

THE UNIVERSITY OF CHICAGO

PARENT OF ORIGIN EFFECTS ON GENE EXPRESSION AND QUANTITATIVE  
TRAITS IN THE HUTTERITES

A DISSERTATION SUBMITTED TO  
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES  
AND THE PRITZKER SCHOOL OF MEDICINE  
IN CANDIDACY FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS

BY  
SAHAR VICTORIA MOZAFFARI

CHICAGO, ILLINOIS

2018

Copyright © 2018 by Sahar Victoria Mozaffari

All Rights Reserved

Freely available under a CC-BY 4.0 International license

## Table of Contents

LIST OF FIGURES . . . . .	v
LIST OF TABLES . . . . .	vi
ACKNOWLEDGMENTS . . . . .	vii
ABSTRACT . . . . .	viii
1 INTRODUCTION . . . . .	1
1.1 Human Genetics and the Genetics of Complex Traits . . . . .	1
1.2 On The Origin of Genomic Imprinting . . . . .	3
1.3 The Search for Parent of Origin Effects . . . . .	5
2 PARENT OF ORIGIN EFFECTS ON QUANTITATIVE PHENOTYPES IN A FOUNDER POPULATION . . . . .	7
2.1 Abstract . . . . .	7
2.2 Introduction . . . . .	7
2.3 Results . . . . .	9
2.3.1 GWAS . . . . .	9
2.3.2 Parent of Origin GWAS . . . . .	10
2.3.3 Maternal and Paternal GWAS . . . . .	10
2.3.4 GWAS for Differential Parent of Origin Effects . . . . .	12
2.3.5 Parent of Origin Effects on Gene Expression . . . . .	16
2.4 Discussion . . . . .	17
2.5 Methods . . . . .	20
2.5.1 Sample Composition . . . . .	20
2.5.2 Genotype Data . . . . .	21
2.5.3 Phenotype Data . . . . .	21
2.5.4 GWAS . . . . .	21
2.5.5 Maternal and Paternal GWAS . . . . .	22
2.5.6 Differential Effect GWAS (PO-GWAS) . . . . .	22
2.5.7 Parent of Origin eQTL studies . . . . .	23
2.5.8 Maternal and Paternal Parent of Origin eQTL . . . . .	24
2.5.9 Differential Parent of Origin eQTL . . . . .	24
2.6 Supplementary Information . . . . .	24
2.6.1 Supplementary Figures . . . . .	24
3 PARENT OF ORIGIN GENE EXPRESSION IN A FOUNDER POPULATION IDENTIFIES TWO NOVEL IMPRINTED GENES AT KNOWN IMPRINTED REGIONS. . . . .	29
3.1 Abstract . . . . .	29
3.2 Introduction . . . . .	29

3.3	Results . . . . .	31
3.3.1	Mapping Transcripts to Parental Haplotypes . . . . .	31
3.3.2	Imprinted Genes in Lymphoblastoid Cell Lines (LCLs) . . . . .	32
3.3.3	Validation of Imprinted Genes in PBLs . . . . .	33
3.3.4	Methylation at Imprinting Control Regions . . . . .	33
3.4	Discussion . . . . .	34
3.5	Methods . . . . .	36
3.5.1	Genotypes . . . . .	36
3.5.2	RNA-seq in Lymphoblastoid Cell Lines (LCLs). . . . .	36
3.5.3	RNA-seq in Peripheral Blood Leukocytes (PBLs) . . . . .	37
3.5.4	Identifying Imprinted Genes . . . . .	37
3.5.5	DNA methylation profiling and processing in PBLs . . . . .	37
	REFERENCES . . . . .	39

## List of Figures

1.1	Low Hanging Fruit. . . . .	2
2.1	Maternal and Paternal GWAS results for Age of Menarche. . . . .	12
2.2	Maternal and Paternal GWAS results for LDL Cholesterol. . . . .	13
2.3	Opposite Effect Parent of Origin GWAS Result for Age of Menarche. . . . .	14
2.4	Opposite Effect Parent of Origin GWAS Result for BMI. . . . .	15
2.5	Manhattan and QQ Plots from Standard GWAS of 21 Quantitative Phenotypes	25
2.6	Maternal and Paternal GWAS results for CIMT. . . . .	26
2.7	Maternal and Paternal GWAS results for LVMI. . . . .	27
2.8	Maternal and Paternal GWAS results for FEV1. . . . .	27
2.9	Manhattan and QQ Plots from Differential Effect GWAS of 21 Quantitative Phenotypes . . . . .	28

## List of Tables <sup>1</sup>

1.1 Description of bacteria. . . . .	4
--------------------------------------	---

---

1. Note: Due to the large size of some tables, the tables have been provided in a supplementary file accompanying the dissertation. In such cases, the page number provided below directs the reader to a table's caption.

## **ACKNOWLEDGMENTS**

## **ABSTRACT**

# CHAPTER 1

## INTRODUCTION

### 1.1 Human Genetics and the Genetics of Complex Traits

A central goal of genetics is to understand the contribution of genetic variation to phenotypic variation. The mechanism by which genetic variants contribute to a phenotype is determined by the genetic architecture of the phenotype, however, the underlying rules that determine how genetic variants contribute to phenotype diversity is still not fully known.

Monogenic traits are determined by genetic variation in one gene and have discrete phenotypes. However, complex traits do not have predictable patterns of inheritance and have a large variety of phenotypes. Genetic variation in many genes, as well as interaction of genes with environmental factors can contribute to complex traits phenotypes.

Genome Wide Association Studies (GWAS) have been effective in detecting associations between common SNPs and common diseases since 2005, with the first large GWAS with good coverage of the genome in 2007 from the Wellcome Trust Case Control Consortium[82]. Although GWAS have resulted in the discovery of thousands of novel associations to hundreds of phenotypes, the loci identified by GWAS explain a small proportion of the estimated heritability of the trait, the fraction of phenotypic variation in a population that is due to genetic variation. There are many explanations that could explain this "missing heritability," or the proportion of heritability not accounted for by these significant GWAS loci: gene-environment interactions, epistatic interactions, inflated heritability estimates, rare variants not genotyped, common variants with small effect sizes, structural variants, and parent of origin effects[83, 23, 27, 86]. Heritability of parent of origin effects has been studied in mice[51] and only beginning to be studied in humans[39].

Additionally, in traditional GWAS, the impact of parental origin of associated alleles has been largely ignored, and the maternal and paternal alleles are treated as equivalent.

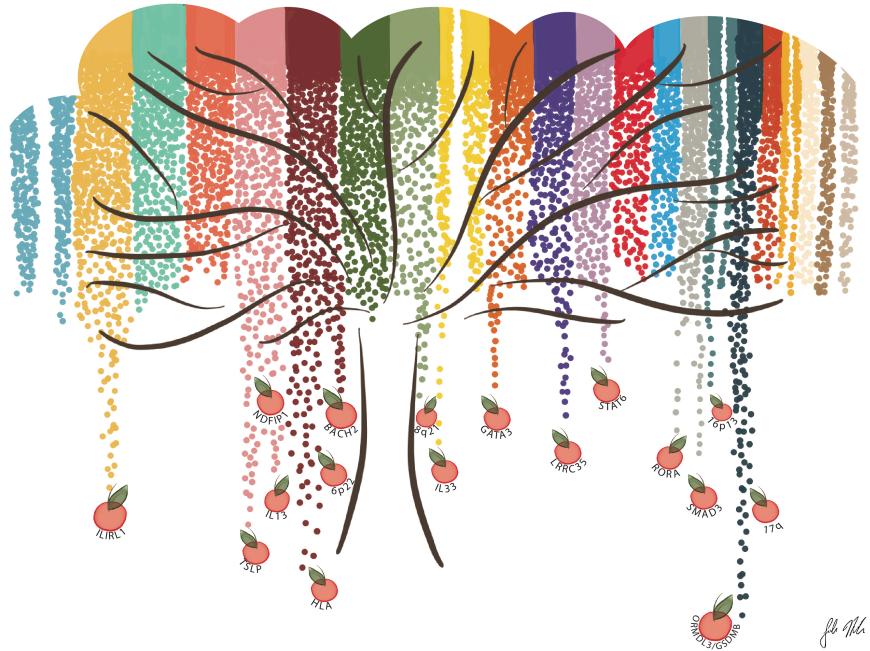


Figure 1.1: **Asthma GWAS - low hanging fruit**

Sequence variants could affect disease susceptibility or a quantitative trait differently depending on whether the variant was inherited from the father or the mother. Parent of origin effects include phenomena such as imprinting where epigenetic modifications determined by parental origin allows for differential gene expression on homologous chromosomes[43, 40].

Parental origin of mutations at imprinted loci have been involved in diseases such as Prader-Willi or Angelman syndrome[40, 60]. The testing of maternal and paternal alleles separately can disentangle parent of origin effects. Parent of origin effects can alter gene expression levels that can ultimately affect other phenotypic traits including disease. Moreover, parent of origin eQTLs can provide insight into the molecular mechanisms that may underlie genetic associations with both rare and common diseases[40, 60, 38, 71, 24].

The classic examples of parent of origin effects are imprinted genes. More than 80% of imprinted genes in humans are found in genomic clusters, and at least thirteen clusters have been identified on eight chromosomes[40, 60, 61, 2]. These clusters contain both mater-

nally and paternally expressed genes as well as genes that encode non-coding RNAs[60, 2]. Parent-specific expression of the genes within a cluster is determined by cis-acting imprinting control regions (ICRs). ICRs show parental allele-specific DNA methylation and chromatin modifications. ICRs acquire methylation in the female germ line during oogenesis. ICRs methylated in females typically contain the promoters of long non-coding RNA that are antisense to a protein- coding gene in the cluster and silence it. In contrast, ICRs that acquire methylation in the male germ line are located in intergenic regions.

## 1.2 On The Origin of Genomic Imprinting

Genomic imprinting in its broadest sense suggests that a phenotype observed for a particular gene or genes depends on the sex of the parent from with the gamete containing that gene or genes originated [67]. It was said that a particular gene is imprinted if it results in a different phenotype when it is maternally inherited versus paternally inherited.

The first use of the term "imprinting" was used in reference to the recognition by the cell of of chromosomes in *Sciara* [19, 67]. "The "imprint" a chromosome bears is unrelated to the genic constitution of the chromosome and is determined only by the sex of the germ line through with the chromosome has been inherited." [19]

The preferential inactivation of the paternally-derived X chromosomes in mouse were the first demonstrations of a functional imprint in mammalian genomes [74, 46, 15]. The first suggestion of imprinting on autosomes was by a deletion on mouse chromosome 17 that showed a different phenotype based on which parent the deletion was inherited from [33, 64]. The development of the pronuclear transplantation technique allowed for the creation of mice zygotes which contained only maternal or only paternal genetic contributions and provided evidence that the maternal and paternal genomes are not equal. The differential imprinting on the parental chromosomes prevented complete embryonic development in these mice with complete uniparental disomy [67, 50].

Human Syndrome	Location	Major features
Pseudo-hypoparathyroidism type 1a and type 1b	20q13.3	Dysmorphism, obesity, cognitive impairment
Prader-Willi syndrome	15q11-13	Developmental delay, obesity, hypogonadism

Table 1.1: **Description of bacteria.** \*Mycobacteria are unable to be gram stained due to the low permeability of their cell walls. They are more closely related evolutionarily to gram-positive bacteria than gram-negative. However, their thick cell walls share features of gram-negative bacteria, e.g. a “pseudoperiplasm” similar to the gram-negative periplasm.

Further experiments suggested that imprinting occurs during gametogenesis and is necessary for full term development; an egg with a male pronucleus developed to term, however, an egg with two female pronuclei (gynogenetic embryos) or two male pronuclei (androgenetic) developed poorly[72, 50]. Non-complementation in genetic crosses of translocated chromosomes provided a way to refine the imprinted regions of the genome[14].

Genetic characterization of Prader-Willi syndrome (PWS) was the first human genetic disease to be associated with maternal heterodisomy of chromosome 15q11-13[52]. It suggested that clinical phenotype of PWS arises from the absence of paternal contribution of 15q11-13 as opposed to a specific genetic mutation. Conversely the absence of maternal contribution to the same region should result in Angelman syndrome (AS)[52, 63]. This provided more evidence that at “imprinted” regions the functional differences depend on the sex of the transmitting parent and genetic input from both parents are required for normal human development[52].

A rare disease, PWS affects between 1 in 10,000 and 30,000 people. PWS arises from loss of paternal genetic contribution at 15q11-13, mostly by chance mutation but also through uniparental disomy, sporadic mutations, chromosome translocations, and gene deletions. (wiki). On the other side of the spectrum, AS is also rare, affecting between 1 in 12,000 and 20,000 people and is caused by the loss of the normal maternal contribution at 15q11-13. (wiki) Various other human imprinted syndromes due to loss or gain of expression of imprinted genes have been characterized (Table 1.1). [60]

### 1.3 The Search for Parent of Origin Effects

Parent of origin effects and imprinted genes have been most elegantly studied in mice, where two inbred strains are bred reciprocally to identify parent of origin effects on gene expression in progeny that have the same genotypes but different patterns of inheritance[6] . Such studies are obviously more challenging in humans. Previous studies have attempted to identify parent of origin alleles using different approaches, addressing parent of origin effects on gene expression and traits, both case-control and quantitative.

For example, in one study of gene expression, investigators examined Hardy Weinberg Equilibrium estimates using genotypes derived from RNA-seq data and considered imprinted loci to be those with no or fewer than expected heterozygotes (C.T. Watson, ASHG 2014). Garg et al. used gene expression in LCLs from 29 CEU and 30 YRI HapMap trios to identify 30 imprinting eQTLs with parent of origin specific effects on expression by first comparing maternal alleles and paternal alleles associated with gene expression, and then comparing reciprocal heterozygotes[26]. A study from the GTEx Consortium used RNA-seq data to determine allele specific expression in 45 different tissues from various numbers of individuals to identify new imprinted genes[9]. By considering genes with monoallelic expression that were evenly distributed to both the reference and alternate alleles across individuals as evidence for imprinting, they identified 42 imprinted genes, both known and novel, and used family studies to confirm imprinting of 5 novel imprinted genes. Most recently, Santoni et al. identified nine novel imprinted genes using single-cell allele-specific gene expression and identifying genes with mono-allelic expression in fibroblasts from 3 unrelated individuals and probands of 2 family trios, and then using the trios to confirm parent of origin of the alleles[66].

