M	10 μ M	Sigma	A9525
Endothelin-1 (Endothelin 1 97% (HPLC), powder)	10 nM	0,1 μ M	1 μ M	sigma	E7764
Insulin (Insulin Insuman rapid 40 ie)	4*10 ⁻⁵ IE/ml	4*10 ⁻⁴ IE/ml	4*10 ⁻³ IE/ml	Sanofi	1843315
Isoproterenol ((-) Isoproterenol hydrochloride)	0,1 μ M	1 μ M	10 μ M	Sigma	I6504
Noradrenaline (Aterenol 1mg/ml)	0,1 μ M	1 μ M	10 μ M	Sanofi	3870227
inhibitor	low concentration	medium concentration	high concentration	source	identifier
AKT (Tricibine Akt V Inhibitor)	0,156 μ M	1,56 μ M	15,6 μ M	Sigma	124038
ERK (ERK/MEK Inhibitor)	0,1 μ M	1 μ M	10 μ M	Promega	U0126
FAK (PF 573228 FAK inhibitor)	10 μ M	0,1 mM	1mM	Tocris	3239
GSK (BIO GSK3b inhibitor)	5 nM	50 nM	500 nM	Tocris	3194
PI3K (Ly294002, PI3K Inhibitor)	650 nM	6,5 μ M	65 μ M	Millipore	440202

Methods S1. Detailed description and manual for C-MORE functions. Related to Figure 1,2,3 and STAR methods

1. Calculation of intracluster variability

Intra-cluster variability was computed using Euclidean distance and ward.D2 clustering method. Intra-cluster variability values for different numbers of clusters (2 to 100) were used to fit loess models. Second derivative of loess fit model predictions were evaluated to determine an intra-cluster variability cutoff. Starting from the highest number of clusters, the change between this value and its predecessor was calculated as fraction of the maximal observed difference per substance. The maximum intra-cluster variability detected in any of the substances for a given threshold was retained (0.1, 0.5, 1, 2, 5, 10 and 20%), the median of all values was 24.75. Thus a cutoff of 25 was used for analysis.

2. Benchmarking of cmoRe against CellProfiler raw data implemented with naïve linear models

To evaluate cmoRe performance, we tested differences in NRCM cell size data obtained with CellProfiler (CP) between non-treated and INS/PE treated cells by omitting single cmoRe (pre-)processing steps.

Therefore, we tested for differences between control (non-treated, ctrl) and INS/PE treated cells (per dose level, i.e. ctrl vs INS high data, ctrl vs. PE intermediate) on well aggregated data.

First, we assessed non-filtered data (no exclusion of dead cells and fibroblasts, Supplementary Figure 6A, left [CellProfiler, no preprocessing]) using linear models. For the comparison ctrl vs INS, we only found a significant difference ($p < 0.05$) between highest dose of INS and ctrl, and for all comparisons between PE and ctrl. Next, we used data from which dead cells and fibroblasts were filtered prior to aggregation per well (Supplementary Figure 6A, preprocessed). Again, only the comparison ctrl vs INS high was the only significant finding within the comparisons of ctrl and INS treated cells. T-values (p-values), corresponding to the respective tests, however, were larger (smaller) (Supplementary Figure 6B).

Next, we assessed the effect of using linear mixed models for analyses (labeled cmoRe), to adjust for variation between repetitions of experiments. All comparisons between ctrl and INS irrespective of dose level yielded significant results. For visualization, z-transformed residuals $y_i - \hat{y}_i$ (y_i : measured data in well i , \hat{y}_i : predicted data with the mixed model formula $y \sim 1 + (1|experiment/plate)$) are shown.

Finally, we tested the number of significant features from all CP features. cmoRe yielded the highest number of significant results ($FDR < 0.05$), much less features were obtained when using linear models instead of linear mixed models for analysis (CP, no preprocessing and CP, preprocessed). The latter, however, showed slightly higher numbers of features, highlighting the beneficial effect of filtering fibroblasts and dead cells in detecting differences in NRCM morphology induced by differential treatment.

3. Manual for our R-package cmoRe

C-MORE: A high content single cell morphology
assay for cardiovascular medicine:
The cmoRe R package

July 26, 2021

Document version 0.3

Package version 0.3

Contents

1 Introduction	4
2 Cellprofiler data preparation	4
2.1 Cellprofiler output	4
3 The R package <i>cmoRe</i>	4
3.1 The <i>imgExp</i> class	6
3.2 Data loading	6
3.3 QC plots	6
3.4 Filtering of single cells	6
3.4.1 Detected numbers per well	7
3.4.2 Morphology based	7
3.4.3 Fibroblast filter (<i>fb</i>)	8
3.4.4 Cell-cycle assignment (<i>cc</i>)	9
3.4.5 Debris/detached cells filtering (<i>nc</i>)	10
3.5 Threshold identification	10
3.5.1 Cutoff identification	12
3.5.2 Imputation of missing data	12
3.5.3 Assignment of single cells	12
3.6 Cell-cycle analysis	12
3.7 Additional features	13
3.8 Aggregation	14
3.9 Z-transformation	14
3.10 Removal of single cell data	15
4 Single cell data analysis	15
5 Well-aggregated data analysis	15
5.1 Dose dependent specific alterations	15
5.2 Inhibitors	16
5.2.1 Positive control vs inhibitor	16
5.2.2 Negative control and inhibitor	17
5.3 TAVI	17
5.3.1 Concentration dependent regulation	17
5.3.2 Concentration independent regulation	19
5.3.3 Crossvalidation	19
5.4 IPS cells	19
6 Crossvalidation	20
7 Data prediction	20
8 Feature selection	20
9 Metafeature calculation	20

10 Phenotype visualization

22

1 Introduction

The *cmoRe* R package contains a collection of functions for preprocessing and analysis of morphological analysis obtained with *CellProfiler* (CP) on a single cell basis.

