

Jetset: selecting an optimal Affymetrix probe set to represent a gene

Qiyuan Li, Aron C Eklund

June 28, 2016

Contents

1	Introduction	1
2	Contents of the package	2
2.1	jmap	2
2.2	jscores	2
2.3	scores.hgu95av2	2
3	Probe set selection	3
3.1	Selecting probe sets by Entrez Gene ID	3
3.2	Selecting probe sets by ensembl ID	3
3.3	Selecting by gene symbol	3
3.4	Selecting by gene alias	4
4	View probe set quality scores	5
5	Statistics for this release	6
5.1	How many probe sets could be mapped to a gene?	6
5.2	How many genes are mapped to a probe set?	6
6	Jetset algorithm details	7
7	Reference	8
8	R sessionInfo	8

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 the following objects:

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

[1] "jmap"          "jscores"       "scores.hgu133a"
[4] "scores.hgu133plus2" "scores.hgu95av2" "scores.u133x3p"
```

The functions `jmap` and `jscores` are the intended user-level interface, and `scores.*` are data sets that support these functions.

2.1 jmap

`jmap` is a function that returns the best probe sets matching a list of Entrez GeneIDs, gene symbols, or gene aliases.

2.2 jscores

`jscores` is a function that returns the jetset scores for all probe sets matching a list of Entrez GeneIDs, gene symbols, aliases, or ensembl IDs.

2.3 scores.hgu95av2

`scores.hgu95av2` is a data frame with Entrez IDs and 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)
```

	nProbes	EntrezID	process	specificity	coverage
33323_r_at	16	2810	56.5	1	1
32166_at	16	7094	57.0	1	1
38751_i_at	14	521	60.5	1	1
31811_r_at	16	5481	64.5	1	1
34243_i_at	14	26013	64.5	1	1
37941_at	16	4606	66.0	1	1

- 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.

Note that the robustness score and overall score are not stored in this data; instead these scores are calculated on-the-fly when the score data is retrieved using `jscores`.

- 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.

3 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 probe sets by Entrez Gene ID

The Entrez GeneID is an unambiguous way to specify a gene; however the numeric ID is not particularly descriptive, which may explain why this is not commonly provided in publications.

```
> jmap('hgu95av2', eg = "2099")
      2099
"1681_at"

> jmap('hgu133a', eg = "2099")
      2099
"205225_at"

> jmap('hgu133plus2', eg = "2099")
      2099
"205225_at"

> jmap('u133x3p', eg = "2099")
      2099
"g4503602_3p_at"
```

Entrez GeneID 2099 corresponds to the estrogen receptor (ESR1) gene, which is an important indicator of breast cancer phenotype.

3.2 Selecting probe sets by ensembl ID

The ensembl ID is another unambiguous way to specify a gene.

```
> jmap('hgu95av2', ensembl = "ENSG00000091831")
ENSG00000091831
"1681_at"

> jmap('hgu133a', ensembl = "ENSG00000091831")
ENSG00000091831
"205225_at"
```

3.3 Selecting by gene symbol

Often, we know the official HUGO gene symbols, and want the corresponding probe sets.

```
> jmap('hgu133a', symbol = c("ESR1", "ERBB2", "AURKA"))
      ESR1      ERBB2      AURKA
"205225_at" "216836_s_at" "208079_s_at"
```

Unfortunately, the gene symbol for a given gene can change. Furthermore, in rare cases a HUGO gene symbol can correspond to two distinct genes.

3.4 Selecting by gene alias

If we have gene symbols, but they are not the *official* symbols, the gene aliases might be useful.

```
> jmap('u133x3p', alias = c("P53", "HER-2", "K-RAS"))
```

	P53	HER-2	K-RAS
"g8400737_3p_at"	"Hs.323910.2.A1_3p_a_at"		NA

4 View probe set quality scores

We might want to 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.

```
> jscores('hgu95av2', symbol = 'STAT1')
```

	nProbes	EntrezID	process	specificity	coverage
32860_g_at	16	6772	342.0	0.875	0.6666667
AFFX-HUMISGF3A/M97935_3_at	20	6772	424.5	0.900	0.6666667
32859_at	16	6772	74.5	1.000	0.3333333
33338_at	16	6772	168.5	0.875	0.3333333
AFFX-HUMISGF3A/M97935_MB_at	20	6772	1495.0	0.850	0.6666667
33339_g_at	16	6772	1862.5	0.750	1.0000000
AFFX-HUMISGF3A/M97935_MA_at	20	6772	2347.5	0.950	1.0000000
AFFX-HUMISGF3A/M97935_5_at	20	6772	3505.5	0.850	1.0000000

	robust	overall	symbol
32860_g_at	0.565256579	0.329733005	STAT1
AFFX-HUMISGF3A/M97935_3_at	0.492584017	0.295550410	STAT1
32859_at	0.883141136	0.294380379	STAT1
33338_at	0.754977281	0.220201707	STAT1
AFFX-HUMISGF3A/M97935_MB_at	0.082600019	0.046806678	STAT1
33339_g_at	0.044745860	0.033559395	STAT1
AFFX-HUMISGF3A/M97935_MA_at	0.019925319	0.018929053	STAT1
AFFX-HUMISGF3A/M97935_5_at	0.002887471	0.002454351	STAT1

Note that the probe sets are sorted by decreasing overall score.

