

HumanMethylation27k probes in gene bodies

Tim Triche, Jr.

August 8, 2012

27k probes by design and color

First we have to grab the probes that made it onto the 450k chip.

```
> library('IlluminaHumanMethylation450k.db')
> probes.27k <- IlluminaHumanMethylation450k_get27k()
> lapply(probes.27k, function(x) {
+   if( class(x) == 'list') lapply(x, head)
+   else head(x)
+ })
```

\$I

\$I\$R

```
      Probe_ID
1 cg04037732
2 cg04765675
3 cg04927982
4 cg05113908
5 cg08798116
6 cg09229960
```

\$I\$G

```
      Probe_ID
1 cg02004156
2 cg04344997
3 cg04702045
4 cg05935584
5 cg06700462
6 cg06791102
```

\$II

```
      Probe_ID
1 cg03515901
2 cg08455548
3 cg20401549
4 cg00029931
5 cg00032666
6 cg00060882
```

Annotation

How many probes align to UCSC gene bodies? That's a bit complicated, because each Illumina probe can be mapped to several transcripts, and each transcript to several probes. Normalizing the schema reduced the database size by 100MB.

```
> probes.27k <- unlist(probes.27k, recursive=T)
> head(mget(probes.27k, IlluminaHumanMethylation450kPROBELOCATION, ifnotfound=NA))

$cg04037732
[1] "NM_001166660:1stExon" "NM_181303:1stExon" "NM_018977:5'UTR"
[4] "NM_018977:1stExon" "NM_181303:5'UTR" "NM_001166660:5'UTR"

$cg04765675
[1] "NM_018360:TSS1500" "NM_001168683:TSS1500"

$cg04927982
[1] "NM_001146702:TSS200" "NM_004187:TSS200"

$cg05113908
[1] "NM_001079855:5'UTR" "NM_003918:5'UTR"

$cg08798116
[1] "NM_001448:1stExon"

$cg09229960
[1] "NM_000117:1stExon"
```

Location mapping is many-to-one and emerges as a list of concatenations, so we cannot simply set up a simpleBimap object... UNLESS the concatenation is part of a VIEW (across four or so different tables). So a VIEW is what we use.

```
> probeloc <- mget(probes.27k, IlluminaHumanMethylation450kPROBELOCATION,
+               ifnotfound=NA)
> body.or.exon <- function(x) length( grep('(Body|Exon)', x) ) > 0
> length(which(unlist(lapply(probeloc, body.or.exon))))

[1] 12173

> in.body <- function(x) length( grep('Body', x) ) > 0
> gene.body.probes <- names(which(unlist(lapply(probeloc, in.body))))
> length(which(unlist(lapply(probeloc, in.body))))

[1] 5700

> head(gene.body.probes)

[1] "cg16510010" "cg19963797" "cg20085077" "cg01860753" "cg03085637"
[6] "cg05795157"
```

They are not nearly as scarce as I initially thought.

Versioning

It's important to keep track of where information came from, and who is in charge of keeping it organized, when documenting phenomena like this.

```
> IlluminaHumanMethylation450k_dbInfo()[c(8:10,22:24,31,33),]
      name                                     value
8  EGSOURCEDATE                               2010-Sep7
9  EGSOURCENAME                               Entrez Gene
10 EGSOURCEURL                                ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
22 GPSOURCENAME                               UCSC Genome Bioinformatics (Homo sapiens)
23 GPSOURCEURL                                ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19
24 GPSOURCEDATE                               2010-Mar22
31  MANIFEST                                  HumanMethylation450_15017482_v.1.1.csv
33 MANIFESTDATE                              2010-Dec7

> IlluminaHumanMethylation450kSVNID

[1] "$Id: zzz.R 1291 2011-05-04 23:49:22Z ttriche $"

> IlluminaHumanMethylation450kBLAME

[1] "$Author: ttriche $"
```

Writing to Bioconductor standards simply enforces this.

R session

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

- R version 2.15.0 Patched (2012-05-31 r59485),
x86_64-unknown-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=C, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C,
LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: AnnotationDbi 1.19.28, Biobase 2.17.6,
BiocGenerics 0.3.0, DBI 0.2-5, IlluminaHumanMethylation450k.db 1.4.7,
RSQLite 0.11.1, org.Hs.eg.db 2.7.3
- Loaded via a namespace (and not attached): IRanges 1.15.25,
stats4 2.15.0, tools 2.15.0