Pathway Fingerprinting: a working example

Gabriel Altschuler

May 2, 2019

This document demonstrates how to use the `pathprint` package to analyze a dataset using Pathway Fingerprints. The pathprint package takes gene expression data and processes this into discrete expression scores (+1,0,-1) for a set of 633 pathways. For more information, see the pathprint website.

Contents

1 Summary

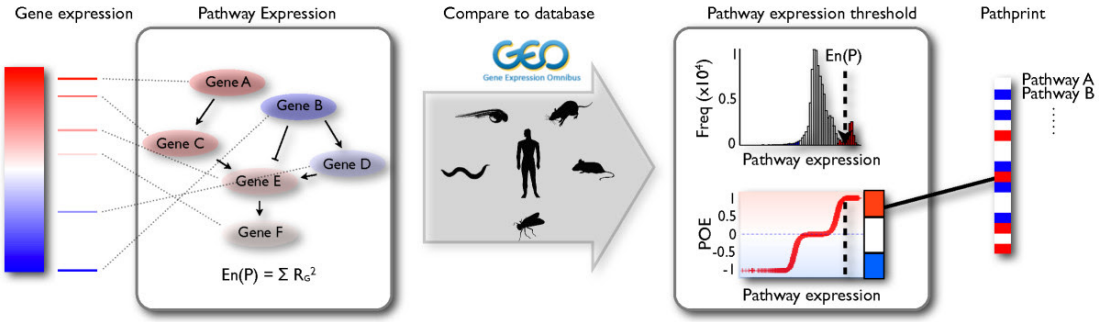
Systems-level descriptions of pathway activity across gene expression repositories are confounded by platform, species and batch effects. Pathprinting integrates pathway curation, profiling methods, and public repositories, to represent any expression profile as a ternary score (-1, 0, +1) in a standardized pathway panel. It provides annotation and a robust framework for global comparison of gene expression data.

2 Background

New strategies to combat complex human disease require systems approaches to biology that integrate experiments from cell lines, primary tissues and model organisms. We have developed Pathprint, a functional approach that compares gene expression profiles in a set of pathways, networks and transcriptionally regulated targets. It can be applied universally to gene expression profiles across species. Integration of large-scale profiling methods and curation of the public repository overcomes platform, species and batch effects to yield a standard measure of functional distance between experiments. A score of 0 in the final pathprint vector represents pathway expression at a similar level to the majority of arrays of the same platform in the GEO database, while scores of 1 and -1 reflect significantly high and low expression respectively.

3 Method

Below we describe the individual steps used to construct the pathway fingerprint.



Rank-normalized gene expression is mapped to pathway expression. A distribution of expression scores across the Gene Expression Omnibus (GEO is used to produce a probability of expression (POE) for each pathway. A pathprint vector is derived by transformation of the signed POE distribution into a ternary score, representing pathway activity as significantly underexpressed (-1), intermediately expressed (0), or overexpressed (+1).

4 Pathway sources

Canonical pathway gene sets were compiled from Reactome, Wikipathways, and KEGG (Kyoto Encyclopedia of Genes and Genomes), which were chosen because they include pathways relating to metabolism, signaling, cellular processes, and disease. For the major signaling pathways, experimentally derived transcriptionally upregulated and downregulated gene sets were obtained from Netpath. We have supplemented the curated pathways with non-curated sources of interactions by including highly connected modules from a functional-interaction network, termed 'static modules.' The modules cover 6,458 genes, 1,542 of which

are not represented in any of the pathway databases. These static modules offer the opportunity to examine the activity of less studied or annotated biological processes, and also to compare their activity with that of the canonical pathways.

