EPIGENETIC MECHANISMS OF FOLIC ACID ACTION

IN MICE MODELS OF

NEURAL TUBE DEFECTS

A Dissertation

Presented to the Faculty of the Weill Cornel Graduate School

of Medical Sciences

in Partial Fulfillment of Requirements for the Degree of

Doctor of Philosophy

by

Dhruva Chandramohan

August 2017

EPIGENETIC MECHANISMS OF FOLIC ACID ACTION IN MICE MODELS OF NEURAL TUBE DEFECTS

Dhruva Chandramohan, PhD.

Weill-Cornell Medicine 2017

Neural tube defects (NTDs) are among the most common serious birth defects, occurring in 0.5-10 per 1000 live births globally. Clinical studies have shown that periconceptional supplementation with folic acid (FA) can reduce NTD's by up to 70%. The mechanism(s) by which FA prevents NTDs is an area of active research.

Lrp6 provides two mice models of NTD that exhibit altered prevalence under FA supplementation. Crooked tail mice (*Lrp6*$^{Cd/+}$) are rescued by FA supplementation, while *Lrp6* deficient mice (*Lrp6*$^{-/-}$) show increased rates of absorption and embryonic lethality under supplementation. We assayed the methylomes of heterozygous Cd and KO E9.5 mice. To analyze this data, we created the multiDiff package. It implements the novel maximum difference estimate for assigning biologically meaningful effects in complex designs. multiDiff displays consistently superior performance while varying simulation parameters as measured by AUC compared to DSS-general, a competing method ($p<2.2^{-16}$, paired t-test).

Integrative analyses found that Rn45s showed FA-associated differential expression and differential promoter methylation on the KO background. The analysis suggests that FA primarily acts in an independent rather than additive or combinatorial manner on methylation and expression. Genes associated with an independent mechanism of action were enriched for transcriptional regulation. On the Cd background, we noted genes affected by FA that had known links to *Lrp6* biology, with the greatest number being associated with RhoA, suggesting involvement in the planar cell polarity Wnt signaling pathway.

Methylation was also assayed in whole blood from P2 animals. We found 43 persistently differentially methylated sites associated with the Cd mutation, and 25 with the KO mutation. Persistently differentially methylation loci associated with FA was identified at 86 sites in the Cd background, and 208 in the KO background.

Before concluding, we discuss preliminary analyses of epigenetic and genetic data in human NTD patients, with a focus on ongoing challenges in the field. The continuing growth of public datasets, especially in other neurological disorders such as autism, combined with advances in sequencing technology, and improvements in analytical methods are giving modern NTD researchers the tools to overcome these challenges.

BIOGRAPHICAL SKETCH

Dhruva Chandramohan was born in Redbank, New Jersey, and graduated from Bridgewater-Raritan High School in 2005. He graduated in 2009 from New York University with a Bachelor of Arts in Mathematics and Philosophy, and a Minor in Psychology, with an Honors thesis on the analysis of mathematical models of binocular rivalry. He then joined the lab of Dr. Kevin Hall at the National Institute for Diabetes, Digestive, and Kidney Disease as a Post-Baccalaureate Fellow, where from 2009 to 2011 he performed obesity research, creating the first version of the NIDDK's Body Weight Simulator.

Dedication:

To my amazing and patient wife, Anita, and my parents and sister.

For putting up with my foibles

Acknowledgements

I would like to thank all the people who have helped me through graduate school. My advisor, Dr. Chris Mason, has given me the opportunity to grow and learn in an ever-widening area of inquiry, encompassing a wide swath of computational genomics and epigenetics, with boundless enthusiasm.

Thanks goes to my special thesis committee members: Dr. Margaret Ross, Dr. Jason Mezey, Dr. Christina Leslie, for their direction, suggestions, and encouragement during our meetings. Professor Ross has consistently pushed me to towards a deeper biological understanding, and to make sure complex analyses are clearly communicated. I really appreciate Prof. Mezey and Prof. Leslie for taking the time to walk through various technical discussions on modeling and presentation.

A huge thanks to Bozena Castaldo of the Ross Lab and Jorge Gandara, formerly of the Mason lab, for raising the animals and generating the data upon which this thesis is built. Without their time, training, and expertise, there would be little say.

I would also like to thank my colleagues from the Mason lab. I would like to thanks Dr. Cem Mayden, Dr. Altuna Akalin, Paul Zumbo, Dr. Yogesh Saletore, Ms. Priyanka Vijay, Mr. Marjan Bozinoski, Ms. Lenore Pipes. I would like to thank PBTECH for their support on the high performance computing systems, including Ms. Vanessa Borcherding. I would also like to Dr. Vanessa Aguiar-Pulido from the Ross lab for her helpful suggestions on the analysis and methods comparison.

I would like to thank my fellow graduate students from the CBM program. In particular, I would like to thank Dr. Kelson Zawack, and Hilary Monoco for their support and ever fun conversations. From the rest of the Tri-Is, I would also like to thank Dr. Michael Levine for the many fun games of Magic: the Gathering, and having the most rigorous mind around. I would like to thank German Sabio and his wife Anu for times well spent.

I would like to thank the various guides I encountered on the road to grad school. Dr. Ramendra Pandey, who was one of my first mentors in science, and patiently guided me at a young age. Professor Ricky Pollack at NYU was a wonderful teacher and friend. I also thank Dr. John Rinzel for his guidance on my undergraduate thesis, Dr. Kevin Hall for mentoring me through my first taste of computational biology, and Dr. Artie Sherman for many fun and interesting lunches about science, politics, and philosophy.

I would like to thank my PhD program, the Tri-Institutional Training Program in Computational Biomedicine at Weill Cornell Medical College. Additionally, I would like to thank Prof. David Christini, Margie Mendoza, and Kathleen Pickering for being wonderfully helpful and caring administrators.

I would like to thank my parents, who have been waiting patiently for this day, and my sister Dhanya.

Finally, thank you to my amazing wife Anita, for bringing so many moments of big and small joy to my life, and for helping push me to the finish line.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

AUC=Area Under Curve
Cd=Crooked Tail
DEG=Differentially Expressed Gene
DMC=Differentially Methylated Cytosine
DMG=Differentially Methylated Gene
E9.5=Embryonic Day 9.5
RRBS=enhanced reduced representation bisulfite sequencing
FA=Folic Acid
FPR=False Positive Rate
KO=Knockout
MMD=Maximum Methylation Difference
NTD=Neural Tube Defect
P2=Post-natal Day 2
ROC=Receiver-Operating Characteristic
RRBS=reduced representation bisulfite sequencing sequencing
SB=Spina Bifida
TPR=True Positive Rate
WGS=Whole genome sequencin

CHAPTER 1

Introduction:
Neural Tube Defects and Folic Acid

1.1 Background on Neural Tube Defects

Neural tube defects (NTDs) affect an estimated 300,000 infants annually worldwide, making them second only to second only to congenital heart anomalies in the U.S. (World Health Organization., 2015; World Health Organization, 2015). This results in approximately 88,000 deaths, as well as 8.6 million disability-adjusted life years, every year. A systematic review of international NTD occurrence reported national levels ranging from 0.3 -199.4 per 10,000 live births from 1990-2014 (Zaganjor et al., 2016). Due to the role played by nutrition in their etiology (see below), increased rates of famine driven by conflict, population growth, and climate change will likely result in increased NTD incidence, particularly in low income countries where NTDs may account for 29% of neonatal deaths with observable causes (Blencowe et al., 2010).

NTDs are the result of failures during embryonic development (Detrait et al., 2005; Jiang et al., 2011; Wallingford et al., 2013). Under normal conditions, the central nervous system starts as a flat sheet of cells, known as the neural plate, with paired neural folds develop along the rostrocaudal axis (Figure 1.1) During development, these folds roll and fuse to create a hollow tube containing the brain and spinal column (by day 28 in humans (O’Rahilly and Müller, 1994), E9.5 in mice). When this fusion fails to occur, parts of the CNS can become exposed. The resulting defect can either be open to the environment, or, in rarer instances, enclosed by skin. Broad classification is primarily based on position: those defects occurring in the cranial region involving absence of the cranial vault and severe defects in the cerebral hemispheres are referred to as anencephaly, which is invariably fatal. If part of the brain or surrounding tissue is pushed through an opening in the skull, the condition is referred to as encephalocele. Defects that occur in the caudal portion of the neural tube are referred to as spina bifida, or meningomyelocele, and are the most frequently observed category (Zaganjor et al., 2016). Finally, defects over the entire body axis are referred to as craniorachischisis, and are also fatal.

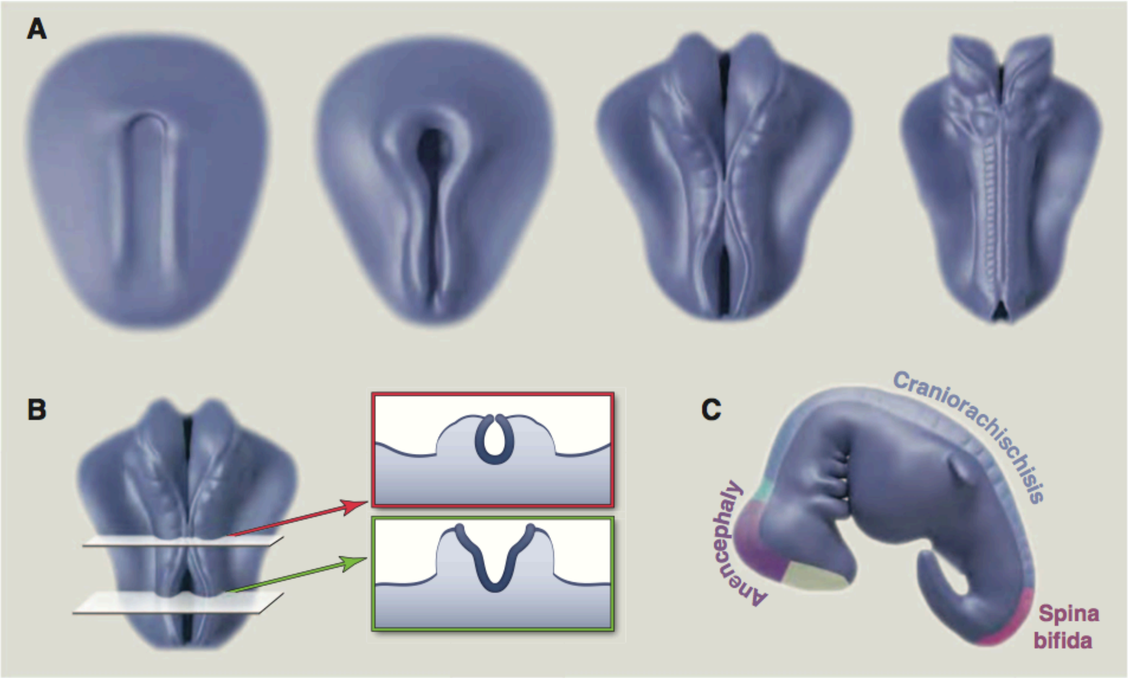


Figure 1.1. Neural Tube Closure. A. Successive images showing the progression of neural tube closure in a a stylize vertebrate embryo (rostral=up). B. Cross section illustrate closed (red) and open (regions) of the neural tube. C. Region-specific NTDs. Reprinted From Wallingford, 2013.

It is important to note that each of the above categories groups together a large number of specific phenotypic traits, and are not mutually exclusive. This is similar to the class of NTDs taken together- the name collects under one heading many different conditions. Partially for this reason, the etiology of NTDs is complex and multifactorial. Twin studies in mice have shown increased concordance in monozygotic twins compared to dizygotic twins (7.7% vs. 4.0%, (Deak et al., 2008)), indicating a heritable genetic component. However recurrence is only 2-5% in human families with two affected siblings (Sebold et al., 2005), demonstrating incomplete penetrance. Numerous studies of the genetics of NTDs have been published focusing on mutations in genes associated with folic acid metabolism, especially the rate limiting enzyme in the methyl cycle, MTHFR (Kirke et al., 2004; van der Put et al., 1995; Wang et al., 2015).

To date, no large-scale study whole genome study of NTD has been published, however sample collection and sequencing is ongoing among various national monitoring centers, such as the California Birth Defects Monitoring Program (CBDMP) (Croen et al., 1991).

1.2 NTDs and Folic Acid Supplementation

Maternal diet has also been shown to have a strong effect on NTD risk, in particular consumption of vitamin B9, or folic acid (FA). The Leeds observational studies linked circulating levels of FA and other B vitamins in maternal blood to NTD risk, which were followed by successful interventional studies in preventing recurrence in mothers of NTD-bearing infants (Smithells, 1984; Smithells et al., 1976, 1980, 1981). Further clinical studies found that maternal intake of FA reduced the incidence of NTDs by between 30-70% (Bower and Stanley, 1989; Mills et al., 1989; Mulinare J et al., 1988; Shaw et al., 1995; Werler MM et al., 1993). In 1996, the US FDA introduced a FA fortifications program for staple cereal grains and flour (1996). Incidence of two forms of NTD, spina bifida and anencephaly, were reduced by 20% and 34% respectively, with greater declines being reported as a result of similar programs in Chile and Canada (Wats et al. 2007, Lopez-Camelo et al). Currently 80 countries around the world have implemented FA fortification programs, with a notable absence of such programs in European countries, excepting the United Kingdom. Governments may prefer education and voluntary dietary supplementation, as dietary folate is a naturally occurring nutrient found in foods such as leafy green vegetables, legumes, egg yolk, liver, and citrus fruit. They may also oppose universal fortification, for fear of potential side effects (Mills and Dimopoulos, 2015).

Evidence for potential negative side affects of FA have come from several fronts. Among mice models of NTD, some show shift towards early embryonic lethality under FA supplementation (Gray et al., 2010). In vitro experiments have found that FA inhibits neurite extension synaptogenesis, and growth cone motility in chick embryos (Wiens, 2016). Finally, a Norwegian study of 6837 ischemic heart patients (Ebbing et al., 2009), linked FA supplementation to increased risk of cancer (Hazard Ratio 1.21) and all-cause mortality (Hazard Ratio 1.18).

1.3 FA and One-Carbon Metabolism

The mechanism by which FA prevents NTDs has been an area of active and ongoing research (Blencowe et al., 2010; Ernest et al., 2002; Mills et al., 1989; Molloy et al., 2017; Werler MM et al., 1993). Identifying said mechanism would both allay concerns about side effects of supplementation, while also potentially allowing for the design of alternative regimens. FA metabolism plays a central role in numerous cellular reactions through the one-carbon cycle (Bodnar et al., 2010; Greenberg et al., 2011; Suh et al., 2001). After being enzymatically reduced to tetrahyrafolate, it can be converted to L-methylfolate. L-methyfolate is biologically active, unlike FA or natural dietary folate, and it utilizes the supplied methyl (one-carbon) group in the synthesis of purines and pyramidines.

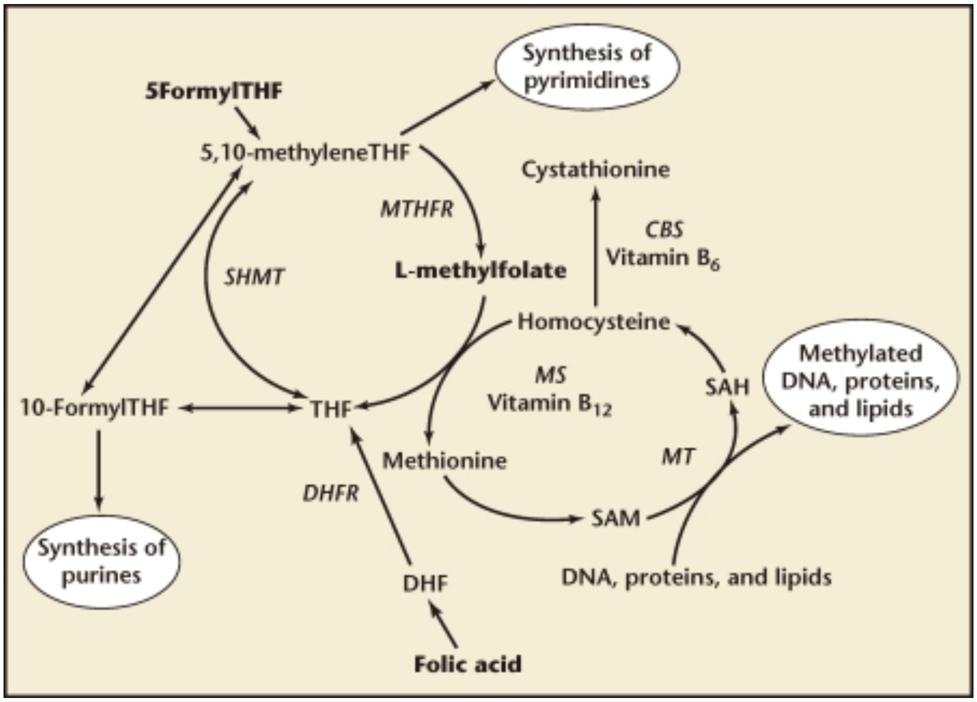


Figure 1.2 Folic Acid and One Carbon Metabolism. Reprinted from Greenberg et al., 2011

1.4 DNA Methylation

DNA methylation describes the addition of methyl groups (CH_3) to nucleotides within DNA, generally by specialized methyltransferase enzymes, such as DNMT1, DNTMT3A and DNMT3B in mammals (Lister and Ecker, 2009; Robertson and Wolffe, 2000). Such methylation is most frequently observed in cytosine, in particular those adjacent to guanine residues, i.e. CpG dinucleotides. Regions of the genome which are enriched for CpG dinucleotides are referred to as CpG islands, and are often co-located with promoters (Fatemi et al., 2005). Hypermethylation in the promoter region of a gene has been linked to decreased expression (Irizarry et al., 2009), through either direct prevention of transcriptional factor binding, or through recruitment of histones to create closed chromatin states (Hashimshony et al., 2003; Jones et al., 1998) . However more complex relationships between methylation and expression have also been observed (Wagner et al., 2014). Methylated cytosines are also at a greater risk for mutation, as they can be deaminated into thymidine. It has been hypothesized that the silencing of retrotranspons may be a key evolutionary reason for the presence of DNA methylation, in accordance with viewing such elements as primarily genetic “parasites”, rather than having important functional significance (Yoder et al., 1997).

DNA methylation plays an important role in many vital biological processes such as embryogenesis (Smith et al., 2012), cellular differentiation (Beerman and Rossi, 2015), chromatin structure (Hashimshony et al., 2003), and imprinting (Li et al., 1993). Aberrant DNA methylation is widespread in cancer (Baylin, 2005) and linked to both disease initiation and progression (Portela and Esteller, 2010). Thus alterations of DNA methylation provides one pathway by which FA may be acting to prevent NTDs.

The study of base modifications, of which methylated cytosine (5mC) is just one instance, has seen rapid advances in recent years. Ten-eleven translocation (TET) enzymes (Iyer et al., 2009) can further oxidized 5mc into 5-hydroxymethycytosine (5hmC), 5-formlcytosine (5fC), and 5-carboxylcytosine (5caC), which can be recognized by base-excision repair mechanisms. Methods have been developed to detect each of these bases through chemical conversion (Booth et al., 2013) or immunological assays (Wheldon et al., 2014). 5hmC has been found at high levels in mouse embryonic cells (Szwagierczak et al., 2010), while 5fmC has been found to have a highly tissue specific distributions during development, while 5caC may be a marker for active de-methylating processes. Another line of research has been into base modifications in RNA, particularly incorporation of N^6-methyladenosine (m^{6}A) (Fu et al., 2014; Lichinchi et al., 2016), which has been linked to numerous cancers.

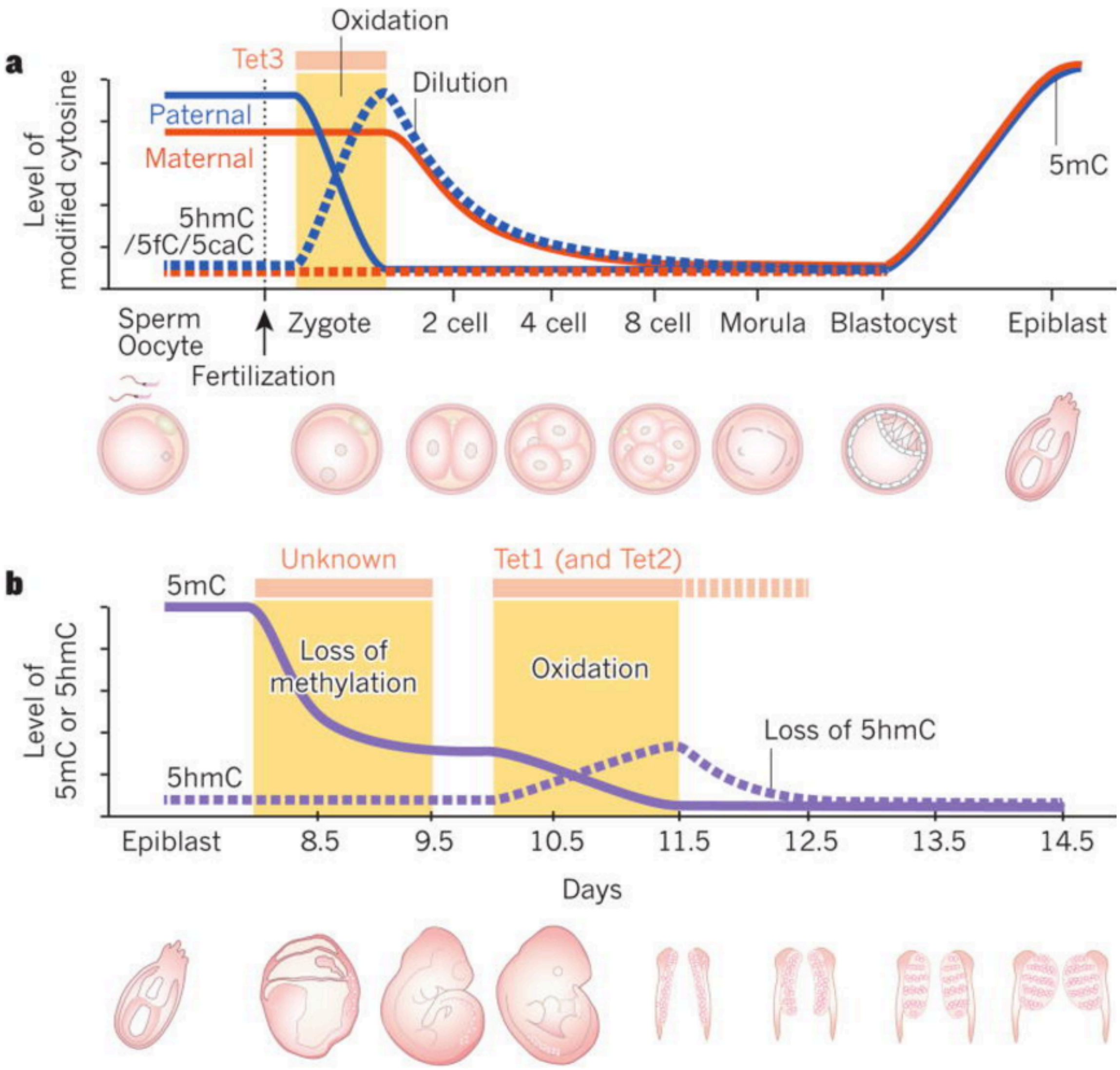


Figure 1.3 Dynamic of DNA Methylation Reprogramming in Mouse Embryos. **a**, Dynamics of 5mC and its oxidation products in pre-implantation embryos.. **b**, Illustration of the 5mC and 5hmC dynamics in primordial germ cells (PGCs) during their reprogramming. Reprinted from Kohli and Zhang (2014)

During gamatogenesis and embryogenesis, DNA methylation patterns are erased and re-established in a process known as reprogramming (Figure 1.3). Reprogramming during embryogenesis involves initial demethylation of the paternal genome by TET enzymes, resulting in the increase of the downstream oxidative products 5hmc, 5faC, and 5caC (Gu et al., 2011b; Inoue et al., 2011; Wossidlo et al., 2011). Afterwards, global methylation levels for modified cytosines in both maternal and paternal genomes are reduced passively via replication (Inoue and Zhang, 2011), until being re-established at the blastocyte stage around E6.5 (Borgel et al., 2010; Smith et al., 2012). At this point, methylation is laid down throughout the genome, with differential patterning amongst cells leading to lineage restriction to certain tissue types (Ji et al., 2010; Mohn et al., 2008). In cells that have been selected to become primordial germline cells (PGCs), a more complex epigenetic program is observed at the epiblast stage, where demethylation occurs via non-oxidative processes (evidenced by unchanged levels of 5hmc), and then undergoing a second stage of oxidative demethylation. In the case of female embryos, one copy of the X chromosomes is randomly selected on a per-cell basis for genome-wide methylation resulting in inactivation of its constituent genes (Hellman and Chess, 2007; Mohn et al., 2008).

Although DNA methylation is generally stable, it’s patterning does change with age (Ahuja and Issa, 2000), allowing the biological age of cells, tissues, and organisms to be inferred, in what has been referred to as the "epigenetic clock." In humans, a prominent example of such a predictor is Horvath's clock, so named for Steve Horvath's seminal study of methylation array data from 8,000 human samples including 51 tissue and cell types (Horvath, 2013). Horvath's clock utilizes 353 CpG sites, selected via elastic-net regression from a pool of over 21,000 sites, and is predictive across tissues, irrespective of their specific methylation patterns, with a reported correlation of r=0.96, which similar correlations reported on independent datasets (Gibbs, 2014). However, positive deviances in predicted epigenetic age from actual age, referred to epigenetic age acceleration, have been observed associated with Alzherimer's disease (Horvath et al., 2015), Parkinson's disease (Ritz and Horvath, 2015), and age-related macular degeneration (Lu et al., 2016). In general, lower levels of methylation are observed as an organism ages (Bjornsson et al., 2008).

1.5 *Lrp6*-Based NTD Models and Wnt Signaling

There are over 200 mouse models of NTD, however only 23 have been tested for their responsiveness to FA, and only 11 have in fact shown a positive effect (Harris and Juriloff, 2007, 2010). In particular, the curly tail mouse, the best studied model of spina bifida (van Straaten and Copp, 2001), is resistant to FA supplementation, though it can be rescued by inositol (Burren et al., 2010). The large number of potential candidates to study is indicative of the complexities of neurological development, with transcriptional differences being found along the neural column during fusion, indicating specialized processes for each region (Colas and Schoenwolf, 2001; Copp and Greene, 2010; Wallingford, 2005; Yamaguchi and Miura, 2013).

Our study focuses on two mutations in the Lrp6 (Low-density lipoprotein-receptor related protein 6) gene, which provide NTD models with opposite responses to maternal FA.

