LUND UNIVERSITY

Parent-of-Origin Effects and Allelic Imbalance in Metabolism - Across the Life Course: Implications for Metabolic Traits, Disease Risk, and Gene Regulation

Andersson, Jonas

2026

Document Version:
Publisher's PDF, also known as Version of record

Parent-of-Origin Effects and Allelic Imbalance in Metabolism

Across the Life Course: Implications for Metabolic Traits, Disease Risk, and Gene Regulation

Jonas Andersson

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Medicine at Lund University to be publicly defended on date of September 3rd
At 13:00 in the Aghard Lecture Hall, CRC.
Address: Jan Waldenströms gata 35, 214 28 Malmö

Faculty opponent
Prof. Ingrid Kockum, Karolinska Institute, Sweden.

Organization: LUND UNIVERSITY

Document name: Doctoral dissertation **Date of issue** 2025-09-03

Author: Jonas Andersson

Title and subtitle: Parent-of-Origin Effects and Allelic Imbalance in Metabolism - Across the Life Course: Implications for Metabolic Traits, Disease Risk, and Gene Regulation

Abstract: Parent-of-origin effects (POEs), whereby the phenotypic impact of a genetic variant depends on parental inheritance, represent a mechanism through which genetic and environmental factors interact in complex traits such as metabolic disorders and cancer. As parental contributions differ, POEs can influence both early-life development and adult metabolic function, with parental imbalances observed in traits such as type 2 diabetes (T2D), including increased maternal transmission. These effects may arise through genomic imprinting or more subtle parental expression biases. Allelic expression imbalance (AEI) analyses can identify such effects by detecting parent-specific differences in gene expression, providing insight into underlying genetic and epigenetic regulation. Allelic imbalance also plays a key role in cancer, where somatic copy number variations alter gene dosage and regulatory elements, leading to tumor-associated changes in gene expression. The aim of this thesis was to identify and characterize genes exhibiting POEs influencing metabolic traits and T2D risk, particularly insulin secretion, and to investigate allelic imbalance in cancer to better understand gene regulatory mechanisms in disease.

Results

Paper 1: Developmentally dynamic POEs on key metabolic traits related to T2D risk were identified in an Indian population, with shifting maternal and paternal influences across life stages and supported by genetic POE at loci such as *KCNQ1.*

Paper 2: AEI analyses in human pancreatic islets identified >30% of genes showing allelic imbalance, including candidates potentially influenced by POEs. AEI was assessed across chronic hyperglycemic, acute hyperglycemic, and normoglycemic states, enabling identification of genes potentially involved in T2D pathogenesis. Selected candidates were then tested for POEs in large family-based cohorts (n > 10,000), leading to the identification and replication of several POE-associated variants.

Paper 3: By analyzing gene expression and AEI across multiple resolutions, including bulk pancreas, islets, and single-cell data from fetal and adult tissues, we identified imprinted genes with potential roles in β-cell development and function, supported by genetic association and POE analyses in a large family-based cohort (n > 10,000).

Paper 4: By integrating genomic and transcriptomic data using the BASE framework, we characterized the allele-specific expression landscape in high hyperdiploid acute lymphoblastic leukemia (HeH ALL). This revealed widespread allelic imbalance driven by somatic copy number variation, providing insight into how gene dosage and cis-regulatory alterations shape gene expression in cancer and highlighting candidate genes and regulatory mechanisms involved in leukemia development.

Conclusion: Collectively, these studies demonstrate that POEs and allelic imbalance are key contributors to the regulation of metabolic traits and disease susceptibility, with dynamic, context-dependent influences on insulin secretion, β-cell function, and cancer-related gene expression.

Key words: Type 2 diabetes, T2D**,** Parent-of-Origin Effects, POE, Allelic expression imbalance, AEI, Allele-specific expression, ASE

Language: English **Number of pages**:74

ISSN: **1652-8220**

ISBN: 978-91-8021-903-7

Signature  Date 2026-07-01

Parent-of-Origin Effects and Allelic Imbalance in Metabolism

Across the Life Course: Implications for Metabolic Traits, Disease Risk, and Gene Regulation

Jonas Andersson

LUND
UNIVERSITY

Published by:

Department of Clinical Sciences, Malmö

Faculty of Medicine

Lund University

Lund 2026

ISBN 978-91-8021-903-7 (print)

Lund University, Faculty of Medicine Doctoral Dissertation Series 2026:105

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University,
Lund, 2026

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

To my dear Mother who unfortunately left us just recently
I promised you to finish this thesis, so here it is 💔

To my beloved ex-wife
If anyone knows the cost of making this thesis happen, it's you.
All work and no play made Jonas a dull boy.

When you thought it was gonna be easy peasy lemon squeezy,
but it was actually difficult, difficult, lemon difficult.

Lord of the diabetes *(the fellowship of Groop)*

Table of Contents

Papers included in the thesis

Paper 1

Wagh R, Hatem G, **Andersson J**, Kunte P, Bandyopadhyay S, Yajnik CS, Prasad RB. Parent-of-origin effects in the life-course evolution of cardiometabolic traits. Diabetologia. 2025 Jun;68(6):1298-1314. doi: 10.1007/s00125-025-06396-5. Epub 2025 Apr 2. PMID: 40175764; PMCID: PMC12069499.

Paper 2

Andersson J, Hatem G, Maziarz M, Karagiannopoulos A, Kovacs G, Vitai M, Lehtovirta M, Wong WKM, Hakaste L, Thangam M, Kryvokhyzha D, Hansson O, Ahlqvist A, Yajnik CS, Koranyi L, Hardikar AA, Eliasson L, Tuomi T, Shekhawat RS, Asplund O, Munoz F, Cataldo R, Mulder H, Fex M, Nilsson P, Prasad RB Allelic expression imbalance in pancreatic islets indicate parent-of-origin effects on type 2 diabetes risk (Manuscript)

Paper 3

Hatem G, **Andersson J**, Asplund O, Ahuja V, Singh T, Maziarz M, Karagiannopoulos A, Lehtovirta M, Hakaste L, Yajnik CS, Eliasson L, Tuomi T, Artner I, Prasad RB. Imprinted genes in beta cell development and T2D risk (Manuscript)

Paper 4

Andersson J, Aydın E, Gunnarsson R, Lilljebjörn H, Fioretos T, Johansson B, Paulsson K, Yang M. Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE. Sci Rep. 2024 Oct 5;14(1):23181. doi: 10.1038/s41598-024-73743-8. PMID: 39369032; PMCID: PMC11455916.

Papers not included in the thesis

Paper 1

Bertonnier-Brouty L, **Andersson J**, Kaprio T, Hagström J, Bsharat S, Asplund O, Hatem G, Haglund C, Seppänen H, Prasad RB, Artner I. E2F transcription factors promote tumorigenicity in pancreatic ductal adenocarcinoma. Cancer Med. 2024 May;13(9):e7187. doi: 10.1002/cam4.7187. PMID: 38686617; PMCID: PMC11058697.

Paper 2

Bertonnier-Brouty L, Bsharat S, Achanta K, **Andersson J**, Pranomphon T, Singh T, Kaprio T, Hagström J, Haglund C, Seppänen H, Prasad RB, Artner I. Homeobox protein B6 and homeobox protein B8 control immune-cancer cell interactions in pancreatic cancer. Mol Biomed. 2025 Jul 7;6(1):48. doi: 10.1186/s43556-025-00292-5. PMID: 40619489; PMCID: PMC12229978.

Paper 3

Hatem G, **Andersson J***, Hjort L*, Minja DTR, Catellani FC, Møller SL, Aspund O, Christensen DL, Nielsen BB, Theander TG, Artner I, Shaat N, Lusingu JPA, Vaag A, Bygbjerg IC, Schmiegelow C, Groth-Grunnet L, Prasad RB. Signatures of fetal programming in cord blood from mothers with gestational diabetes mellitus

Paper 4

Andersson J*, Shekhawat RS*, Madsen A, Nilsson P, Ahlqvist E, Vaag A, Hansen T, Tuomi T, Prasad RB. Parent-of-origin effects on insulin secretion and resistance

Abbreviations

AEI	Allelic expression imbalance
ALL	Acute lymphoblastic leukemia
ASE	Allele-specific expression
BMI	Body mass index
CIR	Corrected insulin response
CNV	Copy number variations
CPM	Counts per million
eQTL	Expression quantitative trait locus
GIP	Gastric inhibitory polypeptide
GLP-1	Glucagon-like peptide-1
GWAS	Genome-wide association studies
HOMA	Homeostatic model assessment
IDF	International Diabetes Federation
OGTT	Oral glucose tolerance test
POE	Parent-of-origin effect
SNP	Single-nucleotide polymorphism
T2D	Type 2 diabetes

Popular science summary

Typ 2-diabetes (T2D) är en vanlig och mycket komplex sjukdom som påverkas av flera olika faktorer. Både arv, miljö, kroppens utveckling och ämnesomsättning spelar roll. Vid T2D har kroppen svårt att hålla blodsockret på en normal nivå. Det beror ofta på två saker: dels att kroppens vävnader blir mindre känsliga för insulin, dels att de insulinproducerande β-cellerna i bukspottkörteln inte kan producera eller frisätta tillräckligt med insulin. Insulin är ett hormon som hjälper socker att tas upp från blodet och användas som energi i kroppens celler.

Stora genetiska studier har hittat många genvarianter som är kopplade till ökad risk för T2D. Dessa studier har varit mycket viktiga, men de förklarar inte hela bilden. En genvariant kan nämligen påverka kroppen på olika sätt beroende på när i livet den är aktiv, i vilken vävnad den verkar och ibland även om den har ärvts från mamman eller pappan.

En central del av denna avhandling handlar om så kallade parent-of-origin-effekter. Det betyder att en genvariant kan ha olika effekt beroende på om den ärvs från mamman eller pappan. Vi ärver två kopior av de flesta gener, en från varje förälder. Ofta används båda kopiorna, men ibland används den ena kopian mer än den andra. I vissa fall kan en genkopia till och med vara helt avstängd. Detta kallas genomisk prägling och är särskilt viktigt under fosterutvecklingen, då organ och vävnader bildas. Ett närliggande begrepp är ”obalans i alleluttryck”. Det innebär att de två genkopiorna inte uttrycks lika mycket. Genom att studera detta kan man få ledtrådar om vilka gener som kan vara inblandade i parent-of-origin-effekter, och därmed vara särskilt viktiga i en viss vävnad eller sjukdom. I detta arbete användes sådana analyser för att förstå hur gener regleras i bukspottkörteln, både under fosterutvecklingen och i vuxna bukspottkörtelöar där β-cellerna finns.

Resultaten visar att parent-of-origin-effekter kan påverka viktiga egenskaper som insulinsöndring, insulinkänslighet och blodfettsnivåer. Effekterna var inte alltid likadana genom livet, utan kunde förändras från barndom till vuxen ålder. Det tyder på att vissa genetiska effekter etableras tidigt och kan bidra till sjukdomsrisk senare i livet. Flera gener med föräldraspecifik reglering var också aktiva i fostrets bukspottkörtel, vilket kan vara viktigt för hur insulinproducerande β-celler utvecklas. I vuxna bukspottkörtelöar hittades obalans i alleluttryck i gener som är relevanta för β-cellernas funktion och T2D. Detta tyder på att skillnader i hur de två genkopiorna används kan påverka både normal insulinproduktion och sjukdomsutveckling. Samma typ av analys kunde även användas i akut lymfatisk leukemi, vilket visar att metoden kan ge viktig kunskap även vid andra sjukdomar.

Sammanfattningsvis visar resultaten att genetiska effekter inte alltid är enkla eller likadana för båda föräldrarna. Genom att ta hänsyn till föräldraursprung, utvecklingsstadium och genreglering i olika vävnader kan vi få en bättre förståelse för varför T2D och andra komplexa sjukdomar uppstår.

Introduction

Complex traits

This thesis will primarily address type 2 diabetes (T2D), a metabolic disorder, but will also cover cancer, specifically acute lymphoblastic leukemia (ALL), which, although not a metabolic disorder itself, exhibits metabolic dysregulation as a central feature (1). Both T2D and cancer are complex traits (2), although cancer involves additional complexity due to acquired somatic mutations.

A complex trait is a phenotype that is influenced by multiple genetic variants and environmental factors, rather than by a mutation in a single gene. The genetic architecture is typically polygenic, meaning many loci contribute small effects that together shape the trait (2). Key characteristics of complex traits include:

1. **Polygenic inheritance**
   Many genes contribute, each with a small effect, as seen in the multiple susceptibility loci associated with T2D and cancer risk (2). Most of these risk variants map to non-coding regions of the genome, suggesting they influence gene regulation rather than protein sequence (3).
2. **Environmental influence**
   Lifestyle, diet, and other exposures significantly influence the development and progression of both T2D and cancer (4, 5).
3. **Non-Mendelian inheritance patterns**
   The risk for these diseases does not follow simple dominant or recessive rules (2).
4. **Gene–gene and gene–environment interactions**
   Interactions between genetic variants and environmental factors influence disease susceptibility (6-8).
5. **Quantitative or multifactorial nature**
   Traits like blood glucose regulation in T2D or overall cancer risk exist along a continuous spectrum rather than discrete categories (9, 10).

These features highlight why both T2D and cancer are considered complex traits and underscore the challenges in predicting, preventing, and treating them.

Diabetes - A Global Health Challenge

Diabetes is not a single disease but a group of metabolic disorders characterized by chronically elevated blood glucose levels (11). Persistent hyperglycemia often leads to long-term damage of multiple organs, causing substantial morbidity and suffering among affected individuals. Common long-term complications include cardiovascular disease, chronic kidney disease, hearing and vision loss, peripheral neuropathy, and neuropsychiatric complications (11, 12). In addition to its clinical impact, diabetes imposes a considerable societal burden, with global healthcare expenditures estimated to exceed 1 trillion USD by 2025 (13).

Type 2 Diabetes

Prevalence

As of 2024, an estimated 589 million adults aged 20-79 years worldwide are living with diabetes, a number projected to increase to 853 million by 2050, making diabetes one of the fastest growing global health challenges (13). T2D is the most common form, accounting for approximately 90% of all cases worldwide (11, 13) (Figure 1). In Sweden, T2D likewise accounts for approximately 90% of diabetes cases, with an estimated 566,600 affected adults (20–79 years) (≈7.5%) (13). Similarly, in Finland, T2D accounts for approximately 90% of diabetes cases, with an estimated 400,400 affected adults (20–79 years) (≈6.9%) (13, 14).

Notably, China and India have experienced a particularly rapid increase in T2D prevalence over the past few decades, coinciding with rapid urbanization, dietary transitions, and reduced physical activity. Despite this sharp rise, individuals in these populations tend to develop T2D at lower body mass index (BMI) levels compared with Western populations, reflecting ethnic differences in body composition, visceral adiposity, and metabolic susceptibility (15-17). Importantly, the true disease burden is likely underestimated because a large fraction of adults with diabetes remains undiagnosed, often linked to limited access to screening and healthcare capacity. According to 2024 estimates from the International Diabetes Federation (IDF), the proportion of undiagnosed diabetes is about 49.7% in China and 43.0% in India (18, 19). Historically, global diabetes projections have frequently been revised upward as new data have accumulated, suggesting that long-term forecasts, including those for 2050, may represent conservative estimates (13).

Summary

Map 1 Number of people with diabetes worldwide and per IDF Region, in 2024–2050 (20–79 years)

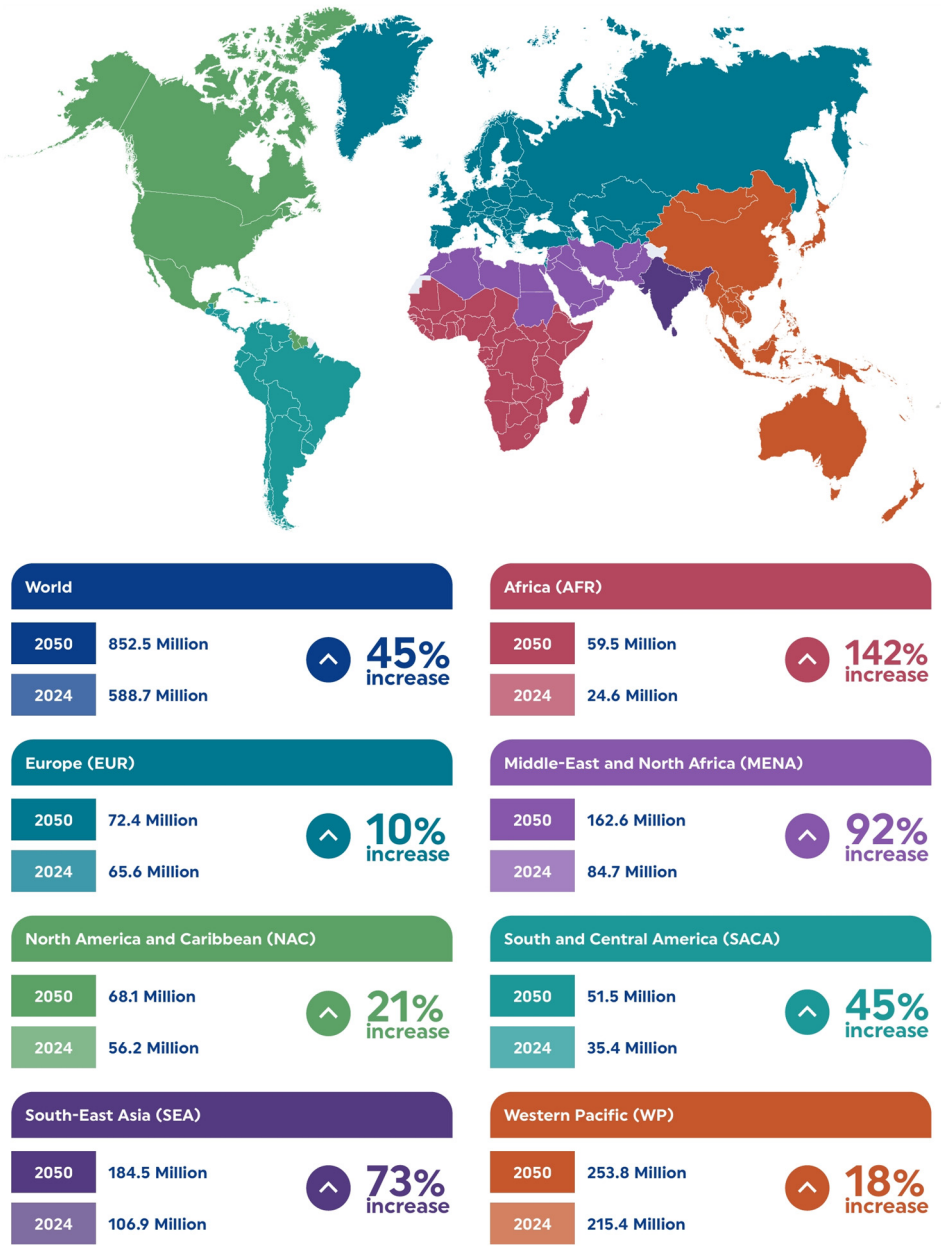

Figure 1. Number of people with diabetes worldwide and per IDF Region, in 2024-2050 (20-79 years)
IDF Diabetes Atlas 11th edition – 2025 | diabetesatlas.org

Pathophysiology

T2D is a complex, lifelong disease that arises from an intricate interplay between genetic susceptibility and environmental factors. While the major environmental contributors such as diet and physical activity are well established, the genetic determinants of T2D remains a challenge (11). The Disease is classically characterized by a combination of insulin resistance and impaired insulin secretion. Disease onset typically occurs when pancreatic β-cells can no longer compensate for increasing insulin resistance, a process often associated with obesity (11).

The "omnious octet"

Beyond this classical view, T2D is now recognized as a multisystem disorder involving multiple organs and regulatory pathways. This concept is encapsulated by the "**ominous octet**", which describes eight interrelated pathophysiological mechanisms contributing to hyperglycemia, including dysfunction of the pancreas, liver, skeletal muscle, adipose tissue, gut, kidney, and central nervous system (20) (Figure 2).

More specifically, the first component is **pancreatic β-cell dysfunction**, leading to inadequate insulin secretion and impaired glycemic control (20, 21).

The second and third components are **insulin resistance in skeletal muscle and liver**, respectively. In skeletal muscle, insulin resistance reduces insulin-stimulated glucose uptake, whereas in the liver it impairs suppression of hepatic glucose production and contributes to increased gluconeogenesis (20, 21).

The fourth component is **adipocyte dysfunction**, where impaired insulin mediated inhibition of lipolysis elevates circulating free fatty acids, further exacerbating insulin resistance and β-cell dysfunction (20, 22);

The fifth component involves **gastrointestinal incretin dysfunction**, characterized by a reduced incretin effect due to impaired GLP-1 action and inadequate GIP mediated stimulation of insulin secretion (20, 23). In this thesis, particular attention is given to the gene ***GLP1R***, which encodes the glucagon-like peptide-1 receptor. This receptor has gained considerable interest in recent years due to its central role in incretin-based therapies for T2D and obesity. Pharmacological agents targeting this pathway, such as the GLP-1 receptor agonist semaglutide and the dual GIP/GLP-1 receptor agonist tirzepatide, mimic the effects of the incretin hormone GLP-1, leading to glucose-dependent insulin secretion, suppression of glucagon release, delayed gastric emptying, reduced appetite, weight loss, and improved cardiometabolic outcomes (24).

The sixth component is **α-cell dysfunction**, leading to inappropriate glucagon secretion, increased hepatic glucose output, and fasting hyperglycemia (20, 25).

The seventh is **renal glucose handling**, where upregulated sodium-glucose cotransporter 2 (SGLT2)-mediated glucose reabsorption limits glucosuria and sustains hyperglycemia (20, 26).

The final component is **central nervous system dysregulation**. In obesity, functional magnetic resonance imaging (fMRI) studies have demonstrated an attenuated and delayed hypothalamic response to oral glucose ingestion, consistent with altered central nutrient sensing in brain regions involved in appetite and energy balance (20, 27).

Together, this framework highlights the complexity of T2D pathogenesis and emphasizes that β-cell dysfunction occurs within a broader context of systemic metabolic dysregulation.

The "egregious eleven"

More recently, this model has been expanded to incorporate additional contributors to disease development and progression, sometimes referred to as an updated ominous octet or the "**egregious eleven**" (28) (Figure 2).

The ninth component is **systemic low-grade inflammation and immune dysregulation**, where chronic activation of inflammatory pathways in adipose tissue, liver, and pancreatic islets promotes insulin resistance and accelerates β-cell dysfunction.

The tenth component is **reduced amylin action secondary to β-cell dysfunction**, which can accelerate gastric emptying and increase intestinal glucose absorption, thereby worsening postprandial hyperglycemia.

The eleventh component involves **alterations in the gut microbiome**, which may influence glucose homeostasis through effects on short chain fatty acid production, bile acid metabolism, intestinal barrier integrity, and systemic inflammatory signaling.

In summary, these additional mechanisms extend the original framework and reinforce the view of T2D as a heterogeneous and multifactorial disease involving metabolic, inflammatory, and neuroendocrine pathways. In this thesis, particular focus is placed on the first component, insulin deficiency due to **pancreatic β-cell dysfunction**, which may arise from reduced β-cell mass, β-cell dedifferentiation in adulthood, or lower β-cell mass already established in early life, such as during fetal development (Figure 3).

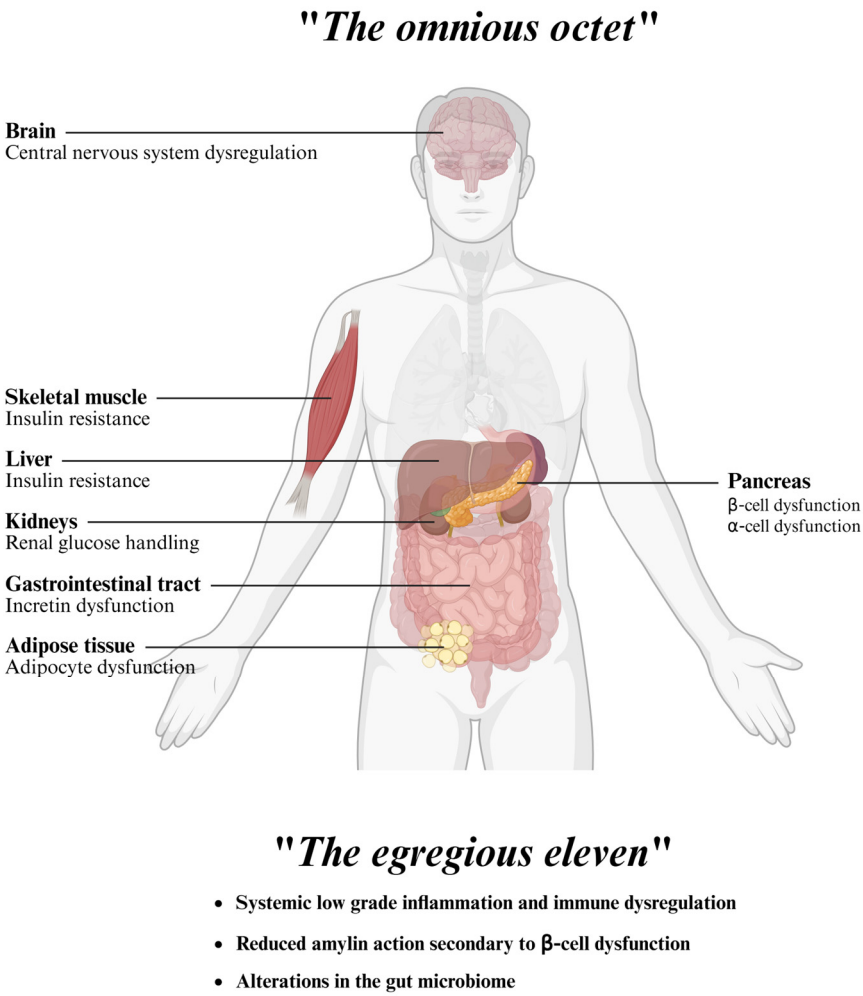

Figure 2. The "Ominous Octet" and extension to the "Egregious Eleven" in type 2 diabetes.
The Ominous Octet illustrates eight key organ-specific defects contributing to hyperglycemia, including pancreatic β-cell and α-cell dysfunction, insulin resistance in skeletal muscle and liver, adipocyte dysfunction, impaired incretin action in the gastrointestinal tract, altered renal glucose handling, and central nervous system dysregulation. Building on this framework, the "Egregious Eleven" incorporates additional mechanisms, including systemic low-grade inflammation and immune dysregulation, reduced amylin action secondary to β-cell dysfunction, and alterations in the gut microbiome, highlighting the increasingly recognized complexity and multisystem nature of type 2 diabetes pathophysiology.

Insulin resistance and obesity are well-established contributors to T2D. However, obesity alone cannot explain disease onset. If obesity were sufficient to cause T2D, most individuals with obesity would develop hyperglycemia. In reality, approximately 80% remain free of diabetes (29), indicating that additional factors are required. Growing evidence suggests that pancreatic β-cell dysfunction is a central determinant of whether insulin resistance progresses to overt disease. One explanation for β-cell compensatory failure is pre-existing β-cell dysfunction, potentially resulting from reduced β-cell mass or impaired functional capacity. This relative insulin deficiency can accelerate metabolic dysregulation and promote further loss of β-cell function. Although the precise mechanisms underlying β-cell failure remain incompletely understood, intrinsic β-cell dysfunction, potentially influenced by hereditary factors, appears to play a central role in T2D pathogenesis (29).

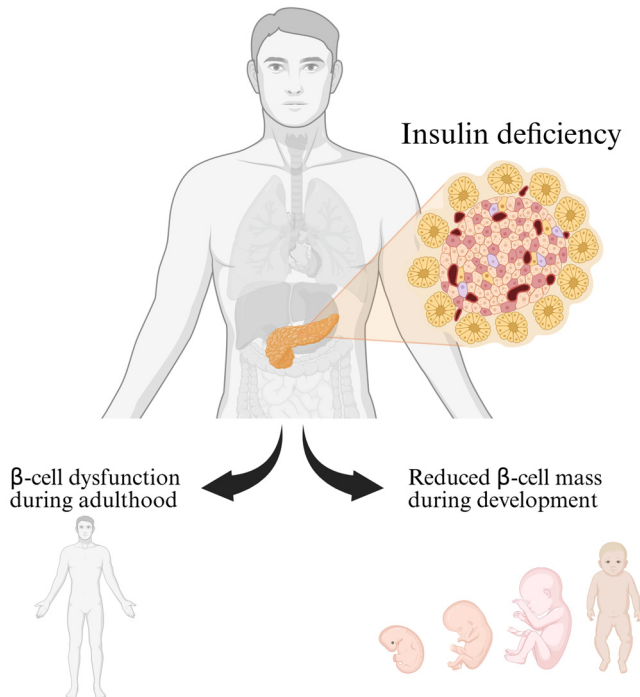

Figure 3. Origins of insulin deficiency in type 2 diabetes.
Insulin deficiency can arise from impairments in pancreatic β-cell function and/or mass. In adulthood, this may result from β-cell dysfunction, including processes such as dedifferentiation and apoptosis, as well as other mechanisms that contribute to a reduction in β-cell mass. Alternatively, insulin deficiency may reflect a lower β-cell mass established earlier in life, including during fetal development. Together, these pathways highlight that insulin deficiency in type 2 diabetes can originate from both developmental programming and acquired β-cell dysfunction across the lifespan.

Assessment of Insulin Sensitivity and β-Cell Function

Given the central role of β-cell dysfunction and insulin resistance in the development of T2D, reliable methods to assess these processes are essential in both clinical and research settings. In clinical practice and epidemiological studies, glucose metabolism is commonly evaluated using fasting measurements of glucose, insulin and/or C-peptide, or by performing an oral glucose tolerance test (OGTT).

Commonly used indices include the fasting-based Homeostatic Model Assessment (HOMA), including its updated version HOMA2, and the OGTT-derived Corrected Insulin Response (CIR) (30, 31). These methods provide indirect but practical measures of glucose metabolism and pancreatic β-cell function and are therefore frequently used in epidemiological and clinical studies of diabetes.

OGTT

OGTT is commonly used to assess glucose metabolism. After an overnight fast, participants ingest a standardized glucose load, typically 75 g of glucose dissolved in water. Blood samples are collected at defined time points, most commonly at baseline and two hours after ingestion, to measure plasma glucose and insulin or C-peptide concentrations (32). These measurements allow evaluation of glucose tolerance and can also be used to derive indices of insulin secretion and insulin sensitivity.

HOMA indices

HOMA is a fasting-based method used to estimate insulin resistance and β-cell function from fasting glucose and insulin concentrations (31). HOMA-β provides an estimate of pancreatic β-cell function under fasting conditions, reflecting basal insulin secretion. The updated HOMA2 model accounts for nonlinear relationships between glucose and insulin and provides estimates of insulin resistance (HOMA2-IR) and β-cell function (HOMA2-β). HOMA2 can also incorporate C-peptide as a measure of endogenous insulin secretion (31). Detailed calculations of these indices are provided in the Methods section.

CIR

In contrast to fasting-based indices, OGTT-derived measures can provide insight into dynamic insulin secretion following glucose stimulation. One such index is the Corrected Insulin Response (CIR), which estimates early-phase insulin secretion based on glucose and insulin concentrations measured during the OGTT. Higher CIR values indicate a greater early insulin secretory response to glucose (33). The calculation of CIR is described in detail in the Methods section.