Pathprinting: An integrative approach to understand the functional basis of disease Gabriel M Altschuler, Oliver Hofmann, Irina Kalatskaya, Rebecca Payne, Shannan J Ho Sui, Uma Saxena, Andrei V Krivtsov, Scott A Armstrong, Tianxi Cai, Lincoln Stein and Winston A Hide Genome Medicine (2013) 5:68 DOI: 10.1186/gm472

5 Initial data processing

An existing GEO sample on the Human Affy ST 1.0 chip will be used as en example. The dataset GSE26946 profiles expression data from iPS and human ES cells. The R package `GEOquery` can be used to retrieve the data. An 'exprs' object, i.e. a dataframe with row names corresponding to probe or feature IDs and column names corresponding to sample IDs is required by `pathprint`. In addition, we need to know the GEO reference for the platform, in this case GPL6244, and the species, which is 'human' or "Homo sapiens' (both styles of name work).

```
> library(GEOquery)
> GSE26946 <- getGEO("GSE26946")
> GSE26946.exprs <- exprs(GSE26946[[1]])
> GSE26946.exprs[1:5, 1:3]
        GSM663450 GSM663451 GSM663452
7892501  8.904383  9.328561  8.760057
7892502  7.217361  9.118137  6.242542
7892503  6.091620  5.620844  5.726464
7892504 11.072690 10.883280 10.714790
7892505  5.777377  4.814570  4.463360
> GSE26946.platform <- annotation(GSE26946[[1]])
> GSE26946.species <- as.character(unique(phenoData(GSE26946[[1]])$organism_ch1))
> GSE26946.names <- as.character(phenoData(GSE26946[[1]])$title)
```

6 Pathway fingerprinting

6.1 Fingerprinting from new expression data

Now the data has been prepared, the `pathprint` function `exprs2fingerprint` can be used to produce a pathway fingerprint from this expression table.

```
> library(pathprint)
> library(SummarizedExperiment)
> library(pathprintGEOData)
> # load  the data
```

```
> data(SummarizedExperimentGEO)
> # load("chipframe.rda")
> ds = c("chipframe", "genesets","pathprint.Hs.gs","platform.thresholds","pluripotents.frame")
> data(list = ds)
> # extract part of the GEO.fingerprint.matrix and GEO.metadata.matrix
> GEO.fingerprint.matrix = assays(geo_sum_data[,300000:350000])$fingerprint
> GEO.metadata.matrix = colData(geo_sum_data[,300000:350000])
> # free up space by removing the geo_sum_data object
> remove(geo_sum_data)
> # Extract common GSMs since we only loaded part of the geo_sum_data object
> common_GSMs <- intersect(pluripotents.frame$GSM,colnames(GEO.fingerprint.matrix))
> GSE26946.fingerprint <- exprs2fingerprint(exprs = GSE26946.exprs,
                                            platform = GSE26946.platform,
                                            species = GSE26946.species,
                                            progressBar = FALSE
                                            )
[1] "Running fingerprint"
> GSE26946.fingerprint[1:5, 1:3]
                                                GSM663450
Glycolysis / Gluconeogenesis (KEGG)                     0
Citrate cycle (TCA cycle) (KEGG)                        1
Pentose phosphate pathway (KEGG)                        1
Pentose and glucuronate interconversions (KEGG)         1
Fructose and mannose metabolism (KEGG)                  0
                                                GSM663451
Glycolysis / Gluconeogenesis (KEGG)                     0
Citrate cycle (TCA cycle) (KEGG)                        1
Pentose phosphate pathway (KEGG)                        1
Pentose and glucuronate interconversions (KEGG)         0
Fructose and mannose metabolism (KEGG)                  0
                                                GSM663452
Glycolysis / Gluconeogenesis (KEGG)                     0
Citrate cycle (TCA cycle) (KEGG)                        1
Pentose phosphate pathway (KEGG)                        1
Pentose and glucuronate interconversions (KEGG)         1
Fructose and mannose metabolism (KEGG)                  0
```