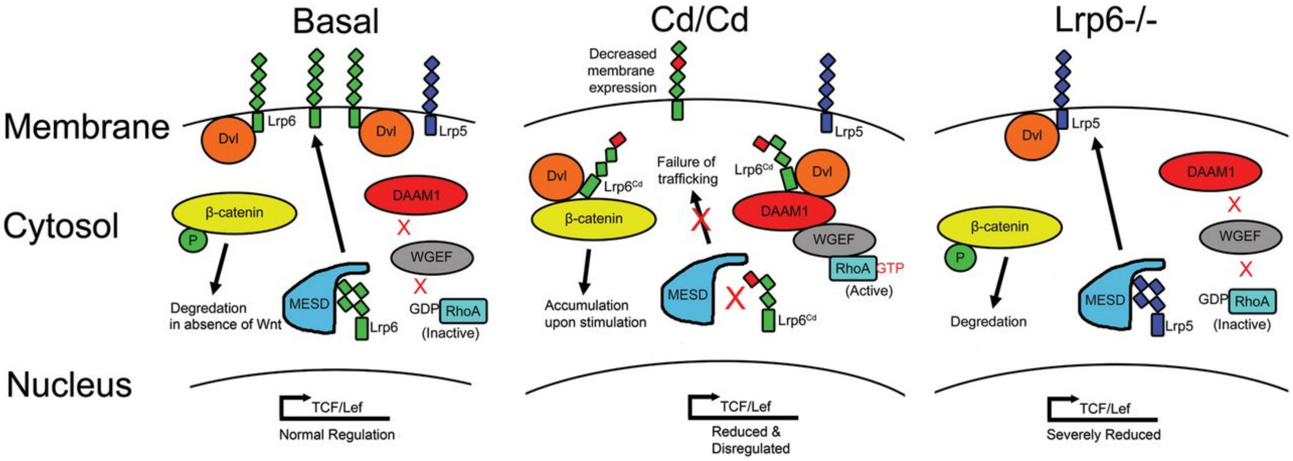


Figure 1.4 Model of *Lrp6* Disregulation within the Cell. Under basal conditions, *Lrp6* is efficiently trafficked and inserted into the cell membrane facilitated by MESD. The Cd mutation prevents *Lrp6* interaction with MESD and leads to cleavage and defective processing of *Lrp6*^{Cd}, all of which reduces localization of *Lrp6*^{Cd} on the surface of the cell. The mutant *Lrp6*^{Cd} accumulates within the cell to alter levels of β-catenin and increase GTP-RhoA levels through complexation with DAAM1. In the *Lrp6*$^{-/-}$ cells, only Lrp5 remains available for signal transduction, and it is insufficient to activate either the canonical or non-canonical pathways to the level necessary for proper neural tube closure. Reprinted from Gray, et al 2013

Under basal conditions, Lrp6 is inserted into the cell membrane by MESD (Figure 1.4), and is a co-receptor for the Wnt (Wingless/Integrated) signaling pathway, which plays a critical role in neural development. Wnts are a family of secreted molecules that regulate numerous developmental events through several signal transduction pathways. Wnt signaling falls into three known pathways: the canonical pathway, the non-canonical planar cell polarity (PCP) pathway, and the non-canonical calcium pathway (Amerongen and Nusse, 2009). *Lrp6* and *Lrp5*, along with their obligate co-receptor Frizzled (*Frz*) bind to *axin* in the presence of Wnt. The removal of axin from the cytosol prevents the formation of the beta catenin destruction complex, causing beta catenin to accumulate and be trafficked into the nucleus where it activates TCF/Lef transcription factors.

Lrp6$^{\mathrm{Cd}}$ consists of a single point mutation in the extracellular domain of the Lrp6 protein (Carter et al., 2005). Heterozygotes display a crooked tail. In *Lrp6*$^{\mathrm{Cd/Cd}}$ embryos, co-localization of *Lrp6* into the membrane is reduced due to complexation with DAAM1, activating GTP-RhoA, while also causing build-up of beta-catenin in the cellular membrane, and reduced activation of TCF/Lef. Untreated, resulting defects include embryonic lethality, exencephaly and runted pups with severe lumbosacral and tail deformities. However, maternal supplementation can enact a true rescue, with normal Mendelian distribution of genotypic ratios (Gray et al., 2013).

In *Lrp6*$^{-/-}$ embryos, there is insufficient signal transduction to activate either the canonical or non-canonical wnt pathways for neural tube closure, and the embryos are not viable. Defects include body axis truncation, limb defects, eye and palate defects and a high incidence of exencephaly and/or spina bifida. Maternal supplementation with FA causes a shift towards early embryonic lethality and exencephaly, though as a percentage of live births NTD is reduced (Gray et al., 2010). Recent genetic studies in humans have linked Lrp6 mutations in humans to NTDs (Lei et al., 2015) and tooth angenesis (Ockeloen et al., 2016), making these backgrounds ever more attractive models.

We hypothesized that the effect of FA on DNA methylation could contain information as to its phenotypic interactions with the Lrp6 *Cd* and null mutations. The most direct method by which this could be occurring would be epigenetic lesions caused by Lrp6 mutation status whose methylation levels were also effected by maternal FA, with corresponding changes in expression patterns in associated genes. For the null mutants, another possibility would be that downstream effects of the removal of Lrp6, in particular the reduction in TCF/Lef signaling, allow the addition of FA to disrupt the normal demethylation process during early embryogenesis. Such disruption, which would display itself as hypermethylation associated with increased maternal FA, could potentially be accompanied by increased expression of TET enzymes in order to compensate for the additional methyl groups. Such compensation would be expected relatively close to implantation to match the timeline set out by Kohli and Zhang.

A similar general mechanism could potentially explain the observations in the Cd background. This would be the case if the Cd mutation resulted in downstream interference with the reestablishment of methylation patterns prior to the epiblast developmental stage, with the additional FA compensating for this effect. If true, increased expression of DNMTs might be observed in the mutants, mirroring the hypothesis above in the nulls. However, the rescue of the crooked tail phenotype indicates a potentially much closer link between FA's effects and Lrp6 biology. Increased methylation and decreased expression of one or more of the proteins the mutant Lrp6 protein complexes with, such as DAAM1, would be of particular interest.

At the time the data for the samples was being collected, there was no published method for analysis of methylation data involving 2x2 or more complex designs as was needed for the study. This necessitated the creation of the multiDiff package (see below). In the discussion that follows, care must be taken to separate two separate uses of the word "interaction". In a biological context, the word interaction generally implies some form of binding, or potentially some form of co-location. However, there is a separate mathematical notion of "interaction," which refers to interaction terms in fitting a model. In brief, such interaction terms are usually the product of two other explanatory variables (i.e. C3=C1 * C2). As they are not a *linear* combination of other terms they do not generally cause any direct issues with model fitting. They correspond to non-linearity in the data, and are appropriate when the effect of one variable mediates the effect of the other, or when they both mediate each other. Hence the term “interaction”. Observation of a non-zero mathematical interaction indicates that the variables biologically interact at or upstream of the relevant data. The converse is not true- if the two individual biological effects “add together” when combined (see discussion of link functions in Chapter 2), then they may not require an interaction term.

CHAPTER 2

Analyzing Methylation Data with Complex Designs

2.1. Assays for DNA Methylation

Current studies of DNA methylation rely on two major classes of assays- those using microarrays, the most popular being the Illumina Infinium Human Methylation 450K BeadCHIP (Morris and Beck, 2015) and those using high-throughput sequencing. Their strengths and weaknesses parallel the use of these platforms for detecting genomic variants. Arrays have the benefit of being lower cost, and consistently collecting identical sets of sites across samples. Sequencing based methods can be more challenging to analyze, due both to the increased amount of data and processing time, as well as the necessity of harmonizing sites across samples, while offering the potential to cover a larger number of sites. By examining the methylation patterning of specific reads, it also becomes possible to look for differences in combinatorial entropy at a given loci, an approach which has been used to detect changes in what have been referred to as epiallelles in leukemia patients (Li et al., 2014).

Both the 450K Methylation arrays and NGS methods detect methylated cytosines by means of sodium bisulfite conversion. Sodium bisulfite converts unmethylated cytosines to uracils (Hayatsu, 2008), which are then treated as thymines during subsequent PCR reactions. For NGS data, specialized aligners, such as BISMARK (Krueger and Andrews, 2011), BSMAP (Xi and Li, 2009), or RMAPBS (Smith et al., 2009), utilize in-silico bisfulfite treated version of the reference genome to align reads. Methylation levels can then be called from the resulting bam or sam file by comparing the number of C and T reads at each covered loci.

Due to the time and expense of the conversion process, whole-genome bisulfite sequencing (WGBS) was, and remains, impractical for many studies. Reduced Representation Bisulfite Sequencing (Gu et al., 2011a; Meissner et al., 2005) allows for a cost effective approach. By using the Mspl digestion enzyme to fragment DNA at C^CGG loci, and performing size selection, it is possible to collect data from CpG islands throughout the genome, and thus the promoters of most genes. An enhanced version of the protocol, eRRBS (Garrett-Bakelman et al., 2015), increases the number of covered loci and allows for coverage of sites further out from CpG islands, referred to as CpG shores.

Methylation can be analyzed using binomial generalized linear modeling (see below). A common package for doing this is methylKit (Akalin et al., 2012) methylKit also allows use of DSS (Feng et al., 2014), a hierarchical Bayesian model using the beta-binomial, as well as the ability to control for covariates in the analysis. However, it currently lacks the ability to simultaneously analyze and visualize multiple covariates and their interactions.

2.2 Introduction to Generalized Linear Models

Generalized linear models (GLMs) are a class of statistical models used to analyze data that share an underlying distribution in a manner that generalizes the approach of standard linear regression (Agresti, 2015). In linear regression, one can write:

$$(1)\quad Y_i = \sum_j \beta_j X_{ij} + \varepsilon_i$$

Where Y_i is the feature being predicted in sample *i*, X_{ij} are the j predictors of samples i, β_j are the weights put on the features, and ε_i is the error for the ith sample. The appropriateness of the model can in part be assessed by how closely the errors, ε, actually follow a normal distribution with a single variance, σ^2 and mean zero. Non-normality of the errors can potentially occur if the variance is not constant across the range of linear predictors $X_{ij}B_{ij}$, in a condition referred to as heteroscedastacity. In such a case, more complex models become appropriate.

We can reformulate the linear model in the following equivalent manner:

$$(2)\quad \begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \sum_j \beta_j X_{ij} \end{aligned}$$

Here, we are imagining the Y_i's being drawn from normal distributions with the same variance, but with means controlled by the remaining features/covariates. The equivalence can be seen in the independence of the mean and variance for the normal. In addition the range of the linear predictor (now defined as μ) is the same as the range of the mean. In both formulations, we are also restricted to having the predictors having a linear effect on the mean.

The first conceptual step for a GLM is to consider other distributions besides the normal from which the independent variable can be drawn. However, both of the previously mentioned properties- the mean with domain equal to the reals, and independence of the mean and variance, may fail to hold. For example, neither is true of the Poisson distribution, one of the most common distributions to model count data. It has mean and variance both equal to a single parameter, λ, which can be any positive real number. Thus, if we tried to set:

$$\begin{aligned} &Y_i \sim Poisson(\lambda_i) \\ &\lambda_i = \sum_j \beta_j X_{ij} \end{aligned} \tag{3}$$

We could potentially run into both limitations. It is theoretically possible for us to have a predicted mean less than zero, outside the allowable range. In addition, it would make little sense to try and use the original formulation of the linear model- with the error now Poisson distributed. it would be unable to have zero mean and non-zero variance in the stochastic part of the model, thus failing to capture the notion of error.

GLMs solve this problem by introducing what is referred to as a *link function*, generally noted as *g*. The link function serves the purpose of both mapping the domain of the distribution to the reals, and allowing more flexibility in related covariates to the mean. Thus we have the standard formulation of a GLM as:

$$\begin{aligned} &Y_i \sim f(\mu_i;\sigma^2) \\ &g(\mu_i) = \sum_j \beta_j X_{ij} \end{aligned} \quad (4)$$

Here, *f* is a member of the exponential family of distributions, or potentially it's extension, the over-dispersed exponential family (Chang et al., 2001; Gelfand and Dalal, 1990). Hypothesis testing is usually done with either a Wald test, which applies a chi-square test to each coefficient, or a likelihood ratio test, which compares the likelihood of the specified model to a reduced one with one or more terms removed. Another alternative is Rao's score test, based on the score and Fisher information of parameter estimate can also be use. All three tests are asymptotically equivalent. For computational performance considerations, the Wald test was utilized in the analysis presented throughout.

GLMs and their extensions have been applied to the analysis of many kinds of biological data, including microarray data (Smyth, 2004), differential expression (Anders and Huber, 2010) and splicing in RNA-seq (Anders et al., 2012), DNA-methylation, neuronal spike patterns (Gerwinn et al., 2010). In the case of methylation data, we can set to f be the binomial distribution, and take the coverage as the (known) number of trials. Here, the link function must take the real line into the interval [0,1], which is done naturally by a sigmoidal function. The logit function, discussed below, is just such a sigmoid, and is the canonical link function for logistic regression. Other choices of link function may include the probit, log-complementary log (Agresti, 2015), or arcsine, discussed below.

2.3 Survey of published methods for base-resolution methylation data

methylKit's (Akalin et al., 2012) main functionality is built around a binomial GLM, a natural fit for binary response data (in this case methylated (C) and unmethylated (T) reads). Other approaches have been implemented in various software packages for methylation analysis. In general, tradeoffs must be made between model complexity, runtime, ability to handle heteroscedasticity, and Type I and Type 2 errors.

Instead of directly dealing with counts, some packages analyze the methylation ratio (methylated reads over coverage). These included Bsmooth (Hansen et al., 2012), IMA (Wang et al., 2012), Minfi (Aryee et al., 2014), COHCAP (Warden et al., 2013), CpGassoc (Barfield et al., 2012), and metilene (Jühling et al., 2016). For count-based methods, which thus retain information about coverage variability between samples, methylKit by default uses logistic regression, contrasting with RADMeth (Dolzhenko and Smith, 2014), MethylSig (Park et al., 2014), MOABS (Sun et al., 2014), DSS (Feng et al., 2014) and DSS-general (Park and Wu, 2016), which all use some form of beta-binomial regression. The beta distribution has support on the domain [0,1], and the variance of the distribution can be used to capture biological variability. A recent comparison of different methylation callers (Zhang et al., 2016), found that the ratio tests were not sensitive to small changes in methylation level, but that the beta-binomial implementations shared similar and superior performance.

2.4 Complex Designs and Assignment of Effect Sizes

Most of the previously mentioned packages are designed to call differential methylation between only two group comparisons- only methylKit, RADMeth, and DSS-general are able to control for covariates or analyze more than one group. Of these, only DSS-general was created to analyze complex or general designs. In their initial paper (Park and Wu, 2016), they find what they believe to be unacceptable performance for pure logistic regression. However they do not directly simulate interaction terms, a key question when one has more than one effect of interest. Also they do not take up the problem of estimating effect size in logistic or beta-binomial regression.

As has been increasingly appreciated (Sullivan and Feinn, 2012), statistical thresholds by themselves are generally insufficient when examining a large number of tests as is common in NGS datasets. A measure of effect size is vital to ensuring that to removing hypotheses that explain only miniscule amounts of the variation of interest.

For logistic regression and its relatives, the standard effect size is the log-odds of the regression coefficient. This corresponds to the fact that the log-odds changes by a constant factor when the linear predictor is moved by the corresponding amount. However, log-odds adds nothing to biological interpretability in the context of methylation data. Figure 2.1 illustrates this problem. Empirical estimates of methylation difference can be made in a straightforward manner with a single binary covariate, in despite of the fact that the same coefficient value may have different empirical effects depending on the methylation level of the baseline condition. As each site is analyzed separately until correction for multiple hypothesis testing, there is no conflict. However, this becomes completely ambiguous when dealing with the case of two or more covariates being analyzed together at a single site.

In order to solve this problem, we devised *the maximum difference estimate.*

2.5 The Maximum Difference Estimate

Let the input data consist of *N* CpG sites and *D* samples, with a design matrix **X** with *J* features after expanding interaction terms. For CpG site I (I = 1,2,...N) in sample d (d =1,2…,D), let Y_{id} and m_{id} be the methylated and total read counts respectively. At each site, logistic regression fits the unknown parameters β_{ij} data to the model:

$$\begin{gathered} Y_{id} \sim Binomial(\mu_{id}, m_{id}) \\ \log it(\mu_{id}) = \beta_{i0} + \sum_{j \geq 1} \beta_j X_{ij} \end{gathered} \quad (5)$$

where logit is the function log(p/(1-p)), β_{i0} is the baseline methylation at site I, and x_d and β_i are the dth row of the design matrix and ith row of the coefficient matrix.

From the estimates of the parameters *β* we define the maximum difference estimates for each term in the following manner. Let *s* be a binary vector with *J* entries, which we call a state vector. A particular *s* captures the patterns of effects operating at a given loci. Let s_k be the indexed set of such vectors (k=1,2,...,2^J). If we arranged the s_k's into a 2^J by J matrix, they can be seen as equivalent to a binary truth table for *J* variables (Figure 2.1B).

We define the maximum difference estimate for the jth predictor at site I, M_{ij} *as*

$$M_{ij} = \operatorname{sgn}(\beta_{ij}) \bullet \max_{1 \le k \le 2^J} \left[\log it^{-1}(\beta_{i0} + s_k \bullet \beta_i) - \log it^{-1}(\beta_{i0} + s_k \bullet \beta_i - \beta_{ij}) \right]$$

Effectively, we consider all possible states the site could be in, and compare the methylation levels with and without the factor of interest. We then use the state that maximizes the effect of the covariate to assign its effect size. We use the maximum due to the saturating nature of the sigmoidal curve, which by its nature makes it difficult to estimate multiple effects with small sample sizes as is common in bioinformatics studies. We also implement another methodology for those who wish to assign more conservative effect sizes, the mean difference estimate. Let L_{di} be the linear predictor for sample *d* at site *i*. Then we can assign the mean difference estimate, $A_{ij,}$ over the observed samples for factor j at site i:

$$A_{ij} = 1/D \cdot \sum_{d=1,,,D} \left[logit(L_{di}) - logit(L_{di} - \beta_j) \right]$$

Although the mean difference estimate has the benefit of making a clean empirical prediction on the data, it is inherently biased if sample sizes among cohorts are unequal.

Note that although the maximum difference estimate was defined above for logistic regression, it can be applied to any underlying model of the sigmoidal curve, i.e. any other link function. By it's design it will tend to be relatively aggressive in assigning effect sizes, and will be sensitive to the region in the domain of the linear predictor that has the greatest dynamic response, which for standard sigmoids will be close to zero in linear space.

In the context of methylation analysis, we will refer to the maximum difference estimate as the *maximum methylation difference* (MMD) of an effect at a given loci.

2.6 Benchmarking of False Positive Rate

For our FPR analysis, we used GEO dataset GSE61163, which was utilized in Zhang et al (2016) in comparing differential methylation callers in the single binary factor context. The samples are from 39 individuals with chronic myelomonocytic leukemia (CML), whose tumors were sequenced with the eRRBS protocol before treatment. Following Zhang et al, we randomly selected 5 subjects per cohort. The only alteration in the procedure was random assignment to four groups instead of two, imitating a 2x2 factorial design. Then both methods were run using the model:

Cov1 +Cov2 +IsBoth

, where IsBoth corresponded to an interaction term, i.e. IsBoth=Cov1*Cov2

This was done for 1000 trials, with a q-value of 0.01 used for all runs. Results are shown in Figure 2.1C & D. We observe that the maximum methylation difference threshold exhibit the ability to strongly control the FPR of multiDiff, dropping to less than one percent at an MMD threshold of 25% or greater. Notably, this thresholding had minimal effect on the already very low FPR of DSS. While demonstrating the power of controlling for biological variability, it is also indicative of the conservative results of utilizing beta-binomial models.

2.7 Benchmarking of Accuracy.

To benchmark accuracy, we followed a similar simulation procedure to the one outlined in the DSS-general paper, with some modifications. First, we added an interaction term to the simulation, since it was vital to ensure that our method could detect such interactions accurately. We then expanded the sites in the simulation by duplicating the baseline condition over eight chromosomes, each of which was associated with a single state vector for generating DMRs. This eliminated the need to specify the covariance of the effects within a given region, instead allowing us to assess the ability of a method to detect each possible differential state against the same background. Finally, we set the differential effect of the first covariate, where active, to be uniformly hypermethylating, the effect of Cov2 to be uniformly hypomethylating, and their interaction to be uniformly hypermethylating. The result of applying this protocol to an artificially flat methylation landscape is depicted in Figure 2.2 A, with the percentage of differential sites in a region set to 100%. The null region/chromosome (no effects, i.e. s=(0,0,0)), is shown in the middle of the landscape. The control cohort has methylation identical to this null region across each chromosomes. The treatment cohorts have differential signal added as appropriate.

Results are shown in Figure 2.3 for simulations using a sine link, and Figure 2.4 for those using logit link. Standard ROC curves are generated by varying p-value or q-value thresholds. Here we instead use varying MMD thresholds, after filtering by q-values <0.01, the standard cutoff, to demonstrate the value of . Each curve is the average of 20 runs. Simulations in the bottom row were down with differential effect sizes set to 10x the original parameters. In all combinations of parameters tested, multiDiff had a higher AUC. Taking all runs over all parameters, the difference in performance was statistically significant ($p<2.2e^{-16}$, paired t-test). Thus, an effect size threshold can make performance of the binomial analysis comparable if not superior to that of beta-binomial

2.8 Application of multiDiff to P2 Cd Male Dataset

We applied multiDiff to a novel eRRBS dataset of Crooked tail mice. The Crooked tail (Cd) mouse is a model of neural tube defects that displays rescue under maternal supplementation of folic acid. Whole blood was collected from 16 male animals in a 2 x 2 factorial design (*Lrp6*$^{Cd/+}$ vs wildtype, 2ppm vs 10ppm FA diet) design. In Table 2.1 we show the result of running the standard methylKit analyses to interrogate the effect of diet in each genotype, and then intersecting them, compared to running multiDiff on the data, with and without interactions. multiDiff without interactions is able to find ~2x more sites at the 25% threshold, and ~90x when the interaction term is included. The interaction term is associated with 70,924 sites at the 25% threshold, which are not possible to recover from the single factor analyses. To attempt to do so, one could take the symmetric difference of the sites found in each genotype. However, the cardinality of the symmetric difference is bounded above by the size of the union of the two sets, achieving that bound only when they are disjoint. in this example analysis, we could thus even at the theoretical maximum find only ~50% of interacting sites by analyzing each genotype separately.

In Figure 2.5 we show the comparison of log p-values of the different analyses. Although there is an overall positive Pearson correlation between the wt and het values (, visually there appears to be a negative relationship. This appears due to the relative depletion of sites that are highly significant in both cohorts. Mathematically, such distinct effects are provided for by interaction terms, in this case the term Genotype:Diet. Biologically, thsThus comparing figure 2.6 A and B, we are able to see via inspection that there is enrichment for high p-values in the expected regions (up and to the right) when we fit the data using model with interactions, compare to without.

Figure 2.6 shows the standard visualizations of multiDiff output on the P2 Cd Male dataset, covering over 1.3 million CpGs, with the MMD threshold set to 25%, and q-value threshold set to 0.01, the standard settings. All loci that are not called as differential under the user's statistical and effect size thresholds are removed by default. The heatmap is designed to be analogous to those for RNA-seq. Each effect is shown in terms of a binary call matrix annotating the diagram, as well as the MMD associated with the effect in the heatmap proper. The violin and bar plots show the MMD distribution and number of DMCs at the input thresholds.

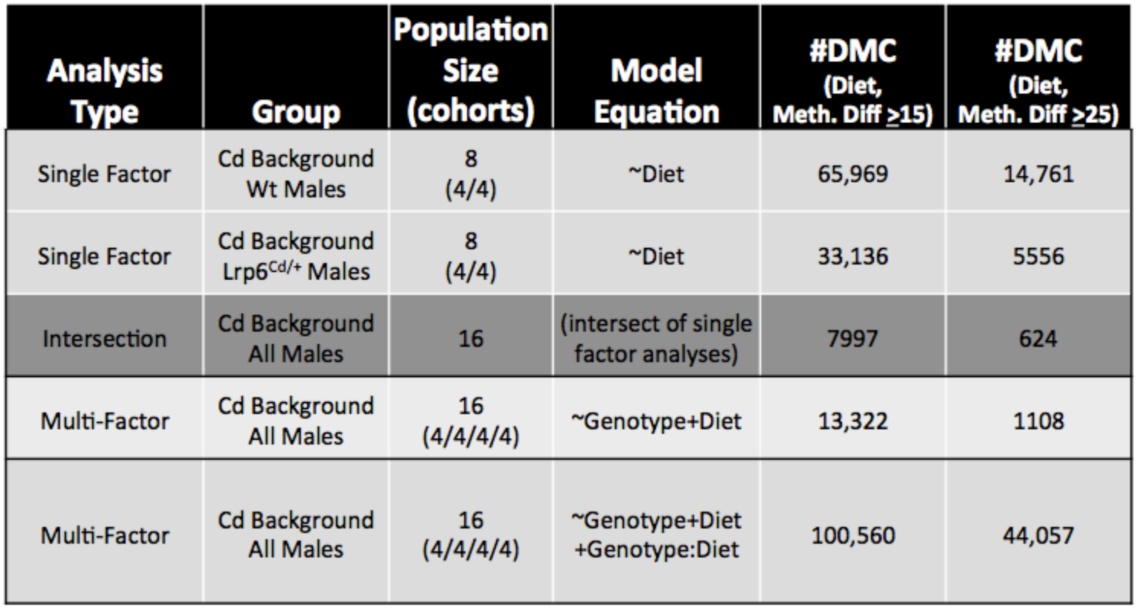
2.9 Additional Comparison to DSS

We further investigated performance characteristics by analyzing the Cd mouse dataset with DSS, in both it's single and general modes. The results are shown in Table 2.2. In single factor analysis, the DSS results were always a strict subset of the multiDiff results, with the number of hits for multiDiff even after MMD filtering more than an order of magnitude greater for both Wt and mutant animals. The only site that was called in all four analyses, before and after MMD thresholding, was in the exonic region of *Egln2* (Egl-9 Family Hypoxia Inducible Factor 2). Egln2 is responsible for post-transcriptional modification of Hypoxia Inducible Factor (HIF), which is involved in oxygen homeostasis (Semenza, 2001). In contrast, the 624 sites found by intersecting the binomial results, although still a fraction of those found when analyzing the full dataset together, contained 32 DMCs in promoters, shown in Table 2.3. However, no significant pathway enrichment was detected in these sites, nor overlap with known wnt signaling genes.

No overlap between the methods was observed in the either the simple two factor model (Diet+ Genotype), or the model with interactions (Diet + Genotype). After controlling for effect size, all of the hits produced by DSS were filtered out in the more complex analyses. We believe this to indicate that beta-binomial regression may be too conservative in such settings with relatively low sample sizes.

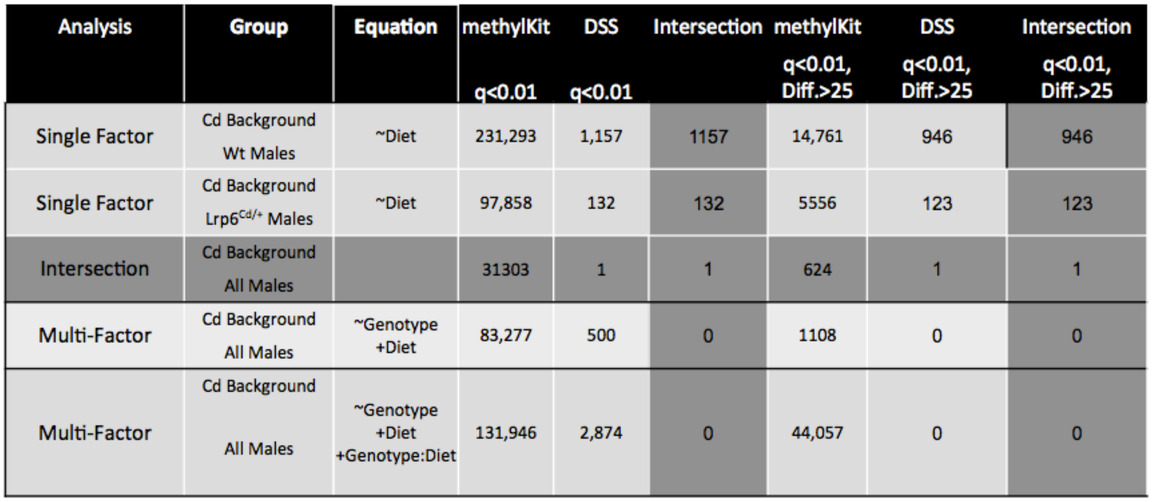
2.10 Conclusion

multiDiff, implementing the maximum difference estimate, is able to control the normally high FDR associated with logistic regression, and displays superior AUC performance to an alternate method for analyzing complex methylation designs.



Analysis Type	Group	Population Size (cohorts)	Model Equation	#DMC (Diet, Meth. Diff ≥15)	#DMC (Diet, Meth. Diff ≥25)
Single Factor	Cd Background Wt Males	8 (4/4)	~Diet	65,969	14,761
Single Factor	Cd Background $Lrp6^{Cd/+}$ Males	8 (4/4)	~Diet	33,136	5556
Intersection	Cd Background All Males	16	(intersect of single factor analyses)	7997	624
Multi-Factor	Cd Background All Males	16 (4/4/4/4)	~Genotype+Diet	13,322	1108
Multi-Factor	Cd Background All Males	16 (4/4/4/4)	~Genotype+Diet +Genotype:Diet	100,560	44,057

Table 2.1. Comparison of multiDiff output to intersection of standard methylKit analyses. Methylation Difference is the empirical difference in the group means for the single factor analysis, MMD for Multi-factor. q-value cutoff was 0.01.

