Jetset: selecting an optimal probe set to represent a gene

Qiyuan Li, Aron C Eklund

September 17, 2010

Contents

1 Introduction

On Affymetrix gene expression microarrays, a given gene may be detected by multiple probe sets which may deliver inconsistent or even contradictory measurements. Therefore, obtaining an unambiguous expression estimate of a pre-specified gene can be nontrivial. We developed scoring methods to assess each probe set for specificity, coverage, and degradation resistance. We used these scores to select the optimal probe set for each gene, and thus create a simple one-to-one mapping between gene and probe set.

`jetset` is a package enabling the selection of optimal probe sets from the HG-U95Av2, HG-U133A, HG-U133 Plus 2.0, or U133 X3P microarray platforms.

```
> library(jetset)
```

2 Contents of the package

The `jetset` package contains three types of objects:

```
> ls("package:jetset")

 [1] "eg2ps.hgu133a"       "eg2ps.hgu133plus2"   "eg2ps.hgu95av2"
 [4] "eg2ps.u133x3p"       "gene2ps.hgu133a"     "gene2ps.hgu133plus2"
 [7] "gene2ps.hgu95av2"    "gene2ps.u133x3p"     "scores.hgu133a"
[10] "scores.hgu133plus2"  "scores.hgu95av2"     "scores.u133x3p"
```

In this document, we demonstrate usage with the `*.hgu95av2` objects.

2.1 `scores.hgu95av2`

`scores.hgu95av2` is a data frame with pre-calculated quality control scores for each probe set ID. All scores range from 0 to 1, and a higher score indicates better (predicted) performance.

```
> head(scores.hgu95av2)

          EntrezID process specificity coverage     robust    overall
1000_at       5595   222.0        0.62        1 0.69052115 0.42812311
1001_at       7075   816.0        0.88        1 0.25636974 0.22560537
1002_f_at     <NA>      NA          NA       NA         NA         NA
1003_s_at      643  1949.5        0.69        1 0.03870148 0.02670402
1004_at        643   315.0        0.56        1 0.59129633 0.33112594
1005_at       1843   941.0        0.69        1 0.20811973 0.14360261
```

- The Entrez GeneID (*EntrezID*) is a unique gene identifier. Note: as in other Bioconductor packages, the GeneID is stored as type *character*.
- The processivity requirement (*process*) is the number of consecutive bases that must be synthesized to generate a target that can be detected by the probe set.
- The *specificity* score is the fraction of the probes in a probe set that are likely to detect the targeted gene and unlikely to detect other genes.
- The *coverage* score is the fraction of the splice isoforms belonging to the targeted gene that are detected by the probe set.
- The robustness score (*robust*) is intended to quantify robustness against transcript degradation. The robustness score uses the processivity requirement to estimate the signal intensity of a probe set, relative to the ideal case of perfect processivity.
- The *overall* score is the product of the specificity score, coverage score, and robustness score.

2.2 `eg2ps.hgu95av2`

`eg2ps.hgu95av2` is an character vector, of which the values are probe set IDs and the names are Entrez GeneIDs.

2.3 `gene2ps.hgu95av2`

`gene2ps.hgu95av2` is a “convenience function” that returns the best probe sets matching a list of Entrez GeneIDs, gene symbols, or gene aliases.

3 Simple probe set selection

A typical application of the jetset packages is to identify probe sets corresponding to a published list of genes.

3.1 Selecting by Entrez Gene ID

The Entrez GeneID is the least ambiguous way to specify a gene, although this is not commonly given in publications.

```
> gene2ps.hgu95av2(eg = "2099")
```

```
     2099 
"1681_at" 
```

Entrez GeneID 2099 corresponds to the ESR1 gene.

3.2 Selecting by gene symbol

Often, the official gene symbol is supplied.

```
> gene2ps.hgu95av2(symbol = c("ESR1", "ERBB2", "AURKA"))
```

```
         ESR1         ERBB2         AURKA 
   "1681_at"   "33218_at" "34852_g_at" 
```

3.3 Selecting by gene alias

If the supplied symbols are not the *official* symbols, the gene aliases might be useful.

```
> gene2ps.hgu95av2(alias = c("p53", "HER-2", "K-RAS"))
```

```
       p53      HER-2      K-RAS 
 "1939_at" "33218_at" "35701_at" 
```

4 Probe set quality scores

We can compare quality scores for all probe sets corresponding to a gene of interest. For this example, the STAT1 gene is used because it is detected by several probe sets.

```
> id <- get("STAT1", org.Hs.egSYMBOL2EG)
> id

[1] "6772"

> scores.hgu95av2[which(scores.hgu95av2$EntrezID == id), ]

                            EntrezID process specificity coverage      robust
32860_g_at                      6772   405.5        0.81      0.5 0.508445547
33338_at                        6772   168.5        0.75      0.5 0.754977281
33339_g_at                      6772  1203.0        0.75      1.0 0.134435239
AFFX-HUMISGF3A/M97935_3_at      6772   451.0        0.80      0.5 0.471284303
AFFX-HUMISGF3A/M97935_5_at      6772  2863.5        0.80      1.0 0.008425592
AFFX-HUMISGF3A/M97935_MA_at     6772  1773.5        0.75      1.0 0.051907122
AFFX-HUMISGF3A/M97935_MB_at     6772  1536.0        0.70      1.0 0.077139816
                                overall
32860_g_at                  0.205920447
33338_at                    0.283116480
33339_g_at                  0.100826429
AFFX-HUMISGF3A/M97935_3_at  0.188513721
AFFX-HUMISGF3A/M97935_5_at  0.006740474
AFFX-HUMISGF3A/M97935_MA_at 0.038930342
AFFX-HUMISGF3A/M97935_MB_at 0.053997871
```

We can confirm that `eg2ps.hgu95av2` and `gene2ps.hgu95av2` return the probe set with the highest overall score:

```
> eg2ps.hgu95av2[id]

      6772
"33338_at"

> gene2ps.hgu95av2(eg = id)

      6772
"33338_at"
```

5 Reference

Qiyuan Li, Nicolai Juul, Balazs Gyorffy, Zoltan Szallasi and Aron C. Eklund (2010). Jetset: selecting an optimal microarray probe set to represent a gene. Manuscript in preparation.

6 R sessionInfo

The results in this file were generated using the following packages:

```
> sessionInfo()

R version 2.11.1 (2010-05-31)
x86_64-apple-darwin9.8.0

locale:
[1] C

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods  
[8] base     

other attached packages:
[1] jetset_0.99.0        org.Hs.eg.db_2.4.1   RSQLite_0.9-0       
[4] DBI_0.2-5            AnnotationDbi_1.10.1 Biobase_2.8.0
```