Sample relationship verification using 450k DNA Methylation data: beta-values

Maarten van Iterson[*1]

[1]Department of Medical Statistics and Bioinformatics, Section Moleculair Epidemiology, Leiden University Medical Center

*mviterson@gmail.com

2017-07-21

Abstract

Here we will show how you could verify sample relationships on publicly available DNA methylation data, often a beta-value matrix from the Gene Expression Omnibus.

Package

omicsPrint

Contents

Sample relationship verification using 450k DNA Methylation data: beta-values

1 Intro

Here we use the beta-value matrix extracted from GEO (GSE52980). The data was generated to perform an epigenome analysis of human epidermal and dermal samples with aging and sun exposure(Vandiver et al. 2015). Since, multiple samples from the same individual have been used we should be able to verify this using genotypes extracted from the beta-values.

2 Extract beta-values from GEO

First we extract the data from GEO and select the platform on which the DNA methylation was measured (GPL13534).

```
library(Biobase)
library(GEOquery)
# load series and platform data from GEO
gset <- getGEO("GSE52980", GSEMatrix = TRUE, getGPL=FALSE)
## https://ftp.ncbi.nlm.nih.gov/geo/series/GSE52nnn/GSE52980/matrix/
## OK
## Found 3 file(s)
## GSE52980-GPL11154_series_matrix.txt.gz
## GSE52980-GPL13534_series_matrix.txt.gz
## GSE52980-GPL570_series_matrix.txt.gz
idx <- grep("GPL13534", attr(gset, "names"))
betas <- exprs(gset[[idx]])
##or just read from ftp
##betas <- read.table("ftp://ftp.ncbi.nlm.nih.gov/geo/platforms/GPL13nnn/GPL13534/suppl/GPL13534%5FHumanMe
dim(betas)
## [1] 485512     91
```

3 Preprocessing metadata

We need to preprocess the metadata a little to obtain a clear picture of which samples belong to which individuals. Furthermore, some additional (carcinoma) samples have been removed.

```
pheno <- pData(gset[[idx]])[, c("title", "characteristics_ch1", "characteristics_ch1.1",  "characteristics
colnames(pheno) <- c("title", "sex", "age", "exposure", "tissue")
pheno <- apply(pheno, 2, function(x) gsub("^.*: ", "", x))
pheno <- as.data.frame(pheno)
pheno <- subset(pheno, grepl("ermis", tissue)) ##ignore carcinoma samples
tbl <- table(pheno$age)
pheno <- subset(pheno, age %in% names(tbl[tbl == 4])) ##for ease of mapping to same individual
```

```
pheno <- pheno[order(pheno$age),]
nrow(pheno)/4
## [1] 15
pheno$sample_id <- rep(1:15, each=4)
head(pheno)
##                                                    title  sex age
## GSM1255796    genomic DNA from old sun protected dermis 8 male >90
## GSM1255806      genomic DNA from old sun exposed dermis 8 male >90
## GSM1255834 genomic DNA from old sun protected epidermis 8 male >90
## GSM1255844   genomic DNA from old sun exposed epidermis 8 male >90
## GSM1255778 genomic DNA from young sun protected dermis 10 male  20
## GSM1255788   genomic DNA from young sun exposed dermis 10 male  20
##                 exposure    tissue sample_id
## GSM1255796 sun protected    dermis         1
## GSM1255806   sun exposed    dermis         1
## GSM1255834 sun protected epidermis         1
## GSM1255844   sun exposed epidermis         1
## GSM1255778 sun protected    dermis         2
## GSM1255788   sun exposed    dermis         2
betas <- betas[, colnames(betas) %in% rownames(pheno)]
mid <- match(colnames(betas), rownames(pheno))
colnames(betas) <- pheno$sample_id[mid]
dim(betas)
## [1] 485512     60
```

4 Convert beta-values to genotypes

If we would have had the raw idats we could extract the 65 SNPs. However, experiences show that the 65 SNPs available on the array are not enough to perform sample verification with high confidence. Fortunately, several probes on the array contain SNPs occuring frequently in different populations(Chen et al. 2013; Zhou, Laird, and Shen 2016). We have extended the results of Zhou *et al.* using information from the dutch population GoNL.

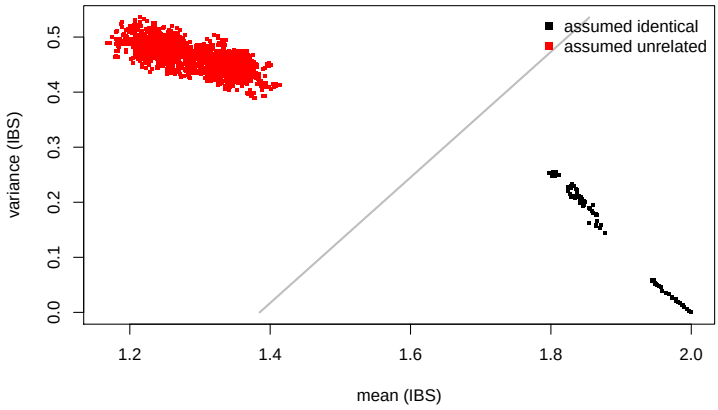
The paper reported that the individuals are from cauciasian origin so we use a european set of SNPs; the GoNL set.

```
library(DNAmArray)
data(hg19.GoNLsnps) ##From DNAmArray
cgs <- hg19.GoNLsnps$probe[hg19.GoNLsnps$distance_c <= 0]
dnamCalls <- beta2genotype(betas[rownames(betas) %in% cgs, ])
dnamCalls[1:5, 1:5]
##            4 7 8 6 3
## cg00004073 2 2 3 2 2
## cg00017157 2 2 2 2 3
## cg00027155 3 2 3 3 3
```