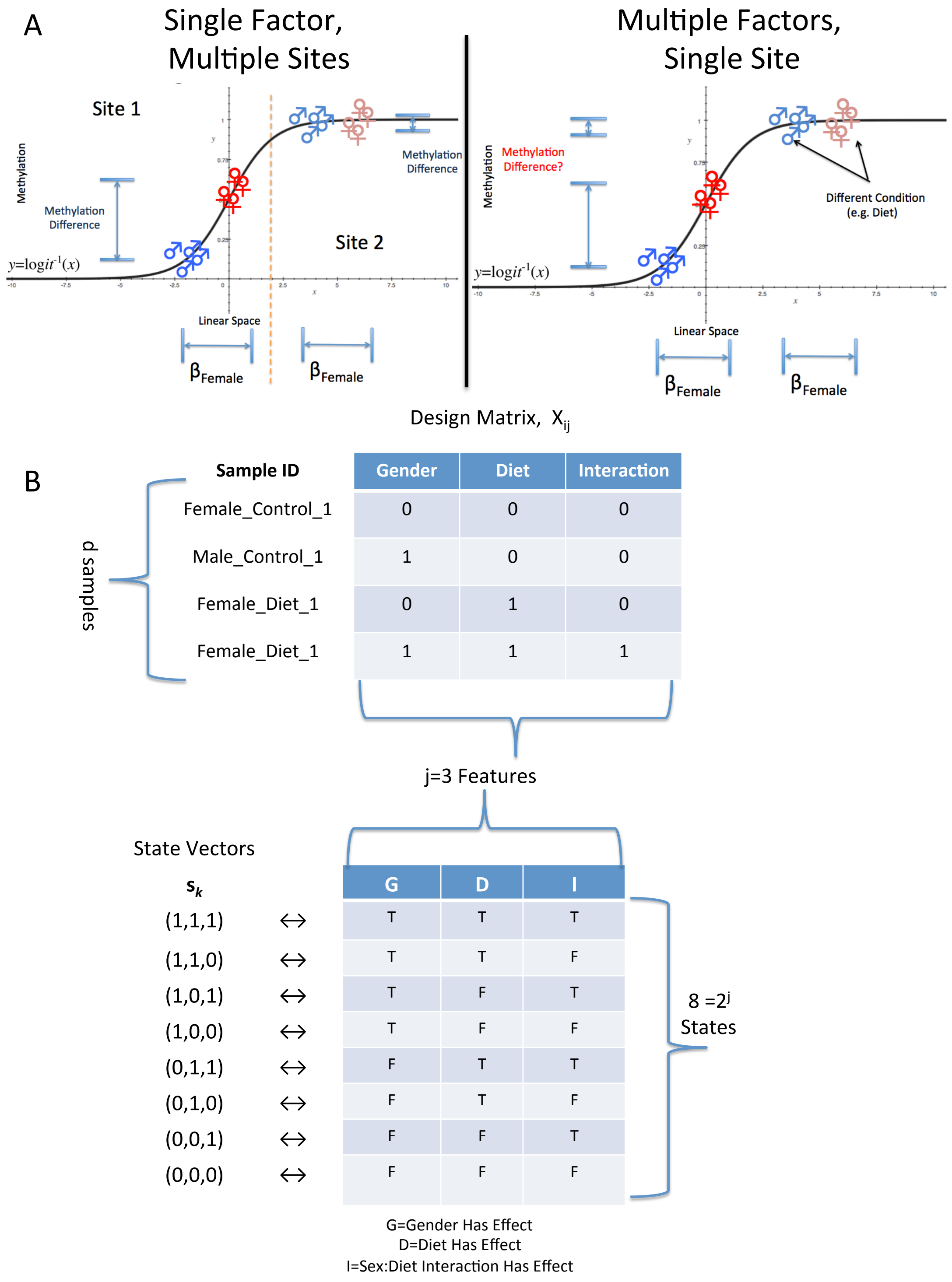


Analysis	Group	Equation	methylKit q<0.01	DSS q<0.01	Intersection	methylKit q<0.01, Diff.>25	DSS q<0.01, Diff.>25	Intersection q<0.01, Diff.>25
Single Factor	Cd Background Wt Males	~Diet	231,293	1,157	1157	14,761	946	946
Single Factor	Cd Background $Lrp6^{Cd/+}$ Males	~Diet	97,858	132	132	5556	123	123
Intersection	Cd Background All Males		31303	1	1	624	1	1
Multi-Factor	Cd Background All Males	~Genotype +Diet	83,277	500	0	1108	0	0
Multi-Factor	Cd Background All Males	~Genotype +Diet +Genotype:Diet	131,946	2,874	0	44,057	0	0

Table 2.2. Comparison of multiDiff and DSS output on Cd Male Dataset. Numbers are the number of DMCs called associated with diet. Diff. is the empirical difference in the group means for the single factor analysis, MMD for Multi-factor. The sole site found in the intersection of the Single Factor analyses is in the exonic region of Egln2.

Nfia,Pinlyp,Slc9a9,Klhl38,Hs3st2,Ino80e,Gm13003,Denr,Fap,Col18a1,Pdzd7,Myh14,Ptgds,Fcgbp,Mydgf,Popdc2,Zrsr1,Alpk3,Nhsl2,Apcdd1,Gnat2,Frk,Psma8,Ltk,Ddx4,Sema3b,Dock2,Ehf,3110099E03Rik,Gm16287,Frmpd1os,5830416I19Rik

Table 2.3 . List of genes with FA-associated DMCs in promoter regions in Lrp6 Cd and Wt Embryos. DMCs were intersected from the single factor methylKit analysis, MMD>25%



Design Matrix, X_{ij}

Sample ID	Gender	Diet	Interaction
Female_Control_1	0	0	0
Male_Control_1	1	0	0
Female_Diet_1	0	1	0
Female_Diet_1	1	1	1

State Vectors

s_k		G	D	I
(1,1,1)	↔	T	T	T
(1,1,0)	↔	T	T	F
(1,0,1)	↔	T	F	T
(1,0,0)	↔	T	F	F
(0,1,1)	↔	F	T	T
(0,1,0)	↔	F	T	F
(0,0,1)	↔	F	F	T
(0,0,0)	↔	F	F	F

G=Gender Has Effect
D=Diet Has Effect
I=Sex:Diet Interaction Has Effect

Figure 2.1 Effects and Effect Sizes in Logistic Regression. A. Similar distributions of data can be readily assigned a biologically meaningful effect size in terms of methylation when originating from different sites, but do not admit ready interpretation with multiple factors. B Binary state vectors capture the combination of effects acting on a given locus. The set of state vectors can be arranged to form a truth table over possible effects.

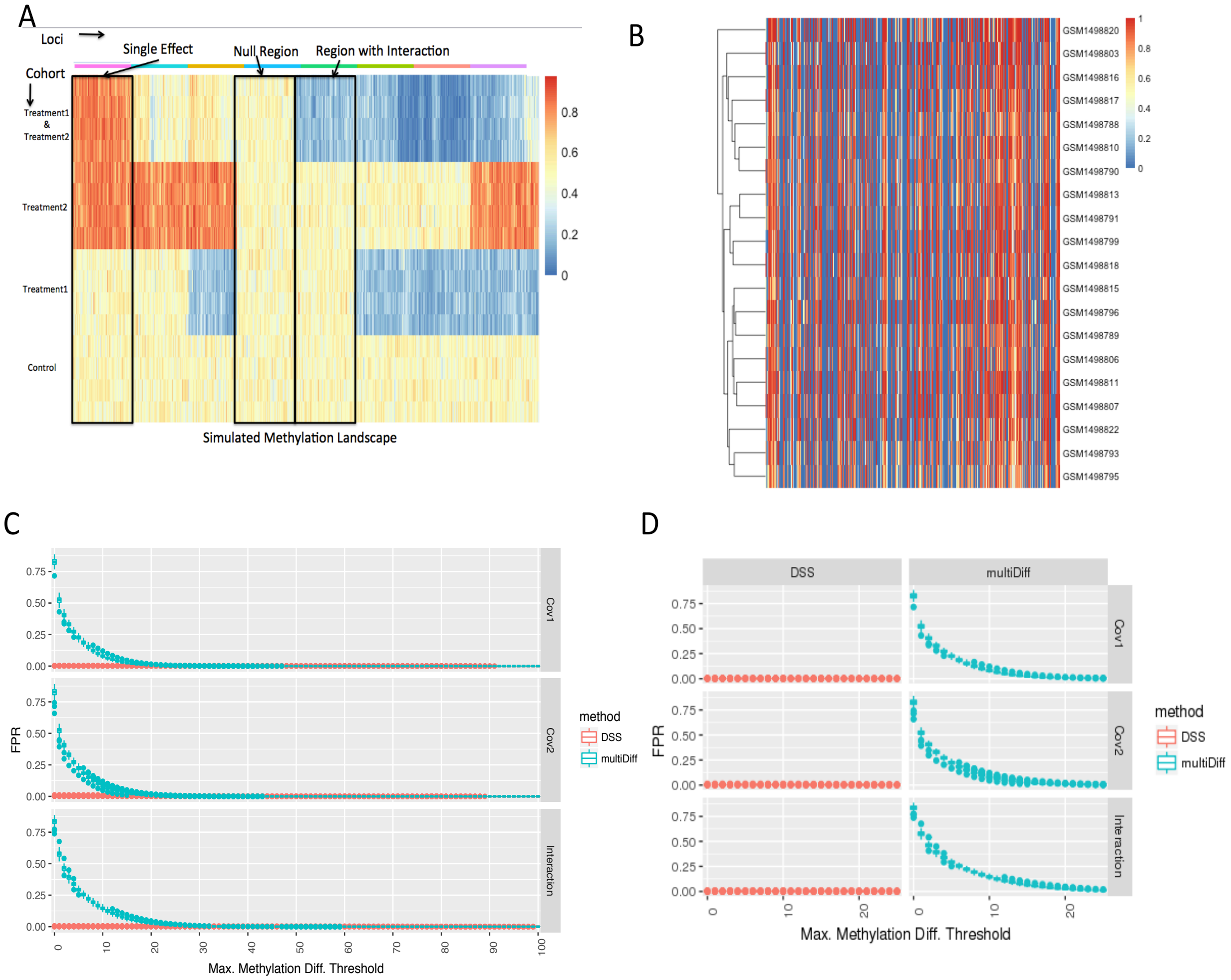


Figure 2.2 FPR Benchmarking Simulations. A. Simplified methylation landscape, with each chromosome corresponding to a region where a different set of effects is active. B. Raw methylation signal on random subsample of the CML cohort C. Plot of confidence intervals of FPR over 100 trails. D. Inset showing performance when the Maximum Methylation Difference (MMD) threshold goes from zero to 25%.

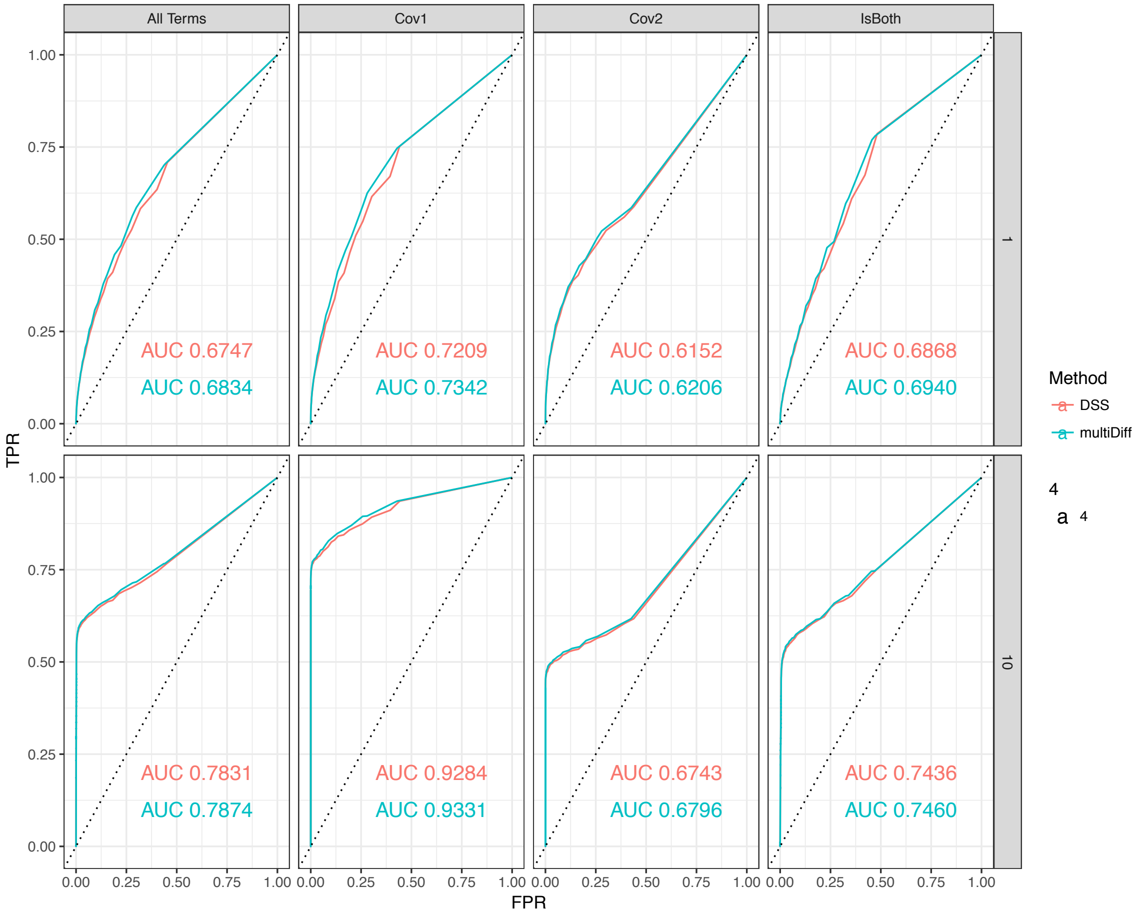


Figure 2.3. Comparison of multiDiff and DSS, simulating with sine link. Top row is simulated using effect size reported in Park and Wu 2016, bottom row has effect sizes increased by 10x.

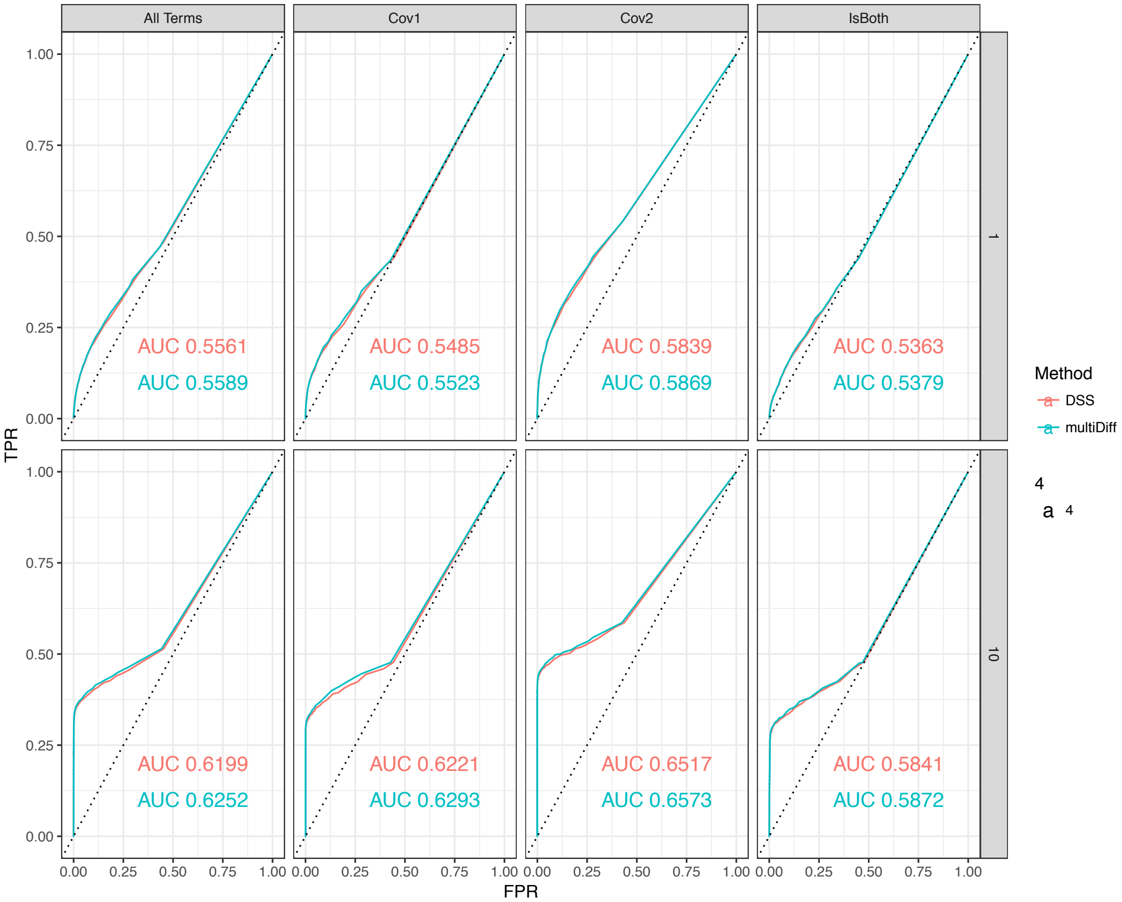


Figure 2.4 Comparison of multiDiff and DSS, simulating with logit link. Top row is simulated using effect size reported in Park and Wu, 2016, bottom row has effect sizes increased by 10x.

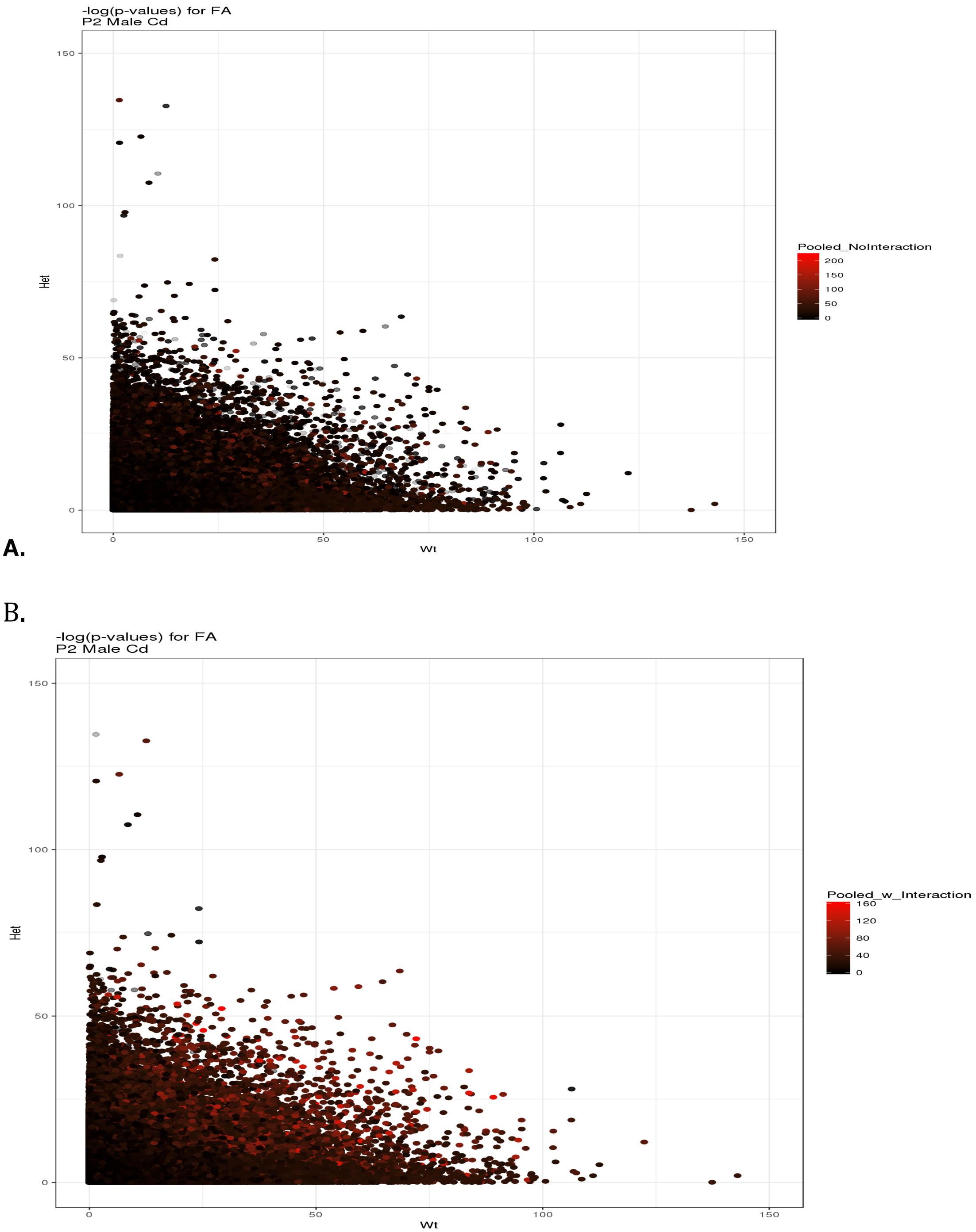


Figure 2.5 Comparison of Single Factor and Pooled P-Values. A. Log p-values for the Wt and Cd mutants, colored by the log p-values in the analysis of the pooled data without the interaction term Genotype:Diet . B. Log p-values for the Wt and Cd mutants, colored by the log p-values in the analysis of the pooled data run with atheinteraction term.

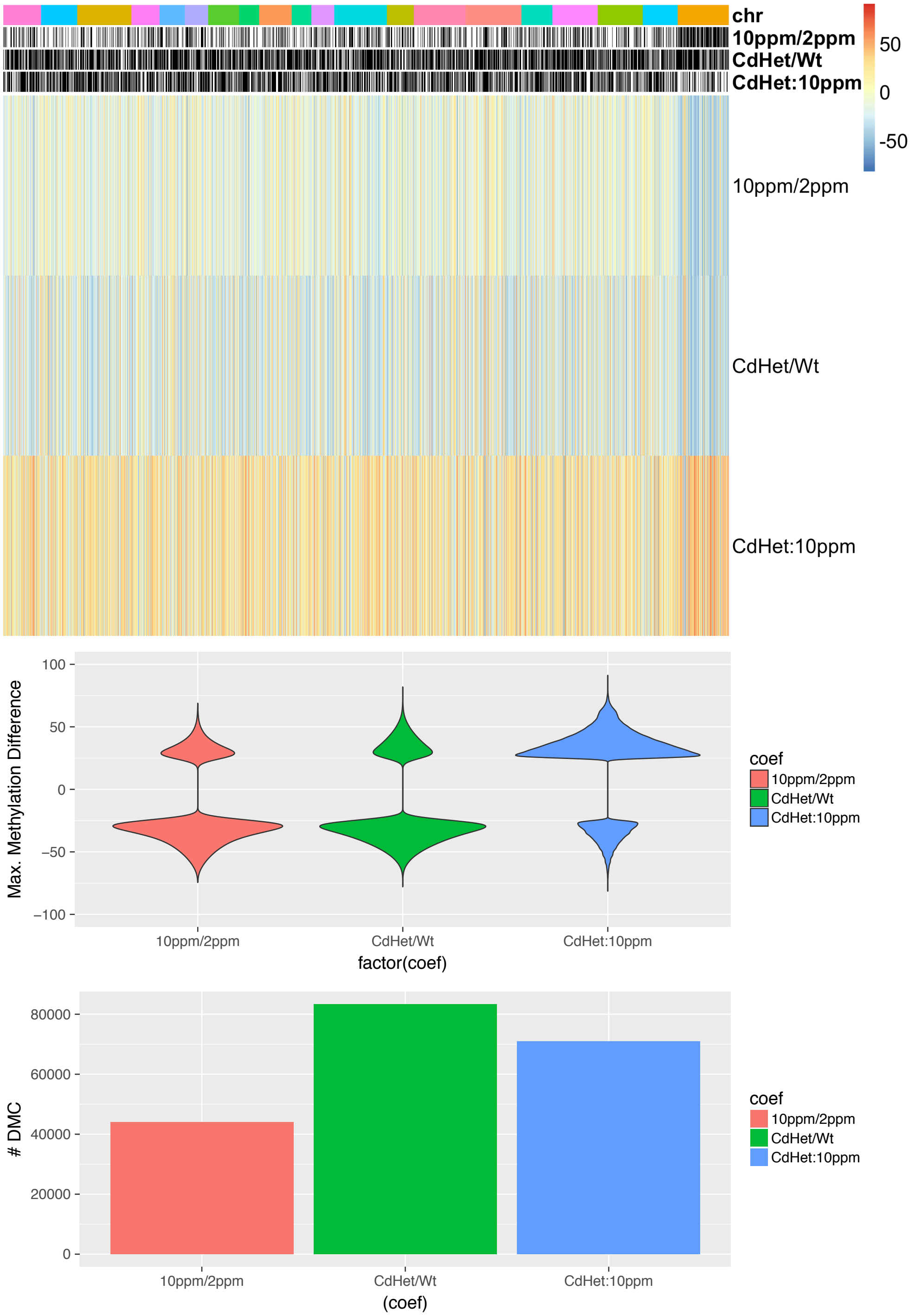


Figure 2.6 Visualization of multiDiff output for P2 Cd Male Dataset. DMCs called with q<0.01, Maximum Methylation Difference ≥ 25

CHAPTER 3

Analysis of FA Action in *Lrp6* NTD Models

3.1 Potential Mechanisms of FA Action

We hypothesized that changes in DNA methylation due to FA supplementation were key to understanding the effect of FA on Lrp6 *Cd* and null mutants. Three more specific hypotheses are depicted in Figure 3.1, illustrating different potential models of FA action, and their associated patterns of methylation. Since the same classification scheme can be applied to expression data, we consider both within the same framework.