6.2 Using existing data

The pathprint package uses the object `compressed_result`, drawn from the data-package `pathprintGEOData`, which was constructed in 2012 and does not contain all the GEO data. When uncompressed yields `GEO.fingerprint.matrix` and `GEO.metadata.matrix`. `GEO.fingerprint.matrix` contains 50001 samples that have already been fingerprinted, along with their associated metadata, in the object `GEO.metadata.matrix`. As the above data record is publically available from GEO it is actually already in the matrix and we can compare this to the fingerprint processed above. It should be noted that occasionally there may

be discrepancies in one or two pathways due to the way in which the threshold is applied.

```
> colnames(GSE26946.exprs) %in% colnames(GEO.fingerprint.matrix)
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> GSE26946.existing <- GEO.fingerprint.matrix[,colnames(GSE26946.exprs)]
> all.equal(GSE26946.existing, GSE26946.fingerprint)
[1] TRUE
```

7 Fingerprint Analysis

7.1 Intra-sample comparisons

The fingerprint vectors can be used to compare the differntially expressed functions within the sample set. The most straight forward method to represent this is using a heatmap, removing rows for which there is no change in functional expression.

```
> heatmap(GSE26946.fingerprint[apply(GSE26946.fingerprint, 1, sd) > 0, ],
        labCol = GSE26946.names,
        mar = c(10,10),
        col = c("blue", "white", "red"))
```

7.2 Using consensusFingerprint and fingerprinDistance, comparison to pluripotent arrays

We can also investigate how far in functional distance, these arrays are from other pluripotent fingerprints. This can be achieved using the set of pluripotent arrays included in the package, from which a consensus fingerprint can be created.

```
> # construct pluripotent consensus
> pluripotent.consensus<-consensusFingerprint(
    GEO.fingerprint.matrix[,common_GSMs], threshold=0.9)
> # calculate distance from the pluripotent consensus for all arrays
> geo.pluripotentDistance<-consensusDistance(
    pluripotent.consensus, GEO.fingerprint.matrix)
[1] "Scaling against max length, 540"
> # calculate distance from pluripotent consensus for GSE26946 arrays
> GSE26946.pluripotentDistance<-consensusDistance(
    pluripotent.consensus, GSE26946.fingerprint)
[1] "Scaling against max length, 540"
>
```

