COHCAP Analysis of CpG Island Methylation for Illumina 450k Methylation Arrays

Charles Warden (cwarden45@gmail.com)
Bioinformatics Core, City of Hope Medical Center

Method Article

Keywords: 450k array, CpG Islands, COHCAP

Posted Date: January 30th, 2014

DOI: https://doi.org/10.1038/protex.2014.002

Abstract

This protocol provides detailed instructions on how to produce quality control metrics, differentially methylated CpG sites, and differentially methylated CpG islands for Illumina 450k methylation array data using the COHCAP Bioconductor package. Scripts for comparing tools to identify differentially methylated regions are also provided, with benchmarks described in the introduction to this protocol.

Introduction

COHCAP[1] \(City of Hope CpG Island Analysis Pipeline) provides a pipeline for analyzing single-nucleotide resolution DNA methylation data produced by either an Illumina methylation arrays or targeted bisulfite sequencing \(BS-Seq). More specifically, COHCAP provides tools to provide quality control metrics, differential methylation analysis \(for CpG sites and CpG islands), and integration with gene expression data. COHCAP was specifically designed in order determine regions showing differential methylation. At the time that the COHCAP paper was published, there was only one other tool that provided this functionality \(Illumina Methylation Analyzer, IMA[2]). Since then, additional tools have been created that provide similar functionality, such as Bumphunter[3] \(in the minfi[4] package), ChAMP[5], and RnBeads[6]. Accordingly, users may be interested in seeing updated benchmarks. Briefly, "Table 1":http://www.nature.com/protocolexchange/system/uploads/2959/original/Table_1.jpg?1390601464 shows that the default parameters for COHCAP can still identify a relatively conservative list of regions with a reasonable run-time, while still providing novel candidates \(Figure "1":http://www.nature.com/protocolexchange/system/uploads/2953/original/Figure_1.jpg?1390345855 - "2":http://www.nature.com/protocolexchange/system/uploads/2955/original/Figure_2.jpg?1390345857 ; IMA comparison is provided in Warden et al. 2013[1]; by default, Bumphunter doesn't provide gene mappings). Among the 4 genes selected for EpiTect validation in the original COHCAP paper, COHCAP shows meets or exceeds the accuracy of all the other algorithms \("Table 2":http://www.nature.com/protocolexchange/system/uploads/2957/original/Table_2.jpg?1390347459). Although this is a limited sample size to access accuracy, we believe the combined evidence from these simple benchmarks indicates that COHCAP is still likely to be a competitive choice to use for differential methylation analysis for 450k array data. When the COHCAP algorithm was initially published, the standalone version of the program could be run either using a GUI or command line. Users had to provide slightly different input files for Illumina array versus targeted BS-Seq. Annotation files had to be provided in all circumstances, and the COHCAP package was initially only compatible with the annotation file for the 450k Illumina methylation array. Since the time of publication, minor feature updates have been added to the standalone version of COHCAP, such as the ability to also read annotation files for 27k Illumina methylation data. More recently, COHCAP has been released as a Bioconductor[2] package. Users familiar with R should find it easier to install the COHCAP package. For example, the Bioconductor pacakge doesn't require Java to be installed. Likewise, it require the user to specify the location of the Rscript file, which was required for the Perl-based standalone version. The Bioconductor version of COHCAP also standardizes the analysis functions for both Illumina array and targeted BS-Seq data, while also providing a new function to help BS-Seq users create the necessary input files \(allowing the use of annotations that are optimized for their own dataset). The majority of COHCAP users have been concerned with analysis of Illumina 450k array data, so this protocol is designed to provide step-by-step instructions on how to analyze such data using the COHCAP Bioconductor package.

Reagents

*Computer with at least 2 GB of RAM \(Linux server required for large data set, but PC and Mac are also OK with the data set used in this protocol)

Equipment

1) R: "http://www.r-project.org/":http://www.r-project.org/ **Please note that COHCAP is currently a 'devel' release** Accordingly, you will need to use R-devel to run COHCAP using the instructions provided on the Bioconductor website. R-devel for Windows: "http://cran.r-project.org/bin/windows/base/rdevel.html":http://cran.r-project.org/bin/windows/base/rdevel.html ****Alternatively, you can download the source files for COHCAP and COHCAPanno and install them using **R CMD INSTALL** \(for any version of R). Please see the instructions below on how to use R CMD INSTALL: "http://cran.r-project.org/doc/manuals/R-admin.html#Installing-packages":http://cran.r-project.org/doc/manuals/R-admin.html#Installing-packages **2) Bioconductor**: "http://www.bioconductor.org/":http://www.bioconductor.org/ ****minfi package: "http://www.bioconductor.org/packages/release/bioc/html/minfi.html ":http://www.bioconductor.org/packages/release/bioc/html/minfi.html ****COHCAP package: "http://www.bioconductor.org/packages/2.14/bioc/html/COHCAP.html":http://www.bioconductor.org/packages/2.14/bioc/html/COHCAP.html **Please note that COHCAP is currently a 'devel' release** This requires a slightly modified installation using R-devel \(see above). This is due to the Bioconductor release schedule. This protocol will be updated when the next version of Bioconductor is released in Spring 2014 **3) Illumina 450k array data**: Click "here":http://sourceforge.net/projects/cohcap/files/Protocol_Exchange_Example.zip/download to download data ****This protocol uses .idat files, but COHCAP can also be used with a table of beta values \(such as those that can be exported from Genome Studio) ****In general, COHCAP is not limited to 450k data. It can also handle targeted BS-Seq data, 27k data, etc. However, this protocol specifically focus on step-by-step analysis for 450k data **4) IGV \(optional)**: used to visualize .wig files Click "here":http://www.broadinstitute.org/igv/ to download IGV **5) UCSC Genome Browser \(optional)**: used to visualize .wig files Click "here":http://genome.ucsc.edu/cgi-bin/hgGateway to view data in the UCSC Genome Browser and/or open custom tracks

Procedure

1) Process Raw Data The input file for COHCAP is a table of beta values. Users should typically be able to produce this table using Genome Studio. Click "here":https://docs.google.com/uc?id=0B1xpw6_kQMKuVm1kS3V6dlJBbmc&export=download&revid=0B1xpw6_kQMKuQ25yS0RDQ3JlYWNKTEk0THRGZmxQQjVGTU40PQ for more detailed instructions on using this strategy. The main benefit to using this input file format is that any sort of pre-processing can be preformed upstream of COHCAP.

