Long-read transcriptome assembly reveals vast isoform diversity in the placenta associated with metabolic and endocrine function

Sean T. Bresnahan[1*], Hannah Yong[2], William H. Wu[3], Sierra Lopez[4], Jerry Kok Yen Chan[5,6], Frédérique White[7], Pierre-Étienne Jacques[7], Marie-France Hivert[8,9], Shiao-Yng Chan[2,10], Michael I. Love[11], Jonathan Y. Huang[4,6,10‡], Arjun Bhattacharya[1,12‡]

[1]The University of Texas MD Anderson Cancer Center, Department of Epidemiology, Houston, TX
[2]A*STAR Institute for Human Development and Potential, Singapore, Singapore
[3]Rice University, Department of BioSciences, Houston, TX
[4]The University of Hawai‘i at Mānoa, Department of Public Health Sciences, Honolulu, HI
[5]KK Women's and Children's Hospital, Singapore, Singapore
[6]Duke-NUS Medical School, Singapore, Singapore
[7]Department of Biology, Sherbrooke University, Sherbrooke, QC, Canada
[8]Diabetes Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA
[9]Department of Medicine, Sherbrooke University, Sherbrooke, QC, Canada
[10]National University of Singapore, Yong Loo Lin School of Medicine, Department of Obstetrics & Gynecology, Singapore, Singapore
[11]University of North Carolina–Chapel Hill, Departments of Genetics & Biostatistics, Chapel Hill, NC
[12]The University of Texas MD Anderson Cancer Center, Institute for Data Science in Oncology, Houston, TX

[‡]Co-last authors
[*]Corresponding author: Sean T. Bresnahan, stbresnahan@mdanderson.org

Author Contributions

S.T.B., J.Y.H and A.B. designed the study. H.Y. and J.K.Y.C. prepared samples for sequencing. M.-F.H. managed the Gen3G study and S.-Y.C. managed the GUSTO study. S.T.B. performed all bioinformatic and statistical analyses and wrote the initial draft of the manuscript. All authors contributed to the writing and editing of the final manuscript.

Competing Interests Statement

The authors have no competing interests to disclose.

Classification

Research Reports

Keywords

Placenta, Maternal-Child Health, Transcriptomics, Long-Read RNA-sequencing, Oxford Nanopore, Transcript Isoforms, Inferential Uncertainty

This PDF file includes:

Main Text
Figures 1 to 4

Abstract

The placenta plays a critical role in fetal development and mediates maternal metabolic effects on offspring health outcomes. Despite its importance, the placenta remains understudied in large-scale genomic initiatives, with most analyses focusing on gene-level expression based on annotations from adult tissue references that obscure isoform diversity particularly vital to understanding function in developmental tissues. Here, we employed largest-in-class long-read RNA sequencing (N = 72) to generate a comprehensive placental transcript reference, yielding 37,661 high-confidence isoforms across 12,302 genes. Our assembly revealed 14,985 novel isoforms of known genes and 2,759 transcripts representing previously unannotated genes, with extensive diversity in genes involved in obesity, placental lactogen production, and growth control. We demonstrate this approach has two immediate practical advantages: First, the improved reference reduced inferential uncertainty in isoform quantification by approximately 37%, reduced the frequency of low-confidence annotations and increased the yield (read depth) of high-confidence annotations. Second, in analyses of gestational diabetes mellitus (GDM) with short-read RNA-seq datasets from two independent, multi-ancestry birth cohorts (N = 344) as an exemplar, we were able to identify novel isoforms of chorionic somatomammotropin hormone 1 (*CSH1*) mediating the effect of maternal hyperglycemia on birth weight that was not apparent in conventional gene-level analyses. These findings demonstrate that isoform-level annotation of placental transcriptomics provides greater sensitivity and specificity than gene-level approaches. This enhanced precision may be essential for understanding placental function in developmental programming of metabolism and reveal ancestry- and context-specific variation in placental function and responses to environmental exposures.

Introduction

The placenta, a transient organ unique to pregnancy, serves as the critical interface between maternal and fetal environments. Beyond its essential roles in nutrient transfer, gas exchange, and waste elimination, the placenta functions as a master regulator of the intrauterine environment through complex neuroendocrine and immunological signaling pathways that influence fetal development and programming of later-life health outcomes[1–9].