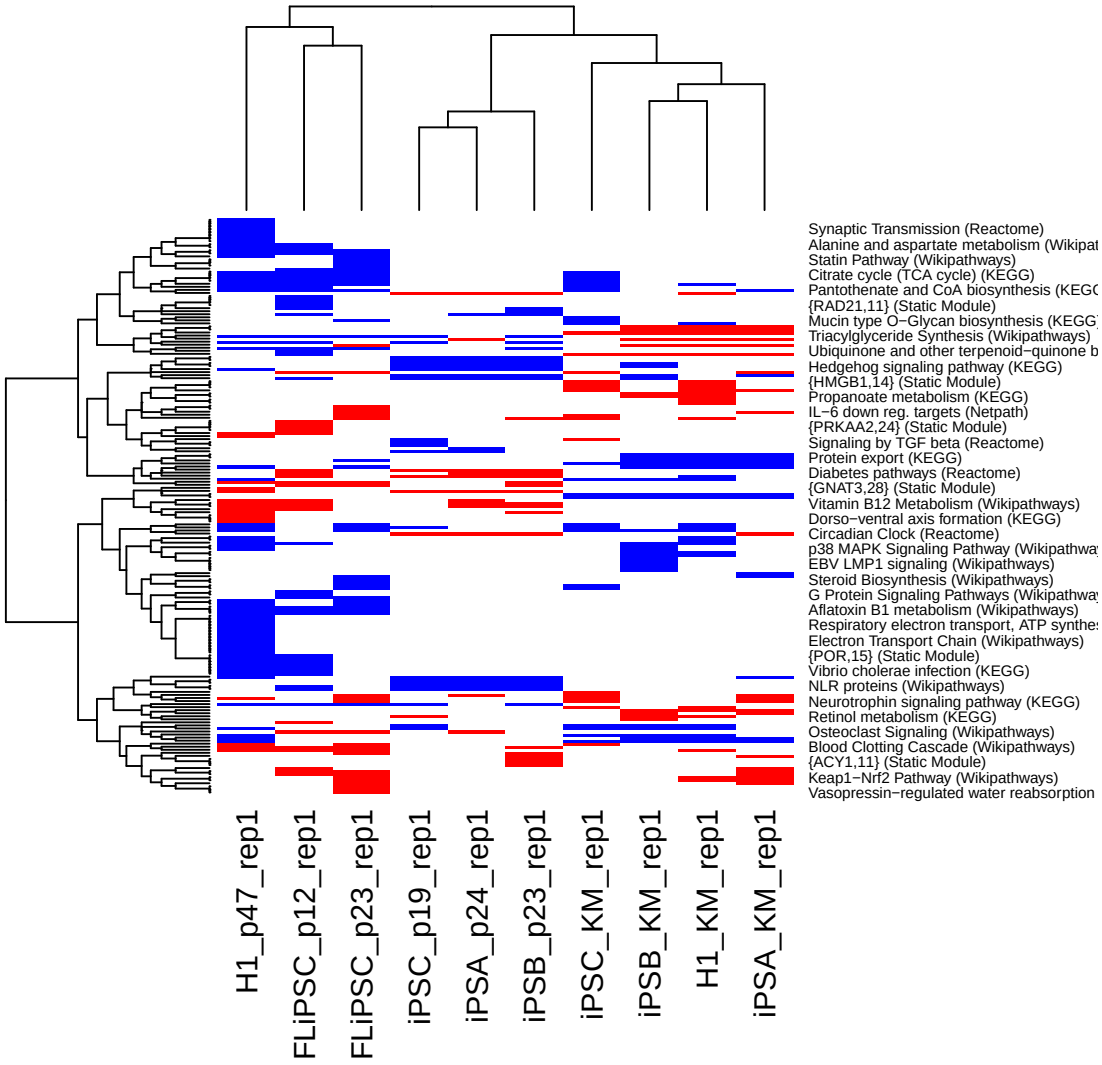


Figure 1: Heatmap of GSE26946 pathway fingerprints, blue = -1, white = 0, red = +1

```
> par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
> geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
     nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
> par(mar = c(7, 4, 4, 2))
> hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
     breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
     main = "", xlab = "")
> hist(GSE26946.pluripotentDistance[, "distance"],
     breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
     main = "", col = "red", add = TRUE)
```

7.3 Identifying similar arrays

We can use the data contained within the GEO fingerprint matrix to order all of the GEO records according to distance from an experiment (or set of experiments, see below). This can be used, in conjunction with the metadata, to annotate a fingerprint with data from the GEO corpus. Here, we will identify experiments closely matched to the H1, embyonic stem cells within GSE26946

```
> GSE26946.H1<-consensusFingerprint(
     GSE26946.fingerprint[,grep("H1", GSE26946.names)], threshold=0.9)
> geo.H1Distance<-consensusDistance(
     GSE26946.H1, GEO.fingerprint.matrix)
[1] "Scaling against max length, 404"
> # look at top 20
> GEO.metadata.matrix[match(head(rownames(geo.H1Distance),20),
                        rownames(GEO.metadata.matrix)),
                    c("GSE", "GPL", "Source")]
DataFrame with 20 rows and 3 columns
                GSE         GPL
          <character> <character>
GSM663458    GSE26946     GPL6244
GSM663459    GSE26946     GPL6244
GSM663455    GSE26946     GPL6244
GSM663453    GSE26946     GPL6244
GSM663454    GSE26946     GPL6244
...               ...         ...
GSM697681    GSE21655     GPL6244
GSM697684    GSE21655     GPL6244
GSM772472    GSE31163     GPL6244
GSM687189    GSE27834     GPL6104
GSM687192    GSE27834     GPL6104
                                          Source
                                     <character>
GSM663458                 Human Embryonic Stem Cells
GSM663459                 Human Embryonic Stem Cells
GSM663455              induced pluripotent stem cells
GSM663453              induced pluripotent stem cells
```