Fewer have searched for parent of origin effects on case control and quantitative phenotype traits. In a study on 38,167 genotyped Icelanders with known status for 7 diseases, investigators identified variants that were associated with breast cancer when paternally in-

herited and variants associated with type 2 diabetes when maternally inherited[38]. Parent of origin associations with height in the same Icelandic population (n=88,835) identified four associations with height of which three were in known imprinted regions one of which was also shown in the Sardinia population[85].

Large pedigrees are ideal for identifying parent of origin effects[9]. The advantages that large family studies have for these studies include: 1) formally proving parent of origin effects detected from allele specific expression, 2) detecting subtle imprinting that does not lead to strictly monoallelic expression, and 3) providing a measure of sensitivity to detect those loci or regions that could not be detected in other studies.[9] The Hutterite population is ideally suited for these studies. The ~1,400 individuals studied by our group are related to each other in a 13-generation pedigree that includes 3,671 individuals, all of whom originated from only 64 founders. Ninety-eight Hutterites were initially selected for whole genome sequencing; alleles were phased using Affymetrix framework markers in the 98 individuals and then imputed to the remaining Hutterites who had been previously genotyped with the Affymetrix framework markers[42]. After quality control, parent of origin was assigned to more than 7 million variants. Of those individuals, 431 also have RNA-seq expression data from lymphoblastoid cell lines (LCLs), and between 600-1300 individuals have been phenotyped for cardiovascular disease (CVD) associated and asthma associated quantitative traits

# CHAPTER 2

## PARENT OF ORIGIN EFFECTS ON QUANTITATIVE PHENOTYPES IN A FOUNDER POPULATION

### 2.1 Abstract<sup>1</sup>

The impact of the parental origin of associated alleles in GWAS has been largely ignored. Yet sequence variants could affect traits differently depending on whether they are inherited from the mother or the father. To explore this possibility, we studied 21 quantitative phenotypes in a large Hutterite pedigree. We first identified variants with significant single parent (maternal-only or paternal-only) effects, and then used a novel statistical model to identify variants with opposite parental effects. Overall, we identified parent of origin effects (POEs) on 11 phenotypes, most of which are risk factors for cardiovascular disease. Many of the loci with POEs have features of imprinted regions and many of the variants with POE are associated with the expression of nearby genes. Overall, our results indicate that POEs, which can be opposite in direction, are relatively common in humans, have potentially important clinical effects, and will be missed in traditional GWAS.

### 2.2 Introduction

Genome-wide association studies (GWAS) typically treat alleles inherited from the mother and the father as equivalent, although variants can affect traits differently depending on whether they are maternal or paternal in origin. In particular, parent of origin effects (POEs) can result from imprinting, where epigenetic modifications allows for differential gene expression on homologous chromosomes that is determined by the parental origin of the chromosome. Mutations in imprinted genes or regions can result in diseases. For example,

---

1. Citation for chapter:

two very different diseases, Prader-Willi Syndrome (PWS) and Angelman Syndrome (AS), are due to loss of function alleles in genes within an imprinted region on chromosome 15q11-13. Inheriting a loss of function mutation for the *SNRPN* gene from the father results in PWS but inheriting a loss of function mutation for the *UBE3A* gene from the mother results in AS[60, 24]. Long noncoding RNA genes at this and other imprinted regions act to silence (i.e. imprint) genes in cis. Imprinted genes are often part of imprinted gene networks, suggesting regulatory links between these genes [58, 25, 76]. More than 200 imprinted loci have been described in humans [11] but there are likely many other, as yet undiscovered, imprinted loci.

Previous studies have utilized pedigrees to test maternal and paternal alleles separately for association with phenotypes or with gene expression to uncover new imprinted loci [38, 9, 26, 59, 11]. Kong *et al* [38] discovered one locus associated with breast cancer risk only when the allele is inherited from the father and another locus associated with type 2 diabetes risk only when the allele is inherited from the mother. Garg et al. reported parent-of-origin cis-eQTLs with known or putative novel imprinted genes affecting gene expression[26]. Two additional studies by Zoledziewsk et al. and Benonisdottir et al. identified opposite POEs on adult height at known imprinted loci[85, 11]. Both studies reported associations with variants at the *KCNQ1* gene, and one showed additional opposite POEs with height at two known imprinted loci (*IGF2-H19* and *DLK1-MEG3*)[11]These studies provide proof-of-principle that alleles at imprinted loci can show POEs, some with opposite effects, with common phenotypes.

Many existing studies and methods identify parent of origin effects use case/parent trios or case/mother duos[17, 32, 3, 80, 79]. Similar to Kong *et al.* [38], our method does not require data on the parent and only uses the parent of origin informative alleles which were assigned and phased using PRIMAL[42]. In contrast to Kong *et al.* [38] which used binary traits, our method tests for parent of origin effects on quantitative traits, similar to

Benonisdottir *et al.* [11] which tested for parent of origin effects on height.

No previous study has included a broad range of human quantitative phenotypes or has studied genome-wide variants with effects in different directions depending on the parent of origin. To address this possibility, we developed a statistical model that directly compares the effects of the maternal and paternal alleles to identify effects that are different, including those that are opposite. We applied this model in a study of 21 common quantitative traits that were measured in the Hutterites, a founder population of European descent for which we have phased genotype data [42]. We identified variants with maternally inherited or paternally inherited effects only and variants with opposite POEs. Some of the identified regions have characteristics similar to known imprinted genes. Overall, we show that this model can identify putative novel imprinted regions with POEs for a broad range of clinically relevant quantitative phenotypes.

## 2.3 Results

### 2.3.1 GWAS

We first performed standard genome-wide association studies (GWAS) of 21 traits in the Hutterites (Table S1). These studies identified one genome wide significant association ( $p < 5 \times 10^{-8}$ ) with each of five of the 21 traits: low density lipoprotein level (LDL)-cholesterol, triglycerides, carotid artery intima media thickness (CIMT), left ventricular mass index (LVMI), and monocyte count. The results of all 21 GWAS are summarized in Table S2 and Supplementary Figure 2.5. Results for all variants for all GWAS are deposited in dbGaP (phs000185 ? submission in progress).

### *2.3.2 Parent of Origin GWAS*

We considered two possible mechanisms of POEs. In the first, the effect size of one parent’s allele is close to zero and the effect size of the other parent’s allele is significantly different from zero. For these cases, we performed a paternal only or maternal only GWAS. In other cases, the maternal and paternal alleles may both have effect sizes different from zero, but the effects are significantly different from each other or opposite in direction. To detect these types of POEs, we developed a model that tests for differences between parental effects (see Methods). This model is especially powerful to identify variants with parental effects in opposite directions.

### *2.3.3 Maternal and Paternal GWAS*

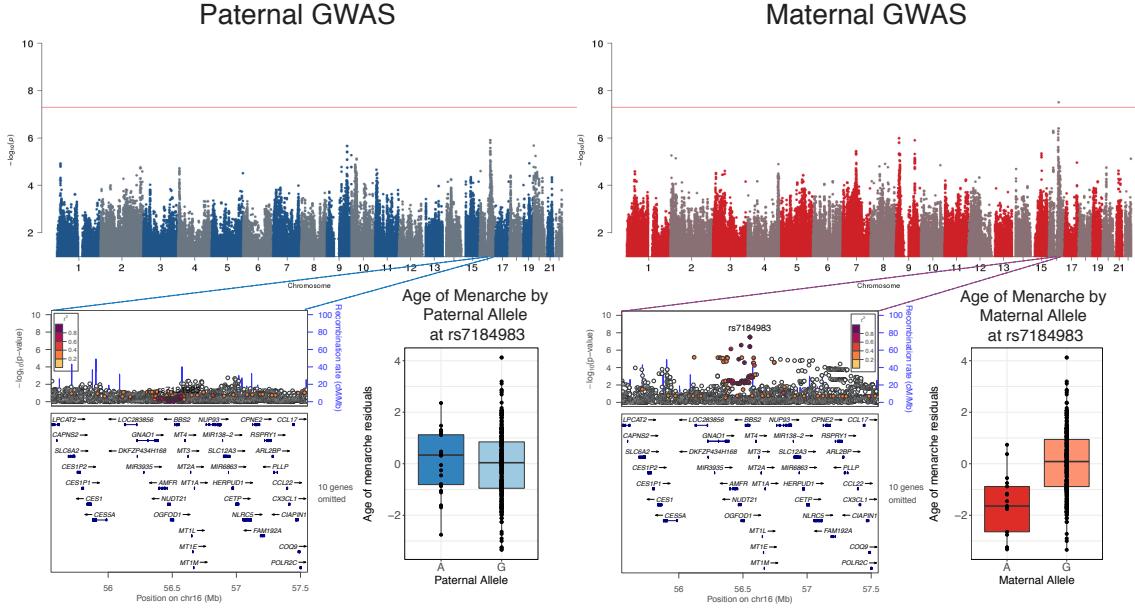
Using the same phenotypes, genotypes, pedigree, and criteria for significance as in the standard GWAS, we tested for maternal and paternal effects on each trait by testing each parentally inherited allele with the trait of interest, similar to previous studies[38, 85, 26]. Variants were considered to have POEs if they had a p-value less than  $5 \times 10^{-8}$  in only one parent and were not significant in the standard GWAS (i.e., the LDL association on chromosome 19 and the triglycerides association chromosome 11 were not considered to have POEs; see Table S1). The most significant parent of origin associations are summarized in Table 1. All significant results of the parent of origin GWAS for all 21 phenotypes are included in Table S5.

Overall, seven phenotypes had genome-wide significant parent of origin associations: four in the maternal only GWAS and three in the paternal only GWAS. Three cardiovascular disease (CVD)-associated phenotypes (age at menarche, CIMT, LVMI) and one lung function phenotype (forced expiratory volume in one second [FEV1]) were associated with maternally-inherited alleles only.

A maternally inherited allele at rs7184983 (G) on chromosome 16 was associated with

younger age of menarche ( $P = 3.11 \times 10^{-8}$ ) (Figure 2.1). This SNP, rs7184983, is located upstream of the *BBS2* gene and is associated with increased expression of *OGFOD1* in transformed fibroblast cells and tibial nerve [28]. The maternally inherited allele at rs4077567 (G) on chromosome 2 was associated with decreased CIMT ( $P = 3.02 \times 10^{-8}$ ) (Figure 2.6). This SNP is in the intron of a long intergenic noncoding gene, *LINC00607*, that is expressed in aorta, coronary, and tibial artery, all tissues potentially relevant to CIMT and atherosclerosis [28]. A maternally inherited allele at rs574232282 (G) in the intron of *SCMH1* on chromosome 1 was associated with increased LVMI ( $P = 1.39 \times 10^{-8}$ ) (Supplementary Figure 2.7). *SCMH1* is expressed in aorta, coronary, and tibial artery [28]. *SCMH1* protein associates with the polycomb group multiprotein complexes required to maintain the transcriptionally repressive state of certain genes[28]. Lastly, maternally inherited alleles at rs9849387 (A) and rs6791779 (C) on chromosome 3 were both associated with reduced FEV1 ( $P= 4.10 \times 10^{-9}$  and  $1.48 \times 10^{-8}$ , respectively) (Supplementary Figure 2.8). The nearest gene to rs9849387 is *ROBO2* (65kb, downstream), which is expressed in the lung as well as in brain, and ovary[28].The nearest gene to rs6791779 is *MIR4444-1*(267kb) whose expression has not been characterized.