Sample relationship verification using 450k DNA Methylation data: beta-values

```
## cg00033213 2 2 1 2 1
## cg00035449 1 2 2 2 1
dim(dnamCalls)
## [1] 895  60
```

```
data <- alleleSharing(dnamCalls)
## Hash relations
## There are 0 SNP dropped because of low call rate!
## There are 0 sample set to NA because too little SNPs called!
## Using 895 polymorphic SNPs to determine allele sharing.
## Running `square` IBS algorithm!
## 61 of 1830 (3.33%) ...
head(data)
##       mean       var colnames.x colnames.y  relation
## 1 2.000000 0.0000000          4          4 identical
## 2 1.355307 0.4463225          7          4 unrelated
## 3 1.335196 0.4736480          8          4 unrelated
## 4 1.356425 0.4488821          6          4 unrelated
## 5 1.362011 0.4303226          3          4 unrelated
## 6 1.340782 0.4598003          5          4 unrelated
mismatches <- inferRelations(data)
##                   Assumed relation
## Predicted relation identical unrelated
##          identical       150         .
##          unrelated         .      1680
```



The related and unrelated pairs are separated very well and no mismatches are detected.

5 SessionInfo

```
sessionInfo()
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.2 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [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] stats4    parallel  methods   stats     graphics  grDevices
## [7] utils     datasets  base
##
## other attached packages:
##  [1] DNAmArray_0.0.2
##  [2] minfi_1.20.2
##  [3] bumphunter_1.14.0
##  [4] locfit_1.5-9.1
##  [5] iterators_1.0.8
##  [6] foreach_1.4.3
##  [7] Biostrings_2.42.1
##  [8] XVector_0.14.1
##  [9] SummarizedExperiment_1.4.0
## [10] FDb.InfiniumMethylation.hg19_2.2.0
## [11] org.Hs.eg.db_3.4.0
## [12] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
## [13] GenomicFeatures_1.26.4
## [14] AnnotationDbi_1.36.2
## [15] GenomicRanges_1.26.4
## [16] GenomeInfoDb_1.10.3
## [17] IRanges_2.8.2
## [18] S4Vectors_0.12.2
## [19] GEOquery_2.40.0
## [20] Biobase_2.34.0
## [21] BiocGenerics_0.20.0
## [22] omicsPrint_0.99.8
## [23] MASS_7.3-47
## [24] BiocStyle_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] httr_1.2.1              nor1mix_1.2-2
```

```
##  [3] bit64_0.9-7              splines_3.3.2
##  [5] doRNG_1.6.6              blob_1.1.0
##  [7] Rsamtools_1.26.2         yaml_2.1.14
##  [9] RSQLite_2.0              backports_1.1.0
## [11] lattice_0.20-35          quadprog_1.5-5
## [13] limma_3.30.13            digest_0.6.12
## [15] RColorBrewer_1.1-2       preprocessCore_1.36.0
## [17] htmltools_0.3.6          Matrix_1.2-10
## [19] plyr_1.8.4               siggenes_1.48.0
## [21] XML_3.98-1.9             pkgconfig_2.0.1
## [23] biomaRt_2.30.0           genefilter_1.56.0
## [25] zlibbioc_1.20.0          xtable_1.8-2
## [27] RevoUtilsMath_10.0.0     BiocParallel_1.8.2
## [29] annotate_1.52.1          openssl_0.9.6
## [31] tibble_1.3.3             beanplot_1.2
## [33] pkgmaker_0.22            survival_2.41-3
## [35] magrittr_1.5             mclust_5.3
## [37] memoise_1.1.0            evaluate_0.10.1
## [39] nlme_3.1-131             data.table_1.10.4
## [41] tools_3.3.2              registry_0.3
## [43] matrixStats_0.52.2       stringr_1.2.0
## [45] rngtools_1.2.4           base64_2.0
## [47] rlang_0.1.1              grid_3.3.2
## [49] RCurl_1.95-4.8           bitops_1.0-6
## [51] rmarkdown_1.6            codetools_0.2-15
## [53] multtest_2.30.0          DBI_0.7
## [55] reshape_0.8.6            R6_2.2.2
## [57] illuminaio_0.16.0        GenomicAlignments_1.10.1
## [59] knitr_1.16               rtracklayer_1.34.2
## [61] bit_1.1-12               rprojroot_1.2
## [63] stringi_1.1.5            Rcpp_0.12.11
```

Reference

Chen, Y. A., M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg. 2013. "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray." *Epigenetics* 8 (2): 203–9.

Vandiver, A. R., R. A. Irizarry, K. D. Hansen, L. A. Garza, A. Runarsson, X. Li, A. L. Chien, et al. 2015. "Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin." *Genome Biol.* 16 (Apr): 80.

Zhou, W., P. W. Laird, and H. Shen. 2016. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes." *Nucleic Acids Res.*, Oct.