In the additive model, both the FA and Lrp6 mutation affect the observed level of a fixed marker (either percent methylation at a loci or expression of a transcript). These effects persist and additively combine when the interventions are paired. In the independent model, FA and the mutations operate at distinct loci, and their interactions occur downstream of the observed marker. Therefore the critical markers underlying FA's effect are not necessarily associated with a genotype-driven change. However, FA's effect continues to be observed under the mutation condition. Finally, in the combinatorial model, both FA and the mutation must be present to create a differential effect. This suggests an interaction upstream of the marker being observed. The combinatorial model can

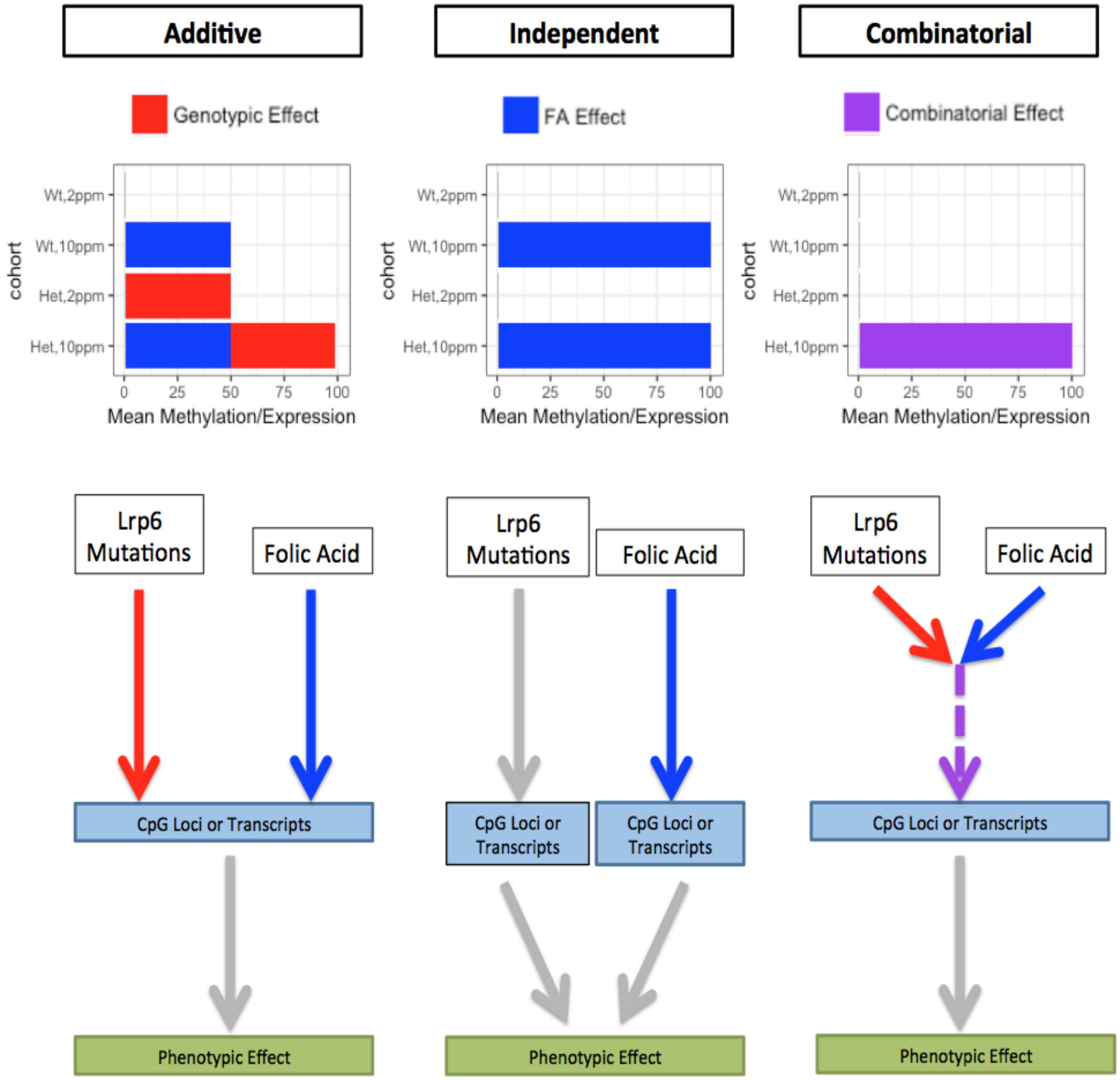


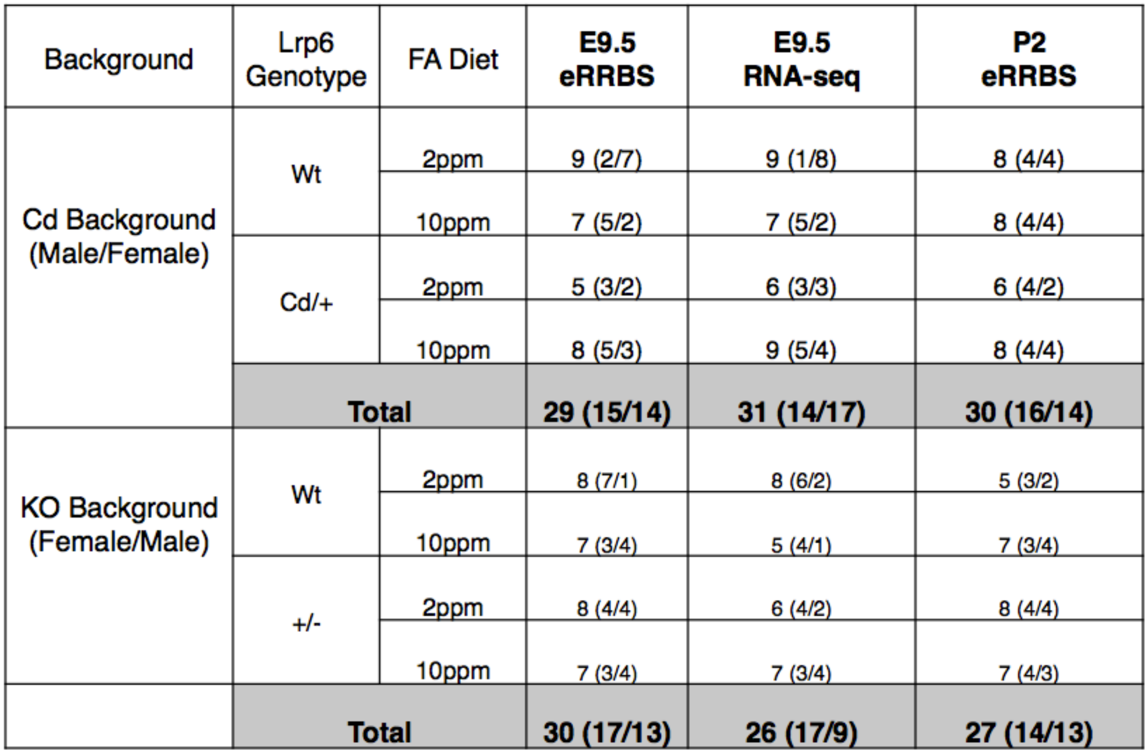
Figure 3.1. Potential Mechanisms of FA Action. In the Additive model, both FA and mutation status affect the given marker, and their effects add together when both are present. In the Independent model, FA and the Lrp6 mutation are affecting distinct loci or transcripts, and interact downstream when present together. In the Combinatorial model, the critical loci are only affected when both FA and the mutation are present

be described as either a Lrp6-dependent effect of FA, or an FA-dependent effect of Lrp6 status, but here they are mathematically equivalent.

To investigate the evidence for these different models, we performed enhanced reduced representation bisulfite sequencing (eRRBS) to profile DNA methylation at base-level resolution and RNA-seq, both assays on Embryonic day 9.5 (E9.5) Lrp6 knockout (KO) or Lrp6 *Cd* mouse lines.. We gathered heterozygous mutants and wildtype siblings whose dams had been fed either control (2ppm) or FA supplemented (10ppm) diets (Carter et al., 1999; Gray et al., 2010). E 9.5 was chosen as this corresponds to the end of neurulation. Heterozygous animals were used so that embryos could be appropriately developmental-stage matched for somite count and to permit comparison with postnatal ages, as *Lrp6*$^{-/-}$ are non-viable. We additionally performed eRRBS on whole blood from P2 mice to look for evidence of sites with persistent changes in methylation associated with maternal FA supplementation and mutation status. These could prove to be useful biomarkers with applications to NTDs human. The design of the full dataset is shown in Table 3.1.

3.2 Maps of Differential Methylation

Figure 3.3 shows the maps of differential methylation in both backgrounds without regard to mechanism. We observe that in both backgrounds the largest number of loci is affected by maternal FA, followed by mutation-FA interactions, and last the mutations themselves. However, in the Cd background, FA's effect had a strong bias towards increasing methylation throughout the genome, even when it wasn't called as having an effect, with the exception of chrX. On the KO background, FA's distribution of effects were more strongly bimodal, with some bias towards hypomethylation.



Background	Lrp6 Genotype	FA Diet	E9.5 eRRBS	E9.5 RNA-seq	P2 eRRBS
Cd Background (Male/Female)	Wt	2ppm	9 (2/7)	9 (1/8)	8 (4/4)
		10ppm	7 (5/2)	7 (5/2)	8 (4/4)
	Cd/+	2ppm	5 (3/2)	6 (3/3)	6 (4/2)
		10ppm	8 (5/3)	9 (5/4)	8 (4/4)
	Total		**29 (15/14)**	**31 (14/17)**	**30 (16/14)**
KO Background (Female/Male)	Wt	2ppm	8 (7/1)	8 (6/2)	5 (3/2)
		10ppm	7 (3/4)	5 (4/1)	7 (3/4)
	+/-	2ppm	8 (4/4)	6 (4/2)	8 (4/4)
		10ppm	7 (3/4)	7 (3/4)	7 (4/3)
	Total		**30 (17/13)**	**26 (17/9)**	**27 (14/13)**

Table 3.1 Summary of Lrp6 NTD Mice Model Dataset. Counts are those after outlier samples were identified and removed during exploratory analysis

3.3.1 Analysis of FA Mechanisms in *Cd* Background

Results of genic annotation of differentially methylated cytosines and differentially expressed genes (DEGs) assigned by mechanistic classification are seen in Table 3.3. We refer to the annotated genes from the methylation analysis as differentially methylated genes (DMGs) No overlap was found between the DEGs (q<0.05, ILFCI≥0.5), and DMGs in the Cd background for any of the mechanisms.

Associated with the independent mechanism was enrichment (using STRING DB (Szklarczyk et al., 2017)) amongst the DEGs for housekeeping pathways in the nucleus such as regulation of nucleobase-containing compound metabolic process (fdr=4.28e^{-5}) and gene expression (fdr=6.05e^{-5}) (Table 3.2). There was no observed enriched pathway in the DMGs.

Since FA rescues *Cd* mice, we looked for connection between Lrp6 and the lists of DEGs and DMGs by submitting them to StringDB along with the following list of Lrp6-associated genes: Ctb1, Daam1, Lrp6, Lrp5, Dvl1, Rhoa, Arhgef19, Mesdc2, Frzb, Axi, and observed which if any genes were found to have connections to this input set. Among the DMGs, only S1pr4, which is involved in cell migration shared a link to Lrp6 via RhoA. Similarly, the DEGs Crtc1, Dot1l, Dvl3, Dyrk1b., Gli2, Mll2, Myh11,and Hipk2 are connected to wnt signaling and Lrp6 as can be seen in Figure 3.3. Dvl3 shares strong similarity to Dvl1, which is normally recruited by Lrp6 to the plasma membrane. Hipk2 is part of a digenic NTD model of exencephaly with Hipk1 (Isono et al., 2006).

3.3.2 Analysis of Combinatorial Mechanisms in *Cd* Background

More DEGs were found associated with the combinatorial mechanism (91) than the independent mechanism (46). Once, again, known NTD gene Hipk2 and Dvl were found to be differentially expressed associated with the combinatorial mechanism. Zic3 is also known to be an NTD gene (Carrel et al. (2000); Klootwijk et al. (2000); Purandareet al. (2002); Lickert et al. (2005)) however it's location on chrX meant that it's signal was highly confounded with sex differences.

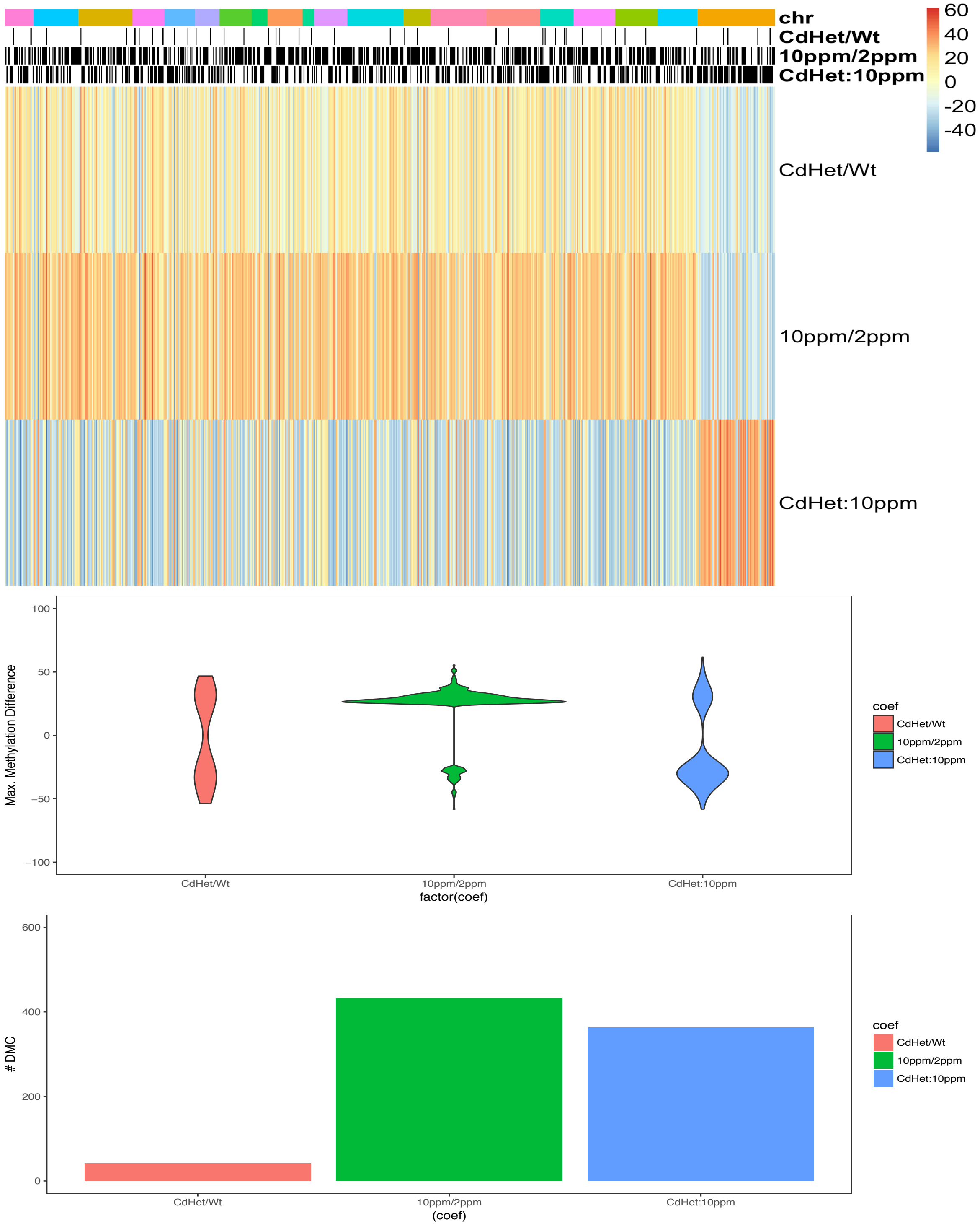


Figure 3.2 Differential Methylation in Lrp6 Cd KO Background. Top Row: MMD Heatmap of effects with differential calls of loci indicated at top of figure. Middle Row: Violin plots showing MMD distribution for each effect. Bottom Row: Number of DMCs associated with each effect

Mechanism	Additive	Independent	Combinatorial
Differentially Methylated Genes (DMC in First Exon or >3 in Promoter, multiDiff: q<0.01, Max. Meth Diff>25)	-	1700008J07Rik,1700018L02Rik,Ccer1,Gm17762,Ier5,Magee1,Ndnl2,Olfr551,Pcdh7,Pgrmc1,Rab33a,S1pr4,Zic3,	Ccer1, Chst7 Pgrmc1, Zic3
Differentially Expressed Genes (Deseq2: abs(lfc)>0.5, qvalue<0.05)	-	5930403L14Rik,Acsf2,Arhgap33,Bahcc1,Col6a2,**Crtc1**,**Dot1l,Dvl3**,Dyrk1b,Fam167a,Foxp4,Gli2,Gltscr1,Gm8615, Hipk2,Hist1h4c,Hist1h4f,Lrrn4,Map1a,Mbd6,Mll2,Msi1,Myh11,Nckap5l,Ncor2,Nfic,Pla2g7,Prdm16,Prrc2a,Rai1,Rfx2,Rpl38,Rpph1,Rps16,Sipa1l3,Srrm4,Tnrc18,Tox2,Wiz,Wnk2,Wnk4,Zfhx3,Zfhx4,Zfp536,Zfp628,Zmiz2,	1700085B03Rik,2610203C20Rik,2810055G20Rik,2900092D14Rik,3110039I08Rik,4930487H11Rik,4933400A11Rik,4933407K13Rik,Adamtsl4,Adat3,Aifm2,Ankle1,Ankrd13b,Ankrd16,Ankrd23,Arhgef4,BC024139,Baiap3,Ccdc74a,Cfp,Clcn1,Cnpy1,Col16a1,Col27a1,Dcp1b,Dennd4b,Dmpk,Dnm3os,Dpm3,Dvl1,Fam167a,Fam193b,Fam196a,Folh1,Fuz,Gigyf1,Gm10578,Gm16907,Gm6878,Gpr35,Gpt,Gramd1a,Hesx1,Hist1h4c,Hoxd3,Hoxd4,Il18bp,Itgb7,Leng8,Lmtk3,Lrrc16b,Mamdc4,Man2c1,Map4k2,Meg3,Miat,Mir1197,Mir3064,Mir3091,Mpp3,Mus81,Mycbpap,Neurod1,Nfic,Nhlh1,Npr2,Phf1,Phf21b,Pkn3,Pla2g4b,Ppp1cc,Ppp4r1l-ps,Psd2,Ptprn2,Rfx1,Rpl10a,Rpl38,Rps16,Rps18,Rps29,Slc2a4rg-ps,Slc9a5,Srrm4,St18,Tcirg1,Tnfrsf25,Trim58,Trpv6,Whrn,Ypel4,Zbtb48,M
Intersection	-	-	-

Table 3.2 Genes Associated with Mechanisms of FA Action in Cd Background

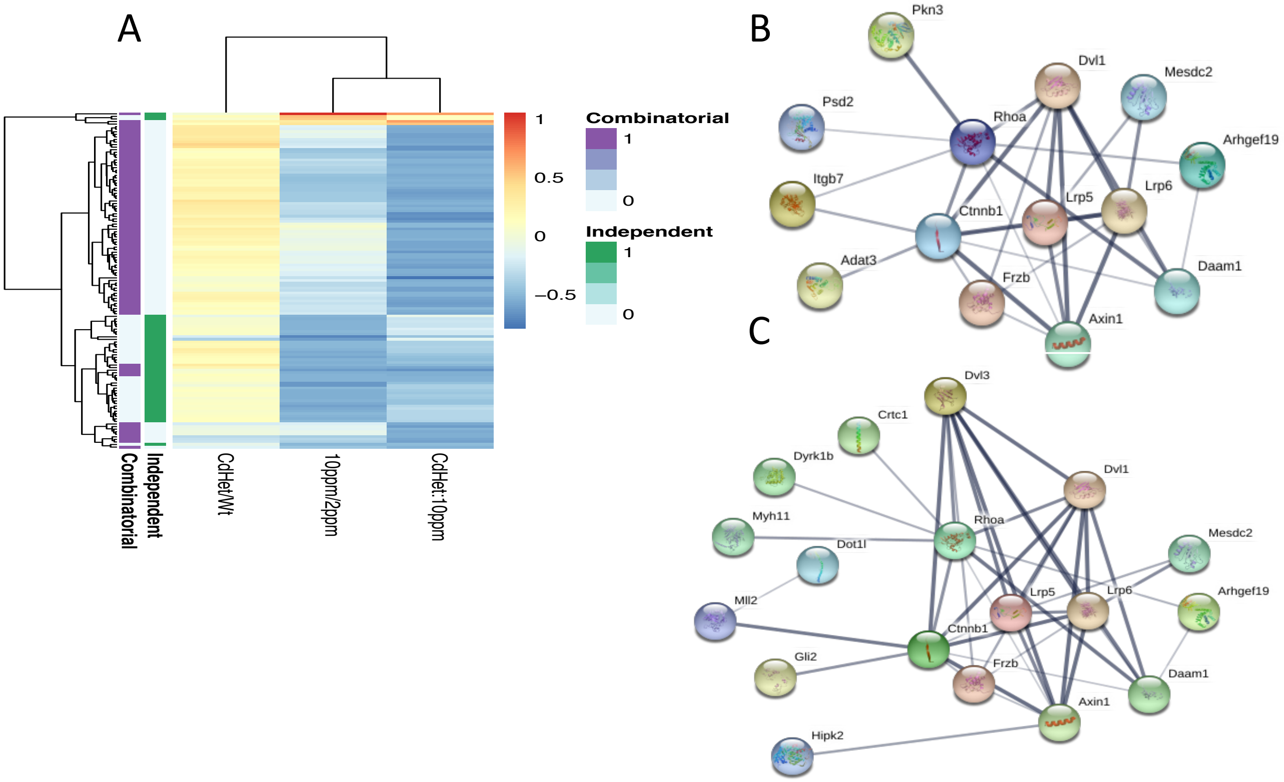


Figure 3.3 Analysis of Differentially Expressed Genes, Cd Background. A. Heatmap of Log2FoldChange, Cd Background. B. Combinatorial DEGs Associated with Wnt Signalng C. Independent DEGs Associated with Wnt Signaling

3.3.3 Discussion of Mechanisms in Cd Background

The data does not support the additive mechanism at all in the *Cd* strain, but there does not seem to be evidence that both the independent and combinatorial mechanisms are active. Most interestingly, even when it doesn't pass the threshold to be significant, the interaction between the diet and mutation is ever present, slightly repressing expression. The simplest interpretation for the main effect of the diet is the additional carbon is being used to counteract the known and observed hyperactivity associated with the Cd mutation through silencing. However, this does not explain the mechanisms driving the interaction effects.

There are only 5 genes up-regulated by the FA:Cd Interaction. The second ranked is Folh (Folate Hydrolase, LFC=0.665). Folh1 acts as a glutamate carboxypeptidase on folate, and is known to be expressed in the central and peripheral nervous system. It appears to be a good candidate for follow up study. If it is playing a role in FA rescue, then the initial hypothesis would that it is playing a role in trafficking beta catenin into the nuclease, directly by some form of recruitment, or indirectly, by activation of the actin destruction complex. Also of interest is Hesx1 (LFC=0.5342759), a homeobox gene which is a transcriptional repressor in the developing forebrain (Dattani et al., 1998). The other up-regulated interaction genes are: Fam167a (LFC=0.676), Pla2g7, (LFC=0.117), and Lrrn4 (LFC=.403).

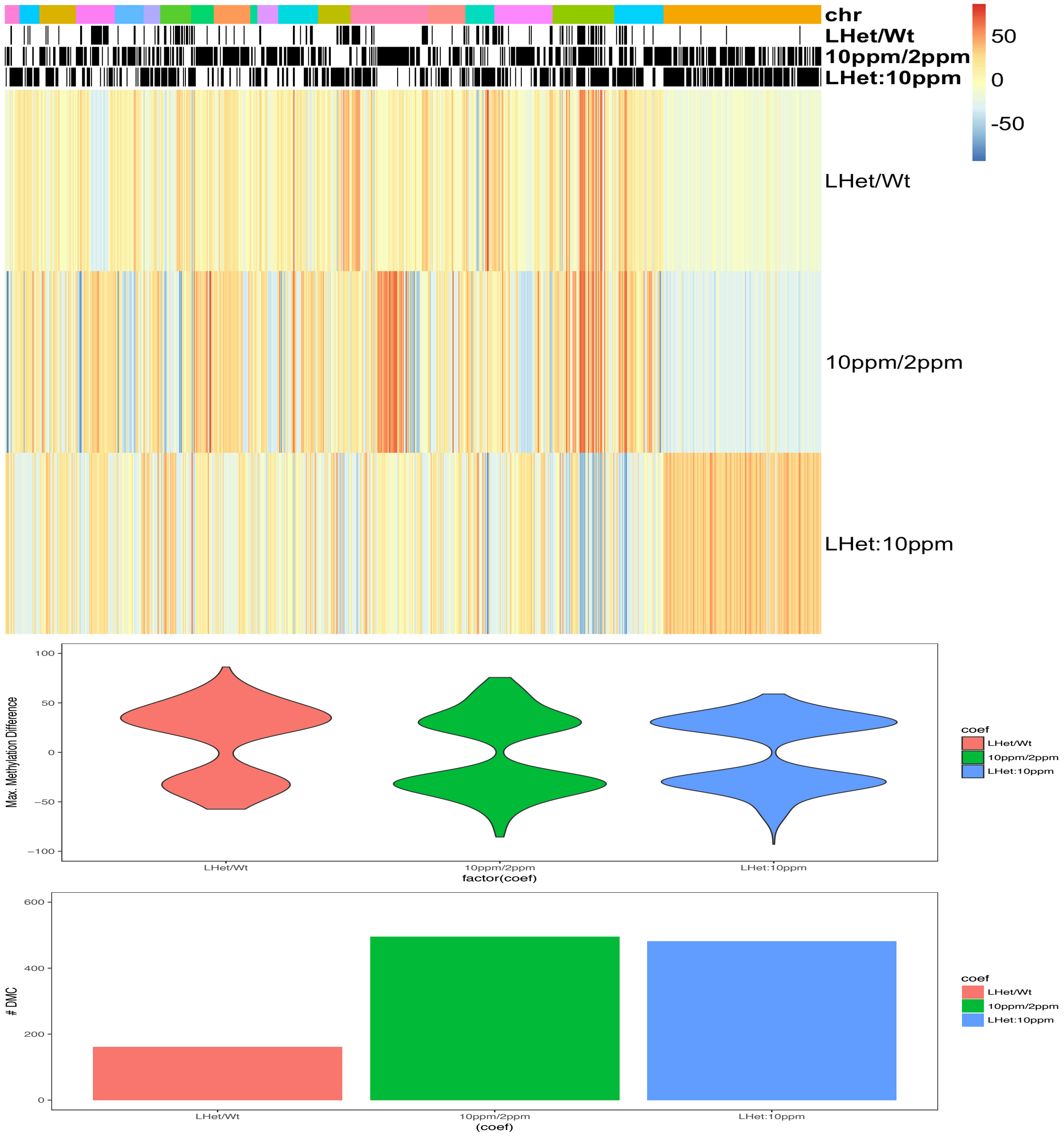


Figure 3.4 Differential Methylation in Lrp6 KO Background. Top Row: Heatmaps of Max. Meth Difference Estimates, with differential calls of loci in black on top. Middle Row: Violin plots showing distribution of Max. Meth Difference Estimates for each effect. Bottom Row: Number of DMCs associated with each effect

3.4 Analysis of FA Mechanisms in *Lrp6* KO

Results of genic annotation of sites assigned by potential mechanism are seen in Table 3.3. Rn45s, a pre-ribosomal protein was found to both be differentially methylated and expressed under the independent mechanism, and has been reported to be differentially methylated in the brains of mice perinatally exposed to lead. (Sánchez-Martín et al., 2015). Ubql2 has been found to be associated with amyotrophic lateral sclerosis and dementia (Deng et al., 2011).

By far the largest number of genes (703), were associated with differential expression under the independent mechanism. These are shown in Table 3.4. Enriched KEGG Pathways were Systemic lupus erythematosus (FDR=0.0009160), Metabolic pathways (FDR=0.0106), and Alcoholism (FDR=0.0106). The set was also enriched for histone domains H2A/H2B/H3/H4 (FDR=1.87e13)

Mechanism	Additive	Independent	Combinatorial
Differentially Methylated Genes (DMC in First Exon or >3 in Promoter, multiDiff: q<0.01, Max. Meth Diff>25)	Ubqln2 Rn45s	1700018B24Rik,AA414768,B630019K06Rik,Chst7,Gm8773,Gsc,Ppp1r3fos,Rn45s,Rnf113a1,Sowahd,Spin4,Ubl4a,Ubqln2*,Zfa-ps,Zic3,	1110012L19Rik,1700018B24Rik,AA414768,B630019K06Rik,Chst7,Crk,Cx3cl1,Gm16617,Gpr50,Mirlet7c2,Mospd4,Ndufb11,Ppp1r3fos,Ptchd1,Rai2,Rbm3os,Rn45s,Rnf113a1,Slitrk4,Sowahd,Spin4,Ubqln2,Zic3,
Differentially Expressed Genes (Deseq2: abs(lfc)>0.5, qvalue<0.05)	-	703 genes	-
Intersection	-	Rn45s	-

Table 3.3 Genes Associated with Mechanisms of FA Action in KO Background.