Three other CVD-related phenotypes (systolic blood pressure, LDL-C, and total cholesterol) had associations with paternally inherited alleles only. The paternally inherited allele at rs12024326 (A) on chromosome 1 was associated with lower LDL-cholesterol levels ( $P = 8.06 \times 10^{-10}$ ) (Figure 2.2). rs12024326 is in the intron of gene *ADCK3*, and the same allele was associated with increased expression of *ADCK3* in whole blood, as well as decreased expression of a neighboring gene, *CDC42BPA* in brain (cerebellum), heart (left ventricle), esophagus, and tibial artery[28].The paternal G allele at rs4843650 on chromosome 16 was associated with increased LDL-C and is located in the intron of *JPH3*, which is expressed predominantly in the brain[28]. A SNP on chromosome 13 (rs1536182) was associated with systolic blood pressure levels when it was inherited from the father (Figure S5). The pater-

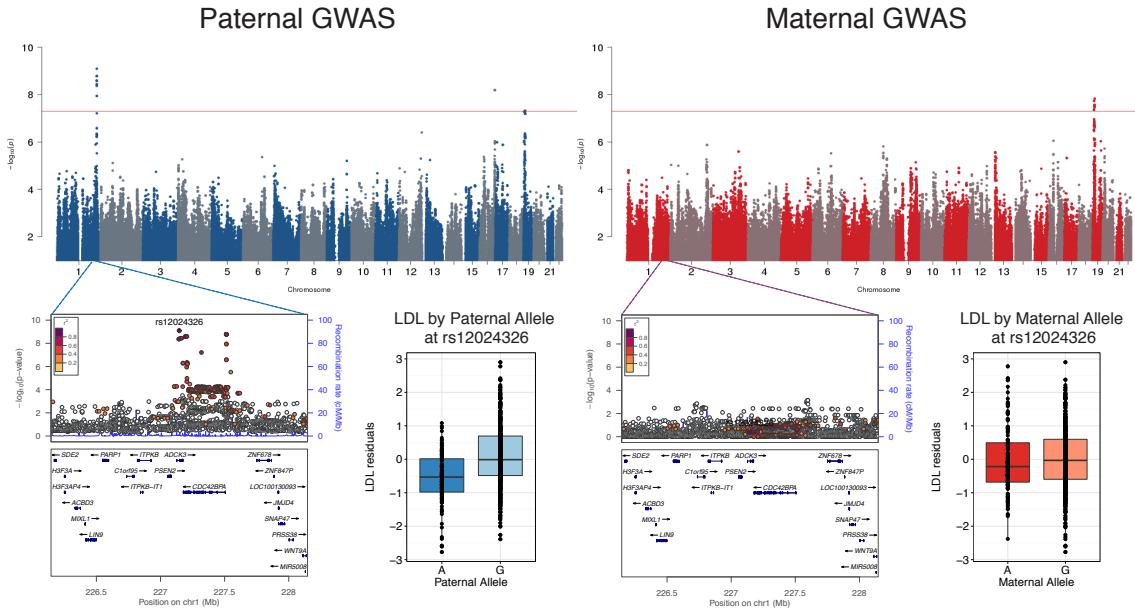


**Figure 2.1: Maternal and Paternal GWAS results for Age of Menarche.** The top panel shows the Manhattan plots from the maternal (left) and maternal (right) GWAS. LocusZoom plots for both GWAS are shown in the lower panel for the associated region in the GWAS. Boxplots show the distribution of age of menarche residuals (y-axes) by the corresponding maternal and paternal alleles at this SNP (x-axes). The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show  $\pm 1.5 \times \text{IQR}$ .

nally inherited A allele at this SNP was associated with decreased systolic blood pressure, as well as decreased expression of its closest gene, *LINC01055*, a long intergenic noncoding gene, in testis[28]. A paternally inherited allele at rs113588203 (G) on chromosome 1 was associated with lower total cholesterol ( $P = 1.76 \times 10^{-8}$ ) (Figure S6). This SNP is intergenic between *RHOU* (96kb, downstream), which is expressed across multiple tissues, and *MIRR4454* (331kb), which is expressed in adipose, kidney and heart tissues[28].

### 2.3.4 GWAS for Differential Parent of Origin Effects

Because some imprinted regions include genes that have both maternal or paternal specific tissue expression, we next tested for such differential effects with these 21 phenotypes. In these analyses, we compared the effect and direction of the association between maternal and

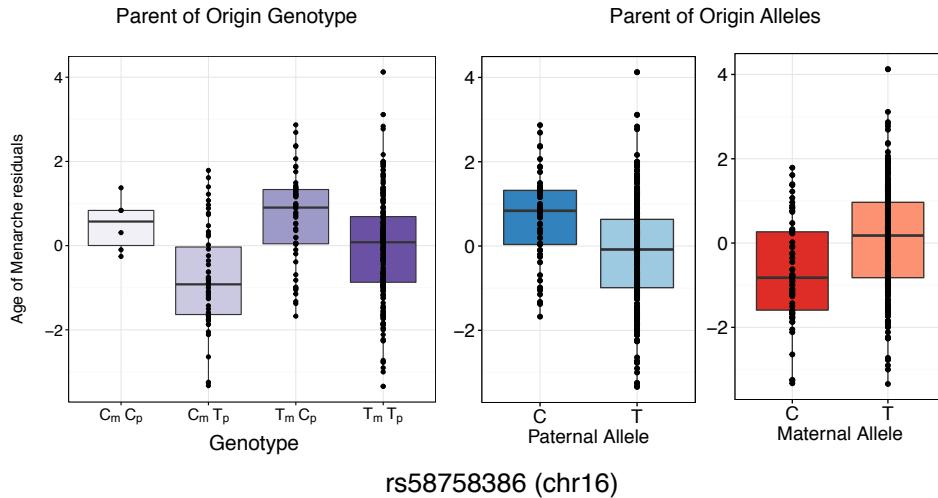


**Figure 2.2: Maternal and Paternal GWAS results for LDL Cholesterol.** The top panel shows the Manhattan plots from the maternal (left) and maternal (right) GWAS. LocusZoom plots for both GWAS are shown in the lower panel for the associated region in the GWAS. Boxplots show the distribution of LDL residuals (y-axes) by the corresponding maternal and paternal alleles at this SNP (x-axes). The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show  $\pm 1.5 \times$  IQR.

paternal alleles to identify variants that have different effects, including opposite effects, on the phenotype. Such loci would be completely hidden in standard GWAS in which paternally and maternally inherited alleles are combined. These opposite effect GWAS revealed 11 independent loci with opposite POEs for nine different traits, at least six of which are associated with CVD risk (Table 3, Figure S7).

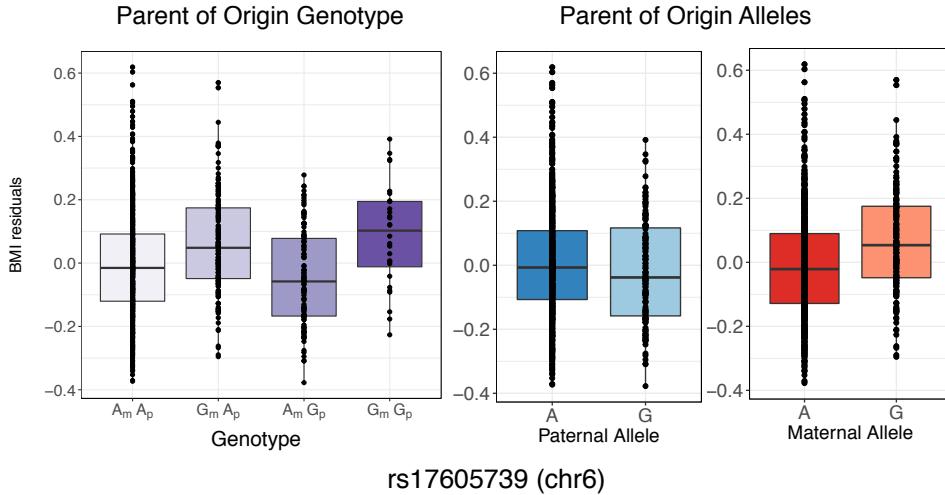
A locus on chromosome 16, near the *CDH8* gene (128kb, upstream), was associated with opposite POEs with age of menarche (Figure 2.3). *CDH8* is highly expressed in the brain, as well as in the aorta artery and pituitary gland. Two loci on chromosomes 5 and 6 were associated with opposite POEs on body mass index (BMI) (Figure 2.4). The most significant variant on chromosome 5 (rs77785972) is near a long intergenic noncoding gene, *LINC01340* (409kb, downstream), whose expression has not been well characterized.

The SNP on chromosome 6 (rs17605739) is also in a long intergenic noncoding gene, *RP1-209A6.1*, which is expressed in low levels in the tibial artery, bladder, spleen, lung, pituitary gland, as well as testis.



**Figure 2.3: Opposite Effect Parent of Origin GWAS Result for Age of Menarche.** Box plots of age of menarche residuals (y-axes) are shown for each of the four genotypes (left panel; x-axis), and for paternal (center panel; x-axis) and maternal (right panel; x-axis) alleles. The maternal C allele is associated with decreased and maternal T allele with increased age of menarche. The paternal C allele is associated with increased and the paternal T allele with decreased age of menarche. The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show +/-1.5 x IQR.

A SNP on chromosome 16 (rs1032596) was associated with opposite POEs on LDL-cholesterol (Figure S8). This SNP lies in the intron of another long noncoding RNA gene, *LINC01081*, which has been suggested to be imprinted because its downstream genes have also been shown to have parent- and tissue-specific activity[73]. A region on chromosome 2 has opposite effects associated with LVMI (Figure S9). The associated SNPs are in the intron of *XIRP2*, a cardiomyopathy associated protein that is expressed in skeletal muscle and heart left ventricle, suggesting that this gene could play a role in determining left ventricular mass[77, 54, 28]. In addition, the most significant SNP at this region, rs17616252 (and multiple SNPs in LD) is a strong eQTL ( $P = 1.8 \times 10^{-13}$ ) for the gene *XIRP2* in skele-



**Figure 2.4: Opposite Effect Parent of Origin GWAS Result for BMI.** Box plots of two significant loci plot BMI residuals (y-axes) for each of the four genotypes (left panel; x-axis), and for paternal (center panel; x-axis) and maternal (right panel; x-axis) alleles. For the (A) SNP on chromosome 5 the maternal A allele is associated with decreased and maternal G allele with increased BMI. The paternal A allele is associated with increased and the paternal G allele with decreased BMI. For the (B) SNP on chromosome 6 the maternal A allele is associated with decreased and maternal G allele with increased BMI. The paternal A allele is associated with increased and the paternal G allele with decreased BMI. The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show  $\pm 1.5 \times \text{IQR}$ .

tal muscle, *XIRP2-AS1* in testis, and *B3GALT1* in transformed fibroblast cells[28]. Four variants in a region on chromosome 1 in a microRNA gene, *MIR548F3*, were associated with opposite POEs on triglyceride levels (Figure S10). The expression of *MIR548F3* has not been characterized. SNP rs7033776 near *MELK* (27kb, downstream) on chromosome 9 was associated with opposite effects on total cholesterol (Figure S11). *MELK* is expressed in the colon and esophagus in addition to transformed lymphocytes and fibroblasts[28].

Nine linked variants on chromosome 1 were associated with opposite POEs of blood eosinophil count (Figure S12). These variants are near the gene *IGSF21* (27kb, downstream) which is a member of the immunoglobulin superfamily and likely acts as a receptor in immune response pathways[56]. A variant on chromosome 3, rs12714812, was associated with opposite POEs for FEV1 (Figure S13). This variant has been shown to regulate the expression of

a gene *CNTN3* (45kb, upstream) in heart and brain[28]. Studies in mice have suggested that this gene is imprinted and maternally expressed in the murine placenta[12]. Variant rs142030841 in the intron of the gene *TPGS2* on chromosome 18 has opposite POEs with neutrophil levels (Figure S14). This SNP is an expression quantitative trait locus (eQTL) for the noncoding RNA gene *RP11-95O2.5* in skin, testis, breast, thyroid and adipose tissue, for *CELF4* in tibial nerve and lung, and for *TPGS2* in tibial artery and transformed fibroblast cells[28].

### 2.3.5 Parent of Origin Effects on Gene Expression

To determine if any of the associated variants also showed POEs on gene expression in the Hutterites, we used RNA-seq gene expression data from lymphoblastoid cell lines (LCLs) collected from 430 of the individuals in the GWAS sample. We first tested for association of maternal and paternal variants with genes detected as expressed in the LCLs and whose transcript start site was within 1Mb of each associated SNP. There were no significant associations after multiple testing correction, similar to a previous study[11]. However, because we considered this to be exploratory analyses, we show results for the five most significant parent of origin eQTLs (Table 3). We next used the opposite effect model for each SNP in Table 2 and expression of all genes that were detected as expressed in LCLs and whose transcript start site was within 1Mb of the associated SNP. This resulted in 57 tests (1 SNP for each of 8 phenotypes, and 57 genes). The five most significant opposite effect eQTLs, none of which passed the Bonferroni threshold of  $8.77 \times 10^{-4}$ , are shown in Table 4. The most significant opposite effect eQTL was for *POLR1E* expression with the SNP on chromosome 9 (rs7033776) that was associated with total cholesterol (opposite effect eQTL  $P = 9.86 \times 10^{-4}$ ) (Figure S15). *POLR1E* is involved in the purine metabolism pathway as well as DNA-directed polymerase activity. The same SNP, rs7033776, had modest opposite effects with the expression of three other genes in the region (*PAX5*, *FBXO10*, and

FRMPD1), a signature consistent with an imprinted region. Another SNP with opposite POEs on LVMI, rs16853098, was an opposite effect eQTL for STK39, a gene that has been previously associated with hypertension[78].

## 2.4 Discussion

In this study, we introduced a novel statistical method that allows assessment of standard GWAS signals along with measures of differential POEs on common quantitative phenotypes. Similar to previous parent of origin studies of fewer phenotypes[38, 11, 26], we tested for associations of maternally- or paternally-derived alleles with each phenotype. We then extended this method to identify variants for which maternally- and paternally-derived alleles have different, including opposite, effects on phenotypic values. The focus on 21 common disease-associated phenotypes in a single large pedigree allowed us to broadly survey physiological effects of putative imprinted regions and the candidate genes at each associated locus. In contrast to previous studies, our new model can identify variants with opposite POEs that would be missed in traditional GWAS (Table 2).

Our studies of  $\approx$ 1,000 Hutterites who are related to each other in a single pedigree allowed us to detect POEs, even when few genome-wide significant associations were detected in standard GWAS of the same phenotypes. Our method revealed parent of origin specific genome-wide significant associations for seven of the 21 phenotypes examined, with maternally-inherited alleles associated with four phenotypes, paternally-inherited alleles with three phenotypes (Table 1), and opposite parent of origin alleles with nine phenotypes, of which five also showed single POEs at different loci (Table 2). Overall, 11 of the 21 phenotypes examined showed genome-wide significant evidence of POEs with alleles at one or more loci. In contrast, standard GWAS of these same phenotypes and using the same markers in these same individuals revealed genome-wide significant association for only five traits.

It is notable that four of the nine significant opposite parent of origin effects (one each

with LDL-C and triglycerides, and two with BMI) lie in or near long intergenic non-coding RNA genes (lincRNAs). LincRNAs are a feature of imprinted regions[60], where they can silence the expression of genes on the opposite chromosome[10, 58]. One of the variants, rs1032596, with an opposite parent of origin effect on LDL-C is located in the *LINC01081* gene. This noncoding RNA, along with *LINC01082*, regulates the *FOXF1* enhancer resulting in *FOXF1* parent- and tissue-specific activity[73] providing experimental support for tissue specific expression, a feature of imprinted regions.

