**This document contains the notes about the different models and decisions taken over the year to perform the analysis and contains potential research papers to use in the discussion**

**Notes over the year:**

CpGs dropped:

* Cg21051031 and cg23513930 dropped with detection p value
* Cg26094004 dropped using difference between means filter

Checked all models

* 4 models already checked
  + Best is DMPFindar as continuous with shrinkvar=TRUE
* DMPFinder as categrorical with shrinkvar=F
  + More inflated than other models
* Linear mixed effects regression
  + More inflated than other models too

DMRcate:

* Set fdr to 1 to check for total number of regions
  + Therefore all probes are significant
  + Using no threshold or betacutoff there are 106722 DMRs from all probes
* Using threshold of min 10 cpgs and betacutoff
  + Ranges increase from 2256 to 4259
  + 4259 is the max number of ranges determined from significant CpGs
  + The size of the ranges become bigger as more CpG sites are significant
* Testing fdr 0.01 and different min number of cpgs
  + 5: 6307
  + 10: 2082
  + 15: 711
  + 20: 238
  + 25: 117
* Testing fdr significance levels:
  + Fdr 0.01: 122569 probes
  + Fdr 0.001: 49621 probes
    - Returned 632 ranges with same filters
  + Fdr 0.0001: 22102 probes
    - Returned 324 ranges with same filters
  + Fdr 0.00001: 9714 probes
    - Returned 163 ranges
    - Similar number of significant probes compared to methylation probe analysis
* Testing using p < 1e-7 like in DMP analysis
  + Returns 9110 significant probes and 75 regions with min cpgs is 15 and betacutoff of 0.05
    - 260 DMRs if min cpgs is 10
  + Total number of regions possible is 106722 with size as small as 1 cpg

Next Step for MEME Suite:

* Need primary sequences in FASTA format file
* Which files are from knockout vs control?

Notes Feb 10th meeting:

* Lead with num of cpgs diff methylated and tolerating some inflation due to cell culture with annotation
* Look up size of dmr
* How many regions meet same parameters significant or not significant?
* Test fdr =1 , apply same filters, find num regions and apply bonferroni
  + Number of regions is 4259 which meet filters
  + Bonferroni correction is 0.05/4259 = 1.17e-5
  + 228 differentially methylated regions remain
  + Check the plots to make sure they look good

Lab meeting next week:

* Concise, but detailed presentation; troubleshooting can leave some parts of this out
* Facilitate a discussion at the end
* What are the next steps for this project?
* Other ideas and next steps forward?

Checked mitochondria related genes overrepresented

Feb 27th notes:

STREME single file from RNA-Seq is too big to run program

ELMER:

* Output with empirical p-value gives 19132 probe gene pairs as having an FDR < 0.05
* Total probe gene-pairs tested in 81466
* 8895 total probe-gene pairs FDR < 0.01
  + Check number of genes in 8895
    - 1252 unique genes
  + Check for total num of input genes
  + How many unique genes in 81466 this is denominator
    - 13111 unique genes tested
  + Divide 1st by third
    - About 9.5% of the genes tested are being shown as significant
* Archive a couple pairs of significant genes from sleuth
* Rerun elmer with only significant genes from slueth (164 genes)
  + The top 30 gene-probe pairs are the same as from all results
  + Check number of pairs that meet significance

Mitocarta:

* Find genes that overlap with cpgs
* This will reduce the number of genes in the database
* Need to define how genes overlap

MEME:

* Potentially collapse controls and knockouts to consensus
  + Look into literature for RNA-seq and MEME
* <https://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pmc/articles/PMC8329593/>
  + Extract DNA sequences representing 500 bp upstream of TSS (transcriptional start site) of up and down-regulated genes
  + Use with streme
  + Find relatively enriched ungapped motifs in the sequences of interest compared to null distribution--+
  + Differentially expressed find TSS in reference pull the bps upstream of start