Although users can almost always find data matrices with beta values, it may not always be possible to find raw .idat files in the precise format that can be processed using Genome Studio. To avoid this need, this protocol will use the _minfi_ package for pre-processing. The cell line data set used in the COHCAP publication will be used as an example for this protocol. **1a) Download raw data**: Because the "GEO entry":http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42308 doesn't provide raw .idat files, please download the data from this link: "http://sourceforge.net/projects/cohcap/files/Protocol_Exchange_Example.zip/download":http://sourceforge.net/projects/cohcap/files/Protocol_Exchange_Ex Please note that this file also provides a template for running this protocol as well as the benchmarks described in the introduction. **1b) Run minfi to create COHCAP input file** First, open R. Next, set the working directory to the extracted "Protocol_Exchange_Example" folder. This can be accomplished using the **setwd\()** command in R. The template script assumes that the .zip file has been extracted on the Desktop of a Windows computer \(but it still requires some modification to the file path to work on your own computer). Using the example data set, run the following code: **==== start code====** **library\(minfi)** **idat.folder <- "5958154021"** **RG.raw <- read.450k.exp\(idat.folder)** **methyl.norm <- preprocessIllumina\(RG.raw, bg.correct = TRUE, normalize = "controls")** **beta.table <- getBeta\(methyl.norm)** **probes <- rownames\(beta.table)** **output.table <- data.frame\(SiteID=probes, beta.table)** **beta.file <- "minfi.txt"** **write.table\(output.table, file=beta.file, sep="\t", quote=F, row.names=F)** **====end code====** The purpose of the code above is to create a text file \(minfi.txt) that will be used as the input file for COHCAP. **2) Annotation and Quality Control** **==== start code====** **library\(COHCAP)** **sample.file <- "COHCAP_sample_description.txt"** **project.folder <- getwd\()** **project.name <- "COHCAP_450k_Protocol_Exchange"** **beta.table <- COHCAP.annotate\(beta.file, project.name, project.folder, platform="450k-UCSC")** **COHCAP.qc\(sample.file, beta.table, project.name, project.folder)** **====end code====** The above code should produce a table of annotated beta values in the "Raw_Data" folder as well as the following quality control files in the "QC" folder: 2a) "COHCAP_450k_Protocol_Exchange_cluster.pdf":http://www.nature.com/protocolexchange/system/uploads/2961/original/COHCAP_450k_Protocol_Express_c 1390603806 : allows user to see if samples within the same group cluster together 2b) **COHCAP_450k_Protocol_Exchange_descriptive statistics.txt**: contains minimum, maximum, median and 25th/75th percentile beta values; intented to complement the sample histogram 2c) "COHCAP_450k_Protocol_Exchange_hist.pdf":http://www.nature.com/protocolexchange/system/uploads/2963/original/COHCAP_450k_Protocol_Express_hist 1390603831 : allows use to see if any samples appear to have an abnormal beta distribution \(in which case, we would recommend those samples should be removed as outliers); the descriptive statistics file can help users identify the specific outliers\(s) 2d) "COHCAP_450k_Protocol_Exchange_pca.pdf":http://www.nature.com/protocolexchange/system/uploads/2965/original/COHCAP_450k_Protocol_Express_pca 1390603854 : allows user to see if samples within the same group cluster together 2e) **COHCAP_450k_Protocol_Exchange_pca.txt**: includes all principal components, allowing use to produce additional figures it they are not satisfied with the default 2D PCA plot **3) CpG Site Analysis** **==== start code====** **filtered.sites <- COHCAP.site\(sample.file, beta.table, project.name, project.folder, ref="parental")** **====end code====** The code produces the following files: 3a) **COHCAP_450k_Protocol_Exchange_CpG_site_filter.xlsx**: table of differentially methylated CpG sites 3b) **COHCAP_450k_Protocol_Exchange_wig** folder: contains .wig files for visualization using IGV or the UCSC Genome Browser Visualizing these files will be described in the next section, but you can at least confirm the following files have been created: **mutant.avg.beta.wig**: average beta values per CpG site across all samples in the "mutant" group **mutant.vs.parental.delta.beta.wig**: average delta beta values per CpG sites for the average beta across all samples in the "mutant" group subtracted by the average beta across all samples in the "parental" group **parental.avg.beta.wig**:average beta values per CpG site across all samples in the "parental" group **4) CpG Island Analysis** **==== start code====** **filtered.islands <- COHCAP.avg.by.island\(sample.file, filtered.sites, beta.table, project.name, project.folder, ref="parental")** **====end code====** The code produces the following files: 4a) **COHCAP_450k_Protocol_Exchange_CpG_island_filtered-Avg_by_Island.xlsx**: statistics for differentially methylated CpG islands 4b) **COHCAP_450k_Protocol_Exchange_Box_Plots**: folder containing box-plots for differentially methylated CpG islands. For example, the box-plot for chr18:55862653-55862873 \(mapped to gene NEDD4L, corresponding to the file "NEDD4L_chr18_55862653_55862873.pdf":http://www.nature.com/protocolexchange/system/uploads/2967/original/NEDD4L_chr18_55862653_55862873.jp 1390852235) is shown below: This CpG island will be used as an example to demonstrate how you can use the .wig files \(in the "CpG_Site" folder) to visualize methylation patterns anywhere in the genome. **Visualizing CpG Site Methylation via IGV[8] \(Optional)** Step #1) Download IGV: "http://www.broadinstitute.org/software/igv/download":http://www.broadinstitute.org/software/igv/download Step #2) Open IGV Step #3) Make sure the genome reference is set to hg19 Step #4) Import the .wig files using "File –> Load from File". If desired, you can also load other tracks \(such as CpG Islands) the same way. Step #5) Enter the desired region of interest \(can be coordinates in the form chrA:pos1-pos2 or a gene symbol). As an example, we will visualize the regions summarized in the box plot above \(chr18:55862653-55862873) Step #6) To make visulization easier, you may wish to expand the tracks by doing to "Tracks –> Fit Data to Window" Step #7) At any time, you can export a screenshot of what you view in IGV using "File –> Save Image". Here is a screenshot for the region listed above \(also see "Figure 7":http://www.nature.com/protocolexchange/system/uploads/2969/original/IGV_chr18_55862653_55862873.png?1390853127): **Visualizing CpG Site Methylation via UCSC Genome Browser[9] \(Optional)** Step #1) Go to the UCSC Genome Browser web portal: "http://genome.ucsc.edu/cgi-bin/hgGateway":http://genome.ucsc.edu/cgi-bin/hgGateway Make sure you are using hg19 as your human reference sequence. Step #2) Click the button for "add custom tracks" Step #3) Upload your .wig files by clicking "Select File" and then "Submit". This will need to be done separately for each .wig file. Keep clicking "add custom tracks" until all necessary files are uploaded. Unlike IGV, you do not need to do anything special to upload tracks like the UCSC CpG Island locations. Step #4) Once all of the .wig files are uploaded, click "go to genome browser" Step #5) Enter the desired region of interest \(can be coordinates in the form chrA:pos1-pos2 or a gene symbol). As an example, we will visualize the regions summarized in the box plot above \(chr18:55862653-55862873) Step #6) The UCSC Genome Browser contains a lot of tracks to compare to your differentially methylated region. You can add tracks \(listed below the genome view) by setting the pull-down value to anything except "hide" and clicking "refresh". For example, the UCSC CpG Islands track is in the "Regulation" section \(and likely not set to be viewed by default) Step #7) If you wish to view your genomic region more clearly, right-click on the image and select "View Image". This will open up a new window with a .png image that can be saved. For example, see a screenshot of the region below \(also see "Figure 8":http://www.nature.com/protocolexchange/system/uploads/2971/original/UCSC_chr18_55862653_55862873.png?1390854865): This region was selected because it was falsely identified by all the programs presented in "Table 2":http://www.nature.com/protocolexchange/system/uploads/2957/original/Table_2.jpg?1390347459 . As you can see from the figures above \(and