Another variant with POEs in our study has been suggested to be imprinted in previously published work. The variant associated with opposite POEs for FEV1 is an eQTL for the gene *CNTN3*. *CNTN3* was shown to have exclusive maternal allele-specific expression in murine placentas[12], although this finding may have been due to contaminating maternal cells[55, 62].

Other regions associated with POEs harbor genes involved in transcriptional repression (e.g., *SCMH1* with LVMI on chromosome 1) or the associated SNPs are reported as eQTLs in GTEx with expression in tissues relevant to the phenotype under investigation (e.g., the LVMI-associated SNPs are eQTLs for *XIRP2*, which is expressed in skeletal muscle and heart left ventricle)[28]. Overall, these patterns of expression provide additional support that the parent of origin associations in our study are flagging imprinted regions or regions involved in the regulation of gene expression. Finally, we used gene expression in LCLs from the Hutterites to directly test for parent of origin eQTLs among SNPs associated with phenotypes in the parent of origin GWAS. Although none of the parent of origin eQTLs met criteria for significance after correcting for multiple testing, the SNP on chromosome 9 with opposite POEs on total cholesterol levels was borderline significant as an opposite parent of origin eQTL for *POLR1E*, and possible for three other genes at the same locus (*PAX5*, *FBXO10*, and *FRMPD1*). The presence of multiple genes with potential parent of origin expression patterns is further supportive of an imprinted locus. The availability

of gene expression only in LCLs from the Hutterites limits the inferences we can draw about effects on expression because imprinted regions are often tissue-specific and sometimes developmentally regulated[60, 24]. Despite this limitation, the fact that many of the SNPs associated with POEs on phenotypes are themselves eQTLs in relevant tissues (GTEx) and some are suggestive of having POEs on expression in LCLs from the Hutterites is generally supportive of the suggestion that some of the regions identified in this study are imprinted or have network interactions with imprinted genes[16] in humans. Additionally, our data suggest that loci with POEs influence a broad spectrum of quantitative phenotypes that are themselves risk factors for common diseases.

In particular, the discovery of POEs for eight traits that are associated cardiovascular disease risk is intriguing. These include metabolic phenotypes, such as BMI, total cholesterol, triglycerides, LDL, and age of menarche, that have indirect effects on cardiac health, as well as LVMI and CIMT, which more directly reflect cardiac health. Some of these phenotypes showed associations with paternally inherited alleles only (systolic blood pressure, LDL-C, total cholesterol), maternally inherited alleles only (LVMI, CIMT, and age at menarche), and/or with opposite effect variants (BMI, LDL-C, triglycerides, total cholesterol, LVMI, age at menarche). It has been suggested that genomic imprinting evolved in the mammalian lineage as a way to regulate maternally and paternally expressed genes in the placenta during pregnancy and modulate metabolic functions related to growth, where the parental interests may be in conflict ? paternal alleles favoring growth of the fetus at the expense of the mother while maternal alleles favor restricting resources to the fetus to ensure preservation of her nutritional needs [29, 10, 58]. Our data show some effects that are consistent with this theory. For example, three independent paternally inherited alleles on chromosome 1 are associated with increased LDL-C (Fig. 2) and total cholesterol (Figure S7); a paternal allele on chromosome 13 is also associated with increased systolic blood pressure (Figure S6). However, it is not always possible to interpret our results in light of this model, such

as the association of maternal allele on chromosome 2 with decreased CIMT (Figure S3), or the maternal allele on chromosome 16 associated with decreased age of menarche (Figure 2.1), which confers increased cardiovascular risk [13]. However, because many of the traits associated with POEs in this study were measured in adults, and none were measured in neonates, we are likely observing the downstream effects of processes that occurred in utero. Nonetheless, this kinship theory, or parent-conflict hypothesis, could account for the enrichment of parent of origin associations, particularly those with opposite effects, among metabolic and CVD-associated traits [60].

Finally, we note that the parent of origin GWAS for 21 phenotypes in the Hutterites revealed overall twice as many genome-wide significant loci compared to standard GWAS of the same phenotypes in the same individuals, suggesting that variation at imprinted loci may represent some of the "missing heritability" of these phenotypes and potentially for the disease for which they confer risk. This idea is consistent with observations in both mice and humans [39]. POEs in mice contribute disproportionately to the heritability of 97 traits, including those related to total cholesterol, weight, HDL, and triglycerides [51]. Exactly how much loci with POEs in humans contribute to phenotypic variation and disease risk overall remains to be determined, but our study provides compelling evidence that it is likely to be significant for many important traits.

## 2.5 Methods

### 2.5.1 Sample Composition

The individuals in this study have participated in one or more of our studies on the genetics of complex traits in the Hutterites [20, 81, 1]. The more than 1,500 Hutterites in our study are related to each other in a 13-generation pedigree including 3,671 individuals.

### *2.5.2 Genotype Data*

Variants detected in the whole genome sequences of 98 Hutterites were previously imputed to an additional 1,317 individuals using PRIMAL, a high-accuracy pedigree based imputation method [42]. PRIMAL was used to phase alleles and assign parent of origin for 83% of about 12 million autosomal SNPs. For these studies, we selected SNPs that had a MAF 1% and genotype call rate 85%. This yielded 5,891,982 autosomal SNPs. Parent of origin allele call rates differed among individuals and between phenotypes (Table S1

### *2.5.3 Phenotype Data*

We included 21 quantitative phenotypes that were previously measured in the Hutterites. Descriptions for each phenotype, as well as exclusion criteria, transformations, and covariates used with each phenotype in the GWAS, are available in the Supplementary Methods (Table S1

Descriptions for 18 of the 21 phenotypes can be found in Cusanovich et al [20]. The remaining three are described here. Height was measured in cm on a stadiometer with shoes removed. BMI was calculated using weight (kg, measured on scale) divided by height (m) squared. Age at menarche was collected retrospectively by interview.

### *2.5.4 GWAS*

We used a linear mixed model as implemented in GEMMA to test for genome wide association with 21 phenotypes using an additive model. We corrected for relatedness, as well as relevant covariates (Table S1

### 2.5.5 Maternal and Paternal GWAS

To evaluated the evidence for POEs, we tested maternal and paternal alleles separately with each phenotype, comparing phenotypic differences between the maternally inherited alleles and between the paternally inherited alleles. We used a linear mixed model as implemented in GEMMA, which allows us to correct for relatedness as a random effect, as well as sex, age, and other covariates as fixed effects [84].The linear mixed model for the parent of origin GWAS for testing maternal alleles and paternal alleles is shown in Equation 2.1 and Equation 2.2, respectively.

$$Y = W\alpha + X_M\beta_M + g + \epsilon \quad (2.1)$$

$$Y = W\alpha + X_P\beta_P + g + \epsilon \quad (2.2)$$

$n$  is the number of individuals,  $Y$  is an  $n \times 1$  vector of quantitative traits,  $W$  is an  $nc$  matrix of covariates (fixed effects) including intercept 1.  $\alpha$  is a  $c \times 1$  vector of covariate coefficients.  $X_M$  is an  $n \times 1$  vector of maternal alleles, and  $X_P$  an  $n \times 1$  vector of paternal alleles.  $\beta_M$  and  $\beta_P$  are the effect sizes of maternal and paternal alleles, respectively.  $g$  is a vector of genetic effects with  $g \sim N(0, A(\sigma_g)^2)$  where  $A$  is the genetic relatedness matrix;  $\epsilon$  is a vector of non-genetic effects with  $\epsilon \sim N(0, I(\sigma_e)^2)$ .

### 2.5.6 Differential Effect GWAS (PO-GWAS)

To test for a difference in the same allele inherited from each parent, including opposite effects, we re-parameterized the test model (Equation 2.3) from Garg et al[26]. The null model (Equation 2.4) is a standard GWAS model, ignoring parent of origin of alleles. The test model (Equation 2.3) is more significant when maternal and paternal alleles have differential effects on gene expression.

$$Y = W\alpha + X_M\beta_M + X_P\beta_P + g + \epsilon \quad (2.3)$$

This new model allows us to measure the difference in parental effect of the same allele when the genotype is a covariate in Equation 2.5.

$$Y = W\alpha + X_{MP}\beta_{MP} + g + \epsilon \quad (2.4)$$

$X_{PM}$  is a  $n \times 1$  vector of genotypes with possible values [ 0,1,2 ], equivalent to  $X_P + X_M \cdot (\beta_M - \beta_P)$  is the difference in parental effect size. If the difference in parental effect size is large and significantly different from 0 it suggests a parent of origin effect exists at this variant.  $((X_M - X_P))/2$  is a  $n \times 1$  vector of genotypes with possible values [-1,0,1].  $((\beta_P + \beta_M))/2$  is the average parental effect size that is captured in normal GWAS using genotypes. The average genotypes are added in as a covariate, with the average parental effect size the corresponding covariate coefficient. This differential effect GWAS was tested in GEMMA using BIMBAM format to use average genotype values[68].

### 2.5.7 Parent of Origin eQTL studies

RNA-seq data from LCLs were available from a previous study in the Hutterites[20]. For this study, sequencing reads were reprocessed as follows. Reads were trimmed for adaptors using Cutadapt (with reads  $\leq 5$  bp discarded) then remapped to hg19 using STAR indexed with gencode version 19 gene annotations[21, 48]. To remove mapping bias, reads were processed using WASP mapping pipeline[75]. Gene counts were collected using HTSeq-count[4]. VerifyBamID was used to identify sample swaps to include individuals that were previously excluded[36]. Genes mapping to the X and Y chromosome were removed; genes with a Counts Per Million (CPM) value of 1 (expressed with less than 20 counts in the sample with lowest sequencing depth) were also removed. Limma was used to normalize

and convert counts to log transformed CPM values [65]. Technical covariates that showed a significant association with any of the top principal components were regressed out (RNA Integrity Number and RNA concentration).

### *2.5.8 Maternal and Paternal Parent of Origin eQTL*

LCL RNA-seq data was used to test the single parent model for the most significant SNP from the maternal or paternal only GWAS for each phenotype. We selected all genes detected as expressed in the LCLs and residing within 1Mb of each most significant associated SNP. Summary of the SNPs and genes tested are in Table S3.

### *2.5.9 Differential Parent of Origin eQTL*

LCL RNA-seq data was used to test the opposite effect model for the most significant SNP in each region that was associated with a phenotype in the parent of origin opposite effects GWAS. We selected all genes detected as expressed in the LCLs and residing within 1Mb of each associated SNP. Summary of the SNPs and genes tested are in Table S4.