Cancer

Cancer is a complex, multifactorial disease that arises from the interaction of genetic susceptibility, somatic mutations, and environmental influences. A hallmark of many cancers is profound metabolic dysregulation, reflecting the reprogramming of cellular metabolism required to sustain uncontrolled proliferation and survival (1). Increasing evidence suggests that alterations in gene dosage and regulatory variation can contribute to these processes through mechanisms such as allelic expression imbalance (AEI), in which the two alleles of a gene inherited from each parent are expressed at unequal levels. AEI may arise from cis-regulatory variation, epigenetic mechanisms, or chromosomal dosage effects, thereby influencing disease phenotypes (34-36).

Acute Lymphoblastic Leukemia

These regulatory perturbations are particularly relevant in hematological malignancies such as acute lymphoblastic leukemia (ALL), the most common childhood cancer, accounting for approximately 25% of all childhood cancers (37). ALL is characterized by widespread chromosomal alterations that affect gene dosage and transcriptional regulation, leading to uncontrolled proliferation of immature lymphoid cells, bone marrow failure and potentially life-threatening disease (34, 38, 39). Studies have shown that inherited variants and chromosomal abnormalities such as copy number variations (CNVs) can generate AEI in leukemic cells, suggesting that allele-specific regulatory effects may represent a general mechanism linking genetic variation to altered transcriptional output in both cancer and metabolic tissues.

Genetic architecture of type 2 diabetes

Polygenic Risk and Heritability

As mentioned earlier, T2D arises from an interplay between genetic and environmental factors and represents a complex polygenic disease with a substantial heritable component. Family and twin studies have estimated the heritability of T2D to range from approximately 30% to 72%, indicating a substantial genetic contribution to disease susceptibility (40-42). Twin studies have played a central role in estimating the heritability of T2D. Studies consistently show that T2D occurs more frequently in monozygotic twins, who share nearly identical genomes, than in dizygotic twins, who share approximately half of their segregating genetic variation (11, 43, 44). These findings support a substantial genetic contribution to disease risk. Further evidence comes from population studies showing marked differences

in the prevalence of T2D among ancestral groups. Importantly, such differences have been observed even among populations living in similar environments, suggesting that inherited genetic variation contributes to disease susceptibility beyond environmental influences alone (45, 46).

Common and Rare Variant Architecture

Genome-wide association studies (GWAS) based on multi-ancestry populations have identified many genetic variants associated with T2D. As of 2024, 1,289 independent association signals mapping to 611 loci have been reported (47-50). Although some variants are rare and confer relatively large effect sizes, most identified loci are represented by common variants with modest individual effects. Collectively, these variants explain only approximately 50% of the estimated heritability of T2D (11, 49).

The substantial proportion of unexplained heritability is commonly referred to as "missing heritability" and has prompted considerable debate regarding the underlying genetic architecture of T2D (11). One hypothesis proposes that the remaining heritability resides in a very large number of additional common variants with small additive effects, consistent with a highly polygenic model in which T2D represents the extreme ends of a continuous liability distribution. In contrast, an alternative hypothesis suggests that rare variants with larger effects may underlie signals detected at common loci and thereby account for a substantial share of the missing heritability (11, 51, 52). Importantly, the predominance of common variants in GWAS findings largely reflects statistical constraints. Association testing of low-frequency alleles requires substantially larger sample sizes than are needed to detect common variants with small effects. Consequently, rare variants may be underrepresented in current datasets due to limited statistical power rather than absence of biological relevance.

An argument frequently raised against both common and rare deleterious risk variants is that such alleles would be expected to undergo purifying selection and therefore be eliminated from the population. However, this reasoning is less compelling in the context of T2D, where disease penetrance depends strongly on environmental exposures. Genetic variants that increase susceptibility in today's obesogenic and sedentary environment may previously have been neutral or even advantageous under historical conditions characterized by food scarcity and higher energy expenditure (11).

Together, these observations support a highly polygenic architecture in which numerous common variants of small effect and rarer variants of larger effect contribute to disease susceptibility within an environment-dependent framework of risk.

Importantly, both common and rare variants are now thought to exert much of their effect through regulatory mechanisms rather than through alterations in protein-coding sequences. Indeed, the majority of T2D-associated variants map to non-coding regions of the genome and frequently overlap regulatory elements such as enhancers and promoters that are active in metabolically relevant tissues, particularly pancreatic islets. Consequently, understanding how genetic variation influences gene regulation has become a central challenge in deciphering the molecular basis of T2D susceptibility.

Tissue-Specific Gene Regulation and Disease Pathways

Although all cells in the body share an identical DNA sequence, they differ markedly in structure and function due to tightly regulated, tissue-specific gene expression programs established during development and differentiation. These programs enable cells to acquire specialized functions, in contrast to the pluripotent state of embryonic stem cells and are maintained through cell-type–specific regulatory landscapes. Consequently, the functional impact of a genetic variant depends not only on its genomic location but also on the regulatory context in which it acts.

Genetic risk for T2D is mediated through multiple biological pathways, including insulin secretion, insulin resistance, adipogenesis, and β-cell survival. Notably, variants influencing insulin secretion and β-cell function appear to play a central role in disease pathogenesis, consistent with evidence that β-cell failure is a key determinant of disease onset and progression. This diversity of genetic mechanisms likely contributes to the pronounced heterogeneity observed among individuals with T2D and may account for differences in clinical presentation, disease progression, and treatment response.

GWAS typically identify statistical associations at the level of individual variants rather than the genes or biological mechanisms they influence. As many disease-associated variants reside in non-coding regions, interpreting GWAS signals therefore requires an understanding of how regulatory variation affects gene expression in relevant tissues. In the context of T2D, this is particularly important in pancreatic islets, where genetic variation can influence β-cell development, insulin secretion, and cellular function.

Regulatory Variation

Overview of Regulatory Elements

In addition to alterations in the coding regions of the DNA sequence itself, gene expression is strongly influenced by regulatory variation. Regulatory variation refers to differences in genomic elements that control when, where, and to what extent genes are expressed. These elements include promoters, enhancers, silencers, insulators, non-coding RNAs, and epigenetic modifications, all of which contribute to the fine-tuning of transcriptional regulation (53-55).

Promoters

Promoters are regulatory DNA regions located upstream of genes and serve as the primary sites for transcription initiation. They contain specific sequence motifs recognized by transcription factors and the basal transcriptional machinery, including RNA polymerase II. Promoters define the transcription start site and direction of transcription and largely determine baseline levels of gene expression (53-55).

Enhancers

Enhancers are distal regulatory DNA elements that increase gene expression by stimulating promoter activity. They can be located upstream, downstream, or within introns of their target genes and may act over large genomic distances. Enhancers function by binding transcription factors and co-activators and by facilitating chromatin looping, enabling physical interactions with promoters that enhance transcriptional efficiency (53-55).

Silencers

Silencers are regulatory DNA elements that repress gene expression. Like enhancers, they act through the binding of regulatory proteins and can be located at variable distances from the promoter. Silencers reduce transcription by inhibiting promoter activation, interfering with transcription factor binding, or promoting chromatin states that are less accessible to the transcriptional machinery. Silencers are enriched for expression quantitative trait loci (eQTLs) and disease-associated variants (53-55).

Together, promoters, enhancers, and silencers form an integrated regulatory network that enables precise spatial, temporal, and quantitative control of gene expression. Genetic variation within these regulatory elements can lead to

substantial differences in gene expression without altering the underlying protein-coding sequence, thereby contributing to phenotypic variation and disease susceptibility.

Gene Expression Analysis

Linking Genetic Variation to Function

While GWAS have identified numerous genetic variants associated with T2D, these associations alone are insufficient to explain the molecular mechanisms underlying disease development. To better understand how genetic variation contributes to phenotypic variation, it is essential to investigate how regulatory variants influence gene expression in disease-relevant tissues, such as pancreatic islets. Gene expression analyses provide a critical link between genetic variation and biological function by revealing which genes are transcribed and how their expression modulated by genetic and environmental factors.

eQTL Analysis

One widely used approach to study regulatory variation is expression quantitative trait locus (eQTL) analysis, which identifies genetic variants associated with changes in total gene expression in nearby genes (56, 57).

Allele-specific Expression Analysis

Complementary to eQTL analysis, allele-specific expression (ASE) analysis quantifies the expression of the two alleles of a gene within the same individual, enabling the detection of allelic expression imbalance (AEI) (58, 59).

AEI refers to unequal expression of the two alleles of a gene in heterozygous individuals, ranging from partial allelic imbalance to complete monoallelic expression, and reflects cis-acting regulatory variation or gene dosage effects (Figure 4). Because both alleles are measured within the same cellular context, AEI provides a direct link between genetic variation and gene regulation. AEI can arise from multiple regulatory mechanisms, including classical eQTLs, splice QTLs, non-coding RNA-mediated regulation, copy number variation, and genomic imprinting or parent-of-origin effects.

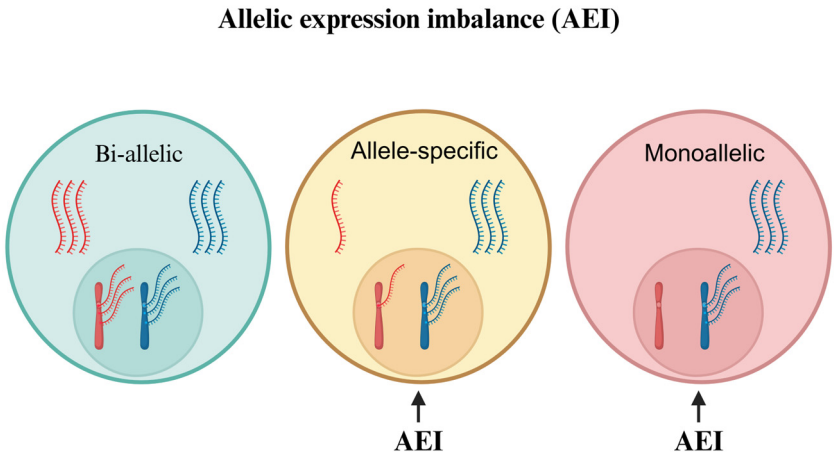

Figure 4. Schematic representation of allelic expression imbalance (AEI).
Allelic expression imbalance refers to unequal expression of maternal and paternal alleles within the same cell. Under bi-allelic expression, both alleles are expressed at comparable levels. In contrast, allele-specific expression reflects preferential expression of one allele over the other, while monoallelic expression represents complete silencing of one allele. AEI encompasses both partial (allele-specific) and complete (monoallelic) imbalance, providing a framework for identifying cis-regulatory variation and gene dosage effects that influence transcriptional output.

Parent-of-Origin Effects

Definition and Mechanisms

Parent-of-origin effects (POEs) refer to situations in which the phenotypic impact of a genetic variant depends on whether it is inherited from the mother or the father. Early embryological studies in mice demonstrated that maternal and paternal genomes are not functionally equivalent, as embryos containing only maternal or only paternal genomes fail to develop normally (60, 61). Consistent with these observations, studies of imprinted loci, most notably insulin-like growth factor 2 (***IGF2***), showed that paternal expression is essential for normal fetal growth, while the maternal allele is silenced (62). In contrast, the insulin-like growth factor 2 receptor (***IGF2R***) is maternally expressed and limits IGF2-mediated fetal growth by binding and promoting the degradation of IGF2, illustrating the reciprocal and often antagonistic effects of maternally and paternally expressed imprinted genes during development (63). Together, these findings contributed to the identification of genomic imprinting, an epigenetic mechanism that typically results in parent-specific monoallelic expression and, in some cases, biased expression of one parental allele (64, 65).

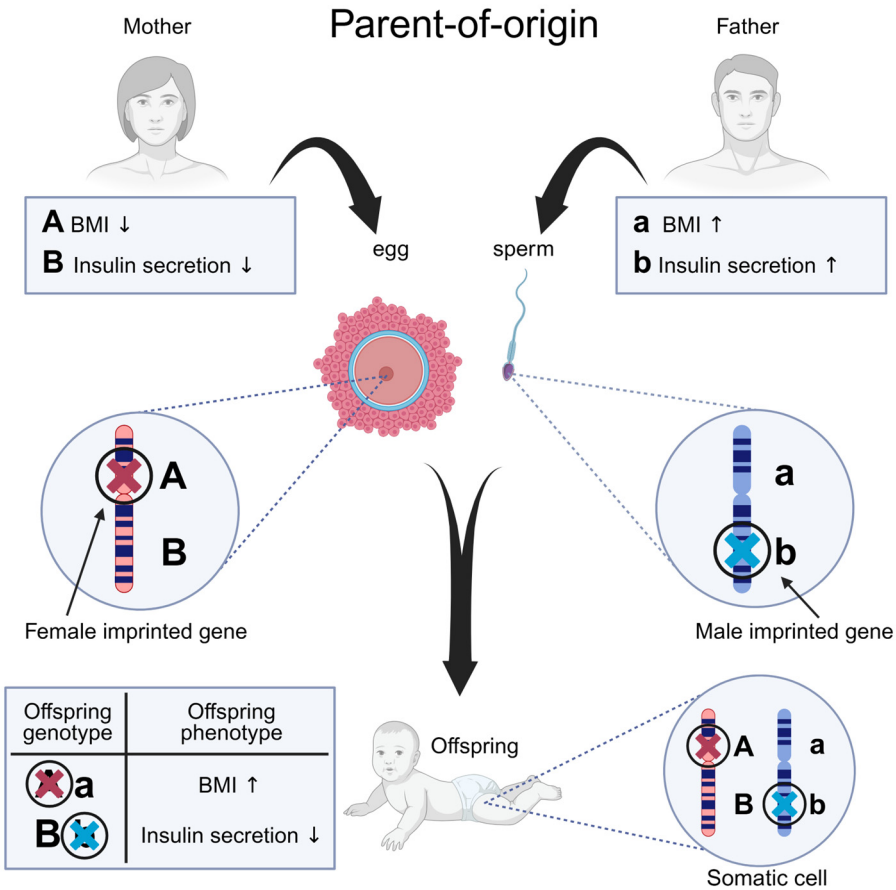

Figure 5. Illustration of parent-of-origin effects mediated by genomic imprinting.
Offspring inherit one allele from each parent at each locus; however, for imprinted genes, only one parental allele is expressed while the other is silenced. In this example, BMI is determined by the paternally expressed allele, whereas insulin secretion is determined by the maternally expressed allele. As a result, the offspring exhibits increased BMI and reduced insulin secretion. Uppercase and lowercase letters represent alternative alleles and do not imply dominance.

As illustrated in Figure 5, the mother and father carry alternative alleles associated with opposing metabolic traits, including differences in body mass index (BMI) and insulin secretion. Although the offspring inherits one allele from each parent at each locus, only one parental copy is expressed for a given imprinted gene, while the other is silenced. In this example, BMI is determined by the paternally expressed allele, whereas insulin secretion is determined by the maternally expressed allele. Consequently, the offspring exhibits increased BMI and reduced insulin secretion, reflecting POEs arising from parent-specific gene expression. This illustrates how differential expression of maternal and paternal alleles can lead to parent-of-origin-

dependent effects on metabolic phenotypes relevant to type 2 diabetes and provides a biological explanation for why disease risk may vary depending on parental inheritance.

While this example of genomic imprinting represents the classical mechanism underlying POEs, such effects can also arise when both parental alleles are expressed but at different levels. In these cases, one allele is preferentially expressed over the other, resulting in a parent-specific expression bias rather than complete silencing. This can still lead to phenotypic differences depending on parental origin and may contribute to variation in complex traits, including type 2 diabetes (66).

POEs are of particular interest in type 2 diabetes, as they may contribute to the well-documented preferential maternal transmission of the disease. Epidemiological studies have shown that the lifetime risk of type 2 diabetes is higher in individuals whose mothers are affected compared to those with unaffected mothers or affected fathers (40, 41, 67, 68), suggesting a role for imprinting or other parent-specific regulatory mechanisms in disease susceptibility.

Imprinted genes

Imprinted genes represent a small subset of the genome that exhibit parent-specific expression, whereby only one parental allele is expressed while the other is silenced. In humans, approximately 100–200 imprinted genes have been identified, although this number may vary depending on tissue type and developmental stage (64, 65). Their expression is regulated by epigenetic mechanisms, most notably DNA methylation at imprinting control regions established during gametogenesis and maintained throughout development (64). Imprinted genes often display tissue-specific and temporally restricted expression patterns and are enriched in pathways related to growth, development, and metabolic regulation (69).

Beyond known and predicted imprinted genes

While imprinted genes represent the classical mechanism underlying POEs, emerging evidence suggests that POEs may extend beyond known imprinted regions. Studies in mouse models have demonstrated that a substantial proportion of phenotypes exhibit parent-of-origin-dependent variation, even when they cannot be explained by known imprinted genes, although such effects may be partly explained by shared genetic background and environmental factors within families (70). One explanation for this broader landscape is that parent-specific effects may be context-dependent, arising only in particular tissues, developmental stages, or physiological conditions, and are therefore not consistently captured in standard analyses. In line with this, recent large-scale human genetic studies have demonstrated that such effects can be detected across diverse traits, suggesting that

POEs may be more widespread than previously appreciated, and may arise from additional, potentially context-dependent regulatory mechanisms beyond classical imprinting (71, 72). Experimental evidence further suggests that interactions between imprinted and non-imprinted genes can generate complex POEs, supporting a network-based mechanism underlying these effects (73). However, other studies have reported that POEs are largely restricted to known imprinted regions, highlighting ongoing uncertainty in the field (74). Notably, many weak parent-of-origin signals detected outside these regions are located near known imprinted domains, suggesting that they may reflect local chromatin context and tissue-specific regulatory effects rather than independent imprinting (74). Consistent with this, our group had previously identified a POE at the ***THADA*** locus, a gene implicated in type 2 diabetes susceptibility, where the phenotypic impact of genetic variation differed depending on parental inheritance. Importantly, this locus is not located within a canonical imprinted region, suggesting that POEs can occur outside classical imprinted domains and may reflect alternative regulatory mechanisms influencing metabolic traits (75). Comprehensive characterization of POEs remains challenging, in part due to the limited availability of large family-based cohorts with sufficient statistical power across diverse traits (76). As parental origin cannot be inferred without parental genotypes, many large-scale genetic studies are not well suited to detect such effects. Nevertheless, recent large-scale analyses have begun to overcome these limitations and have identified POEs at loci outside known imprinted regions, suggesting that additional regulatory mechanisms beyond classical imprinting contribute to these effects (71, 76).

Parent-of-Origin Effects and Developmental Programming

POEs may also intersect with developmental programming, whereby early-life exposures influence long-term metabolic health. In addition to genetic inheritance, maternal factors can shape offspring phenotype through the intrauterine environment, including nutrient availability, hormonal milieu, and maternal metabolic status (77). This concept of developmental plasticity is central to the Developmental Origins of Health and Disease (DOHaD) hypothesis, which proposes that environmental conditions during critical periods of development can have lasting effects on disease risk (77). For example, exposure to nutrient scarcity in utero may induce adaptive responses that optimize survival under limited resource conditions. However, these adaptations may become maladaptive later in life when the postnatal environment is characterized by nutrient abundance, thereby increasing the risk of metabolic disorders such as type 2 diabetes. Similarly, exposure to a hyperglycemic intrauterine environment, such as in gestational diabetes mellitus (GDM), can lead to fetal hyperinsulinemia and altered metabolic programming, predisposing offspring to obesity and impaired glucose metabolism later in life (78). Interestingly, in the context of type 2 diabetes, epidemiological studies have consistently reported a stronger maternal than paternal transmission of

disease risk (40, 67, 68), suggesting that both maternal genetic effects and intrauterine influences contribute to susceptibility. However, paternal factors may also play a role, as genetic variation as well as environmental exposures such as age, smoking, and BMI can influence epigenetic marks in sperm and thereby affect offspring metabolic phenotypes (79-81). These paternal effects highlight that developmental programming is not exclusively maternal and may involve contributions from both parents.

Evidence for POE in insulin deficiency and T2D

Evidence for POEs in type 2 diabetes and related traits has been demonstrated at several loci, most notably ***KCNQ1***, which was initially identified through GWAS as a susceptibility gene for T2D (82). Subsequent studies revealed a strong POE at this locus, with risk alleles conferring increased disease susceptibility depending on parental inheritance (72, 83). These findings have been replicated in multiple populations, including in our own studies (75).

Mechanistically, variants at the ***KCNQ1*** locus have been shown to be monoallelically expressed and differentially methylated in fetal pancreatic islets, while exhibiting biallelic expression in adult islets, indicating a temporally regulated imprinting pattern (84). This suggests that parent-of-origin-dependent gene regulation during fetal development may influence the establishment of β-cell mass. Supporting this, paternal mutations at the ***KCNQ1*** locus have been shown to reduce pancreatic β-cell mass through epigenetic modification of downstream targets (85).

Beyond ***KCNQ1***, additional loci such as ***GRB10*** have demonstrated POEs on glucose metabolism, with paternal inheritance associated with increased T2D risk and maternal inheritance associated with reduced fasting glucose (86). More recently, a large-scale genome-wide analysis by Hofmeister et al. identified POEs across multiple growth- and metabolism-related traits, including T2D, glucose-related phenotypes, and HbA1c, further supporting a broader contribution of parent-specific genetic regulation to metabolic disease susceptibility (71). Furthermore, our studies of gene expression in human pancreatic islets have identified AEI in the ***PEG3*** gene, as well as a POE at the ***THADA*** locus, characterized by excess maternal transmission and differential methylation (75, 87).

Together, these findings provide strong evidence that POEs contribute to the genetic architecture of insulin secretion, β-cell function, and the pathogenesis of T2D.

Evolutionary perspective

From an evolutionary perspective, genomic imprinting and the resulting POEs are thought to arise from selective asymmetries between maternally and paternally inherited alleles. The most widely accepted explanation is the **kinship (or parental**

conflict) theory, proposed by Haig and Moore, which suggests that imprinting evolved to modulate gene expression in a way that reflects differing evolutionary interests of maternally and paternally inherited alleles. This conflict arises because genes expressed in the fetus can influence how much nutrient is transferred from the mother during pregnancy, thereby affecting fetal growth and long-term metabolic health. According to this model, paternally expressed genes tend to promote fetal growth and nutrient acquisition, whereas maternally expressed genes act to limit resource use (88, 89).

Alternative, non-mutually exclusive hypotheses have also been proposed. The **sexual antagonism theory** suggests that imprinting may evolve due to sex-specific differences in optimal trait expression, thereby helping to resolve intralocus sexual conflict (90, 91). In contrast, the **maternal-offspring co-adaptation theory** proposes that imprinting facilitates coordinated interactions between maternal and offspring traits, potentially favoring expression of the parental allele that optimizes fitness within this interaction (92). In this model, imprinting helps ensure that offspring traits are better matched to the maternal environment.

A common feature of these models is that they invoke selective asymmetry between maternally and paternally inherited alleles at a given locus, leading to the evolution of differential, parent-specific gene expression (93). Together, these evolutionary frameworks provide a basis for understanding why POEs are enriched in pathways related to growth and metabolism and how their dysregulation may contribute to metabolic diseases such as T2D.

Overall aim and objectives

This thesis aimed to investigate the role of parent-of-origin effects (POEs) in the regulation of metabolic disease risk, with a particular focus on type 2 diabetes (T2D) susceptibility and insulin secretion. The methodological framework centered on allele-specific expression (ASE) analysis as a means of detecting allelic expression imbalance (AEI) through large-scale screening of genes in adult and embryonic pancreas. Identified candidate genes were subsequently evaluated in large family-trio and population-based genotype cohorts to assess POEs on T2D risk and insulin secretion, with the latter serving as a measure of β-cell function and its developmental regulation. In addition, one study demonstrated that ASE-based approaches are applicable beyond T2D, including acute lymphoblastic leukemia (ALL).

Paper 1

This study aimed to investigate parent-of-origin effects on metabolic traits related to type 2 diabetes risk across the life course in an Indian population, and to evaluate whether previously identified POE variants from European populations show similar effects.

Paper 2

This study aimed to characterize allelic expression imbalance in human pancreatic islets and investigate its relevance to type 2 diabetes pathogenesis, including the contribution of cis-regulatory variation, imprinted genes, and parent-of-origin effects to β-cell dysfunction and disease risk.

Paper 3

This study aimed to explore AEI in fetal pancreatic tissue in comparison with adult islets and to investigate genetic POE and its relevance to β-cell development and type 2 diabetes risk.

Paper 4

This study aimed to characterize the allelic expression imbalance landscape in high hyperdiploid acute lymphoblastic leukemia and investigate how copy number variations influence allele-specific gene expression in cancer.

Material and methods

Study cohorts

Type 2 diabetes (T2D) is a heterogeneous disorder that arises through interactions between inherited susceptibility and environmental exposures. To disentangle these contributing factors, large and well-phenotyped population cohorts have been established. Such cohorts constitute the empirical basis for genetic, epidemiological, and metabolic investigations aimed at understanding T2D and related cardiometabolic traits.

The present thesis utilizes data from several family- and population-based cohorts, as well as tissue biobanks. These cohorts differ in terms of demographic composition, genetic background, geographical origin, and overall sample size, thereby providing complementary perspectives for studying disease mechanisms. The cohorts included are:

Family cohorts:

1. The Botnia cohort
2. Hungarian Transdanubian biobank
3. Pune Maternal Nutrition Study

Unrelated and population-based cohorts:

1. ANDIS
2. PPP

Tissue biobanks:

1. HTL
2. Fetal
3. HeH ALL

Within each project included in this thesis, specific participant subsets were selected from these broader cohorts. The inclusion criteria varied according to the scientific question being addressed and the analytical strategy employed. A concise overview of the cohorts is provided below, while more detailed descriptions of cohort characteristics and study-specific selection procedures are presented in the corresponding project sections.

Family cohorts

The Botnia cohort

The Botnia Study was initiated in 1990 with the aim of investigating non-insulin dependent diabetes mellitus. Recruitment initially targeted all individuals diagnosed with T2D attending five primary healthcare centers in the Botnia region of western Finland. Family members of these individuals were subsequently invited to participate, creating a family-based study design that enabled extended genetic and metabolic analyses (40).

Over time, the cohort was expanded beyond the original recruitment area to include additional regions in Finland as well as southern Sweden. The study population is predominantly of European ancestry, with Finnish and Swedish individuals comprising most participants.

Participants underwent detailed clinical and metabolic assessments, including biomarker measurements. In addition to diabetes-related variables, broader phenotypic and molecular data were collected to support investigations of metabolic dysfunction and genetic variation. Owing to its recruitment of individuals with T2D together with their relatives, the Botnia Study is regarded as a large family-based cohort.

The Botnia dataset includes previously performed imputed genome-wide association data from 9,160 individuals organized in a family trio design, comprising 5,746 parents and 3,414 offspring. Genotyping was performed using the Illumina Global Screening Array (GSA-24v1), and imputation was carried out on the Michigan Imputation Server using the HRC reference panel (GRCh38).

For the purposes of this thesis, a defined subset of participants from the Botnia cohort was selected and used for genetic investigations and association analyses across several included projects.

All participants provided informed consent, and ethical approvals were obtained from the ethics committees of Helsinki University Central Hospital and Lund University.

Hungarian Transdanubian biobank

The Hungarian Transdanubian Biobank (HTB) was initiated in 1992 at the Hungarian Heart Center in Balatonfüred. The biobank comprises clinical data and biological samples from individuals with type 2 diabetes and their family members. It includes more than 9,000 participants from the western region of Hungary, primarily organized as family trios.

Genotyping data were previously generated for 9,194 individuals using the Sequenom MassARRAY iPLEX platform, including 4,592 parents and 4,602 offspring. Among the genotyped individuals, 16.6 percent had a diagnosis of T2D according to the WHO 1998 criteria.

All participants provided informed consent, and ethical approvals were obtained from the local ethics committees.

Pune Maternal Nutrition Study

The Pune Maternal Nutrition Study (PMNS) is a rural, community-based preconception birth cohort established in the early 1990s in six villages near Pune, India. The study was initiated to examine how maternal nutrition and physical activity influence fetal development and long-term metabolic risk including diabetes. The region was selected because of its agrarian economy, limited irrigation, and recurrent drought conditions that contributed to widespread undernutrition at the time of recruitment. Since then, the area has undergone substantial socioeconomic development and urbanization.

Between 1993 and 1996, 2,466 married, non-pregnant women aged 15 to 40 years were enrolled prior to conception. Women who conceived a singleton pregnancy before 21 weeks of gestation were followed prospectively, together with their spouses and offspring. The cohort initially included mothers who experienced nutritional constraints during pregnancy, with longitudinal follow-up of their children. Over time, the study expanded to include subsequent generations, now comprising F0 mothers, their F1 offspring, and the F2 generation. Most participating families were Hindu, and the prevalence of gestational diabetes at recruitment was below 1 percent according to diagnostic criteria used at that time.

Participants have undergone repeated clinical, anthropometric, metabolic, cognitive, and epigenetic assessments from birth through adulthood, with major follow ups at 6, 12, 18, and 24 years. This multigenerational prospective design provides a unique opportunity to study developmental and life course influences on metabolic health.

Genome wide genotyping of PMNS family trios were previously performed using Affymetrix SNP 6.0 arrays. Variants were excluded for more than 5 percent missingness, minor allele frequency below 1 percent, or deviation from Hardy

Weinberg equilibrium at $p < 0.05$. Additional quality control included call rate thresholds, allele consistency checks, and Mendelian error assessment. Imputation was performed using the TOPMed reference panel release 3 GRCh38 with Eagle v2.4 for phasing and Minimac4 for imputation, retaining variants with imputation scores above 0.4. The analyses in this thesis included 914 individuals at the 24 year follow up, comprising 425 parents and 489 offspring in a trio design.

All participants provided informed consent, and ethical approvals were obtained from King Edward Memorial Hospital Research Centre Ethics Committee as well as village leadership authorities.

Unrelated and population-based cohorts:

ANDIS

The All New Diabetics in Scania (ANDIS) project is an ongoing population-based cohort established in Scania County, southern Sweden. Individuals newly diagnosed with diabetes across the region are invited to participate through 194 regional healthcare centers within the first year after diagnosis. The initial phase of the study, ANDIS1, enrolled participants between January 1, 2008, and November 3, 2016, whereas the subsequent phase, ANDIS2, included individuals recruited from November 4, 2016, to April 6, 2022. Approximately 90% of the participants are of European ancestry. Participants were excluded if they were younger than 18 years at the time of diagnosis, had a history of pancreatitis, or lacked complete data for the clustering variables.

Genotyping data from the ANDIS cohorts were generated using the Illumina Infinium CoreExome-24 v1.1 array (ANDIS1) and the Illumina Global Screening Array (GSA-MD v3.0) (ANDIS2). Quality control excluded samples with sex discrepancies, duplicate samples, and excessive relatedness. Prior to imputation, samples and variants failing quality control criteria were excluded based on missingness thresholds (>0.02), deviation from Hardy-Weinberg equilibrium ($p < 1\times10^{-8}$), and low minor allele frequency (<0.001). Genotype imputation was performed using the TOPMed reference panel (GRCh38/hg38), and variants with imputation quality Rsq > 0.3 were retained for downstream analyses.

The study was conducted in accordance with relevant ethical guidelines and regulations. Ethical approval for the ANDIS project was obtained from the Regional Ethics Review Committee in Lund (approval numbers 584/2006, 2011/354, 2011/367, 2012/676, 2014/198, and 904/2017) as well as from the Swedish Ethical Review Authority (2021/04647 and 2023/08012). Written informed consent was obtained from all participants, and no financial compensation was provided.

