# Chapter 13

# High-Dimensional Profiling: The Theta Comparative Cell Scoring Method

**Scott J. Warchal, John C. Dawson, and Neil O. Carragher**

## Abstract

Principal component analysis enables dimensional reduction of multivariate datasets that are typical in high-content screening. A common analysis utilizing principal components is a distance measurement between a perturbagen—such as small-molecule treatment or shRNA knockdown—and a negative control. This method works well to identify active perturbagens, though it cannot discern between distinct phenotypic responses. Here, we describe an extension of the principal component analysis approach to multivariate high-content screening data to enable quantification of differences in direction in principal component space. The theta comparative cell scoring method can identify and quantify differential phenotypic responses between panels of cell lines to small-molecule treatment to support in vitro pharmacogenomics and drug mechanism-of-action studies.

**Key words** Phenotypic screening, High-content analysis, Cell-based profiling

## 1 Introduction

Phenotypic screening allows the identification of treatments that modify a disease model without prior knowledge of the molecular target. This re-emerging method can generate hypotheses for the etiology behind poorly understood diseases, in addition to the discovery of potential therapeutics that act through novel biological mechanisms [1].

One form of phenotypic screening is high-content image-based screening which uses multiple measurements to create a detailed multivariate profile of a perturbagen. This can make screens less biased to prenominated target or therapeutic class hypothesis and also create a phenotypic fingerprint to inform mechanism of action [2–5].

A distinct phenotypic response between cell types which represent the broad heterogeneity of human disease and/or more defined clinical subtypes can highlight differences in cellular signaling, metabolic, and biochemical transporter mechanisms that

Bridget Wagner (ed.), *Phenotypic Screening: Methods and Protocols*, Methods in Molecular Biology, vol. 1787,
https://doi.org/10.1007/978-1-4939-7847-2_13, 

explain the variation of drug efficacy between patients often observed in the clinic. Correlation of distinct phenotypic response and drug sensitivity across genetically distinct cell types with genomic, transcriptomic, and proteomic data can help elucidate compound mechanism of action and identify molecular biomarkers which predict drug sensitivity and clinical outcomes [6, 7]. We can also use phenotypic similarity between different perturbagens to infer mechanistic similarities. One such example is that small molecules which elicit similar cellular phenotypes are likely to have similar mechanisms of action [8]. Phenotypes can also be used to model disease biology where the underlying signaling pathways and molecular targets associated with disease progression are lacking or poorly understood [9].

In order to quantify complex phenotypes, high-content screening generates multivariate datasets in which multiple phenotypic measurements are taken from a single cell or image. These datasets are usually subjected to some form of dimensionality reduction technique in order to make analysis more manageable. A common dimensional reduction method is principal component analysis, which creates new features (principal components) through orthogonal linear combinations of the original features in order to maximize variation. As principal components are ranked in order of variation, a subset of the principal components can be taken as a replacement for the original feature measurements—with the aim of reducing the number of variables while still retaining as much information as possible. This approach helps visualize complex multivariate data points by plotting them in 2D or 3D principal component space [10, 11].

A simple method used to identify active perturbagens in multivariate datasets is a distance measurement such as Euclidean or Mahalanobis distance between the perturbagen and the negative control in principal component space. This can be used to create a threshold distance to separate the active from inactive, as well as rank perturbations on phenotypic activity [11]. However, this distance metric cannot readily discern between different active phenotypes. Two perturbations may produce very different phenotypes and coordinates in principal component space, and yet have similar distances from the negative control.

In order to discern between perturbations such as these we need a measure of directionality. The idea behind the theta comparative cell scoring (TCCS) method is that different directions in phenotypic space indicate different phenotypes. Therefore measuring the angle between perturbagen-induced phenotypes can be used as a phenotypic similarity score independent of potency. This is very similar to the use of cosine similarity, though the TCCS method centers measurements on the negative control and removes inactive perturbagens that may otherwise produce inaccurate measures of directionality.