Gestational diabetes mellitus (GDM), characterized by glucose intolerance with onset during pregnancy, affects approximately 16% of pregnancies worldwide and significantly impacts both maternal and fetal health[9]. While decreasing insulin sensitivity is characteristic of physiological pregnancy, GDM is characterized by insulin insufficiency relative to the often exaggerated decline in maternal insulin sensitivity and glucose metabolism potentially through altered placenta-mediated endocrine signaling such as the pro-diabetogenic human placental lactogen (hPL) encoded by *CSH1*[10]. In turn, GDM-exposed fetuses experience a placenta-mediated alterations in signals that regulate the intrauterine environment promoting macrosomia, altered body composition with increased adiposity, and heightened risk for metabolic disorders later in life[11,12]. These GDM associations with later-in-life phenotypes are in line with the Developmental Origins of Health and Disease (DOHaD) hypothesis, which posits that the intrauterine environment shapes developmental trajectories with lasting implications for offspring health[13–17].

Limitations of short-read sequencing and current annotations restrict understanding of placental biology

Despite the placenta's central role in prenatal development, it has been notably underrepresented from large-scale tissue-specific genomic and transcriptomic initiatives such as the Genotype-Tissue Expression Project (GTEx), the NIH Roadmap Epigenomics Consortium, and the Encyclopedia of DNA Elements Consortium (ENCODE)[18–20]. These limited resources have restricted our understanding of placenta-specific gene regulation and transcript diversity relevant to fetal growth and metabolic programming.

The transcriptomes of developing tissues, including the placenta, exhibit distinct patterns of alternative splicing and isoform usage compared to adult tissues[20–23]. Consequently, utilizing adult tissue or species-level aggregate transcriptome references like GENCODE[24] for placental transcriptomics obscures tissue-specific regulatory mechanisms critical to understanding placental function. This limitation is compounded by the inherent constraints of short-read RNA sequencing approaches, which face significant challenges in accurately resolving transcript isoforms due to read-to-transcript mapping ambiguity[25]. When sequenced fragments are compatible with the exons of multiple isoforms—a common scenario for genes with complex splicing patterns—the probabilistic assignment of reads introduces significant uncertainty in estimating isoform-level expression[26]. This inferential uncertainty is particularly problematic for placental transcriptomics because the placenta exhibits uniquely high levels of alternative splicing and isoform diversity, especially in genes crucial for placental development and function[22].

Furthermore, analyses of existing placental RNA-sequencing datasets have focused overwhelmingly on aggregated gene-level measurements[5,14,21,23,27–34], an approach that misses critical biological information by obscuring transcript-isoforms that may be particularly susceptible to genetic or environmental effects relevant to complex traits[35–41]. This gene-centric view fails to capture scenarios where pathological exposures may be mediated through changes in specific transcript isoforms that could be masked, compensated for by expression of other, potentially non-functional isoforms from the same gene, or have opposing functions on the same biological process[35]. For example, previous studies that challenged the role of classical reproductive hormones in GDM, based on findings that circulating hPL levels do not correlate with insulin sensitivity changes during pregnancy[42], measured aggregate hormone levels without distinguishing between functionally distinct isoforms that may have differential effects on maternal metabolism.

Placental long-read sequencing to improve functionally relevant expression annotations

Long-read RNA sequencing technologies generate individual reads of up to 15-20 kilobases—sufficient to capture entire transcripts from end to end—providing direct evidence of exon combinations and revealing comprehensive transcript isoform diversity[43–47]. Recent studies demonstrate that over half of the isoforms detected using long-read RNA sequencing were missed by traditional short-read approaches in the same tissue samples[48–50], highlighting a significant gap in our current understanding of transcriptome complexity. Previous work, including from our group, has demonstrated that alternative splicing patterns and isoform usage explain substantially more variance in complex phenotypes than gene expression, with modeling of isoforms rather than genes increasing discovery of transcriptome-mediated genetic associations with complex traits by approximately 60%[35,37,51].