PPP

The Prevalence, Prediction and Prevention of Diabetes (PPP–Botnia) Study is a population-based cohort established in western Finland between 2004 and 2008 to examine the occurrence and determinants of T2D, impaired glucose regulation, and metabolic syndrome, with the broader aim of informing preventive strategies. Individuals aged 18 to 75 years were randomly recruited from the Finnish Population Registry, representing approximately 6–7% of the regional population, and 5,208 participants were included in the study (54.7% response rate). Written informed consent was obtained from all participants, and ethical approval was granted by the Ethics Committee of Helsinki University Hospital, Finland. Genotyping was carried out using the ThermoFisher (Affymetrix) Axiom FinnGen1 array. Samples were subjected to quality control procedures excluding individuals with discordant sex information, non-Finnish ancestry, excessive heterozygosity (±4 SD), duplicate samples, or non-canonical chromosomes. Variants were filtered before imputation by excluding markers with sample missingness above 5%, variant missingness above 2%, or deviation from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$). Genotype imputation was performed using the SISu v3 reference panel aligned to GRCh38/hg38, and poorly imputed variants ($R^2 < 0.4$) were removed. In the current analysis, only participants without diabetes were included, resulting in a final cohort of 4,599 genotyped individuals.

Tissue biobanks

Human pancreatic islets (HTL)

Pancreatic islets from 261 donors, including individuals with T2D (n = 44) and non-diabetic controls (n = 217), were obtained through the EXODIAB network and the Nordic Network for Clinical Islet Transplantation. Ethical approval for all procedures was granted by the Lund University ethics committee.

Total RNA, including miRNA, was extracted using miRNeasy or AllPrep DNA/RNA Mini Kits (QIAGEN) according to established protocols. RNA quality and integrity were evaluated using Agilent Bioanalyzer or TapeStation systems, while RNA concentration was measured using NanoDrop or Qubit instruments. RNA-seq libraries were generated from 1 µg of high-quality RNA using TruSeq library preparation kits and sequenced on Illumina HiSeq 2000 or NextSeq 500 platforms.

Glucose treated islets (GTI)

Pancreatic islets from 45 donors, including 31 normoglycemic and 14 hyperglycemic individuals, were obtained through the EXODIAB network via the Nordic Network for Clinical Islet Transplantation (a subset of HTL). For each donor, 3,000 to 5,000 islets were divided into two aliquots and cultured for 24 hours at 37°C in CMRL 1066 medium (ICN Biomedicals) containing either 5.5 mmol/L or 18.9 mmol/L glucose. All donors had provided informed consent for organ donation for medical research, and the study was approved by the Ethics Committee at Lund University, Malmö, Sweden (permit no. 2011/263).

Fetal tissue dataset

Fetal tissue samples were obtained from terminated fetuses between 5 and 14 gestational weeks, including pancreas (n = 16), kidney (n = 10), liver (n = 10), and skin (n = 6). All participating women provided informed consent, and the study was approved by the Regional Ethics Committee in Lund (Dnr 2019/593, 2015/241, and 2018-579).

RNA was extracted using TRIzol and RNeasy protocols, and RNA quality was assessed using the Agilent 2200 TapeStation system. RNA libraries were prepared using the TruSeq RNA Library Preparation Kit (Illumina), followed by paired-end sequencing (101 bp) on Illumina HiSeq 2000 or NextSeq 550 platforms.

RNA-seq data were processed using the GTEx V8 expression quantification pipeline. Reads were aligned to the human reference genome GRCh38/hg38 with Gencode v26 annotation using STAR v2.5.3a, and gene expression quantification was performed using RSEM and RNASeQC. Batch correction was performed using ComBat-seq.

Australian dataset

RNA sequencing of human pancreatic islets (N = 66; 1 T2D) was performed on the Illumina HiSeq 4000 platform using 150 bp paired-end sequencing. The dataset, generated by our collaborators Anandwardhan A. Hardikar and Wilson K. M. Wong from the Diabetes and Islet Biology Group, School of Medicine, Western Sydney University, Australia, was publicly available and retrieved from the NCBI Gene Expression Omnibus (GEO) under accession number GSE152111. Samples S150 and S418 originated from the same individual; therefore, only S418, which had slightly higher sequencing coverage, was retained for downstream analyses, resulting in a final dataset of 65 samples.

HeH ALL dataset

Copy number calling data for cancer cell lines were obtained from the Cancer Cell Line Encyclopedia (CCLE; https://sites.broadinstitute.org/ccle/). This work was supported by grants from the Swedish Childhood Cancer Fund (PR2020-0033 and TJ2020-0024) and the Crafoord Foundation (20230778 and 20240747).

This study was based exclusively on publicly available datasets. Next-generation sequencing data for 18 extensively characterized cancer cell lines from the CCLE were retrieved from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP186687. Whole-genome sequencing (WGS) data for the HepG2 cell line were accessed through ENCODE (https://www.encodeproject.org/) under accession number ENCBS760ISV. HepG2 RNA-sequencing data were obtained from ENCODE under accession number ENCLB707CLT and from the NCBI SRA under accession numbers SRR8616129 and SRR629573. WGS and RNA-seq data for the 12 high hyperdiploid acute lymphoblastic leukemia (HeH ALL) samples included in this study were downloaded from the European Genome-phenome Archive (EGA; https://ega-archive.org/) under accession numbers EGAD00001008988 and EGAD00001002112, respectively. All analysis scripts and code are publicly available on GitHub (https://github.com/biomedswe/BASE).

Definition of Diabetes in study cohorts

World Health Organization (WHO)

"A chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar)."

Extended inhouse T2D definition in HTL dataset

T2D status was defined based on either a clinical diagnosis (n = 40) or HbA1c levels above 6.5% (n = 4).

Calculation of β-cell function indices

HOMA2

The updated Homeostatic Model Assessment (HOMA2) model was used to estimate β-cell function (HOMA2-β). This index was calculated using the HOMA2 calculator (Diabetes Trials Unit, University of Oxford, UK), which accounts for nonlinear relationships between glucose and insulin and incorporates variations in

hepatic and peripheral glucose resistance (31, 94). Fasting plasma glucose and fasting serum insulin were used as inputs.

CIR

Early-phase insulin secretion was assessed using the Corrected Insulin Response (CIR), calculated from glucose and insulin concentrations measured 30 minutes after oral glucose ingestion during the OGTT according to the following formula:

$$\text{CIR} = \frac{100 \times I_{30}}{G_{30} \times (G_{30} - 3.89)}$$

where I_{30} and G_{30} represent plasma insulin (μU/mL) and glucose (mmol/L), respectively (33).

POE genetic analysis

In family cohorts

Parent-of-origin effects (POEs) were investigated using genotyped family trios from the Botnia, PMNS and HTB cohorts. Genotype data for candidate variants were extracted from the full dataset using PLINK (v.2.0), and the parental origin of alleles in the offspring was determined through phasing based on parental genotypes using a custom in-house script. This enabled assignment of maternally and paternally inherited alleles for each variant.

Statistical analyses were performed to evaluate POEs by comparing the phenotypic effects of maternally versus paternally inherited alleles. Generalized estimating equations (GEE), as implemented in the *geepack* R package, were used to account for familial correlations. The appropriate distribution family was specified depending on the trait, using a Gaussian model for continuous outcomes and a binomial model for binary outcomes. An independence correlation structure was applied for unrelated individuals, whereas an exchangeable correlation structure was used for related individuals. Analyses were conducted by modeling the effects of the reference (allele 1) and alternative (allele 2) alleles inherited from each parent.

Several complementary models were applied to disentangle parental effects.

Maternal and paternal effects were assessed by comparing genotype groups while holding the allele inherited from the other parent constant (Figure 6). Specifically,

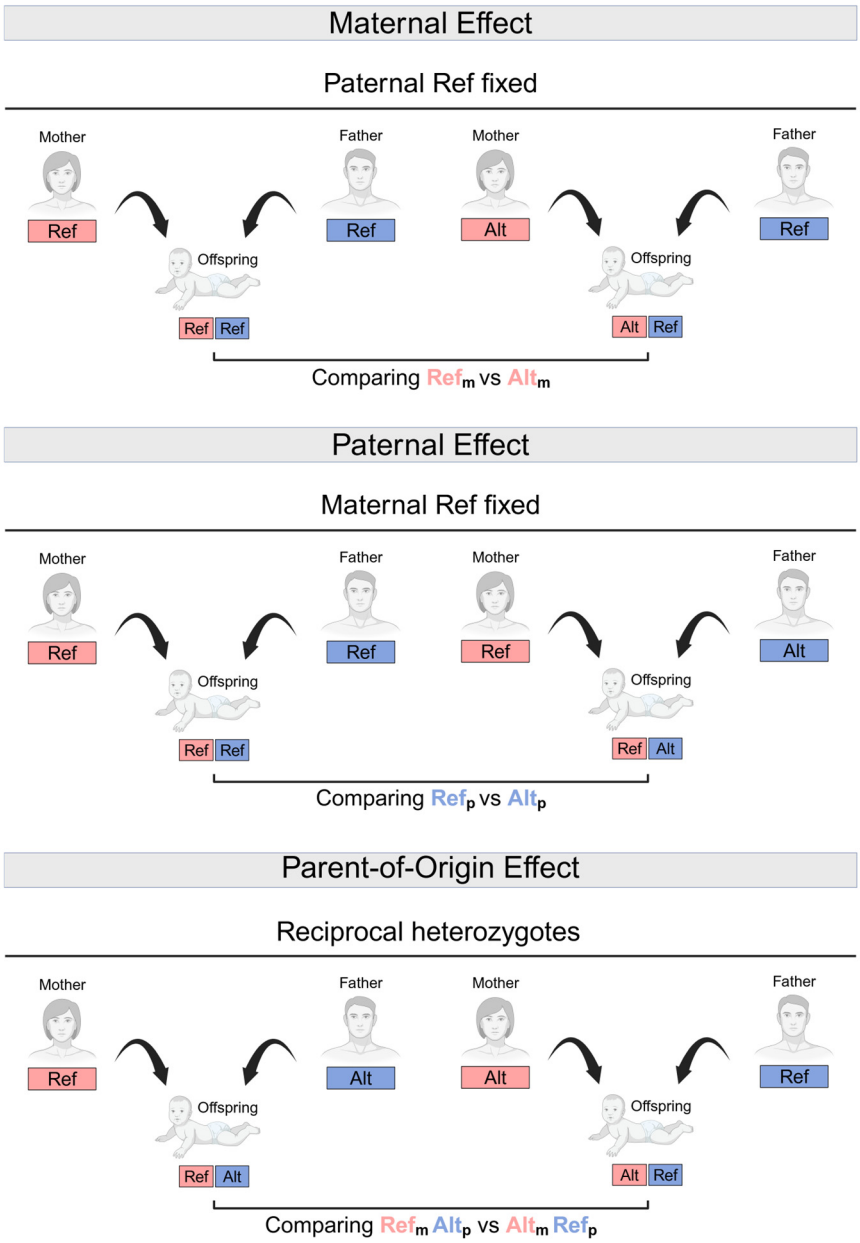

Figure 6. Schematic overview of parent-of-origin effect (POE) analyses in family-based cohorts. Maternal and paternal effects were assessed by comparing offspring genotypes while holding the allele from the other parent constant (Ref_m vs Alt_m with paternal Ref fixed; Ref_p vs Alt_p with maternal Ref fixed). Parent-of-origin effects were evaluated by comparing reciprocal heterozygotes (Ref_mAlt_p vs Alt_mRef_p). Red indicates maternally inherited alleles and blue indicates paternally inherited alleles.

maternal effects were evaluated by comparing offspring carrying different maternal alleles under fixed paternal background, and vice versa for paternal effects.

In addition, POEs were assessed by comparing reciprocal heterozygotes, allowing direct evaluation of differential effects depending on whether the alternative allele was inherited from the mother or the father (Figure 6).

All models were adjusted for relevant covariates: age, age squared, and sex. For specific variants with known associations with additional traits, such as ***GLP1R*** (rs6923761), BMI was included as an additional covariate. Statistical significance was defined at a threshold of nominal $p < 0.05$.

In unrelated individuals

Parent-of-origin effects in unrelated individuals were assessed using Quicktest, a statistical framework designed to detect such effects without requiring family-based data (95). The method exploits differences in phenotypic variance between genotype groups, specifically testing whether heterozygous individuals exhibit increased variance compared to homozygous groups, which can indicate parent-of-origin effects (Figure 7).

Quicktest was applied to replicate candidate loci identified in the primary analyses in independent datasets. This approach complements family-based methods by enabling detection of parent-of-origin signals in population-based cohorts.

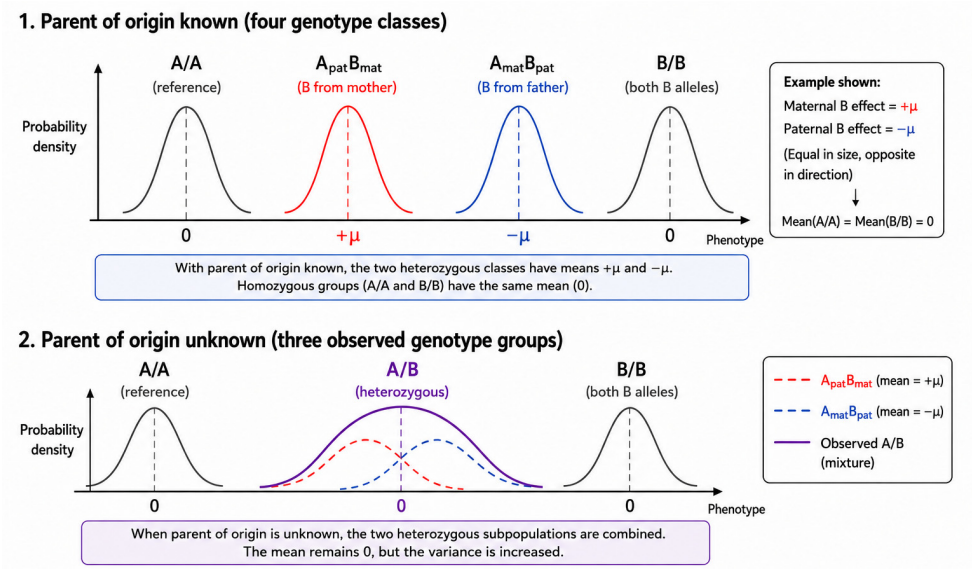

Figure 7. Simplified illustration of how parent-of-origin effects can be detected in unrelated individuals using variance differences.
Parent-of-origin effects are illustrated for a scenario in which maternal and paternal inheritance of the same allele produce effects of equal magnitude but opposite direction. When parental origin is known (top panel), the two heterozygous parent-of-origin classes have different phenotypic means, while the homozygous groups have the same mean. When parental origin is unknown (bottom panel), as is typically the case in population-based cohorts, the two heterozygous classes are combined into a single observed heterozygous group. Although the average phenotype remains unchanged, the mixture of subpopulations results in increased phenotypic variance among heterozygotes relative to homozygotes. Quicktest exploits this variance difference to identify candidate parent-of-origin effects in unrelated individuals. The figure is a simplified conceptual illustration and is not intended to represent exact statistical distributions.

Transcriptomic profiling

RNA sequencing and variant calling

RNA-seq data were processed using an in-house, automated and reproducible Nextflow pipeline with containerized software environments (96). Raw sequencing reads underwent quality control, adapter trimming, and filtering to remove low-quality bases and rRNA contamination. Cleaned paired-end reads were then aligned to the human reference genome (GRCh38) using STAR in two-pass mode.

Post-alignment processing followed GATK best practices (97), including duplicate removal, handling of spliced reads, and base quality score recalibration. Genetic variants were identified using GATK HaplotypeCaller and filtered to retain high-quality biallelic SNPs.

To minimize reference mapping bias in allele-specific analyses, reads overlapping known polymorphisms were processed using the WASP pipeline (98), which remaps reads and retains only those that align consistently. Finally, allele-specific read counts at heterozygous SNPs were quantified using GATK ASEReadCounter.

Gene expression quantification

Gene-level expression was quantified from WASP-processed and deduplicated alignments using featureCounts (99). Library strandedness was assessed to ensure appropriate counting settings. Expression values were normalized to counts per million (CPM) using edgeR package in R following standard protocol (100), and lowly expressed genes were filtered out by requiring CPM ≥ 1 in at least 80% of samples, retaining robustly expressed genes for downstream analyses.

Allele-specific expression analysis

In human pancreatic islets to investigate T2D pathogenesis

To assess allelic expression imbalance, allele-specific expression (ASE) was quantified using a pipeline combining allele-specific read counting with gene-level statistical inference. As mentioned above, allele-specific read counts at heterozygous transcribed SNPs were obtained using GATK ASEReadCounter (97), generating per-variant reference and alternative allele counts, which were subsequently aggregated into gene-level count tables.

ASE was analyzed using the ASEP framework, an R-based method for detecting gene-level allelic expression imbalance from RNA-seq data (59, 101). As the original ASEP package is no longer actively maintained and is incompatible with recent R versions, a modified implementation was used (102). Input count data were reformatted according to ASEP requirements, and gene coordinates (hg38) were retrieved from Ensembl. ASE was assessed both across all samples and within specific conditions, including stratification by disease status and comparisons between experimental treatments.

ASEP applies a mixture model with resampling-based inference to detect consistent allelic imbalance across individuals while accounting for heterogeneity in effect size and direction. An initial genome-wide analysis was performed using 1,000 resampling iterations. Genes showing nominal evidence of imbalance ($p < 0.1$) were subsequently reanalyzed using 100,000 iterations to obtain more stable and precise p-value estimates.

To ensure robust ASE detection, stringent filtering criteria were applied prior to analysis. Specifically:

- Genes were required to meet a minimum expression threshold (CPM ≥ 1 in at least 80% of samples)
- Only heterozygous transcribed SNPs were included
- Genes were required to have heterozygous SNPs present in more than three individuals
- A minimum read depth of >20 reads per SNP was required
- Minor allele count (MAC) >5
- Minor allele frequency (MAF) ≥5%

These criteria, based on established recommendations, reduce technical noise and increase confidence in ASE estimates (59).

In cancer pathogenesis

Allele-specific expression (ASE) was assessed using the in-house made BASE pipeline, which integrates matched whole-genome sequencing (WGS) and RNA-seq data (103). WGS data were used to identify high-confidence heterozygous variants and to estimate allele-specific copy number states, while RNA-seq data were processed to obtain allele-specific read counts at expressed variants following bias correction.

ASE was quantified by comparing reference and alternative allele counts at heterozygous sites, which were aggregated at the gene level. Statistical testing was performed to determine whether observed allelic ratios deviated from expected expression under two models: balanced biallelic expression and expression consistent with underlying copy number variation. This allowed discrimination between ASE driven by regulatory effects and ASE resulting from gene dosage changes.

Stringent filtering criteria were applied to ensure reliable ASE detection, including minimum read depth and expression thresholds. Genes were classified as exhibiting ASE based on deviation from expected allelic ratios, while accounting for copy number states and multiple testing.

Comparison of AEI accounting for unequal gene sets

Differences in the proportion of genes exhibiting allelic expression imbalance between T2D and non-T2D groups were assessed while accounting for unequal gene set sizes, both genome-wide and within imprinted genes. Statistical comparisons were performed using bootstrap resampling and contingency table-based methods, including chi-square and Fisher's exact tests. Effect sizes were summarized as odds ratios with corresponding confidence intervals.

SNP association with islet phenotypes

Associations between the ***GLP1R*** variant rs6923761 and islet phenotypes were evaluated using robust linear regression models to account for heteroscedasticity. Models were adjusted for relevant covariates, including demographic and islet-specific factors, with additional adjustment for BMI where appropriate.

Single cell RNA sequencing

Single cell RNA Sequencing of fetal pancreas

ScRNAseq of 3 fetal pancreata was performed as described previously (23). Briefly, single-cell libraries were generated using Chromium Single Cell 3′ Chemistry v3 (10x Genomics) according to the manufacturer's instructions. The libraries were sequenced on an Illumina NextSeq 500 instrument using 400 million reads flow cells, with the recommended read configuration of 42 bp paired-end sequencing.

Clustering of cell types

Clustering of cell types was made in Lopue browser 6.3.0. using 2 different datasets that had the best quality (gestational age: 8 weeks). Genes that were present in at least 3 cells were included. T-sne plots were then generated by filtering expression on ***RTL1*** using the built-in function "Gene/Feature expression". "Feature max" was selected and expression was presented as Log2.

Correlation with developmental markers

Single-cell RNA-seq data from human fetal pancreas samples were analysed in R 4.4.1 using Seurat 5.2.1. Filtered 10x Genomics feature-barcode matrices were imported, and Seurat objects were generated for each sample. Low-quality cells were removed based on RNA count, detected feature number, and mitochondrial transcript ratio thresholds. The filtered data were then log-normalized, and count matrices were extracted for downstream analysis.

Spearman rank correlation analysis was performed between predefined gene sets, including genes of interest, developmental markers, and imprinted genes. For each query gene, correlations were calculated only in cells with non-zero expression of that gene. Genes expressed in fewer than three cells or with zero variance were excluded. P-values were adjusted using the false discovery rate method. Correlation coefficients were visualized as heatmaps using ggplot2.

Papers in summary

Paper 1
Parent-of-origin effects in the life-course evolution of cardiometabolic traits

Introduction & Aim

POEs, wherein the effect of a genetic variant depends on whether it is inherited from the mother or the father, may contribute to the heritability of metabolic traits and T2D risk. POEs can arise through epigenetic mechanisms such as genomic imprinting, where only one parental allele is expressed while the other is silenced, or through parental expression bias, where one allele is preferentially expressed over the other. These mechanisms can influence gene regulation and, consequently, disease susceptibility.

Most studies of POEs have focused on adult populations of European ancestry; however, it remains unclear to what extent these patterns are established during early development, as well as in other diverse populations. Therefore, this study aimed to identify POEs on metabolic traits relevant to T2D risk in an Indian population, with particular emphasis on how these effects may vary across the life course. In addition, we aimed to investigate whether previously reported POE variants identified in European populations exhibit similar effects in this population.

Results

The study identified significant POEs on key glucose- and insulin-related traits, as well as lipid levels and anthropometric measures across developmental stages, demonstrating that these effects are both trait-specific and dynamic over time. Consistent with findings from European populations, lipid traits predominantly exhibited maternal effects. For fasting insulin and insulin sensitivity (HOMA2S), maternal effects were observed in early childhood (6 years) and were associated with reduced insulin sensitivity, but these effects changed direction during adolescence (12 years), indicating age-dependent modulation of insulin resistance. In contrast, insulin secretion (HOMA2B) exhibited a distinct pattern, with stronger

paternal effects in early life (6 years) that transitioned to maternal influences during later developmental stages and into early adulthood.

Overall, longitudinal analyses using mixed-effects models showed that paternal contributions were more prominent for insulin-related traits, including fasting insulin and HOMA indices, whereas maternal effects were more evident for glycemic and lipid traits, reflecting the cumulative effects across time. This suggests differential parental regulation of components of glucose homeostasis, with potential implications for disease development.

Importantly, genetic analyses revealed POEs at established T2D susceptibility loci, including ***KCNQ1***, where associations were observed for insulin secretion, supporting a role for parent-specific genetic regulation of β-cell function. Notably, while the underlying genetic variants remained constant, the direction and magnitude of their effects changed across developmental stages, suggesting that these relationships may be mediated by epigenetic mechanisms. In addition, POEs were observed for birthweight and lipid traits, further supporting a broader influence of POEs on metabolic phenotypes.

Together, these findings indicate that POEs contribute to variation in insulin secretion and insulin sensitivity, both of which are central to the pathophysiology of T2D. The observed temporal shifts in parental effects further suggest that these mechanisms are developmentally regulated, highlighting the importance of considering both parental origin and life course stage when investigating genetic influences on T2D risk. Collectively, the genetic and phenotypic evidence supports a role for POEs and epigenetic regulation in shaping metabolic traits across the life course.

Paper 2
Allelic expression imbalance in pancreatic islets indicate parent-of-origin effects on type 2 diabetes risk

Introduction & Aim

T2D is a complex metabolic disorder influenced by both genetic and environmental factors, with impaired insulin secretion and action as central features. Although GWAS have identified numerous variants associated with T2D, these explain only part of the heritability and provide limited insight into the regulatory mechanisms linking genetic variation to β-cell dysfunction.

Integrative approaches combining genetic and transcriptomic data can help address this gap. AEI, which reflects unequal expression of the two alleles within an individual, can reveal cis-regulatory effects and provides a functional link between

genetic variation and gene expression. AEI may arise through several mechanisms, including cis-regulatory variants and epigenetic processes such as imprinting or parental expression bias, which can underlie POEs.

Therefore, this study aimed to characterize AEI in human pancreatic islets and assess its relevance to T2D. Specifically, we aimed to distinguish genes potentially involved in disease pathogenesis from those reflecting secondary effects of hyperglycemia, to evaluate whether AEI signals are driven by underlying genetic variation, and to investigate POEs using imprinted genes and genetic analyses in family-based as well as in unrelated population-based cohorts.

Results

AEI in chronic, acute and normoglycemic states

AEI was assessed in human pancreatic islets using bulk RNA sequencing data from 326 organ donors, including 45 individuals with T2D and 281 diabetes-free controls. Of the 19,273 expressed genes in all samples together, 12,854 passed quality control, of which 3,956 (30.8%) exhibited significant AEI (FDR < 0.05).

Stratified analysis of T2D and non-T2D donors separately showed that AEI was more prevalent in T2D islets (36.0%) than in non-T2D islets (27.1%), indicating increased allele-specific regulatory effects in the diabetic state. Classification of AEI patterns identified 1,844 genes with AEI exclusively in T2D, consistent with a gain of AEI, and 1,101 genes with AEI only in non-T2D, indicating a loss of AEI in T2D. A smaller subset showed opposing allelic effects between groups, while 1,800 genes exhibited consistent AEI across all analyses, reflecting stable allele-specific regulation in pancreatic islets.

Identification of potentially causal genes

To distinguish genes potentially involved in T2D pathogenesis from those reflecting secondary effects of hyperglycemia, AEI patterns in chronic T2D were compared with those under acute glucose exposure. Excluding genes with shared or non-specific AEI patterns yielded 2,945 genes with T2D-specific AEI. After removing genes responsive to acute glucose, 2,919 genes remained as potential causal candidates.

Several of these overlapped with previously reported candidate genes, differentially expressed genes in T2D islets, and β-cell-expressed genes. Functional follow-up showed that ***ABCC8*** knockdown reduced glucose-stimulated insulin secretion, whereas ***CORO2B*** had minimal effects. Integration with GWAS and eQTL data further identified overlap with established T2D loci, including ***ADCY5***, ***SIX3***, and ***GLP1R***.

GLP1R links T2D, DEGs, POE and treatment response

GLP1R was prioritized based on its role in insulin secretion, its identification as a T2D GWAS locus, and the presence of AEI specifically in T2D islets. It has also been reported as differentially expressed in T2D and supported by functional knockout data.

The missense variant rs6923761 was associated with insulin content in pancreatic islets, supporting its functional relevance. Given the presence of AEI at this locus, POE was assessed in genetic analyses. Across population-based and family-based cohorts, rs6923761 showed parent-specific effects on β-cell function, with the maternally inherited A allele associated with increased HOMA2-B and reduced T2D risk.

In addition, analyses in individuals with newly diagnosed diabetes suggested an effect on response to GLP-1 receptor agonist therapy, with parental inheritance modifying the association with BMI. Together, these findings link ***GLP1R*** to T2D through combined evidence from AEI, differential expression, and genetic analyses, highlighting a POE mechanism with potential therapeutic relevance.

Imprinted genes

To place these findings in a broader regulatory context, AEI was examined in known and predicted imprinted genes, which exhibit parent-specific expression. Of 275 imprinted genes, 179 were expressed in pancreatic islets, of which 102 passed quality control and 38 showed significant AEI, indicating modest enrichment compared to non-imprinted genes.

Comparison between T2D and non-T2D islets identified imprinted genes with altered AEI patterns, including genes showing gain or loss of AEI in T2D. These changes were not observed under acute hyperglycemic conditions, suggesting relevance to disease mechanisms. Several genes, including ***GLIS3*** and ***TSHZ3***, are located at established T2D loci. However, no enrichment of AEI among imprinted genes was observed when comparing T2D and non-T2D islets. This indicates that imprinting-related AEI is largely conserved between T2D and non-T2D islets, with no evidence of global dysregulation in T2D.

Other imprinted genes, such as ***RHOBTB3*** and ***GRB10***, showed consistent AEI across conditions, indicating stable parent-specific expression. Functional follow-up of ***RHOBTB3*** did not affect insulin secretion, suggesting effects not mediated through gene expression levels. Genetic analyses identified multiple variants in ***TSHZ3*** and ***GLIS3*** with POE on β-cell function and T2D-related traits, with replication across cohorts. Variants in ***RHOBTB3*** also showed POE on β-cell function.

Together, these findings support a role for imprinted genes in mediating AEI and contributing to POE on β-cell function and T2D risk.

Paper 3
Imprinted genes in beta cell development and T2D risk

Introduction & Aim

Both maternal and paternal genomes are required for normal embryonic development, reflecting functional differences between parental alleles that can arise through genomic imprinting and other POEs. Imprinted genes exhibit parent-specific expression, where one parental allele is preferentially or exclusively expressed, and several imprinted loci, including ***IGF2***, ***H19***, ***KCNQ1***, and ***KLF14***, have been implicated in growth regulation, β-cell development, insulin secretion, and diabetes susceptibility. Although more than 150 imprinted genes have been identified in humans and mice, many remain incompletely characterized, particularly in the context of pancreatic development and β-cell biology. Since pancreatic β-cell dysfunction is central to the pathogenesis of diabetes, understanding how imprinted genes and allele-specific regulation contribute to β-cell development and function may provide important insight into disease susceptibility and progression.

Therefore, this study aimed to investigate the role of imprinted genes and AEI during pancreatic development by comparing fetal and adult pancreatic tissues and assessing their contribution to β-cell function and type 2 diabetes risk. We hypothesized that genes expressed specifically in fetal pancreas are more likely to be involved in islet development, whereas genes expressed in both fetal pancreas and adult islets may contribute to both developmental processes and mature islet function. To address this, RNA sequencing and genotype data from fetal pancreas, adult pancreas, human pancreatic islets, and islet cell types were integrated to characterize developmental AEI patterns and evaluate POEs in genetic analyses.

Results

Expression profiling identified widespread activity of imprinted genes during fetal pancreatic development, with 173 known or predicted imprinted genes expressed in fetal pancreas tissue. Several established developmental regulators, including ***IGF2***, ***H19***, ***MEG3***, ***MEST***, and ***GNAS***, were among the most highly expressed genes. Comparison of fetal and adult pancreas identified 75 imprinted genes with fetal-specific and/or fetal-enriched expression, suggesting potential roles in pancreatic

development and β-cell maturation. Several of these genes also remained expressed in adult human pancreatic islets, indicating continued relevance for islet function.

Single-cell analyses of embryonic pancreas further demonstrated that a large proportion of imprinted genes were expressed across pancreatic cell populations during development. ***RTL1*** showed particularly strong and widespread expression, and correlation analyses revealed associations between imprinted genes and key developmental regulators. ***RTL1*** expression positively correlated with developmental markers including ***NANOG*** and ***MAFB***, whereas ***NNAT*** showed negative correlations with endocrine lineage markers such as ***FOXA2***, ***NKX6-1***, ***PDX1***, and ***SOX9***, supporting stage-specific roles for imprinted genes during pancreatic differentiation.