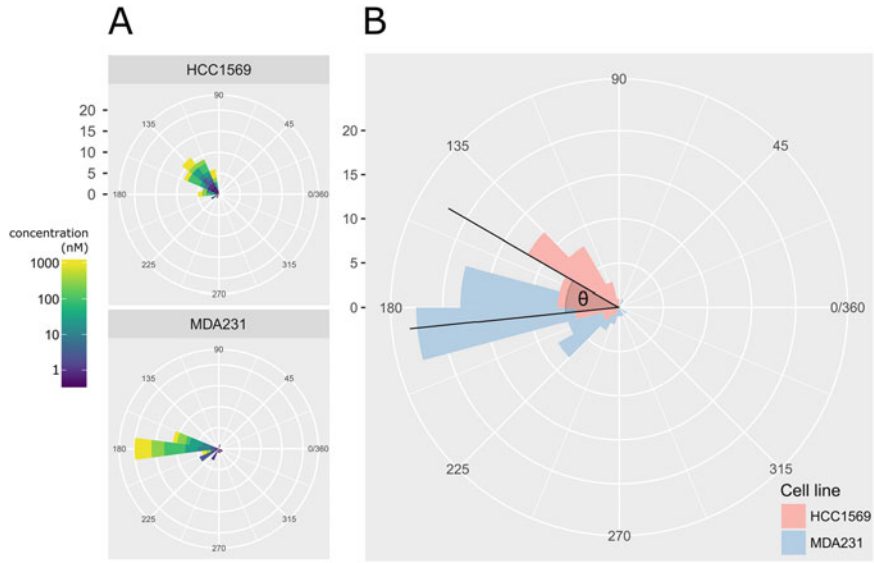


**Fig. 1** Circular histograms showing the similar phenotypic direction of HCC1569 and MDA-MB-231 (MDA231) breast cancer cell lines treated with the aurora kinase inhibitor barasertib. (**a**) Theta values calculated from the first two principal components against a reference vector for both HCC1569 and MDA-MB-231 cell lines treated with barasertib at multiple concentrations. (**b**) Depiction of the $\theta$ value when calculated between a pair of cell lines representing the difference in phenotypic response

The idea of directionality can also be used to produce intuitive and quantitative figures such as circular histograms depicting the direction in phenotypic space or the difference in theta values between two perturbations or samples (Fig. 1).

## 2 Materials

1. Optical-bottom imaging plates (96- or 384-well).
2. Cell culture medium.
3. Trypsin.
4. Perturbagen Library.
5. Paraformaldehyde (PFA).
6. Triton X-100.
7. Wheat-germ agglutinin 594 (WGA), diluted in $dH_2O$.
8. SYTO14 green fluorescent nucleic acid stain.
9. Microtiter plate seals.
10. Aluminum foil.
11. Cell painting stock solution: 10 mg/mL Hoechst 33342, 1 mg/mL concanavalin A (diluted in 0.1 M $NaHCO_3$), 200 U/mL phalloidin-594 (diluted in methanol), 1 mg/mL WGA, 1 mM MitoTracker DeepRed.

12. Blocking buffer: 1% Bovine serum albumin (BSA) in PBS (w/v).
13. Cell painting working solution: 2 μg/mL Hoechst 33342, 11 μg/mL concanavalin A, 3 μM SYTO14, 2.5 U/mL phalloidin-594, 0.25 μg/mL WGA, 600 nM MitoTracker DeepRed.

## 3 Methods

### 3.1 Cell Seeding

Preliminary studies are required to determine the optimal number of cells to seed per well (*see* **Note 1**). This number is dependent on the characteristics of the cell line(s) and the well area in a given plate. Approximate values are provided in Table 1.

1. Using a sub-confluent population of cells, detach the cells by short-term incubation with trypsin and suspend to the desired concentration in cell culture medium.
2. Seed the cells into each well of an optical bottom microtiter 96- or 384-well plate. Make sure that the cells do not settle in the stock of cell suspension by frequently agitating the stock of cell suspension.
3. Incubate the plates containing cells for 24 h.

### 3.2 Compound Addition

1. Make up stock compound plates in DMSO at 1000× the final concentration.
2. Make an intermediate plate by diluting stock compound plate 1:50 in cell culture medium.
3. Remove cell plates from the incubator and transfer a volume from the intermediate plate to the cell plate in a 1:20 dilution.
4. Return cell plates to the humidified, 37 °C, 5% $CO_2$ incubator for an additional 48 h.

### 3.3 Fluorescent Labeling

#### 3.3.1 Fixation

1. Make a solution of 8% paraformaldehyde (PFA) in phosphate-buffered saline (PBS).
2. Add an equal volume of PFA to each well, and incubate at room temperature for 30 min.
3. Wash wells three times with 50 μL of PBS.