"Supplemental Figure S7":http://nar.oxfordjournals.org/content/suppl/2013/03/27/gkt242.DC1/nar-03334-met-n-2012-File006.pdf from the original COHCAP paper), visual inspection would likely confirm this region as a reasonable candidate. Nevertheless, the are some additional criteria that could be applied that would exclude this region: **Possible Criteria #1**: Filter based upon genomic region. This is an intronic CpG island. If visualization of the CpG island shows no overlap with the gene promoter, this candidate may be less interesting \(although, strictly speaking, there isn't a technical reason why this region shouldn't validate successfully) **Possible Criteria #2**: Increase the minimum number of CpG sites per island \(default = 4). A more densely covered region may be less likely to show contradictory results due to lack of consistent methylation changes at all possible CpG sites \(not just the ones included on the 450k array). **In all cases, it is strongly recommended that you visualize your .wig files to identify the optimal start and stop coordinates for your differentially methylated region.** The optimal boundaries often do not exactly correspond to the start and stop coordinates for the official CpG island \(which is how the region is defined in COHCAP, even the CpG shores for a given CpG island).

Timing

1) Process Raw Data: ~2 min **2)** Annotation and Quality Control: ~5 min **3)** CpG Site Analysis: ~9 min **4)** CpG Island Analysis: ~2 min **All statistics are reported as tested with a Windows 7 desktop with 24.0 GB RAM with the data set and parameters specified in the procedure** The following factors influence the run-time for analysis: **A) Data set size**: larger data sets will take longer time to process **B) Parameter stringency**: CpG island analysis will take longer if there are more differentially methylated CpG sites. Therefore, use of more liberal parameters may result in a considerably longer run-time **C) Memory / CPU**: COHCAP may run noticeably quicker on more powerful computers. In fact, sufficiently large cohorts \(for example, patient cohorts with over 100 samples) will likely need to be run remotely on a powerful Linux cluster.

Troubleshooting

1) Process Raw Data *If you see the following error message, you probably need to install the IlluminaHumanMethylation450kmanifest in order to read the raw .idat files. Please click "here":http://www.bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylation450kmanifest.html for detailed installation instrucitons Loading required package: IlluminaHumanMethylation450kmanifest Error in getManifest\(object) : cannot load manifest package IlluminaHumanMethylation450kmanifest *The minfi package recognizes .idat files based upon the extensions _Grn.idat and _Red.idat. However, some publicly available data will use slightly different naming systems. For example, the data files from "E-MTAB-1274":http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1274/E-MTAB-1274.raw.1.zip \(Stricker et al. 2013[10])will require the user to replace _Cy3.idat with _Grn.idat and _Cy5.idat with _Red.idat **2) Annotation and Quality Control** **Please note that COHCAP is currently a 'devel' release** Accordingly, you will need to use R-devel to run COHCAP using the instructions provided on the Bioconductor website. R-devel for Windows: http://cran.r-project.org/bin/windows/base/rdevel.html ****Alternatively, you can download the source files for COHCAP and COHCAPanno and install them using **R CMD INSTALL** \(in the stable version of R). Please see the instructions below on how to use R CMD INSTALL: "http://cran.r-project.org/doc/manuals/R-admin.html#Installing-packages":http://cran.r-project.org/doc/manuals/R-admin.html#Installing-packages Links for source code: COHCAP: "http://bioconductor.org/packages/devel/bioc/html/COHCAP.html":http://bioconductor.org/packages/devel/bioc/html/COHCAP.html COHCAPanno: "http://bioconductor.org/packages/devel/data/experiment/html/COHCAPanno.html":http://bioconductor.org/packages/devel/data/experiment/html/COHCAP ****When running COHCAP in R-devel via biocLite\() installation, you might notice this error when trying to run the COHCAP template using R-devel: Loading required package: bumphunter Error in loadNamespace\(i, c\(lib.loc, .libPaths\()), versionCheck = vI\[\[i]]) : there is no package called ‘registry’ Error: package ‘bumphunter’ could not be loaded You can fix this error by running the following command: install.packages\("registry") **3) CpG Site Analysis** \[no specific suggestions] **4) CpG Island Analysis** When visualizing the .wig files in IGV, you may also want to view the UCSC CpG Island locations. You can download a .bed file for this track here: "http://sourceforge.net/projects/cohcap/files/COHCAP_BSSEQ_anno.zip/download":http://sourceforge.net/projects/cohcap/files/COHCAP_BSSEQ_anno.zip/d This link is also useful for users that wish to analyze targeted BS-Seq data using the COHCAP Bioconductor package. **In general, issues running COHCAP can be addressed on the COHCAP Discussion Group**: "http://sourceforge.net/p/cohcap/discussion/?source=navbar":http://sourceforge.net/p/cohcap/discussion Issues with the Bioconductor package should be listed as "Bioconductor: \[issue description]"