Main functionality of the package spans:

- Preprocessing
 - Quality control
 - Filtering for vital cardiomyocytes
 - Exclusion of non-cardiomyocytes (fibroblasts)
 - Cell-cycle assignment
 - Automatic threshold selection of multimodal distributions¹
- Data analysis
 - Feature selection
 - Meta-feature calculation
 - Visualization

Parts of the analysis steps (modell based testing of single features without crossvalidation) are used in a more general form using the *dataAnalysisMisc* package.

2 Cellprofiler data preparation

2.1 Cellprofiler output

Cellprofiler output files are stored in a standardized folder structure (Fig. [1](#)).

Different experimental runs are stored in a folder (*run_n*), with subfolders for each measured 96-well plate (*plate_m*). Within each subfolder, cell profiler output data (*Cells.txt*, *Primarieswithoutboder.txt*, *Cytoplasm.txt*) as well as metadata information (*Treatment.csv*) file are stored. The latter contains information about treatments applied to each well, see. Tbl. [1](#).

3 The R package *cmoRe*

The following sections give a short overview of the functionality implemented in the *cmoRe* package. Not all parameters are outlined in detail, please refer to the package vignette (`vignette(package="cmoRe")`) and documentation (`?fun`) for further information.

¹Currently only implemented as dichotomization, see below.

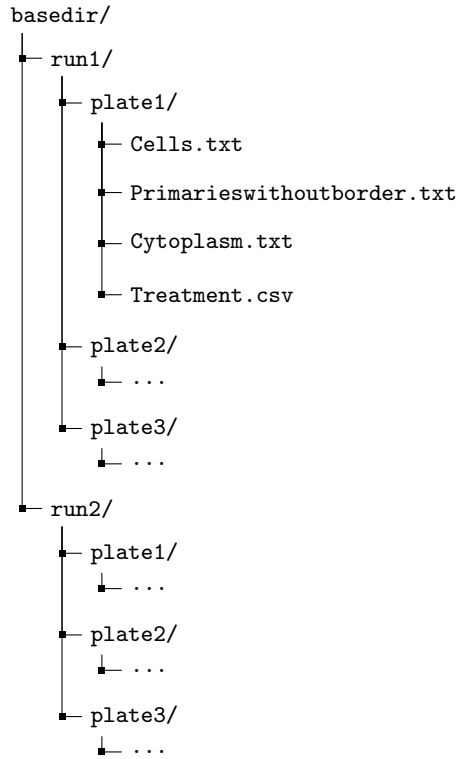


Figure 1: Required structure of *CellProfiler* output data and metadata files.

Well	Treatment	Konzentration
A1	ctrl	1
⋮	⋮	⋮
H12	PE	3

Table 1: Representative information of a *Treatment.csv* file containing the metadata for each plate layout.

Suffix	CP file
.cell	Cells.txt
.nucl	Primarieswithoutborder.txt
.cyto	Cytoplasm.txt

Table 2: Suffixes of features for CellProfiler output files.

3.1 The *imgExp* class

For a given analysis, an *imgExp* class is instantiated (*obj*). Its constructor expects a *list* containing paths to the different experimental runs with the corresponding plates in subfolders and a user specified unique ID (*uid*) (Fig. 1 and Lst. 1).

```

1 # Where is the CP data stored?
  data <- list()
3 data[[1]] <- paste0("run1/", c("plate1/", "plate2/", "plate3/"))
  data[[1]] <- paste0("run2/", c("plate1/", "plate2/", "plate3/"))
5
  # Instantiate imgExp class
7 obj <- new("imgExp", data, "UID1")

```

Listing 1: Instantiating an *imgExp* object.

3.2 Data loading

`getData(obj)` checks completeness, loads and stores data in the `@data` slot of *obj*.

Measurements from different files are merged by *Metadata.Well*, *ImageNumber* and *ObjectNumber* (parameter *mrg* in *loadData()*) per experimental run and plate. Suffixes are added to feature names to denote their origin (Tbl. 2). File-names are specified by the *fn* parameter (CP output files) and *treatF* (metadata file, *Treatment.csv*) in *loadData()*.

3.3 QC plots

Initial quality control plots can be obtained with the `qcPlots(obj)` function to visualize the number of CP recognized cells per well or the distribution of any selected calculated feature per plate, e.g. median cell size (Lst. 2 Fig. 2).

```

1 # Creates a pdf file in folder
  qcPlots(obj, folder="/tmp/", var="AreaShape.Area.cell", fun=median)

```

Listing 2: Representative QC plot for the distribution of median cell size.

3.4 Filtering of single cells

To retain mostly vital, adherent cardiomyocytes with a minimum number of cells per well for subsequent analyses, different filtering steps are implemented.

