

Gene Expression and Methylation from Lung Genomic Research Consortium (LGRC)

J Fah Sathirapongsasuti

October 30, 2021

The data is also available at <https://www.lung-genomics.org/research/>. We provide them here in a processed form to accompany the methods in the package `COPDSexualDimorphism`.

1 Clinical Data

Clinical phenotypes of 254 LGRC samples are given as a `data.frame` named `meta`. It has six fields: `tissueid`, `newid`, `GENDER`, `age`, `cigever`, `pkysrs`, and `diagmaj`. `tissueid` identifies the samples and `newid` identifies the subjects. Some subjects might have more than one sample from left/right/upper/lower lung or blood. These are designated by the last two letters of the tissue ID. The information for these samples have been adjudicated as described in Sathirapongsasuti et al (in review).

```
> library(COPDSexualDimorphism.data)
> `+%` <- function(x,y) paste(x,y,sep=" ")
> data(lgrc.meta)
> head(meta)
```

| | tissueid | newid | GENDER | age | cigever | pkysrs | diagmaj |
|------------|------------|--------|----------|-----|---------------|--------|------------------|
| LT196199RU | LT196199RU | 202158 | 1-Male | 82 | 2-Ever (>100) | 60 | 2-COPD/Emphysema |
| LT073345RU | LT073345RU | 84736 | 1-Male | 74 | 3-Never | 0 | 3-Control |
| LT156041LU | LT156041LU | 299693 | 2-Female | 70 | 2-Ever (>100) | 77 | 2-COPD/Emphysema |
| LT095342LU | LT095342LU | 198904 | 1-Male | 60 | 2-Ever (>100) | 19 | 2-COPD/Emphysema |
| LT155982RU | LT155982RU | 79946 | 2-Female | 48 | 2-Ever (>100) | 28 | 2-COPD/Emphysema |
| LT083759RL | LT083759RL | 221323 | 1-Male | 73 | 2-Ever (>100) | 120 | 2-COPD/Emphysema |

2 Gene Expression

Gene expression profile for 229 LGRC samples are available in two parts. One is `expr`, a matrix of 14497 Ensembl genes (rows) by 229 samples (columns), and the other is `expr.meta`, a `data.frame` of 229 samples (rows) by the subjects' clinical metadata. The subjects are arranged in the same order in the two objects.

```
> data(lgrc.expr)
> data(lgrc.expr.meta)
> dim(expr)
```

```
[1] 14497 229
```

```
> head(expr.meta)
```

| | tissueid | sample_name | newid | GENDER | age | cigever | pkysrs |
|---|------------|-----------------|--------|----------|-----|---------------|--------|
| 1 | LT001098RU | LT001098RU_COPD | 161745 | 2-Female | 46 | 2-Ever (>100) | 35 |
| 2 | LT001796RU | LT001796RU_CTRL | 212671 | 1-Male | 48 | 2-Ever (>100) | 19 |

```

3 LT005419RU LT005419RU_COPD 291396 1-Male 70 2-Ever (>100) 43
4 LT007392RU LT007392RU_COPD 169067 1-Male 46 2-Ever (>100) 45
5 LT009615LU LT009615LU_CTRL 49801 2-Female 49 2-Ever (>100) 45
6 LT010491LL LT010491LL_COPD 180409 1-Male 78 2-Ever (>100) 51
      diagmaj gender
1 2-COPD/Emphysema 2-Female
2      3-Control 1-Male
3 2-COPD/Emphysema 1-Male
4 2-COPD/Emphysema 1-Male
5      3-Control 2-Female
6 2-COPD/Emphysema 1-Male

```

Corresponding to the Ensembl genes in the expression profile is the data frame `genes`. This is a result of a query to BiomaRt database, stored here for convenience.

```

> data(lgrc.genes)
> head(lgrc.genes)

```

```

      ensembl_gene_id hgnc_symbol
ENSG00000000003 ENSG00000000003 TSPAN6
ENSG00000000005 ENSG00000000005 TNMD
ENSG00000000419 ENSG00000000419 DPM1
ENSG00000000457 ENSG00000000457 SCYL3
ENSG00000000460 ENSG00000000460 C1orf112
ENSG00000000938 ENSG00000000938 FGR

```

```

ENSG00000000003 tetraspanin 6 [Source:HGNC Symbol]
ENSG00000000005 tenomodulin [Source:HGNC Symbol]
ENSG00000000419 dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit [Source:HGNC Symbol]
ENSG00000000457 SCY1-like 3 (S. cerevisiae) [Source:HGNC Symbol]
ENSG00000000460 chromosome 1 open reading frame 112 [Source:HGNC Symbol]
ENSG00000000938 Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog [Source:HGNC Symbol]

```

```

      chromosome_name band strand start_position end_position
ENSG00000000003 X q22.1 -1 99883667 99894988
ENSG00000000005 X q22.1 1 99839799 99854882
ENSG00000000419 20 q13.13 -1 49551404 49575092
ENSG00000000457 1 q24.2 -1 169821804 169863408
ENSG00000000460 1 q24.2 1 169631245 169823221
ENSG00000000938 1 p36.11 -1 27938575 27961788

```

```

      ensembl_gene_id.1 entrezgene
ENSG00000000003 ENSG00000000003 7105
ENSG00000000005 ENSG00000000005 64102
ENSG00000000419 ENSG00000000419 8813
ENSG00000000457 ENSG00000000457 57147
ENSG00000000460 ENSG00000000460 55732
ENSG00000000938 ENSG00000000938 2268

```

3 Methylation

Methylation data for 245 LGRC subjects is provided as a data frame `methp` which contains percent methylation for 12094 variably methylated regions (VMRs). Each row provides average median absolute deviation (MAD), length, and the number of probes for a VMR.

```

> data(lgrc.methp)
> methp[1:5, c("name", "ave.mad", "length", "num.probes")]

```

| | name | ave.mad | length | num.probes |
|---|--------------------------|------------|--------|------------|
| 1 | vmr_chr1_932668_932806 | 0.03778364 | 139 | 5 |
| 2 | vmr_chr1_939506_939647 | 0.04619729 | 142 | 5 |
| 3 | vmr_chr1_966705_966843 | 0.05257659 | 139 | 5 |
| 4 | vmr_chr1_989551_989797 | 0.04155331 | 247 | 5 |
| 5 | vmr_chr1_1006424_1006565 | 0.04367978 | 142 | 5 |

4 Session Information

```
> sessionInfo()
```

R version 4.1.1 (2021-08-10)

Platform: x86_64-pc-linux-gnu (64-bit)

Running under: Ubuntu 20.04.3 LTS

Matrix products: default

BLAS: /home/biocbuild/bbs-3.14-bioc/R/lib/libRblas.so

LAPACK: /home/biocbuild/bbs-3.14-bioc/R/lib/libRlapack.so

locale:

```
[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
[3] LC_TIME=en_GB            LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8  LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8     LC_NAME=C
[9] LC_ADDRESS=C             LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

attached base packages:

```
[1] stats      graphics  grDevices  utils      datasets  methods    base
```

other attached packages:

```
[1] COPDSexualDimorphism.data_1.30.0
```

loaded via a namespace (and not attached):

```
[1] compiler_4.1.1 tools_4.1.1
```