Anticipated Results

COHCAP should produce the following subfolders: **QC**: quality control metrics \(histogram, dendrogram, PCA plot, descriptive statistics) **CpG_Site**: differentially methylated CpG Sites, .wig files for visualization **CpG_Island**: differentially methylated CpG Islands, box-plots to visualize per-sample methylation **Raw_Data**: table of annotated beta values for CpG sites \(all probes), CpG site statistics \(all probes), CpG island beta values \(all islands meeting necessary criteria), CpG island statistics \(all islands meeting necessary criteria)

References

1. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, Yu H, Jove R, Yuan YC. \(2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. 41 \(11): e117 2. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S. IMA: \(2012) An R package for high-throughput analysis of Illumina’s 450K Infinium methylation data. Bioinformatics. 28:729-730. 3. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA \(2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 41\(1):200-9 4. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, and Irizarry RA. \(2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinformatics \[Epub ahead of print] 5. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S. \(2014) ChAMP: 450k Chip Analysis Methylation Pipeline. 30 \(3): 428-430 6. Assenov U, Müller F, Lutsik P, Walter J, Lengauer T. and Christoph Bock D. \(2013) Compehensive Analysis of DNA Methylation Data with RnBeads.

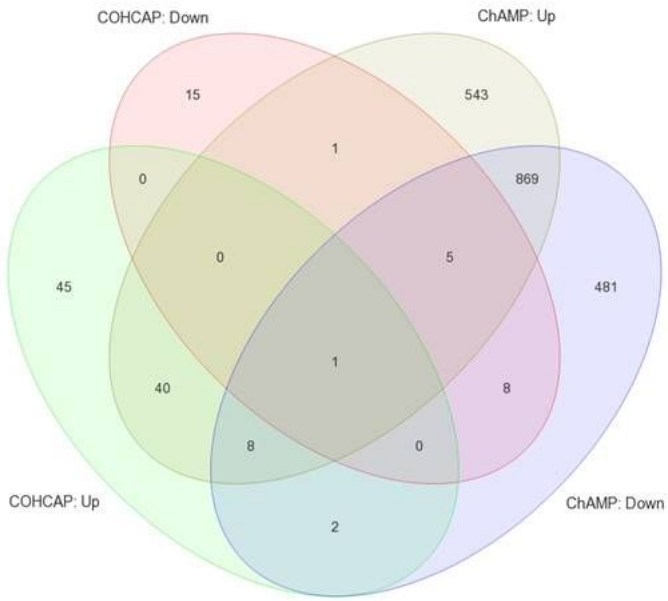
http://rnbeads.mpi-inf.mpg.de 7. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. \(2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5\(10):R80. 8. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. \(2011) Integrative genomics viewer. Nat Biotechnol. 29\(1):24-6 9. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. \(2002) The human genome browser at UCSC. Genome Res. 12\(6):996-1006. 10. Stricker SH, Feber A, Engström PG, Carén H, Kurian KM, Takashima Y, Watts C, Way M, Dirks P, Bertone P, Smith A, Beck S, Pollard SM. \(2013) Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner. Genes Dev. 27\(6):654-69

Acknowledgements

The following Biostar discussion group was helpful in determining updated benchmarks and finding an open-source tool to process raw .idat files: "http://www.biostars.org/p/19220":http://www.biostars.org/p/19220

Figures

Figure 1: Overlapping Differentially Methylated Genes for COHCAP versus ChAMP

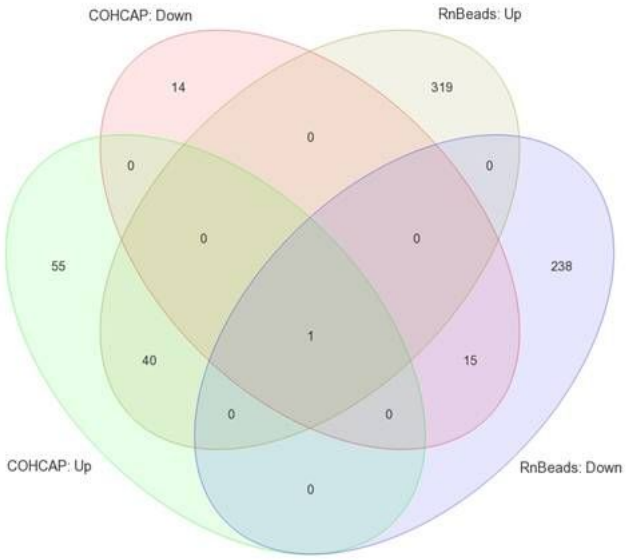


Notice that COHCAP can identify genes not identifed via ChAMP, even with the conservative default settings. Additionally, notice that many genes are listed among both up- down-regulated gene lists for ChAMP but not COHCAP. The gene symbol for all 4 lists is the place holder "NA," which should be ignored. "Up" and "Down" describes the methylation trend in the "mutant" samples compared to the "parental" samples.