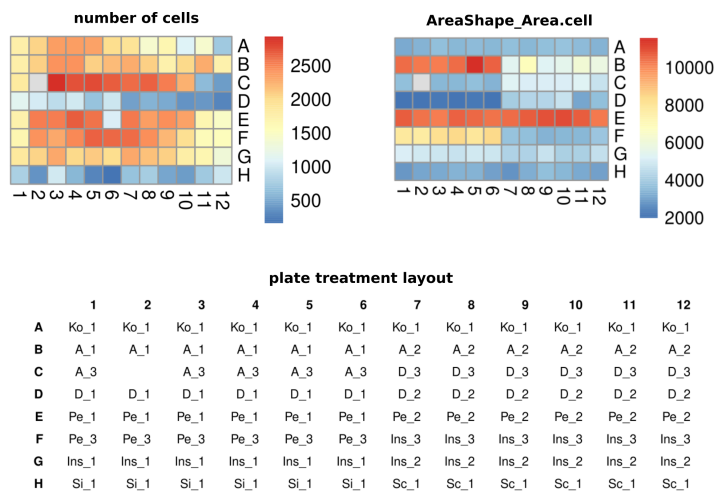


Figure 2: QC plots for numbers of cells, cell size per well (median) and plate treatment layouts.

3.4.1 Detected numbers per well

```
obj <- checkFilter(obj, filter=100)
```

Listing 3: Remove wells with <100 cells per well.

Wells/treatments with only very few identified cells (e.g. for highly cytotoxic substances and too high concentrations) can be removed with the `checkFilter()` function. The `filter` parameter removes all treatment-experimental run-plate combinations containing less than the specified number of cells. Defaults to `NULL` and prints the respective numbers in this case without filtering (Lst. 3).

Furthermore, rows (corresponding to single cells) with unspecified treatments (`is.na(obj@data$TREATMENT) == TRUE`) are removed.

3.4.2 Morphology based

```
1 # Calculate cutoffs
  obj <- calc(obj)
3 # Alternative: calculate only cell cycle cutoffs
  #obj <- calc(obj, fun="cc")
5
7 # Apply cutoffs and filter data
  obj <- filter(obj)
```

Listing 4: Identification of thresholds and assignment per cell.

An alternative, fast method with allows to calculate cutoffs without the need to previously load all data into RAM, is available with the following lines:

```

1 # create new object for demonstration purposes
  obj0 <- new("imageExp", "uid2", data)
3
4 # calculate cutoffs
5 cutoffs <- calcCutoffs(obj0, fun="nc")
6
7 # load data
  obj0 <- getData(obj0)
9
10 #and use the precalculated cutoffs
11 obj0 <- addCutoffs(obj0, cutoffs)

```

Listing 5: Fast identification of thresholds (full CP data) and separated assignment per cell.

Vital, adherent cardiomyocytes are selected by applying three filters on each single cell.

- Differentiation between fibroblasts and cardiomyocytes (fibroblast filter, `fun="fb"`), Fig. 3
- Selection of adherent cells and removal of debris (nuclear-to-cellular ration, `fun="nb"`), Fig. 5
- Selection of vital cardiomyocytes (Assignment of cell-cycle, `fun="cc"`), Fig. 4

The latter two filters rely on the analysis value distributions for identification of cutoffs with given constraints. The general approach is shown in Sec. 3.5.

3.4.3 Fibroblast filter (*fb*)

- feature: *Intensity-MedianIntensity-DNA.nucl*
- transformation: *identity*
- method: identify minimum left of global maximum
- constraints
 - *xMinGlobMax* := 0.2
 - *xMaxGlobMax* := 0.6

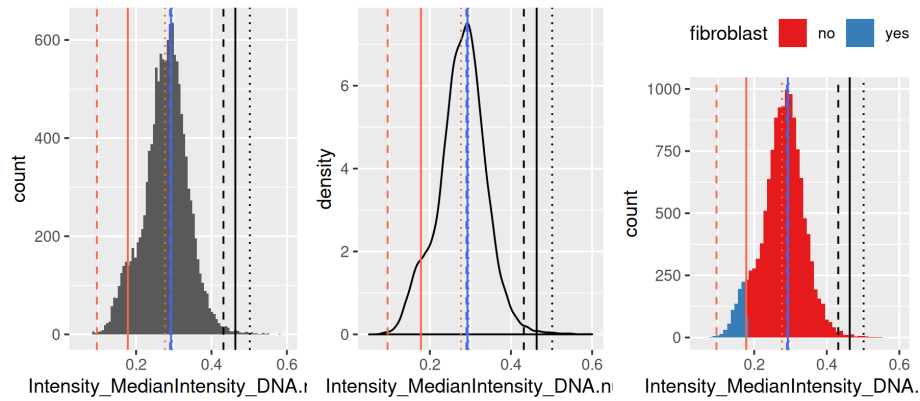


Figure 3: Threshold identification for the differentiation between fibroblasts and cardiomyocytes. Left: histogram with calculated cutoffs (see Sec. 3.5), corresponding density (middle) and histogram with group assignments (right). Solid (median), dashed (10%) and dotted (90%) quantiles. Bold lines: Constraints on cutoff identification ($xCut$, $xMinGlobMax$).

3.4.4 Cell-cycle assignment (cc)

- feature: $Intensity_IntegratedIntensity_DNA.nucl$
- transformation: log
- method: identify global maximum (G1 peak), minimum left: cutoff for dead cells, minimum right: G2 cells
- constraints
 - $xMinGlobMax := log(100)$

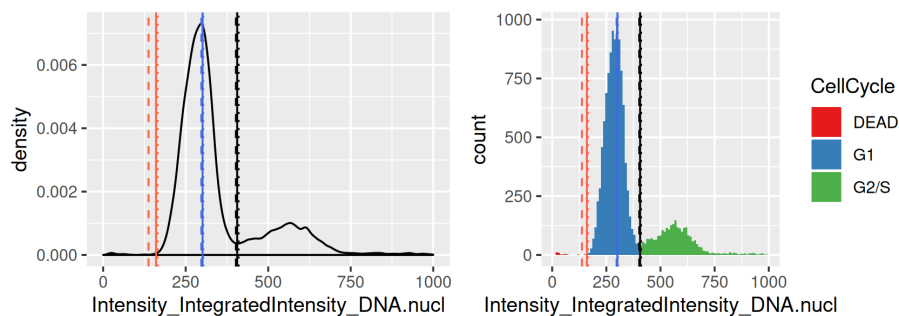


Figure 4: Threshold identification for cell cycle assignment. Left: density with calculated cutoffs (see Sec. 3.5); solid (median), dashed (10%) and dotted (90%) quantiles and histogram of cell-cycle assigned cells (right).