**Table 1**
**Approximate cell seeding densities for different plates**

| Plate | Cells/well | Volume/well (μL) |
|---|---|---|
| 96 | 2000–3000 | 100 |
| 384 | 750–1500 | 50 |

#### 3.3.2 Permeabilization

1. Add 30 μL of 0.1% Triton-X100 solution in PBS to each well, and incubate for 20 min at room temperature.
2. Wash wells three times with 50 μL of PBS.

#### 3.3.3 Cell Labeling

Cell labeling protocol adapted from the cell painting protocol [12, 13].

1. Protect the staining solution from light sources by wrapping in aluminum foil.
2. Add 30 μL of cell painting solution and incubate in a dark place at room temperature for 30 min.
3. Wash plate three times with 50 μL of PBS. Do not aspirate the final volume.
4. Seal the plates. If the plates are not imaged immediately, then store them at 4 °C in the dark or wrapped in aluminum foil.

### 3.4 Imaging

1. Set up the microscope to image five channels at 20× magnification. *See* Table 2 for suggested filter settings.
2. Image multiple sites per well; we recommend a minimum of four.
3. Adjust the focus and exposure settings (*see* **Note 2**). These settings should be kept constant between batches and comparable experiments as intensity measurements are a function of exposure time.

### 3.5 Image Analysis

The following image analysis instructions use CellProfiler [14] nomenclature, though other image analysis software packages may be used to achieve similar results.

1. Extract metadata from either the image or the file path; record the date, plate number, plate name, well, site, and channel for each image.

**Table 2**
**Cell painting reagents and suggested filters**

| | | Filter wavelength | |
|---|---|---|---|
| **Stain name** | **Filter name** | **Excitation (nm)** | **Emission (nm)** |
| Hoechst 33342 | DAPI | 377 ± 40 | 447 ± 60 |
| Con A | FITC | 482 ± 35 | 536 ± 40 |
| SYTO14 | Cy3 | 531 ± 40 | 594 ± 40 |
| Phalloidin & WGA | TxRed | 562 ± 40 | 624 ± 40 |
| MitoTracker DeepRed | Cy5 | 628 ± 40 | 692 ± 40 |

2. Add in external metadata from a .csv file such as compound labels or concentrations and match via plate name and well name/position.
3. Assign each image to a channel name using the extracted channel metadata.
4. Segment the nucleus using IdentifyPrimaryObjects.
5. Segment the cell body/cytoplasm using the nucleus object as a seed in the phalloidin/WGA channel with the IdentifySecondaryObjects module.
6. Measure image quality in the DAPI channel using MeasureImageQuality. Out-of-focus images and any debris can usually be detected in the DAPI channel. Image quality can also be measured in all the channels though the MeasureImageQuality module although this will increase analysis time.
7. Measure object size and shape of both the nucleus and cell body with MeasureObjectSizeShape.
8. Measure intensity of the nucleus in the DAPI channel and intensity of the cell body in the other four channels using MeasureObjectIntensity.
9. Measure texture in the channels for Golgi apparatus and actin staining (WGA channel) in the cell body objects, and the DAPI channel for the nuclei objects using MeasureTexture.
10. Measure object neighbors for both nuclei and cell bodies with MeasureObjectNeighbors.
11. Export measurement data as .csv files or to a database, excluding any feature measurements that may not be relevant such as object number or object *x*-*y* position.

### 3.6 Data Analysis

1. Check the data produced by the CellProfiler analysis for any missing rows or columns; these need to be removed as appropriate (*see* **Note 3**).
2. Using the ImageQuality measurements produced by CellProfiler, identify any images that may be out of focus or contain debris and after visually checking the images remove the data relating to that image if necessary (*see* **Note 4**).
3. If the data is at the object-level, i.e., measurements per cell, then aggregate this to a well median, so each measurement describes the median measurement per feature per well.
4. Remove non-informative features (any measurement columns that are not metadata) such as those with zero or very low variance.
5. Remove redundant features, such as one of a pair of features that are very highly correlated with each other. This can be performed by calculating a correlation matrix of the feature

dataset and finding groups of features that have Spearman's correlations greater than 0.95, and then removing all but one of these features from the dataset.