Figure 1

Overlapping Differentially Methylated Genes for COHCAP versus ChAMP

Figure 2: Overlapping Differentially Methylated Genes for COHCAP versus RnBeads



Notice that COHCAP can identify genes not identified via ChAMP, even with the conservative default settings. Unlike ChAMP (Figure 1), RnBeads makes fewer predictions (with more similar numbers to COHCAP and IMA) and there are no overlapping genes in the lists showing increased or decreased methylation. This is likely due to the fact that the COHCAP algorithm is more similar to IMA and RnBeads (which all use pre-defined annotations to define boundaries) than ChAMP (which, like Bumphunter, defines region boundaries *ab initio*). The gene symbol for all 4 lists is the place holder "NA," which should be ignored. "Up" and "Down" describes the methylation trend in the "mutant" samples compared to the "parental" samples.

Figure 2

Overlapping Differentially Methylated Genes for COHCAP versus RnBeads

Table 2: Predicted Methylation Status for Verified Regions

	chr10:114712115-114712544	chr18:55862653-55862873	chr2:17721537-17722021	chr17:27044168-27045049
Gene Symbol	TCF7L2	NEDD4L	VSNL1	RAB34
True Status (EpiTect)	Increased Methylation	No Change	Increased Methylation	Decreased Methylation
COHCAP (Average by Island)	Increased Methylation	Increased Methylation	Increased Methylation	Decreased Methylation
IMA (ISLAND)	Increased Methylation	Increased Methylation	Increased Methylation	Decreased Methylation
Bumphunter[1] (via minfi)	Increased Methylation (chr10: 114711288-114713187)	Increased Methylation (chr18: 55862577-55862872)	Increased Methylation (chr2: 17721431-17722068)	Decreased Methylation (chr17: 27044169-27045302)
ChAMP	[no overlapping peak]	Increased Methylation (chr18: 55862456-55862762) feat.rel = 5'UTR_island	[no overlapping peak]	Decreased Methylation (chr17: 27044606- 27045326) feat.rel = Body_island
RnBeads (region: cpgislands)	Increased Methylation	Increased Methylation	Increased Methylation	Decreased Methylation

HCT116 dataset used in original COHCAP[1] publication (which already included the COHCAP and IMA analysis) was used for benchmarks. Validation results have already been presented in that publication. Please note that the IMA results in Table 2 are for CpG islands (no gene mapping is provided) whereas the IMA results are for the TSS1500 regions for each gene. Bumphunter and ChAMP region coordinates are not identical, so overlapping regions are reported in the table above. **True positives and true negatives are shown on bold (with colors corresponding to the expression trend) in each row for the respective algorithms: COHCAP (75% accuracy), IMA (75% accuracy), Bumphunter (75% accuracy), ChAMP (25% accuracy), and RnBreads (75% accuracy).**

[1]Bumphunter provides a very large resulting table for region-based stats. In Table 1, region counts were provied using either fwer < 0.05 or fwerArea < 0.05. Accordingly, TCF7L2 CpG island meets the fwer criteria but not the fwerArea critera. For all other regions, fwer and fwerArea are both less than 0.05.

Figure 3

Table 2 Predicted Methylation Status for Verified Regions

Table 1: Size of Differentially Methylated Gene Lists

	Criteria	Increased Methylation	Decreased Methylation	Run Time[1]
COHCAP[2] (Average by Island)	FDR < 0.05 Methylated Beta > 0.7 Unmethylated Beta < 0.3	96 genes (105 regions)	30 genes (35 regions)	7-18 minutes[2]
IMA (TSS1500)	FDR < 0.05 Delta-beta > 0.3	295 genes	431 genes	4 minutes
Bumphunter (via minfi)	fwerArea < 0.05 **OR** fwer < 0.05	(47 regions, fwerArea) (3113 regions, fwer)	(9 regions, fwerArea) (4412 regions, fwer)	3 hours, 59 min
ChAMP	Dmr.p < 0.05	1467 genes (2398 regions)	1374 genes (2334 regions)	10 minutes
RnBeads (region: promoters)	comb.p.adj.fdr < 0.05 mean.mean.diff > 0.3	532 genes	409 genes	1 hour, 37 min

HCT116 dataset used in original COHCAP publication (which already included the COHCAP and IMA analysis) was used for benchmarks.

[1]Run time was determined with a Windows 7 desktop with 24.0 GB RAM. COHCAP and RnBeads run times are for the entire pipeline. IMA, Bumphunter, and ChAMP run times refer to the minimum number of steps required to call differentially methylated regions. Templates for running benchmarks (with the exact same parameters) are provided in the example dataset download: http://sourceforge.net/projects/cohcap/files/Protocol_Exchange_Example.zip/download

[2]COHCAP and IMA stats come from the original COHCAP publication, which was analyzed using methylation data exported from Genome Studio. The COHCAP benchmark templates use data processed by minfi (in order to match this protocol), so the results (and total run-time) vary. Thus, COHCAP has a range for run times (7 minutes as previously published, 18 minutes for entire protocol described here), whereas all other programs have a single run time.