Further analyses identified subsets of imprinted genes potentially involved in either β-cell development alone or both β-cell development and mature β-cell function. Fourteen genes showed fetal pancreas-specific expression without detectable expression in adult islets, suggesting developmental roles. Among these, ***RTL1*** and ***WT1-AS*** exhibited significant AEI in fetal pancreas ($p < 0.05$), and ***RTL1*** variants were associated with HOMA-B in the Botnia cohort. Moreover, the same ***RTL1*** loci showed nominal associations with corrected insulin response (CIR) in the Botnia cohort, while an additional nominal association was identified for one ***WT1-AS*** locus. ***RTL1*** expression also strongly correlated with ***DLK1***, and both genes displayed coordinated expression during stem cell differentiation toward β-cells, supporting a role in endocrine development.

Additional genes, including ***MEG3***, ***PHACTR2*** and ***SFMBT2***, exhibited AEI in fetal pancreas and/or adult islets, suggesting temporally regulated imprinting or allele-specific regulation. Variants in several of these genes showed associations with insulin secretion traits and parent-of-origin effects on HOMA-B across family-based and population-based cohorts. Notably, ***PHACTR2*** variants demonstrated significant parent-of-origin associations with HOMA-B, although functional knockdown of ***PHACTR2*** in EndoC-βH1 cells did not alter glucose-stimulated insulin secretion.

Finally, parent-of-origin analyses of imprinted genes displaying altered AEI between T2D and non-T2D islets identified multiple variants associated with T2D risk and β-cell function across independent cohorts. Replicated associations were observed for variants in genes including ***CTNNA3***, ***NTM***, ***SIM2***, ***RTL1***, ***DLK1***, and ***KLF14***, supporting a broader contribution of imprinted genes and parent-specific regulation to β-cell biology and T2D susceptibility.

Paper 4
Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE

Introduction & Aim

ALL is the most common childhood cancer and is characterized by extensive chromosomal abnormalities that alter gene dosage and transcriptional regulation. High hyperdiploid ALL (HeH ALL), a major subtype of childhood ALL, is defined by recurrent chromosomal gains that are thought to arise early during leukemogenesis and contribute to disease development through altered gene expression. AEI, reflecting unequal expression of the two alleles of a gene, can arise from CNV, cis-regulatory effects, or epigenetic mechanisms, and provides a framework for investigating the functional consequences of genomic alterations in cancer. However, the impact of chromosomal CNVs on allele-specific gene expression in HeH ALL remains incompletely understood. Therefore, this study aimed to characterize the AEI landscape in high hyperdiploid ALL and investigate how CNVs contribute to allele-specific transcriptional regulation in leukemia.

Results

Genome-wide analyses revealed widespread AEI in high hyperdiploid ALL, demonstrating that chromosomal gains substantially influence allele-specific gene expression. Many genes located on gained chromosomes exhibited expression bias toward the duplicated allele, consistent with gene dosage effects contributing to transcriptional dysregulation in leukemic cells. However, AEI patterns could not be fully explained by copy number status alone, indicating that additional cis-regulatory mechanisms also contribute to allele-specific transcription.

Integration of ASE and copy number data identified recurrent AEI across multiple chromosomal regions characteristic of HeH ALL. Several genes with strong allelic imbalance were implicated in pathways related to proliferation, transcriptional regulation, and leukemogenesis, supporting a functional role for AEI in disease biology.

The study further demonstrated that ASE-based analyses can detect regulatory effects that are not apparent from total gene expression analyses alone. In several cases, genes displayed pronounced allelic imbalance despite relatively modest changes in overall expression, highlighting the sensitivity of AEI approaches for identifying functional regulatory perturbations in cancer.

Together, these findings demonstrate that HeH ALL exhibits extensive AEI largely driven by chromosomal copy number alterations, while also revealing additional allele-specific regulatory effects that may contribute to leukemogenesis and disease progression. The study further supports ASE analysis as a valuable framework for understanding the functional consequences of genomic alterations in cancer.

Discussion

Complex traits such as T2D are shaped by interactions between genetic, epigenetic, developmental, and environmental factors. Classical genetic studies, including GWAS, have provided important insight into loci and variants associated with disease risk. However, these approaches often treat inherited alleles as equivalent, regardless of whether they are inherited from the mother or the father. This thesis expands beyond this classical framework by using advanced genetic approaches to investigate how AEI and POEs contribute to metabolic trait variation, β-cell function, and disease susceptibility. Together, the included studies show that genetic effects are not always static or purely additive, but may depend on parental origin, tissue context, developmental stage, and disease state.

A central concept in this thesis is that maternal and paternal genomes can have distinct biological effects. This is most clearly demonstrated by genomic imprinting, where one parental allele is selectively silenced, but POEs may also arise through more subtle parental expression biases or other allele-specific regulatory mechanisms. Since epigenetic marks such as DNA methylation can differ depending on parental inheritance, these mechanisms provide a link between inherited genetic variation, parental contribution, and environmental influences. In this way, POEs represent an additional regulatory layer through which genetic variation may influence complex traits and metabolic disease risk.

Paper 1 demonstrated that POEs contribute to variation in cardiometabolic traits across the life course. Importantly, these effects were not uniform across traits or ages. Lipid traits showed robust maternal effects, consistent with findings from European populations, whereas insulin-related traits displayed more dynamic patterns. For fasting insulin and insulin sensitivity, maternal effects were observed in early childhood but changed direction during adolescence. In contrast, insulin secretion showed stronger paternal effects in early life, which later shifted toward maternal influences. These findings suggest that POEs are not fixed but may be developmentally regulated and modified across different stages of life.

The observation that POEs were detected already in childhood is particularly important, as it suggests that these mechanisms are established early and may contribute to long-term metabolic susceptibility. The replication of previously reported POEs in an Indian population further strengthens the evidence that these mechanisms are not restricted to European ancestry populations. Rather, they may

reflect broader biological principles involved in the regulation of metabolic traits. At the same time, differences in the magnitude and timing of effects across populations and developmental stages highlight the importance of studying diverse cohorts and longitudinal data when investigating genetic contributions to complex disease.

These findings also suggest that POEs can be used as a strategy to identify genes involved in metabolic regulation. If the phenotypic effect of a genetic variant depends on whether it is inherited maternally or paternally, this may point toward regulatory mechanisms such as imprinting, parental expression bias, or allele-specific gene dosage effects. This provided a rationale for the analyses in Paper 2, where allelic expression imbalance was used to identify genes and regulatory mechanisms relevant to T2D in human pancreatic islets.

Paper 2 showed that AEI is widespread in human pancreatic islets and more prevalent in islets from individuals with T2D than in non-diabetic donors. This suggests that allele-specific regulatory control may be altered in the diabetic state. By comparing AEI patterns in chronic T2D with those observed under acute glucose exposure, the study aimed to distinguish genes potentially involved in disease pathogenesis from genes reflecting secondary effects of hyperglycemia. This approach identified a set of candidate genes with T2D-specific AEI, several of which overlapped with known T2D loci, differentially expressed genes, and β-cell-expressed genes.

A key strength of ASE analysis is that it compares the expression of two alleles within the same cellular environment. Since both alleles are exposed to the same trans-acting regulatory factors, differences in expression are more likely to reflect cis-regulatory variation, gene dosage effects, or parent-specific regulatory mechanisms. Thus, AEI provides a functional link between genetic variation and gene expression that is not captured by classical association studies alone. In this thesis, AEI was therefore used not only as a descriptive measure of allele-specific transcription, but also as a tool to prioritize genes with potential relevance to β-cell dysfunction and T2D risk.

One of the most notable findings in Paper 2 was the identification of ***GLP1R*** as a candidate gene linking AEI, differential expression, genetic association, β-cell function, and treatment response. ***GLP1R*** is biologically relevant because of its role in insulin secretion and as the target of GLP-1 receptor agonist therapy. The finding that a ***GLP1R*** variant showed POEs on β-cell function and T2D risk suggests that parent-of-origin mechanisms may influence not only disease susceptibility, but potentially also therapeutic response. Although these translational implications remain exploratory, they raise the possibility that allele-specific and parent-specific regulation could eventually contribute to more individualized treatment strategies in a foreseeable future.

Paper 2 also investigated imprinted genes as a biologically relevant subset of genes with potential parent-specific regulation. Imprinted genes showed evidence of AEI in pancreatic islets, and several displayed altered AEI patterns in T2D. However, there was no evidence for global dysregulation of imprinting-related AEI in T2D islets. This suggests that imprinting-related regulation is largely conserved between diabetic and non-diabetic islets, but that specific imprinted genes may still contribute to disease-relevant mechanisms. Genetic analyses further supported POEs at several loci, including genes implicated in β-cell function and T2D-related traits.

Paper 3 extended these findings into fetal development, addressing whether imprinted genes and allele-specific regulation may contribute to pancreatic development, β-cell mass and subsequently early programming of β-cell function. This is important because imprinting is strongly linked to embryonic and fetal growth, and many imprinted genes have established roles in development. The study showed that a large number of known or predicted imprinted genes are expressed in fetal pancreas, including well-established developmental regulators such as ***IGF2***, ***H19***, ***MEG3***, ***MEST***, and ***GNAS***. Comparison with adult pancreas identified imprinted genes with fetal-specific or fetal-enriched expression, suggesting that these genes may have developmental roles in the pancreas.

The fetal analyses are particularly relevant in the context of diabetes because β-cell mass and β-cell function are shaped during development. If POEs influence β-cell development, they may contribute to differences in insulin secretion capacity later in life. This is consistent with the broader concept that metabolic disease risk can be programmed early in development. Insulin itself is not only a metabolic hormone, but also an important growth factor during fetal life. Therefore, parent-specific regulation of genes involved in β-cell development or insulin-related pathways could have consequences for both fetal growth and later metabolic health.

Single-cell analyses in Paper 3 provided additional support for a developmental role of imprinted genes. Several imprinted genes were expressed in embryonic pancreatic cell populations and correlated with markers of pancreatic development. ***RTL1*** showed particularly strong expression in embryonic pancreas and correlated with developmental markers, while ***NNAT*** showed negative correlations with several endocrine lineage markers. These findings suggest that imprinted genes may be active at specific stages of pancreatic differentiation and may contribute to the regulation of endocrine lineage development.

Further analyses identified genes with fetal pancreas-specific expression and ASE, including ***RTL1*** and ***WT1-AS***. ***RTL1*** variants were associated with HOMA-B in the Botnia cohort, with nominal associations also observed for corrected insulin response. ***RTL1*** expression correlated with ***DLK1***, and both genes showed coordinated expression during differentiation from induced pluripotent stem cells

toward β-cells. Together, these findings support a potential role for ***RTL1*** and related imprinted loci in pancreatic endocrine development and β-cell formation.

Paper 3 also identified imprinted genes that may be relevant to both β-cell development and mature β-cell function. Genes such as ***MEG3***, ***PHACTR2*** and ***SFMBT2*** showed allele-specific expression in fetal pancreas and/or adult islets. Several variants in these genes were associated with insulin secretion traits, and parent-of-origin analyses identified signals for HOMA-B across family-based and population-based cohorts. Although functional knockdown of ***PHACTR2*** did not alter glucose-stimulated insulin secretion in EndoC-βH1 cells, this does not exclude a role in β-cell biology. Instead, it may suggest that ***PHACTR2*** acts through mechanisms other than direct regulation of acute insulin secretion, or that its function is context-dependent and not fully captured in this experimental model. In line with this, genetic data show POEs on ***PHACTR2***.

Together, Papers 1 to 3 support a model in which allele-specific regulation and POEs contribute to metabolic disease risk across the life course. Paper 1 shows that parent-specific genetic effects are dynamic and detectable from childhood. Paper 2 shows that AEI in adult pancreatic islets can reveal disease-relevant regulatory mechanisms and candidate genes for T2D. Paper 3 suggests that these mechanisms may begin during fetal development, influencing pancreatic development and β-cell biology before metabolic disease becomes apparent. This provides a coherent biological framework in which parent-specific regulation may influence β-cell development, β-cell function, insulin secretion, and ultimately T2D susceptibility.

Although the main focus of this thesis is metabolic disease, Paper 4 demonstrated that AEI and ASE analyses are also valuable beyond T2D. In high hyperdiploid ALL, ASE analysis revealed widespread allele-specific transcriptional imbalance associated with chromosomal gains and copy number alterations. This showed that AEI can capture functional consequences of structural genomic variation in cancer, including gene dosage effects that may not be fully apparent from total gene expression analyses alone. Thus, Paper 4 supports the broader methodological relevance of ASE-based approaches for studying disease mechanisms across different biological contexts.

The inclusion of Paper 4 broadens the perspective of the thesis by showing that allelic imbalance is not limited to imprinting or metabolic disease. Instead, it represents a general mechanism through which genetic and genomic variation can affect transcriptional output. In leukemia, chromosomal abnormalities can alter gene dosage and contribute to leukemogenesis, whereas in pancreatic islets and fetal pancreas, allele-specific regulation may arise from cis-regulatory variation or genomic imprinting, thereby contributing to parent-of-origin effects. Despite these different disease contexts, the shared principle is that analysing AEI provides insight into regulatory mechanisms that conventional gene expression analysis may miss.

Overall, this thesis highlights the importance of moving beyond classical additive genetic models when studying complex traits. While traditional genetic studies are essential for identifying disease-associated loci, they often do not explain how these variants act, in which tissues they are relevant, or whether their effects depend on parental origin or developmental timing. By integrating genetic data, transcriptomic data, ASE analyses, family-based cohorts, population-based cohorts, fetal tissue, adult islets, and cancer models, this thesis provides a broader view of how inherited variation can influence gene regulation and disease risk.

A major strength of this thesis is the integration of complementary genetic and transcriptomic approaches across multiple biological contexts. The use of both family-based and population-based cohorts enabled the investigation of POEs in different study designs, while the availability of deep metabolic phenotyping in the family cohorts allowed detailed assessment of diabetes-relevant traits such as insulin secretion and insulin sensitivity. In addition, the inclusion of fetal pancreatic tissue, adult human islets, and single-cell datasets made it possible to study allele-specific and parent-specific regulation across developmental stages and tissue contexts. This multi-layered approach strengthens the biological interpretation of the findings and provides insight into how genetic variation may influence β-cell development, mature islet function, and disease susceptibility.

However, several limitations should also be considered. Most findings are based on association analyses and therefore do not establish causality. Parent-of-origin analyses are particularly challenging because they require large sample sizes, reliable family structure or phasing, and sufficient statistical power to detect effects that may be modest and context-specific. Similarly, ASE analyses can be sensitive to technical biases, including sequencing depth, mapping bias, genotyping accuracy, and differences in sample composition. Distinguishing true biological AEI from random allelic variation is therefore a key challenge. This was partly addressed by applying quality control filters, integrating eQTL and genetic association analyses, and investigating parent-of-origin effects in family trios, where parental transmission can be more directly assessed. Nevertheless, additional replication and functional validation will be required to confirm the causal relevance of the identified genes and variants.

Despite these limitations, the findings emphasize that genetic effects are highly context-dependent. A variant may have different consequences depending on whether it is inherited from the mother or the father, whether it acts during fetal development or adulthood, and whether it is studied in healthy or diseased tissue. This is particularly relevant for T2D, where disease risk reflects long-term interactions between inherited susceptibility, β-cell development, insulin secretion capacity, insulin resistance, and environmental exposures. POEs may therefore represent one mechanism through which early developmental processes and inherited genetic variation interact to shape metabolic health.

In conclusion, the studies included in this thesis demonstrate that AEI and POEs are important contributors to gene regulation, β-cell biology, and metabolic disease susceptibility. The work supports a model in which parent-specific genetic regulation begins early in development, changes across the life course, and contributes to variation in insulin secretion and T2D risk. Furthermore, the application of ASE analysis in acute lymphoblastic leukemia demonstrates that allele-specific approaches can provide broader insight into the functional consequences of genetic and genomic variation. Collectively, these findings show that considering parental origin, developmental timing, and ASE can improve our understanding of complex disease biology beyond what is possible through classical genetic approaches alone.

Future perspective

Future studies should build on these findings by clarifying the causal mechanisms linking AEI and POEs to β-cell development, β-cell function, and type 2 diabetes risk. Although this work identifies several candidate genes and variants, many findings are based on association analyses. Functional validation will therefore be essential. This could include targeted perturbation of candidate genes in β-cell models, stem cell-derived islets, and pancreatic organoids to assess their effects on gene expression, β-cell maturation, and insulin secretion.

However, one challenge is that the parental origin of alleles is often difficult to determine in human β-cell models, stem cell-derived islets, and organoids. Therefore, mouse models could provide an important complementary approach. By breeding mice with specific maternal or paternal inheritance of variants or gene disruptions of interest, it would be possible to directly test whether the phenotypic effect depends on parental origin. Such models could help distinguish true parent-of-origin effects from general gene dosage effects and provide stronger evidence for causality.

Another important future direction will be to investigate these mechanisms across additional developmental stages and tissue contexts. The fetal pancreas analyses suggest that parent-specific regulation may be established early in life, but longitudinal molecular data from human pancreatic development remain limited. Integrating single-cell transcriptomics, epigenomics, and genotype data across fetal, childhood, and adult stages could help define when POEs arise, whether they persist over time, and how they influence β-cell maturation and metabolic disease susceptibility.

Larger and more diverse cohorts will also be required. POE analyses are statistically challenging and require family-based data or accurate phasing, making replication essential. Expanding studies to non-European populations and deeply phenotyped cohorts will improve power and help determine whether parent-specific effects are shared across populations or shaped by ancestry, environment, and lifestyle.

Finally, future work should explore the translational potential of allele-specific regulation. If specific AEI or POE patterns influence β-cell function, disease progression, or treatment response, they may eventually help identify biomarkers of diabetes risk or therapeutic response. While such applications remain exploratory, AEI and POEs provide important frameworks for uncovering regulatory mechanisms that are missed by classical genetic approaches.

Use of Generative AI

Generative AI tools (ChatGPT Edu) were used to assist with language editing, improving clarity, and refining the structure and flow of parts of this thesis. The research, scientific content, analysis, interpretations, and conclusions are entirely my own. All AI-assisted text was critically reviewed, revised, and approved by the author. The cover image was created by the author with assistance from OpenAI ChatGPT. The author selected, edited, and takes full responsibility for the final image.

Acknowledgements

Rashmi, Thank you for all your patience, support, and guidance over the years. I am incredibly grateful for the opportunity to learn so much, from our conversations and meetings to the exciting courses I was fortunate enough to attend.

Most of all, thank you for believing in me, even at times when I didn't believe in myself. Your encouragement and trust have meant more than I can express, and they have played a significant role in my growth, both professionally and personally.

I will always be grateful for everything you have taught me and for the confidence you have helped me build. Thank you, too, for standing by me during some of the most difficult periods of my PhD, including the loss of my mother. Your kindness, understanding, and unwavering support meant more than I can ever express. I truly don't think I would have made it through without you. Thank you!

Isabella, although changes in the planning meant that we did not have many one-on-one meetings, I would still like to thank you for all your support. You played an important role in creating opportunities for me to learn and develop, especially in the field of single-cell RNA sequencing and I truly appreciate that. Thank you.

Marlena, thank you so much for all the invaluable guidance you've given me in biostatistics. I've learned a great deal from our discussions, and I'm also grateful for the excellent statistics books you've lent me over the years. Beyond the technical support, you've been an important person to turn to for advice on everything from my perfectionist tendencies and striving for the best possible results to more general challenges and day-to-day matters at work. Your insight, support, and encouragement have meant a great deal to me. Thank you!

Johan, Matthias J and Mattias B, thank you for all your IT support whenever the systems decided to misbehave, for your help with the servers, and for always coming to the rescue with last-minute technical emergencies, whether it was fixing urgent issues or sorting out permissions and access.

A special thank you to **Johan** for somehow managing to recover a large amount of data after I accidentally deleted it from the server including the backup :D, for helping download RNA-seq datasets, and for countless other things that were well above and beyond your pay grade. You deserve a very long Christmas wish list!

Mikael, thank you for all your help with discussions on genotype data, for teaching me how to use PLINK, and for introducing me to Nextflow. Your support has been invaluable throughout this work.

Francesca, thank you for being such an amazing office neighbor and friend. It was always a little brighter coming to work knowing you'd be there. Thank you for letting me practice my teaching skills, although with such a brilliant student, you made it far too easy! Thank you for everything, from keeping my plants alive to helping me put up my beloved Christmas decorations. Most of all, thank you for your warmth, your kindness, and for always looking out for me, especially during my difficult times. Your friendship has meant more to me than you know.

Emma, thank you for your support, particularly during our development discussions. Your encouragement and thoughtful advice helped me through moments of self-doubt, and I am truly grateful for your guidance.

Ola, thank you for all the times you asked if I wanted to join you and the guys for lunch. It may have seemed like a small gesture, but it meant a great deal to me and made me feel included. Thank you also for all the fun floorball games we played in Lund. I especially appreciated that you took on the role of delivering the traditional Christmas dinner speech in true Leif Groop style. Those moments are memories I will always look back on with a smile.

Kristoffer, you've always brought a great sense of humor to the workplace, and it's been a pleasure talking with you. Thank you so much for all the lunches, coffee breaks, and laughs we've shared along the way. Your good humor and positive energy have made work more enjoyable, and I truly appreciate it. Thank you!

Olof, thank you for all the help you've given me, from answering small questions to helping retrieve data from the server. Your support has been invaluable, and I really appreciate all the time and effort you've put in. Thank you!

Rajveer, thank you so much for all the help you gave me, from turning on my computer when it had shut down so I could connect remotely, to helping with figures and countless other things that made my life a little easier. I especially appreciate all the urgent little problems you solved during the final stages of writing my thesis. Your willingness to help, no matter how busy things were, meant a great deal to me.

Ulrika, you deserve a special mention here for all the work you do behind the scenes. From handling administrative tasks to helping with countless other things I've asked of you, you're always there to make things happen. LUDC wouldn't last long without someone like you. Thank you for everything you do!

Tiinamaija, thank you for the valuable collaboration and for opening up opportunities that resulted in excellent replications of our findings. It has been a pleasure working with you.

Claes, it has been a real privilege getting to know you. You have been a true inspiration, showing that medical limitations do not have to define what a person can achieve. I sometimes wish I had your strength, confidence, and fearless approach to life. Thank you for all the fascinating conversations we've had, both at CRC and during the retreat we attended together. Your knowledge, insight, and thoughtful perspective have had a profound impact on me, and your input has been invaluable throughout the time we've known each other.

Umur, Hanna, Ece and Felipe, you have been my closest friends at CRC, and I am truly grateful to have gotten to know each of you. Thank you for all the enjoyable lunches, fun after-work gatherings, and the many great moments we shared together. You made my time at CRC both more enjoyable and truly memorable.

Poya, I would also like to extend my sincere thanks to you. Without your expertise in restoring my vision, completing this work would have been significantly more difficult, if not impossible. Thank you so much for everything you have done.

Jasmina, thank you for all your help with the countless details and questions regarding the Botnia cohort. Your support and expertise have been greatly appreciated.

Nicholas, although your time as a visiting researcher in our group was relatively short, we quickly became good friends. It makes me happy that we've stayed in touch ever since, and I truly hope our friendship continues for many years to come.

Isabell, thank you for always being there for me, no matter what. I would not be where I am today without your unwavering support, and I will always be grateful for everything you have done for me.

To my family and friends, thank you for your patience, for listening to my endless complaints about analyses and thesis writing, and for always being there for me. I can't wait to spend much more time with you now that this chapter has finally come to an end.

References

1. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. 2009;324(5930):1029-33.
2. Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. 2021;37(11):995-1011.
3. Flannick J, Florez JC. Type 2 diabetes: genetic data sharing to advance complex disease research. Nature Reviews Genetics. 2016;17(9):535-49.
4. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC Med. 2017;15(1):131.
5. Fund WCR. New major cancer prevention report on dietary and lifestyle patterns 2026 [Available from: https://www.wcrf.org/about-us/news-and-blogs/new-major-cancer-prevention-report-on-dietary-and-lifestyle-patterns/.
6. Herrera-Luis E, Benke K, Volk H, Ladd-Acosta C, Wojcik GL. Gene–environment interactions in human health. Nature Reviews Genetics. 2024;25(11):768-84.
7. Hunter DJ. Gene–environment interactions in human diseases. Nature Reviews Genetics. 2005;6(4):287-98.
8. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nature Reviews Genetics. 2009;10(6):392-404.
9. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
10. Gibson G. Rare and common variants: twenty arguments. Nature Reviews Genetics. 2012;13(2):135-45.
11. Prasad R, Groop L. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. Genes. 2015;6(1):87-123.
12. (CDC) Cfdcap. Prevent Diabetes Complications: Centers for disease control and prevention (CDC); 2022 [updated November 3, 2022. Available from: https://www.cdc.gov/diabetes/managing/problems.html.
13. International Diabetes Federation. IDF Diabetes Atlas 11th Edition. 2025.
14. Terveyskylä (Health Village Finland) FUH. How many have diabetes? https://www.terveyskyla.fi/en: Terveyskylä; 2025 [Available from: https://www.terveyskyla.fi/en/diabeteshub/diabetes/diabetes-as-a-disease/what-is-diabetes/how-many-have-diabetes.
15. Xu Y, Wang L, He J, Bi Y, Li M, Wang T, et al. Prevalence and Control of Diabetes in Chinese Adults. JAMA. 2013;310(9):948-59.

16. Zhou B, Lu Y, Hajifathalian K, Bentham J, Di Cesare M, Danaei G, et al. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. The Lancet. 2016;387(10027):1513-30.
17. Ma RC, Chan JC. Type 2 diabetes in East Asians: similarities and differences with populations in Europe and the United States. Ann N Y Acad Sci. 2013;1281(1):64-91.
18. Atlas D. Diabetes data in China: IDF; 2024 [Available from: https://diabetesatlas.org/data-by-location/country/china/?utm_source=chatgpt.com.
19. Atlas D. India Diabetes Statistics & Healt Data: IDF; 2024 [Available from: https://diabetesatlas.org/data-by-location/country/india/?utm_source=chatgpt.com.
20. DeFronzo RA. From the Triumvirate to the Ominous Octet: A New Paradigm for the Treatment of Type 2 Diabetes Mellitus. Diabetes. 2009;58(4):773-95.
21. DeFronzo RA. The Triumvirate: β-Cell, Muscle, Liver: A Collusion Responsible for NIDDM. Diabetes. 1988;37(6):667-87.
22. Kashyap S, Belfort R, Gastaldelli A, Pratipanawatr T, Berria R, Pratipanawatr W, et al. A Sustained Increase in Plasma Free Fatty Acids Impairs Insulin Secretion in Nondiabetic Subjects Genetically Predisposed to Develop Type 2 Diabetes. Diabetes. 2003;52(10):2461-74.
23. Meier JJ, Nauck MA. Incretins and the development of type 2 diabetes. Curr Diab Rep. 2006;6(3):194-201.
24. Collins L CR. Glucagon-Like Peptide-1 Receptor Agonists. StatPearls [Internet]: tatPearls Publishing; 2024.
25. Baron AD, Schaeffer L, Shragg P, Kolterman OG. Role of Hyperglucagonemia in Maintenance of Increased Rates of Hepatic Glucose Output in Type II Diabetics. Diabetes. 1987;36(3):274-83.
26. Abdul-Ghani MA, DeFronzo RA. Inhibition of renal glucose reabsorption: a novel strategy for achieving glucose control in type 2 diabetes mellitus. Endocr Pract. 2008;14(6):782-90.
27. Matsuda M, Liu Y, Mahankali S, Pu Y, Mahankali A, Wang J, et al. Altered hypothalamic function in response to glucose ingestion in obese humans. Diabetes. 1999;48(9):1801-6.
28. Schwartz SS, Epstein S, Corkey BE, Grant SF, Gavin JR, 3rd, Aguilar RB. The Time Is Right for a New Classification System for Diabetes: Rationale and Implications of the β-Cell-Centric Classification Schema. Diabetes Care. 2016;39(2):179-86.
29. Meier JJ, Bonadonna RC. Role of Reduced β-Cell Mass Versus Impaired β-Cell Function in the Pathogenesis of Type 2 Diabetes. Diabetes Care. 2013;36(Supplement_2):S113-S9.
30. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia. 1985;28(7):412-9.
31. Wallace TM, Levy JC, Matthews DR. Use and abuse of HOMA modeling. Diabetes Care. 2004;27(6):1487-95.

32. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2013;36 Suppl 1(Suppl 1):S67-74.
33. Hanson RL, Pratley RE, Bogardus C, Narayan KM, Roumain JM, Imperatore G, et al. Evaluation of simple indices of insulin sensitivity and insulin secretion for use in epidemiologic studies. Am J Epidemiol. 2000;151(2):190-8.
34. Andersson J, Aydın E, Gunnarsson R, Lilljebjörn H, Fioretos T, Johansson B, et al. Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE. Sci Rep. 2024;14(1):23181.
35. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet. 2010;11(8):533-8.
36. Robles-Espinoza CD, Mohammadi P, Bonilla X, Gutierrez-Arcelus M. Allele-specific expression: applications in cancer and technical considerations. Curr Opin Genet Dev. 2021;66:10-9.
37. Bhojwani D, Yang JJ, Pui CH. Biology of childhood acute lymphoblastic leukemia. Pediatr Clin North Am. 2015;62(1):47-60.
38. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet. 2013;381(9881):1943-55.
39. Pui CH, Nichols KE, Yang JJ. Somatic and germline genomics in paediatric acute lymphoblastic leukaemia. Nat Rev Clin Oncol. 2019;16(4):227-40.
40. Groop L, Forsblom C, Lehtovirta M, Tuomi T, Karanko S, Nissén M, et al. Metabolic Consequences of a Family History of NIDDM (The Botnia Study): Evidence for Sex-Specific Parental Effects. Diabetes. 1996;45(11):1585-93.
41. Hemminki K, Li X, Sundquist K, Sundquist J. Familial Risks for Type 2 Diabetes in Sweden. Diabetes Care. 2010;33(2):293-7.
42. Willemsen G, Ward KJ, Bell CG, Christensen K, Bowden J, Dalgård C, et al. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. Twin Res Hum Genet. 2015;18(6):762-71.
43. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia. 1999;42(2):139-45.
44. Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J, et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia. 1992;35(11):1060-7.
45. Bennett P, Burch T, Miller M. DIABETES MELLITUS IN AMERICAN (PIMA) INDIANS. The Lancet. 1971;298(7716):125-8.
46. Knowler WC, Pettitt DJ, Saad MF, Bennett PH. Diabetes mellitus in the Pima Indians: incidence, risk factors and pathogenesis. Diabetes Metab Rev. 1990;6(1):1-27.

47. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics. 2018;50(11):1505-13.
48. Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nature Genetics. 2022;54(5):560-72.
49. Suzuki K, Hatzikotoulas K, Southam L, Taylor HJ, Yin X, Lorenz KM, et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. Nature. 2024;627(8003):347-57.
50. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. Nature Genetics. 2020;52(7):680-91.
51. Plomin R, Haworth CMA, Davis OSP. Common disorders are quantitative traits. Nature Reviews Genetics. 2009;10(12):872-8.
52. Mitchell KJ. What is complex about complex disorders? Genome Biol. 2012;13(1):237.
53. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Chapter 6: How Cells Read the Genome: From DNA to Protein. Molecular Biology of The Cell, 6th edtion2015.
54. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Chapter 7: Control of Gene Expression. Molecular Biology of the Cell, 6th edition2015.
55. Segert JA, Gisselbrecht SS, Bulyk ML. Transcriptional Silencers: Driving Gene Expression with the Brakes On. Trends Genet. 2021;37(6):514-27.
56. Viñuela A, Varshney A, Van De Bunt M, Prasad RB, Asplund O, Bennett A, et al. Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. Nature Communications. 2020;11(1).
57. Alonso L, Piron A, Morán I, Guindo-Martínez M, Bonàs-Guarch S, Atla G, et al. TIGER: The gene expression regulatory variation landscape of human pancreatic islets. Cell Reports. 2021;37(2):109807.
58. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for data processing in allelic expression analysis. Genome Biology. 2015;16(1).
59. Fan J, Hu J, Xue C, Zhang H, Susztak K, Reilly MP, et al. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. PLOS Genetics. 2020;16(5):e1008786.
60. McGrath J, Solter D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell. 1984;37(1):179-83.
61. Surani MA, Barton SC, Norris ML. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature. 1984;308(5959):548-50.