3.4.5 Debris/detached cells filtering (*nc*)

- feature: *AreaShape_Area.cell*, *AreaShape_Area.nucl*
- transformation: *identity*
- method: identify minimum left of global maximum within the interval (*xMinGlobMax*, *xCut*)
- constraints
 - *bw* := 0.1
 - *xCut* := 15
 - *xMinGlobMax* := 1.5

3.5 Threshold identification

The general approach for threshold selection is based on a density analysis of the respective value distribution with given constraints (e.g. cutoff within a pre-specified interval), transformations (*log*, *identity*) and a specific method (e.g. identifying the minimum left of the global maximum)

To account for variability, cutoffs are calculated (and applied) for each experimental run, plate and treatment separately. Robustness of cutoffs is assured by repeatedly performing analyses on resampled data and aggregating the obtained results. 10, 50 (median) and 90% quantiles are calculated. If no cutoffs could be identified for a given combination, data is imputed based on the remaining information.

An overview of utilized constraints is shown in Tbl. 3.

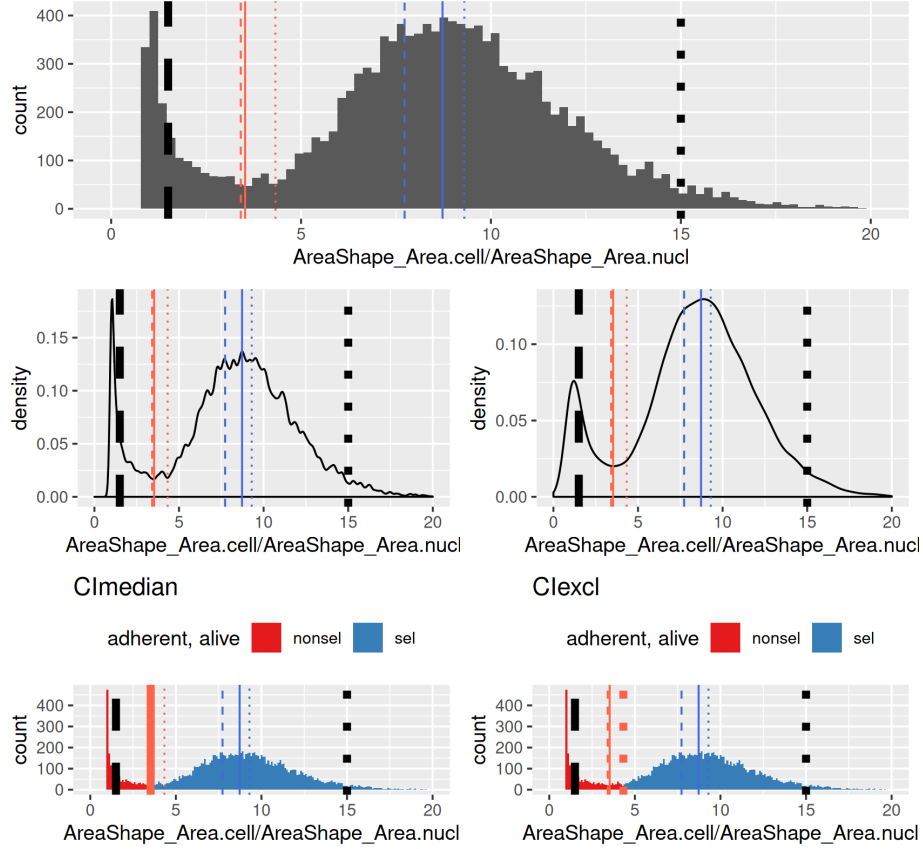


Figure 5: Threshold identification for identification of attached, vital cardiomyocytes. Upper row: histogram of ratio distribution. Middle: different bandwidths for density estimation (left: 0.1, right: auto). Bottom: Differently selected cutoffs (left: median CI_{median} , right: 90% CI_{excl} quantile) of the left minimum. Solid (median), dashed (10%) and dotted (90%) quantiles.

Parameter	Description
bw	Bandwith for R's density function
$xCut$	Cutoff to remove all values above $xCut$ to focus on the relevant data below this threshold,
$xMinGlobMax$	Lower border of the interval in which the global maximum is expected.
$xMaxGlobMax$	Upper border of the interval in which the global maximum is expected.

Table 3: Parameters and constraints for threshold identification.

3.5.1 Cutoff identification

Cutoffs are identified by repeated (e.g. $n=100$) sampling from the available data (with replacement, number of observations equals number of drawings).

First, the density of the resulting data is calculated with default parameters. Obtaining the first and second derivative leads to identification of extremal points from which all maxima and minima are retrieved and the global maximum is selected.

If no value could be identified, *NA* is saved for this iteration, and a warning is printed. Quantiles (default: 10, 50 and 90%) of all relevant extremal points (usually global maximum, left and right minimum) are calculated (while removing *NAs*) and returned.