## **2.6 Supplementary Information**

### *2.6.1 Supplementary Figures*

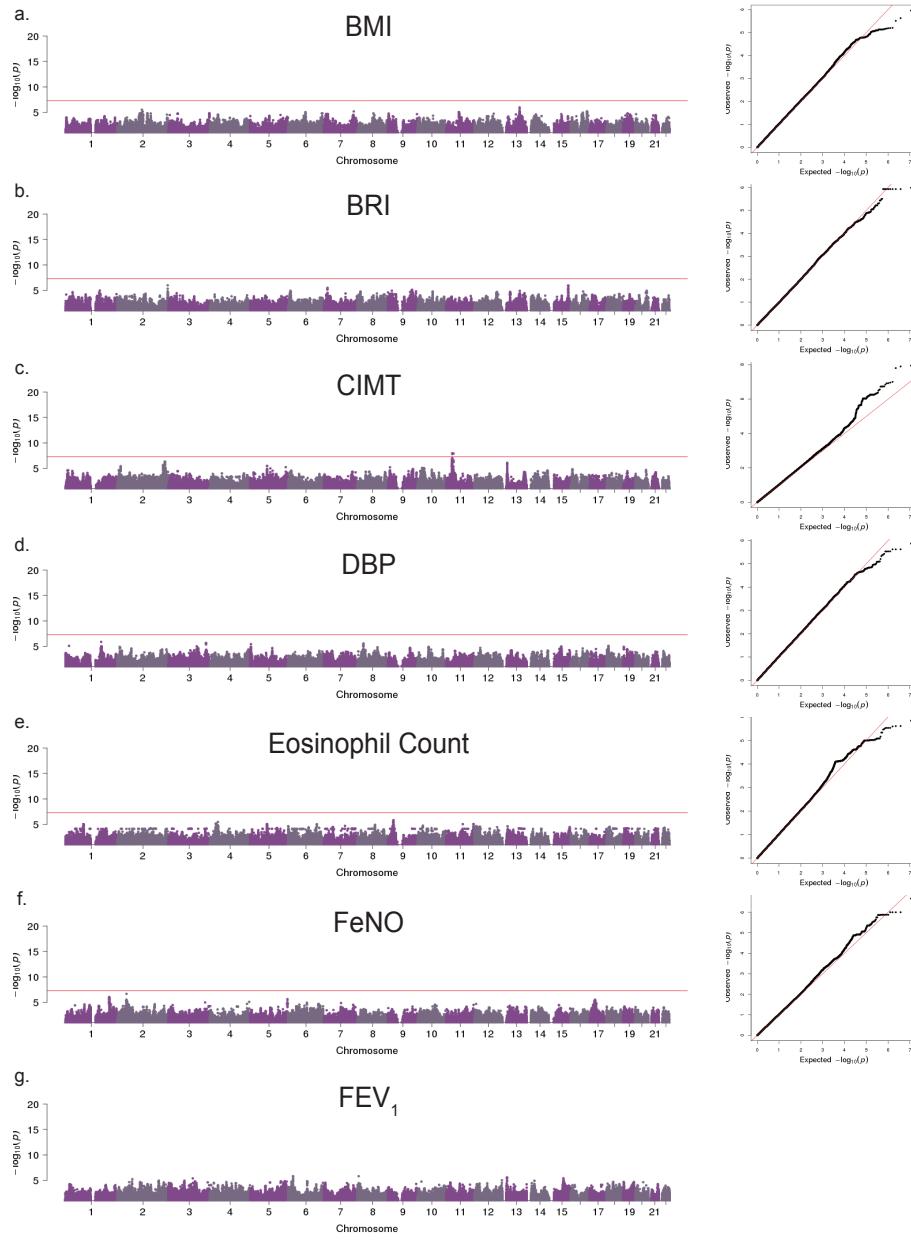
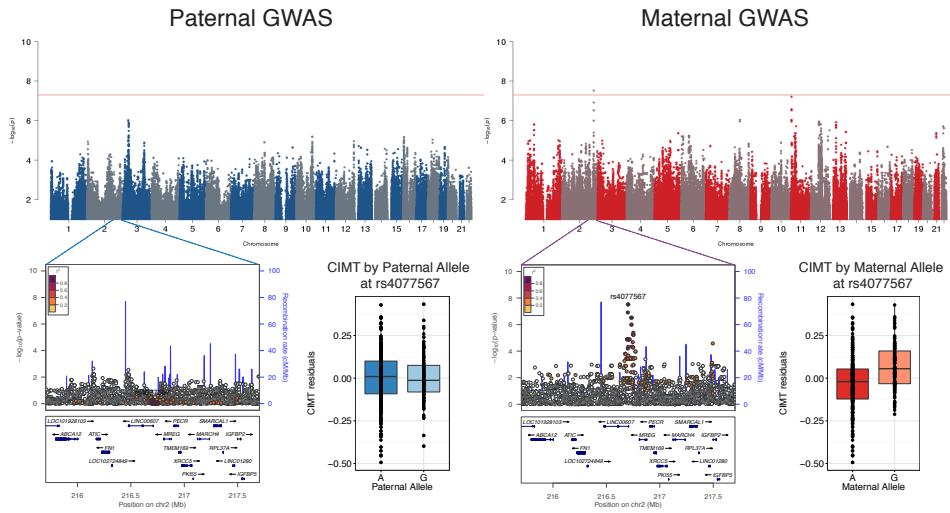
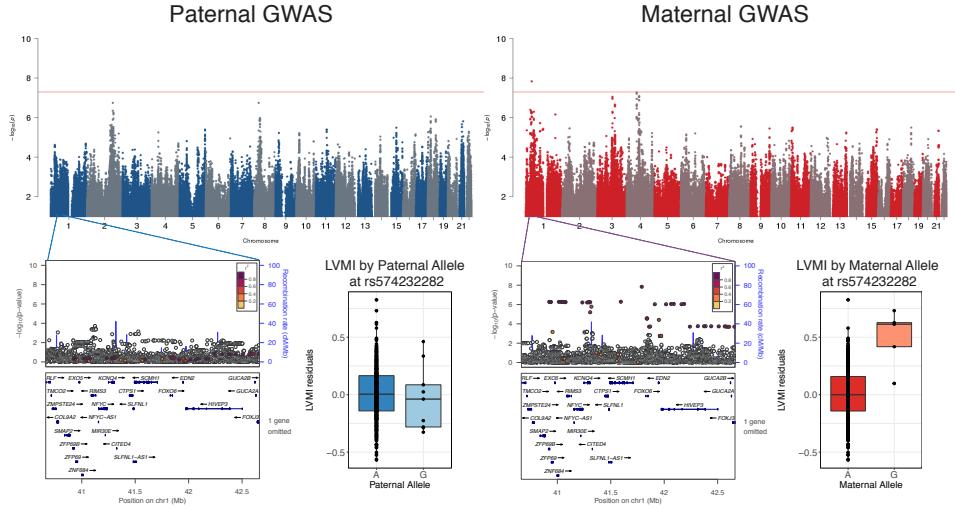


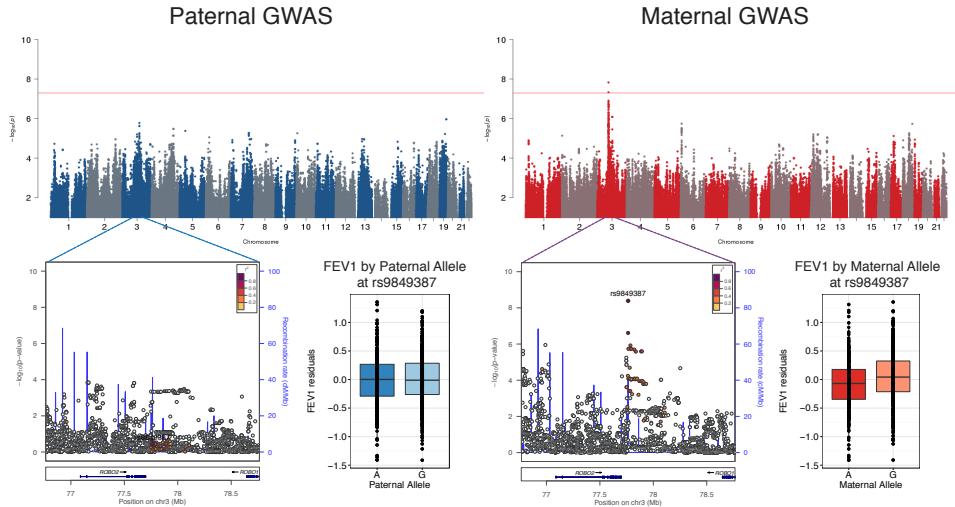
Figure 2.5: Manhattan and QQ Plots from Standard GWAS of 21 Quantitative Phenotypes



**Figure 2.6: Maternal and Paternal GWAS results for CIMT.** The top panel shows the Manhattan plots from the paternal (left) and maternal (right) GWAS. LocusZoom plots are shown in the lower panel for the associated region in the GWAS. Boxplots show the distribution of CIMT residuals (the residuals correspond to the inverse of raw CIMT values) (y-axes) by the corresponding maternal and paternal alleles at this SNP (x-axes). The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show  $\pm 1.5 \times$  IQR.



**Figure 2.7: Maternal and Paternal GWAS results for LVMI.** The top panel shows the Manhattan plots from the paternal (left) and maternal (right) GWAS. LocusZoom plots are shown in the lower panel for the associated region in the GWAS. Boxplots show the distribution of LVMI residuals (the residuals correspond to the inverse of raw CIMT values) (y-axes) by the corresponding maternal and paternal alleles at this SNP (x-axes). The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show  $\pm 1.5 \times \text{IQR}$ .



**Figure 2.8: Maternal and Paternal GWAS results for FEV1.** The top panel shows the Manhattan plots from the paternal (left) and maternal (right) GWAS. LocusZoom plots for both GWAS are shown in the lower panel for the associated region in the GWAS. Boxplots show the distribution of FEV1 residuals (y-axes) by the corresponding maternal and paternal alleles at this SNP (x-axes). The horizontal bar of the boxplot shows the median, the box delineates the first and third quartile, and the whiskers show  $1.5 \times \text{IQR}$ .

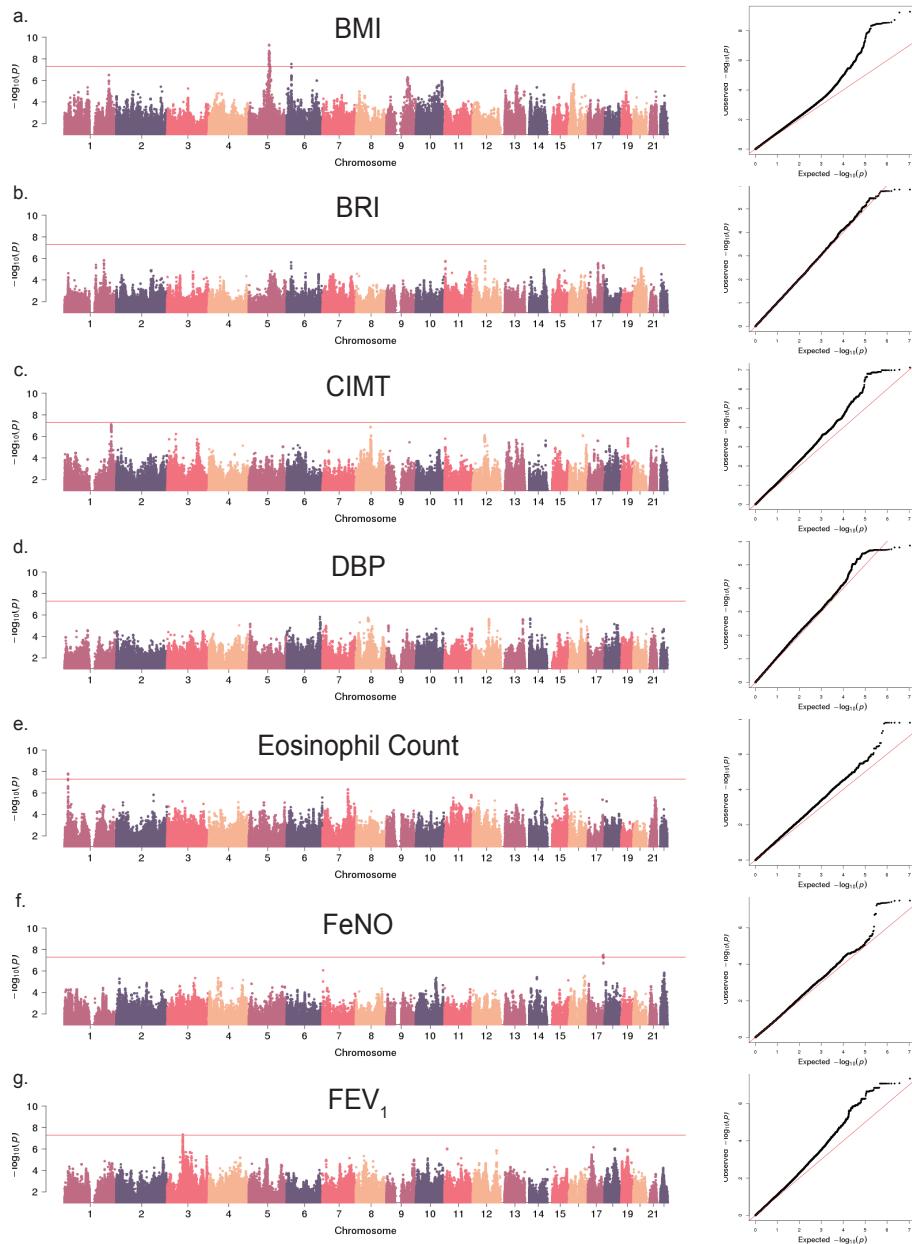


Figure 2.9: Manhattan and QQ Plots from Differential Effect GWAS of 21 Quantitative Phenotypes

# **CHAPTER 3**

## **PARENT OF ORIGIN GENE EXPRESSION IN A FOUNDER POPULATION IDENTIFIES TWO NOVEL IMPRINTED GENES AT KNOWN IMPRINTED REGIONS.**

### **3.1 Abstract<sup>1</sup>**

Genomic imprinting is the phenomena that leads to silencing of one copy of a gene inherited from a specific parent. Mutations in imprinted regions have been involved in diseases showing parent of origin effects, such as Prader-Willi and Angelman syndrome, among others. Identifying genes with evidence of parent of origin expression patterns in family studies allows the detection of more subtle imprinting. Here we use allele specific expression in lymphoblastoid cell lines from 306 Hutterites related in a single pedigree to provide formal evidence for parent of origin effects. Our approach identified known imprinted genes, two putative novel imprinted genes, and 12 genes with asymmetrical parent of origin gene expression. We used gene expression in peripheral blood leukocytes (PBL) to validate our findings, and then confirmed imprinting control regions (ICRs) using DNA methylation levels in the PBLs.

### **3.2 Introduction**

Imprinted genes have one allele silenced in a parent of origin specific manner. In humans, approximately 105 imprinted loci have been identified, many of which play important roles in development and growth[24, 60]. Dysregulation of imprinted genes or regions can cause diseases that show parent of origin effects, such as Prader-Willi or Angelman syndrome, among others[60]. Imprinted regions have also been associated with complex traits, such as height and age of menarche [11, 85], as well as common diseases such as obesity and some

---

1. Citation for chapter:

cancers [60]. More than 80% of imprinted genes in humans are clustered in genomic regions that contain both maternally and paternally expressed genes, as well as genes that encode non-coding RNAs. Parent-specific expression of the genes within a cluster are maintained by complex epigenetic mechanisms at cis-acting imprinting control regions (ICRs) [37], which show parent of origin specific DNA methylation patterns and chromatin modifications.

Using RNA-seq and allele specific expression (ASE) we can map genes to parental haplotypes and identify those that are expressed when inherited from only the father or only from the mother, a hallmark feature of imprinted loci. Parent of origin effects and imprinted genes have been most elegantly studied in mice, where two inbred strains are bred reciprocally to identify parent of origin effects on gene expression in progeny that have the same genotypes but different patterns of inheritance[7]. Additionally, uniparental inheritance of imprinted regions in mice were associated with abnormal developmental phenotypes[14] before it was shown that imprinting defects are associated with human disease[52]. One approach to identifying imprinted loci in humans has been to test for parent of origin effects on gene expression and phenotypes in pedigrees[38, 11]. For example, Garg et al. used gene expression in LCLs from HapMap trios to identify 30 imprinting eQTLs with parent of origin specific effects on expression including two imprinted genes[26]. A study from the GTEx Consortium used RNA-seq data and allele specific expression to identify allelic imbalance in 45 different tissues. By considering genes with monoallelic expression that was evenly distributed to both the reference and alternate alleles across individuals as evidence for imprinting, they identified 42 imprinted genes, both known and novel, and used family studies to confirm imprinting of 5 novel imprinted genes[9]. Santoni et al. identified nine novel imprinted genes using single-cell allele-specific gene expression and identifying genes with mono-allelic expression in fibroblasts from 3 unrelated individuals and probands of 2 family trios, and then using the trios to confirm parent of origin of the alleles[66].