62. DeChiara TM, Robertson EJ, Efstratiadis A. Parental imprinting of the mouse insulin-like growth factor II gene. Cell. 1991;64(4):849-59.
63. Barlow DP, Stöger R, Herrmann BG, Saito K, Schweifer N. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature. 1991;349(6304):84-7.
64. Tucci V, Isles AR, Kelsey G, Ferguson-Smith AC. Genomic Imprinting and Physiological Processes in Mammals. Cell. 2019;176(5):952-65.
65. Baran Y, Subramaniam M, Biton A, Tukiainen T, Tsang EK, Rivas MA, et al. The landscape of genomic imprinting across diverse adult human tissues. Genome Res. 2015;25(7):927-36.
66. Lawson HA, Cheverud JM, Wolf JB. Genomic imprinting and parent-of-origin effects on complex traits. Nat Rev Genet. 2013;14(9):609-17.
67. Cox NJ. Maternal Component in NIDDM Transmission: How large an effect? Diabetes. 1994;43(1):166-8.
68. Alcolado JC, Laji K, Gill-Randall R. Maternal transmission of diabetes. Diabetic Medicine. 2002;19(2):89-98.
69. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet. 2011;12(8):565-75.
70. Mott R, Yuan W, Kaisaki P, Gan X, Cleak J, Edwards A, et al. The architecture of parent-of-origin effects in mice. Cell. 2014;156(1-2):332-42.
71. Hofmeister RJ, Cavinato T, Karimi R, van der Graaf A, Pajuste F-D, Kronberg J, et al. Parent-of-origin effects on complex traits in up to 236,781 individuals. Nature. 2025;646(8085):647-56.
72. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, et al. Parental origin of sequence variants associated with complex diseases. Nature. 2009;462(7275):868-74.
73. Macias-Velasco JF, St Pierre CL, Wayhart JP, Yin L, Spears L, Miranda MA, et al. Parent-of-origin effects propagate through networks to shape metabolic traits. eLife. 2022;11:e72989.
74. Edwards CA, Watkinson WMD, Telerman SB, Hulsmann LC, Hamilton RS, Ferguson-Smith AC. Reassessment of weak parent-of-origin expression bias shows it rarely exists outside of known imprinted regions. Elife. 2023;12.
75. Prasad RB, Lessmark A, Almgren P, Kovacs G, Hansson O, Oskolkov N, et al. Excess maternal transmission of variants in the THADA gene to offspring with type 2 diabetes. Diabetologia. 2016;59(8):1702-13.
76. Hofmeister RJ, Rubinacci S, Ribeiro DM, Buil A, Kutalik Z, Delaneau O. Parent-of-Origin inference for biobanks. Nature Communications. 2022;13(1).
77. Barker DJ. The developmental origins of adult disease. J Am Coll Nutr. 2004;23(6 Suppl):588s-95s.
78. Dabelea D, Hanson RL, Lindsay RS, Pettitt DJ, Imperatore G, Gabir MM, et al. Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: a study of discordant sibships. Diabetes. 2000;49(12):2208-11.

79. Soubry A. POHaD: why we should study future fathers. Environ Epigenet. 2018;4(2):dvy007.
80. Donkin I, Barrès R. Sperm epigenetics and influence of environmental factors. Mol Metab. 2018;14:1-11.
81. Fullston T, McPherson NO, Owens JA, Kang WX, Sandeman LY, Lane M. Paternal obesity induces metabolic and sperm disturbances in male offspring that are exacerbated by their exposure to an "obesogenic" diet. Physiol Rep. 2015;3(3).
82. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet. 2008;40(9):1092-7.
83. Hanson RL, Guo T, Muller YL, Fleming J, Knowler WC, Kobes S, et al. Strong Parent-of-Origin Effects in the Association of KCNQ1 Variants With Type 2 Diabetes in American Indians. Diabetes. 2013;62(8):2984-91.
84. Travers ME, Mackay DJ, Dekker Nitert M, Morris AP, Lindgren CM, Berry A, et al. Insights into the molecular mechanism for type 2 diabetes susceptibility at the KCNQ1 locus from temporal changes in imprinting status in human islets. Diabetes. 2013;62(3):987-92.
85. Asahara S, Etoh H, Inoue H, Teruyama K, Shibutani Y, Ihara Y, et al. Paternal allelic mutation at the Kcnq1 locus reduces pancreatic β-cell mass by epigenetic modification of Cdkn1c. Proc Natl Acad Sci U S A. 2015;112(27):8332-7.
86. Prokopenko I, Poon W, Mägi R, Prasad BR, Salehi SA, Almgren P, et al. A central role for GRB10 in regulation of islet function in man. PLoS Genet. 2014;10(4):e1004235.
87. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proceedings of the National Academy of Sciences. 2014;111(38):13924-9.
88. Haig D. Genomic imprinting and kinship: how good is the evidence? Annu Rev Genet. 2004;38:553-85.
89. Moore T, Haig D. Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet. 1991;7(2):45-9.
90. Day T, Bonduriansky R. A Unified Approach to the Evolutionary Consequences of Genetic and Nongenetic Inheritance. The American Naturalist. 2011;178(2):E18-E36.
91. Day T, Bonduriansky R. Intralocus sexual conflict can drive the evolution of genomic imprinting. Genetics. 2004;167(4):1537-46.
92. Wolf JB, Hager R. A maternal-offspring coadaptation theory for the evolution of genomic imprinting. PLoS Biol. 2006;4(12):e380.
93. Patten MM, Ross L, Curley JP, Queller DC, Bonduriansky R, Wolf JB. The evolution of genomic imprinting: theories, predictions and empirical tests. Heredity (Edinb). 2014;113(2):119-28.
94. Innovation OU. Oxford - HOMA Calculator 2025 [Available from: https://www.rdm.ox.ac.uk/about/our-facilities-and-units/DTU/software/homa.

95. Hoggart CJ, Venturini G, Mangino M, Gomez F, Ascari G, Zhao JH, et al. Novel Approach Identifies SNPs in SLC2A10 and KCNK9 with Evidence for Parent-of-Origin Effect on Body Mass Index. PLoS Genetics. 2014;10(7):e1004508.
96. Biomedswe. RNA_variant_calling. 1.0 ed: Github; 2025.
97. Derek Caetano-Anolles. RNAseq short variant discovery (SNPs + Indels) www.broadinstitute.org: Broadinstitute; 2022 [Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-.
98. Asiimwe R, Alexander D. STAR+WASP reduces reference bias in the allele-specific mapping of RNA-seq reads. 2024.
99. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30(7):923-30.
100. Bioconductor. edgeR http://www.bioconductor.org/: Bioconductor; 2023 [Available from: http://www.bioconductor.org/packages/release/bioc/html/edgeR.html.
101. Fan J. ASEP - GitHub www.github.com2021 [Available from: https://github.com/Jiaxin-Fan/ASEP.
102. Andersson J. fix pvalue=NA bug in ASEP www.github.com2023 [binomial(link=logit) leads to the error "object logit not found". This is caused by a syntax error, the right syntax is binomial(link="logit") This has been fixed in this commit]. Available from: https://github.com/biomedswe/ASEP/tree/biomedswe-fix-Only-%22NAs%22-for-p-values.
103. Yang-mj B. BASE. Github; 2024.

Paper 1

Diabetologia (2025) 68:1298–1314
https://doi.org/10.1007/s00125-025-06396-5

ARTICLE

Parent-of-origin effects in the life-course evolution of cardiometabolic traits

Rucha Wagh[1,2] · Gad Hatem[3] · Jonas Andersson[3] · Pooja Kunte[1] · Souvik Bandyopadhyay[4] · Chittaranjan S. Yajnik[1] · Rashmi B. Prasad[3,5]

Received: 13 December 2024 / Accepted: 22 January 2025 / Published online: 2 April 2025

Abstract

Aims/hypothesis Cardiometabolic traits are heritable, and some display parent-of-origin effects, which indicates preferential inheritance from one parent or parental bias. Most studies of these phenomena have focused on adult populations. We aimed to investigate the heritability and parent-of-origin effects on cardiometabolic traits in a birth cohort with serial measurements to determine whether these patterns emerged early in life.

Methods The Pune Maternal Nutrition Study comprises a birth cohort in which offspring and parents were studied from birth and followed up for 24 years. We investigated parent-of-origin effects on cardiometabolic traits cross-sectionally at available timepoints using linear regression, and longitudinally across the life course using mixed-effect regression. Maternal and paternal effects on offspring phenotype were modelled after adjusting for age, sex and BMI. Parent-of-origin effects were calculated based on the difference between maternal and paternal effects. We also investigated these effects in another birth cohort, that of the Pune Children's Study. Genetic parent-of-origin effects were assessed using generalised estimating equations after taking the parental origin of the alleles into account.

Results Birthweight showed a maternal parent-of-origin effect. At 24 years, maternal bias was seen for some obesity-related traits for daughters, while paternal bias was seen for WHR in sons. A shift from paternal bias at 6 years to maternal bias at 24 years for the skinfold thickness was observed in daughters. Fasting glucose and lipids showed maternal bias at 6, 12 and 24 years. For fasting insulin and HOMA2-S, a negative maternal effect at 6 years transitioned to a positive one at 12 years. For HOMA2-B, a paternal effect at 6 years transitioned to a maternal one at 12 years, and this remained so at 24 years. Some of these findings were also observed in the cohort from the Pune Children's Study. Longitudinal modelling revealed stronger paternal effects over time for fasting insulin and HOMA indices but maternal effects for glucose and lipids, reflecting their cumulative effect over time. Genetic variants at the *KCNQ1* locus showed a maternal parent-of-origin effect on birthweight, on HOMA2-B at 12 years, and on lipids at 6 and 12 years.

Conclusions/interpretation Our study provides proof of concept of the existence of parent-of-origin effects on cardiometabolic traits from birth, through childhood and puberty, until adult age. Our results indicate a predominantly maternal influence on intrauterine, pubertal and reproductive-age metabolism in the offspring. While the longitudinal analysis indicated a maternal bias for the macronutrients (glucose and lipids), and a paternal bias for glucose–insulin metabolism, the cross-sectional analysis revealed a transition between parental influence across physiological stages. This dynamic relationship may have its origins in the life-history theory of evolution, and could inform strategies for primordial prevention aimed at curbing the rising burden of cardiometabolic disease. Further studies are needed to determine the mechanisms underlying such effects.

Keywords Genetics · Life-course programming · Macronutrients · Parent-of-origin · Sex-specific parental effects · Type 2 diabetes

Abbreviations

GWAS	Genome-wide association study
PCS	Pune Children's Study
PMNS	Pune Maternal Nutrition Study

Chittaranjan S. Yajnik and Rashmi B. Prasad contributed equally to this study.

Extended author information available on the last page of the article

Research in context

What is already known about this subject?

- Cardiometabolic traits such as lipids, glucose and insulin levels have previously been reported to have parent-of-origin effects as well as sex-specific parental effects
- These observations were based on cross-sectional studies in adult offspring from European family-based cohorts
- Parent-of-origin effects on type 2 diabetes and related traits have been reported for *KNCQ1*, *KLF14* and *MOB2* variants

What is the key question?

- Do cardiometabolic traits show parent-of-origin effects from early life and a dynamic life-course evolution?

What are the new findings?

- Parent-of-origin effects were observed for glycaemic traits in early life, with an early paternal effect at 6 years transitioning to a late maternal effect at 12 years; for insulin and HOMA2-S, the negative maternal effect transitioned to a positive effect at 12 years, whereas for HOMA2-B, the paternal effect at 6 years transitioned to a maternal effect at 12 years
- Maternal bias was seen for lipid traits (both sexes), birthweight (both sexes) and anthropometric traits (daughters only), paternal bias was seen for WHR (sons only), and there was a shift from a paternal effect at 6 years to a maternal effect at 24 years for the skinfold thickness
- The parental phenotype associations were supported by genetic parent-of-origin analyses

How might this impact on clinical practice in the foreseeable future?

- This study provides proof of concept regarding the dynamic life-course evolution of parent-of-origin effects for cardiometabolic traits, and will guide future research in epigenetic mechanisms, potentially providing targets for primordial prevention

Introduction

Human traits and diseases are the consequence of a complex interplay between genetics and environment. Heritability measures how much of the variation in a trait within a population is due to genetic factors. Anthropometric and metabolic traits have been shown to be heritable to varying degrees [1, 2]. Genetic association studies have identified a number of variants that are associated with these traits; however, the proportion of heritability attributed to these variants is rather limited [3, 4].

In a classic Mendelian pattern of transmission, a trait may be inherited from both the parents equally; it is also possible that it may be inherited preferentially from one of the parents, while the contribution of the other parent may be low, neutral or even opposite. Such effects (whereby the expression of the phenotype in the offspring depends upon which parent they are inherited from) are termed parent-of-origin effects. These effects may be attributed to genetic imprinting, intrauterine effects or maternally inherited mitochondrial genes [5]. The significance of such effects in the aetiology of type 2 diabetes and obesity has been emphasised previously [2, 6–8]. Type 2 diabetes shows a preferential maternal transmission [2, 7], and a substantial component may originate in the intrauterine period. Several studies have demonstrated that early-life exposures can influence developmental programming and increase the risk of cardiometabolic disorders in later life [9–11].

Parent-of-origin as well as offspring sex-specific parental effects have been reported for anthropometric measurements, insulin secretion and cholesterol levels [1, 2, 12], and these results are supported by genetic associations [8, 13, 14]. For instance, the sons of mothers with type 2 diabetes had lower insulin concentrations compared with the sons of fathers with the same condition, while the daughters of mothers with type 2 diabetes had lower HDL-cholesterol levels compared with the daughters of fathers with type 2 diabetes [2].

The previous studies showing parent-specific effects on cardiometabolic outcomes were performed in adult offspring of European origin and comprised a cross-sectional design. In the present study, we aimed to investigate the role of early-life programming and life-course tracking on cardiometabolic traits from early life to young adulthood, to examine whether these parental effects manifested at an early age. Our objective was to gain insights into

intergenerational transmission of metabolic traits and chronic disease risk to identify possible earlier points of intervention, with the overarching aim of primordial prevention.

Primordial prevention may target the preconception period, pregnancy or early childhood to reduce risk factors for non-communicable diseases. The preferred period is preconception, and classic success stories include the use of folic acid (with or without other micronutrients) in prevention of neural tube defects, and strict control of maternal hyperglycaemia to prevent congenital anomalies in pregnancies affected by diabetes [15, 16]. Such interventions are effective largely due to epigenetic modifications during intergenerational transmission [15, 16]. Recently, there has been growing acknowledgement of the role of epigenetic mechanisms occurring in sperm cells [17–19]. Epigenetic programming and reprogramming continue in postnatal life, although the time windows become progressively narrower. Modifying the family environment to reduce obesity and related disorders is an example of postnatal intervention. Overall, preconception measures primarily target mothers (and increasingly fathers), while postnatal efforts focus on families.

The Pune Maternal Nutrition Study (PMNS), which studies a well-characterised prospective birth cohort from India, provides a unique opportunity to investigate such effects in parent–offspring trios in a life-course model. In this study, we investigated the heritability and parent-of-origin effects for anthropometric, glycaemic, insulin-related and lipid traits in the PMNS birth cohort, with follow-up from birth through puberty till adulthood. Parent–offspring associations and transitions of parent-specific effects on offspring measures across childhood were assessed cross-sectionally (at each available timepoint) as well as longitudinally (across time). We also assessed these effects in another birth cohort, that of the Pune Children's Study (PCS), with similar follow-up. Genetic variants that were previously shown to have parent-of-origin effects on cardiometabolic traits were assessed for similar effects at each available timepoint in the PMNS to better understand genetic contributions vs non-genetic contributions (e.g. the intrauterine environment).

Methods

Cohort characteristics for the PMNS

The PMNS (Fig. 1 and electronic supplementary material [ESM] Fig. 1) was a community-based preconceptional rural birth cohort study established in 1993 in six villages near Pune, India, to prospectively study associations of maternal nutritional status with fetal growth and later diabetes risk in the offspring. Married, non-pregnant women aged 15–40 years (N=2466) were invited to participate, and those who consented and became pregnant between 1994 and 1996 (singleton pregnancy <21 weeks of gestation) were recruited into the study. They were followed up, together with their spouses (F0 generation), and subsequently their offspring (F1 generation). The majority of the families belonged to the Hindu religion. The prevalence of gestational diabetes was <1% based on WHO 1985 criteria (the criteria that were applicable at the time of the visit [20]). Growth and metabolism-related measurements were performed on the parents and the offspring at birth and every 6 years thereafter (ESM Table 1). Final analysis for heritability and parent-specific effects in this study was performed on the participants with complete data at each timepoint. The cohort was representative of the population from which it was derived.

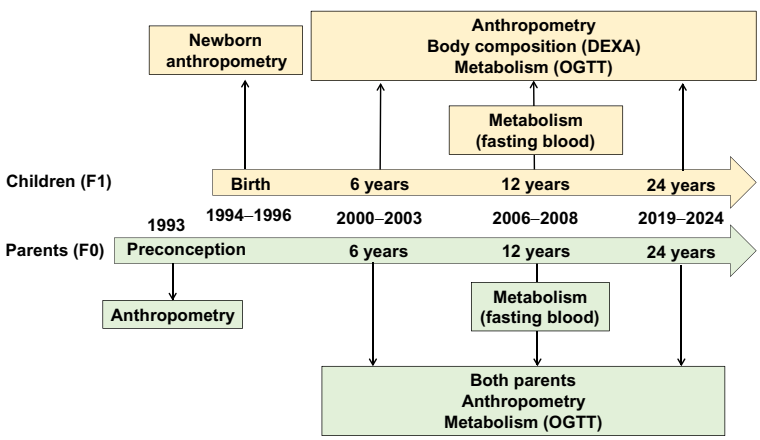

Fig. 1 Study design for the PMNS. DEXA, dual-energy x-ray absorptiometry

Anthropometric and clinical measurements in parents and offspring

Body size (anthropometric traits), glucose tolerance (as assessed using a 75 g OGTT) and corresponding insulin concentrations were measured in both parents at the time of the initial visit and at the 6- and 24-year follow-up visits. Only body size measurements and a fasting blood test were available at the 12-year follow-up visit.

Newborn anthropometric measurements comprising weight, length, abdominal circumference and skinfold thickness were obtained within 72 h birth. Comprehensive assessments of body composition and glucose and insulin concentrations in the offspring were performed at 6, 12 and 24 years. Total fat and lean mass and body fat percentage were measured using dual-energy x-ray absorptiometry (Lunar DPX-IQ 240 pencil beam machine, Lunar Corporation). BMI was calculated as weight (kg)/square of height (m^2), WHR was calculated as waist circumference (cm)/hip circumference (cm), and the body roundness index was calculated as described previously [21].

For the biochemistry measurements, participants (both parents and offspring) arrived at the Diabetes Unit of the King Edward Memorial Hospital, Pune, India, the evening before, had a standardised dinner, and fasted overnight. In the morning, a fasting blood sample was collected. At 6 years, an OGTT was performed, using 1.75 g/kg anhydrous glucose. At 12 years, only a fasting sample was collected. At 24 years, a full OGTT (using 75 g anhydrous glucose) was repeated.

Glucose was measured using the glucose oxidase/peroxidase method, and specific insulin by ELISA (ESM Table 1). HOMA2-S and HOMA2-B were calculated using data from the fasting samples and the calculation provided on the iHOMA2 website (https://www.phc.ox.ac.uk/research/technology-outputs/ihoma2) [22]. Lipid measurements (triglycerides, total cholesterol and HDL-cholesterol) were made using standardised enzymatic assays, and LDL-cholesterol was calculated using Friedewald's formula.

PCS birth cohort

The PCS was a hospital-based urban birth cohort of approximately 400 children born in the King Edward Memorial Hospital, Pune, India, between 1987 and 1989, and their parents. All children born in the above time frame based on hospital registers were invited to participate. The majority of the families belonged to the Hindu religion. Serial measurements for anthropometric traits were obtained at 4, 8 and 21 years, and for cardiometabolic traits at 8 and 21 years. The methods of measurements were the same as described above for the PMNS. The cohort was representative of the population from which it was derived.

Statistical analysis

Heritability and parent-of-origin effects for cardiometabolic traits were assessed between the F0 and F1 generations across the various timepoints in the PMNS cohort (6, 12 and 24 years), and also for anthropometric traits at birth. Skewed variables were transformed using logarithmic or inverse normal transformations. Heritability was estimated using regression models, with offspring phenotype as the outcome and mid-parental phenotype (mean) as the predictor, adjusted for age and sex of both parents and offspring, and presented as β coefficients (with SE) and corresponding p values. Maternal- and paternal-specific effects were investigated cross-sectionally (at each timepoint) using regression models to assess evolution of the effect over time, and longitudinally using mixed-effect regression models to assess the overall effect on the offspring, using intra-family and inter-observation correlations as random effects. All models were adjusted for age and sex for anthropometric traits and for age, sex and BMI of both parents and offspring for cardiometabolic traits. All reported associations from the regression models were significant after multiple testing using the Bonferroni method (adjusted p value threshold ≤0.001). Parent-of-origin effects were tested by calculating the difference in maternal and paternal regression coefficients using the Wald test: $Z = (b1 - b2)/\sqrt{[\mathrm{SE}b1^2 + \mathrm{SE}b2^2 - cov(b1 \times b2)]}$, where $b1$ and $b2$ are regression coefficients, SE are the corresponding standard errors and cov refers to corresponding covariances, and are expressed as Z values with the corresponding p value using the cumulative probability function $p=2\times[1-pnorm(abs(|Z|))]$. Valuesof $p<0.05$ were considered statistically significant.

Genetic parent-of-origin analysis

GWAS quality control and imputation Genome-wide genotyping data were generated for the PMNS trios (mother, father and offspring) using Affymetrix SNP 6.0 chips. SNPs were excluded if the missingness was >5%, the minor allele frequency was <1% and the Hardy–Weinberg equilibrium p value was <0.05. Further quality control included allele checks, call rate and Mendelian errors. Imputation was performed on the TOPMED imputation server using the TOPMED r3 (GRCh38/hg38) as a reference panel with Eagle version 2.4 phasing and Minimac4 imputation [23–25]. SNPs of interest were extracted based on having imputation scores >0.4.

Parent-of-origin analysis The rs2237892 SNP at the *KCNQ1* locus, the rs4731702 SNP at the *KLF14* locus and the rs2334499 SNP at the *MOB2* locus have previously been shown to display parent-of-origin effects on type 2 diabetes risk [8, 14, 26]. The *KCNQ1* variants have also been shown to be associated with insulin secretion, fasting glucose, birthweight and placental weight, whereas variants at *KLF14*

show associations with lipid levels and variants at *MOB2* show associations with insulin sensitivity [26–30]. These SNPs were assessed for their association with the respective traits (where available) in a parent-of-origin manner for each allele at available timepoints. First, the parental origin was determined using a custom R script taking into account the parental genotypes. Then the generalised estimating equations R package geepack [31] was used, with independence as the correlation structure for major allele A and minor allele B, where m is the maternally inherited allele and p is the paternally inherited allele, for the following models: maternal, ApBm vs AA; paternal, AmBp vs AA; parent-of-origin effects, ApBm vs AmBp. Offspring sex and age were used as covariates where applicable, and all traits of interests were first log-transformed and then converted to z scores. A p value <0.05 was considered to be statistically significant.

Ethics statement

The study has been approved by the village leaders and the institutional committee (King Edward Memorial Hospital Research Centre Ethics Committee) at all timepoints for the PMNS, and by the institutional committee (King Edward Memorial Hospital Research Centre Ethics Committee) at all timepoints for the PCS. Participants ≥ 18 years of age signed an informed consent form. Children <18 years of age provided assent, together with parental consent.

Results

Parent-of-origin and offspring sex-specific parental effects on anthropometric and cardiometabolic traits

The PMNS birth cohort comprises approximately 700 parent–offspring trios with serial measurements for anthropometric traits at birth and at 6, 12 and 24 years old, and for cardiometabolic traits at 6, 12 and 24 years old (Fig. 1). The parents were on average short and underweight (based on WHO criteria for stunting and obesity [32–34]) before and during pregnancy. The offspring achieved greater height and weight than their parents, and had substantially higher levels of circulating macronutrients as adults (ESM Tables 2A and 2B). To investigate the proportion of offspring phenotypic traits attributable to parent phenotype variation through the life course, we calculated heritability estimates for cardiometabolic traits at available timepoints (ESM Tables 3 and 4). We next examined whether there was an association between the trait of the offspring and the trait of the mother (maternal effect) or between the trait of the offspring and the trait of the father (paternal effect). If offspring traits showed a significantly stronger association with the mother's traits compared with the father's, this would indicate a maternal bias or maternal parent-of-origin effect, and similarly for paternal bias (paternal parent-of-origin effect).

Anthropometric traits Most of the traits were heritable at all timepoints, with an increasing trend in the parental effect from birth to 24 years (Fig. 2a and ESM Tables 3 and 4). A significant maternal parent-of-origin effect was observed for birthweight when male and female offspring were analysed together (Table 1, Figs 2a and 3a). Daughters showed a significant paternal bias for the sum of skinfolds at 6 years, but a stronger maternal effect at 24 years. At 24 years, daughters also showed a maternal bias for weight, BMI, waist and hip circumference and WHR, but a paternal bias for WHR was observed for sons (Table 1 and ESM Table 5).

Glucose and insulin indices For fasting glucose concentrations, the maternal effects were significantly stronger than the paternal effects at 6, 12 and 24 years (Figs 2b and 3e, f and Table 2), in all offspring combined as well as sons only. These effects were seen only at 12 years old in daughters.

A contrasting shift from 6 years to 12 years was observed for the paternal and maternal effects in relation to insulin and its indices. For fasting insulin and HOMA2-S, there was a significant negative maternal association at 6 years, which shifted to a significant positive one at 12 years, reflecting a change in the direction of the parent-of-origin effect. For HOMA2-B, there was a stronger positive paternal effect at 6 years, which changed to a stronger positive maternal effect at 12 years and continued to show a maternal effect at 24 years (Figs 2b and 3g, h and Table 2).

In sons, fasting insulin and HOMA2-S showed a significant negative maternal association at 6 years, which shifted to a positive one at 12 years, and the parent-of-origin effects were significant only at 6 years. HOMA2-B showed a paternal bias at 6 years, which shifted to a maternal effect at 12 years. In daughters, a positive maternal bias was seen for fasting insulin and HOMA2-S at 12 years and for HOMA2-B at 24 years (Table 2).

Lipid levels A strong and consistent maternal effect was seen for triglycerides, total cholesterol, HDL-cholesterol and LDL-cholesterol at all timepoints (Figs 2b and i–l). A similar association was seen when analyses were performed separately for sons and daughters (Table 2).