3.5.2 Imputation of missing data

If no cutoffs could be detected in the previous step, missing data is imputed by calculating the median of all additional thresholds detected for the respective analysis.

3.5.3 Assignment of single cells

In the faster cutoff calculations using `calcCutoffs()` and `addCutoffs()`, only a simplified assignment method is currently provided (*CI*median, see below).

Two methods to assign single cells to their respective category are implemented, *CI*median (default) and *CI*excl. The former assigns each cell based on the 50% quantile (median) cutoff, the latter uses the boundaries of the calculated interval and marks values within the intervals with *NA*.

3.6 Cell-cycle analysis

If we used `calc()` to compute cutoffs, we can use the following code.

```
1 # Plot cell cycle distributions per plate
  analyzeCC(obj)
3
4 # Plot cell cycle distributions aggregated per treatment
5 analyzeCC(obj, agg=T)
```

Listing 6: Obtain cell cycle information.

Alternatively, if we used `calcCutoffs()` and `addCutoffs()`, the relevant fractions are calculated as follows (Lst. [7](#)):

```
1 cc <- sumAgg(obj0)
```

Listing 7: Calculation of cell cycle fractions.

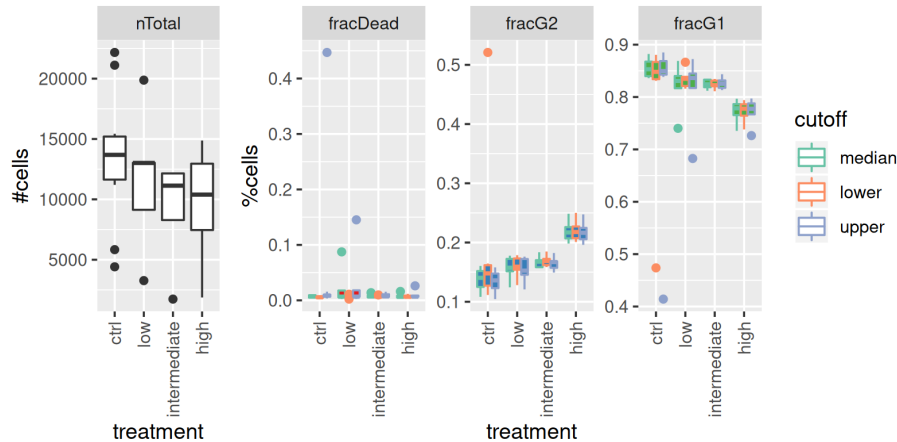


Figure 6: Cell cycle distribution for increasing concentrations for a given treatment. Left: Numbers of cells, right: fractions of cells in the respective cellcycles with different cutoffs used to separate populations.

Cell cycle classification is performed as described in Sec. 3.4.4. All cells, which are included in the `obj@data` *data.frame* at the time of running the `calc(obj)` function² are evaluated. Aggregated visualization and values can be obtained with the `analyzeCC()` function³. The latter returns a *data.frame* with the the total number/fraction of cells per as signed cell-cycle state (three values each, estimate [median], lower and upper interval bounds) (Lst. 6), `sumAgg` a *data.frame* with the fraction of dead, G1 and G2/S cells per well (Fig. 6).

3.7 Additional features

```
1 # vars: vector of feautres (colnames in obj@data)
  obj <- addCutoffVars(obj, vars, fun=log, nBoot=100)
```

Listing 8: Automatic threshold identification for additional features.

The previously applied method to identify cutoffs can be applied to any feature, with an arbitrary transformation (often log). The `addCutoffVars()` function is used to detect cutoffs similarly to the cell cycle analysis.

Currently, only dichotomization based on the identification of the first minimum right of the global maximum is implemented. For identification of the left minimum multiply the respective data by -1. A function for transformation is specified with the `fun` parameter (Lst. 8).

²and all cells present in the CellProfiler output files for `calcCutoffs()` and `addCutoffs()`

³only in combination with `calc()`

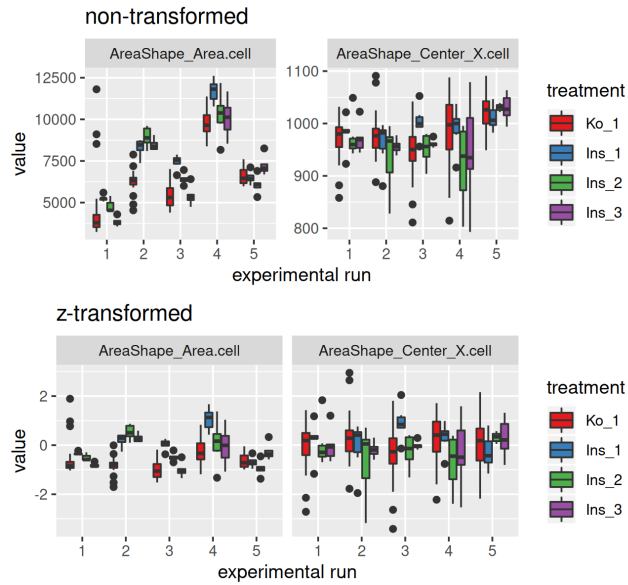


Figure 7: Non- and z-transformed data.

3.8 Aggregation

```
# median aggregation
obj@dataAgg <- medianAgg(obj)
```

Listing 9: Aggregation of single cell data per well

The `medianAgg()` function aggregates data by well (median, Lst. 9). Fractions of binary features can be calculated and added to the aggregated data as shown in Lst. 10.

```
#add fractions of binary features
exp@dataAgg <- cbind(exp@dataAgg, fractAgg(exp))
```

Listing 10: Calculation of dichotomized feature fractions.