Differentially Expressed Genes Assoc. With Independent Mechanism of FA Action in *Lrp6* KO Mice
0610010K14Rik,1110001J03Rik,1110007C09Rik,1110008J03Rik,1110065P20Rik,1700037C18Rik,1810019J16Rik,2010001M06Rik,2010320M18Rik,2310003H01Rik,2310011J03Rik,2310045N01Rik,2310067B10Rik,2610305D13Rik,2810428I15Rik,4930432K21Rik,5730408K05Rik,8430429K09Rik,9130017N09Rik,9430076G02Rik,A230056P14Rik,A430005L14Rik,A530016L24Rik,AI413582,Aacs,Aamp,Aatk,Abcc10,Abhd16a,Abhd17a,Abhd8,Abtb1,Acaa1a,Acads,Acsf3,Acta1,Actn3,Actr5,Acy1,Adam23,Adc,Adck5,Adprhl2,Adssl1,Aes,Agap2,Ahdc1,Akr1e1,Akr7a5,Aldh16a1,Alkbh2,Anapc13,Ankrd23,Anks6,Ano8,Ap5z1,Apba2,Apba3,Apbb1,Apc2,Aprt,Arhgef10l,Arl10,Armc6,Arrdc1,Arvcf,Asb2,Asb6,Asic4,Atg2a,Atg4d,Atoh8,Atp13a2,Atp5g1,Atpbd4,Atxn7l2,Azi1,B3gat3,B930041F14Rik,BC005764,Bai2,Baiap2,Bcam,Bcat2,Bckdha,Bcl7c,Beta-s,Bola2,Brat1,Btbd2,Cacnb3,Cactin,Capn10,Car11,Car2,Card10,Caskin2,Casr,Cc2d1a,Ccdc106,Ccdc124,Ccdc22,Ccdc64b,Ccdc8,Ccdc85c,Ccdc88c,Ccnb1ip1,Ccrl1,Cd14,Cd4,Cdc42ep1,Cdk2ap2,Cdk5rap3,Cep170b,Chkb,Chpf,Chrd,Chtf18,Cited2,Ckb,Cldn4,Clec2l,Clgn,Clpp,Cntrob,Col18a1,Col2a1,Cope,Coro7,Cox6b2,Crip1,Crip2,Crocc,Csf2ra,Csnk1g2,Csrnp1,Ctu2,Cuedc2,Cul7,Cul9,Cx3cl1,Cyb561,Cyp26c1,D2Wsu81e,D330041H03Rik,Dak,Dalrd3,Dapk3,Dbp,Dcaf15,Dctn1,Dcxr,Dda1,Ddah2,Ddn,Des,Dhx34,Dhx38,Dnalc4,Doc2g,Dohh,Dos,Dpm3,Dpysl4,Drap1,Dtx2,Dus3l,Dvl2,E030030I06Rik,E130309D14Rik,E4f1,Ecsit,Edf1,Eef2,Efs,Egln2,Ell2,Emilin1,Eml2,Emp3,Engase,Enho,Eno1,Eno3,Ephb3,Ephb6,Ercc1,Erf,Esrra,Etfb,Evpl,Exoc3l4,Exosc5,Eya2,Fam129c,Fam131a,Fam173a,Fam181a,Fam195a,Fam195b,Fam69b,Fam83h,Fance,Fasn,Fastk,Fau,Fbf1,Fbxl18,Fbxl6,Fbxo31,Fbxw9,Fcho1,Fdxr,Fes,Fgd2,Fitm1,Fkbp8,Flot1,Flywch2,Foxi2,Frzb,Fsd1,Ftl1,Fzr1,Galk1,Gamt,Gatsl3,Gck,Gcn1l1,Ggt7,Gins2,Gipc1,Gjb3,Gli1,Gm13154,Gm13212,Gm16119,Gm1943,Gm4349,Gmppa,Gnaz,Gnb2,Gnl1,Gpaa1,Gpr137,Gpr162,Gpr179,Gpt,Grcc10,Grrp1,Gstp1,Gstt2,Gtpbp6,Gtse1,Gypa,H60b,Haghl,Hap1,Hapln3,Hbb-bh1,Hdac5,Hdgfrp2,Hecw2,Hes5,Hhipl1,Hip1r,Hist1h1c,Hist1h1d,Hist1h2ak,Hist1h2bc,Hist1h2be,Hist1h2bf,Hist1h2bg,Hist1h2bh,Hist1h2bj,Hist1h2bk,Hist1h2bl,Hist1h2bm,Hist1h2bn,Hist1h2bp,Hist1h3a,Hist1h3b,Hist1h3c,Hist1h3d,Hist1h3e,Hist1h3f,Hist1h3g,Hist1h3h,Hist1h3i,Hist1h4a,Hist1h4b,Hist1h4c,Hist1h4d,Hist1h4f,Hist1h4h,Hist1h4i,Hist1h4j,Hist1h4k,Hist1h4n,Hist2h2ac,Hist3h2a,Hist3h2ba,Hist4h4,Hmg20b,Hmgn5,Hmha1,Hook2,Hps1,Hps4,Hspb7,Hspbp1,Hyal2,Hyal3,Ier2,Igsf9,Igsf9b,Ikzf1,Ints1,Irs2,Irx1,Isg15,Isoc2a,Isyna1,Itga3,Jag2,Josd2,Jph2,Jrk,Kank3,Kazald1,Kcnj14,Kdm4b,Kel,Kif2a,Kifc5b,Klhl17,Klhl36,Klk8,Krt19,Lamb2,Lars2,Lemd2,Leprel2,Lipe,Llgl1,Llgl2,Lmna,Lmnb2,Lmtk3,Lpcat2,Lrfn4,*Lrp6*,Lrrc16b,Lrrc29,Lrrc45,Lrrc4b,Lrrc56,Lzts2,Man2c1,Map1lc3a,Map2k2,Map3k10,Map3k11,Map3k14,Map4k2,Mapk13,Mapk8ip1,March9,Mbd3,Mblac1,Mcf2l,Mcrs1,Mdfi,Mdk,Med16,Meis3,Metrn,Mettl21d,Mettl22,Mfap2,Mgmt,Mib2,Mier2,Miip,Mipol1,Mir5109,Mlycd,Mocs1,Mpnd,Mpv17l2,Mroh1,Mrpl12,Msto1,Mus81,Mvd,Mvk,Mybpc3,Myh14,Myh7,Myh7b,Myl1,Myl3,Myl4,Myl6,Myl7,Myo18b,Myo7a,N4bp3,Naa10,Nacad,Nacc2,Naglu,Nat6,Nat8l,Nat9,Ncdn,Ncln,Ndufa11,Ndufb7,Ndufs7,Ndufs8,Neurl2,Nfatc4,Nkiras2,Nkx6-2,Nle1,Nol12,Nosip,Nphp4,Nppa,Nppb,Nr1h2,Nrip3,Nt5c,Nthl1,Nudc,Nudt14,Nudt22,Nudt8,Nxpe2,Nxph3,Ogdhl,Oxsm,P4htm,Pacsin3,Pafah1b3,Palm,Pcbp4,Pcif1,Pdgfa,Pex14,Pex16,Pex6,Pfkl,Pgls,Phb,Phf15,Phldb1,Pick1,Pih1d1,Pik3cd,Pkdcc,Pkhd1l1,Pkmyt1,Pkn1,Pkn3,Pla2g6,Plcd1,Pld3,Plec,Plekhm2,Plekho1,Pmm1,Pnkp,Pofut2,Pold2,Poll,Polr2e,Polr2f,Polr2i,Pop7,Ppan,Ppdpf,Ppp1r14b,Ppp1r16a,Ppp1r37,Prdx5,Prelp,Prex1,Prrg2,Prrx2,Psmb10,Psmd3,Psmg3,Ptov1,Ptprs,Pus1,Qtrt1,Rabac1,Rabep2,Rac3,Rai1,Ralgds,Rapgef3,Rara,Rarres1,Rcn3,Rcor2,Rexo1,Rfx1,Rhag,Rhbdd3,Rhbdf1,Rhpn1,Rims3,Ring1,Rmrp,Rn45s,Rnaseh2c,Rnf126,Rnf220,Rnf31,Rnu12,Rpl11,Rpl27,Rpl28,Rpl34,Rpl35,Rpl36,Rpl38,Rpl8,Rpph1,Rprl3,Rps15,Rps15a-ps4,Rps15a-ps6,Rps16,Rps19,Rps26,Rps3a1,Rps7,Rps9,Rpusd1,Rrp9,Rtbdn,Rtkn,Ruvbl2,Rxrb,Saa1,Sars2,Sbf1,Sbno2,Scarf1,Scarf2,Scoc,Scrib,Scrn2,Sema5b,Setd4,Sgsm2,Sh3gl1,Sh3glb2,Shank1,Shd,Sirt6,Sirt7,Skiv2l,Slc12a9,Slc25a10,Slc25a22,Slc25a38,Slc25a42,Slc26a10,Slc2a8,Slc30a10,Slc35c2,Slc35e4,Slc39a3,Slc39a8,Slc4a3,Slc9a3r2,Snora17,Snora78,Sox3,Sphk2,Spint2,Sppl2b,Spry1,Spsb2,Spta1,Src,Srf,Ssbp4,Ssh3,Sssca1,Stk19,Stoml1,Sugp1,Svopl,Syne4,Sypl,Syt3,Tab1,Tada3,Taf1c,Taldo1,Tarbp2,Tbc1d10a,Tbx1,Tcf3,Telo2,Tesc,Tesk1,Tff3,Tfip11,Thap7,Thop1,Timm13,Tjp3,Tk1,Tmem121,Tmem132a,Tmem134,Tmem143,Tmem160,Tmem205,Tmem219,Tmem56,Tmtc3,Tnks1bp1,Tnni1,Tnni3,Tomm40,Tor2a,Tpgs1,Tppp3,Tpt1,Trabd,Traf4,Trappc9,Trim46,Triobp,Trmt61a,Trp53i11,Trp53i13,Trpv2,Tsen34,Tsen54,Tssc4,Tssk6,Twf2,Txnip,U2af1l4,Ubb,Ubtd1,Ulk4,Unc45a,Unk,Upk3bl,Uqcr10,Uqcr11,Usp19,Vill,Vps18,Wbp1,Wdr18,Wdr24,Wdr25,Wdr34,Wdr6,Wdr65,Wdr86,Wdr90,Wdtc1,*Wnt*4,Wtip,Xab2,Xk,Ydjc,Zbtb12,Zbtb17,Zbtb48,Zdhhc1,Zfand2b,Zfp13,Zfp219,Zfp414,Zfp459,Zfp523,Zfp574,Zfp651,Zfp653,Zfp668,Zfp688,Zfp707,Zfp710,Zfp97,Zfpl1,Zfyve28,Znhit2,Zscan10,Zscan25,Zswim4,

Table 3.4 Differentially Expressed Genes Associated with Independent Mechanism of FA Action in *Lrp6* KO Background. Independent genes were defined as those with q<0.05, abs(LFC)>0.5 under the 10ppm/2ppm contrast.

3.5. Persistent Biomarkers of FA and *Lrp6* Mutation Status

As previously discussed, “neural tube defects” is an umbrella term that groups many distinct syndromes under a single heading, ranging from inevitably lethal conditions such as exencephaly, to others that may have no noticeable impact on health and can easily go undiagnosed, such as in spina bifida occuluta, which as the name implies is often hidden. Nevertheless, if biomarkers can be found that correlate with NTD status, both parents and caregivers can be better prepared for the full range of outcomes, and be given the opportunity to intervene at an earlier stage. Proteomic analysis has been able to identify ADP-ribosylation factor 1, a protein similar to cold agglutinin FS-1 antibody light-chain, vitamin K3 protein and another unknown protein as biomarkers of NTD status in expecting mothers, with a 90

Ongoing advances in biotechnology are making it possible to extract and sequence fetal genetic material from circulating blood (Kitzman et al., 2012). Utilization of handheld DNA sequencers, such as the MinION from Oxford Nanopore (Jain et al., 2016), which are already capable of detecting base modifications, may soon allow for the assaying of fetal methylation levels on an ongoing basis. The combination of biomarkers with such a monitoring system could be a powerful tool, and the ability for such assessments to be done non-invasively could lower costs and improve quality of care.

3.6. P2 eRRBS Dataset

As shown in Table 3.1, whole blood from P2 mice on both mutant backgrounds was collected, with mothers fed control (2ppm) or elevated (10ppm) amounts of FA. Extraction and data analysis was performed as previously described. Samples were collected and sequenced in three groups. Lack of batch of effects was confirmed using visual inspection of clustering and PCA analysis.

3.7 Filtering of CpG's Associated with Tissue/Developmental Differences

It is well known that DNA methylation has tissue–specific patterning (Chen et al., 2016; Lehmann-Werman et al., 2016; Lokk et al., 2014). In addition, it has been shown that methylation data can be used to accurately estimate the age of a sample- though it can also be used to estimate a separate quantity, "epigenetic age" that may match or diverge from chronological age. Such differences complicate direct comparison between the methylation data in the embryonic and P2 samples.

To control for these differences, cross-timepoint cohorts were constructed for each background, with all *Lrp6* mutants removed. These control cohorts used only wildtype mice on the control diet, while controlling for gender and gender-specific tissue differences by running multiDiff with the following model:

Tissue (P2/E9.5) + Gender (Male/Female) + Tissue:Gender

After filtering both sites associated with primary and gender specific tissue differences, 87% (546,389 sites) of embryonic sites were retained on the Cd background, while 90% (559,937) of sites were retained on the KO background

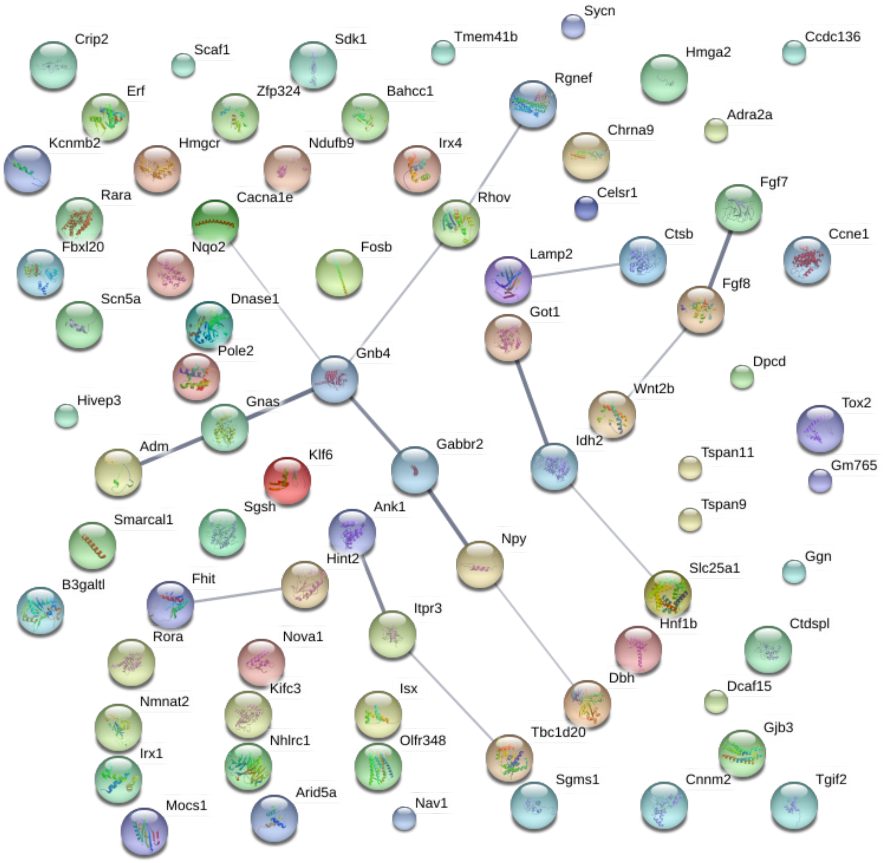
3.8 Analysis of Persistent Methylation Across Time Points

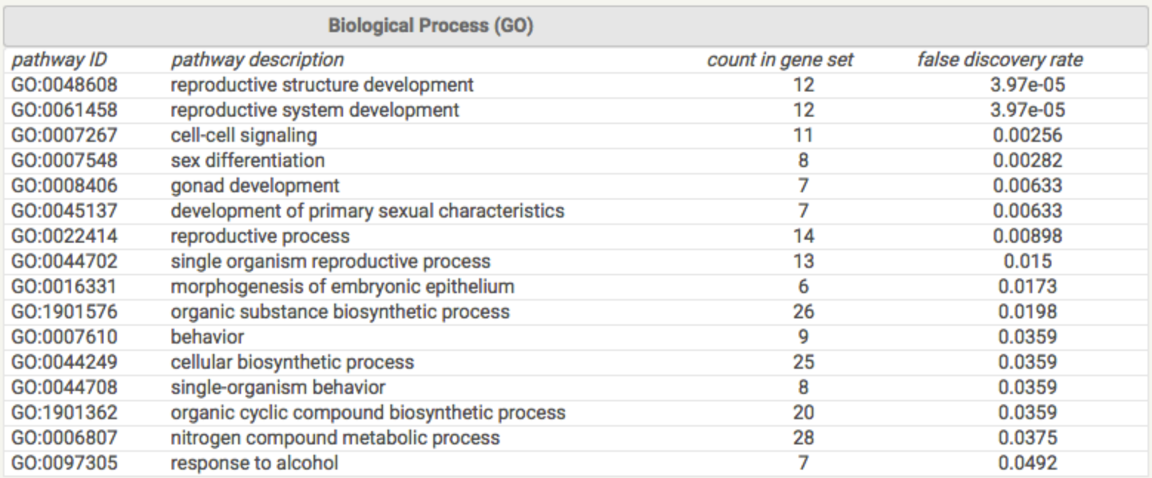
After following the filtering procedure previously described, the remaining differential sites at each timepoint were intersected. The results can be seen in Table 3.4. Larger numbers of sites were found to be associated with persistent effects of FA diet than either mutation. After annotation to the nearest gene, no genes with more than a single persistent DMC were found in the Cd background for either effect. In the KO background, Rn45s and Fktn were persistently associated with FA. The *Fktn* gene encodes a type II transmembrane protein that is targeted to the Golgi apparatus through an N-terminal signal anchor(Esapa et al., 2002), and has been linked to lissencephaly(Deak et al., 2008; Puckett et al., 2009), a condition where parts of the brain appear smooth.

Rn45s was previously discussed as the sole gene showing differential expression and differential promoter methylation associated with the independent mechanism of FA action. Neither Fktn nor Rn45s show significant changes in expression related to diet (q-values:0.78, 0.40, LFC: 0.208254, -0.0987), or diet-genotype interactions at E9.5.

3.9 Pathway Analysis of Persistently Methylated Genes

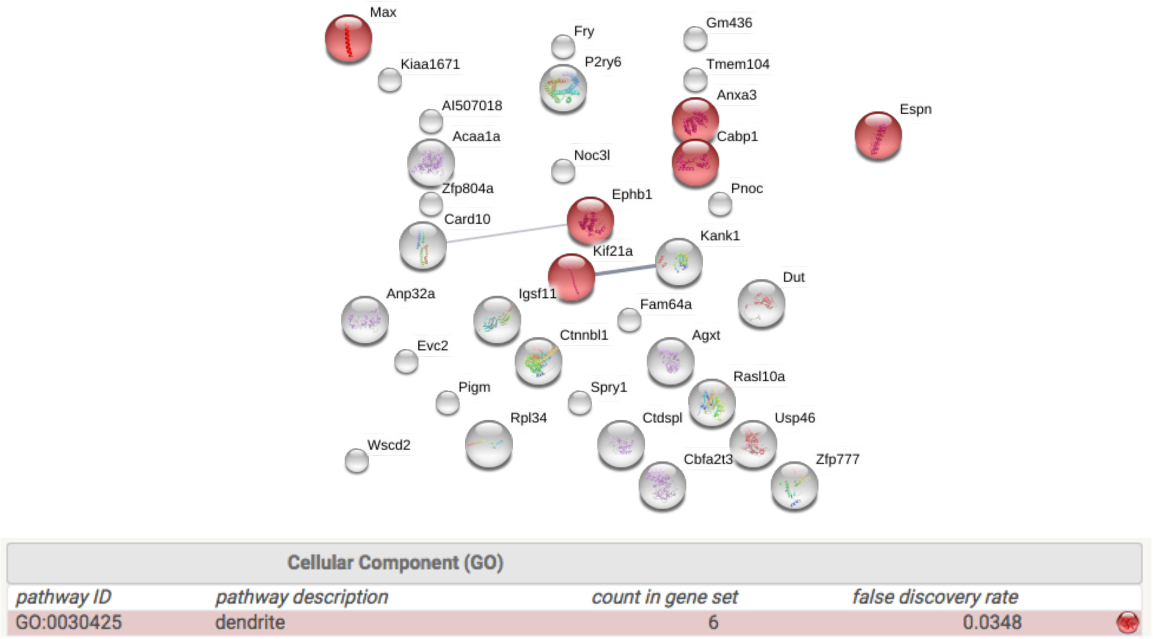
In the Cd background, persistent methylation associated with dietary FA was significantly enriched for reproductively and sex associated developmental pathways (Figure 3.5). We speculate this may be a partial artifact of the gender imbalances in the cohort, or potentially be linked to the known elevation of NTDs in female mice. On the other hand, the genes associated with the Cd mutation are only enriched for dendritic function. (fdr=0.03), This is quite intriguing, and may provide basis for hope of a novel biomarker if it can be found to match human data. The associated genes with dendritic function are Max, Espn, Ephb1, Anxa3, Cabp1, Kif21a.





Biological Process (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0048608	reproductive structure development	12	3.97e-05
GO:0061458	reproductive system development	12	3.97e-05
GO:0007267	cell-cell signaling	11	0.00256
GO:0007548	sex differentiation	8	0.00282
GO:0008406	gonad development	7	0.00633
GO:0045137	development of primary sexual characteristics	7	0.00633
GO:0022414	reproductive process	14	0.00898
GO:0044702	single organism reproductive process	13	0.015
GO:0016331	morphogenesis of embryonic epithelium	6	0.0173
GO:1901576	organic substance biosynthetic process	26	0.0198
GO:0007610	behavior	9	0.0359
GO:0044249	cellular biosynthetic process	25	0.0359
GO:0044708	single-organism behavior	8	0.0359
GO:1901362	organic cyclic compound biosynthetic process	20	0.0359
GO:0006807	nitrogen compound metabolic process	28	0.0375
GO:0097305	response to alcohol	7	0.0492

Figure 3.5 Pathway Analysis of Genes with Persistent DMCs Assoc. with FA. On Cd Background. Persistent genes were defined by annotating the intersection of differential sites associated with FA at E9.5 and P2. Visualization and enrichment analysis performed using STRING.



Cellular Component (GO)			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0030425	dendrite	6	0.0348 ●

Figure 3.6 Pathway Analysis of Genes with Persistent DMCs Assoc. with Cd Mutation On Cd Background. Persistent genes were defined by annotating the intersection of differential sites associated with Cd mutation at E9.5 and P2. Visualization and enrichment analysis performed using STRING.

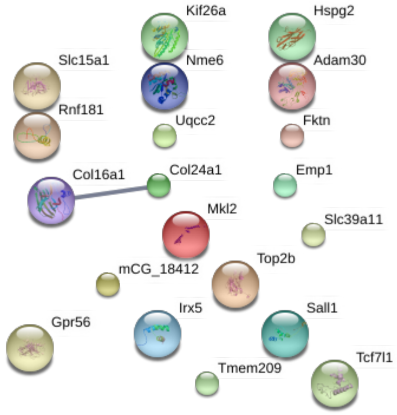
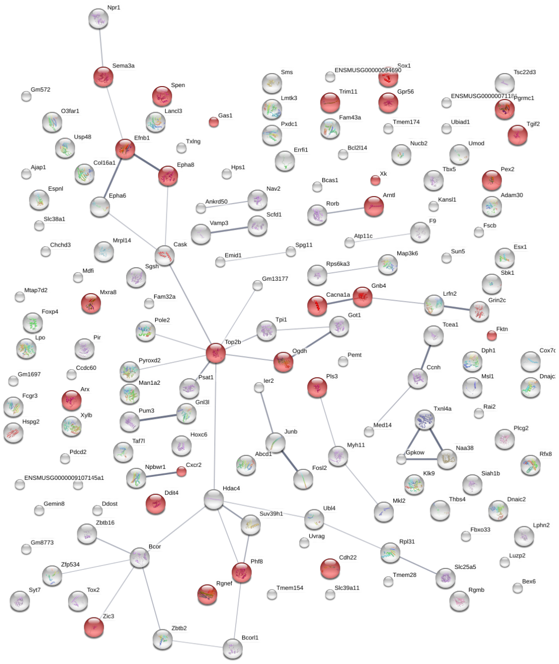
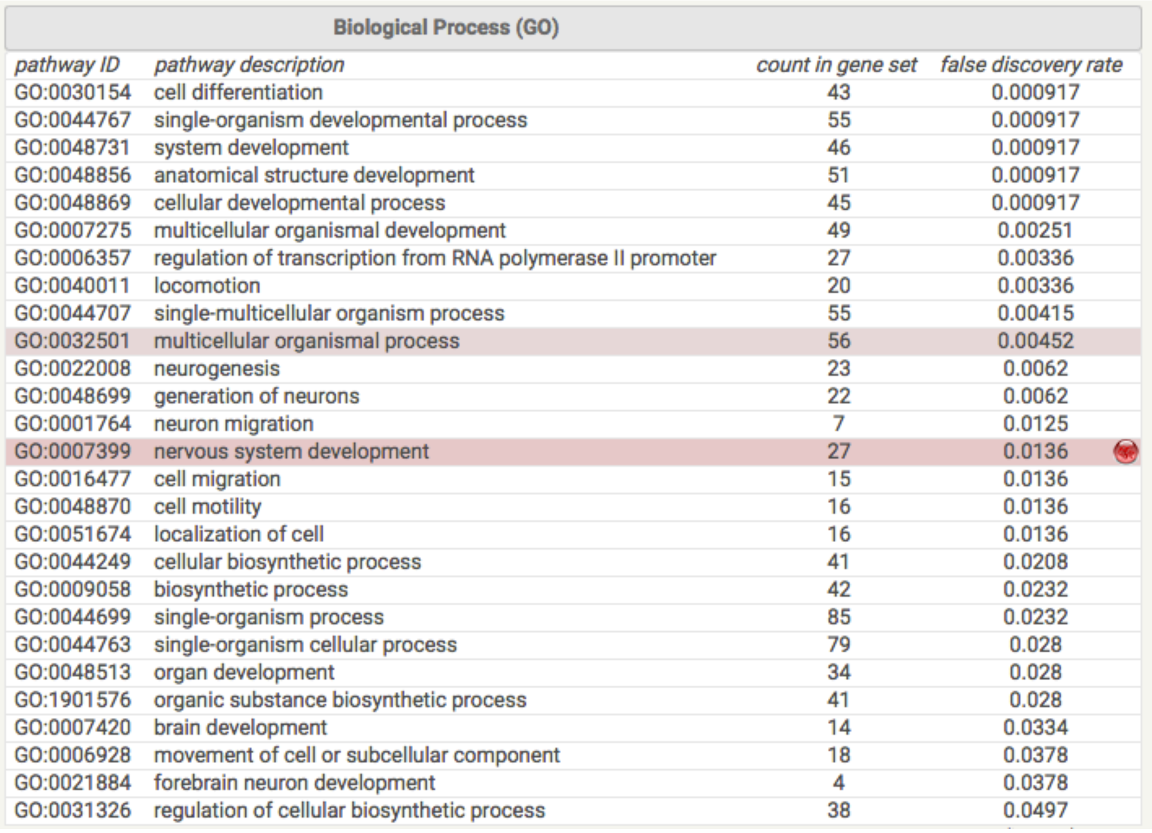


Figure 3.7 Pathway Analysis of Genes with Persistent DMCs Assoc. with Lrp6 KO. Persistent genes were defined by annotating the intersection of differential sites associated with Lrp6 KO mutation (+/-) at E9.5 and P2. Visualization and enrichment analysis performed using STRING.