Here, we perform a parent of origin ASE study in a large pedigree to characterize parent

of origin specific gene expression in the Hutterites, a founder population of European descent, for which we have phased genotype data[42]. We use RNA-seq from lymphoblastoid cell lines (LCLs) to map transcripts to parental haplotypes and identify known and two not previously reported imprinted genes. We validated the two putative imprinted genes by showing the same patterns of parent of origin expression PBLs from different Hutterite individuals, and show DNA methylation signatures of imprinting in the PBLs at these imprinted regions.

### 3.3 Results

#### 3.3.1 *Mapping Transcripts to Parental Haplotypes*

For each of 306 individuals, the total number of transcripts at each gene was assigned as maternally inherited, paternally inherited, or unknown parent of origin. The last group included transcripts without heterozygote SNPs or SNPs without parent of origin information. Transcripts were assigned to the parentally inherited categories using SNPs in the reads and matching alleles to either the known maternally or paternally inherited alleles. All the genes analyzed had some transcripts of unknown origin (average 97.8%, range 8.3-100%). For each gene we assigned parental origin to an average of 1.8% of transcripts (range: 0-34.7%), and for each individual we assigned parental origin to an average of 1.4% of transcripts (range: 0-1.7%). On average, about 40 SNPs per gene were used to assign the transcripts of a gene to parent (range 1-1839 SNPs).

After quality control (see Methods), transcripts in 15,889 genes were detected as expressed in 306 individuals. Transcripts for 14,791 of those genes could be assigned to a parent. Of these, 75 genes were only expressed on the paternally-inherited allele in at least one individual and not on the maternally inherited allele in anyone. Similarly, 64 genes were only expressed on the maternally-inherited allele in at least one individual and not on the paternally inherited allele in any individuals (S1 Table).

### 3.3.2 Imprinted Genes in Lymphoblastoid Cell Lines (LCLs)

Among the 139 genes with only paternally inherited expression or only maternally inherited expression, there are three known imprinted genes (*CDKN1C*, *NDN*, *SNRPN*) and one predicted to be imprinted (*IFITM1*)[44]. *CDKN1C* showed patterns opposite of what has been reported[30, 49], which could be due to the small sample (only three individuals showed expression from one parent) or to the different cell types used here (LCLs) and in previous studies (developing brain and embryonal tumors for *CDKN1C*).

We expect some imprinted genes to have ?leaky? expression, such that there is some expression from the parental chromosome that is mostly silenced. To detect these genes, we used a binomial test to find patterns of gene expression asymmetry by parental transcript levels. This analysis identified 28 genes with an FDR  $\geq 5\%$  (Table 2). The 11 genes that showed the most asymmetry are known imprinted genes: *ZDBF2*, *PEG10*, *SNHG14*, *NHP2L1*, *L3MBTL1*, *ZNF331*, *LPAR6*, *FAM50B*, *KCNQ1*, *NAP1L5*, and *IGF1R*. Parent of origin expression for *ZDBF2* and *KCNQ1* are shown in Fig 1A and 1B, respectively. We identified two additional genes that showed asymmetry in parental expression from mostly one parent (*PXDC1*, *PWAR6*), which we consider potentially novel imprinted genes. The remaining fifteen genes showed significant patterns of asymmetry but had expression from both maternal and paternal transcript levels. These genes are likely not imprinted but could have asymmetry in expression that could be due to an expression quantitative trait loci (eQTL).

Two genes showed gene expression signatures consistent with imprinting but have not previously been recognized as imprinted genes. The first potentially novel imprinted gene is *PXDC1*, which is in the same region and next to ( $\geq 100\text{kb}$ ) a known imprinted gene, *FAM50B*. The second potentially novel imprinted gene is *PWAR6*, or Prader Willi Angelman Region RNA6, a gene encoding regulatory class of RNA. Although this gene is located within the intron of a known imprinted region, *SNHG14*, this noncoding RNA has not previously been

recognized as having parent of origin specific expression (Fig 1C).

The remaining fifteen genes show significant asymmetry using the binomial test but do not have expression from mostly one parental chromosome. One of these genes, *SNHG17*, is a noncoding RNA. Another gene with parent of origin asymmetry, *ZNF813*, is next to a known imprinted gene, *ZNF331*. The remaining genes with asymmetrical parent origin expression have almost equal expression on both parental chromosomes. These genes include *DAAM1*, which is involved in cytoskeleton, specifically filopodia formation [31, 45], and has a suggested role for cytoskeleton organization during Mammalian testis morphogenesis and gamete progression [57]; *RP11-379H18.1*, a noncoding RNA gene; *HMGN1P38* [70]; *MTX2*, a nuclear gene that interacts with mitochondrial membrane protein metaxin 1 and is involved in mitochondrial protein import and metabolism of proteins in mice; *MAF1* a negative regulator of RNA polymerase 2; *ZNF714*, *CPNE1*, *IL16*, *ATP6V0D1*, *FAHD1*, *HSP90AB3P*, and *CNN2* are the remaining genes that show parent of origin asymmetry but not with a pattern consistent with imprinting (S1 Figure).

### 3.3.3 Validation of Imprinted Genes in PBLs

Using the same methods described above, we assigned parent of origin to transcripts in PBLs from 99 Hutterite individuals not included in the LCL studies. Maternal and paternal expression in PBLs for all 28 genes identified in LCLs showed similar trends of asymmetry as in LCLs (Figure 2).

### 3.3.4 Methylation at Imprinting Control Regions

One of the mechanisms underlying parent of origin effects on expression at imprinted loci is differential methylation at cis-acting imprinting control regions (ICRs). DNA methylation from the Illumina HumanMethylation 450K array was available in PBLs from the same individuals included in the validation studies described above. To determine the expected

patterns of methylation at known imprinted loci, we first looked at previously characterized methylated regions at known imprinted regions from Court et al. and Joshi et al. [18, 35].

The methylation patterns at the two potentially novel imprinted genes identified in this study, *PXDC1* and *PWAR6*, lie in or near known imprinted regions that contain previously characterized ICRs. Previously characterized ICRs near show about 50% methylation (beta value of between 0.25 and 0.75) in our DNA methylation data, which likely reflect methylation at only one parental chromosome in all the cells in the sample. Methylation patterns in PBLs at these two ICRs fall within this hemi-methylation range, further suggesting that these two genes are indeed imprinted (Fig 3).

### 3.4 Discussion

Discussion Dysregulation of imprinted genes can have a large impact on mammalian development and has been associated with significant diseases in humans. Studies aimed at identifying imprinted genes at genome-wide levels have used allele specific expression and imbalance to infer parent of origin. Here we used a large pedigree with assigned parent of origin alleles to map transcripts to chromosomes with known parent of origin and identify imprinted genes.

Using this approach, we found genes with expression primarily from either the maternal or paternal haplotype. Because gene silencing at imprinted loci may be incomplete, we used a binomial test on parent of origin gene expression and identified 11 known imprinted genes and two potentially novel imprinted genes. Both of these novel genes, *PWAR6* and *PXDC1*, lie in known imprinted regions but have not themselves been characterized as imprinted. The remaining genes that have significant parent of origin asymmetry in gene expression do not show clear imprinting expression patterns. To validate these findings, we mapped gene expression in PBLs from Hutterite individuals not included in the LCL study. The same genes showed similar patterns of asymmetry in these different cell sources (transformed B

cells and peripheral blood leukocytes) from different individuals.

In addition to validating gene expression, we characterized methylation patterns near genes showing asymmetry. Using results from studies that had previously characterized ICRs in patients with uniparental disomy at many imprinted regions [35, 18], we estimated regions for defining hemi-methylation near the genes identified in our study. Using this approach, we were able to provide additional supportive data for the two potentially novel imprinted genes to be true imprinted genes regulated by previously characterized ICRs.

Although our study is the largest pedigree-based study to date to search genome-wide for imprinted genes, it has limitations. First, we are able to determine the parent of origin for many transcripts in the Hutterites but we could not assign every RNA sequencing read to a parent due to lack of heterozygous sites or missing parent of origin information for alleles. Second, we conducted these studies in lymphoblastoid cell lines, and therefore could only study genes imprinted in this cell type and would miss the many imprinted genes that are tissue-specific and/or developmentally regulated. Third, while we can verify previously characterized ICRs, our study is not designed to identify novel ICRs because DNA methylation values from an array cannot be assigned to parental haplotype. Lastly, although we characterized the gene expression and methylation patterns for two potentially novel imprinted genes, replication of these genes in a different population and in different tissues, and functional characterization of these genes are required to confirm their status as imprinted genes. Similarly, some of the other genes with parent of origin asymmetry in the blood cells examined in this study may show more clear-cut evidence for imprinting in other tissues or at specific periods of development.

In summary, we have identified novel imprinted genes using gene expression from a founder population. The genes with asymmetrical parental expression had similar patterns of asymmetry in a different source of blood cells and in different individuals, and we were able to replicate the methylation patterns in known ICRs near the known and novel im-

printed genes in this study. Our method and study population allowed us to map reads to parental haplotypes and uncovered *PWAR6* and *PXDC1* as novel imprinted genes that could potentially impact disease risk and development.

## 3.5 Methods

### 3.5.1 Genotypes

Hutterite individuals (n=1,653) were genotyped using one of three Affymetrix genotype arrays, as previously described[? ], of which 121 underwent whole genome sequencing by Complete Genomics, Inc (CGI) (n=98) or Illumina whole genome sequencing (n=27). A total of 10,235,233 variants present in the sequenced individuals were imputed and phased to the remaining 1532 genotyped individuals using PRIMAL[? ]. Parent of origin was assigned to 89.85% of the alleles with call rate 81.6842% after QC. For this study, we included individuals with genotyped parents in the primary analyses in LCLs. Written consents for these studies were obtained from the adult participants and parents of children under 18; written assents were obtained from all children. This study was approved by the University of Chicago Institutional Review Board.

### 3.5.2 RNA-seq in Lymphoblastoid Cell Lines (LCLs).

RNA-seq was performed in LCLs as previously described [20]. For this study, sequencing reads were reprocessed as follows. Reads were trimmed for adaptors using Cutadapt (reads less than 5 bp discarded) then remapped to hg19 using STAR indexed with gencode version 19 gene annotations[21, 48]. To remove mapping bias, reads were processed and duplicate reads removed using WASP [75]. We used a custom script modified from WASP to separate reads that overlap maternal alleles or paternal alleles. Reads without informative SNPs (homozygous, or no parent of origin information) were categorized as unknown where the

unknown, maternal, and paternal make up the total gene expression. Gene counts were quantified using STAR for each category. VerifyBamID was used to identify sample swaps [36]. Genes mapping to the X and Y chromosome were removed; genes with a CPM log transformed value less than 1 in less than 20 individuals were also removed.

### *3.5.3 RNA-seq in Peripheral Blood Leukocytes (PBLs)*

RNA-seq was performed in whole blood as previously described [69]. For this study, sequencing reads were reprocessed as described above for the studies in LCLs. For all analyses, we excluded 32 individuals who were also in the LCL study.

### *3.5.4 Identifying Imprinted Genes*

We used a binomial test to detect asymmetry in parent of origin gene expression. We generated a binomial Z-score for each individual for each gene ( $Z_i$ ) and excluded those where  $Z_i = 0$ . For each gene, the number of subjects with  $Z_i > 0$  can be modeled by a Binomial distribution with probability 1/2. For imprinted genes that show patterns of asymmetry, we expect a distribution of Z-scores that are skewed to one direction: right-skewed for genes asymmetrically maternally expressed and left-skewed for genes asymmetrically paternally expressed. Because we are only asking whether there are more individuals with more maternal expression or more paternal expression and not gene expression measures there is no need to model over-dispersion.

### *3.5.5 DNA methylation profiling and processing in PBLs*

One milliliter of whole blood from 145 Hutterites was drawn into TruCulture (Myriad RBM; Austin, Texas) tubes containing proprietary TruCulture media. DNA was extracted using AllPrep DNA/RNA Mini Kits (Qiagen). DNA samples were bisulfite converted and hybridized to the Illumina HumanMethylation 450K array at the University of Chicago

Functional Genomics Center. Samples were processed using default parameters using the R package minfi [5], normalized using SWAN (subset within-array normalization [47]) and quantile normalized similar to previous methylation studies [53]. Probes were removed if: (1) mapped non-uniquely to a bisulfite-converted genome; (2) mapped to sex chromosomes; (3) had a probe detection p-value  $\geq 0.01$  in at least 25% of samples; and (4) contained common SNPs within the probe sequence, as previously described[8]. Principal components analysis (PCA) was used to identify significant technical covariates, and the ComBat function [34] within the R package sva [41] was used to correct for chip effect. Analyses of DNA methylation levels were conducted using beta values, which were converted from M-values using the lumi R package[22].

## References

- [1] Mark Abney, Mary Sara McPeek, and Carole Ober. Broad and Narrow Heritabilities of Quantitative Traits in a Founder Population. *The American Journal of Human Genetics*, 68(5):1302–1307, May 2001.
- [2] Lara K Abramowitz and Marisa S Bartolomei. Genomic imprinting: recognition and marking of imprinted loci. *Current opinion in genetics & development*, 22(2):72–78, April 2012.
- [3] Holly F Ainsworth, Jennifer Unwin, Deborah L Jamison, and Heather J Cordell. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genetic Epidemiology*, 35(1):19–45, December 2010.
- [4] S Anders, P T Pyl, and W Huber. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015.
- [5] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, May 2014.
- [6] Tomas Babak. Identification of imprinted loci by transcriptome sequencing. *Methods in molecular biology (Clifton, N.J.)*, 925:79–88, 2012.
- [7] Tomas Babak, Brian DeVeale, Emily K Tsang, Yiqi Zhou, Xin Li, Kevin S Smith, Kim R Kukurba, Rui Zhang, Jin Billy Li, Derek van der Kooy, Stephen B Montgomery, and Hunter B Fraser. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature genetics*, 47(5):544–549, May 2015.
- [8] Nicholas E Banovich, Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F Degner, John D Blischak, Julien Roux, Jonathan K Pritchard, and Yoav Gilad. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS genetics*, 10(9):e1004663, September 2014.
- [9] Yael Baran, Meena Subramaniam, Anne Biton, Taru Tukiainen, Emily K Tsang, Manuel A Rivas, Matti Pirinen, Maria Gutierrez-Arcelus, Kevin S Smith, Kim R Kukurba, Rui Zhang, Celeste Eng, Dara G Torgerson, Cydney Urbanek, GTEx Consortium, Jin Billy Li, Jose R Rodriguez-Santana, Esteban G Burchard, Max A Seibold, Daniel G MacArthur, Stephen B Montgomery, Noah A Zaitlen, and Tuuli Lappalainen. The landscape of genomic imprinting across diverse adult human tissues. *Genome Research*, 25(7):927–936, July 2015.
- [10] D P Barlow and M S Bartolomei. Genomic Imprinting in Mammals. *Cold Spring Harbor Perspectives in Biology*, 6(2):a018382–a018382, February 2014.