3.9 Z-transformation

```
obj <- zTrans(obj)
```

Listing 11: Z-transformation of data.

The `zTrans()` function performs an experimental-run wise z -transformation of data from treatments present in all treatments and for aggregated data (Lst. 11). A minimum of two experimental runs is required (Fig. 7).

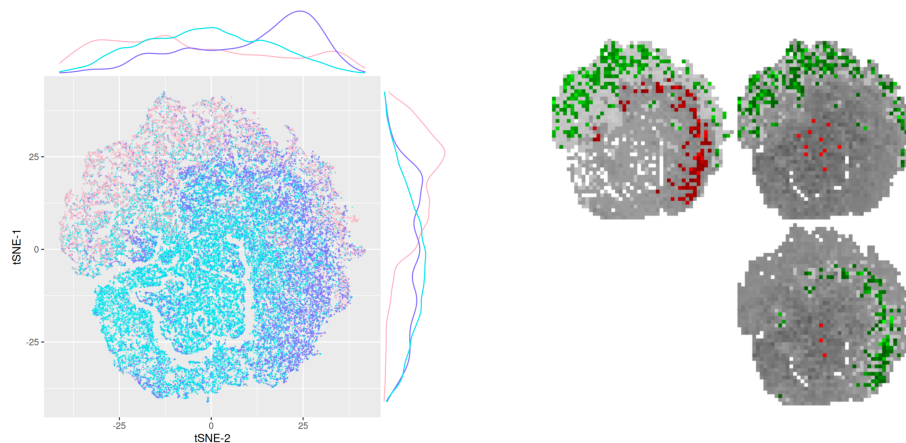


Figure 8: Representative single-cell analysis results. Left: t-SNE of data treated with three treatments. Right: difference maps with 2.5% and 97.5% quantiles used for color code thresholding to highlight differences in population composition.

3.10 Removal of single cell data

```
1 obj <- removeRawData(obj)
```

Listing 12: Remove single cell data.

If no further single cell data analysis is planned, data can be removed to keep the object sizes with the calculated metadata and aggregated data small (Lst. [12](#)), Fig. [8](#)

4 Single cell data analysis

Single cell data can be accessed by the `obj@data` slot as `data.frame` and might be used e.g. in a variety of dimension reduction methods (t-SNE, umap).

5 Well-aggregated data analysis

Analysis of well-aggregated data by mixed model analysis, following feature selection and phenotype visualization is depicted.

5.1 Dose dependent specific alterations

For the assessment of effects for varying dose treatments of NRCMs, a linear mixed effect model was utilized. Three doses were assessed together with non-treated data and a linear relationship was assumed (Tbl. [4](#) Lst. [13](#)).

Dose	Comment
0	control
1	low concentration
2	intermediate concentration
3	high concentration

Table 4: Recoding of concentrations for equidistant assumptions to test for differences.

Dose was evaluated as fixed effect, *treatment* nested under plate (*plate*) and experimental run (*expRun*) were included as random effects (see Lst. 13) using the *lme4* package.

The effect of dose as independent variable on each morphological feature was assessed by likelihood ratio tests of the null and full (differing in *dose* as covariate) model.

Crossvalidation (see Sec. 6) and the testing method are implemented in the `analyze()` function.

treatment: substance, e.g. PE.

dose: assumed equidistant dose concentrations, $dose \in \{0, 1, 2, 3\}$

expRun: experimental runs, $expRun \in \{1, 2, 3, 4, 5\}$

plate: plate per experimental run, $plate \in \{1, 2\}$

val: measurement of a given morphological feature.

```
1 val ~ dose + (1|expRun/plate/treatment)
```

Listing 13: Formula used to identify dose dependent differences.

5.2 Inhibitors

The same approach as outlined in Sec. 5.1 was used to calculate substance specific differences (positive control).

5.2.1 Positive control vs inhibitor

Differences between positive controls and inhibitor data was calculated from meta-features (see Sec. 9) as follows.

i : index of meta features, $i \in \mathbb{N}^+$

m_i^{ctrl} : median aggregated meta feature i value (positive control).

m_i^{tr} : values of meta feature i for treatment tr

The differences d between meta features of treated and control data are defined as follows:

$$d_i^{tr,ctrl} := m_i^{tr} - m_i^{ctrl} \quad (1)$$

d^2 values are then evaluated using a linear mixed effects model to test for differences between controls and inhibitor treated measurements separate for each concentration while adjusting for the respective meta-features (Lst. 5.2.1). Experimental run and well are included as random effects.

meta_feature: meta-feature label (e.g. $\{MF1, MF2, MF3\}$ if three meta features were selected / calculated).

treat: treatment factors to be evaluated for a significant effect on differences (e.g. $\{ins_1, pe_2\}$).

```

1 # fixed effect formula
  d^2 ~ meta_feature + treat
3 # random effect formula
  ~ 1 | expRun/well

```

5.2.2 Negative control and inhibitor

In addition to the the testing for differences between positive control vs positive control substance + inhibitor combinations, the similarity/ equivalence between negative controls and positive control substance + inhibitor was evaluated as follows: Per treatment and dose, a mixed model was fitted with (negative) control data and data from treated cells (substance + inhibitor in a specific dose) with a random effects *experimentalRun* and *treatment* (multiple wells per treatment, Lst. 14) for each meta feature (*SYN_VAR_n / VAL*, Fig. 10).

```
lmer(VAL~1+treatment+(1|experimentalRun/treatment), data=data)
```

Listing 14: Model estimates for controls and substance + inhibitor treated data.