We can confirm that jmap returns the probe set with the highest overall score:

```
> jmap('hgu95av2', symbol = 'STAT1')
```

STAT1
"32860_g_at"

We can also retrieve quality scores for all probesets:

```
> allscores <- jscores('hgu95av2')
> str(allscores)
```

```
'data.frame':      12625 obs. of  8 variables:
 $ nProbes      : int  16 16 14 16 14 16 16 14 16 16 ...
 $ EntrezID     : chr  "2810" "7094" "521" "5481" ...
 $ process      : num  56.5 57 60.5 64.5 64.5 66 67 70.5 71 74.5 ...
 $ specificity   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ coverage     : num  1 1 1 1 1 1 1 1 1 1 ...
 $ robust       : num  0.91 0.909 0.904 0.898 0.898 ...
 $ overall      : num  0.91 0.909 0.904 0.898 0.898 ...
 $ symbol       : chr  "SFN" "TLN1" "ATP5I" "PPID" ...
 - attr(*, "version.Refseq")= chr "2016-06-22"
 - attr(*, "version.org.Hs.eg.db")= chr "3.3.0"
```

5 Statistics for this release

5.1 How many probe sets could be mapped to a gene?

```
> table(!is.na(scores.hgu95av2$EntrezID))

FALSE  TRUE
 1764 10861

> table(!is.na(scores.hgu133a$EntrezID))

FALSE  TRUE
 3852 18431

> table(!is.na(scores.hgu133plus2$EntrezID))

FALSE  TRUE
18758 35917

> table(!is.na(scores.u133x3p$EntrezID))

FALSE  TRUE
22664 38695
```

5.2 How many genes are mapped to a probe set?

```
> length(na.omit(unique(scores.hgu95av2$EntrezID)))

[1] 8528

> length(na.omit(unique(scores.hgu133a$EntrezID)))

[1] 12292

> length(na.omit(unique(scores.hgu133plus2$EntrezID)))

[1] 20206

> length(na.omit(unique(scores.u133x3p$EntrezID)))

[1] 20094
```

6 Jetset algorithm details

We downloaded probe sequences corresponding to four human gene expression microarrays from Affymetrix: U95Av2, U133A, U133 Plus 2.0, and X3P. We used NCBI BLASTN to search the 25-base probe sequences for matches to the Refseq human RNA database (Pruitt, et al., 2005). The BLASTN search was run with the default parameters, except that filtering was turned off, the word size was set to 8 to increase sensitivity, and the expectation value was set to 1 to reduce output size.

```
blastall -p blastn -d refseq.human.rna -i probe.hgu133a.fa -o hgu133a.refseq.20110817.bls  
-F F -m 8 -e 1 -W 8 -a 8
```

We used the alignment score (bit score) between each probe and cDNA as an indication of probe sensitivity. We defined three levels of alignment: a strong alignment has a score between 48 and 51, indicating that at least 24 bases are identical and that the probe is very likely to detect the target. A moderate alignment has a score between 32 and 47, corresponding to an uninterrupted alignment of length 16 to 23 bases; the probe may or may not respond to the target. A weak alignment has a score less than 32 and is unlikely to respond to the target.

Specificity. A probe was considered to specifically detect a given gene if it aligned strongly to at least one transcript of the gene, but did not have a strong or moderate alignment to a transcript from another gene. The gene specifically detected by the largest number of probes in a probe set was considered the targeted gene of the probe set. We defined the specificity score S_s of a probe set as the fraction of its probes that specifically detect the targeted gene.

Coverage. A transcript of the targeted gene was considered detected by a probe set if the transcript has a strong alignment to the majority of the probes in the probe set. The coverage score S_c of a probe set is defined as the fraction of all transcripts belonging to the targeted gene that are detected by the probe set.

Robustness. The processivity requirement for a probe-transcript alignment is the number of bases between the 5' end of the alignment and the 3' end of the transcript sequence; this corresponds to the length of labeled target that must be synthesized by in vitro transcription to reach the query region. The overall processivity requirement N of a probe set is the median processivity requirement for all strong alignments between probes in the probe set and transcripts in the targeted gene. We define the robustness score S_r of a probe set as the probability that synthesis of the target up to the processivity requirement is achieved without interruption: $S_r = (1 - p)^N$

Here, p is the probability of the IVT synthesis being interrupted at each base, due to either transcript degradation or lack of enzyme processivity. The value of p is likely to be variable in clinical specimens, but for simplicity we use a value corresponding to the manufacturer's design criteria: 1/300 for the X3P array, or 1/600 for the other arrays.

Overall score. We define the overall score S_o as the product of the three scores described above: $S_o = S_s * S_c * S_r$

For a given gene, the probe set targeting this gene with the highest overall score is selected to represent the gene.

7 Reference

Qiyuan Li, Nicolai J. Birkbak, Balazs Gyorffy, Zoltan Szallasi and Aron C. Eklund (2011). Jetset: selecting the optimal microarray probe set to represent a gene. BMC Bioinformatics. 12:474.

8 R sessionInfo

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

```
> sessionInfo()

R version 3.3.0 (2016-05-03)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.11.5 (El Capitan)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

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

other attached packages:
[1] jetset_3.3.0

loaded via a namespace (and not attached):
[1] org.Hs.eg.db_3.3.0   IRanges_2.6.0        parallel_3.3.0
[4] DBI_0.4-1            tools_3.3.0          Biobase_2.32.0
[7] AnnotationDbi_1.34.0 RSQLite_1.0.0        S4Vectors_0.10.1
[10] BiocGenerics_0.18.0 stats4_3.3.0
```