6. Normalize the data to the negative control values on each plate. This is performed by subtracting the median of the negative control for each feature, per plate (*see* **Note 5**).
7. Scale the features. For each feature: subtract the feature mean from each individual value, and then divide by the standard deviation of the feature. This standardizes the features to have a mean of zero and unit variance. This is done otherwise features with large values/small units—such as object area which is measured in pixels—will skew the subsequent statistical methods.
8. Calculate the principal components of the feature data and determine the number of principal components needed to account for a proportion of the variance in the dataset, typically 80–90% (*see* **Note 6**).
9. Remove those principal components that fall outside of this subset.
10. Calculate the negative control medoid, which is the median value for each feature of the negative controls.
11. Adjust the principal component values so that the negative control medoid is centered on the origin (*see* **Note 7**).
12. Calculate the l1-norm (AKA city-block or Manhattan distance) from the negative control medoid to each data point in principal component space.
13. Calculate the l1-norm of each negative control point from the origin and calculate a distance threshold as 2 standard deviations of these negative control distances. Any compound that has a distance less than this threshold from the medoid of the negative control can be labeled as inactive.
14. Once the inactive compounds have been removed, perturbagen similarity can be determined by the angle between perturbagen vectors ($\theta$). In two dimensions—using the first two principal components—this can be visualized on a scatter plot. The $\theta$ value can be calculated in any number of dimensions, although visualization becomes more difficult. The similarity angle can be calculated by the cosine similarity converted to degrees (*see* Eq. (1)). Note that 180° is the value of maximum dissimilarity, where two perturbagens having completely different directions in phenotypic space, with values greater than 180° becoming increasingly similar as they approach 360°. Therefore $\theta$ values are constrained between 0 and 360 by subtracting from 360 any value greater than 180, i.e., $\theta > 180 \rightarrow \theta \coloneqq 360 - \theta$:

$$\theta = \cos^{-1}\left(\frac{u \cdot v}{\|u\|\|v\|}\right) \times \frac{180}{\pi} \quad (1)$$

where $u$ and $v$ are the vectors of principal components for each compound.

15. If two principal components capture a large proportion of the variance in the dataset then a visualization can be made by calculating $\theta$ for every perturbagen against a common reference vector, and then plotting a circular histogram of the $\theta$ values (*see* **Note 8**).
16. Identify cell line pairs treated with the same compound that have significantly different theta values (*see* **Note 9**), indicating a distinct phenotypic response between cell lines to a compound treatment.
17. *See* **Notes 10** and **11** for additional troubleshooting steps.

## 4 Notes

1. Too few cells will provide fewer replicates and may run the risk of having no cells contained in an image if a perturbagen reduces the cell number. Seeding too many cells can mean cells do not form a single monolayer which makes image analysis more difficult. We advise seeding the number of cells to result in approximately 60–70% confluence.
2. After setting the focus for the first channel (DAPI/Hoechst), all additional channel's focus settings are based on these measurements. Therefore adjusting the focus settings for the first channel will also affect all of the other channels, so it is advised to set this first and check a few different wells to ensure that the settings are robust.
3. It is recommended to remove columns containing large amounts of missing numbers. This can often be caused by missing metadata in certain samples, or some features that remain constant between samples—such as Euler number—that may be transformed to missing data entries after scaling or aggregation. Once columns of largely missing data have been removed, rows containing missing values can be removed. Without first removing the missing data columns it is often possible to erroneously remove the entire or large proportions of the dataset when using missing rows as the first step.
4. Out-of-focus images can be detected using ImageQuality_PowerLogLogSlope measurements in the nuclei channel. Images with very low values are likely to be out of focus [15]. Debris such as dust or fibers typically show up in the nuclei channel,

and can be detected by identifying images with a large percentage of saturated pixels.

5. Normalizing to the negative control is a useful step in any plate-based screen to remove any batch effects between plates that may influence the results. It is especially important when comparing effects between cell lines as this converts the values to changes from the negative control for that particular plate; as we expected to have a single cell line per plate this also removes any inherent phenotypic differences between the cell lines, and allows the compound-induced changes to be comparable.
6. The number of principal components required to capture a specified proportion of the variance in the data can be calculated in *R* (assuming that data is numeric feature data), to calculate the value for 80% of the variance:

```
threshold <- 0.8
pca_output <- prcomp(data)
pc_variance <- pca_output$sdev^2
cumulative_proportion_variance <- cumsum(pc_variance) / sum(pc_variance)
n_components <- min(which(cumulative_proportion_variance >= threshold))
```

