Functional analysis of sites prioritized by limma and glmnet

Nivretta Thatra

The [COHCAP](https://www.bioconductor.org/packages/devel/bioc/manuals/COHCAP/man/COHCAP.pdf) (City of Hope CpG Island Analysis Pipeline) package has annotations available for 450k-UCSC, 450k-HMM and 27k array probes. Specifically, these annotations contain which chromosome, location, gene and CpG island each CpG site maps to. We mapped two lists of CpG sites (those prioritized by glmnet and those prioritized by limma) to 450k annotations.

# Step 1: Glmnet CpG site mapping to chromosome, gene, location, and CpG Island

# Mapped to 450k Hmm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SiteID | Chr | Loc | Gene | Island |
| cg24673385 | 1 | 21051448 | SH2D5 | NA |
| cg15486123 | 1 | 152881815 | IVL | NA |
| cg22853943 | 2 | 4703822 | NA | NA |
| cg13921903 | 2 | 241290446 | NA | 2:240939049-240939317 |
| cg08704934 | 3 | 194826585 | C3orf21 | 3:196307679-196307875 |
| cg14581129 | 12 | 53358946 | NA | 12:51645202-51645774 |
| cg05393297 | 12 | 53359155 | NA | 12:51645202-51645774 |
| cg25025879 | 12 | 53359317 | NA | 12:51645202-51645774 |
| cg16329197 | 12 | 53359506 | NA | 12:51645202-51645774 |
| cg12011926 | 15 | 29037099 | NA | NA |
| cg12602405 | 17 | 79456824 | NA | NA |

**Mapped to 450k UCSC:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SiteID | Chr | Loc | Gene | Island |
| cg24673385 | 1 | 21051448 | SH2D5 | NA |
| cg15486123 | 1 | 152881815 | IVL | NA |
| cg22853943 | 2 | 4703822 | NA | NA |
| cg13921903 | 2 | 241290446 | NA | chr2:241293841-241294408 |
| cg08704934 | 3 | 194826585 | C3orf21 | NA |
| cg14581129 | 12 | 53358946 | NA | chr12:53359192-53359507 |
| cg05393297 | 12 | 53359155 | NA | chr12:53359192-53359507 |
| cg25025879 | 12 | 53359317 | NA | chr12:53359192-53359507 |
| cg16329197 | 12 | 53359506 | NA | chr12:53359192-53359507 |
| cg12011926 | 15 | 29037099 | NA | chr15:29033878-29034710 |
| cg12602405 | 17 | 79456824 | NA | chr17:79454734-79455823 |

## Summary:

The 11 CpG sites are found on chromosomes 1, 2, 3, 12, 15 or 17. The only difference between the two annotations is the CpG islands to which they map.

The CpG sites map to three genes: SH2D5, IVL and C3orf21. Next we looked at which GO terms the three identified genes map to.

# Step 2: Glmnet CpG site-gene GO terms

We used the package mygene for GO term queries.

## Summary:

Gene SH2D5 is involved in postsynaptic density (presumably in neurons), cell junction, and the postsynaptic membrane. Gene IVL is involved in cornfication (hard layer of skin formation), peptide cross-linking, and other protein binding. Gene C3orf21 is a key component of ER membrane, is involved in enzyme transport, and in ion binding.

# Step 3: Limma sites' annotation to chromosome, location, gene and CpG Islands

Repeat step 1 but for sites prioritized by limma, mapping only to 450k-HMM for brevity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SiteID | Chr | Loc | Gene | Island |
| cg16329197 | 12 | 53359506 | NA | 12:51645202-51645774 |
| cg25025879 | 12 | 53359317 | NA | 12:51645202-51645774 |
| cg05393297 | 12 | 53359155 | NA | 12:51645202-51645774 |
| cg14581129 | 12 | 53358946 | NA | 12:51645202-51645774 |
| cg26513180 | 16 | 89883248 | FANCA | 16:88409682-88410944 |
| cg19041462 | 8 | 17146201 | VPS37A | NA |
| cg12011926 | 15 | 29037099 | NA | NA |
| cg05795554 | 16 | 89883212 | FANCA | 16:88409682-88410944 |
| cg16808927 | 16 | 710921 | WDR90 | 16:650649-651454 |
| cg06903451 | 1 | 1323192 | CCNL2 | 1:1312116-1313099 |
| cg12419862 | 22 | 24373484 | LOC391322 | 22:22702803-22703712 |
| cg01172526 | X | 2826829 | ARSD | NA |
| cg10265016 | 2 | 18061319 | KCNS3 | NA |

## Summary:

The 13 CpG sites are found on chromosomes 1, 2, 8, 12, 15, 16, 22 or the X chromosome.

The CpG sites map to seven genes: FANCA, VPS37A, WDR90, CCNL2, LOC391322, ARSD and KCNS3. Now let's look at which GO terms the seven identified genes map to.

# Step 4: Limma CpG site-gene GO terms

Hiding these results because they are quite lengthy. Please see summary below.

## Summary:

FANCA is involved in many processes, most notably in DNA repair, gonad development, and inflammation VPs37A is involved in protein transportation and viral life cycles. WDR90 is involved in protein binding. CCNL2 is involved in transcription and regulation of RNA polymerases. LOC391322 did not have any GO terms. ARSD is involved in many process, most notably lipid metabolic processin and ion binding. KCN3 is involved in ion transport and voltage-gated channel regulation.

There does not seem to be any clear patterns in the functions of genes identified by the two methods. Interestingly, limma found one important gene associated with gonad development, suggesting the importance of the interaction effect between gender and ancestry.

# Notes:

The easiest way to get all annotation information (gene, chromosome, GO ID, GO term, etc) would be to use the package IlluminaHumanMethylation450k.db - but I can't get it to download correctly. I get this error: "ERROR: loading failed \* removing ‘/home/nivretta/R/x86\_64-redhat-linux-gnu-library/3.3/IlluminaHumanMethylation450k.db’ The downloaded source packages are in ‘/tmp/RtmpK8whuD/downloaded\_packages’ installation path not writeable, unable to update packages: cluster, lattice, Matrix, mgcv, nlme, survival Warning message: In install.packages(pkgs = doing, lib = lib, ...) : installation of package ‘IlluminaHumanMethylation450k.db’ had non-zero exit status".

And this seems to be [a documented problem](https://support.bioconductor.org/p/62068/), so I gave up and moved on.

A function from the mygene package which allows many genes to be queried at once – queryMany - does not seem to be working.