Longitudinal modelling by mixed-effect models

Given the availability of the measurements across multiple timepoints, we assessed the overall parent-specific effects using longitudinal mixed-effect models. A significant association of anthropometric traits with each of the parents was seen; however, no parent-of-origin effects

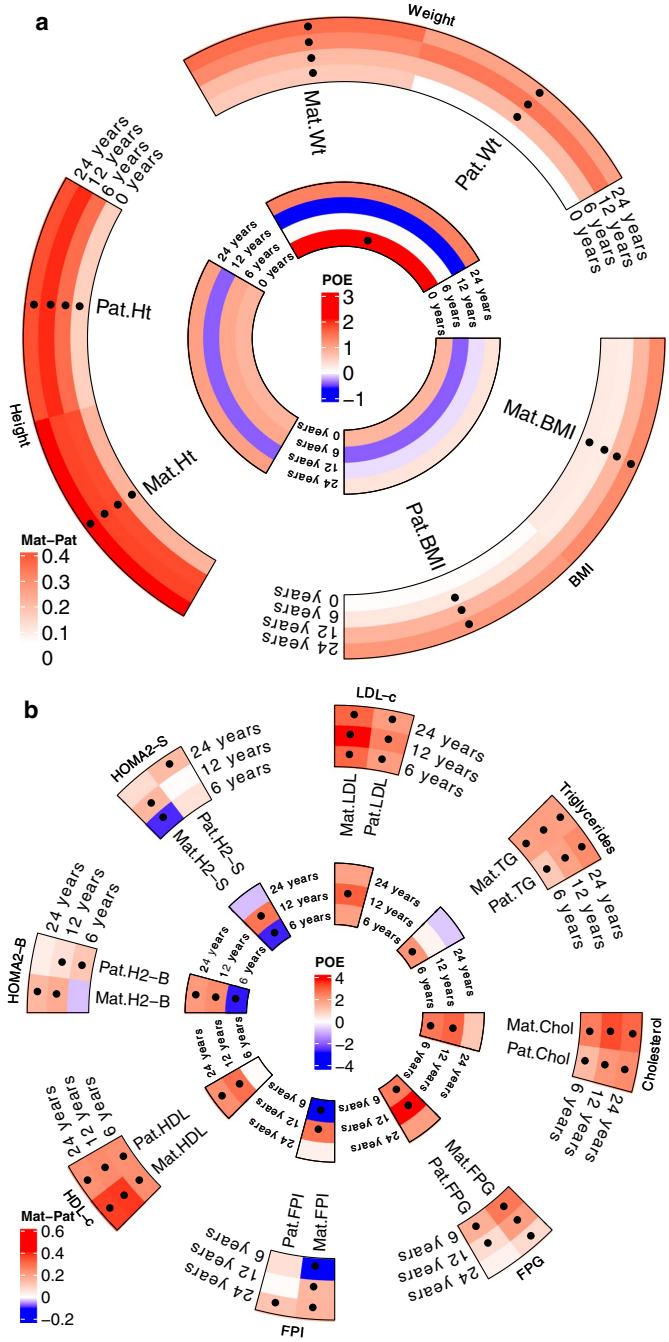

Fig. 2 Circular heatmaps for (**a**) anthropometric traits and (**b**) metabolic traits, representing the phenotype associations between maternal traits or paternal traits and the offspring traits at the various timepoints. The outer circle represents the β coefficients for mother–offspring associations and father–offspring associations. The inner circle represents the parent-of-origin effects expressed as *Z* values. The black dots indicate a significant *p* value ($p<0.05$). Wt, weight; Ht, height; Chol, cholesterol; FPG, fasting plasma glucose; FPI, fasting plasma insulin; H2-S, HOMA2-S; H2-B, HOMA2-B; HDL-c and HDL, HDL-cholesterol; LDL-c and LCL, LDL-cholesterol; TG, triglycerides; Mat, maternal; Pat, paternal; POE, parent-of-origin effects

Table 1 Parent-of-origin test for offspring anthropometric traits

	Mother–child associations		Father–child associations		
	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)
At birth					
All offspring	*N*=745		*N*=705		
Weight (kg)	0.176 (0.045)	0.000104191	0.058 (0.034)	0.094857416	2.10 (0.036)
Height (cm)	0.223 (0.053)	3.42×10^{-05}	0.165 (0.048)	0.000645348	0.81 (0.418)
BMI (kg/m^2)	0.106 (0.042)	0.012537531	0.063 (0.033)	0.054005535	0.80 (0.421)
Sons	*N*=388		*N*=368		
Weight (kg)	0.158 (0.058)	0.006646421	0.032 (0.043)	0.463051919	1.75 (0.081)
Height (cm)	0.227 (0.075)	0.00257224	0.155 (0.064)	0.01625536	0.74 (0.460)
BMI (kg/m^2)	0.118 (0.057)	0.037943635	0.039 (0.044)	0.374421164	1.10 (0.272)
Daughters	*N*=357		*N*=337		
Weight (kg)	0.235 (0.072)	0.001322832	0.100 (0.056)	0.07317818	1.48 (0.139)
Height (cm)	0.230 (0.074)	0.00219911	0.203 (0.073)	0.005667585	0.26 (0.793)
BMI (kg/m^2)	0.113 (0.064)	0.080784399	0.094 (0.048)	0.05364251	0.23 (0.818)
At 6 years					
All offspring	*N*=690		*N*=655		
Weight (kg)	0.207 (0.027)	1.81×10^{-13}	0.206 (0.025)	2.36×10^{-15}	0.02 (0.981)
Height (cm)	0.421 (0.042)	2.53×10^{-22}	0.366 (0.041)	2.51×10^{-18}	0.95 (0.345)
BMI (kg/m^2)	0.101 (0.019)	7.06×10^{-08}	0.112 (0.017)	8.81×10^{-11}	−0.44 (0.658)
Sons	*N*=363		*N*=342		
Weight (kg)	0.193 (0.035)	9.11×10^{-08}	0.206 (0.033)	1.18×10^{-09}	−0.28 (0.782)
Height (cm)	0.458 (0.058)	5.89×10^{-14}	0.338 (0.054)	1.61×10^{-09}	1.50 (0.133)
BMI (kg/m^2)	0.093 (0.024)	0.00012994	0.096 (0.022)	2.63×10^{-05}	−0.10 (0.923)
Daughters	*N*=327		*N*=313		
Weight (kg)	0.224 (0.044)	5.54×10^{-07}	0.208 (0.040)	2.88×10^{-07}	0.26 (0.792)
Height (cm)	0.380 (0.060)	8.12×10^{-10}	0.415 (0.062)	1.10×10^{-10}	−0.41 (0.684)
BMI (kg/m^2)	0.113 (0.029)	0.000122813	0.131 (0.026)	7.15×10^{-07}	−0.47 (0.638)
At 12 years					
All offspring	*N*=658		*N*=597		
Weight (kg)	0.312 (0.039)	4.93×10^{-15}	0.345 (0.037)	1.32×10^{-19}	−0.63 (0.531)
Height (cm)	0.430 (0.049)	2.08×10^{-17}	0.460 (0.047)	7.59×10^{-21}	−0.44 (0.660)
BMI (kg/m^2)	0.217 (0.031)	5.02×10^{-12}	0.221 (0.030)	5.98×10^{-13}	−0.09 (0.929)
Sons	*N*=342		*N*=314		
Weight (kg)	0.263 (0.047)	4.57×10^{-08}	0.332 (0.044)	4.10×10^{-13}	−1.08 (0.280)
Height (cm)	0.522 (0.067)	7.76×10^{-14}	0.423 (0.063)	9.97×10^{-11}	1.09 (0.276)
BMI (kg/m^2)	0.152 (0.037)	5.03×10^{-05}	0.208 (0.036)	1.53×10^{-08}	−1.09 (0.278)
Daughters	*N*=316		*N*=283		
Weight (kg)	0.364 (0.063)	2.22×10^{-08}	0.360 (0.061)	1.01×10^{-08}	0.04 (0.966)
Height (cm)	0.326 (0.072)	9.56×10^{-06}	0.486 (0.071)	6.09×10^{-11}	−1.58 (0.115)
BMI (kg/m^2)	0.284 (0.050)	2.79×10^{-08}	0.238 (0.048)	1.49×10^{-06}	0.66 (0.511)
At 24 years					
All offspring	*N*=462		*N*=357		
Weight (kg)	0.365 (0.056)	2.56×10^{-10}	0.259 (0.057)	6.75×10^{-06}	1.34 (0.182)
Height (cm)	0.482 (0.048)	1.02×10^{-20}	0.415 (0.043)	1.10×10^{-19}	1.03 (0.303)
BMI (kg/m^2)	0.309 (0.058)	2.36×10^{-07}	0.283 (0.063)	1.20×10^{-05}	0.30 (0.761)
Sons	*N*=253		*N*=205		
Weight (kg)	0.328 (0.073)	1.23×10^{-05}	0.353 (0.071)	1.51×10^{-06}	−0.24 (0.808)
Height (cm)	0.505 (0.064)	2.05×10^{-13}	0.446 (0.056)	1.39×10^{-13}	0.69 (0.489)
BMI (kg/m^2)	0.253 (0.072)	0.000642941	0.390 (0.075)	6.88×10^{-07}	−0.57 (0.571)

Table 1 (continued)

	Mother–child associations		Father–child associations		
	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)
Daughters	*N*=209		*N*=152		
Weight (kg)	0.401 (0.086)	6.77×10^{-06}	0.094 (0.092)	0.31322184	2.44 (0.015)
Height (cm)	0.447 (0.074)	1.43×10^{-08}	0.376 (0.068)	1.41×10^{-07}	0.71 (0.480)
BMI (kg/m^2)	0.381 (0.094)	9.00×10^{-05}	0.094 (0.110)	0.392542665	1.98 (0.048)

Associations between offspring anthropometric measurements with that of each of the parents are presented as β values and SE. Differences between maternal and paternal effects are presented as *Z* values and *p* values. The *Z* value shows the difference between regression coefficients. All the regression models were adjusted for age and sex

were observed (ESM Table 6). For cardiometabolic traits, stronger maternal bias was observed for fasting glucose, triglycerides and cholesterol, while stronger paternal bias was observed for fasting insulin, HOMA2-B and HOMA2-S, when all offspring were analysed together, reflecting the pooled effect for each parent for all timepoints combined. Similar associations were also seen for sons separately, with addition of stronger maternal effects for HDL-cholesterol. However, for daughters, stronger maternal bias was seen only for cholesterol (ESM Table 7 and ESM Fig. 2).

Parent-of-origin effects in the PCS birth cohort

To assess whether similar patterns of parent-of-origin effects could be observed in another birth cohort, we used the data from the cohort in the PCS, comprising approximately 400 trios with follow-up at 4, 8 and 21 years. We used the models described above to assess heritability and parent-of-origin effects on cardiometabolic traits. Anthropometry data were available at 4, 8 and 21 years, and data for other cardiometabolic traits were available at 8 and 21 years.

For all offspring combined, as well as sons and daughters separately, heritability estimates were concordant with those seen in the PMNS for anthropometric and lipid traits. For anthropometric measures, the increasing trend in the paternal effects from 4 to 21 years in the PCS mirrored that in the PMNS from 6 to 24 years. For glucose–insulin traits, heritability estimates were significant for fasting glucose, HOMA2-B and HOMA2-S at 8 years, and that for fasting plasma glucose was significant at 21 years for all offspring combined. For the analysis of sons and daughters separately, heritability estimates were significant for fasting glucose at 8 and 21 years (ESM Table 8).

Some of the parent-of-origin effects in the PMNS were also seen in the PCS. Similar to the PMNS, no parental bias was seen for anthropometric traits, with the exception of a maternal bias for height at 8 years (ESM Table 9). We observed a maternal bias for fasting glucose and fasting insulin at 8 years, similar to that seen in the PMNS at 12 years for all offspring. A similar trend was seen for HOMA2-B and HOMA2-S at 8 years (ESM Table 10).

A consistent maternal bias was seen at 8 and 21 years for cholesterol, and a stronger maternal effect was seen at both timepoints for triglyceride levels for all offspring combined. A stronger maternal effect was also seen for HDL-cholesterol at 21 years for all offspring combined, similar to that seen in the PMNS at 24 years. Similar effects on cholesterol and HDL-cholesterol were also seen for sons and daughters separately (ESM Table 10).

Genetic parent-of-origin effects in PMNS

***KCNQ1* locus** The rs2237892 SNP at the *KCNQ1* locus was previously shown to display parent-of-origin effects on type 2 diabetes risk and insulin secretion [8, 14, 26]. Here we examined whether the same variant showed similar effects on HOMA2-B at 6, 12 and 24 years, and found that the *KCNQ1* variant showed significant parent-of-origin specific associations at 12 and 24 years, with the previously reported type 2 diabetes risk and insulin secretion-lowering maternal allele C lowering (and the alternate T allele increasing) HOMA2-B at 12 and 24 years. Although the same association was not statistically significant at 6 years, the direction of effect was the opposite, similar to that seen in the phenotype correlations (Table 3).

We also found significant parental effects on birthweight, triglycerides at 6 years and HDL-cholesterol at 12 years (Table 3).

***KLF14* locus** A maternal parent-of-origin effect was seen for the rs4731702 variant at the *KLF14* locus for triglyceride levels at 6 and 24 years. The type 2 diabetes risk-increasing maternal allele C [14] in the *KLF14* variant increased triglyceride levels while the T allele decreased them. The maternal T allele also increased the sum of skinfolds at 6 years (Table 3).

***MOB2* locus** The rs2334499 variant at the *MOB2* locus showed significant maternal effects at 6, 12 and 24 years, with significant parent-of-origin effects at 12 years.

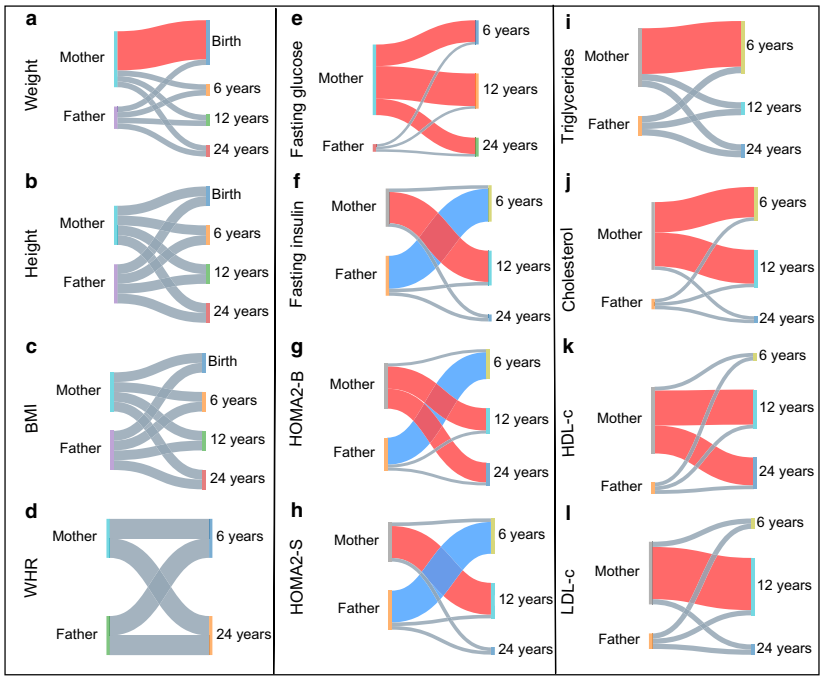

Fig. 3 Individual Sankey diagrams representing the parent-of-origin effects (expressed as *Z* values) separately for each phenotype at each available timepoint: (**a**–**d**) anthropometric traits; (**e**–**h**) glycaemic traits; (**i**–**l**) lipids. A significant maternal bias is indicated in red, and a significant paternal bias is indicated in blue. Insignificant parent-of-origin effects are represented in grey. HDL-c, HDL-cholesterol; LDL-c, LDL-cholesterol

Previous studies showed that the T allele of the *MOB2* variant increased type 2 diabetes risk when inherited from the father but decreased type 2 diabetes risk when inherited from the mother [14]. In our study, the maternal T allele decreased HDL-cholesterol levels at 6, 12 and 24 years (Table 3).

Discussion

By harnessing the potential of a birth cohort, we observed strong parent-of-origin effects for birthweight and metabolic traits but only occasionally for postnatal anthropometric traits in a life-course model. These parent-specific effects on clinically relevant phenotypes changed over time for specific cardiometabolic traits, and may have potentially important implications for life-history theory and clinical practice. Analysis of serial measurements cross-sectionally at available timepoints revealed changing parental influences on glucose and insulin metabolism, including insulin secretion and sensitivity, in the offspring. For insulin secretion, there was a transition from a predominantly paternal association in early childhood to a maternal association at pubertal age, whereas, for insulin sensitivity, a significant negative maternal association transitioned to a significantly positive one. Thus, both insulin secretion and action at pubertal age were predominantly associated with maternal phenotype. The maternal effects on lipid traits remained consistent from childhood to adulthood. Longitudinal modelling showed stronger paternal bias for fasting insulin and its indices but maternal bias for glucose and lipids, indicating their cumulative parental influence over time. Previously reported genetic parent-of-origin effects of variants in *KCNQ1* on insulin secretion [26] were replicated in this study at 12 and 24 years, and the direction of effect mirrored that seen in the phenotype associations. The same variants consistently showed maternal effects on lipid levels and on birthweight (which is partly dependent on gestational age at birth). Parent-of-origin effects were also observed for variants at *KLF14* for triglycerides at 6 years and *MOB2* for lipid levels at 6 and 12 years.

Mendelian genetics stipulates an equal contribution from each of the parents to human traits. However, parent-specific influences on metabolic traits, including the beta cell response

Table 2 Parent-of-origin effects for cardiometabolic traits

	6 years					12 years					24 years				
	MC		FC			MC		FC			MC		FC		
	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)
All offspring	*N*=690		*N*=655			*N*=658		*N*=597			*N*=462		*N*=357		
Fasting glucose	0.287 (0.034)	1.99×10^{-16}	0.174 (0.028)	7.58×10^{-10}	2.59 (0.010)	0.212 (0.030)	4.89×10^{-12}	0.078 (0.018)	1.71×10^{-05}	3.84 (0.000)	0.082 (0.021)	1.19×10^{-04}	0.030 (0.017)	0.066	1.92 (0.055)
Fasting insulin	−0.090 (0.041)	0.0275	0.069 (0.042)	0.1	−2.72 (0.007)	0.158 (0.041)	1.333×10^{-04}	0.010 (0.037)	0.784	2.68 (0.007)	0.165 (0.075)	0.029	0.132 (0.060)	0.034	0.31 (0.758)
HOMA2-B	−0.023 (0.040)	0.57	0.128 (0.040)	0.001	−2.66 (0.008)	0.183 (0.040)	4.98×10^{-06}	0.068 (0.033)	0.040	2.25 (0.025)	0.156 (0.047)	0.0011	0.038 (0.035)	0.277	2.00 (0.046)
HOMA2-S	−0.080 (0.040)	0.048	0.065 (0.041)	0.111	−2.52 (0.012)	0.157 (0.042)	1.885×10^{-04}	0.007 (0.038)	0.844	2.65 (0.008)	0.081 (0.068)	0.233	0.151 (0.059)	0.011	−0.78 (0.436)
Triglycerides	0.208 (0.031)	2.44×10^{-11}	0.120 (0.027)	1.10×10^{-05}	2.16 (0.031)	0.213 (0.034)	4.29×10^{-10}	0.204 (0.028)	7.21×10^{-13}	0.22 (0.827)	0.208 (0.047)	1.06×10^{-05}	0.249 (0.039)	8.97×10^{-10}	−0.66 (0.507)
Cholesterol	0.279 (0.033)	5.23×10^{-16}	0.151 (0.036)	3.58×10^{-05}	2.58 (0.010)	0.364 (0.032)	4.56×10^{-27}	0.230 (0.033)	5.00×10^{-12}	2.93 (0.003)	0.319 (0.045)	1.07×10^{-11}	0.250 (0.040)	1.35×10^{-09}	1.13 (0.257)
HDL-cholesterol	0.244 (0.036)	4.28×10^{-11}	0.242 (0.035)	1.28×10^{-11}	0.05 (0.956)	0.390 (0.039)	8.14×10^{-22}	0.251 (0.033)	6.07×10^{-14}	2.76 (0.006)	0.382 (0.049)	8.90×10^{-14}	0.237 (0.044)	1.58×10^{-07}	2.20 (0.028)
LDL-cholesterol	0.319 (0.037)	2.36×10^{-16}	0.215 (0.050)	1.14×10^{-07}	1.89 (0.059)	0.425 (0.035)	1.67×10^{-29}	0.269 (0.037)	2.96×10^{-12}	3.02 (0.002)	0.324 (0.046)	8.19×10^{-12}	0.214 (0.039)	8.50×10^{-08}	1.84 (0.065)
Sons	*N*=363		*N*=342			*N*=342		*N*=314			*N*=253		*N*=205		
Fasting glucose	0.255 (0.040)	5.55×10^{-10}	0.137 (0.032)	2.250×10^{-05}	2.32 (0.020)	0.148 (0.037)	7.43×10^{-05}	0.034 (0.021)	0.117	2.68 (0.007)	0.112 (0.028)	0.0001	0.028 (0.025)	0.208	2.32 (0.020)
Fasting insulin	−0.152 (0.059)	0.01	0.074 (0.060)	0.219	−2.69 (0.007)	0.182 (0.067)	0.007	0.053 (0.062)	0.397	1.41 (0.158)	0.151 (0.089)	0.092	0.099 (0.069)	0.151	0.46 (0.648)
HOMA2-B	−0.047 (0.055)	0.392	0.139 (0.057)	0.014	−2.36 (0.019)	0.203 (0.063)	0.001	0.057 (0.050)	0.256	1.82 (0.068)	0.118 (0.063)	0.063	0.052 (0.053)	0.331	0.80 (0.425)
HOMA2-S	−0.153 (0.059)	0.009	0.066 (0.058)	0.261	−2.64 (0.008)	0.175 (0.067)	0.01	0.054 (0.064)	0.402	1.30 (0.195)	0.139 (0.086)	0.107	0.101 (0.069)	0.146	0.35 (0.729)
Triglycerides	0.241 (0.042)	1.97×10^{-08}	0.060 (0.038)	0.11	3.22 (0.001)	0.249 (0.042)	6.97×10^{-09}	0.141 (0.037)	0.0001	1.94 (0.052)	0.214 (0.061)	5.597×10^{-04}	0.321 (0.057)	9.08×10^{-08}	−1.27 (0.205)
Cholesterol	0.304 (0.048)	6.13×10^{-10}	0.099 (0.047)	0.034	3.07 (0.002)	0.382 (0.044)	1.91×10^{-16}	0.206 (0.045)	7.28×10^{-06}	2.81 (0.005)	0.311 (0.058)	3.04×10^{-07}	0.236 (0.054)	2.45×10^{-05}	0.94 (0.348)
HDL-cholesterol	0.262 (0.049)	1.970×10^{-07}	0.230 (0.049)	4.350×10^{-06}	0.46 (0.647)	0.410 (0.056)	3.73×10^{-12}	0.236 (0.048)	1.34×10^{-06}	2.35 (0.019)	0.444 (0.066)	2.18×10^{-10}	0.185 (0.060)	0.002	2.91 (0.004)
LDL-cholesterol	0.297 (0.053)	4.21×10^{-08}	0.141 (0.054)	0.009	2.06 (0.040)	0.435 (0.050)	3.43×10^{-16}	0.304 (0.054)	5.89×10^{-08}	1.77 (0.076)	0.300 (0.056)	2.46×10^{-07}	0.221 (0.052)	3.07×10^{-05}	1.04 (0.299)

Table 2 (continued)

	6 years					12 years					24 years				
	MC		FC			MC		FC			MC		FC		
	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)	β (SE)	*p* value	β (SE)	*p* value	*Z* value (*p* value)
Daughters	*N*=327		*N*=313			*N*=316		*N*=283			*N*=209		*N*=152		
Fasting glucose	0.319 (0.064)	1.30×10^{-06}	0.250 (0.056)	1.27×10^{-05}	0.81 (0.420)	0.359 (0.054)	1.34×10^{-10}	0.187 (0.032)	2.28×10^{-08}	2.75 (0.006)	0.044 (0.032)	0.173	0.042 (0.026)	0.113	0.06 (0.954)
Fasting insulin	−0.035 (0.058)	0.549	0.052 (0.059)	0.382	−1.05 (0.296)	0.141 (0.050)	0.004	−0.028 (0.043)	0.517	2.57 (0.010)	0.180 (0.120)	0.137	0.213 (0.111)	0.058	−0.20 (0.840)
HOMA2-B	0.006 (0.060)	0.923	0.114 (0.058)	0.051	−1.29 (0.196)	0.171 (0.050)	0.0007	0.082 (0.043)	0.059	1.36 (0.175)	0.254 (0.070)	4.22×10^{-04}	0.037 (0.045)	0.412	2.62 (0.009)
HOMA2-S	−0.018 (0.056)	0.751	0.049 (0.058)	0.402	−0.82 (0.410)	0.143 (0.052)	0.006	−0.030 (0.045)	0.5	2.53 (0.011)	0.017 (0.102)	0.864	0.276 (0.101)	0.007	−1.81 (0.071)
Triglycerides	0.159 (0.046)	6.169×10^{-04}	0.197 (0.039)	8.28×10^{-07}	−0.62 (0.538)	0.150 (0.055)	0.007	0.275 (0.042)	3.67×10^{-10}	−1.79 (0.073)	0.246 (0.071)	6.942×10^{-04}	0.186 (0.050)	2.956×10^{-04}	0.69 (0.487)
Cholesterol	0.252 (0.047)	2.08×10^{-07}	0.227 (0.058)	1.097×10^{-04}	0.33 (0.741)	0.343 (0.047)	4.28×10^{-12}	0.281 (0.047)	8.26×10^{-09}	0.93 (0.353)	0.332 (0.073)	1.210×10^{-05}	0.251 (0.060)	5.160×10^{-05}	0.86 (0.389)
HDL-cholesterol	0.219 (0.055)	8.47×10^{-05}	0.261 (0.050)	3.73×10^{-07}	−0.56 (0.572)	0.381 (0.054)	2.04×10^{-11}	0.270 (0.045)	6.14×10^{-09}	1.56 (0.118)	0.261 (0.074)	5.399×10^{-04}	0.338 (0.066)	1.230×10^{-06}	−0.77 (0.441)
LDL-cholesterol	0.363 (0.055)	1.86×10^{-10}	0.316 (0.060)	3.08×10^{-07}	0.57 (0.567)	0.420 (0.050)	2.56×10^{-15}	0.233 (0.051)	7.35×10^{-06}	2.62 (0.009)	0.380 (0.081)	6.62×10^{-06}	0.176 (0.060)	0.004	2.02 (0.044)

Association between offspring measures with that of each of the parents are presented as β values and SE. Differences between maternal and paternal effects are presented as *Z* values and *p* values. The *Z* value shows the difference between regression coefficients. All regression models were adjusted for age, sex and BMI

MC, mother–child association; FC, father–child association

Table 3 Genetic parent-of-origin effects of variants at *KCNQ1*, *KLF14* and *MOB2* on relevant cardiometabolic traits

Trait	rsID	Locus	Chr	Position	Ref. allele	Alt.allele	Maternal estimate	Maternal SE	Maternal *p* value	Paternal estimate	Paternal SE	Paternal *p* value	POE estimate	POE SE	POE *p* value
[a]HOMA2-B, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.200	0.286	0.484	0.240	0.264	0.364	−0.578	0.434	0.183
[a]HOMA2-B, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.238	0.209	0.254	−0.304	0.094	0.001	0.618	0.186	8.597×10^{-04}
[a]HOMA2-B, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.333	0.0387	0	0.190	0.157	0.226	0.359	0.020	0
Birthweight	rs2237892	*KCNQ1*	11	2818521	C	T	0.383	0.140	0.006	0.079	0.053	0.134	0.304	0.148	0.040
[b]Birthweight	rs2237892	*KCNQ1*	11	2818521	C	T	0.391	0.143	0.006	0.081	0.050	0.104	0.299	0.171	0.079
[a]Fasting glucose, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	−2.226	2.675	0.405	−2.59	2.940	0.378	−0.513	5.083	0.920
[a]Fasting glucose, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.038	0.0275	0.169	0.038	0.045	0.404	−0.079	0.043	0.063
[a]Fasting glucose, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.081	0.076	0.285	−0.022	0.082	0.784	−0.038	0.131	0.771
[a]Fasting insulin, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.357	0.352	0.310	0.276	0.297	0.352	−0.875	0.453	0.053
[a]Fasting insulin, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.246	0.359	0.493	−0.343	0.159	0.031	0.694	0.369	0.060
[a]Fasting insulin, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.252	0.209	0.228	0.203	0.137	0.140	0.411	0.358	0.250
[a]Triglycerides, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.0360	0.120	0.764	0.173	0.098	0.077	−0.289	0.062	2.710×10^{-06}
[a]Triglycerides, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.256	0.203	0.209	−0.023	0.149	0.876	−0.288	0.231	0.213
[a]Triglycerides, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.168	0.179	0.347	0.057	0.169	0.738	−0.353	0.385	0.360
[a]Cholesterol, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.0023	0.105	0.983	0.112	0.095	0.239	−0.227	0.122	0.062
[a]Cholesterol, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.157	0.088	0.075	0.089	0.099	0.367	−0.235	0.128	0.067
[a]Cholesterol, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.112	0.086	0.191	−0.013	0.038	0.734	−0.082	0.092	0.372
[a]HDL-C, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.0174	0.143	0.903	0.036	0.139	0.795	−0.213	0.110	0.052
[a]HDL-C, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	−0.103	0.071	0.149	0.109	0.106	0.307	−0.172	0.067	0.010
[a]HDL-C, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.078	0.078	0.316	−0.020	0.111	0.860	−0.072	0.074	0.333

Table 3 (continued)

Trait	rsID	Locus	Chr	Position	Ref. allele	Alt.allele	Maternal estimate	Maternal SE	Maternal *p* value	Paternal estimate	Paternal SE	Paternal *p* value	POE estimate	POE SE	POE *p* value
BMI, birth	rs2237892	*KCNQ1*	11	2818521	C	T	0.918	0.342	0.007	0.367	0.444	0.409	0.486	0.606	0.422
BMI, 6 years	rs2237892	*KCNQ1*	11	2818521	C	T	1.041	0.495	0.035	0.479	0.459	0.297	0.601	0.722	0.406
[a]BMI, 12 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.131	0.056	0.019	0.062	0.110	0.570	0.070	0.116	0.543
[a]BMI, 24 years	rs2237892	*KCNQ1*	11	2818521	C	T	0.163	0.060	0.007	−0.035	0.130	0.785	0.106	0.179	0.555
[a]Triglycerides, 6 years	rs4731702	*KLF14*	7	130748625	C	T	−0.084	0.069	0.224	0.123	0.071	0.083	−0.205	0.078	0.009
[a]Triglycerides, 12 years	rs4731702	*KLF14*	7	130748625	C	T	−0.075	0.070	0.286	0.064	0.061	0.295	−0.141	0.076	0.065
[a]Triglycerides, 24 years	rs4731702	*KLF14*	7	130748625	C	T	−0.132	0.096	0.170	0.104	0.097	0.286	−0.232	0.108	0.032
[a]HDL-C, 6 years	rs4731702	*KLF14*	7	130748625	C	T	0.136	0.050	0.007	0.019	0.050	0.705	0.108	0.058	0.062
[a]HDL-C, 12 years	rs4731702	*KLF14*	7	130748625	C	T	0.100	0.045	0.027	0.020	0.039	0.618	0.084	0.051	0.100
[a]HDL-C, 24 years	rs4731702	*KLF14*	7	130748625	C	T	0.087	0.044	0.048	0.007	0.041	0.857	0.076	0.048	0.112
[a]Cholesterol, 6 years	rs4731702	*KLF14*	7	130748625	C	T	0.027	0.043	0.525	0.020	0.035	0.566	0.007	0.047	0.878
[a]Cholesterol, 12 years	rs4731702	*KLF14*	7	130748625	C	T	0.033	0.033	0.320	−0.004	0.037	0.923	0.035	0.043	0.414
[a]Cholesterol, 24 years	rs4731702	*KLF14*	7	130748625	C	T	0.026	0.036	0.473	0.042	0.041	0.311	−0.018	0.046	0.703
[a]Fat mass, 6 years	rs4731702	*KLF14*	7	130748625	C	T	−0.00	0.057	0.958	0.062	0.061	0.305	−0.055	0.068	0.422
[a]Skinfold, 6 years	rs4731702	*KLF14*	7	130748625	C	T	−0.074	0.031	0.019	0.007	0.032	0.828	−0.075	0.038	0.045
[a]HDL-C, 6 years	rs2334499	*MOB2*	11	1675619	C	T	−0.119	0.056	0.034	−0.029	0.050	0.567	−0.087	0.072	0.223
[a]HDL-C, 12 years	rs2334499	*MOB2*	11	1675619	C	T	−0.104	0.042	0.012	0.027	0.046	0.554	−0.131	0.055	0.018
[a]HDL-C, 24 years	rs2334499	*MOB2*	11	1675619	C	T	−0.110	0.043	0.011	−0.012	0.044	0.789	−0.089	0.053	0.093

All analyses were adjusted for age, sex and BMI

[a]Calculated as $\log_2$ value

[b]Adjusted for gestational age

rsID, SNP ID; Ref., reference; Alt. alternate; maternal/paternal estimate, β value for the association of maternal allele against reference homozygous genotype for a given trait; maternal/paternal SE, SE value for the association of maternal/paternal allele against reference homozygous genotype for a given trait; maternal/paternal *p* value, *p* value for the association of maternal/paternal allele against reference homozygous genotype for a given trait; POE estimate, parent-of-origin estimate, β value for the association of reciprocal heterozygotes (maternal minor allele vs paternal minor allele) for a given trait; POE SE, SE value for the association of reciprocal heterozygotes (maternal minor allele vs paternal minor allele) for a given trait; POE *p* value, *p* value for the association of reciprocal heterozygotes (maternal minor allele vs paternal minor allele) for a given trait; HDL-C, HDL-cholesterol

to oral glucose and insulin action in target tissues, as well as lipid levels, have been previously described in adult offspring from families of patients with type 2 diabetes and cardiovascular diseases [1, 2, 12]. Since these studies were published, the developmental origins of these disorders have been well established, and the strongest window for epigenetic programming is now thought to be periconceptional [35] and before the three germ layers are established (gastrulation) [36]. This suggests that parental influences should be obvious from early life. To this end, we determined the heritability and parent-of-origin effects in a birth cohort that was followed up at regular intervals into young adulthood. As previously reported, these traits were robustly heritable from early childhood as observed when studied in adults [1, 37–44]. Some of these associations were also present in the smaller PCS cohort.

Our results show maternal effects for total and HDL-cholesterol levels, supporting the well-established previous findings [2, 12], and also show that the effects manifest from childhood, albeit with some differences. A stronger correlation for triglyceride levels was reported between mother and daughter in previous studies compared with our study [1, 12]; however, a significant maternal effect was seen in sons at 6 years and in daughters at 12 and 24 years in the PMNS. These maternal effects for lipid traits in the PMNS were also seen in the PCS cohort. Variants in *KLF14* and *MOB2*, which previously were shown to have parent-of-origin effects on type 2 diabetes [14], here showed parent-specific effects on lipid levels, with the risk-increasing allele being associated with lipid levels. These findings provide insights into parental programming of offspring risk of type 2 diabetes.