Here, we develop the first comprehensive placenta-specific transcriptome reference derived from N = 72 long-read RNA-sequencing libraries. Our analysis identified 37,661 high-confidence transcript isoforms representing 12,302 genes, including 448 novel genes and 14,957 novel isoforms of known genes, with pregnancy-related genes exhibiting extraordinary isoform diversity compared to other human tissues. We show that this assembly offers two complementary advantages: first, improved specificity and accuracy of isoform quantification in short-read datasets by reducing mapping ambiguity; and second, the ability to identify novel isoform-level regulatory mechanisms that may be critical for understanding GDM-associated modulation of fetal development but remain hidden in gene-level analyses. We then leveraged this placenta-specific transcriptome reference to study GDM and fetal growth as an exemplar application, demonstrating how improved isoform resolution can reveal novel molecular mechanisms underlying complex pregnancy outcomes. Our approach uncovered previously uncharacterized placenta-specific isoforms of genes involved in glucose metabolism, placental lactogen production and fetal growth regulation that mediate the effects of maternal hyperglycemia on birth weight. Importantly, these insights were undetectable in conventional gene-level transcriptomics analyses using standard annotations. These findings demonstrate that long-read sequencing-derived placental transcript isoform annotations are essential for progress in understanding placental pathophysiology and enhancing the detection of mechanistic signals in observational human studies examining environmental and other intrauterine exposures. Given the improved quantification precision and novel mechanistic insights demonstrated here, our placenta-specific reference provides a valuable resource for reanalyzing existing short-read placental transcriptomics datasets to uncover previously missed isoform-level regulatory mechanisms.

Results

Long-read sequencing reveals extensive transcript diversity in human placenta

In a subset of two pregnancy follow-up studies[28,52], we generated long-read transcriptome data from villous placental tissue of N = 72 term, live births without known placentation pathologies—such as intrauterine growth restriction and pre-eclampsia—including 50% from pregnancies affected by gestational diabetes mellitus (GDM) (Figure 1A). Using the Oxford Nanopore platform, we sequenced cDNA libraries to obtain approximately 310.8 million reads across all libraries. After stringent quality control and alignment to the human genome assembly (hg38, Methods), we retained approximately 72 million high-quality reads with an average read length of 1.07 kilobases (kb) (SD = 1.25 kb, range = 0.25–22.2 kb) (Table S1). These reads mapped to 68,002 distinct transcript isoforms with a mean length of 1.36 kb (SD = 1.67 kb, range = 0.06–205 kb) (Figure S1A) and an average of 4.5 exons per transcript (Figure S1B). These transcripts represented 23,069 genes, with an average of 2.95 isoforms per gene (SD = 3.84) (Figure S1C). Of these, our analysis captured 20,803 genes annotated in GENCODE v45 (which contains 63,187 genes) and identified 2,266 novel genes not previously annotated (Table S2) including fusion products of multiple genes (n = 102).

Nearly one-quarter of placental transcripts represent novel isoforms of known genes

Among transcripts annotated to known genes (n = 65,386 [96.15%]) we identified three major categories based on comparison to GENCODE v45: (1) transcripts matching known annotations, (2) novel isoforms of known genes, and (3) transcripts of novel genes (Figure 1B). Most transcripts of known genes matched existing annotations, classified as either full splice match (FSM: n = 45,082 [68.95%]), for which all splice junctions exactly matched a

reference transcript, or incomplete splice match (ISM: n = 5,176 [7.91%]), for which a subset of splice junctions matched a reference transcript (Figure 1C).

Notably, we identified 14,957 novel transcripts (22.87% of transcripts expressed in placenta) associated with 5,630 known genes (26.93% of genes expressed in placenta) (Figure 1B). These novel isoforms averaged 0.79 kb in length (SD = 0.47 kb, range = 0.08–4.4 kb) with a mean of 4.5 exons per transcript. The novel transcripts are categorized as: (1) "novel not in catalog" (NNC, n = 9,578 [63.9% of novel transcripts]), with at least one previously unannotated splice donor or acceptor site; (2) "novel in catalog" (NIC, n = 5,190, [34.6%]), combining known splice sites in previously unobserved patterns; or (3) "genic" transcripts (n = 217 [1.45%]), which partially mapped to 152 known genes but with insufficient splice junction matches to be classified as NIC or NNC. Additionally, we identified 2,759 transcripts representing novel genes, categorized as: antisense (n = 1,421, [51.5%]), intergenic (n = 1,006 [36.5%]), originating within introns of known genes (n = 189 [6.8%]), or fusion products spanning multiple genes (n = 143 [5.2%]) (Figure 1C).