7. To center the principal component data so that the medoid of the negative control lies on the origin, find the medoid for the negative control values, which is the median value for each feature column for the negative control values; find how much this differs from the origin for each feature; shift all values for each feature by this difference, e.g., in R:

```
centre_control <- function(df, feature_cols, cmpd_col, neg_control = "DMSO") {
    # 1. the median value for the DMSO values for each measured feature
   medioid <- apply(df[df[, cmpd_col] == neg_control, feature_cols], 2, median)
   # 2. calculate the difference from the origin for each medioid position
    delta <- 0 - medioid
   # 3. iterate along columns and adjust to centre the DMSO data
    for (i in seq_along(feature_cols)) {
        feature <- feature_cols[i]
        df[, feature] <- df[, feature] + d[i]
    }
    return(df)
}
```

8. Creating circular histograms: If the principal component vector only contains information regarding two principal components, then we can calculate a $\theta$ value for each perturbagen against a common reference such as (0, 1). This generates a $\theta$ value for each perturbagen which can be plotted as a histogram. Without constraining them, the $\theta$ values are ranged between

0 and 360. As either end of this range is equivalent to the *x*-axis of this histogram can be wrapped round into a circle which can be used to visualize the phenotypic direction induced by a perturbagen (Fig. 1).

9. To identify distinct phenotypic responses between cell lines treated with a perturbagen, a theta value has to be calculated for each pair of cell lines per perturbagen. Cell lines that elicit a similar response to a given perturbagen will produce a low $\theta$ value, indicating that they produce similar phenotypic trajectories, whereas a $\theta$ value approaching 180 indicates opposite phenotypic directions. In our experience a histogram of all measured $\theta$ values produces a log-normal distribution, indicating that most perturbagens produce similar phenotypic response between cell lines.
10. Image analysis can take considerable time for large numbers of images. We recommended using either a computing cluster or a cloud computing service to process many images in parallel.
11. Large .csv files can also cause problems. If files exceed several GBs we recommend users switch to a database format such as SQLite.

## Acknowledgments

This work was supported by a Cancer Research UK Ph.D. Studentship award to the Cancer Research UK Edinburgh Centre.

### References

1. Swinney DC, Anthony J (2011) How were new medicines discovered? Nat Rev Drug Discov 10:507–519
2. Ljosa V, Caie PD, Ter Horst R, Sokolnicki KL, Jenkins EL, Daya S et al (2013) Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J Biomol Screen 18:1321–1329
3. Singh S, Carpenter AE, Genovesio A (2014) Increasing the content of high-content screening: an overview. J Biomol Screen 19:640–650
4. Reisen F, Sauty de Chalon A, Pfeifer M, Zhang X, Gabriel D, Selzer P (2015) Linking phenotypes and modes of action through high-content screen fingerprints. Assay Drug Dev Technol 13:150810081821009
5. Kümmel A, Selzer P, Siebert D, Schmidt I, Reinhardt J, Götte M et al (2012) Differentiation and visualization of diverse cellular phenotypic responses in primary high-content screening. J Biomol Screen 17:843–849
6. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW et al (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483:570–575
7. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S et al (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483:603–607
8. Perlman Z, Slack M, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ (2004) Multidimensional drug profiling by automated microscopy. Science 306:1194–1199
9. Vincent F, Loria P, Pregel M, Stanton R, Kitching L, Nocka K et al (2015) Developing predictive assays: the phenotypic screening "rule of 3". Sci Transl Med 7:293ps15
10. Tanaka M, Bateman R, Rauh D, Vaisberg E, Ramachandani S, Zhang C et al (2005) An unbiased cell morphology-based screen for

new, biologically active small molecules. PLoS Biol 3:0764–0776

11. Caie PD, Walls RE, Ingleston-Orme A, Daya S, Houslay T, Eagle R et al (2010) High-content phenotypic profiling of drug response signatures across distinct cancer cells. Mol Canc Ther 9:1913–1926
12. Gustafsdottir SM, Ljosa V, Sokolnicki KL, Wilson JA, Walpita D, Kemp MM et al (2013) Multiplex cytological profiling assay to measure diverse cellular states. PLoS One 8:e80999
13. Warchal SJ, Dawson JC, Carragher NO (2016) Development of the theta comparative cell scoring method to quantify diverse phenotypic responses between distinct cell types. Assay Drug Dev Technol 14:395–406
14. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O et al (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7:R100
15. Bray M-A, Fraser AN, Hasaka TP, Carpenter AE (2012) Workflow and metrics for image quality control in large-scale high-content screens. J Biomol Screen 17:266–274