Figure 4

Table 1 Size of Differentially Methylated Gene Lists

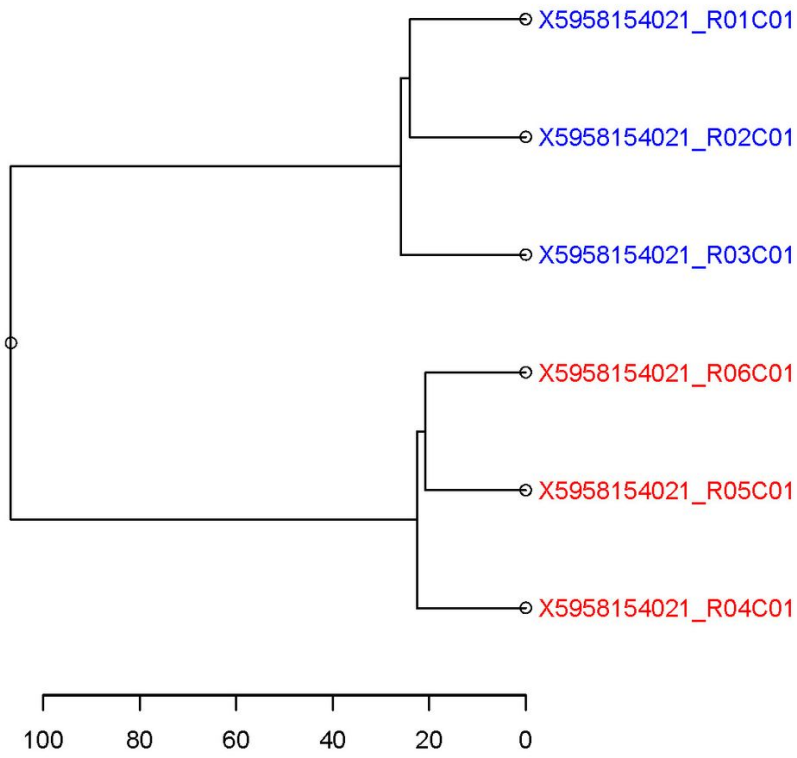


Figure 5

Figure 3 COHCAP_450k_Protocol_Express_cluster.pdf

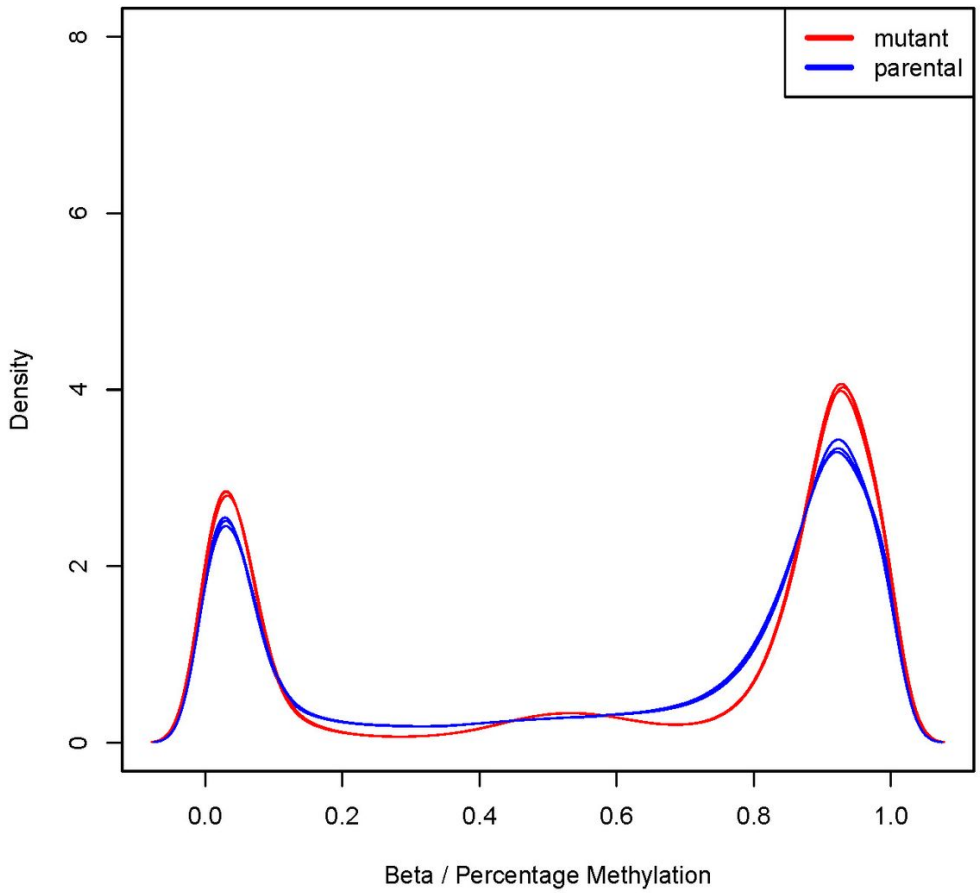


Figure 6