Check sleuth to make it TMM instead of TPM

* Determine way to convert TPM to TMM
* If not need to look into Kallisto/sleuth

Sleuth uses the counts or TPM. Can use norm\_factors in Sleuth to normalize similar to DESeq size factors

EdgeR package implements TMM as the standard normalization procedure

* Try this out and compare the summary stats
  + Aggregated transcript expression for different isoforms to same gene instead of taking highest
* Summary stats show 367 significant genes using bonferonni correction (0.05/14527)
  + Uses a likelihood ratio test
  + Should there be a logFC filter, I was thinking it could be logFC > abs(2)
* Summary stats using LRT show 310 significant genes using bonferonni correction and logFC > 1.5
  + With logFC > 2 summary stats show 185 significant genes
  + The model used is a Negative Binomial Generalized log-Linear Model
* Of top 15 significant genes from sleuth and edger 7 genes overlap
* Of all significant genes (185 from edger, 164 from sleuth) 78 gene overlap

EdgeR using max counts:

* Summary stats using LRT show 179 significant genes using bonferonni correction and logFC > 2
* Of all significant genes (179 from edger, 164 from sleuth) 75 genes overlap
* Of top 15 significant genes from sleuth and edger 8 genes overlap
* Check qq-plot between sleuth and edger, if they are similar proceed
  + Plot top 2/3
* Checked lambda value between EdgeR and Sleuth results
  + Sleuth: 6.9
  + EdgeR: 5.6

**ELMER with Edger max counts:**

* Total probe gene-pairs tested in 81466
  + Each probe tests twenty genes
  + Bonferronni: 6.13753e-07 (0.05/81466)
* 9505 total probe-gene pairs FDR < 0.01
  + Check number of genes in 6709
    - 1344 unique genes
  + How many unique genes in 81466 this is denominator
    - 12860 unique genes tested
  + Filter for significant expression differences
    - Students t test; p < 0.001
    - Reduces to 2732 pairs and 339 unique genes

**ELMER with Edger max counts only significant genes:**

* Total probe-gene pairs tested is 35347
  + Bonferonni: 1.414547e-06 (0.05/35347)
* 26575 probe-gene pairs have FDR < 0.01
  + Students t test; p < 0.001
  + Reduces to 10652 pairs and 56 unique genes

**ELMER with EdgeR max counts adjusted filters:**

* Checks for 81469 pairs using Mann-Whitney U test
* 7000 pairs have FDR < 0.01
  + 1166 unique genes in 7000 pairs
* Filter based on gene expression using students t-test between groups
  + Significance threshold p < 0.001
  + 2314 gene probe pairs
    - 336 unique genes
  + 61 of 225 significant genes from EdgeR
* Run elmer with all to get num pairs
* Run elmer only with significant genes

Redo the fisher combined p using nominal significance in both p <0.05

* completed

Keep on updating slides with new stuff

Package dmricher

* Start take the position
* End add or substract one
* To create images
* Look at this <https://github.com/ben-laufer/DMRichR>

Next Steps: March 7th

* DmRichr to create images for DMRs
  + Installation issues on compute canada
* Mitocarta
* REVIGO
* MEME
  + Perform GO on significant groups
  + Fisher combined with Meth maybe
* Check edgeR and sleuth to see what the differences are
  + Read both documentation for different options
  + Check difference in LRT ratio
  + Use logFC > 2 if not check > 1.5

Search for depletion in methods of paper on slack

* <https://pubmed.ncbi.nlm.nih.gov/34859289/>
* Check for how the fishers exact test was implemented

Next steps:

* Focus dmrichr; goal is to find functional elements between KO and NC
* Check other paper in proposal folder for eforge
* Getting something dmrichr or eforge
* Do feature overrepresentation analysis for following functional elements
* Kegg pathviewer up and down regulated with different colours

Notes:

DMRichR:

* Cannot install on compute canada; different error from local installation
  + I am using the default bioconductor 3.14 installation loaded on compute canada
  + The installation does not allow for changes so DMRichR cannot be installed
* DMRichR requires CpG count matrices from Bismark genome-wide cytosine reports
  + This requires raw reads of WGBS
  + Not sure if it is possible to convert beta values to CpG count matrix
  + Skip to the DMR part use Julia's code

Next Steps:

* Give Dr. Castellani question for Ben
* Refseq or getMapped EntrezIds in miss methyl

List of Items to complete:

* DMRichR: chromHMM
  + Need to figure out how to set up database to run
* Code to create Manhattan plots with axis labels

List of remaining tasks for me to complete:

* Debug code to create Manhattan plots with all x-axis labels
* DMRichR
  + Have the same as the pic julia sent
  + chromHMM
* **Maybe**: MEME suite tools use CpGs mapped to Genes 1000 bp upstream
  + If there is a motif that is common across that controls expression

List of items I won't be able to complete:

* WGCNA

Lab hypothesis:

* Metabolites (Acetyl-CoA, etc.) decreasing facilitates this function

**Genes to investigate further:**

Biological Mechanism: Metabolites (Acetyl-CoA, etc.) decreasing facilitates this function

SFRP2: This gene encodes a member of the SFRP family that contains a cysteine-rich domain homologous to the putative Wnt-binding site of Frizzled proteins. SFRPs act as soluble modulators of Wnt signaling. Methylation of this gene is a potential marker for the presence of colorectal cancer.

“SFRP2 modulates non‑small cell lung cancer A549 cell apoptosis and metastasis by regulating mitochondrial fission via Wnt pathways”

**GABRB1** is a subunit of GABAA Receptor

CXCL12 acts as ligand for chemokine receptor 4; important in cellular functions (e.g. embryogenesis, inflammation, metastasis)

**EIF4EBP3** is an initiation factor prevents assembly of EIF4E which is the rate limiter of protein synthesis

**AC005280.1**: is actually RIOX1 Gene: protein coding gene, Predicted to enable histone H3-methyl-lysine-36 demethylase activity; histone H3-methyl-lysine-4 demethylase activity; and iron ion binding activity. Predicted to be involved in histone lysine demethylation

* In heatmap from roadmap epigenomics, H3K36me and H3K4me3 is underrepresented; RIOX1 gene specifically demethylates H3K4me and H3K36me

LINGO2: Predicted to act upstream of or within positive regulation of synapse assembly. Predicted to be integral component of membrane. Predicted to be active in extracellular matrix and extracellular space

**DOK5:** The protein encoded by this gene is a member of the DOK family of membrane proteins, which are adapter proteins involved in signal transduction. The encoded protein interacts with phosphorylated receptor tyrosine kinases to mediate neurite outgrowth and activation of the MAP kinase pathway

* <https://www.sciencedirect.com/science/article/pii/S0898656806000702?via%3Dihub>

CSAG4: pseudogene

CFAP47: Cilia and Flagella Associated Protein 47

FLRT2: Functions in cell-cell adhesion, cell migration and axon guidance. Mediates cell-cell adhesion via its interactions with ADGRL3 and probably also other latrophilins that are expressed at the surface of adjacent cells.

KCNK15: This gene encodes one of the members of the superfamily of potassium channel proteins containing two pore-forming P domains

**KCTD8:** Auxiliary subunit of GABA-B receptors that determine the pharmacology and kinetics of the receptor response. Increases agonist potency and markedly alter the G-protein signaling of the receptors by accelerating onset and promoting desensitization (By similarity).

FBXL7: Substrate recognition component of a SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex (PubMed:25778398). During mitosis, it mediates the ubiquitination and subsequent proteasomal degradation of AURKA, causing mitotic arrest (By similarity). It also regulates mitochondrial function by mediating the ubiquitination and proteasomal degradation of the apoptosis inhibitor BIRC5

PDGFD: Growth factor that plays an essential role in the regulation of embryonic development, cell proliferation, cell migration, survival and chemotaxis. Potent mitogen for cells of mesenchymal origin.