- [11] Stefania Benonisdottir, Asmundur Oddsson, Agnar Helgason, Ragnar P Kristjansson, Gardar Sveinbjornsson, Arna Oskarsdottir, Gudmar Thorleifsson, Olafur B Davidsdottir, Gudny A Arnadottir, Gerald Sulem, Brynjar O Jensson, Hilma Holm, Kristjan F Alexandersson, Laufey Tryggvadottir, G Bragi Walters, Sigurjon A Gudjonsson, Lucas D Ward, Jon K Sigurdsson, Paul D Iordache, Michael L Frigge, Thorunn Rafnar, Augustine Kong, Gisli Masson, Hannes Helgason, Unnur Thorsteinsdottir, Daniel F Gudbjartsson, Patrick Sulem, and Kari Stefansson. Epigenetic and genetic components of height regulation. *Nature Communications*, 7:13490, November 2016.
- [12] C M Brideau, K E Eilertson, J A Hagarman, C D Bustamante, and P D Soloway. Successful Computational Prediction of Novel Imprinted Genes from Epigenomic Features. *Molecular and Cellular Biology*, 30(13):3357–3370, June 2010.
- [13] D Canoy, V Beral, A Balkwill, F L Wright, M E Kroll, G K Reeves, J Green, B J Cairns, and for the Million Women Study Collaborators\*. Age at Menarche and Risks of Coronary Heart and Other Vascular Diseases in a Large UK Cohort. *Circulation*, 131(3):237–244, January 2015.
- [14] B M Cattanach and M Kirk. Differential activity of maternally and paternally derived chromosome regions in mice. *Nature*, 315(6019):496–498, June 1985.
- [15] H S Chandra and S W Brown. Chromosome imprinting and the mammalian X chromosome. *Nature*, 253(5488):165–168, January 1975.
- [16] Andrew Chess. Monoallelic Gene Expression in Mammals. *Annual Review of Genetics*, 50(1):317–327, November 2016.
- [17] Trees-Juen Chuang, Yu-Hsiang Tseng, Chia-Ying Chen, and Yi-Da Wang. Assessment of imprinting- and genetic variation-dependent monoallelic expression using reciprocal allele descendants between human family trios. *Scientific Reports*, 7(1):1–12, July 2017.
- [18] Franck Court, Chiharu Tayama, Valeria Romanelli, Alex Martin-Trujillo, Isabel Iglesias-Platas, Kohji Okamura, Naoko Sugahara, Carlos Simón, Harry Moore, Julie V Harness, Hans Keirstead, Jose Vicente Sanchez-Mut, Eisuke Kaneki, Pablo Lapunzina, Hide-nobu Soejima, Norio Wake, Manel Esteller, Tsutomu Ogata, Kenichiro Hata, Kazuhiko Nakabayashi, and David Monk. Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Research*, 24(4):554–569, April 2014.
- [19] H V Crouse. The Controlling Element in Sex Chromosome Behavior in *Sciara*. *Genetics*, 45(10):1429–1443, October 1960.
- [20] Darren A Cusanovich, Minal Caliskan, Christine Billstrand, Katelyn Michelini, Claudia Chavarria, Sherryl De Leon, Amy Mitrano, Noah Lewellyn, Jack A Elias, Geoffrey L Chupp, Roberto M Lang, Sanjiv J Shah, Jeanne M Decara, Yoav Gilad, and Carole Ober. Integrated analyses of gene expression and genetic association studies in a founder population. *Human Molecular Genetics*, 25(10):2104–2112, May 2016.

- [21] Alexander Dobin and Thomas R Gingeras. *Mapping RNA-seq Reads with STAR*, volume 29 of *Mapping RNA-seq Reads with STAR*. John Wiley & Sons, Inc., Hoboken, NJ, USA, August 2002.
- [22] Pan Du, Warren A Kibbe, and Simon M Lin. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13):1547–1548, July 2008.
- [23] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11(6):446–450, June 2010.
- [24] J Greg Falls, David J Pulford, Andrew A Wylie, and Randy L Jirtle. Genomic Imprinting: Implications for Human Disease. *The American Journal of Pathology*, 154(3):635–647, March 1999.
- [25] A Gabory, M A Ripoche, A Le Digarcher, F Watrin, A Ziyyat, T Forne, H Jammes, J F X Ainscough, M A Surani, L Journot, and L Dandolo. H19 acts as a trans regulator of the imprinted gene network controlling growth in mice. *Development*, 136(20):3413–3421, September 2009.
- [26] Paras Garg, Christelle Borel, and Andrew J Sharp. Detection of Parent-of-Origin Specific Expression Quantitative Trait Loci by Cis-Association Analysis of Gene Expression in Trios. *PLoS ONE*, 7(8):e41695, August 2012.
- [27] Greg Gibson. Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–145, January 2012.
- [28] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.
- [29] David Haig. The Kinship Theory of Genomic Imprinting. *Annual Review of Ecology and Systematics*, 31(1):9–32, November 2000.
- [30] I Hatada and T Mukai. Genomic imprinting of p57KIP2, a cyclin-dependent kinase inhibitor, in mouse. *Nature genetics*, 11(2):204–206, October 1995.
- [31] Ann-Kathrin Hoffmann, Xenia Naj, and Stefan Linder. Daam1 is a regulator of filopodia formation and phagocytic uptake of *Borrelia burgdorferi* by primary human macrophages. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 28(7):3075–3089, July 2014.
- [32] Richard Howey and Heather J Cordell. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics*, 13(1):149, June 2012.
- [33] D R Johnson. Hairpin-tail: a case of post-reductional gene action in the mouse egg. *Genetics*, 76(4):795–805, April 1974.

- [34] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, 8(1):118–127, January 2007.
- [35] Ricky S Joshi, Paras Garg, Noah Zaitlen, Tuuli Lappalainen, Corey T Watson, Nidha Azam, Daniel Ho, Xin Li, Stylianos E Antonarakis, Han G Brunner, Karin Buiting, Sau Wai Cheung, Bradford Coffee, Thomas Eggermann, David Francis, Joep P Geraedts, Giorgio Gimelli, Samuel G Jacobson, Cedric Le Caignec, Nicole de Leeuw, Thomas Liehr, Deborah J Mackay, Stephen B Montgomery, Alistair T Pagnamenta, Peter Papenhausen, David O Robinson, Claudia Ruivenkamp, Charles Schwartz, Bernhard Steiner, David A Stevenson, Urvashi Surti, Thomas Wassink, and Andrew J Sharp. DNA Methylation Profiling of Uniparental Disomy Subjects Provides a Map of Parental Epigenetic Bias in the Human Genome. *American journal of human genetics*, 99(3):555–566, September 2016.
- [36] Goo Jun, Matthew Flickinger, Kurt N Hetrick, Jane M Romm, Kimberly F Doheny, Goncalo R Abecasis, Michael Boehnke, and Hyun Min Kang. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics*, 91(5):839–848, November 2012.
- [37] Jennifer M Kalish, Connie Jiang, and Marisa S Bartolomei. Epigenetics and imprinting in human disease. *The International journal of developmental biology*, 58(2-4):291–298, 2014.
- [38] Augustine Kong, Valgerdur Steinthorsdottir, Gisli Masson, Gudmar Thorleifsson, Patrick Sulem, Soren Besenbacher, Aslaug Jonasdottir, Asgeir Sigurdsson, Kari Th Kristinsson, Adalbjorg Jonasdottir, Michael L Frigge, Arnaldur Gylfason, Pall I Olason, Sigurjon A Gudjonsson, Sverrir Sverrisson, Simon N Stacey, Bardur Sigurgeirsson, Kristrun R Benediktsdottir, Helgi Sigurdsson, Thorvaldur Jonsson, Rafn Benediktsdottir, Jon H Olafsson, Oskar Th Johannsson, Astradur B Hreidarsson, Gunnar Sigurdsson, DIAGRAM Consortium, Anne C Ferguson-Smith, Daniel F Gudbjartsson, Unnur Thorsteinsdottir, and Kari Stefansson. Parental origin of sequence variants associated with complex diseases. *Nature*, 462(7275):868–874, December 2009.
- [39] Charles A Laurin, Gabriel Cuellar-Partida, Gibran Hemani, George Davey Smith, Jian Yang, and David M Evans. Partitioning Phenotypic Variance Due To Parent-Of-Origin Effects Using Genomic Relatedness Matrices. *bioRxiv*, pages 1–41, May 2017.
- [40] Heather A Lawson, James M Cheverud, and Jason B Wolf. Genomic imprinting and parent-of-origin effects on complex traits. *Nature reviews. Genetics*, 14(9):609–617, August 2013.
- [41] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, March 2012.

- [42] Oren E Livne, Lide Han, Gorka Alkorta-Aranburu, William Wentworth-Sheilds, Mark Abney, Carole Ober, and Dan L Nicolae. PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population. *PLOS Computational Biology*, 11(3):e1004139, March 2015.
- [43] Isabel Lokody. Gene expression: Consequences of parent-of-origin effects. *Nature reviews. Genetics*, 15(3):145–145, January 2014.
- [44] Philippe P Luedi, Fred S Dietrich, Jennifer R Weidman, Jason M Bosko, Randy L Jirtle, and Alexander J Hartemink. Computational and experimental identification of novel human imprinted genes. *Genome Research*, 17(12):1723–1730, December 2007.
- [45] Weiwei Luo, Zi Zhao Lieu, Ed Manser, Alexander D Bershadsky, and Michael P Sheetz. Formin DAAM1 Organizes Actin Filaments in the Cytoplasmic Nodal Actin Network. *PLoS ONE*, 11(10):e0163915–22, October 2016.
- [46] Mary F Lyon and Sohaila Rastan. Parental source of chromosome imprinting and its relevance for X chromosome inactivation. *Differentiation*, 26(1):63–67, June 1984.
- [47] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, 13(6):R44, June 2012.
- [48] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, August 2011.
- [49] S Matsuoka, J S Thompson, M C Edwards, J M Bartletta, P Grundy, L M Kalikin, J W Harper, S J Elledge, and A P Feinberg. Imprinting of the gene encoding a human cyclin-dependent kinase inhibitor, p57KIP2, on chromosome 11p15. *Proceedings of the National Academy of Sciences*, 93(7):3026–3030, April 1996.
- [50] James McGrath and Davor Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37(1):179–183, May 1984.
- [51] Richard Mott, Wei Yuan, Pamela Kaisaki, Xiangchao Gan, James Cleak, Andrew Edwards, Amelie Baud, and Jonathan Flint. The Architecture of Parent-of-Origin Effects in Mice. *Cell*, 156(1-2):332–342, January 2014.
- [52] R D Nicholls, JHM Knoll, M G Butler, and S Karam Nature. Genetic imprinting suggested by maternal heterodisomy in non-deletion Prader-Willi syndrome. *nature.com*, 1989.
- [53] Jessie Nicodemus-Johnson, Rachel A Myers, Noburu J Sakabe, Debora R Sobreira, Douglas K Hogarth, Edward T Naureckas, Anne I Sperling, Julian Solway, Steven R White, Marcelo A Nobrega, Dan L Nicolae, Yoav Gilad, and Carole Ober. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI insight*, 1(20):e90151, December 2016.

