Phenotypic distance measures for image-based high-throughput screening

Xian Zhang, Grégoire Pau, Wolfgang Huber, Michael Boutros
xianzhang@gmail.com

October 17, 2016

Contents

1 Introduction

High-throughput image-based screening (also termed high-content screening) has become a popular method in systems biology, functional genomics and drug discovery. Data analysis of high-content screening can be divided into two steps: image quantification and phenotypic analysis. Previously we have developed two R packages, EBImage and imageHTS for image quantification. The current R package phenoDist is designed for measuring the phenotypic distance between treatments (e.g., RNAi, small molecular), in order to identify phenotypes and to group treatments into functional clusters. The package implements various methods to compute phenotypic distance including scaling, principle component analysis, factor analysis [6], Kolmogorov-Smirnov statistics [5], SVM (Support Vector Machine) supervised classification [2], SVM weight vector [3], and SVM classification accuracy [8]. The package also provides functions for phenotype identification, treatment clustering and gene enrichment analysis. In this vignette, we will demonstrate how phenoDist can be used for phenotypic distance calculation, phenotype identification and phenotypic clustering in high-content screening data analysis.

2 Cell feature extraction with imageHTS

Before being analyzed with phenoDist, an image-based screen has to be setup as an *imageHTS* object and analyzed with segmentation and cell feature extraction. A human kinome siRNA screen for HeLa cell morphology is used as an example. Detailed description of the screen can be found in [4, 2]. The screen has been previously analyzed; screen information and data can be accessed remotely through imageHTS at `http://www.ebi.ac.uk/huber-srv/cellmorph/kimorph/`. We first initialize an *imageHTS* object to store the screen information.

```
> library('imageHTS')
```

```
> localPath <- file.path(tempdir(), 'kimorph')
> serverURL <- 'http://www.ebi.ac.uk/huber-srv/cellmorph/kimorph/'
> x <- parseImageConf('conf/imageconf.txt', localPath=localPath, serverURL=serverURL)
```

```
File "conf/imageconf.txt" read.
Number of plates= 3
```

```
Number of replicates= 2
Number of wells= 384
Number of channels= 3
Number of spots= 1
```

```
> x <- configure(x, 'conf/description.txt', 'conf/plateconf.txt', 'conf/screenlog.txt')
> x <- annotate(x, 'conf/annotation.txt')
```

The cell images are then processed with segmentation and feature extraction. The following code is not run in the vignette due to time constraints; we will later download the analysis results remotely.

```
> unames <- setdiff(getUnames(x), getUnames(x, content='empty'))
> segmentWells(x, uname=uname, segmentationPar='conf/segmentationpar.txt')
> extractFeatures(x, uname, 'conf/featurepar.txt')
```

3 Phenotypic distance calculation

Based on the cell feature data, one can design a phenotypic distance measure, to quantitatively indicate how similar two phenotypes (when treated by RNAi for example) are. The phenotypic distance measurement can subsequently be used to identify phenotypes based on the phenotypic distance between samples and negative controls, and perform clustering analysis based on the phenotypic distance between samples [8]. Multiple methods to compute phenotypic distance are implemented in this package. Here we show two examples: one by PCA transformation and euclidean distance; the other by SVM classification accuracy.

```
> library('phenoDist')
```

```
> profiles <- summarizeWells(x, unames, 'conf/featurepar.txt')
> load(system.file('kimorph', 'selectedFtrs.rda', package='phenoDist'))
> pcaPDM <- PDMByWellAvg(profiles, selectedWellFtrs=selectedWellFtrs, transformMethod='PCA',
+ distMethod='euclidean', nPCA=30)
> svmAccPDM <- PDMBySvmAccuracy(x, unames, selectedCellFtrs=selectedCellFtrs, cross=5, cost=1,
+ gamma=2^-5, kernel='radial')
```

The above calculations are not run in the vignette due to time constraints, instead, we load a subset of the pre-calculated svmAccPDM for demonstration purposes.

```
> load(system.file('kimorph', 'svmAccPDM_Pl1.rda', package='phenoDist'))
> dim(svmAccPDM_Pl1)
```

```
[1] 704 704
```

```
> svmAccPDM_Pl1[1:5,1:5]
```

```
           001-01-A03 001-01-A04 001-01-A05 001-01-A06 001-01-A07
001-01-A03         NA       0.94       0.94       0.92       0.94
001-01-A04       0.94         NA       0.77       0.77       0.78
001-01-A05       0.93       0.78         NA       0.63       0.70
001-01-A06       0.91       0.78       0.64         NA       0.70
001-01-A07       0.93       0.78       0.70       0.67         NA
```

Rows and columns of the distance matrix are non-empty wells from the first plate of the three-plate library with two technical replicates. Each value is the phenotypic distance measurement between cell populations from the two corresponding wells, calculated by the SVM classification accuracy method. The distance matrix is not completely symmetric due to random sampling in the SVM cross validation process, but the fluctuation is negligible [8].

With the phenotypic distance matrix, we can assess the reproducibility of the two technical replicates of the screen by comparing the distance between replicates and distance between non-replicates.

```
> ranking <- repDistRank(x, distMatrix=svmAccPDM_Pl1)
> summary(ranking)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    1.00    5.00   34.54   30.62  621.00
```

Lower ranking suggests better reproducibility. When ranking equals 1, the treatment is most similar to its technical replicates.