Model estimates for controls and substance+inhibitor treated cells were evaluated, using their 95% confidence intervals. Two qualitative metrics were used to evaluate similarity between model estimates: the number of meta features which show distinct estimate distributions (*different*) and the number of metafeatures for which the treatment estimates lay within the confidence interval of the control estimate (*estim in ref CI*), see Fig. 9. Representative data is shown in Fig. 10.

5.3 TAVI

5.3.1 Concentration dependent regulation

For the TAVI/aortic stenosis experiment, multiple serum concentrations were tested. As it was unknown which serum concentration might yield appropriate results (and an appropriate concentration might differ between features), it was tested for a significant interaction term between the two factors *concentration* and *treatment* (Lst. 15).

treatment $\in \{control, aorticStenosis, postTAVI\}$
concentration $\in \{0.025, 0.5, 1.0, 2.0\}$ *val*: measurement of a given morphologi-

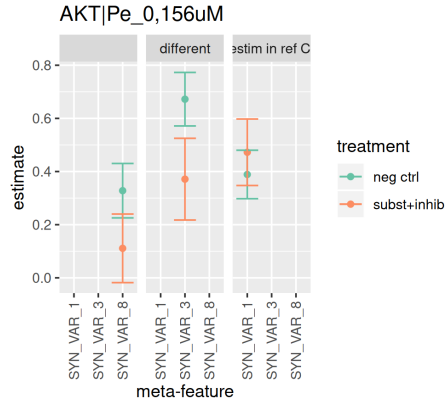


Figure 9: Qualitatively different metrics for the evaluation of similarity between negative controls and substance+inhibitor treated cells on meta-feature level.

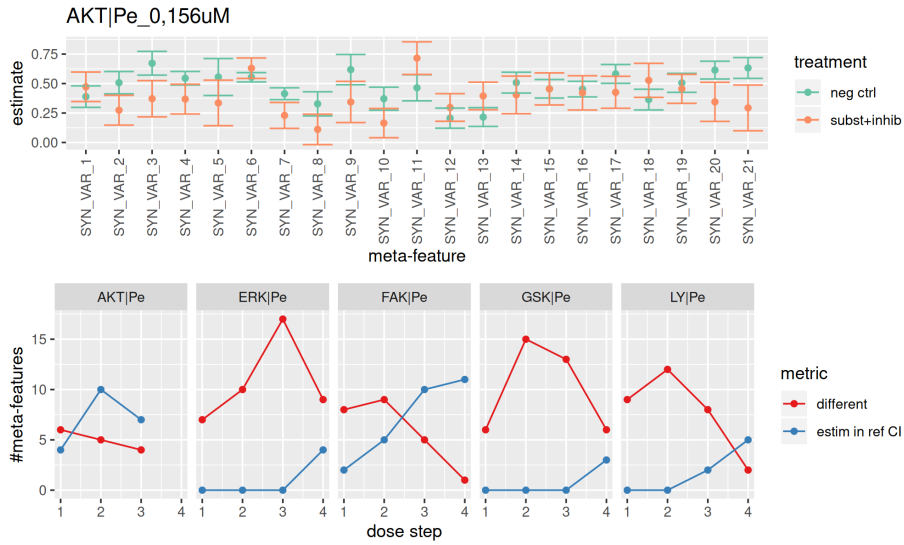


Figure 10: Representative data showing the quantification of similarity between negative controls and substance + inhibitor treated cells for meta-features.

State	Category
0	healthy
1	heterozygous
2	homozygous

Table 5: Encoding of evaluated IPS types for an analysis analogously to the dose-dependent substance approach.

cal feature.

```
1 val ~ treatment*concentration
```

Listing 15: Formula to test for differences in TAVI data.

For the evaluation, features showing a significant interaction term *treatment : concentration* for the comparison *ctrl : concentration* vs *aorticStenosis : concentration* were selected as being specific for aortic stenosis. Meta-features with significant interaction effects *aorticStenosis : concentration* vs *postTAVI : concentration* and inverted effect sized were considered reversible after TAVI.

5.3.2 Concentration independent regulation

Features not showing a significant interaction effect, and thus being regulated independently of the tested concentration, were also included into the evaluation. Analysis was performed as described in Sec. ??, but excluding all features with a non-adjusted interaction p-value above 0.05 or 0.1.

5.3.3 Crossvalidation

Crossvalidation (see Sec. 6) and the testing method are implemented in the `analyzeTAVI()` function.

5.4 IPS cells

For the analysis of (genetically altered) IPS cells, equidistant differences were assumed and differences were assessed with a linear mixed effect model using the *dataAnalysisMisc* package (Tbl. 5, Lst. 16).

```
1 #Patient: patient ID
  res <- randEffAnalysis(dat[, , drop=F], pheno[, ],
3                       frm0=as.formula(VAL~1+(1|Patient)),
                       frm=as.formula(VAL~state+(1|Patient)))
```

Listing 16: Testing for differences in the IPS dataset.

For visualization of phenotypic changes, batch adjusted data was used (see Sec. 7).

6 Crossvalidation

Crossvalidation was performed as follows: measurements (aggregated per well) were left out per treatment and experimental run to train a model, the left out data was predicted with the latter.