- [54] Mats I Nilsson, Aliyah A Nissar, Dhuha Al-Sajee, Mark A Tarnopolsky, Gianni Parise, Boleslav Lach, Dieter O Fürst, Peter F M van der Ven, Rudolf A Kley, and Thomas J Hawke. Xin Is a Marker of Skeletal Muscle Damage Severity in Myopathies. *The American Journal of Pathology*, 183(6):1703–1709, December 2013.
- [55] Hiroaki Okae, Hitoshi Hiura, Yuichiro Nishida, Ryo Funayama, Satoshi Tanaka, Hatsune Chiba, Nobuo Yaegashi, Keiko Nakayama, Hiroyuki Sasaki, and Takahiro Arima. Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. *Human Molecular Genetics*, 21(3):548–558, October 2011.
- [56] Nuala A O’Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S Joardar, Vamsi K Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M McGarvey, Michael R Murphy, Kathleen O’Neill, Shashikant Pujar, Sanjida H Rangwala, Daniel Rausch, Lillian D Riddick, Conrad Schoch, Andrei Shkeda, Susan S Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E Tully, Anjana R Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D Murphy, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, January 2016.
- [57] Paolo Pariante, Raffaele Dotolo, Massimo Venditti, Diana Ferrara, Aldo Donizetti, Francesco Aniello, and Sergio Minucci. First Evidence of DAAM1 Localization During the Post-Natal Development of Rat Testis and in Mammalian Sperm. *Journal of cellular physiology*, 231(10):2172–2184, October 2016.
- [58] Manus M Patten, Michael Cowley, Rebecca J Oakey, and Robert Feil. Regulatory links between imprinted genes: evolutionary predictions and consequences. *Proceedings of the Royal Society B: Biological Sciences*, 283(1824):20152760, February 2016.
- [59] John Rb Perry, Felix Day, Cathy E Elks, Patrick Sulem, Deborah J Thompson, Teresa Ferreira, Chunyan He, Daniel I Chasman, Tõnu Esko, Gudmar Thorleifsson, Eva Albrecht, Wei Q Ang, Tanguy Corre, Diana L Cousminer, Bjarke Feenstra, Nora Franceschini, Andrea Ganna, Andrew D Johnson, Sanelia Kjellqvist, Kathryn L Lunetta, George McMahon, Ilja M Nolte, Lavinia Paternoster, Eleonora Porcu, Albert V Smith, Lisette Stolk, Alexander Teumer, Natalia Tšernikova, Emmi Tikkanen, Sheila Ulivi, Erin K Wagner, Najaf Amin, Laura J Bierut, Enda M Byrne, Jouke-Jan Hottenga, Daniel L Koller, Massimo Mangino, Tune H Pers, Laura M Yerges-Armstrong, Jing Hua Zhao, Irene L Andrulis, Hoda Anton-Culver, Femke Atsma, Stefania Bandinelli, Matthias W Beckmann, Javier Benitez, Carl Blomqvist, Stig E Bojesen, Manjeet K Bolla, Bernardo Bonanni, Hiltrud Brauch, Hermann Brenner, Julie E Buring, Jenny Chang-Claude, Stephen Chanock, Jinhui Chen, Georgia Chenevix-Trench, J Margriet Collée, Fergus J

Couch, David Couper, Andrea D Coveillo, Angela Cox, Kamila Czene, Adamo Pio D'adamo, George Davey Smith, Immaculata De Vivo, Ellen W Demerath, Joe Dennis, Peter Devilee, Aida K Dieffenbach, Alison M Dunning, Gudny Eiriksdottir, Johan G Eriksson, Peter A Fasching, Luigi Ferrucci, Dieter Flesch-Janys, Henrik Flyger, Tatiana Foroud, Lude Franke, Melissa E Garcia, Montserrat García-Closas, Frank Geller, Eco Ej de Geus, Graham G Giles, Daniel F Gudbjartsson, Vilmundur Gudnason, Pascal Guénel, Suiqun Guo, Per Hall, Ute Hamann, Robin Haring, Catharina A Hartman, Andrew C Heath, Albert Hofman, Maartje J Hooning, John L Hopper, Frank B Hu, David J Hunter, David Karasik, Douglas P Kiel, Julia A Knight, Veli-Matti Kosma, Zoltán Kutalik, Sandra Lai, Diether Lambrechts, Annika Lindblom, Reedik Mägi, Patrik K Magnusson, Arto Mannermaa, Nicholas G Martin, Gisli Masson, Patrick F McArdle, Wendy L McArdle, Mads Melbye, Kyriaki Michailidou, Evelin Mihailov, Lili Milani, Roger L Milne, Heli Nevanlinna, Patrick Neven, Ellen A Nohr, Albertine J Oldehinkel, Ben A Oostra, Aarno Palotie, Munro Peacock, Nancy L Pedersen, Paolo Peterlongo, Julian Peto, Paul Dp Pharoah, Dirkje S Postma, Anneli Pouta, Katri Pylkäs, Paolo Radice, Susan Ring, Fernando Rivadeneira, Antonietta Robino, Lynda M Rose, Anja Rudolph, Veikko Salomaa, Serena Sanna, David Schlessinger, Marjanka K Schmidt, Mellissa C Southey, Ulla Sovio, Meir J Stampfer, Doris Stöckl, Anna M Storniolo, Nicholas J Timpson, Jonathan Tyrer, Jenny A Visser, Peter Vollenweider, Henry Völzke, Gerard Waeber, Melanie Waldenberger, Henri Wallaschofski, Qin Wang, Gonneke Willemse, Robert Winqvist, Bruce Hr Wolffenduttel, Margaret J Wright, Australian Ovarian Cancer Study, GENICA Network, kConFab, LifeLines Cohort Study, InterAct Consortium, Early Growth Genetics (EGG) Consortium, Dorret I Boomsma, Michael J Econs, Kay-Tee Khaw, Ruth Jf Loos, Mark I McCarthy, Grant W Montgomery, John P Rice, Elizabeth A Streeten, Unnur Thorsteinsdottir, Cornelia M van Duijn, Behrooz Z Alizadeh, Sven Bergmann, Eric Boerwinkle, Heather A Boyd, Laura Crisponi, Paolo Gasparini, Christian Gieger, Tamara B Harris, Erik Ingelsson, Marjo-Riitta Jarvelin, Peter Kraft, Debbie Lawlor, Andres Metspalu, Craig E Pennell, Paul M Ridker, Harold Snieder, Thorkild Ia Sørensen, Tim D Spector, David P Strachan, André G Uitterlin-den, Nicholas J Wareham, Elisabeth Widén, Marek Zygmunt, Anna Murray, Douglas F Easton, Kari Stefansson, Joanne M Murabito, and Ken K Ong. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514(7520):92–97, October 2014.

- [60] Jo Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nature reviews. Genetics*, 15(8):517–530, August 2014.
- [61] Nuno D Pires and Ueli Grossniklaus. Different yet similar: evolution of imprinting in flowering plants and mammals. *F1000Prime Reports*, 6, August 2014.
- [62] C Proudhon and D Bourc'his. Identification and resolution of artifacts in the interpretation of imprinted gene expression. *Briefings in Functional Genomics*, 9(5-6):374–384, January 2011.

- [63] Wolf Reik. Genomic imprinting and genetic disorders in man. *Trends in Genetics*, 5:332–336, January 1989.
- [64] DR Johnson Genetics Research and 1974. Further observations on the hairpin-tail (Thp) mutation in the mouse. *cambridge.org*.
- [65] M E Ritchie, B Phipson, D Wu, Y Hu, C W Law, W Shi, and G K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, April 2015.
- [66] Federico A Santoni, Georgios Stamoulis, Marco Garieri, Emilie Falconnet, Pascale Ribaux, Christelle Borel, and Stylianos E Antonarakis. Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. *American journal of human genetics*, 100(3):444–453, March 2017.
- [67] C Sapienza. Genome imprinting and dominance modification. *Annals of the New York Academy of Sciences*, 564:24–38, 1989.
- [68] Bertrand Servin and Matthew Stephens. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS genetics*, 3(7):e114, 2007.
- [69] Michelle M Stein, Cara L Hrusch, Justyna Gozdz, Catherine Igartua, Vadim Pivniouk, Sean E Murray, Julie G Ledford, Mauricius Marques dos Santos, Rebecca L Anderson, Nervana Metwali, Julia W Neilson, Raina M Maier, Jack A Gilbert, Mark Holbreich, Peter S Thorne, Fernando D Martinez, Erika von Mutius, Donata Vercelli, Carole Ober, and Anne I Sperling. Innate Immunity and Asthma Risk in Amish and Hutterite Farm Children. *New England Journal of Medicine*, 375(5):411–421, August 2016.
- [70] Liora Z Strichman-Almashanu, Michael Bustin, and David Landsman. Retroposed copies of the HMG genes: a window to genome dynamics. *Genome Research*, 13(5):800–812, May 2003.
- [71] Pernilla Stridh, Sabrina Ruhrmann, Petra Bergman, Mélanie Thessén Hedreul, Sevasti Flytzani, Amennai Daniel Beyeen, Alan Gillett, Nina Krivosija, Johan Öckinger, Anne C Ferguson-Smith, and Maja Jagodic. Parent-of-Origin Effects Implicate Epigenetic Regulation of Experimental Autoimmune Encephalomyelitis and Identify Imprinted Dlk1 as a Novel Risk Gene. *PLoS genetics*, 10(3):e1004265, March 2014.
- [72] M A H Surani, S C Barton, and M L Norris. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature*, 308(5959):548–550, April 1984.
- [73] Przemyslaw Szafranski, Tomasz Gambin, Avinash V Dharmadhikari, Kadir Caner Akdemir, Shalini N Jhangiani, Jennifer Schuette, Nihal Godiwala, Svetlana A Yatsenko, Jessica Sebastian, Suneeta Madan-Khetarpal, Urvashi Surti, Rosanna G Abellar, David A Bateman, Ashley L Wilson, Melinda H Markham, Jill Slamon, Fernando Santos-Simarro, María Palomares, Julián Nevado, Pablo Lapunzina, Brian Hon-Yin

Chung, Wai-Lap Wong, Yoyo Wing Yiu Chu, Gary Tsz Kin Mok, Eitan Kerem, Joel Reiter, Namasivayam Ambalavanan, Scott A Anderson, David R Kelly, Joseph Shieh, Taryn C Rosenthal, Kristin Scheible, Laurie Steiner, M Anwar Iqbal, Margaret L McKinnon, Sara Jane Hamilton, Kamilla Schlade-Bartusiak, Dawn English, Glenda Hendson, Elizabeth R Roeder, Thomas S DeNapoli, Rebecca Okashah Littlejohn, Daynna J Wolff, Carol L Wagner, Alison Yeung, David Francis, Elizabeth K Fiorino, Morris Edelman, Joyce Fox, Denise A Hayes, Sandra Janssens, Elfride De Baere, Björn Menten, Anne Loccuifier, Lieve Vanwalleghem, Philippe Moerman, Yves Sznajer, Amy S Lay, Jennifer L Kussmann, Jasneek Chawla, Diane J Payton, Gael E Phillips, Erwin Brosens, Dick Tibboel, Annelies de Klein, Isabelle Maystadt, Richard Fisher, Neil Sebire, Alison Male, Maya Chopra, Jason Pinner, Girvan Malcolm, Gregory Peters, Susan Arbuckle, Melissa Lees, Zoe Mead, Oliver Quarrell, Richard Sayers, Martina Owens, Charles Shaw-Smith, Janet Lioy, Eileen McKay, Nicole de Leeuw, Ilse Feenstra, Liesbeth Spruijt, Frances Elmslie, Timothy Thiruchelvam, Carlos A Bacino, Claire Langston, James R Lupski, Partha Sen, Edwina Popek, and Paweł Stankiewicz. Pathogenetics of alveolar capillary dysplasia with misalignment of pulmonary veins. *Human Genetics*, 135(5):569–586, April 2016.

- [74] N Takagi and M Sasaki. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature*, 256(5519):640–642, August 1975.
- [75] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12(11):1061–1063, November 2015.
- [76] Annie Varrault, Charlotte Gueydan, Annie Delalbre, Anja Bellmann, Souheir Houssami, Cindy Aknin, Dany Severac, Laetitia Chotard, Malik Kahli, Anne Le Digarcher, Paul Pavlidis, and Laurent Journot. Zac1 Regulates an Imprinted Gene Network Critically Involved in the Control of Embryonic Growth. *Developmental Cell*, 11(5):711–722, November 2006.
- [77] Qinchuan Wang, Jenny Li-Chun Lin, Albert J Erives, Cheng-I Lin, and Jim Jung-Ching Lin. New insights into the roles of Xin repeat-containing proteins in cardiac development, function, and disease. *International review of cell and molecular biology*, 310:89–128, 2014.
- [78] Y Wang, J R O’Connell, P F McArdle, J B Wade, S E Dorff, S J Shah, X Shi, L Pan, E Rampersaud, H Shen, J D Kim, A R Subramanya, N I Steinle, A Parsa, C C Ober, P A Welling, A Chakravarti, A B Weder, R S Cooper, B D Mitchell, A R Shuldiner, and Y P C Chang. Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proceedings of the National Academy of Sciences*, 106(1):226–231, January 2009.
- [79] C R Weinberg, A J Wilcox, and R T Lie. A Log-Linear Approach to Case-Parent-Triad Data: Assessing Effects of Disease Genes That Act Either Directly or through Maternal

- Effects and That May Be Subject to Parental Imprinting. *The American Journal of Human Genetics*, 62(4):969–978, April 1998.
- [80] Clarice R Weinberg. Methods for Detection of Parent-of-Origin Effects in Genetic Studies of Case-Parents Triads. *The American Journal of Human Genetics*, 65(1):229–235, July 1999.
- [81] Lauren A Weiss, Mark Abney, Edwin H Cook Jr., and Carole Ober. Sex-Specific Genetic Architecture of Whole Blood Serotonin Levels. *The American Journal of Human Genetics*, 76(1):33–41, January 2005.
- [82] Wellcome Trust Case Control Consortium, Peter Donnelly, Willem H Ouwehand, Nick Craddock, Nilesh J Samani, John A Todd, Mark I McCarthy, Jeffrey C Barrett, Andrew P Morris, Lon R Cardon, Chris C A Spencer, Matthew A Brown, Panos Deloukas, Martin Farrall, Alistair S Hall, Andrew T Hattersley, Christopher G Mathew, and Miles Parkes. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.
- [83] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS genetics*, 9(5):e1003520, May 2013.
- [84] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, June 2012.
- [85] Magdalena Zoledziewska, Carlo Sidore, Charleston W K Chiang, Serena Sanna, Antonella Mulas, Maristella Steri, Fabio Busonero, Joseph H Marcus, Michele Marongiu, Andrea Maschio, Diego Ortega Del Vecchyo, Matteo Floris, Antonella Meloni, Alessandro Delitala, Maria Pina Concas, Federico Murgia, Ginevra Biino, Simona Vaccargiu, Ramaiah Nagaraja, Kirk E Lohmueller, UK10K Consortium, Nicholas J Timpson, Nicole Soranzo, Ioanna Tachmazidou, George Dedoussis, Eleftheria Zeggini, Understanding Society Scientific Group, Sergio Uzzau, Chris Jones, Robert Lyons, Andrea Angius, Goncalo R Abecasis, John Novembre, David Schlessinger, and Francesco Cucca. Height-reducing variants and selection for short stature in Sardinia. *Nature genetics*, 47(11):1352–1356, November 2015.
- [86] O Zuk, E Hechter, SR Sunyaev Proceedings of the, and 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *National Acad Sciences*.