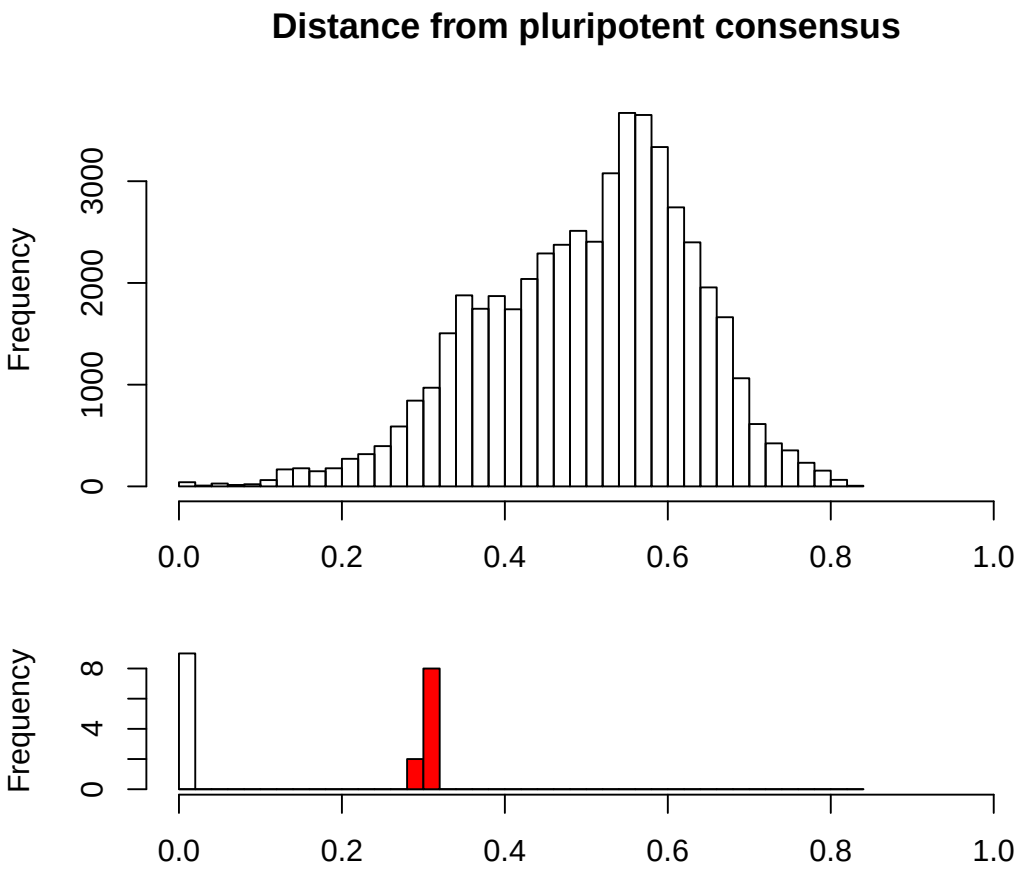


Figure 2: Histogram representing the distance from the pluripotent consensus fingerprint for all GEO (above), curated pluripotent samples (below), and GSE26946 samples (below, red)

```
GSM663454              induced pluripotent stem cells
...                                               ...
GSM697681 hESC maintained under feeder-free conditions
GSM697684 iPSC maintained under feeder-free conditions
GSM772472                           AD-specific iPSC1
GSM687189                           hESCs with SSEA4+
GSM687192                           hESCs with SSEA4+
```