Biological Process (GO)

pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
GO:0030154	cell differentiation	43	0.000917
GO:0044767	single-organism developmental process	55	0.000917
GO:0048731	system development	46	0.000917
GO:0048856	anatomical structure development	51	0.000917
GO:0048869	cellular developmental process	45	0.000917
GO:0007275	multicellular organismal development	49	0.00251
GO:0006357	regulation of transcription from RNA polymerase II promoter	27	0.00336
GO:0040011	locomotion	20	0.00336
GO:0044707	single-multicellular organism process	55	0.00415
GO:0032501	multicellular organismal process	56	0.00452
GO:0022008	neurogenesis	23	0.0062
GO:0048699	generation of neurons	22	0.0062
GO:0001764	neuron migration	7	0.0125
GO:0007399	nervous system development	27	0.0136
GO:0016477	cell migration	15	0.0136
GO:0048870	cell motility	16	0.0136
GO:0051674	localization of cell	16	0.0136
GO:0044249	cellular biosynthetic process	41	0.0208
GO:0009058	biosynthetic process	42	0.0232
GO:0044699	single-organism process	85	0.0232
GO:0044763	single-organism cellular process	79	0.028
GO:0048513	organ development	34	0.028
GO:1901576	organic substance biosynthetic process	41	0.028
GO:0007420	brain development	14	0.0334
GO:0006928	movement of cell or subcellular component	18	0.0378
GO:0021884	forebrain neuron development	4	0.0378
GO:0031326	regulation of cellular biosynthetic process	38	0.0497

Figure 3.8 Pathway Analysis of Genes with Persistent DMCs Assoc. with FA. On KO Background Persistent genes were defined by annotating the intersection of differential sites associated with FA mutation at E9.5 and P2. Visualization and enrichment analysis performed using STRING.

3.10 Discussion

Most exciting was the identification of Folh1 as upregulated by a combinatorial mechanism, along with the associations with nervous system development and dendrite development noted with FA on the KO background and the Cd mutation on the crooked tail background. Intriguingly, they're both found interventions on two different backgrounds that cause negative consequences. This may indicate that by causing excess demethylation in neuronal development genes, normal function can best not be distrupted. Further oxidatizes states of methylation, in particular 5faC, can be used to check active demethylation.

The integrative analysis of methylation and expression provides little evidence for the additive model of FA action in embryonic mice, in which either the Crooked tail or null mutation generated epigenetic lesions that are then acted upon by FA. Rather, FA most probably acts in either an independent or combinatorial manner, with more differentially expressed genes supporting the independent model in the KO background, and the combinatorial model in the Cd background. The lack of additive genes is driven not only by the low numbers of sites found linked to the mutation, but also the genes associated with either mutation.

Of the observed genes linked to *Wnt* signaling, more were observed to be associated with RhoA compared to beta-catenin, pointing towards a greater role for the non-canonical *Wnt* PCP pathway.

Rn45s warrants further study in the null background, as it was the only gene that showed differential methylation and expression correlated with FA on the null background, with the differential methylation showing persistence over developmental time.

Cd Background	10ppm/2ppm	*Lrp6*$^{Cd/+}$/Wt
# Persistent DMCs	86	43
Genes with DMCs	Rara,Adra2a,Sdk1,Sgsh,Dcaf15,Tspan9,Gnb4,Cacna1e,Hmgcr,Zfp324,Bahcc1,Ggn,Ccne1,Tunar,Got1,Nmnat2,Hint2,*Wnt*2b,Sgms1,Nav1,Klf6,Smarcal1,Itpr3,Nova1,Rora,Sycn,Scn5a,Mocs1,Tspan11,Kcnmb2,Pole2,Olfr348,Dbh,Hmga2,Adm,Fgf7,Irx1,B3glct,Dnase1,Ccdc136,Arhgef28,Gm765,Erf,Fendrr,Hnf1b,Gabbr2,Gm11468,Isx,Hivep3,Ctdspl,Gm14207,Tox2,Npy,Arid5a,Cnnm2,Celsr1,Gm38426,Fgf8,Chrna9,Slc25a1,Irx4,Tmem41b,Ctsb,Kifc3,Mir701,Gjb3,Rhov,Nqo2,Gnas,Fhit,Crip2,Ndufb9,Nhlrc1,Tbc1d20,Lamp2,Fosb,Idh2,Scaf1,Tgif2,Dpcd,Fbxl20,4933411E08Rik,Ank1,	Pigm,P2ry6,Lnpk1,Fry,Zfp804a,Pnoc,1700095B10Rik,1700052K11Rik,Cbfa2t3,Spry1,Espn,Card10,Noc3l,Kank1,Anxa3,Gm436,Zfp777,Anp32a,1700013G24Rik,Cabp1,Rpl34-ps1,Ctdspl,Dut,Wscd2,Max,Acaa1a,Kif21a,Usp46,2900026A02Rik,Ctnnbl1,Fam64a,Igsf11,Ephb1,Agxt,2610028E06Rik,Tmem104,Evc2,4930465M20Rik,Rasl10a,Gm14204,3200001D21Rik,1700092K14Rik,1700123O21Rik,
Genes with ≥2 DMCs	-	-

KO Background	10ppm/2ppm	*Lrp6*$^{+/-}$/Wt
# Persistent DMCs	208	25
Genes with Persistent DMCs	Smad6,Casz1,Slc6a1,Mir290a,Tnfrsf11a,Foxn3,Gm5069,Lnpk1,Atg5,Phldb1,Clcn2,Fry,Gpbp1,Gm53,Hint2,Zswim7,Fasn,Sgms1,Nav1,Nkx6-3,Gapdh,Slc22a19,Fanci,Oxsm,Hdc,Nova1,Mir499,Uck2,Ggnbp1,S1pr4,Gm436,Mpnd,Zap70,Ucp1,Lta4h,Txndc9,Rbks,Ago2,Bhmt2,Adrm1,Dnase1,Nudt6,Msl3,Zfp444,Gm6602,Fendrr,Rpl34-ps1,Kbtbd11,Hnf1b,Pdcd4,Gm11468,Ctdspl,Npas3,Vps53,Dut,Wscd2,Trex2,Acaa1a,Selk,Fam69b,Nckap5,Fgf8,Alad,Anln,Tyk2,Ulk4,Stk25,Sec63,Igsf11,Phb,Rtkn,Hapln1,Ephb1,Ndufb9,Nhlrc1,Noct,Dgkh,Idh2,Bdnf,4930465M20Rik,Fbxl20,Six3os1,Tor2a,1700092K14Rik,Ank1,Ifnlr1,	Tmem209,Hspg2,Col16a1,Rn45s,Top2b,Sall1,Nme6,4930415O20Rik,Irx5,Fktn,Emp1,Mkl2,Adam30,Tcf7l1,Adgrg1,Slc15a1,Slc39a11,Uqcc2,Col24a1,Kif26a,Rnf181,
Genes with ≥2 DMCs	Rn45s,Fktn	Ddost,Hspg2,Tpi1,Cox7c,Rn45s,Phf8,Ubl4a,Suv39h1,Pole2,Sbk1,Arhgef28,Fktn,Rps6ka3,Lmtk3,Gnl3l,Nucb2,Rai2,1700030C10Rik,Bex6

Table 3.4 Persistent Differential Methylation Associated with *Lrp6* Mutations and FA. Persistent DMCs were defined by he intersection of differential sites associated with either FA diet or Lrp6 mutation status at E9.5 and P2. Genes were defined by annotation of persistent DMC to the closest gene.

3.11 Materials and Methods

Animals

All procedures involving animals were carried out in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee at Weill Cornell Medicine. Mice were housed in climate-controlled Thoren units with a 12 h light–dark cycle.

Two strains of mice were used:

Lrp6 K/O:

Gene-trap mice in which the Lrp6 locus was inactivated have been backcrossed more than 12 generations to the C3H/HeJ background.

Lrp6Cd:

Crooked tail (*Cd*) mice bear a gain-of-function naturally occurring mutation in Lrp6, a co-receptor for canonical *WNT* signaling.

Colony maintenance, embryo harvest and whole blood collection:

Mating pairs of *Lrp6*$^{+/-}$ and *Lrp6*$^{Cd+/-}$ mice were maintained on a defined diet containing 2ppm or 10ppm FA (Research Diets Inc., New Brunswick, NJ, USA) for two generations prior to tissue or embryo collection. Embryos from timed pregnant females were harvested at E9.5 and scored based on the somites count. For *Lrp6*$^{+/-}$ 18-21 somites and for *Lrp6*$^{Cd/+}$ 18-20 somites, gDNA and RNA was extracted from E9.5 whole embryos simultaneously, (AllPrep DNA/RNA

Micro Kit Cat.No 80284 Qiagen). Whole blood from P2 pups was collected by cardiac puncture. gDNA from whole blood was extracted , (QIAamp DNA Blood Mini Kit Cat. No. 51104 Qiagen)

Genotyping and sexing

Genotyping was done by PCR. Sex-determination of the animals was done by PCR using specific primers for Y chromosome. RNA-seq was used to confirm presence of mutation in Cd embryos.

Modeling Interactions

The following models were applied to both the ERRBS and RNA data, to calculate differential methylation and differential expression respectively:

Diet (10ppm/2ppm) + Genotype (Het/Wt)+ Sex (M/F)

+ Diet:Sex + Genotype:Sex + Diet:Sex

The first three terms capture primary effects (with the relevant contrast noted in parenthesis), while the second three capture potential interactions between them. Next the analyses were performed with gender non-corrected, using the following model:

Diet (10ppm/2ppm) + Genotype (Het/Wt)

+ Diet:Genotype (Het:10ppm)

In the final analysis sites and genes associated with sex, sex-diet, or sex-genotype were removed.

RNA-seq

Samples were aligned and count matrices generated using the Mason lab's in house processing pipeline, r-make. Outlier samples were detected and removed in exploratory analysis using PCA and correlation analysis. Differential expression analysis was done using DESeq2, with each background being analyzed separately.

Gender Control

An effort was made to harvest both male and female samples. This served two purposes. First, by examining the output of Male/Female contrasts, we were able to validate multiDiff ability to detect expected effects. Second, it allowed us to check for interactions of mutation status and maternal diet with gender. For final maps of differential methylation, we removed sites that were affected by sex, sex-genotype, or sex-diet interactions.

CHAPTER 4

Challenges in Genetic and Epigenetic Analyses Related to Human NTDs

4.1 Introduction

The study of model systems functions first as an area of gaining basic biological understanding, and ultimately of providing actionable information for human decisions-making, in a medical, research, economic, or ecological context. In our examination of the biology of folic acid (FA) in Lrp6 mouse mutants, our goal is ultimately to gain such insight into the nature of FA and NTD biology in humans. Ongoing national efforts to register NTD affected families and recruit them for study has resulted in an expanding pool of genetic and epigenetic data for researchers to analyze.

As discussed above, the study of NTD-FA interactions within a single gene, within two of the over 400 mouse models (Harris and Juriloff, 2007, 2010), must control for factors such as sex, genetic background, and interactions. The complex nature of human NTDs makes it all the more vital that care is taken to assess and control for confounders when performing study design, analysis, and interpretation.. Below we represent examples that illustrate these challenges with respect to sex, population ancestry, and genome interpretation.

4.2 Sex Differences in NTD Incidence and DNA Methylation

Human and mouse females are known to be at higher risk for NTDs, though it is speculated that it may reflect 'epigenetic drag' from the X chromosome in which the recruitment of methyl groups for the maintenance of imprinting results in hypomethylation in other areas of the genome (Juriloff and Harris, 2012). In mice, the female-to-male ratio amongst exencephalic mice models where known is approximately 2:1, including in the *Cd* strain.

Although we currently we do not have sufficient data to confidently call sites associated with gender differences, we did use the result of such calls to filter our analysis. As can be seen in the figure below, both mutations show the highest number of sites being affected by Gender:Genotype interactions., with the largest number of sites being observed with CdHet:Male. Removing the sex associated sites from the analyses presented in the previous chapter, although necessary, may have removed some of the most dynamic sites. When unsure whether biological sex is a relevant factor in an ongoing study, one approach is to investigate what occurs when sex and it's interactions are included as covariates.

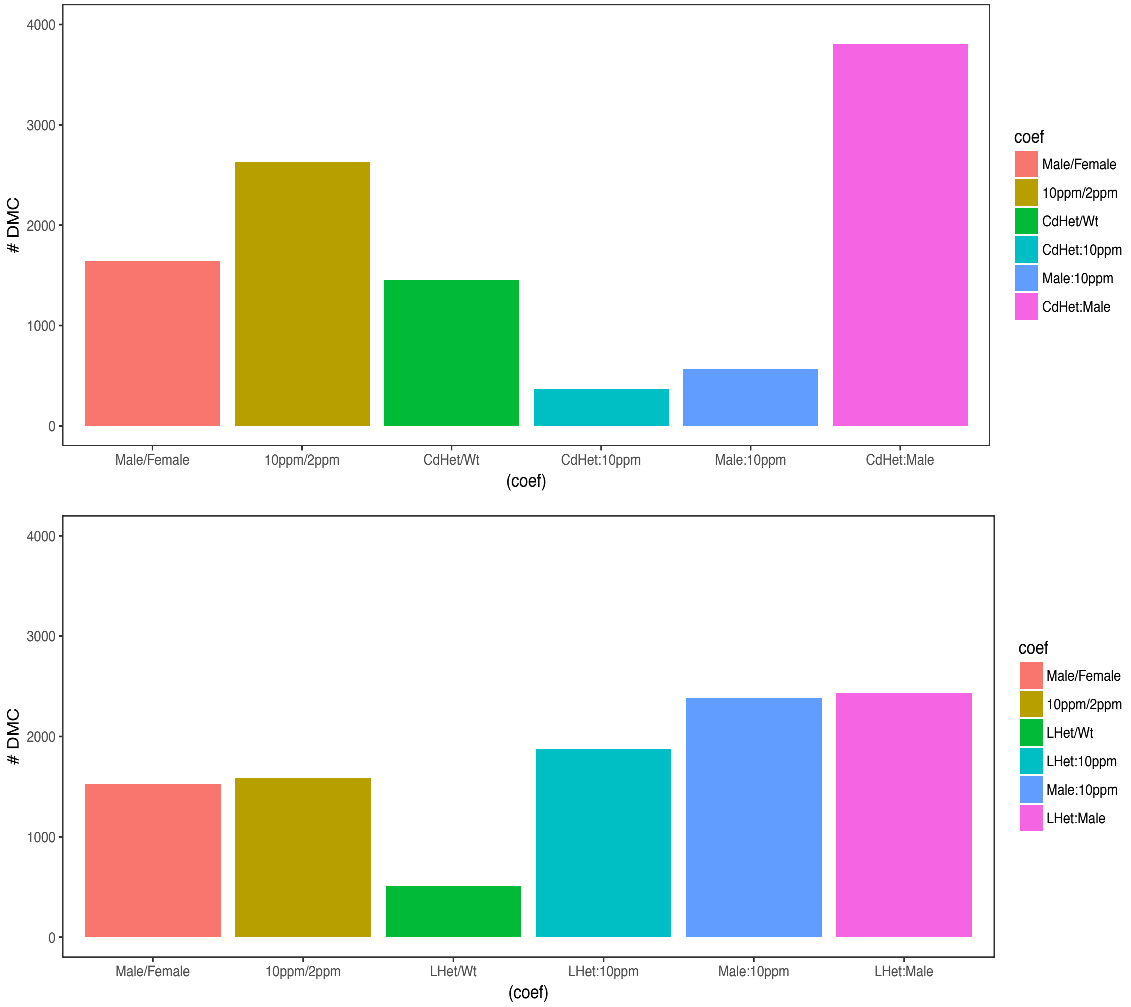


Figure 4.1 DMCs Called for Use in Sex Filtering. In order to remove sites associated with gender, DMCs were called on both Lrp6 mutant backgrounds, including sex (Male/Female), and sex-diet (Male:10ppm) and sex-mutation (LHet:Male, CdHet:Male) interactions (q-value<0.01, MMD$\geq$25). In both backgrounds, these terms were associated with amongst the highest numbers of DMCs.

4.3 Importance of Genetic Ancestry in Methylation and Whole Genome Study of Human Data

Although the primary topic here has been on methylation, the study of whole genomes sequence (WGS) data in NTDs is starting to bear preliminary results (Lei et al., 2015). A difficult problem (discussed below) has been the importance of controlling for ancestry, as variants may display different frequencies within different subpopulations, altering the results of analyses (Polimanti et al., 2015). Similarly, human methylation data has been discovered to mirror ancestry information (Rahmani et al., 2017; Sánchez-Martín et al., 2015). This has the potential to complicate the data collection and interpretation of methylation studies in human NTDs, which are starting to be released (Price et al., 2016; Rochtus et al., 2015; Stolk et al., 2013) . Below is a section of currently unpublished work, reporting a portion of the results of a WGS analyses in Spina Bifida patients, that illustrates the challenges of ancestry within that context..

The sample cohort encompassed 283 human subjects including 125 NTD cases and 158 healthy unrelated controls, derived from Caucasian, US, and Middle Eastern backgrounds. Additional control genomes derive from publically available datasets from the 1000 Genomes Project (containing genomes from over 2,500 individual), the NHBLI Exome Project (containing over 6,500 exomes) and the Exome Aggregation Consortium (ExAC) (Karczewski et al., 2017), (containing over 60,500 exomes). Stringent criteria were used to find polymorphisms that are likely to alter gene function or regulation, looking for SB case-enrichment of these changes in coding regions and intergenic regions that are associated with enhancers and gene regulatory sites. We identified novel genes and gene families that are likely to contribute toward NTD risk.

GPR161 encodes a ciliary G-protein coupled receptor recently shown to be a key regulator of sonic hedgehog signaling, and specifically promotes the processing of the GLI3 ligand to Gli3 repressor (Gli3R) (Mukhopadhyay et al., 2013). Mouse embryos that are homozygous knockouts of *Gpr161* die in mid gestation and display extensive NTDs. We found an insertion in GPR161 that caused a frame shift in the transcript (Q-value = 4 x 10^{-4} using a permutation test, and 3 x 10^{-3} using SKAT). Confirmed using Sanger sequencing, this mutation was identified in seven SB patients from the US sample collection, all of which were heterozygous for the detected insertion.

The discovery of the same mutation in seven cases, and its absence in the controls, was highly unusual and prompted us to examine whether these individuals might be related despite having been randomly collected. o determine if there was any direct or cryptic relatedness among these individuals and the other members of the patient cohort, we used several tools including BEAGLE (IBD) coupled with PRIMUS (relatedness). BEAGLE (Browning and Browning, 2011) identifies regions of the region that are identical by descent between pairs on individuals using a HMM based-method, while PRIMUS (Staples et al., 2014)identifies individuals who are up to third-degree relatives by evaluating possible pedigrees.

We checked the imputed haplotypes in the GPR161 locus using GATK to re-call the genomes both locally around *GPR161*, and globally at sites in the 1000 genomes project. The local variants were converted using PLINK (version 1.90b3w) to PLINK format, and Haploviewer (version 4.1) was used to calculate and visualize haplotype blocks.

Analysis of all SNPs across the genome using PLINK confirmed that these seven individuals were not first order relatives, since they all showed p_HAT identity by state (IBS) scores in the expected range for random individuals[44], different haplotypes, and were estimated to not be close relatives. Similarly, the local analysis using PLINK identified 314 SNPs in a 60KB window around *GPR161* in the seven cases, and they all showed distinct haplotypes for each individual. We next examined the distribution of IBD lengths in the region surrounding *GPR161*, comparing the mean values for the cases with those of the single Puerto Rican (PUR) trio available in the 1K data set, two quartets from NTD-affected families with NTD, and PUR plus three additional populations from the 1K database (**Figure 4.2.1.**). Although cryptic relatedness cannot be completely ruled out, they are no closer than fourth degree relatives (based on BEAGLE plus PRIMUS results). Thus, in the aggregate, a founder effect is suggested for this population subset.

GPR161 intolerance of variants in healthy individuals supports its pathogenicity in NTD. Our genomic data are the strongest clinical genetic connection yet between NTD and this G-protein coupled receptor, whose localization to the primary cilium is regulated by TULP3-ITF-A, where it functions in a positive feedback network to activate the SHH pathway in a cAMP-dependent manner (Hwang and Mukhopadhyay, 2015). Recently, retinoic acid signaling and the canonical *Wnt* pathways were identified as additional downstream targets of GPR161, through transcription factor Cdx1, in the genesis of NTDs in the mouse (Li et al., 2015).

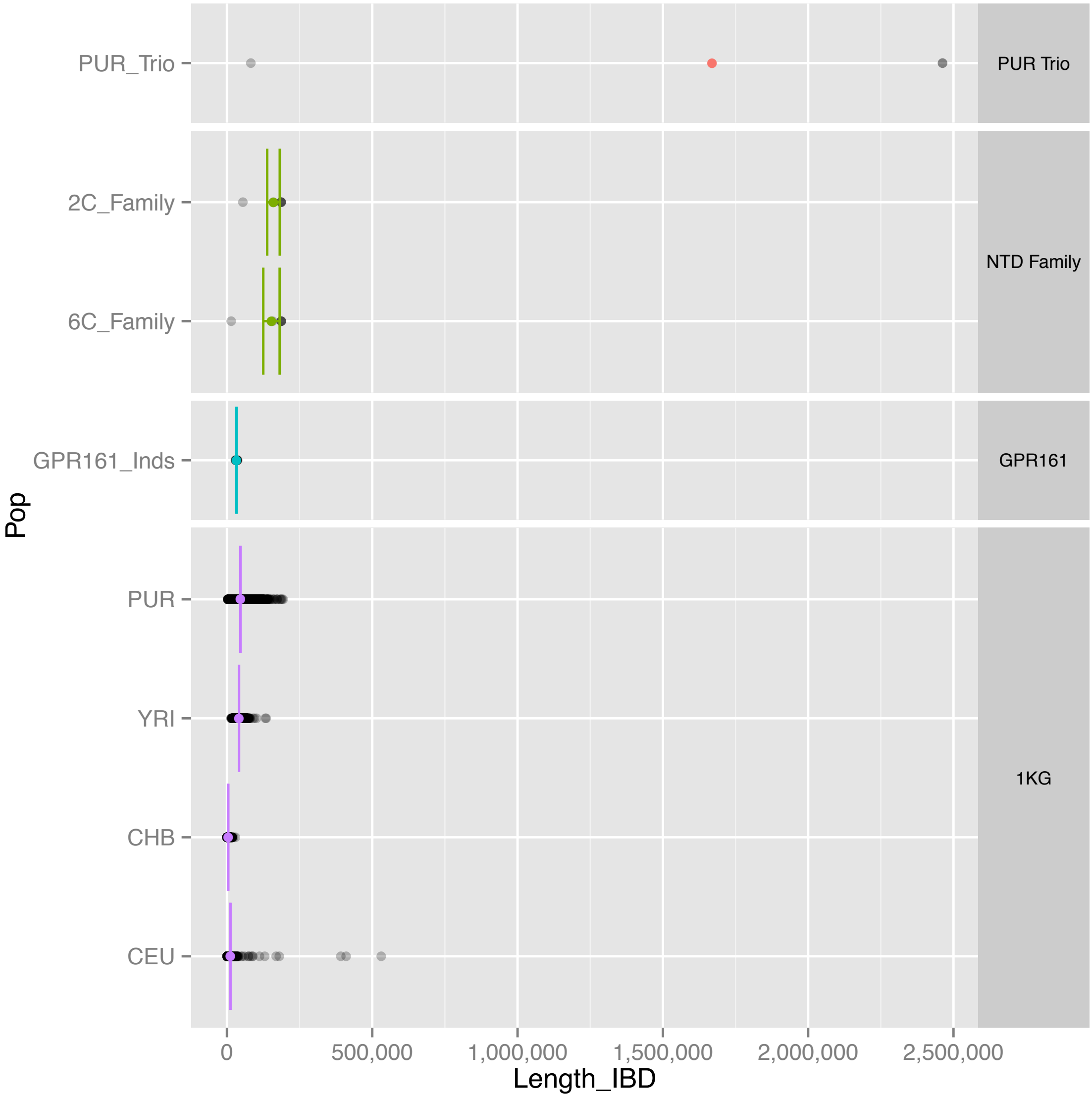


Figure 4.2. Inferences of ancestry and familial relatedness from genomic data.. Distribution of the length of IBD regions compared between the 7 cases with GPR161 mutation and 1000 Genomes populations (1KG). IBD was estimated using BEAGLE. Error bars indicate SE of the

mean (not shown for PUR Trio as the number is small). The mean length of IBD of the 7 cases was well outside of the range of first degree relatives (PUR Trio and NTD Families) and was comparable to the means of the 1KG. Population Abbreviations: PUR-Puerto Ricans from Puerto Rico, YRI-Yoruba in Ibadan, Nigeria, CHB-Han Chinese in Beijing, China, CEU-Utah Residents (CEPH) with Northern and Western Ancestry. 2C and 6C are quartets of families with NTD, used to compare relatedness here, but were not included in the cohort of 125 cases or 158 controls.

4.4.4. From Validating Variants to Annotations

The preceding analysis illustrates in the necessity of thinking about ancestry and population stratification from the beginning of performing large scale WGS studies. To validate the finding, ancestry, relatedness and inferred location had to be computed, sometimes with multiple tools and compared. Part of the difficulty was not having direct access to all the patients to attempt to verify. However, self report of ancestry may not be reliable, especially in areas with highly admixed populations (Burnett et al., 2006; Lins et al., 2011), such as in the U.S. or Brazil. Another difficulty, common in bioinformatics, is that the relevant software and statistical tools may be novel, and thus the output difficult to interpret or subtle analytical mistakes easily made. Expert consultation can make a significant difference in ensuring accuracy.

The final section of this chapter shows the result of comparing different variant annotation services. This is an area that has seen continual growth and innovation, however there is still a great deal unknown about the genome. Following the discussion of ancestry, interpreting a variant cannot be done without knowing what population it's being studied in, hence if a service has difficulty assigning the input ancestry, potential skepticism in their remaining results may be warranted.

4.5 Comparison of Genomic Annotation Services

The age of next-generation sequencing has brought with large amounts of ever-cheaper data, whose rate of growth continues to accelerate, even as the cost of generating it continues to fall. Although the $1000 genome has been announced, the process of annotating and analyzing a genome to find clinically relevant variants presents a much more difficult task, and often dominates the cost of clinical genomics (Mardis, 2010). This demand for genomic annotation has lead to the creation of several commercial services for this purpose. We compared the output of three such services: Ingenuity's Variant Analysis Service, GenomeQuest's GQ-IP, and Omica's Opal. All three offer the ability to annotate and filter variants in order to detect ones that may be deleterious. Using comparable filtering processes (see Table 5.1), we generated lists of genes with potentially deleterious variants from all three services for twelve (12) sample genomes collected from several ongoing studies in our lab. Omicia also offered the ability to perform a Variant Annotation, Analysis and Search Tool (VAAST) (Yandell et al., 2011) Solo analysis. The VAAST analyses were used as a separate set for comparison. We then submitted these genes to the National Institute of Allergy and Infectious Diseases's (NIAID's) Database for Annotation, Visualization, and Integrated, Discovery (DAVID) (Dennis et al., 2003) functional annotation clustering tool, using its default setting (see Methods), and compared the resulting lists of annotations.

4.5.1 Results

The results of comparing the genes identified from each service are shown in Table 4.3. The most immediate striking feature is the lack of genes with consensus between all three services, even in the less stringent case where the Omicia VAAST analysis was not used. Also note the large among of variance in the number of genes returned between each service, with GenomeQuest consistently returning over approximately 200 variants, compared to ~20 for GenomeQuest. Using the additional VAAST analysis offered by Omicia, only two individuals (out of 12, 16.6%) had overlap between all three filtered lists, consisting of a single gene in both cases.

The initial comparison of the DAVID annotations, shown in Table 5.4, were more promising, with 58% (7/12) of the samples having some consensus annotations between all three services in the VAAST comparison, 92% (11/12) in the non-VAAST comparison. However, inspection immediately showed that this approach has limited clinical utility, due to the lack of specificity of many of these consensus annotations, such as the gene ontology (GO) terms “olfaction” and “membrane”.