Several theories have been proposed to explain the evolutionary origins of parent-of-origin effects, which are a consequence of genomic imprinting. Haig and Moore's kinship theory suggests that imprinting evolved to adjust gene activity, with different allele dosages benefitting maternal and paternal relatives in terms of evolutionary fitness [45, 46]. Day and Bonduriansky's sexual antagonism and co-adaptation theory suggests that imprinting evolved to shape offspring resemblance to each parent [47, 48]. Wolf and Hager's maternal–offspring co-adaptation theory proposes that imprinting evolved to favour expression of the fitter allele at a specific gene location [49]. The common feature of these hypotheses is that some processes create a selective asymmetry between the maternally and paternally inherited allele copies at a specific locus, and this causes selection to favour differential expression of the alleles at the same locus [50]. Our results showing a predominantly maternal influence on birthweight and metabolism of major macronutrients (glucose and lipids) during puberty and reproductive age suggests an intergenerational influence on resource allocation towards fecundity, which is a major component of the life-history theory [51].

The transitioning of parental–offspring associations across time from paternal to maternal (supported by genetic associations, e.g. for *KCNQ1* and HOMA2-B) suggests that the effects may be mediated by epigenetics, as the genome remains constant. Epigenetics is a link between the genes and the environment that facilitates the modulation of expression of a particular trait [52, 53]; this kind of programming superimposed on top of the genetic material occurs by means of chemical moieties (e.g. DNA methylation, histone modifications and miRNA), which can alter the way the DNA is read and expressed. Early-life exposures can bring about such epigenetic reprogramming and alter development and function of organs, which, in later life, can increase susceptibility to cardiometabolic disorders [53]. Waterland, Gunasekara and colleagues recently described 'CoRSIVs' (correlated regions of systemic interindividual variation) for methylation consistency across tissues representing the three germ layers. These regions were strongly influenced by the periconceptional environment, with some showing epigenetic metastability, and these alterations programmed the risk for future cardiometabolic diseases [54]. *KCNQ1* may well represent an example of such a region [54], as supported by the maternal parent-of-origin effect on birthweight. Furthermore, parent-of-origin effects have been shown to have spatial and temporal effects. Such effects in early life can have implications for development, while in later life they may have implications for function. For example, *KCNQ1* has been shown to have monoallelic expression in fetal islets but expression is biallelic in adults [55]. Functional studies are required to dissect the implications of such dynamic parent-specific effects.

For insulin secretion, the paternal effect in early childhood (6 years) gives way to a maternal effect at 12 years that continues to adulthood (24 years). For fasting insulin and insulin sensitivity, the negative maternal effect at 6 years changed to a positive effect at 12 years. The timing of this transition seemingly spans pubertal age (generally considered to be between 8 and 14 years of age). Metabolism and puberty are strongly interlinked; the link between nutrition and pubertal development requires maintenance of a minimum positive energy balance, especially in girls [56–59], and undernutrition or overnutrition can have a significant impact on the timing and progress of pubertal development and indeed fertility [56, 60]. It may therefore be speculated that parent-of-origin effects accompany pubertal changes given the increased developmental plasticity at this time period [61]. Nevertheless, it remains to be seen whether the shift in parental programming is a cause, consequence or by-product linked to pubertal processes.

It is interesting to note that anthropometric measures do not show a parent-specific association, unlike that seen for metabolic traits. Anthropometric traits are a consequence of genetics and environment, with heritability estimates increasing over time, which may be partially attributed to environmental influences. The reason for the discrepancy between anthropometric and metabolic traits in terms of

intergenerational transmission of risk remains uncertain. The life-course theory may explain parents' changing roles in the allocation of resources for maintenance, growth, reproduction and immunity. Further studies in the purview of evolutionary biology could shed more light on this.

Our study has several limitations. The findings are observational, whereby associations between parental and offspring phenotypes across trajectories of early childhood are examined, and are therefore not causal. Furthermore, this study is based on two birth cohorts from a single city in India, and validation is required in other similar as well as diverse populations to substantiate the findings. Nevertheless, our study is based on birth cohorts with robust study power and extensive follow-up, and therefore has the potential to answer novel questions, as well as providing a context to findings in adult offspring from other populations. The results of genetic and epigenetic studies in family cohorts as well as target tissues will be very useful in unravelling the mechanisms underlying these parental biases and the evolution of parental programming states.

This proof-of-concept study demonstrates parent-of-origin effects on cardiometabolic traits, from birth through childhood and puberty and into adulthood. We extend the life-history theory to demonstrate the life-course evolution of maternal influences on macronutrient metabolism of the offspring. We hypothesise that this indicates serial epigenetic reprogramming (of glucose metabolism) during pubertal and reproductive windows in the life course against the background of persistent maternal genetic effects (on lipid levels). These effects have implications for fecundity and survival of the offspring. Our results have the potential to provide a novel lead into primordial prevention of cardiometabolic diseases.

Supplementary Information The online version contains peer-reviewed but unedited supplementary material available at https://doi.org/10.1007/s00125-025-06396-5.

Acknowledgements We are grateful to all study participants and their family members for their cooperation over many years. We thank C. Fall, B. Coyaji, V. N. Rao (all KEM Hospital Research Centre, India) and the late D. J. P. Barker (Southampton General Hospital, UK) for their support in establishing the PMNS. We also thank the staff of the Diabetes Unit at the King Edward Memorial Hospital Research Centre for their help in conducting the study over a 30 year period. We are grateful to the Indian Council of Medical Research, the Department of Biotechnology of the Government of India, the Wellcome Trust and the Medical Research Council, UK for their funding support. We thank J. Postma for grant management, I. Artner and N. Wierup for critical reading of the manuscript, M. Marziaz for statistical advice and Leif Groop for critical insights and advice on the parent-of-origin concepts (all Lund University, Sweden). Some of the data formed part of GH's PhD thesis. In addition, some of the data were presented as an abstract at the 60th EASD Annual Meeting in 2024.

Data availability The data that support the findings of this study are not openly available for reasons of sensitivity, and are available from the corresponding author upon reasonable request. Data are stored in controlled-access data storage at the King Edward Memorial Hospital, Pune, India, and Lund University, Malmö, Sweden.

Funding Open access funding provided by Lund University. This study was supported by Network grant for Indo-Swedish collaboration from the Swedish Research Council (2015-06722) and the Department of Science and Technology of the Government of India (DST/INT/SWD/VR/P-04/2016) to RBP and CSY. The PMNS and PCS cohorts were funded by the Wellcome Trust, UK (grants 038128/Z/93, 059609/Z/99, 079877/Z/06/Z, 098575/B/12/Z and 083460/Z/07/Z), the Medical Research Council, UK (grant MR/J000094/1) and the Department of Biotechnology, Government of India (grant BT/PR-6870/PID/20/268/2005). The PMNS cohort was also funded intramurally by the King Edward Memorial Hospital Research Centre.

RBP was supported by the Crafoord Foundation (grant number 20200891), the Åke Wibergs Stiftelse (grant number M20-0214), the Swedish Heart Lung Foundation (grant number 20180522), the Hjelt Foundation, the Swedish Research Council (2021-02623), VINNOVA (2023-04234), the Direktör Albert Påhlssons Stiftelse, the EFSD and the Lilly European Diabetes Research Programme. CSY was a visiting professor at the Danish Diabetes Academy (supported by Novo Nordisk Foundation) and the Southern University of Denmark during 2016-2019.

Authors' relationships and activities SB is employed at Cytel Inc. All authors declare that there are no other relationships or activities that might bias, or be perceived to bias, their work.

Contribution statement RW, GH, JA and RBP analysed the data. RBP developed the study concept and design. RW and RBP wrote the manuscript. All authors took part in interpretation of the results, reviewed the manuscript critically for important intellectual content, and gave final approval of the version to be published. RBP and CSY are the guarantors of this work, and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Almgren P, Lehtovirta M, Isomaa B et al (2011) Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 54:2811–2819. https://doi.org/10.1007/s00125-011-2267-5

2. Groop L, Forsblom C, Lehtovirta M et al (1996) Metabolic consequences of a family history of NIDDM (the Botnia study): evidence for sex-specific parental effects. Diabetes 45:1585–1593. https://doi.org/10.2337/diab.45.11.1585
3. Suzuki K, Hatzikotoulas K, Southam L et al (2024) Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. Nature 627:347–357. https://doi.org/10.1038/s41586-024-07019-6
4. Mahajan A, Spracklen CN, Zhang W et al (2022) Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. Nat Genet 54:560–572. https://doi.org/10.1038/s41588-022-01058-3
5. Lawson HA, Cheverud JM, Wolf JB (2013) Genomic imprinting and parent-of-origin effects on complex traits. Nat Rev Genet 14:609–617. https://doi.org/10.1038/nrg3543
6. Lyssenko V, Groop L, Prasad RB (2015) Genetics of type 2 diabetes: it matters from which parent we inherit the risk. Rev Diabet Stud 12:233–242. https://doi.org/10.1900/RDS.2015.12.233
7. Hemminki K, Li X, Sundquist K, Sundquist J (2010) Familial risks for type 2 diabetes in Sweden. Diabetes Care 33:293–297. https://doi.org/10.2337/dc09-0947
8. Prasad RB, Lessmark A, Almgren P et al (2016) Excess maternal transmission of variants in the THADA gene to offspring with type 2 diabetes. Diabetologia 59:1702–1713. https://doi.org/10.1007/s00125-016-3973-9
9. Barker DJ, Hales CN, Fall CH, Osmond C, Phipps K, Clark PM (1993) Type 2 (non-insulin-dependent) diabetes mellitus, hypertension and hyperlipidaemia (syndrome X): relation to reduced fetal growth. Diabetologia 36:62–67. https://doi.org/10.1007/BF00399095
10. Jaquet D, Gaboriau A, Czernichow P, Levy-Marchal C (2000) Insulin resistance early in adulthood in subjects born with intrauterine growth retardation. J Clin Endocrinol Metab 85:1401–1406. https://doi.org/10.1210/jcem.85.4.6544
11. Nilsson PM, Ostergren PO, Nyberg P, Soderstrom M, Allebeck P (1997) Low birth weight is associated with elevated systolic blood pressure in adolescence: a prospective study of a birth cohort of 149378 Swedish boys. J Hypertens 15:1627–1631. https://doi.org/10.1097/00004872-199715120-00064
12. Predazzi IM, Sobota RS, Sanna S et al (2015) Sex-specific parental effects on offspring lipid levels. J Am Heart Assoc 4:e001951. https://doi.org/10.1161/JAHA.115.001951
13. Lessmark A, Hatem G, Kovacs G et al (2021) Lipid-associated variants near ANGPTL3 and LPL show parent-of-origin specific effects on blood lipid levels and obesity. Genes (Basel) 13:91. https://doi.org/10.3390/genes13010091
14. Kong A, Steinthorsdottir V, Masson G et al (2009) Parental origin of sequence variants associated with complex diseases. Nature 462:868–874. https://doi.org/10.1038/nature08625
15. Greene ND, Stanier P, Moore GE (2011) The emerging role of epigenetic mechanisms in the etiology of neural tube defects. Epigenetics 6:875–883. https://doi.org/10.4161/epi.6.7.16400
16. Antony AC, Vora RM, Karmarkar SJ (2022) The silent tragic reality of Hidden Hunger, anaemia, and neural-tube defects (NTDs) in India. Lancet Reg Health Southeast Asia 6:100071. https://doi.org/10.1016/j.lansea.2022.100071
17. Lismer A, Kimmins S (2023) Emerging evidence that the mammalian sperm epigenome serves as a template for embryo development. Nat Commun 14:2142. https://doi.org/10.1038/s41467-023-37820-2
18. Hardikar AA, Satoor SN, Karandikar MS et al (2015) Multigenerational undernutrition increases susceptibility to obesity and diabetes that is not reversed after dietary recuperation. Cell Metab 22:312–319. https://doi.org/10.1016/j.cmet.2015.06.008
19. Su L, Patti ME (2019) Paternal nongenetic intergenerational transmission of metabolic disease risk. Curr Diab Rep 19:38. https://doi.org/10.1007/s11892-019-1163-0
20. WHO Study Group on Diabetes Mellitus & World Health Organization (1985) Diabetes mellitus. Report of a WHO study group. World Health Organ Tech Rep Ser 727:1–113
21. Zhang X, Ma N, Lin Q et al (2024) Body roundness index and all-cause mortality among US adults. JAMA Netw Open 7:e2415051. https://doi.org/10.1001/jamanetworkopen.2024.15051
22. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC (1985) Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 28:412–419. https://doi.org/10.1007/BF00280883
23. Taliun D, Harris DN, Kessler MD et al (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590:290–299. https://doi.org/10.1038/s41586-021-03205-y
24. Das S, Forer L, Schonherr S et al (2016) Next-generation genotype imputation service and methods. Nat Genet 48:1284–1287. https://doi.org/10.1038/ng.3656
25. Fuchsberger C, Abecasis GR, Hinds DA (2015) minimac2: faster genotype imputation. Bioinformatics 31:782–784. https://doi.org/10.1093/bioinformatics/btu704
26. Hanson RL, Guo T, Muller YL et al (2013) Strong parent-of-origin effects in the association of KCNQ1 variants with type 2 diabetes in American Indians. Diabetes 62:2984–2991. https://doi.org/10.2337/db12-1767
27. Schriever SC, Kabra DG, Pfuhlmann K et al (2020) Type 2 diabetes risk gene Dusp8 regulates hypothalamic Jnk signaling and insulin sensitivity. J Clin Invest 130:6093–6108. https://doi.org/10.1172/JCI136363
28. Chen Z, Yin Q, Ma G, Qian Q (2010) KCNQ1 gene polymorphisms are associated with lipid parameters in a Chinese Han population. Cardiovasc Diabetol 9:35. https://doi.org/10.1186/1475-2840-9-35
29. Beaumont RN, Flatley C, Vaudel M et al (2023) Genome-wide association study of placental weight identifies distinct and shared genetic influences between placental and fetal growth. Nat Genet 55:1807–1819. https://doi.org/10.1038/s41588-023-01520-w
30. Warrington NM, Beaumont RN, Horikoshi M et al (2019) Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nat Genet 51:804–814. https://doi.org/10.1038/s41588-019-0403-1
31. Højsgaard S, Halekoh U, Yan J (2005) The R package geepack for generalized estimating equations. J Stat Softw 15:1–11. https://doi.org/10.18637/jss.v015.i02
32. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ 85:660–667. https://doi.org/10.2471/BLT.07.043497
33. WHO Consultation on Obesity & World Health Organization (2000) Obesity: preventing and managing the global epidemic. Report of a WHO consultation. World Health Organ Tech Rep Ser 894:i-xii, pp 1–253. Available from: https://iris.who.int/handle/10665/42330. Accessed 17 Jan 2025
34. World Health Organization (2025) Growth reference data for 5-19 years. Available from: https://www.who.int/tools/growth-reference-data-for-5to19-years. Accessed 17 Jan 2025
35. Stephenson J, Heslehurst N, Hall J et al (2018) Before the beginning: nutrition and lifestyle in the preconception period and its importance for future health. Lancet 391:1830–1841. https://doi.org/10.1016/S0140-6736(18)30311-8
36. Fleming TP, Watkins AJ, Velazquez MA et al (2018) Origins of lifetime health around the time of conception: causes and consequences. Lancet 391:1842–1852. https://doi.org/10.1016/S0140-6736(18)30312-X
37. Vujkovic M, Keaton JM, Lynch JA et al (2020) Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry

meta-analysis. Nat Genet 52:680–691. https://doi.org/10.1038/s41588-020-0637-y
38. Vogelezang S, Bradfield JP, Ahluwalia TS et al (2020) Novel loci for childhood body mass index and shared heritability with adult cardiometabolic traits. PLoS Genet 16:e1008718. https://doi.org/10.1371/journal.pgen.1008718
39. Mahajan A, Taliun D, Thurner M et al (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 50:1505–1513. https://doi.org/10.1038/s41588-018-0241-6
40. Frahnow T, Osterhoff MA, Hornemann S et al (2017) Heritability and responses to high fat diet of plasma lipidomics in a twin study. Sci Rep 7:3750. https://doi.org/10.1038/s41598-017-03965-6
41. Whitfield JB (2014) Genetic insights into cardiometabolic risk factors. Clin Biochem Rev 35:15–36
42. Prasad RB, Groop L (2015) Genetics of type 2 diabetes-pitfalls and possibilities. Genes (Basel) 6:87–123. https://doi.org/10.3390/genes6010087
43. Poulsen P, Levin K, Petersen I, Christensen K, Beck-Nielsen H, Vaag A (2005) Heritability of insulin secretion, peripheral and hepatic insulin action, and intracellular glucose partitioning in young and old Danish twins. Diabetes 54:275–283. https://doi.org/10.2337/diabetes.54.1.275
44. Woo JG, Morrison JA, Stroop DM, Aronson Friedman L, Martin LJ (2014) Genetic architecture of lipid traits changes over time and differs by race: Princeton Lipid Follow-up Study. J Lipid Res 55:1515–1524. https://doi.org/10.1194/jlr.M049932
45. Haig D (2004) Genomic imprinting and kinship: how good is the evidence? Annu Rev Genet 38:553–585. https://doi.org/10.1146/annurev.genet.37.110801.142741
46. Moore T, Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet 7:45–49. https://doi.org/10.1016/0168-9525(91)90040-W
47. Day T, Bonduriansky R (2004) Intralocus sexual conflict can drive the evolution of genomic imprinting. Genetics 167:1537–1546. https://doi.org/10.1534/genetics.103.026211
48. Day T, Bonduriansky R (2011) A unified approach to the evolutionary consequences of genetic and nongenetic inheritance. Am Nat 178:E18-36. https://doi.org/10.1086/660911
49. Wolf JB, Hager R (2006) A maternal-offspring coadaptation theory for the evolution of genomic imprinting. PLoS Biol 4:e380. https://doi.org/10.1371/journal.pbio.0040380
50. Patten MM, Ross L, Curley JP, Queller DC, Bonduriansky R, Wolf JB (2014) The evolution of genomic imprinting: theories, predictions and empirical tests. Heredity (Edinb) 113:119–128. https://doi.org/10.1038/hdy.2014.29
51. Nettle D, Frankenhuis WE (2019) The evolution of life-history theory: a bibliometric analysis of an interdisciplinary research area. Proc Biol Sci 286:20190040. https://doi.org/10.1098/rspb.2019.0040
52. Waddington CH (1968) Towards a theoretical biology. Nature 218:525–527. https://doi.org/10.1038/218525a0
53. Waterland RA, Michels KB (2007) Epigenetic epidemiology of the developmental origins hypothesis. Annu Rev Nutr 27:363–388. https://doi.org/10.1146/annurev.nutr.27.061406.093705
54. Gunasekara CJ, Scott CA, Laritsky E et al (2019) A genomic atlas of systemic interindividual epigenetic variation in humans. Genome Biol 20:105. https://doi.org/10.1186/s13059-019-1708-1
55. Travers ME, Mackay DJ, Dekker Nitert M et al (2013) Insights into the molecular mechanism for type 2 diabetes susceptibility at the KCNQ1 locus from temporal changes in imprinting status in human islets. Diabetes 62:987–992. https://doi.org/10.2337/db12-0819
56. Frisch RE, McArthur JW (1974) Menstrual cycles: fatness as a determinant of minimum weight for height necessary for their maintenance or onset. Science 185:949–951. https://doi.org/10.1126/science.185.4155.949
57. Bergendahl M, Perheentupa A, Huhtaniemi I (1991) Starvation-induced suppression of pituitary-testicular function in rats is reversed by pulsatile gonadotropin-releasing hormone substitution. Biol Reprod 44:413–419. https://doi.org/10.1095/biolreprod44.3.413
58. Kile JP, Alexander BM, Moss GE, Hallford DM, Nett TM (1991) Gonadotropin-releasing hormone overrides the negative effect of reduced dietary energy on gonadotropin synthesis and secretion in ewes. Endocrinology 128:843–849. https://doi.org/10.1210/endo-128-2-843
59. Aloi JA, Bergendahl M, Iranmanesh A, Veldhuis JD (1997) Pulsatile intravenous gonadotropin-releasing hormone administration averts fasting-induced hypogonadotropism and hypoandrogenemia in healthy, normal weight men. J Clin Endocrinol Metab 82:1543–1548. https://doi.org/10.1210/jcem.82.5.3947
60. Rittmaster RS, Deshwal N, Lehman L (1993) The role of adrenal hyperandrogenism, insulin resistance, and obesity in the pathogenesis of polycystic ovarian syndrome. J Clin Endocrinol Metab 76:1295–1300
61. Hochberg Z (2011) Developmental plasticity in child growth and maturation. Front Endocrinol (Lausanne) 2:41. https://doi.org/10.3389/fendo.2011.00041

Authors and Affiliations

Rucha Wagh[1,2] · Gad Hatem[3] · Jonas Andersson[3] · Pooja Kunte[1] · Souvik Bandyopadhyay[4] · Chittaranjan S. Yajnik[1] · Rashmi B. Prasad[3,5]

✉ Rashmi B. Prasad
rashmi.prasad@med.lu.se

[1] Diabetes Unit, Kamalnayan Bajaj Diabetology Research Centre, King Edward Memorial Hospital and Research Centre, Pune, India

[2] Symbiosis School of Biological Sciences, Symbiosis International (Deemed University), Pune, India

[3] Department of Clinical Sciences, Diabetes and Endocrinology, Lund University, Malmö, Sweden

[4] Cytel Inc., Cambridge, MA, USA

[5] Institute of Molecular Medicine Finland, Helsinki University, Helsinki, Finland

Paper 4

scientific reports

Check for updates

OPEN

Characterizing the allele-specific gene expression landscape in high hyperdiploid acute lymphoblastic leukemia with BASE

Jonas Andersson[1,2], Efe Aydın[1], Rebeqa Gunnarsson[1], Henrik Lilljebjörn[1], Thoas Fioretos[1,3], Bertil Johansson[1,3], Kajsa Paulsson[1] & Minjun Yang[1]✉

Somatic copy number variations (CNVs), including abnormal chromosome numbers and structural changes leading to gain or loss of genetic material, play a crucial role in initiation and progression of cancer. CNVs are believed to cause gene dosage imbalances and modify cis-regulatory elements, leading to allelic expression imbalances in genes that influence cell division and thereby contribute to cancer development. However, the impact of CNVs on allelic gene expression in cancer remains unclear. Allele-specific expression (ASE) analysis, a potent method for investigating genome-wide allelic imbalance profiles in tumors, assesses the relative expression of two alleles using high-throughput sequencing data. However, many existing methods for gene-level ASE detection rely on only RNA sequencing data, which present challenges in interpreting the genetic mechanisms underlying ASE in cancer. To address this issue, we developed a robust framework that integrates allele-specific copy number calls into ASE calling algorithms by leveraging paired genome and transcriptome data from the same sample. This integration enhances the interpretability of the genetic mechanisms driving ASE, thereby facilitating the identification of driver events triggered by CNVs in cancer. In this study, we utilized BASE to conduct a comprehensive analysis of ASE in high hyperdiploid acute lymphoblastic leukemia (HeH ALL), a prevalent childhood malignancy characterized by gains of chromosomes X, 4, 6, 10, 14, 17, 18, and 21. Our analysis unveiled the comprehensive ASE landscape in HeH ALL. Through a multi-perspective examination of HeH ASEs, we offer a systematic understanding of how CNVs impact ASE in HeH, providing valuable insights to guide ASE studies in cancer.

In diploid organisms, the maternal and paternal gene alleles are generally expressed equally. However, there is a group of genes in the genome where transcription originates predominantly from one allele, a phenomenon known as allele-specific expression (ASE). In humans, ASE is often associated with epigenetic inactivation in the X-chromosome or autosomal imprinting, an epigenetic phenomenon resulting in differential parent-origin gene expression[1]. ASE may also be driven by genetic variation. For example, ASE can occur due to genetic variants that lead to alternative splicing or cis-regulatory mutations that cause the two alleles to be activated by different upstream regulators. Additionally, ASE analysis of somatic mutations has unveiled a preferential selection for the expression of some alleles with well-known driver mutations, contributing to the clonal outgrowth of tumors[2]. In tumor cells, an important class of events is somatic copy number variation (CNV). The pathogenetic impact of CNVs is believed to be caused by gene dosage imbalances and allelic expression imbalance of genes that affect cell proliferation. For instance, the partial tandem duplication of the *KMT2A* gene is linked to upregulated *KMT2A* expression in acute myeloid leukemia (AML), contributing to AML development by promoting uncontrolled cell proliferation and inhibiting normal cellular differentiation[3]. Aneuploidy, *i.e.*, an abnormal number of chromosomes, is commonly observed in many cancer cells[4]. Evidence from RNA sequencing (RNA-seq) of aneuploid cells suggests upregulation of trisomic genes (those localized on trisomic chromosomes) at the mRNA level when compared to their disomic counterparts[4,5]. However, the extent to which aneuploidy alters the allelic expression of genes in cancer and the overall role of ASE in CNVs remain unknown. ASE analysis generally involves counting RNA-seq reads carrying reference and alternative alleles over heterozygous sites and

[1]Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, Lund, Sweden. [2] Lund University Diabetes Centre, Department of Clinical Sciences Malmö, Lund University, Malmö, Sweden. [3]Department of Clinical Genetics, Pathology, and Molecular Diagnostics, Office for Medical Services, Laboratory Medicine, Region Skåne, Lund, Sweden. ✉email: minjun.yang@med.lu.se

combining read count information from heterozygous loci within a gene to determine the ASE extent[6,7]. The major problem with these RNA-seq-only techniques is that the allele frequency of heterozygous variant sites in a specific chromosome may be skewed due to genetic imbalance, leading to errors if the genetic background information is missing. This is particularly important for ASE genes identified in aneuploid cancers, where the statistical approaches could fail to resolve the unbalanced heterozygous genotypes present in these samples. To measure accurately CNV states and to study ASE in the context of cancer, we developed a computational tool called Biomedswe Allele-Specific Expression-analyzer (BASE), a method consisting of functionalities to estimate the allele-specific copy number (asCN) states and to identify ASE genes by combining analysis of asCN and RNA-seq data. Specifically, our approach can detect ASE genes in aneuploid cancer samples, irrespective of the availability of whole genome sequencing (WGS) data from matched normal samples. Furthermore, we present a case study where BASE was employed to screen ASE genes using published data on high hyperdiploid acute lymphoblastic leukemia (HeH ALL), one of the most common childhood malignancies, which is characterized by gains of chromosomes X, 4, 6, 10, 14, 17, 18, and 21 in over 80% of cases[8].

Methods

WGS data analysis and variants annotation in BASE

The BASE pipeline is designed to generate a gene-based ASE analysis using matched WGS and RNA-seq data generated by next-generation sequencing (NGS) technologies (https://github.com/biomedswe/BASE). The pipeline is structured so that WGS and RNA-seq data are analyzed independently (Fig. 1). SeqPurge is used to trim the adapter sequences and the low-quality bases from raw Illumina paired-end sequencing reads[9]. The resultant trimmed WGS reads are aligned to the human genome reference build hg38 (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/) using BWA[10] with sambamba[11] used for sorting, marking duplicate reads, and indexing of alignment files. Single nucleotide variants (SNVs) are called by GATK HaplotypeCaller[12] and only heterozygous SNVs meeting specific criteria (DP > 15, QD > 2, MQ > 35, MQRankSum > -12.5, ReadPosRankSum > -8, and FS < 60) within exonic regions of the APPRIS principal isoforms of protein-coding genes are retained. Annotation of SNVs is performed by SnpEff[13].

Reference panel for copy number analysis

BASE utilizes a copy number analysis reference panel compiled from unphased WGS data from 196 in-house normal samples, obtained from the population in Southern Sweden, and comprising 104 males and 92 females. This panel encompasses normalized sequencing depth data spanning around 3.1 million SNV sites across all 23 human chromosomes. For each normal sample, we performed WGS read alignment and bam file processing and called the sequencing depth of defined loci by the GATK HaplotypeCaller gvcf model[12]. We kept only SNVs with

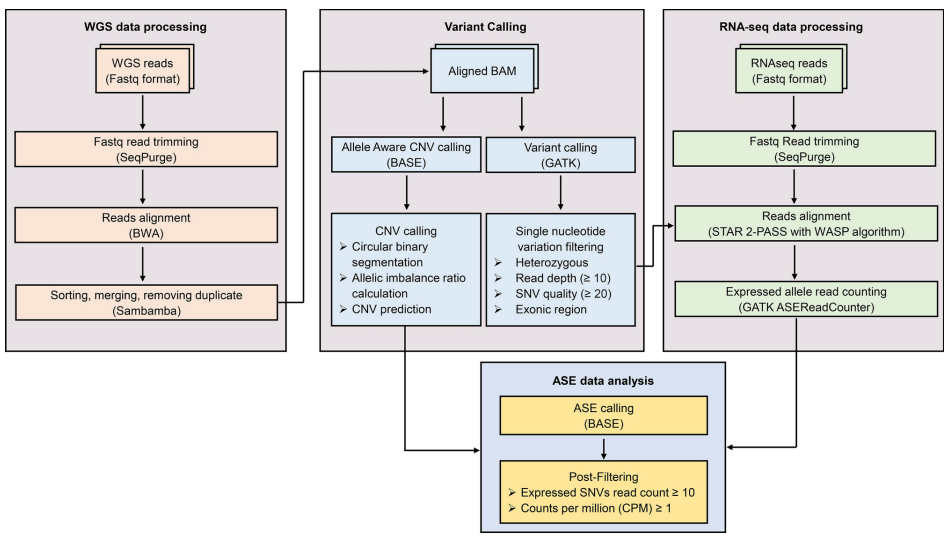

Fig. 1. The BASE workflow is based on whole genome sequencing (WGS) and RNA sequencing (RNA-seq) paired-end fastq files from the same sample. SeqPurge filters the raw sequencing reads, whereas BWA and STAR are used for read alignment. GATK is utilized for single nucleotide variants (SNVs) calling and for determining expressed heterozygous SNV read counts. BASE performs copy number variations (CNV) calling. Finally, ASE genes are identified by integrating heterozygous SNVs read counts and copy number states.

sequencing coverage over 10 and normalized the sequencing-depth using the average sequencing depth of each sample. We multiplied the normalized coverage by 2 for the X and Y chromosomes, respectively, in male samples and calculated the median value of normalized coverage across all samples for each SNV site.

CNV calling

For WGS data from the input sample, the reference and alternative allele read depths of SNV loci present in the built-in reference panel is called using the GATK HaplotypeCaller gvcf model and only informative sites with sequencing coverage over 10 are retained. Subsequently, read counts of SNV loci are normalized to the average sequencing depth. The Log R ratio (LRR) for each SNV locus is calculated as the log-odds ratio of the sequencing depth from the investigated sample over that in the normal reference panel. The circular binary segmentation algorithm in DNAcopy[14], using LRR data as inputs, is then employed for copy number calling. The resulting copy number levels are categorized into five classes: hemizygous deletions (1 copy), wild type (2 copies), 3 copies, 4 copies, and high amplification ($\geq$ 5 copies).

Next, we calculate the segment allelic imbalance ratio (AI_{seg}) for segments exceeding 100 kb. Briefly, for a given segment containing at least 20 heterozygous SNPs, we define the reference allele frequency (RAF) of each SNP locus as the reference allele count versus total sequencing depth in the investigated sample and calculate the AI value (AI_{snp}) for each site by

$$AI_{snp} = abs\,(RAF - 0.5)$$

Subsequently, the unsupervised k-means clustering algorithm with a squared Euclidean distance measure is applied to all AI_{snp} values within each segment. This process divides heterozygous SNPs into two groups, and we define the mean AI_{snp} value of each group as AF_h and AF_L, representing respectively the higher and lower mean AI_{snp} values of the given segment. Finally, AI_{seg} is calculated as

$$AI_{seg} = AF_h/AF_L$$

and the asCN states are estimated by combining the copy number levels and AI_{seg}.