Figure 4 COHCAP_450k_Protocol_Express_hist.pdf

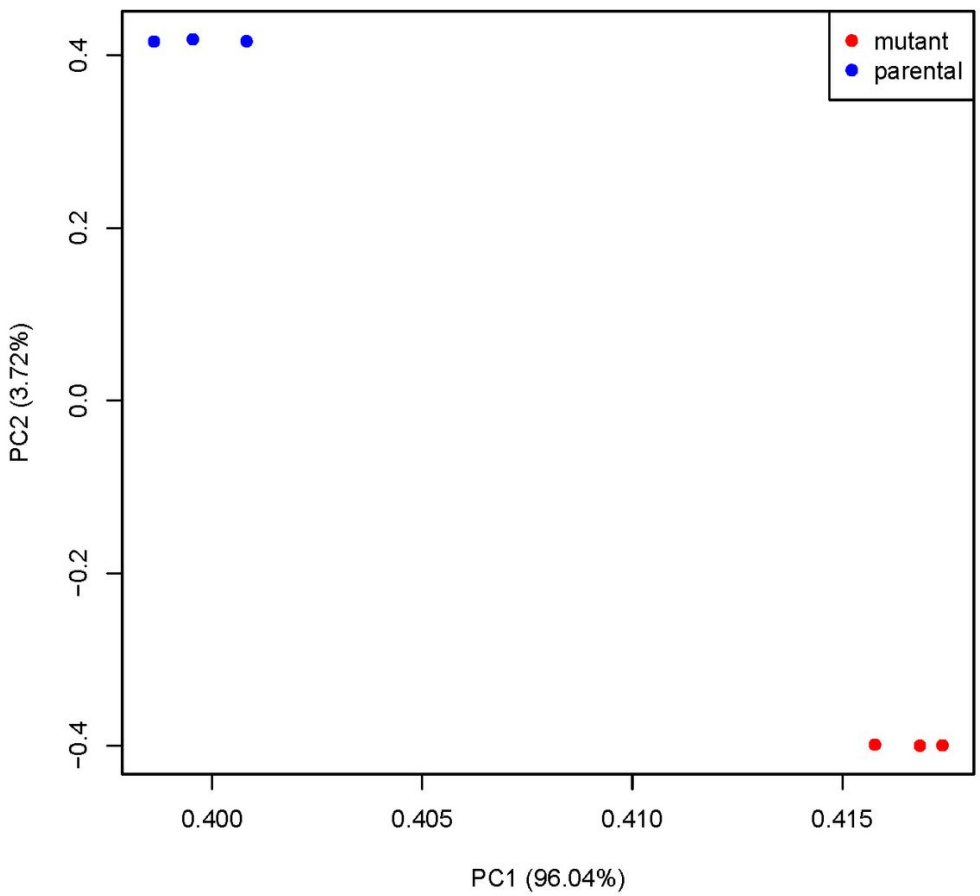


Figure 7

Figure 5 COHCAP_450k_Protocol_Express_pca.pdf

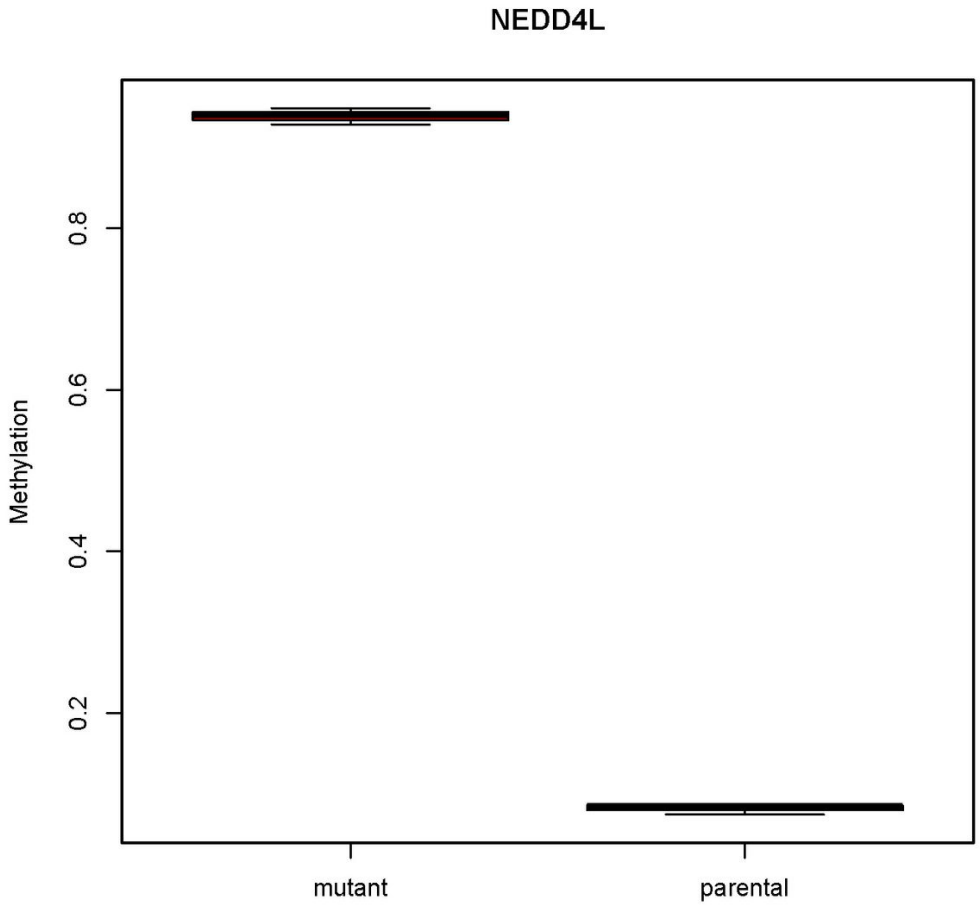


Figure 8

Figure 6 NEDD4L_chr18_55862653_55862873.pdf