We chose to further analyze the similarity of the functional annotation clusters output by DAVID, which included enrichment scores for each individual annotation. We represented the data as a network, using the annotation clusters as nodes, and creating edges consisting of annotation terms shared between clusters, with edge weights determined by the cumulative enrichment of all such common terms. We describe this as a network-based comparison of functional annotation clusters (NET-COFAC). To assess the robustness of the networks, we successively removed edges with higher and higher weights, and measured the number of subgraphs generated at each step, and then performed a two-parameter exponential fit (Figure 4.4). This allowed us to visually assess the degree of similarity in the annotation clusters across samples both through the underlying curves and by examining the parameter values. We generated three benchmarks sets- a negative control with random lists, a positive control with identical lists, and another negative control in which intra-individual gene lists were compared by permuting them between samples. Plotting these the NET-COFAC output parameters for these benchmark, we were able to visually identify regions in the parameter space, which corresponded to each individual control group. Plotting the actual data for the VAAST and non-VAAST comparisons (Figure 4.5), we visually found that their NET-COFAC parameters were fairly similar to the intra-sample control, and thus while generally distinguishable from comparing random lists, ultimately did not carry much signal.

As a positive control, we added ten (10) lines to the NA12878 vcf, introducing homozygous mutations known to cause common conditions, and submitted the file to both the Omicia and Ingenuity services. The list of genes and the RefSeq IDs of the introduced variants are given in Table 5(a). The result of running this file through the analysis pipeline used for the other files is shown in Table 5.5(b). For each service 30% (3/10) of the introduced variants survived the filtering process, with a single gene (HEXA) overlapping. The remaining five variants were verified to be in the uploaded data, but not in the final filtered variant list.

4.5.2 Genome Annotation Discussion

Fundamental questions about current approaches to genomics, both with respect to the reference genome (Rosenfeld et al., 2012) and to our ability to accurately call variants using a single chemical or software pipeline have already been raised .(O'Rawe et al., 2013) Our results indicate that similar uncertainty exists when doing downstream analysis and variant annotation. Just as O'Rawe et al indicated the need to use multiple approaches to achieve a greater accuracy in calling variants, based on our results, we recommend caution when using any genome annotation service on a single individual to identify deleterious variants. If one has access to multiple services, then their output should certainly be compared, and potentially aggregated in order to attempt to generate higher confidence in diagnosing a variant or pathway as a potential target. Of the services studied, Omicia's Opal was the least costly, offering free variant annotation, and $100 per sample for further analysis, and thus offers a cost-effective way to generate a comparison for another pipeline or service.

The source of the variation in the identified genes is unknown. As the filtering processes were not completely uniform, the slight differences may have been the cause of the divergence. Another possibility is that different services used different quality filters when importing the data, thus creating a non-uniform pool of initial variants. An additional possibility would be differences in algorithmic implementation, or even perhaps an erroneous implementation in one or more of the services. Finally, it is possible that the variance would be reduced in duo or trios studies, which all three services also offer.

NET-COFAC was designed as a way to quantify similarity in functional annotation clusters. We note that for an actual case, one could for example, use the aggregate enrichment scores of all consensus annotations to provide a ranking of terms, thus helping to locate terms or pathways that are potentially actionable. The variance in the NET-COFAC fit parameters may indicate that some genomes are fundamentally more complex than others – “complexity” referring here to our ability to identify the functions that are being affected by genetic variants. It is not immediately apparent whether such complexity is fundamental in nature, or dependent on the reference genome being utilized. Further study is needed to probe this question, as well as to how great the variance in genome complexity might be within and across populations.

4.5.3 Methods

4.5.3.1 Samples

Twelve (12) whole genome VCF files were used from ongoing projects within the Mason lab. Seven (7) of the genomes were of individuals with medical condition, and five (5) were controls. Their descriptions were: four (4) affected children from a neural tube-defect study from two separate families (2C1, 2C2, 6C1, 6C2), four (4) control genomes of members of the Mason lab (LPA2, LPB2, LPC2, LPF1), three (3) subject from the NIH Office of Rare Disease Research's Undiagnosed Diseases Project (UDP) (UDP441, UDP3427, UDP4823), and the 1000 Genomes Hi-Seq Whole Genome file for NA12878. The list of sample identifiers and their sex and group are given in Table 4.1.

4.5.3.2 Variant Annotation

The VCF files were submitted to Ingenuity's Variant Analysis, Omicia's Opal, and GenomeQuest via their online interfaces. The full list of filters for each service is described below, and shown in Table 2. The filtering processes where kept as similar as possible, and used criteria for filtering suitable for analyzing an individual with no prior information about their condition. Liberal thresholds for each parameter were generally used. In this way short lists of potentially deleterious variants were generated for each service. These lists were compared against each other to find consensus genes.

A summary of the settings used to filter the VCF files and their associated genetic variants follows: Variants with a frequency in the population of more than 10% were filtered out. Their SIFT (Ng and Henikoff, 2003) score, a method based on sorting intolerant from tolerant amino acids to predict damaging substitutions based on conservation, was required to be less than 0.1. Note that this is higher than the standard threshold of 0.05 for predicting a damaging mutation, in keeping with the liberal approach to the filtering described above. The similar PolyPhen [6] score also predicts damaging substitutions, this time based on physical considerations. Each service had a filter that screened for protein impact, which we applied, again using liberal parameters when possible. At this point, we let the procedure diverge to use features unique to each service. For Omicia, we required that the variant's Omicia Score, which is a meta-classifier combining SIFT, PolyPhen , MutationTester [8], and phyloP values [9], be above 75. Omicia, also offered a VAAST (Variant Annotation, Analysis, and Search Tool) analysis, which probabilistically identifies damaging genes. Genes that were included in the VAAST solo report were used to conduct an additional comparison with the other services. For GenomeQuest, we required that the variant's Clinical Significance was not benign, and did not involve drug-response. Finally for Ingenuity, we used one of the pre-built filters, which required that the observed variant be observed to be or possibly be pathogenic.

4.5.3.3 DAVID Functional Annotation

The resulting gene lists were submitted to the DAVID annotation tool for functional annotation and clustering, using the default setting. The resulting annotation terms were then extracted and compared. The DAVID default settings consist of the following annotations:

Disease: OMIM_DISEASE

Functional Categories: COG_ONTOLOGY, SP_OIR_KEYWORDS, UP_SEQ_FEATURE

Gene Ontology: GOTERM_BP_FAT, GOTERM_CC_FAT, GOTERM_MF_FAT

Pathways: BIOCARTA, KEGG_PATHWAY

Protein Domain: INTERPRO, PIR_SUPERFAMILY, SMART.

4.5.3.4 Network-based Comparison of Functional Annotation Clustering (NET-COFAC)

The DAVID output was used generate the networks for NET-COFAC. Analysis was done in Python using the iPython interface and utilizing the NetworkX package to programmatically create the graphs and visualize the data.

4.5.3.5 Ethics Approval

Subject participation was obtained through IRB approved protocols reviewed by the state of California and Stanford University, the University of Texas at Austin, Weill Cornell Medical College (NY and Qatar) and Hamad Medical Corporation. Consent was obtained from members of the Mason lab for participation in the study. Research was carried out in compliance with the Helsinki Declaration.

4.5.3.6 Competing Interests

Dr. Mason is the co-founder of a genetic testing company, Genome Liberty, but this company does not have a competing product to the commercial entities used in this manuscript.

4.5.3.7 Acknowledgements

Support was provided by the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant 1T32GM083937)

Sample ID	Group
NA12878	Standard Control
LP_A2	Mason Lab Member
LP_B2	Mason Lab Member
LP_C2	Mason Lab Member
LP_F1	Mason Lab Member
2C1	Neural Tube Defect, affected child, Family 1
2C2	Neural Tube Defect, affected child, Family 1
6C1	Neural Tube Defect, affected child, Family 2
6C2	Neural Tube Defect, affected child, Family 2
UDP_441	Undiagnosed Diseases Program
UDP_3427	Undiagnosed Diseases Program
UDP_4823	Undiagnosed Diseases Program

Table 4.1 Sample ID and Information. Whole genome files were used from two different ongoing studies: one involving neural tube defects in children, the others from the Office of Rare Diseases Research's Undiagnosed Diseases Program. Five (5) controls were also used: the standard control sample, NA12878, and four genomes from members of the Mason lab.

Omicia Filters	GenomeQuest Filters	Ingenuity Filters
Frequency: ≤0.1	Minor Allele Frequency: ≤0.10	Common Variants: ≤0.10 in 1000 Genomes, Complete Genomics, OR ESP genomes
SIFT score : ≤0.1	SIFT Score: ≤0.1	Not tolerated by SIFT
Polyphen Prediction: Probably damaging	Polyphen: Damaging, Probably Damaging	Not tolerated by PolyPhen-2
Protein Impact: All	Predicted Impact: is not SILENT	Genetic Analysis: Inferred gain –or loss-of-function variants (default settings)
Omicia Score: ≥0.75	Clinical Significance: is not benign, drug-response	Predicted Deleterious: Experimentally observed Pathogenic OR Possibly Pathogenic
Present in VAAST Solo Report		

Table 4.2. Filters Used with Genome Analysis Services. Filters were selected to be as similar as possible. Analysis was done both with the above settings, and excluding the results of the Omicia VAAST Solo report. An explanation of the various acronyms and terms follows: SIFT (Sorting Intolerant From Tolerant) scores variants based on their effect on conserved amino acid substitutions. PolyPhen (Polymorphism Phenotyping) scores variants based on their effect on structure and function of proteins. phyloP (Phylogentic P values) assigns a p-value based on base-pair resolution conservation and selection-detection. The Omicia Score is a meta-classifier combining SIFT, PolyPhen, MutationTester, and phyloP values. VAAST (Variant Analysis,

Annotation, and Search Tool) probabilistically predicts damaging genes based on prioritizing predicted amino substitutions.

A.

Sample ID	GenomeQuest	Omicia	Ingenuity	GenomeQuest & Ingenuity	GenomeQuest & Omicia	Ingenuity & Omicia	GenomeQuest, Ingenuity, & Omicia
2C1	266	10	29	2	3	3	0
2C2	260	14	35	1	3	4	0
6C1	323	99	45	3	46	8	1
6C2	295	108	37	2	47	8	0
LPA2	262	7	35	1	1	3	0
LPB2	243	11	30	0	2	3	0
LPC2	266	12	35	2	5	4	0
LPF1	295	11	30	1	5	2	0
NA12878	11	4	7	0	0	0	0
UDP3427	256	6	23	2	2	1	0
UDP441	247	22	60	4	6	9	0
UDP4823	193	11	35	1	1	4	1

B.

Sample ID	GenomeQuest	Omicia	Ingenuity	GenomeQuest & Ingenuity	GenomeQuest & Omicia	Ingenuity & Omicia	GenomeQuest, Ingenuity, & Omicia
2C1	266	92	29	2	53	6	1
2C2	260	117	35	1	49	8	0
6C1	323	99	45	3	46	8	1
6C2	295	108	37	2	47	8	0
LPA2	262	97	35	1	12	2	0
LPB2	243	97	30	0	51	6	0
LPC2	266	94	35	2	49	9	1
LPF1	295	95	30	1	52	6	1
NA12878	11	101	7	0	0	1	0
UDP3427	256	94	23	2	56	5	1
UDP441	247	94	60	4	51	9	0
UDP4823	193	87	35	1	37	9	1

Table 4.3 Consensus of Gene Lists from Omicia, GenomeQuest and Ingenuity Genome Analysis Services. (A) Comparison including Omicia VAAST Solo Report. (B) Comparison excluding Omicia VAAST solo report (grey columns are redundant). Each column gives the number of genes in each group, which are not disjoint; the first three columns contain the total number of genes returned from each service after filtering. Note the low and often non-existent pair wise overlap, and the negligible census amongst all three lists. It is also of note that the non-VAAST Omicia and GQ lists share significant overlap, indicating the VAAST analysis is an important source of divergence. The two genes that were found with triple consensus were HYDIN and CDK3, for samples 2C1 and UDP4823 respectively.

A.

Sample ID	GenomeQuest	Omicia	Ingenuity	GenomeQuest & Ingenuity	GenomeQuest & Omicia	Ingenuity & Omicia	GenomeQuest, Ingenuity, & Omicia
2C1	626	0	120	76	0	0	0
2C2	573	17	79	47	17	17	17
6C1	678	280	188	133	208	87	83
6C2	586	313	137	99	223	93	80
LPA2	818	0	104	67	0	0	0
LPB2	696	0	59	38	0	0	0
LPC2	608	11	141	99	9	9	9
LPF1	714	24	89	66	22	22	22
NA12878	12	9	0	0	5	0	0
UDP3427	734	4	60	34	0	0	0
UDP441	537	51	153	98	41	32	30
UDP4823	487	12	162	105	11	10	10

B.

Sample ID	GenomeQuest	Omicia	Ingenuity	GenomeQuest & Ingenuity	GenomeQuest & Omicia	Ingenuity & Omicia	GenomeQuest, Ingenuity, & Omicia
2C1	626	205	120	76	175	56	56
2C2	573	364	79	47	247	43	42
6C1	678	280	188	133	208	87	83
6C2	586	313	137	99	223	93	80
LPA2	818	257	104	67	216	49	45
LPB2	696	257	59	38	223	40	38
LPC2	608	309	141	99	263	100	93
LPF1	714	252	89	66	208	67	65
NA12878	12	234	0	0	12	0	0
UDP3427	734	354	60	34	274	34	33
UDP441	537	202	153	98	174	62	62
UDP4823	487	286	162	105	211	82	78

Table 4.4. Consensus of DAVID Functional Annotation Clustering Terms. (A) Comparison including Omicia VAAST Solo Report. (B) Comparison excluding Omicia VAAST solo report (grey columns are redundant). The gene lists generated from the filtering process were passed through the DAVID functional annotation clustering tool, and the resulting annotations extracted into a single list for each service. Each column gives the number of annotations found in each category. Note the high degree of variation in consensus.

GO:0000166~nucleotide binding	GO:0043169~cation binding
GO:0000278~mitotic cell cycle	GO:0043228~non-membrane-bounded organelle
GO:0000279~M phase	GO:0043232~intracellular non-membrane-bounded organelle
GO:0001882~nucleoside binding	GO:0043233~organelle lumen
GO:0001883~purine nucleoside binding	GO:0043549~regulation of kinase activity
GO:0003677~DNA binding	GO:0044430~cytoskeletal part
GO:0004672~protein kinase activity	GO:0045449~regulation of transcription
GO:0005524~ATP binding	GO:0046872~metal ion binding
GO:0005654~nucleoplasm	GO:0046914~transition metal ion binding
GO:0005856~cytoskeleton	GO:0051174~regulation of phosphorus metabolic process
GO:0005886~plasma membrane	GO:0051252~regulation of RNA metabolic process
GO:0006350~transcription	GO:0051338~regulation of transferase activity
GO:0006355~regulation of transcription, DNA-dependent	GO:0070013~intracellular organelle lumen
GO:0006468~protein amino acid phosphorylation	IPR000719:Protein kinase, core
GO:0006793~phosphorus metabolic process	IPR007110:Immunoglobulin-like
GO:0006796~phosphate metabolic process	IPR013783:Immunoglobulin-like fold
GO:0007049~cell cycle	IPR017441:Protein kinase, ATP binding site
GO:0008270~zinc ion binding	Immunoglobulin domain
GO:0009890~negative regulation of biosynthetic process	Transcription
GO:0009991~response to extracellular stimulus	active site:Proton acceptor
GO:0010033~response to organic substance	atp-binding
GO:0010558~negative regulation of macromolecule biosynthetic process	binding site:ATP
GO:0010604~positive regulation of macromolecule metabolic process	disulfide bond
GO:0010605~negative regulation of macromolecule metabolic process	dna-binding
GO:0016021~integral to membrane	glycoprotein
GO:0016310~phosphorylation	glycosylation site:N-linked (GlcNAc...)
GO:0017076~purine nucleotide binding	kinase
GO:0019220~regulation of phosphate metabolic process	membrane
GO:0022402~cell cycle process	metal-binding
GO:0022403~cell cycle phase	nucleotide phosphate-binding region:ATP
GO:0030554~adenyl nucleotide binding	nucleotide-binding
GO:0031224~intrinsic to membrane	nucleus
GO:0031327~negative regulation of cellular biosynthetic process	signal
GO:0031667~response to nutrient levels	signal peptide
GO:0031974~membrane-enclosed lumen	topological domain:Cytoplasmic
GO:0031981~nuclear lumen	topological domain:Extracellular
GO:0032553~ribonucleotide binding	transcription regulation
GO:0032555~purine ribonucleotide binding	transferase
GO:0032559~adenyl ribonucleotide binding	transmembrane
GO:0042127~regulation of cell proliferation	transmembrane region
GO:0042325~regulation of phosphorylation	Zinc
GO:0042995~cell projection	zinc-finger
GO:0043167~ion binding	

Table 4.5. Consensus DAVID Annotations for Sample LPC2. LPC2 had by far the most agreement in its annotations from each service, even though they did not agree in their filtered gene lists. Nonetheless, the resulting consensus list does little do point to a clear clinical focus, due to the generality of the terms involved, although this could also be attributed to the sample being from a healthy control subject. Annotations were done with the DAVID default setting, which

includes GO cellular component, biological process, and molecular function annotations, as well as KEGG pathways.

(A)

Condition	Gene	Chr	Position (hg19)	RefSeq ID	Ref	Alt
Cystic Fibrosis	CFTR	7	117199646	rs113993960	CTT	-
Hemophilia	F8	X	154132090	rs4898352	A	T
Sickle Cell	HBB	11	5248232	rs334	T	A
Tay Sachs	HEXA	15	72637869	rs121907952	C	T
Lactose Intolerance	LCT	2	136564701	rs121908936	A	T
Familial Medit. Fever	MEFV	16	3293310	rs28940579	A	G
Color Blindness	OPN1MW	X	153461425	rs104894916	G	A
α1-antitrypsin deficiency	SERPINA1	14	94847262	rs17580	T	A
Spinal Muscular Dsytrophy	SMN1	5	70241990	rs76871093	C	T
Cancer (p53)	TP53	17	7578406	rs28934578	C	T

(B)

Ingenuity	Omicia	Consensus	Missed
HEXA	SERPINA1	HEXA	CFTR
LCT	HEXA		F8
SMN1	TP53		HEXA
			LCT
			MEFV

Table 4.6. NA12878 Positive Control. The NA12878 vcf file was used to generate a positive control file by adding homozygous variants associated with well-studied genetic diseases and conditions. (A). Information on the ten (10) introduced variants, including condition, position, and associated gene and RefSeq ID. (B) Genes recovered post-filtering from Ingenuity and Omicia variant analysis services. In each, 30% (3/10) of the introduced variants survived the filtering process, with one gene (HEXA) overlapping. The other five variants were verified to be present in the uploaded files, but not in the filtered variant lists.

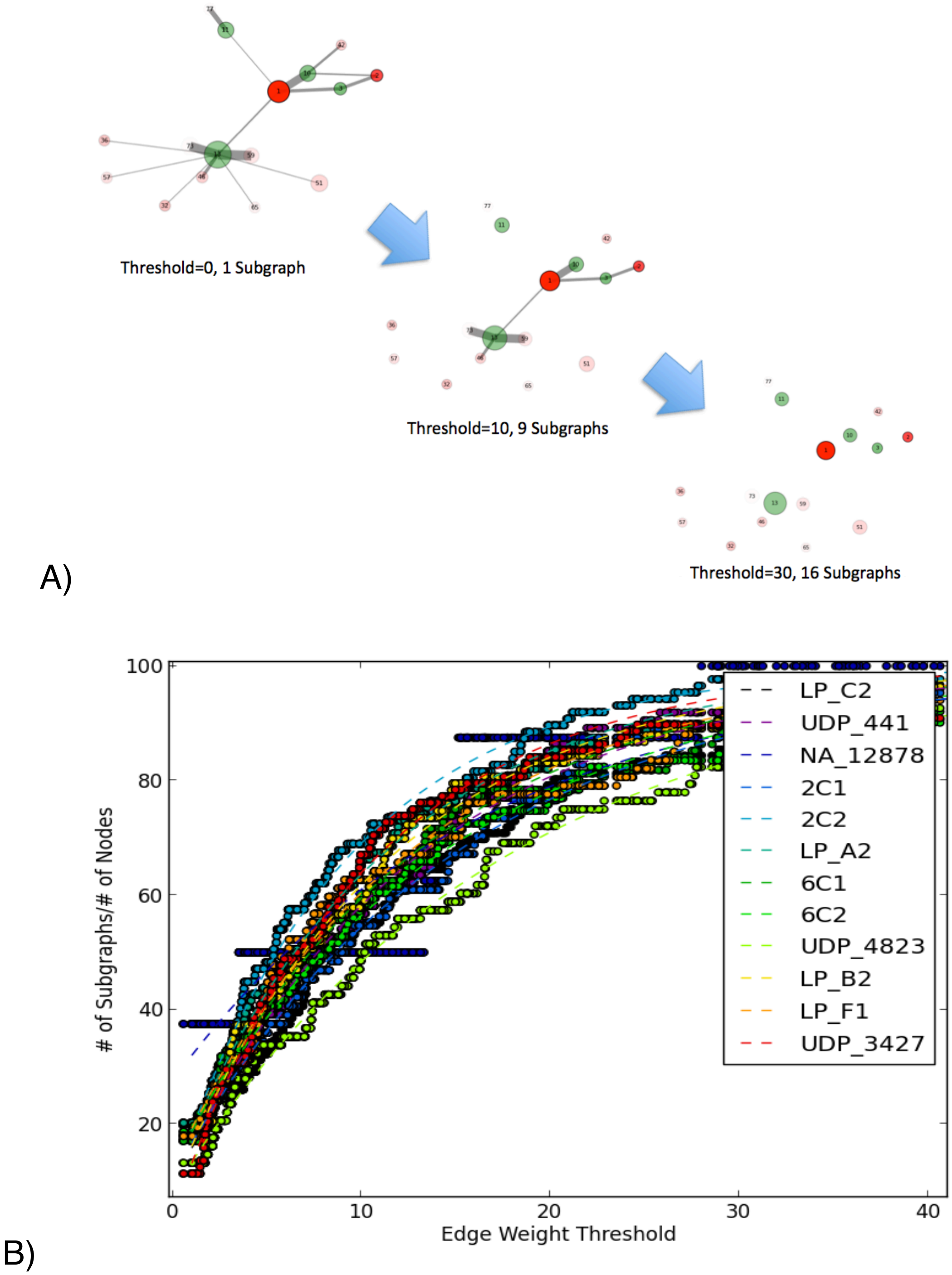


Figure 4.3 Network Analysis of Functional Annotation Clustering (NET-COFAC) (A). Schematic diagram of a NET-COFAC analysis. Nodes represent functional annotation clusters generated by DAVID, edges represent annotation terms they share, edge weights are given by the sum of the enrichment of those terms. In NET-COFAC, edges are removed from the network at higher and higher thresholds, and the number of resulting subgraphs is counted, until all nodes are isolated. Nodes are colored based on the service they belong to;. (B) Network robustness as measured using NET-COFAC. The y-axis shows a normalized metric for the degree the network has been separated. The x-axis is the edge weight threshold, edges below that threshold have been removed. Plots were fitted to a logistic curve.

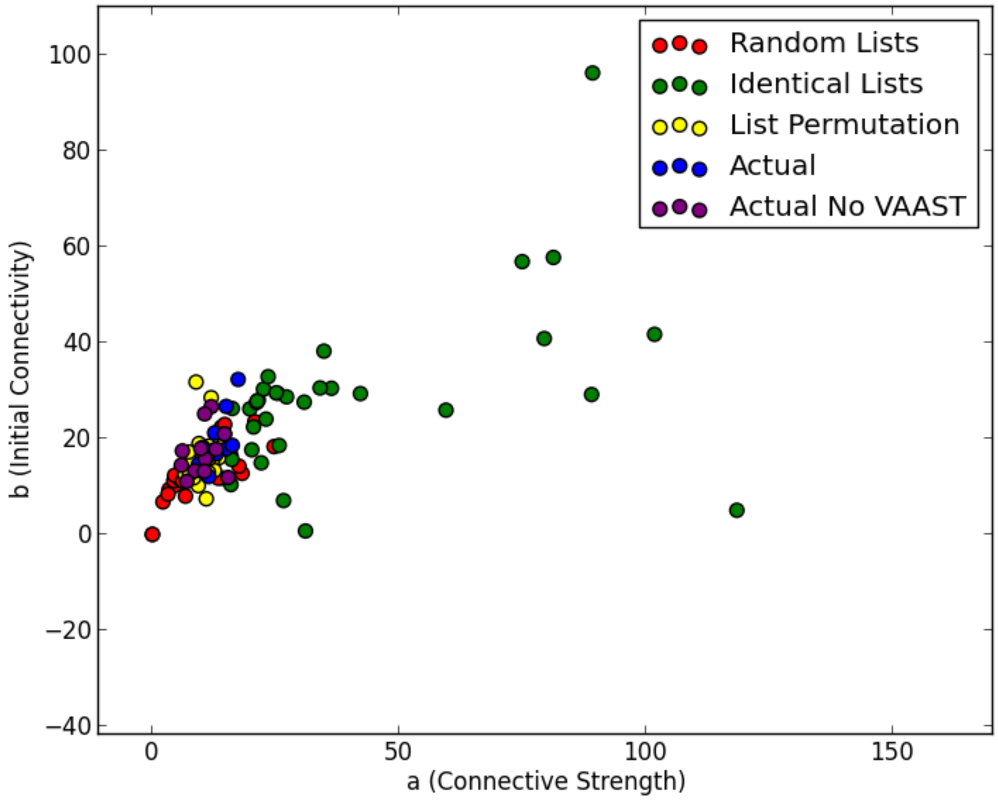


Figure 4.4. Testing NET-COFAC Validity. The parameters generated for the logistic fit from NET-COFAC are plotted above for the actual data, and several generated control sets. Identical lists form a manner of 'positive control' setting an upper bound on how robust the network could be. Random lists serve as a negative control. A final intermediate control was generated by permuting the gene lists among the samples, for example, comparng the Omicia list from UDP441, the GenomeQuest list from NA12878, and the Ingenuity list from LP_C2.

CHAPTER 5

Conclusion

5.1 Overview

The work done for this thesis had the primary goal of gaining understanding of FA's opposing mechanism of action in Lrp6 *Cd* and KO mutants. The multiDiff package was created to help answer this question, and displays superior performance compared to the competing DSS-general method. We were able to use it to initially identify thousands of loci associated with the Lrp6$^{Cd/+}$, FA supplementation, and Lrp6$^{Cd/+}$-FA interactions in P2 Male mice on the *Cd* background.

We then performed an Integrative analysis of methylation and expression, which showed that there was little evidence for an additive model of FA action, in which epigenetic lesions generated as a result of Lrp6 dysregulation are acted upon. Rather, FA appears to act in either an independent or combinatorial manner, with more differentially expressed genes supporting the independent model in the Lrp6 KO background, and the combinatorial model in the Cd background. We identified Rn45s as a gene of interest in the KO background, and Folh1 as a gene of interest in *Cd* and also noted various differentially expressed and methylated genes with known associations with partners of Lrp6 in the Cd background, while emphasizing those with known connections to mouse NTDs in the KO background..

Ongoing challenges related to WGS and methylation studies in human NTDs, were discussed, especially the difficulties of handling sex differences. , genetic ancestry, and genomic annotation

5.2 Future Work in Mouse

Due to the constraints of the current Lrp6 dataset we are unable to robustly estimate and report gender specific effects, though current evidence suggests that the number of sites affected by sex and interactions is relatively large, and is not confined to chrX. Future data collection may allow us probe this question.

A deeper structural challenge with the current dataset is that the two Lrp6 mutations are raised on different genetic backgrounds. When analyzed together, far more DMCs are detected as being associated with background differences than any other effect of interest, which is consonant with the observed variability of the effect of mutation according to mouse background (Doetschman, 2009; Yoshiki and Moriwaki, 2006). In addition, each additional primary effect adds several associated interactions terms. Thus theoretically we would need at minimum model with 8 terms if we confined ourselves to binary interactions, to do a cross-background analysis together- four primary effects including background, and 4 interactions, with therefore 256 possible states. An additional complication is that the model would have to be *nested*, due to the fact that each mutation only appears on a single background. It seems possible that that the relatively aggressive calling of the MMD estimate would be highly oversensitive in such a use case.