\mathbf{v}^{full} : predicted values from the model trained with all data.
 $i \in \mathbb{N}^+$: numbers of models fitted with incomplete data.
 \mathbf{v}_i^{sub} : predicted values of the left out data not used for model training.

$$R_{cv}^2 := 1 - \sum_i (\mathbf{v}^{full} - \mathbf{v}_i^{sub})^2 \cdot \frac{\text{Var}(\mathbf{v}_i^{sub})}{i} \quad (2)$$

$$R_{cv}^2 = \begin{cases} R_{cv}^2 & \text{if } R_{cv}^2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

7 Data prediction

Model based feature selection utilized mixed effects model to estimate random intercepts for specific factors as e.g. different batches. For visualization of meta-feature calculation, non-adjusted data might be utilized or predicted data. The latter was performed by fitting a mixed effects model for each feature with all observed data, and using this model to predict values.

8 Feature selection

Features were selected with a given significance threshold after adjustment for multiplicity of likelihood-ratio test p-values p^* (Bonferroni, Benjamini-Hochberg), R_{cv}^2 and effect size Δ cutoffs.

9 Metafeature calculation

Similar features were aggregated to meta features mf_i based on a hierarchical cluster analysis and the selection of a maximal intracluster variability (Fig. 11).

For metafeature calculation, either directly measured, filtered and transformed data (as described above) was used. Alternatively, for data with even higher variability (TAVI, IPS cells) linear mixed models were fitted for data of each feature, and predicted values were used for following analysis (cluster selection, meta-feature calculation, visualization in radarplots, see Lst. 17 for IPS analyses).

```
fit <- lmer(VAL~Mut_num + (1|Batch), data=data)
prd <- predict(fit, newdata=data)
```

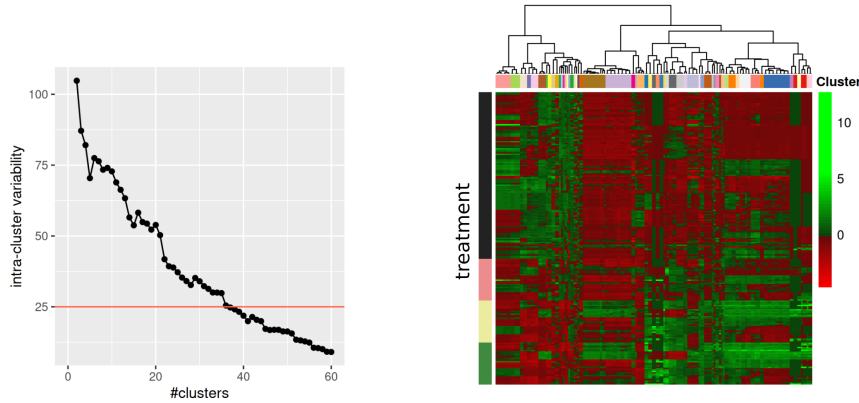


Figure 11: Intra-cluster variability for increasing numbers of clusters (left) and heatmap of selected features with a representative clustering.

Listing 17: Prediction of data per feature.

Using the `findClusters()` function, a range of potential clusters can be evaluated. The latter returns a intra-cluster variability value as a function of assumed clusters (Lst. 18):

$\mathbf{v}_i^{F_j}$: measurement i of feature set F_j corresponding to cluster j .
 $mad(\mathbf{x})$: median absolute deviance of \mathbf{x} .
 F : set of all selected features.
 k : number of clusters with $k := \{i | i \in \mathbb{N}^+ 1 \leq k \leq |F|\}$.

For each number of clusters k to be tested (`nClust` parameter), the following metric m_k is calculated:

$$m_k := \frac{\sum_{j=1}^k mad(\mathbf{v}_i^{F_j})}{k} \quad (3)$$

Ward.D2 clustering (`clustering_method` parameter) and Euclidean distance are used for hierarchical cluster analysis to derive the respective number of clusters (Fig. 11).

```
##calculate metric for 2 to 60 clusters
2 cluster <- findClusters(clusterDat, nClust = c(2,60))
```

Listing 18: Calculation of intracuster variability.

After selection of a desired intra-cluster variability cutoff, such defined similar features are aggregated to meta-features by median aggregation (Lst. 19).

```
## select intra-cluster variability cutoff
```

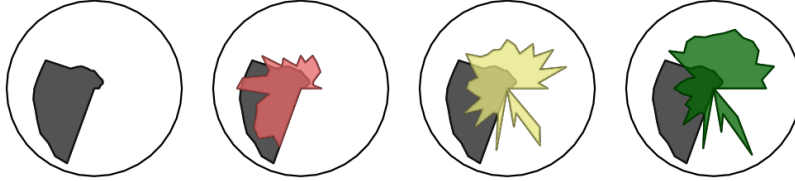



Figure 12: Radarplots of representative data showing meta-features for increasing substance concentrations.

```

2 cutoff <- 100
4 cm <- cluster[order(unlist(cluster[, "dispSum"])),]
# list with feature names and cluster assignment
6 map <- cm[which(cm[, "dispSum"] > cutoff),,drop=F][1,,drop=F][1,"pm.
  clust"][[1]]
8 # calculate meta features
synthMetr <- calculateSyntheticVars(clusterDat, map)

```

Listing 19: Calculation of metafeatures for a specific intra-cluster variability cutoff.

10 Phenotype visualization

Changes in meta-features can be visualized using radar plots (Lst. [20](#), Fig. [12](#)).

```

1 drawRadarplots(plotDat, vars=names(synthMetr$anno)[order(ret
  [[1]][[1]])], labels = F, ctrlLevel="Ko.1", pTest=NULL, agg
  =mean, col=col, main="")

```

Listing 20: Visualization of phenotype changes using radarplots.