4 Phenotype identification

Phenotypic distance between a treatment and the negative control indicates how strong the phenotype is. With the phenotype information, we can assess screen quality by calculating replicate correlation and separation between positive and negative controls.

```
> pheno <- distToNeg(x, distMatrix=svmAccPDM_Pl1, neg='rluc')
> df <- data.frame(pheno=pheno,gene=getWellFeatures(x, uname=rownames(svmAccPDM_Pl1),
+ feature='GeneID'))
> df <- df[order(pheno, decreasing=T),]
> head(df)

             pheno  gene
001-02-B10 0.94875 ERBB4
001-02-O17 0.94250 AURKB
001-02-O04 0.93500   UBC
001-02-D04 0.93250 KIF23
001-02-P04 0.93000   UBC
001-02-B11 0.92500 COPB2
```

Shown are the five siRNA treatments with the most significant phenotypes (i.e., highest phenotypic distance to the negative control). With imageHTS, one can view the cell images for certain siRNA treatments.

Replicate reproducibility can also be assessed by calculating correlation coefficient between replicates, and by calculating Z'-factor, which indicates the separation between positive and negative controls [7].

```
> repCorr(x, pheno)

[1] 0.7709275

> ctlSeparatn(x, pheno, neg='rluc', pos='ubc', method='robust')

[1] 0.4221891
```

These two quality control metrics can be used to assess screen quality, or to evaluate different data analysis methods.

5 Phenotypic clustering analysis

Treatments can be clustered based on the phenotypic distance matrix, which will help us understand their functional relationship. Here we clusters the genes with hierarchical clustering.

```
> phenoCluster <- clusterDist(x, distMatrix=svmAccPDM_Pl1, clusterFun='hclust', method='ward')
```

The clustering can be analyzed for GO term enrichment to identify significant gene clusters with the R package GOstats [1]. The following code is not run in the vignette due to time constraints.

```
> library('GOstats')
> GOEnrich <- enrichAnalysis(x, cl=cutree(phenoCluster, k=5), terms='GO', annotation='org.Hs.eg.db',
+ pvalueCutoff=0.01, testDirection='over', ontology='BP', conditional=TRUE)
```

6 Session info

This document was produced using:

```
> toLatex(sessionInfo())
```

- R version 3.3.1 (2016-06-21), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, utils
- Other packages: Biobase 2.34.0, BiocGenerics 0.20.0, EBImage 4.16.0, RColorBrewer 1.1-2, cellHTS2 2.38.0, e1071 1.6-7, genefilter 1.56.0, hwriter 1.3.2, imageHTS 1.24.0, locfit 1.5-9.1, phenoDist 1.22.0, splots 1.40.0, vsn 3.42.0
- Loaded via a namespace (and not attached): AnnotationDbi 1.36.0, BiocInstaller 1.24.0, Category 2.40.0, DBI 0.5-1, DEoptimR 1.0-6, GSEABase 1.36.0, IRanges 2.8.0, MASS 7.3-45, Matrix 1.2-7.1, RBGL 1.50.0, RCurl 1.95-4.8, RSQLite 1.0.0, Rcpp 0.12.7, S4Vectors 0.12.0, XML 3.98-1.4, abind 1.4-5, affy 1.52.0, affyio 1.44.0, annotate 1.52.0, bitops 1.0-6, class 7.3-14, cluster 2.0.5, colorspace 1.2-7, fftwtools 0.9-7, ggplot2 2.1.0, graph 1.52.0, gtable 0.2.0, jpeg 0.1-8, lattice 0.20-34, limma 3.30.0, munsell 0.4.3, mvtnorm 1.0-5, pcaPP 1.9-61, plyr 1.8.4, png 0.1-7, prada 1.50.0, preprocessCore 1.36.0, robustbase 0.92-6, rrcov 1.4-3, scales 0.4.0, splines 3.3.1, stats4 3.3.1, survival 2.39-5, tiff 0.1-5, tools 3.3.1, xtable 1.8-2, zlibbioc 1.20.0

References

[1] S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[2] F. Fuchs, G. Pau, D. Kranz, O. Sklyar, C. Budjan, S. Steinbrink, T. Horn, A. Pedal, W. Huber, and M. Boutros. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol Syst Biol*, 6:370, 2010.

[3] L. H. Loo, L. F. Wu, and S. J. Altschuler. Image-based multivariate profiling of drug responses from single cells. *Nat Methods*, 4(5):445–453, 2007.

[4] G. Pau, X. Zhang, M. Boutros, and W. Huber. Automated analysis of high-throughput imaging screens with imageHTS. In preparation.

[5] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198, 2004.

[6] D. W. Young, A. Bender, J. Hoyt, E. McWhinnie, G. W. Chirn, C. Y. Tao, J. A. Tallarico, M. Labow, J. L. Jenkins, T. J. Mitchison, and Y. Feng. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat Chem Biol*, 4(1):59–68, 2008.

[7] J. H. Zhang, T. D. Chung, and K. R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen*, 4(2):67–73, 1999.

[8] X. Zhang, G. Pau, W. Huber, and M. Boutros. Phenotype identification and clustering in image-based screening with a novel phenotypic distance measure. In preparation.