RNA-seq data analysis and ASE

The trimmed RNA-seq reads are aligned to the human genome reference build hg38 (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/) using STAR 2-pass mapping pipeline with WASP parameters with sambamba used for bam file processing[15,16]. Only reads with mapping quality $\geq$ 20 and passed WASP filtering are kept. Reference and alternative allele read counts of expressed heterozygous SNVs are obtained using the GATK ASEReadCounter[12]. Heterozygous SNVs are excluded if the total read count of SNV sites is < 10. For genes within unbalanced chromosome segments, RNA-seq read counts are aggregated based on allele frequencies derived from matched WGS data for the corresponding disomic genes within the same chromosome segments. For the latter genes, we implement a pseudo-phasing procedure originally applied by Mayba et al.,[6] where the major and minor haplotypes are assigned by alleles with higher/lower RNA-seq read counts, respectively. The binomial test is subsequently applied to examine whether the allele ratio in the RNA-seq data significantly deviates from two null hypothesis models; model 1: ratio 1 (both alleles are equally expressed, assuming the sample exhibits biallelic copy number equivalence derived from both maternal and paternal genomes) and model 2: the alleles are not equally expressed, but the expression ratios align with the allele-specific copy numbers derived from genome-wide assessments. Only genes that display ASE under the first model but not the second are defined as ASE genes driven by somatic CNVs. A false discovery rate of 0.05 and an additional threshold of allelic imbalance ratio between WGS and RNA-seq (> 2 or < 0.5) data are considered for the identification of genes associated with ASE. For gene expression quantification, the RSEM[17] algorithm is applied using GENCODE v43 (https://www.gencodegenes.org/human/release_43.html) as the reference and only genes with CPM (counts per million) $\geq$ 1 are kept. To evaluate the performance of BASE in ASE calling, we compared it against two established ASE calling programs: MBASED and aScan[6,7]. For MBASED, we adhered to the guidelines from the original paper, applying an estimated MAF cutoff of $\geq$ 0.7 and an adjusted P-value threshold of $\leq$ 0.05. In the case of aSCAN, we considered only informative genes with more than 10 covered reads and used an adjusted P-value threshold of $\leq$ 0.05 to identify ASE genes.

Performance and resource usage

The pipeline uses GNU parallel as the parallelization mechanism to optimize runtime[18]. All genuine tests were conducted on a computer running Linux CentOS 7 with 20 physical cores and 192 GB of RAM. The pipeline peak memory usage was ~80 Gb for WGS data alignment using BWA. The runtime varies depending on multiple factors, such as sequencing depth and variants number of samples.

Results

asCN calling evaluation

By using a built-in copy number analysis reference panel, BASE provides a solution to estimate asCN changes using WGS data from tumor samples. To evaluate the precision of BASE in asCN calling, we first collected WGS data from fifteen extensively studied cancer cell lines (Supplementary Table 1) and compared the asCN calls made by BASE with those derived from matched SNP arrays, performed by the Cancer Cell Line Encyclopedia (CCLE)[19]. On average, BASE detected asCN changes with a recall rate of 98–100% in the eleven cell lines with ploidy levels ranging from 1.6 to 2.7 (Table 1). For the remaining four cell lines, exhibiting ploidy levels from

WGS dataset	Sample name	Ploidy	WGS coverage	Informative region (Mb)	Recall (Mb)	Overall accuracy (%)
Cell line WGS						
	A2058	2.21	33.10	2757.71	2740.17	99.36
	COLO 829 T	3.22	37.84	2768.17	2743.39	99.11
	HEPG2	2.26	42.04	2783.24	2776.96	99.77
	HT115	2.10	35.73	2759.91	2742.70	99.38
	KNS81	1.96	45.91	2761.61	2729.79	98.85
	LAMA84	3.11	42.52	2788.78	2777.25	99.59
	LOVO	2.20	37.91	2760.97	2749.19	99.57
	LS513	2.28	31.91	2758.90	2749.62	99.66
	MKN74	2.30	42.41	2712.25	2669.85	98.44
	MM1S	1.97	42.91	2759.18	2756.67	99.91
	MONO-MAC-1	2.09	38.35	2762.58	2744.30	99.34
	NCIH69	1.64	40.17	2717.08	2700.69	99.40
	OVCAR4	2.99	39.84	2728.81	2685.77	98.42
	SW1990	2.68	34.77	2745.19	2735.80	99.66
	T47D	2.89	40.32	2766.71	2721.87	98.38
Simulated cell line with normal cell contamination						
	COLO 829 T 75%	N.A	96.42	2754.77	2651.55	96.25
	COLO 829 T 50%	N.A	95.87	2747.81	2455.02	89.34
	COLO 829 T 25%	N.A	91.49	2779.28	1726.19	62.11

Table 1. Copy number calling results of 15 cancer cell lines and three simulated COLO 829 cancer cell lines with different tumor cell contents. *WGS*, whole genome sequencing. *N.A.* not applicable

Cell lines	BASE	aScan	MBASED	ASE found in all three programs	ASE found in both BASE and either aSCAN or MBASED
HepG2 (phased SNVs)	639	721	364	324	600
HepG2 (Unphased SNVs)	600	495	336	214	455
NCIH2122	1710	2557	1191	1086	1658
KNS62	1044	1516	721	617	1474
SW900	1362	2075	956	869	1328

Table 2. Overview of ASE Gene calling results across three programs in cell lines.

2.8 to 3.3, the BASE asCN calling result initially exhibited a low overall agreement (< 5%) with CCLE calls using default settings. However, after adjusting the tumor ploidy parameter in BASE, the recall rate was restored to > 98% (Table 1).

Tumor samples never contain 100% tumor cells but also normal cells, *e.g.*, infiltrating immune cells. This poses a problem for the accurate detection of asCN, as the reads from the tumor containing copy number aberrations are diluted with the reads from the normal diploid cells. To evaluate BASE's asCN calling performance in samples with normal cell contamination, we obtained WGS data from the COLO 829 cell line and modelled normal cell contamination by mixing WGS reads from COLO 829R, a B lymphoblast cell line derived from the same patient. In this way, samples containing 75%, 50%, and 25% tumor cell content were modeled. For samples with a 75% tumor purity, BASE detected most asCN events with a calling accuracy rate of 96%. However, it is worth noting that chromosomal regions with loss of heterozygosity (LOH) in COLO 829 were not called correctly in the simulated samples with less than 50% tumor cell fraction due to the increase of heterozygosity within those segments (Table 1; Supplementary Fig. 1).

Evaluation of ASE calling based on phased and unphased SNV Information

In the evaluation of BASE's precision in identifying allelic expression events, we analyzed three replicates of HepG2 cell line RNA-seq data, where heterozygous SNPs data had been obtained from using either a phasing or non-phasing WGS analysis pipeline[20]. By using the BASE, we identified 4,933 informative expressed genes shared by the ASE calling results using either phased or un-phased genomic data as input. Of these, 639 and 600 ASE genes, respectively (Table 2), were identified, with 599 ASE genes being consistently detected in both calling results (Supplementary Fig. 2); this suggests substantial equivalence between the ASE calling results based on unphased and phased data. The relationship between the ASE state and gene expression level was investigated, revealing no gene expression profile difference between ASE and non-ASE genes, indicating that our pipeline can detect ASE regardless of the gene expression level (Fig. 2a). To further BASE's performance in ASE calling, we compared its results with those from MBASED and aSCAN, revealing a strong correlation between the methods (Fig. 2b,c). MBASED identified 5,497 informative genes, with 364 and 336 exhibiting ASE under phased and

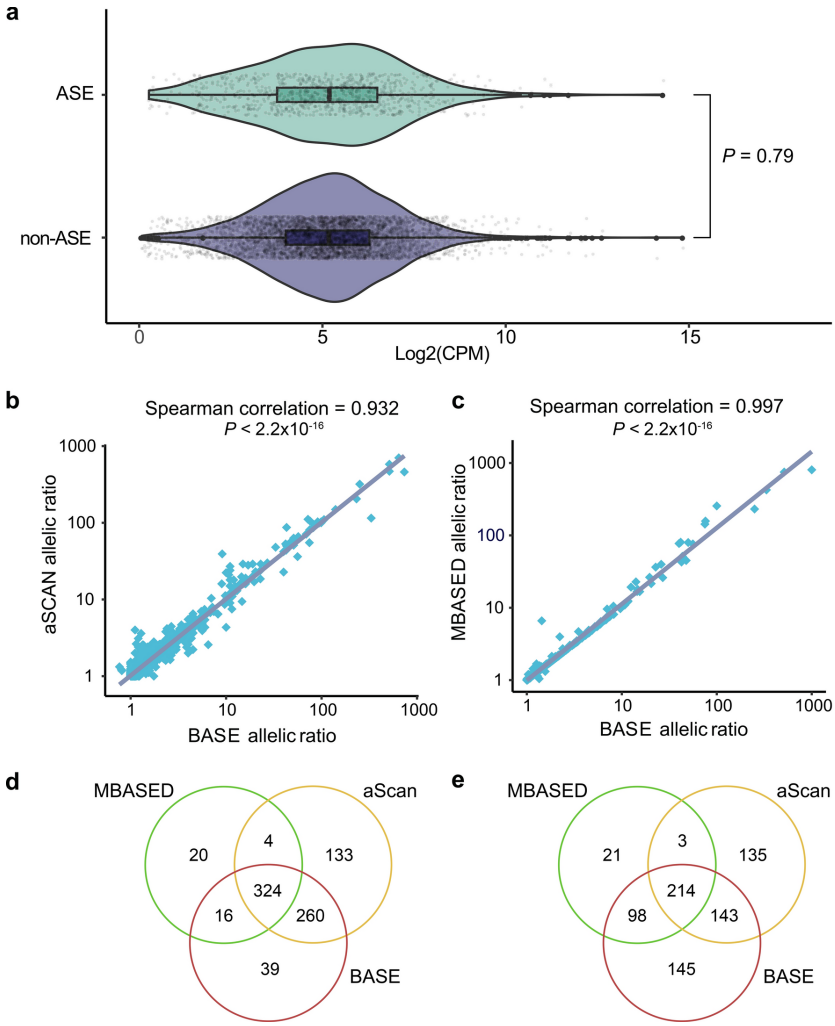

Fig. 2. **a.** Violin plots displaying the expression of ASE genes (top) and non-ASE genes (bottom). No significant differences in gene expression levels between ASE genes and non-ASE genes were observed ($P = 0.79$, Mann–Whitney two-sided test). Gene expression levels were normalized to CPM values and then log2-transformed as shown on the x-axis. The median is denoted by the center of the boxplot, while the lower and upper hinges correspond to the first and third quartiles, respectively. Whiskers represent 1.5 times the interquartile range, with data beyond this range plotted as individual points. Spearman's rank correlation between the allelic expression ratio derived from aSCAN and BASE (**b**), and from MBASED and BASE (**c**), respectively. The Venn diagram displays the overlap of ASE genes identified using phased (**d**) and unphased (**e**) genotype data by three tools: MBASED, aScan and BASE, respectively.

unphased conditions, respectively. aSCAN detected 5,546 informative genes, with ASE identified in 721 genes using phased data and 495 with unphased data (Table 2). Our analysis showed that over 93% of ASE genes identified by BASE were confirmed by MBASED or aSCAN with phased data, and over 75% with unphased data (Fig. 2d,e). These findings demonstrate that BASE's ASE calling is highly consistent with the published methods.

Notably, ASE detection was slightly more accurate when based on phased SNV data compared to unphased data across all three programs, underscoring the potential benefits of using phased genetic information to enhance the precision and reliability of ASE gene detection.

A similar approach was applied to three lung cancer cell lines, KNS62, NCIH2122, and SW900, with matched WGS and RNA-seq data derived from the CCLE (Supplementary Table 1). Our results show that the BASE program identifies over 59% of ASE genes detected across all programs, with more than 95% of the ASE genes called by BASE also recognized by at least one of the other published methods (Table 2). This suggests that BASE is robust in ASE gene calling across different tumor types.

Application: ASE in HeH ALL

To explore the impact of CNVs on allelic expression imbalance in aneuploid tumors, we applied the BASE framework to a previously generated HeH ALL dataset (Supplementary Table 2)[5,8,21]. This dataset comprised 12 samples with matched diagnosis/remission WGS data, along with RNA-seq data from the diagnostic samples. The selection of these 12 samples was based on the criteria that all samples were sequenced using the Illumina system, ensuring consistency in sequencing technology. Additionally, to minimize batch effects and enhance comparability, all selected samples had RNA-seq data generated from sequencing libraries constructed using the same kit. Details of sample characteristics and NGS library preparations are described in the original articles. To examine ASE genes in HeH ALL, we considered only expressed genes (CPM ≥ 1) that had at least one heterozygous site covered by 10 or more RNA-seq reads. In total, we identified 8,210 expressed genes in the twelve HeH ALL samples, with 3,653 (44.5%) genes showing ASE. The majority of genes exhibited ASE in only one sample (2,372/3,653; 64.9%; Fig. 3a; Supplementary Table 3). No significant difference in the percentage of ASE genes was found between informative samples with < 80% and ≥ 80% blast cells ($P = 0.143$, two-tailed

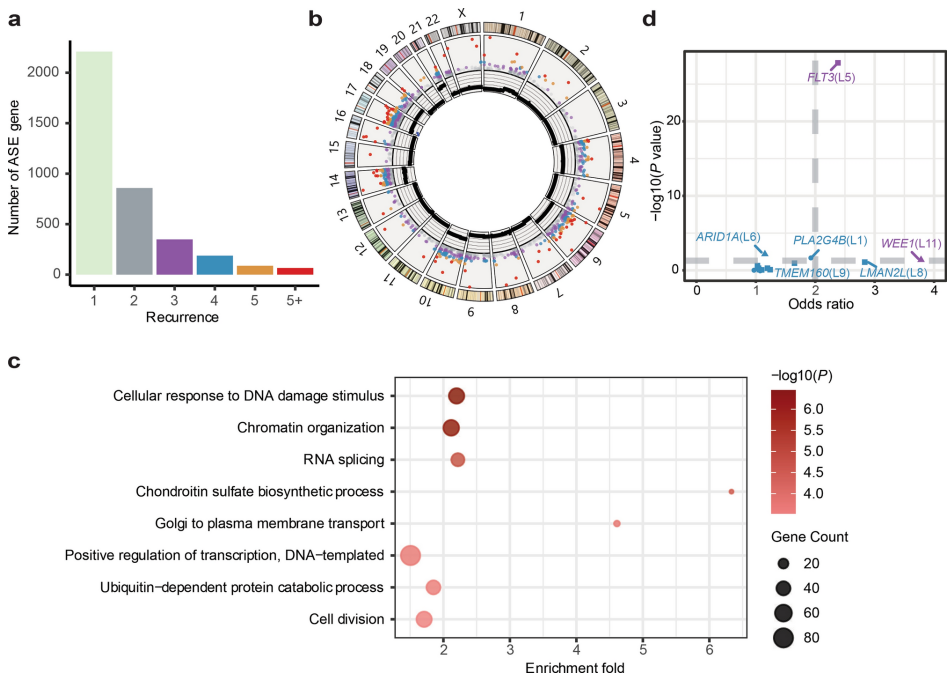

Fig. 3. **a.** Recurrence of ASE genes across 12 HeH ALL samples. **b.** The Circos plot shows ASE genes and copy number alterations of HeH ALL samples. From outer to inner circle: chromosome numbers, chromosome ideograms, scatter plot of recurrent ASEs genes across 12 HeH ALL samples (grey, recurrence = 2; purple, recurrence = 3; blue, recurrence = 4; orange, recurrence = 5; red, recurrence > 5), and average copy number of a particular segment of 12 HeH ALL samples. **c.** Gene Ontology analysis of 1,549 recurrent ASE genes in 12 HeH ALL samples. **d.** ASE states of mutated genes in HeH ALL. ASE genes are marked in purple. The circle, square and triangle indicate the missense, nonsense and silent mutations, respectively. The x-axis indicates the odds ratio of aggregated allelic ratios from RNA-seq compared to those from WGS. The y-axis indicates the log10 transformed binomial test *P* value.

unpaired Mann–Whitney U-test; Supplementary Table 2). Additionally, we found 1,281 genes that displayed ASE in at least two HeH ALL cases, *i.e.*, were recurrent (Fig. 3b; Supplementary Table 3). Gene ontology analysis of these recurrent ASE genes revealed enrichment of the biological processes cellular response to DNA damage stimulus (GO: 0006974; $P = 3.5 \times 10^{-7}$), chromatin organization (GO: 0006325, $P = 6.3 \times 10^{-7}$), and RNA splicing (GO: 0008380, $P = 1.7 \times 10^{-5}$) (Fig. 3c).

Proportion of ASE genes on trisomic versus disomic chromosomes

To investigate the relationship between non-random chromosomal gains and ASE in HeH ALL, we restricted our analysis to the 4,089 informative genes located within gained chromosomes (trisomy, tetrasomy with a 3:1 allele ratio, and pentasomy). We adapted the BASE framework to identify ASE genes that were associated with the chromosomal gains using binomial exact test under the null hypothesis model 1) each allele is equally expressed; or the alternative hypothesis model 2) two alleles are unequally expressed with the odds ratio based on asCN analysis result (Supplementary Fig. 3). Approximately 1,574 ASE genes were significantly correlated with gained chromosomes, with 22.8% (932/4,089) exhibiting recurrent ASE in at least two samples. Chromosomes 4, 6, 10, 14, 17, 18, and 21 were gained in more than 70% of HeH ALL samples and over 30% of the informative genes on these chromosomes showed allelic expression imbalance (Fig. 3b, Supplementary Table 4 and 5), indicating that the genomic imbalance caused by the copy number gains lead to widely unequal transcription of the two gene alleles in HeH ALL.

We have previously shown that genes in the gained chromosomes X, 14, and 21 are associated with the strongest dosage effects and that gains of chromosome 21 had the highest selective advantage in HeH ALL[5,8]. In line with this, we observed a higher frequency of ASE in chromosomes 14 and 21 (over 48%) in contrast to the other commonly gained chromosomes (range 38.4% to 45.9%), indicating that allelic imbalance in the expression of genes on these two chromosomes may be a key hallmark in HeH ALL (Supplementary Table 5). In contrast, when examining informative genes outside of the gained chromosomes, only 4.6% (294/6,441) showed recurrent ASE, highlighting the predominant influence of copy number gain events on ASE in HeH ALL (Fig. 3b, Supplementary Table 3).

To explore further whether the leukemia-related ASE events could be driven by chromosomal gains, we investigated ASE events involving known cancer driver genes reported in OncoKB (https://www.oncokb.org/cancer-genes). A total of 54 oncogenes, including four that are known to be mutated in leukemia (*NSD2*, *CCND3*, *U2AF1*, and *JAK2*)[22], showed ASE in 7/12 HeH ALL samples in trisomic chromosomes, suggesting that the expression of a set of well-known leukemia-related genes could also be altered by gene dosage in HeH ALL, although other causes of ASE cannot be excluded. However, only a few leukemia-related ASE events were observed in disomic chromosomes. For example, *KMT2C* is frequently mutated in multiple human cancers and heterozygous germline mutations in *KMT2C* have been reported to be enriched in infants with leukemia[23]. Here we observed ASE of *KMT2C* in three samples with disomy 7, indicating that the selective expression of specific alleles of *KMT2C* may play a role in the development and progression of HeH ALL.

ASE associated with somatic mutations

We next examined ASE involving mutated genes in HeH ALL by analyzing somatic SNVs from paired tumor/normal WGS data. Only coding region mutations (annotated as "silent" "missense" or "nonsense") were included. We observed 14 somatic mutation sites that were expressed across 12 genes. Among these, the genes *FLT3* and *WEE1* showed ASE in two HeH cases (Fig. 3d). *FLT3* overexpression is a known phenomenon in several leukemias, such as HeH ALL, irrespective of mutational status[24]. In this study, we found the HeH sample L5 that had a hemizygous deletion near the *FLT3* promoter region, displayed cis-dysregulation of *FLT3* expression and increased expression of the mutated *FLT3* allele as previously reported[25]. Overexpression of *WEE1*, a nuclear kinase that is involved in cell cycle G2-M checkpoint signaling, is associated with a poor outcome in various hematological malignancies and inhibition of WEE1 is considered as a potential therapeutic target in AML[26]. Here we identified a heterozygous G > T mutation in HeH sample L11, resulting in a premature stop codon and triggering nonsense-mediated decay in *WEE1*. In line with this, a significant allelic expression imbalance favoring the overexpression of the wild-type allele of *WEE1* was observed. However, the functional consequence of this nonsense mutation in HeH is unknown and further investigations of larger cohorts are needed to elucidate the function of *WEE1* mutations in HeH ALL. Collectively, these results suggest that ASE can also be associated with somatic mutations in HeH, in line with previous observations in other types of tumors[2].

Discussion

Detecting ASE is crucial for understanding the impact of cis-regulatory elements on RNA expression in tumors. Because ASE is assessed within the same cellular context, it naturally mitigates variability caused by experimental conditions such as sample preparation and sequencing depth. Consequently, ASE serves as an outstanding tool for analyzing allelic gene expression differences that arise from somatic aberrations and for investigating the pathogenic consequences of these aberrations in cancers. However, existing methods that solely rely on RNA-seq data for identifying ASE have certain limitations. For example, not using matched genomics data, such as asCN, may lead to incorrect biological interpretations of gene expression regulation caused by CNVs, especially in the context of aneuploidy, which is common in cancer. Population-based ASE analysis is another approach to ASE calling that is often considered robust for identifying ASE genes across large cohorts[27]. However, its utility in oncology is limited due to the heterogeneous and complex nature of genetic alterations in cancers, such as mutations, recombinations, and CNVs, which often surpass the complexity found in healthy tissues or less aggressive diseases. Since many oncogenic alterations are specific to individual samples, these population-based approaches may miss ASE events that are present in only a small number of cancer samples, even when those share the same cancer type.

Here, we integrated asCN data into ASE calling algorithms. Our aim was to discern asCN-derived ASE events by comparing asCN-aware ASE calling results to allelic balanced ASE calling results. This integration helps distinguish passenger events from driver events around recurrent CNVs. Additionally, the BASE pipeline incorporates an asCN calling approach that utilizes a comprehensive copy number analysis reference panel, enabling the estimation of asCN alterations using WGS data extracted from tumor samples, regardless of the availability of matched normal samples.

Using publicly available cell line data, we demonstrated that BASE is able to detect asCN with a low rate of false calls. However, it is important to note that BASE is not optimal to identify asCN in tumor samples with high levels of noise, such as highly heterogeneous samples or tumor samples with high normal cell contamination. The asCN calling performance in regions with LOH could be compromised by the heterozygosity and allelic imbalance ratio changes within those regions. Similar problems are commonly encountered in various CNV calling tools, and specialized methods are required to call CNVs in complex tumor samples[28].

Previous studies have shown that integrating phased variants from long-read sequencing and RNA-seq read counts can enhance the power to detect ASE[29,30]. However, due to the high requirements of DNA sample quality for long-read sequencing, this approach is not universally available for clinical samples. Moreover, precise haplotype phasing across entire chromosomes is challenging, especially in cancer cells with numerous CNVs. In our study, we investigated the ASE calling performance using heterozygous SNP data obtained from both unphased and phased WGS data from the HepG2 cell line. Our results indicated a similar accuracy for both types of data, underlining the robustness of the ASE calling workflow.

Furthermore, we demonstrated the credibility of ASE effects uncovered by BASE through the analysis of paired WGS and RNA-seq datasets from HeH ALL, one of the most common childhood malignancies, which generally shows a similar pattern of chromosomal gains in the majority of samples[8]. This makes HeH ALL a valuable model system for studying the effects of aneuploidy on ASE. Our analysis identified a general trend of ASE in relation to chromosomal gains in HeH ALL, with a significant number of genes on gained chromosomes exhibiting recurrent ASE events. Notably, chromosomes 14 and 21 showed stronger effects on ASE compared to other commonly gained chromosomes. This agrees well with previous findings of these two chromosomes being associated with particularly strong dosage effects in ALL, since the latter would be expected to arise from the homologue that is present in two copies in trisomic chromosomes, leading to ASE, and underscores their importance in the pathogenesis of HeH ALL[5,8].

In HeH ALL, non-CNV derived recurrent ASE events are rare. In some samples, ASE in the absence of a CNV can be attributed to mutations causing leukemia. For instance, our analysis observed the overexpression of a mutated *FLT3* allele in one HeH ALL through the recently described enhancer mechanism[25]. Furthermore, besides the well-known leukemia-related genes, our analysis also identified novel potential driver mutations, such as *WEE1*, associated with selective expression of mutated alleles in HeH ALL. However, given the small number of samples in this study, further investigations involving larger cohorts of HeH ALL are essential for determining novel tumor suppressor and oncogenes.

In summary, we have developed BASE, an automated ASE analysis framework employing combined genome and transcriptome sequencing data from human samples. BASE provides an asCN calling algorithm applying NGS data from tumor-only samples for which no normal samples are available by taking advantage of a built-in reference panel for CNV analysis. By integrating the asCN data into ASE calling algorithms, BASE provides a simple and effective way to interpret the molecular mechanisms underlying ASE and makes it possible to identify driver events that are caused by CNVs in cancer.

Data availability

The present study analyzes existing, publicly available data. The Cancer Cell Line Encyclopedia (CCLE) next generation sequencing data of 18 extensively studied cancer cell lines used in this study were obtained from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP186687. The WGS data generated of the HepG2 cell line are available via ENCODE (https://www.encodeproject.org/) under accession numbers ENCBS760ISV. HepG2 RNA-seq data can be accessed via ENCODE under accession numbers ENCLB707CLT and NCBI SRA under accession numbers SRR8616129 and SRR629573. WGS and RNA-seq data of the 12 HeH ALL samples analyzed in this study are available from the European Genome-phenome Archive (EGA; https://ega-archive.org/) under accession numbers EGAD00001008988 and EGAD00001002112, respectively. The code used to perform the analysis is available on Github (https://github.com/biomedswe/BASE).

Received: 21 March 2024; Accepted: 20 September 2024
Published online: 05 October 2024

References

1. Cleary, S. & Seoighe, C. Perspectives on allele-specific expression. *Annu. Rev. Biomed. Data Sci.* **4**, 101–122. https://doi.org/10.1146/annurev-biodatasci-021621-122219 (2021).
2. Bielski, C. M. et al. Widespread selection for oncogenic mutant allele imbalance in cancer. *Cancer Cell* **34**, 852-862 e854. https://doi.org/10.1016/j.ccell.2018.10.003 (2018).
3. Tsai, H. K. et al. Allelic complexity of KMT2A partial tandem duplications in acute myeloid leukemia and myelodysplastic syndromes. *Blood Adv.* **6**, 4236–4240. https://doi.org/10.1182/bloodadvances.2022007613 (2022).
4. Ben-David, U. & Amon, A. Context is everything: aneuploidy in cancer. *Nat. Rev. Genet* **21**, 44–62. https://doi.org/10.1038/s41576-019-0171-x (2020).
5. Yang, M. et al. Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Commun.* **10**, 1519. https://doi.org/10.1038/s41467-019-09469-3 (2019).

6. Mayba, O. et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* **15**, 405. https://doi.org/10.1186/s13059-014-0405-3 (2014).
7. Zambelli, F. et al. aScan: a novel method for the study of allele specific expression in single individuals. *J. Mol. Biol.* **433**, 166829. https://doi.org/10.1016/j.jmb.2021.166829 (2021).
8. Woodward, E. L. et al. Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukaemia. *Nat. Commun.* **14**, 1658. https://doi.org/10.1038/s41467-023-37356-5 (2023).
9. Sturm, M., Schroeder, C. & Bauer, P. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinform.* **17**, 208. https://doi.org/10.1186/s12859-016-1069-7 (2016).
10. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 (2009).
11. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034. https://doi.org/10.1093/bioinformatics/btv098 (2015).
12. Pei, S. et al. Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Brief Bioinform.* https://doi.org/10.1093/bib/bbaa148 (2021).
13. Cingolani, P. Variant annotation and functional prediction: SnpEff. *Methods Mol. Biol.* **2493**, 289–314. https://doi.org/10.1007/978-1-0716-2293-3_19 (2022).
14. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572. https://doi.org/10.1093/biostatistics/kxh008 (2004).
15. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21. https://doi.org/10.1093/bioinformatics/bts635 (2013).
16. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063. https://doi.org/10.1038/nmeth.3582 (2015).
17. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323. https://doi.org/10.1186/1471-2105-12-323 (2011).
18. Tange, O. *GNU parallel 2018*. (Lulu. com, 2018).
19. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607. https://doi.org/10.1038/nature11003 (2012).
20. Zhou, B. et al. Haplotype-resolved and integrated genome analysis of the cancer cell line HepG2. *Nucleic Acids Res* **47**, 3846–3861. https://doi.org/10.1093/nar/gkz169 (2019).
21. Lilljebjörn, H. et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 11790. https://doi.org/10.1038/ncomms11790 (2016).
22. Brady, S. W. et al. The genomic landscape of pediatric acute lymphoblastic leukemia. *Nat. Genet* **54**, 1376–1389. https://doi.org/10.1038/s41588-022-01159-z (2022).
23. Valentine, M. C. et al. Excess congenital non-synonymous variation in leukemia-associated genes in MLL- infant leukemia: a Children's Oncology Group report. *Leukemia* **28**, 1235–1241. https://doi.org/10.1038/leu.2013.367 (2014).
24. Poubel, C. P., Mansur, M. B., Boroni, M. & Emerenciano, M. FLT3 overexpression in acute leukaemias: New insights into the search for molecular mechanisms. *Biochim. Biophys. Acta Rev. Cancer* **80–88**, 2019. https://doi.org/10.1016/j.bbcan.2019.06.001 (1872).
25. Yang, M. et al. 13q12.2 deletions in acute lymphoblastic leukemia lead to upregulation of FLT3 through enhancer hijacking. *Blood* **136**, 946–956. https://doi.org/10.1182/blood.2019004684 (2020).
26. Vakili-Samiani, S. et al. Targeting Wee1 kinase as a therapeutic approach in Hematological Malignancies. *DNA Repair (Amst)* **107**, 103203. https://doi.org/10.1016/j.dnarep.2021.103203 (2021).
27. Fan, J. et al. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet* **16**, e1008786. https://doi.org/10.1371/journal.pgen.1008786 (2020).
28. Yu, Z., Li, A. & Wang, M. CLImAT-HET: detecting subclonal copy number alterations and loss of heterozygosity in heterogeneous tumor samples from whole-genome sequencing data. *BMC Med Genomics* **10**, 15. https://doi.org/10.1186/s12920-017-0255-4 (2017).
29. Deonovic, B., Wang, Y., Weirather, J., Wang, X. J. & Au, K. F. IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Res* **45**, e32. https://doi.org/10.1093/nar/gkw1076 (2017).
30. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* **7**, 12817. https://doi.org/10.1038/ncomms12817 (2016).

Acknowledgements

We used cancer cell lines copy number calling data generated by the Cancer Cell Line Encyclopedia (CCLE; https://sites.broadinstitute.org/ccle/). This study was supported by grants from the Swedish Childhood Cancer Fund, grant numbers PR2020-0033 (MY) and TJ2020-0024 (MY); the Crafoord Foundation, grant numbers 20230778 (MY) and 20240747 (MY).

Author contributions

JA, EA, RG, HL, TF, BJ, KP and MY, analyzed data. MY planned the study and wrote the manuscript with input from all authors.

Funding

Open access funding provided by Lund University. Crafoordska Stiftelsen (Grant no: 20240747 and 20230778); Barncancerfonden (Grant no: PR2020-0033 and TJ2020-0024).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73743-8.

Correspondence and requests for materials should be addressed to M.Y.