MAGEH1: This gene belongs to the non-CT (non cancer/testis) subgroup of the melanoma-associated antigen (MAGE) superfamily. The encoded protein is likely associated with apoptosis, cell cycle arrest, growth inhibition or cell differentiation.

GLIS1: Acts as both a repressor and activator of transcription (PubMed:21654807). Binds to the consensus sequence 5'-GACCACCCAC-3' (By similarity). By controlling the expression of genes involved in cell differentiation inhibits the lineage commitment of multipotent cells (PubMed:21654807). Prevents, for instance, the differentiation of multipotent mesenchymal cells into adipocyte and osteoblast (By similarity).

**ABCD1:** ATP-dependent transporter of the ATP-binding cassette (ABC) family involved in the transport of very long chain fatty acid (VLCFA)-CoA from the cytosol to the peroxisome lumen (PubMed:11248239, PubMed:15682271, PubMed:16946495, PubMed:18757502, PubMed:21145416, PubMed:23671276, PubMed:29397936, PubMed:33500543). Coupled to the ATP-dependent transporter activity has also a fatty acyl-CoA thioesterase activity (ACOT) and hydrolyzes VLCFA-CoA into VLCFA prior their ATP-dependent transport into peroxisomes, the ACOT activity is essential during this transport process (PubMed:33500543, PubMed:29397936). Thus, plays a role in regulation of VLCFAs and energy metabolism namely, in the degradation and biosynthesis of fatty acids by beta-oxidation, mitochondrial function and microsomal fatty acid elongation (PubMed:23671276, PubMed:21145416). Involved in several processes; namely, controls the active myelination phase by negatively regulating the microsomal fatty acid elongation activity and may also play a role in axon and myelin maintenance. Controls also the cellular response to oxidative stress by regulating mitochondrial functions such as mitochondrial oxidative phosphorylation and depolarization. And finally controls the inflammatory response by positively regulating peroxisomal beta-oxidation of VLCFAs (By similarity).

**POU6F2 (maybe look into this):** Probable transcription factor likely to be involved in early steps in the differentiation of amacrine and ganglion cells. Recognizes and binds to the DNA sequence 5'-ATGCAAAT-3'.

* Amacrine cells changes synaptic release patterns of 4-aminobutyrate (GABA)

MID1: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=MID1>

TMEM35: Enables acetylcholine receptor regulator activity. Involved in chaperone-mediated protein complex assembly and positive regulation of protein localization to cell surface. Located in endoplasmic reticulum.

**GRIA4:** Glutamate receptors are the predominant excitatory neurotransmitter receptors in the mammalian brain and are activated in a variety of normal neurophysiologic processes

DPYD: The protein encoded by this gene is a pyrimidine catabolic enzyme and the initial and rate-limiting factor in the pathway of uracil and thymidine catabolism.

NCAM2: May play important roles in selective fasciculation and zone-to-zone projection of the primary olfactory axons.

TMBIM1: Enables death receptor binding activity. Involved in negative regulation of Fas signaling pathway; negative regulation of extrinsic apoptotic signaling pathway via death domain receptors; and negative regulation of protein localization to plasma membrane. Located in Golgi apparatus; endosome membrane; and lysosomal membrane.

SLC16A2: This gene encodes an integral membrane protein that functions as a transporter of thyroid hormone. The encoded protein facilitates the cellular importation of thyroxine (T4), triiodothyronine (T3), reverse triiodothyronine (rT3) and diidothyronine (T2).