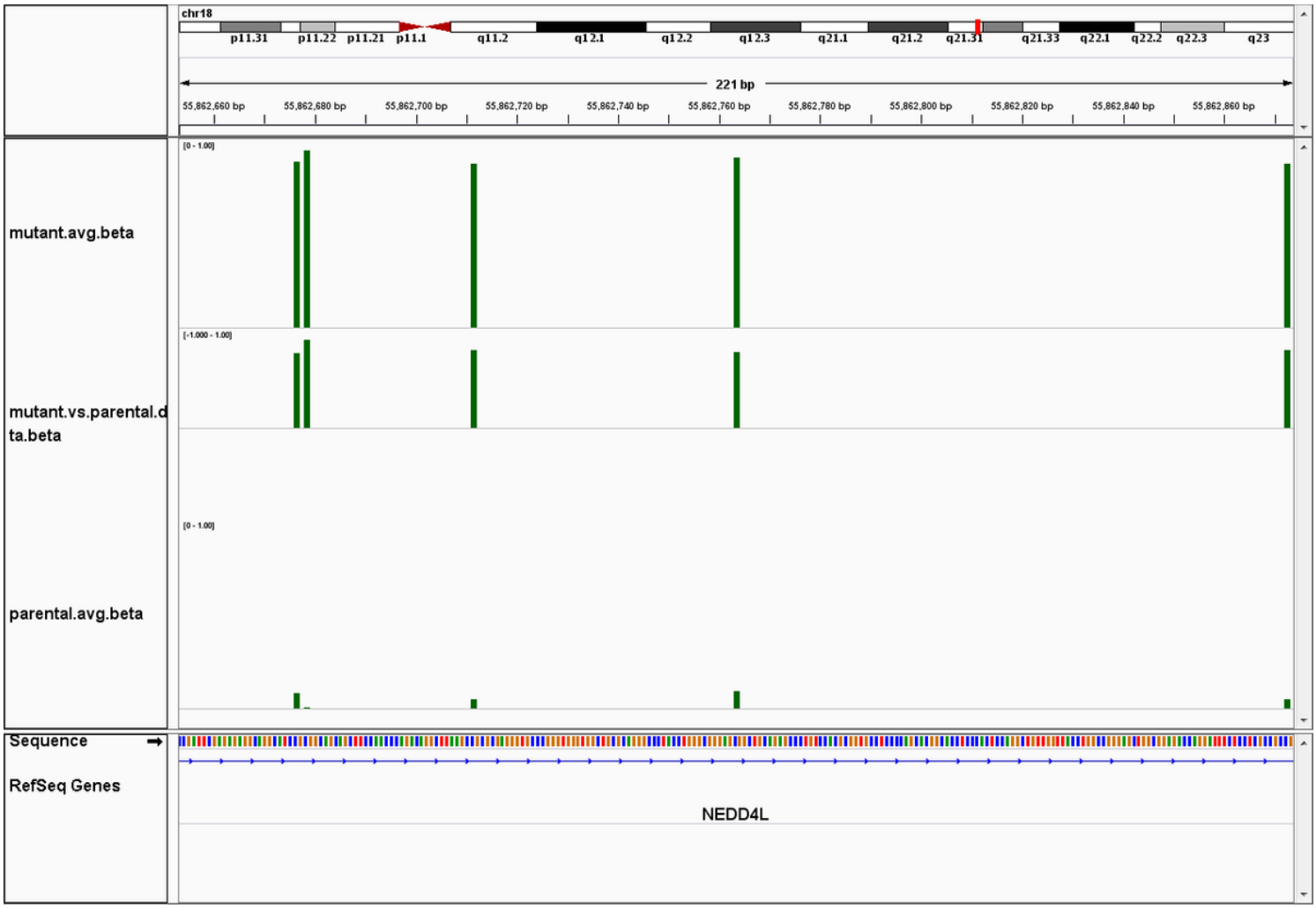


Figure 9

Figure 7 chr18:55862653-55862873 visualized in IGV

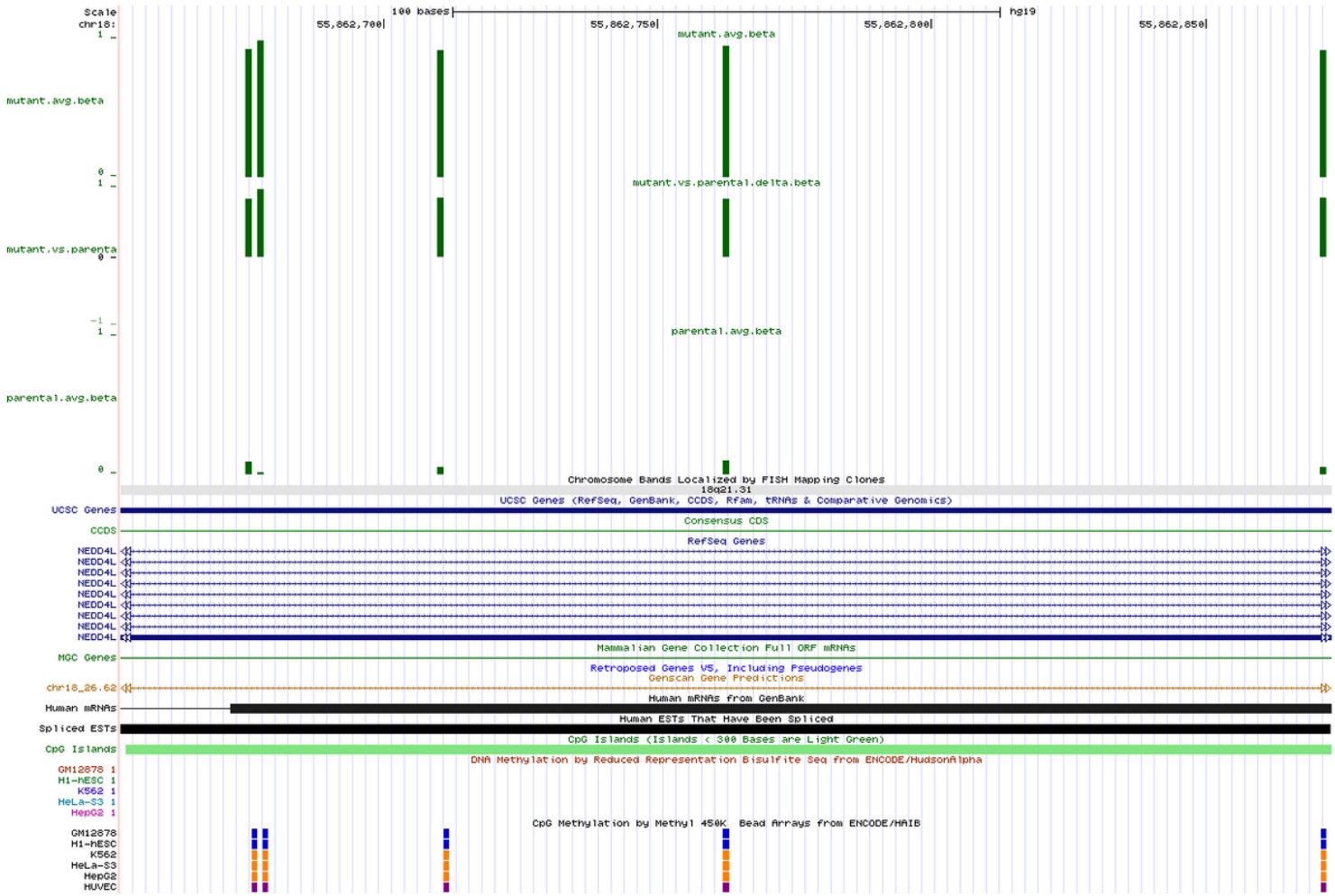


Figure 10

Figure 8 chr18:55862653-55862873 visualized in the UCSC Genome Browser