Guide to BECon: A tool for interpreting DNA methylation findings from blood in the context of brain

Rachel D Edgar[12]

[1]Department of Medical Genetics, University of British Columbia
[2]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute

April 6, 2017

CpG and Gene Entry

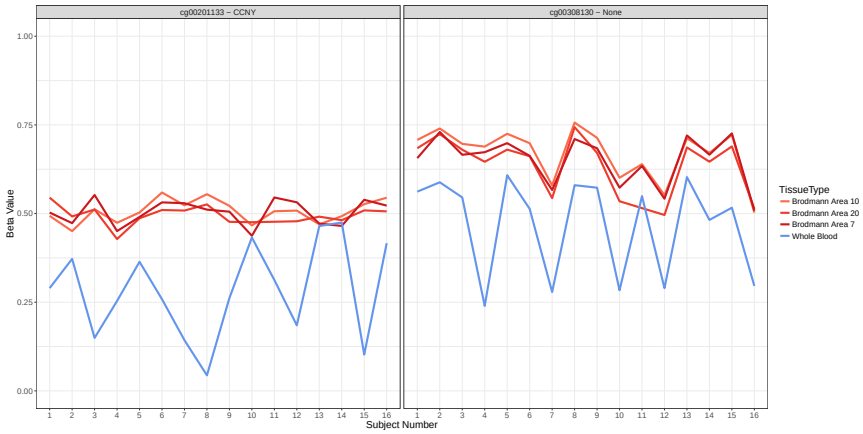
Both CpG IDs and gene names can be queried in BECon. Type the name of the CpG or gene and then select the correct name when it comes up in the drop down menu (note CpG IDs will load slowly). If you want to query many CpG IDs, you may also upload them from a text or csv file. Please format as a list of CpG IDs (one per line) with no header see Table 1.

Comethylation Plots

The comethylation plot generated in the application show the methylation level in each individual. The lines represent the interindividual variation in each tissue, and the lines are coloured by tissue. CpGs are presented as ordered by chromosome then genomic coordinate. The Max CpG Number to Plot will change the number of CpGs shown.

Table 1: Example table for upload to BECon

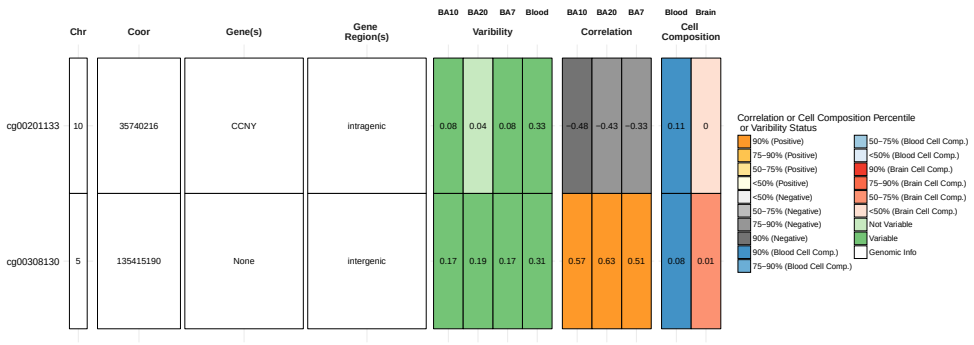
cg05490712
cg14037837
cg27597956
cg15778437
cg26315277
cg05091519
cg01867395



Summary of the Correlations and Variability of Selected CpGs

The summary table gives metrics on each CpG examined. Variability and correlation metrics are calculated from cell composition adjusted data. The columns are as follows:

- CpG ID - The Illumina identifier for the CpG entered. Ordered by chromosome then genomic coordinate.
- Chr (hg19) - Chromosome of the CpG
- Coor -Genomic Coordinate (hg19) of the CpG.
- Gene(s) - The gene(s) associated with the CpG. The associations are explained below (Section). Some CpGs have multiple genes, or isoforms of the same gene so a gene name may appear twice.
- Gene Region(s) - The feature of the gene with which the CpG is associated. There may be multiple features listed if the CpG is associated with multiple gene of isoforms of a gene. This will be listed respectively to the gen names listed.
- Variability - These columns give the reference range variability in blood and brain samples. The reference range variability is the range of the methylation beta values between the 10th percentile and the 90th percentile of all samples at CpG (Lemire et al., 2015). This reference range is intended capture the bulk of the samples to limit the effect of outlier samples at a CpG, giving a falsely high estimate of variability. These columns are colored to in darker green if they exceed the commonly used 0.05 range in beta values in a given tissue.



- Correlation - Spearman correlation values of methylation between blood and the listed brain region. In these columns colors are used to show the percentile of all correlations each value reaches (see legend). Darker grey and orange represent lower or higher correlations, repetitively.
- Cell Composition - The cell composition metrics are the delta betas between data unadjusted for cell composition and data adjusted for cell composition. So how much the beta values change on average at a CpG with cell composition adjustment.

CpG to Gene Annotation

There are multiple approaches to associating a CpG with a gene, such as the closest TSS (Price et al., 2013), presence in a genes body or promoter (Bibikova et al., 2011). We have used a CpG to gene association definition that allows for mulitple gene features, as well as multiple genes. Our inclusive associations is an attempt to capture all possible roles of a CpG in gene regulation. Refseq genes were downloaded from UCSC, including all isoforms of a gene. The gene list included 24,047 genes and a total of 33,431 unique transcription units. If the CpG is 1500bp upstream of a gene's TSS or 300 downstream it is considered to be in the genes promoter. If the CpG is 300bp downstream of the genes TSS to 300 bp upstream of the genes end then it is considered in a genes body (intragenic). If the CpG is 300 bp +- the gene end it is in the 3' region. Importantly, there is a 4th category which closest TSS does not take into account, if a CpG is in none of the above categories it is considered intergenic and not directly associated with a gene. The 485,512 CpGs on the 450K array associated with 23,018 genes (43.8% intragenic, 34.2% promoter, 2.5% 3 region, 19.5% intergenic).

References

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B., and Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295.

Lemire, M., Zaidi, S. H. E., Ban, M., Ge, B., Assi, D., Germain, M., Kassam, I., Wang, M., Zanke, B. W., Gagnon, F., Morange, P.-E., Trgout, D.-A., Wells, P. S., Sawcer, S., Gallinger, S., Pastinen, T., and Hudson, T. J. (2015). Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature Communications*, 6:6326.

Price, M. E., Cotton, A. M., Lam, L. L., Farr, P., Emberly, E., Brown, C. J., Robinson, W. P., and Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin*, 6(1):4.