IGFBPL1: Predicted to enable insulin-like growth factor binding activity. Involved in cellular response to tumor cell. Located in extracellular space

GALNT14: This gene encodes a Golgi protein which is a member of the polypeptide N-acetylgalactosaminyltransferase (ppGalNAc-Ts) protein family. These enzymes catalyze the transfer of N-acetyl-D-galactosamine (GalNAc) to the hydroxyl groups on serines and threonines in target peptides.

CYP7B1: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=CYP7B1>

ARSJ: Sulfatases (EC 3.1.5.6), such as ARSJ, hydrolyze sulfate esters from sulfated steroids, carbohydrates, proteoglycans, and glycolipids. They are involved in hormone biosynthesis, modulation of cell signaling, and degradation of macromolecules

BMP2: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=BMP2>

CLMP: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=CLMP>

SCG5: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=SCG5>

PTGER3: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTGER3>

RBM24: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=RBM24>

C4orf19: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=C4orf19>

COL21A1: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=COL21A1>

RBP7: intracellular transport of retinol:

* <https://www.genecards.org/cgi-bin/carddisp.pl?gene=RBP7>

AFF2: <https://www.genecards.org/cgi-bin/carddisp.pl?gene=AFF2>

**Final Thesis Outline:**

Background:

* Mitochondria DNA/Function
* Nuclear Epigenome
  + Methylation
  + Histone markers?
* Mitochondria on nuclear DNA methylation
* Hypothesis
* Objectives

Methods:

* Cell Line Model
* EPIC Array
* RNA-Seq
* Differential methylation/expression
* ELMER: Integration
* Functional Enrichments

Results:

* Same points as above
* Top RNA genes: in order
  + SFRP2
  + KCTD8
  + FLRT2
  + FBXL7
  + ABCD1
  + RIOX1
  + PGR
  + EIF4EBP3
  + IGFBPL1
  + LINGO2
* Previously identified CpGs
  + Castellani:
    - Cg03964851 is surrogate for cg21051031
    - Removed by detection p-value: poor performing probe
      * Cg23513930 no longer on EPIC array
    - Removed by filter for difference between NC means
      * Cg26094004
    - Significant CpGs:
      * No matching significant CpGs
  + Wang:
    - Removed by detection p-value
      * Cg21848084
    - Removed by filter for difference between NC means
      * cg26094004
* CpG Sites:
  + Hyper: 1113 (t > 0)
  + Hypo: 3129 (t < 0)
* RNA:
  + Over: 41 (logFC > 2)
  + Under: 138 (logFC < -2)

Discussion:

* GABAergic Synapse
  + <https://pubmed.ncbi.nlm.nih.gov/17132215/>
  + <https://pubmed.ncbi.nlm.nih.gov/35335130/>
  + <https://www.pnas.org/doi/10.1073/pnas.1906251116>
* GIRK- inwardly rectifying potassium channels (top in reactome methylation)
* **GRIA4:** Glutamate receptors are the predominant excitatory neurotransmitter receptors in the mammalian brain and are activated in a variety of normal neurophysiologic processes
* **ABCD1:** lok at discussion notes; gene in mitoCarta3.0 definitely talk about this
  + <https://pubmed.ncbi.nlm.nih.gov/24076127/>
  + <https://pubmed.ncbi.nlm.nih.gov/22366764/>
  + <https://pubmed.ncbi.nlm.nih.gov/12509471/>
* **SFRP2**: top gene probe pair and top RNA gene
  + <https://pubmed-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/32071544/>
* **KCTD8:**
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6405316/>
* **AC005280.1**: is actually RIOX1 Gene:
  + Good candidate for discussion because can relate to histone marks
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6525012/>
* **EIF4EBP3** is an initiation factor prevents assembly of EIF4E which is the rate limiter of protein synthesis
  + <https://pubmed.ncbi.nlm.nih.gov/31074051/>
  + <https://pubmed.ncbi.nlm.nih.gov/31853750/>
  + <https://pubmed.ncbi.nlm.nih.gov/34541613/>
* Zinc fingers:
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6222495/>
  + <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5683310/>