A cleaner approach would be to use a CRISPR-Cas9 (Cong et al., 2013; Sander and Joung, 2014) system to edit the Lrp6 locus, so both mutations could be studied in the same background. CRISPR-Cas9 is an extremely powerful gene editing tool, which has been used to both knock in and knockout genes in mice. If the FA-responsiveness of *Cd* is in fact background dependent, the current dataset would have be able to provide even more value, as it would allow us to look for signals associated with background-mutation interactions. In general CRISPR has the ability to eventually consolidate the field of NTD mouse models into a greatly reduced number of background. The precisely manner in which such consolidation is unclear, but ideally of course the more human physiology a model animal captures the more useful it becomes.

To shed more light on FA-*Lrp6* interactions, the performance of CRISP or knock down experiments on the genes of interest fruitful line of pursuit in the Crooked tail background, as there are far fewer targets to consider. Doing such experiments will help clarify whether the methylation hypothesis of FA action is correct.

5.3 Future Work in Human

The fundamental complexities of NTD biology, in particular the large known environmental effects, continue to make WGS studies difficult until cohorts are assembled of large enough size for robust and replicable identification of candidate genes. However, in tension with this is the relatively low prevalence of NTD compared to other major neurological disorders , and the efficacy of folic acid as a preventative measure. In addition, on many occasions sequencing data is anonymized before being given to an analyst. Although this process may not be perfect, it still removes more detailed information about the nature of the defect. This hinders more focused analyses that might fruitfully be pursued to identify genes associated with specific subclasses of NTDs

However, cheaper and more mobile sequencing technology may eventually help overcome some of these difficulties by being able to consistently track intra-individual changes in circulating methylation and metabolites. The deployment of scale of such omics technologies may prove to be a revolution. Just as Facebook has changed attitudes toward privacy, companies such a 23 and Me may help make it possible for researchers to access more details phenotypic inclination and make such analyses possible.

5.4 Closing Thoughts

In the age of so-called “Big Data,” biology is simultaneously on the frontline and the back foot, as biological NGS data can be accumulated so quickly that it overruns both infrastructure and analysts. The data is often big “in the wrong direction,” meaning that the number of variables that can be observed within a given sample, ranging from expression (~10^5), methylation (~10^7), and

SNPs (~10^7), to name the first on an ever growing list, vastly outnumber the samples that can be collected, especially when the features vary temporally. Due to the complexities of biological systems, it is impossible to “let the data speak for itself”. Thus, even in the age of deep learning there will still be considerable need to exercise biological insight in the selection of assays, construction of datasets and the creation and fine-tuning of analysis. With the combination of strong collaborations throughout the field to increase cohorts to robust levels and achieve and maintain best analytic practices, the field of NTD research has the ability to make great progress in the 21st century.

References

Agresti, A. (2015). Wiley: Foundations of Linear and Generalized Linear Models.

Ahuja, N., and Issa, J.P. (2000). Aging, methylation and cancer. Histol. Histopathol. *15*, 835–842.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. *13*, R87.

Amerongen, R. van, and Nusse, R. (2009). Towards an integrated view of Wnt signaling in development. Development *136*, 3205–3214.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. *22*, 2008–2017.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

Barfield, R.T., Kilaru, V., Smith, A.K., and Conneely, K.N. (2012). CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinformatics *28*, 1280–1281.

Baylin, S.B. (2005). DNA methylation and gene silencing in cancer. Nat. Clin. Pract. Oncol. *2*, S4–S11.

Beerman, I., and Rossi, D.J. (2015). Epigenetic Control of Stem Cell Potential During Homeostasis, Aging, and Disease. Cell Stem Cell *16*, 613–625.

Bjornsson, H.T., Sigurdsson, M.I., Fallin, M.D., Irizarry, R.A., Aspelund, T., Cui, H., Yu, W., Rongione, M.A., Ekström, T.J., Harris, T.B., et al. (2008). Intra-individual change in DNA methylation over time with familial clustering. JAMA J. Am. Med. Assoc. *299*, 2877–2883.

Blencowe, H., Cousens, S., Modell, B., and Lawn, J. (2010). Folic acid to reduce neonatal mortality from neural tube disorders. Int. J. Epidemiol. *39*, i110–i121.

Bodnar, L.M., Himes, K.P., Venkataramanan, R., Chen, J.-Y., Evans, R.W., Meyer, J.L., and Simhan, H.N. (2010). Maternal serum folate species in early pregnancy and risk of preterm birth123. Am. J. Clin. Nutr. *92*, 864–871.

Booth, M.J., Ost, T.W.B., Beraldi, D., Bell, N.M., Branco, M.R., Reik, W., and Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat. Protoc. *8*, 1841–1851.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nat. Genet. *42*, 1093–1100.

Bower, C., and Stanley, F.J. (1989). Dietary folate as a risk factor for neural-tube defects: evidence from a case-control study in Western Australia. Med. J. Aust. *150*, 613–619.

Browning, B.L., and Browning, S.R. (2011). A Fast, Powerful Method for Detecting Identity by Descent. Am. J. Hum. Genet. *88*, 173–182.

Burnett, M.S., Strain, K.J., Lesnick, T.G., de Andrade, M., Rocca, W.A., and Maraganore, D.M. (2006). Reliability of self-reported ancestry among siblings: implications for genetic association studies. Am. J. Epidemiol. *163*, 486–492.

Burren, K.A., Scott, J.M., Copp, A.J., and Greene, N.D.E. (2010). The Genetic Background of the Curly Tail Strain Confers Susceptibility to Folate-Deficiency-Induced Exencephaly. Birt. Defects Res. A. Clin. Mol. Teratol. *88*, 76–83.

Carter, M., Ulrich, S., Oofuji, Y., Williams, D.A., and Ross, M.E. (1999). Crooked tail (Cd) models human folate-responsive neural tube defects. Hum. Mol. Genet. *8*, 2199–2204.

Carter, M., Chen, X., Slowinska, B., Minnerath, S., Glickstein, S., Shi, L., Campagne, F., Weinstein, H., and Ross, M.E. (2005). Crooked tail (Cd) model of human folate-responsive neural tube defects is mutated in Wnt coreceptor lipoprotein receptor-related protein 6. Proc. Natl. Acad. Sci. U. S. A. *102*, 12843–12848.

Chang, H.-Y., Suchindran, C.M., and Pan, W.-H. (2001). Using the overdispersed exponential family to estimate the distribution of usual daily intakes of people aged between 18 and 28 in Taiwan. Stat. Med. *20*, 2337–2350.

Chen, Y., Breeze, C.E., Zhen, S., Beck, S., and Teschendorff, A.E. (2016). Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. Epigenetics Chromatin *9*, 10.

Colas, J.-F., and Schoenwolf, G.C. (2001). Towards a cellular and molecular understanding of neurulation. Dev. Dyn. *221*, 117–145.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Copp, A.J., and Greene, N.D.E. (2010). Genetics and development of neural tube defects. J. Pathol. *220*, 217–230.

Croen, L.A., Shaw, G.M., Jensvold, N.G., and Harris, J.A. (1991). Birth defects monitoring in California: a resource for epidemiological research. Paediatr. Perinat. Epidemiol. *5*, 423–427.

Dattani, M.T., Martinez-Barbera, J.P., Thomas, P.Q., Brickman, J.M., Gupta, R., Mårtensson, I.L., Toresson, H., Fox, M., Wales, J.K., Hindmarsh, P.C., et al. (1998). Mutations in the homeobox gene HESX1/Hesx1 associated with septo-optic dysplasia in human and mouse. Nat. Genet. *19*, 125–133.

Deak, K.L., Siegel, D.G., George, T.M., Gregory, S., Ashley-Koch, A., and Speer, M.C. (2008). Further evidence for a maternal genetic effect and a sex-influenced effect contributing to risk for human neural tube defects. Birt. Defects Res. A. Clin. Mol. Teratol. *82*, 662–669.

Deng, H.-X., Chen, W., Hong, S.-T., Boycott, K.M., Gorrie, G.H., Siddique, N., Yang, Y., Fecto, F., Shi, Y., Zhai, H., et al. (2011). Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. Nature *477*, 211–215.

Dennis, G., Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, P3.

Detrait, E.R., George, T.M., Etchevers, H.C., Gilbert, J.R., Vekemans, M., and Speer, M.C. (2005). Human neural tube defects: developmental biology, epidemiology, and genetics. Neurotoxicol. Teratol. *27*, 515–524.

Doetschman, T. (2009). Influence of Genetic Background on Genetically Engineered Mouse Phenotypes. Methods Mol. Biol. Clifton NJ *530*, 423–433.

Dolzhenko, E., and Smith, A.D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics *15*, 215.

Ebbing, M., Bønaa, K.H., Nygård, O., Arnesen, E., Ueland, P.M., Nordrehaug, J.E., Rasmussen, K., Njølstad, I., Refsum, H., Nilsen, D.W., et al. (2009). Cancer incidence and mortality after treatment with folic acid and vitamin B12. JAMA J. Am. Med. Assoc. *302*, 2119–2126.

Ernest, S., Christensen, B., Gilfix, B.M., Mamer, O.A., Hosack, A., Rodier, M., Colmenares, C., McGrath, J., Bale, A., Balling, R., et al. (2002). Genetic and molecular control of folate-homocysteine metabolism in mutant mice. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *13*, 259–267.

Esapa, C.T., Benson, M.A., Schröder, J.E., Martin-Rendon, E., Brockington, M., Brown, S.C., Muntoni, F., Kröger, S., and Blake, D.J. (2002). Functional requirements for fukutin-related protein in the Golgi apparatus. Hum. Mol. Genet. *11*, 3319–3331.

Fatemi, M., Pao, M.M., Jeong, S., Gal-Yam, E.N., Egger, G., Weisenberger, D.J., and Jones, P.A. (2005). Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res. *33*, e176.

Feng, H., Conneely, K.N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. *42*, e69.

Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m6A RNA methylation. Nat. Rev. Genet. *15*, 293–306.

Garrett-Bakelman, F.E., Sheridan, C.K., Kacmarczyk, T.J., Ishii, J., Betel, D., Alonso, A., Mason, C.E., Figueroa, M.E., and Melnick, A.M. (2015). Enhanced Reduced Representation Bisulfite Sequencing for Assessment of DNA Methylation at Base Pair Resolution. J. Vis. Exp. JoVE.

Gelfand, A.E., and Dalal, S.R. (1990). A Note on Overdispersed Exponential Families. Biometrika *77*, 55–64.

Gerwinn, S., Macke, J.H., and Bethge, M. (2010). Bayesian inference for generalized linear models for spiking neurons. Front. Comput. Neurosci. *4*.

Gibbs, W.W. (2014). Biomarkers and ageing: The clock-watcher. Nat. News *508*, 168.

Gray, J.D., Nakouzi, G., Slowinska-Castaldo, B., Dazard, J.-E., Sunil Rao, J., Nadeau, J.H., and Elizabeth Ross, M. (2010). Functional interactions between the LRP6 WNT co-receptor and folate supplementation. Hum. Mol. Genet. *19*, 4560–4572.

Gray, J.D., Kholmanskikh, S., Castaldo, B.S., Hansler, A., Chung, H., Klotz, B., Singh, S., Brown, A.M.C., and Ross, M.E. (2013). LRP6 exerts non-canonical effects on Wnt signaling during neural tube closure. Hum. Mol. Genet. *22*, 4267–4281.

Greenberg, J.A., Bell, S.J., Guan, Y., and Yu, Y.-H. (2011). Folic Acid supplementation and pregnancy: more than just neural tube defect prevention. Rev. Obstet. Gynecol. *4*, 52–59.

Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011a). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat. Protoc. *6*, 468–481.

Gu, T.-P., Guo, F., Yang, H., Wu, H.-P., Xu, G.-F., Liu, W., Xie, Z.-G., Shi, L., He, X., Jin, S., et al. (2011b). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606–610.

Hansen, K.D., Langmead, B., and Irizarry, R.A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. *13*, R83.

Harris, M.J., and Juriloff, D.M. (2007). Mouse mutants with neural tube closure defects and their role in understanding human neural tube defects. Birt. Defects Res. A. Clin. Mol. Teratol. *79*, 187–210.

Harris, M.J., and Juriloff, D.M. (2010). An update to the list of mouse mutants with neural tube closure defects and advances toward a complete genetic perspective of neural tube closure. Birt. Defects Res. A. Clin. Mol. Teratol. *88*, 653–669.

Hashimshony, T., Zhang, J., Keshet, I., Bustin, M., and Cedar, H. (2003). The role of DNA methylation in setting up chromatin structure during development. Nat. Genet. *34*, 187–192.

Hayatsu, H. (2008). Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis--a personal account. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. *84*, 321–330.

Hellman, A., and Chess, A. (2007). Gene Body-Specific Methylation on the Active X Chromosome. Science *315*, 1141–1143.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

Horvath, S., Bennett, D.A., Lu, A.T., and Levine, M.E. (2015). Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer’s disease related cognitive functioning.

Hwang, S.-H., and Mukhopadhyay, S. (2015). G-protein-coupled receptors and localized signaling in the primary cilium during ventral neural tube patterning. Birt. Defects Res. A. Clin. Mol. Teratol. *103*, 12–19.

Inoue, A., and Zhang, Y. (2011). Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. Science *334*, 194.

Inoue, A., Shen, L., Dai, Q., He, C., and Zhang, Y. (2011). Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. Cell Res. *21*, 1670–1676.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186.

Isono, K., Nemoto, K., Li, Y., Takada, Y., Suzuki, R., Katsuki, M., Nakagawara, A., and Koseki, H. (2006). Overlapping Roles for Homeodomain-Interacting Protein Kinases Hipk1 and Hipk2 in the Mediation of Cell Growth in Response to Morphogenetic and Genotoxic Signals. Mol. Cell. Biol. *26*, 2758–2771.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle Georget. Tex *8*, 1698–1710.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. *17*, 239.

Ji, H., Ehrlich, L.I.R., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). A comprehensive methylome map of lineage commitment from hematopoietic progenitors. Nature *467*, 338–342.

Jiang, M., Stanke, J., and Lahti, J.M. (2011). The Connections Between Neural Crest Development and Neuroblastoma. Curr. Top. Dev. Biol. *94*, 77–127.

Jones, P.L., Veenstra, G.J., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J., and Wolffe, A.P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat. Genet. *19*, 187–191.

Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., and Hoffmann, S. (2016). metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res. *26*, 256–262.

Juriloff, D.M., and Harris, M.J. (2012). Hypothesis: The female excess in cranial neural tube defects reflects an epigenetic drag of the inactivating x chromosome on the molecular mechanisms of neural fold elevation. Birt. Defects Res. A. Clin. Mol. Teratol. *94*, 849–855.

Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. *45*, D840–D845.

Kirke, P.N., Mills, J.L., Molloy, A.M., Brody, L.C., O'Leary, V.B., Daly, L., Murray, S., Conley, M., Mayne, P.D., Smith, O., et al. (2004). Impact of the MTHFR C677T polymorphism on risk of neural tube defects: case-control study. BMJ *328*, 1535–1536.

Kitzman, J.O., Snyder, M.W., Ventura, M., Lewis, A.P., Qiu, R., Simmons, L.E., Gammill, H.S., Rubens, C.E., Santillan, D.A., Murray, J.C., et al. (2012). Non-invasive whole genome sequencing of a human fetus. Sci. Transl. Med. *4*, 137ra76.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinforma. Oxf. Engl. *27*, 1571–1572.

Lehmann-Werman, R., Neiman, D., Zemmour, H., Moss, J., Magenheim, J., Vaknin-Dembinsky, A., Rubertsson, S., Nellgård, B., Blennow, K., Zetterberg, H., et al. (2016). Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc. Natl. Acad. Sci. *113*, E1826–E1834.

Lei, Y., Fathe, K., McCartney, D., Zhu, H., Yang, W., Ross, M.E., Shaw, G.M., and Finnell, R.H. (2015). Rare LRP6 variants identified in spina bifida patients. Hum. Mutat. *36*, 342–349.

Li, B.I., Matteson, P.G., Ababon, M.F., Nato, A.Q., Lin, Y., Nanda, V., Matise, T.C., and Millonig, J.H. (2015). The orphan GPCR, Gpr161, regulates the retinoic acid and canonical Wnt pathways during neurulation. Dev. Biol. *402*, 17–31.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362–365.

Li, S., Garrett-Bakelman, F., Perl, A.E., Luger, S.M., Zhang, C., To, B.L., Lewis, I.D., Brown, A.L., D'Andrea, R.J., Ross, M.E., et al. (2014). Dynamic evolution of clonal epialleles revealed by methclone. Genome Biol. *15*, 472.

Lichinchi, G., Gao, S., Saletore, Y., Gonzalez, G.M., Bansal, V., Wang, Y., Mason, C.E., and Rana, T.M. (2016). Dynamics of the human and viral m^{6}A RNA methylomes during HIV-1 infection of T cells. Nat. Microbiol. *1*, nmicrobiol201611.

Lins, T.C., Vieira, R.G., Abreu, B.S., Gentil, P., Moreno-Lima, R., Oliveira, R.J., and Pereira, R.W. (2011). Genetic Heterogeneity of Self-Reported Ancestry Groups in an Admixed Brazilian Population. J. Epidemiol. *21*, 240–245.

Lister, R., and Ecker, J.R. (2009). Finding the fifth base: Genome-wide sequencing of cytosine methylation. Genome Res. *19*, 959–966.

Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., Koltšina, M., Nilsson, T.K., Vilo, J., Salumets, A., et al. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol. *15*, r54.

Lu, A.T., Hannon, E., Levine, M.E., Hao, K., Crimmins, E.M., Lunnon, K., Kozlenkov, A., Mill, J., Dracheva, S., and Horvath, S. (2016). Genetic variants near MLST8 and DHX57 affect the epigenetic age of the cerebellum. Nat. Commun. *7*.

Mardis, E.R. (2010). The $1,000 genome, the $100,000 analysis? Genome Med. *2*, 84.

Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. *33*, 5868–5877.

Mills, J.L., and Dimopoulos, A. (2015). Folic acid fortification for Europe? BMJ *351*, h6198.

Mills, J.L., Rhoads, G.G., Simpson, J.L., Cunningham, G.C., Conley, M.R., Lassman, M.R., Walden, M.E., Depp, O.R., and Hoffman, H.J. (1989). The absence of a relation between the periconceptional use of vitamins and neural-tube defects. National Institute of Child Health and Human Development Neural Tube Defects Study Group. N. Engl. J. Med. *321*, 430–435.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. Mol. Cell *30*, 755–766.

Molloy, A.M., Pangilinan, F., and Brody, L.C. (2017). Genetic Risk Factors for Folate-Responsive Neural Tube Defects. Annu. Rev. Nutr.

Morris, T.J., and Beck, S. (2015). Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. Methods *72*, 3–8.

Mukhopadhyay, S., Wen, X., Ratti, N., Loktev, A., Rangell, L., Scales, S.J., and Jackson, P.K. (2013). The ciliary G-protein-coupled receptor Gpr161 negatively regulates the Sonic hedgehog pathway via cAMP signaling. Cell *152*, 210–223.

Mulinare J, Cordero JF, Erickson J, and Berry RJ (1988). PEriconceptional use of multivitamins and the occurrence of neural tube defects. JAMA *260*, 3141–3145.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Ockeloen, C.W., Khandelwal, K.D., Dreesen, K., Ludwig, K.U., Sullivan, R., van Rooij, I.A.L.M., Thonissen, M., Swinnen, S., Phan, M., Conte, F., et al. (2016). Novel mutations in LRP6 highlight the role of WNT signaling in tooth agenesis. Genet. Med. Off. J. Am. Coll. Med. Genet. *18*, 1158–1162.

O'Rahilly, R., and Müller, F. (1994). Neurulation in the normal human embryo. Ciba Found. Symp. *181*, 70-82-89.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. *5*, 28.

Park, Y., and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. Bioinforma. Oxf. Engl. *32*, 1446–1453.

Park, Y., Figueroa, M.E., Rozek, L.S., and Sartor, M.A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics *30*, 2414–2422.

Polimanti, R., Yang, C., Zhao, H., and Gelernter, J. (2015). Dissecting ancestry genomic background in substance dependence genome-wide association studies. Pharmacogenomics *16*, 1487–1498.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nat. Biotechnol. *28*, 1057–1068.

Price, E.M., Peñaherrera, M.S., Portales-Casamar, E., Pavlidis, P., Van Allen, M.I., McFadden, D.E., and Robinson, W.P. (2016). Profiling placental and fetal DNA methylation in human neural tube defects. Epigenetics Chromatin *9*, 6.

Puckett, R.L., Moore, S.A., Winder, T.L., Willer, T., Romansky, S.G., Covault, K.K., Campbell, K.P., and Abdenur, J.E. (2009). Further evidence of Fukutin mutations as a cause of childhood onset limb-girdle muscular dystrophy without mental retardation. Neuromuscul. Disord. NMD *19*, 352–356.

van der Put, N.M.J., Trijbels, F.J.M., van den Heuvel, L.P., Blom, H.J., Steegers-Theunissen, R.P.M., Eskes, T.K.A.B., Mariman, E.C.M., den Heyer, M., Frosst, P., and Rozen, R. (1995). Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. The Lancet *346*, 1070–1071.

Rahmani, E., Shenhav, L., Schweiger, R., Yousefi, P., Huen, K., Eskenazi, B., Eng, C., Huntsman, S., Hu, D., Galanter, J., et al. (2017). Genome-wide methylation data mirror ancestry information. Epigenetics Chromatin *10*, 1.

Ritz, B.R., and Horvath, S. (2015). Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients.

Robertson, K.D., and Wolffe, A.P. (2000). DNA methylation in health and disease. Nat. Rev. Genet. *1*, 11–19.

Rochtus, A., Jansen, K., Van Geet, C., and Freson, K. (2015). Nutri-epigenomic Studies Related to Neural Tube Defects: Does Folate Affect Neural Tube Closure Via Changes in DNA Methylation? Mini Rev. Med. Chem. *15*, 1095–1102.

Rosenfeld, J.A., Mason, C.E., and Smith, T.M. (2012). Limitations of the human reference genome for personalized genomics. PloS One *7*, e40294.

Sánchez-Martín, F.J., Lindquist, D.M., Landero-Figueroa, J., Zhang, X., Chen, J., Cecil, K.M., Medvedovic, M., and Puga, A. (2015). Sex- and Tissue-Specific Methylome Changes in Brains of Mice Perinatally Exposed to Lead. Neurotoxicology *46*, 92–100.

Sander, J.D., and Joung, J.K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol. *32*, 347–355.

Sebold, C.D., Melvin, E.C., Siegel, D., Mehltretter, L., Enterline, D.S., Nye, J.S., Kessler, J., Bassuk, A., Speer, M.C., George, T.M., et al. (2005). Recurrence risks for neural tube defects in siblings of patients with lipomyelomeningocele. Genet. Med. Off. J. Am. Coll. Med. Genet. *7*, 64–67.

Semenza, G.L. (2001). HIF-1, O(2), and the 3 PHDs: how animal cells signal hypoxia to the nucleus. Cell *107*, 1–3.

Shaw, G.M., Schaffer, D., Velie, E.M., Morland, K., and Harris, J.A. (1995). Periconceptional vitamin use, dietary folate, and the occurrence of neural tube defects. Epidemiol. Camb. Mass *6*, 219–226.

Smith, A.D., Chung, W.-Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics *25*, 2841–2842.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339–344.

Smithells, R.W. (1984). Can vitamins prevent neural tube defects? Can. Med. Assoc. J. *131*, 273–276.

Smithells, R.W., Sheppard, S., and Schorah, C.J. (1976). Vitamin dificiencies and neural tube defects. Arch. Dis. Child. *51*, 944–950.

Smithells, R.W., Sheppard, S., Schorah, C.J., Seller, M.J., Nevin, N.C., Harris, R., Read, A.P., and Fielding, D.W. (1980). Possible prevention of neural-tube defects by periconceptional vitamin supplementation. Lancet Lond. Engl. *1*, 339–340.

Smithells, R.W., Sheppard, S., Schorah, C.J., Seller, M.J., Nevin, N.C., Harris, R., Read, A.P., and Fielding, D.W. (1981). Apparent prevention of neural tube defects by periconceptional vitamin supplementation. Arch. Dis. Child. *56*, 911–918.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., Nickerson, D.A., and Below, J.E. (2014). PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent. Am. J. Hum. Genet. *95*, 553–564.

Stolk, L., Bouwland-Both, M.I., van Mill, N.H., Verbiest, M.M.P.J., Eilers, P.H.C., Zhu, H., Suarez, L., Uitterlinden, A.G., and Steegers-Theunissen, R.P.M. (2013). Epigenetic Profiles in Children with a Neural Tube Defect; A Case-Control Study in Two Populations. PLoS ONE *8*.

van Straaten, H.W., and Copp, A.J. (2001). Curly tail: a 50-year history of the mouse spina bifida model. Anat. Embryol. (Berl.) *203*, 225–237.

Suh, J.R., Herbig, A.K., and Stover, P.J. (2001). New perspectives on folate catabolism. Annu. Rev. Nutr. *21*, 255–282.

Sullivan, G.M., and Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. J. Grad. Med. Educ. *4*, 279–282.

Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A., and Li, W. (2014). MOABS: model based analysis of bisulfite sequencing data. Genome Biol. *15*, R38.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368.

Szwagierczak, A., Bultmann, S., Schmidt, C.S., Spada, F., and Leonhardt, H. (2010). Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. Nucleic Acids Res. *38*, e181.

Wagner, J.R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. *15*, R37.

Wallingford, J.B. (2005). Neural tube closure and neural tube defects: Studies in animal models reveal known knowns and known unknowns. Am. J. Med. Genet. C Semin. Med. Genet. *135C*, 59–68.

Wallingford, J.B., Niswander, L.A., Shaw, G.M., and Finnell, R.H. (2013). The Continuing Challenge of Understanding, Preventing, and Treating Neural Tube Defects. Science *339*, 1222002.

Wang, D., Yan, L., Hu, Q., Sucheston, L.E., Higgins, M.J., Ambrosone, C.B., Johnson, C.S., Smiraglia, D.J., and Liu, S. (2012). IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinforma. Oxf. Engl. *28*, 729–730.

Wang, Y., Liu, Y., Ji, W., Qin, H., Wu, H., Xu, D., Turtuohut, T., and Wang, Z. (2015). Variants in MTHFR gene and neural tube defects susceptibility in China. Metab. Brain Dis. *30*, 1017–1026.

Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. *41*, e117.

Werler MM, Shapiro S, and Mitchell AA (1993). PEriconceptional folic acid exposure and risk of occurrent neural tube defects. JAMA *269*, 1257–1261.

Wheldon, L.M., Abakir, A., Ferjentsik, Z., Dudnakova, T., Strohbuecker, S., Christie, D., Dai, N., Guan, S., Foster, J.M., Corrêa, I.R., et al. (2014). Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. Cell Rep. *7*, 1353–1361.

Wiens, D. (2016). Could folic acid influence growth cone motility during the development of neural connectivity? Neurogenesis Austin Tex *3*, e1230167.

World Health Organization. (2015). Global health estimates (GHE)–Cause-specific mortality.

World Health Organization (2015). Global health estimates (GHE)–Disease burden.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C.J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat. Commun. *2*, 241.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Yamaguchi, Y., and Miura, M. (2013). How to form and close the brain: insight into the mechanism of cranial neural tube closure in mammals. Cell. Mol. Life Sci. CMLS *70*, 3171–3186.

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Res. *21*, 1529–1542.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. *13*, 335–340.

Yoshiki, A., and Moriwaki, K. (2006). Mouse phenome research: implications of genetic background. ILAR J. *47*, 94–102.

Zaganjor, I., Sekkarie, A., Tsang, B.L., Williams, J., Razzaghi, H., Mulinare, J., Sniezek, J.E., Cannon, M.J., and Rosenthal, J. (2016). Describing the Prevalence of Neural Tube Defects Worldwide: A Systematic Literature Review. PLOS ONE *11*, e0151586.

Zhang, Y., Baheti, S., and Sun, Z. (2016). Statistical method evaluation for differentially methylated CpGs in base resolution next-generation DNA sequencing data. Brief. Bioinform.

(1996). Food Standards: Amendment of Standards of Identity for Enriched Grain Products to Require Addition of Folic Acid; Correction.