Orthogonal data sources validate high-quality placental transcript models

For downstream analyses, we filtered out low-quality isoforms using criteria derived from previous long-read transcriptome assembly studies in other tissues (Methods)[36,45–49] (Figure 1C). We used orthogonal short-read RNA-seq, DNase I hypersensitivity site sequencing (DNase-seq), Cap Analysis of Gene Expression sequencing (CAGE-seq), and sequencing of alternative polyadenylation sites (SAPAS) data to validate splice junction coverage and TSS/TES. Transcript models in our assembly showed robust support from sequencing data, with an average of 554.23 long-reads (Nanopore; SD = 3.89×10^4) (Figure S2A), and an average minimum splice junction support of 2.33×10^4 short-reads (Illumina; SD = 2.47×10^5) (Figure S2B). We observed strong enrichment of TSS near CAGE-seq or DNase-seq peaks (median distance = 90 bp with 23,740 [34.9%] of transcripts located within 50 bp) (Figure S2C). Similarly, TTS showed enrichment near alternative polyadenylation sites (Figure S2D) or PolyA motifs (Figure S2E) (median distance = 15 bp with 26,549 [40.5%] of transcripts located within 5 bp). Both TSS and TTS were also frequently located proximally to their corresponding annotated positions in the reference transcripts (Table S2).

Our quality filtering identified relatively few problematic transcripts: n = 2,243 isoforms (3.3%) showed evidence of nonsense mediated decay (NMD), n = 483 (0.7%) exhibited patterns consistent with RT-switching artifacts, and n = 4 (0.005%) showed evidence of PolyA intrapriming (Figure S2F). After removing 30,341 low-quality isoforms (44.62%; Table S3), our placenta-specific transcript assembly contained 37,661 high-confidence transcript isoforms (Figure 1C) with a mean length of 1.60 kb (SD = 1.52 kb, range = 0.08–34.63 kb; Figure S3A) and an average of 5.62 exons per transcript (Figure S3B). Unless explicitly stated, all subsequent analyses reported here refer to this filtered high confidence set, which represent 12,302 genes (mean = 3.06, SD = 3.43 isoforms per gene), including 11,854 genes in the GENCODE v45 reference and 448 novel genes (Figure S3C).

Compared to the currently available long-read defined transcriptomes generated by GTEx[47] v9 for 11 adult tissues, the placenta demonstrates substantially higher isoform complexity (Figure 1D). GTEx tissues averaged 1.55 transcripts per gene (SD = 1.06, range: 1–21), while the placenta shows nearly twice this level of transcript diversity (mean = 3.06, SD = 3.43, range: 1–108) (Figure 1E). Leveraging tissue-specific expression data from the Human Protein Atlas database[53], genes corresponding to the most abundant transcripts in our long-read data were significantly enriched for genes expressed in the placenta and not in other tissues across multiple abundance thresholds (Table S4; Figure S3D). For example, among the 500 most abundant placental long-read transcripts, we observed a 4.3-fold enrichment for placenta-specific genes (adjusted $p = 7.85 \times 10^{-7}$). This result suggests that our sequencing successfully captured genes with established placenta-restricted expression patterns.

Placenta-specific annotations significantly reduce quantification uncertainty in short-read datasets

To evaluate the impact of our placenta-specific transcript annotations on short-read based transcript quantification, we compared quantification results using our high-confidence long-read transcript models versus GENCODE v45 annotations. We analyzed Illumina cDNA libraries from villous placental tissue collected in two independent birth cohort studies: the GUSTO Study[28] (n = 192) and the Genetics of Glucose regulation in Gestation and Growth

(Gen3G) Study[29] (n = 152), both with approximately 20% of samples from GDM-affected pregnancies (Figure 1F; Table S5). Following quality control, we removed two GUSTO samples due to low read quality (Table S5) and two additional GUSTO samples as statistical outliers (Figure S4A).

Our placenta-specific assembly demonstrated significantly improved quantification specificity, with more reads being attributed to fewer, presumably more tissue-relevant transcripts (Figure 1G). This resulted in a 5,426 TPM increase in total quantified expression when using the long-read assembly (13.3% increase; $p < 2 \times 10^{-16}$; Table S6), reflecting less diluted isoform quantification. Transcripts that failed to meet our stringent quality criteria—either undetected or filtered—showed significantly lower mean expression and variance in short-read data compared to FSM and novel transcripts retained in our assembly (Figure 1H, Figure S4B). Statistical analysis revealed a clear hierarchy in transcript expression levels and variance (Table S7): shared (FSM) transcripts showed the highest expression, followed by novel transcripts, then filtered transcripts, and finally absent transcripts. In GUSTO, shared transcripts had 1.16-1.71 log-fold higher expression than other categories (all $p < 1 \times 10^{-4}$), with expression variance showing similar patterns (1.98-2.85 log-fold differences, all $p < 1 \times 10^{-4}$). Importantly, novel transcripts showed expression levels only slightly lower than shared transcripts, confirming that our assembly captures functionally relevant novel isoforms rather than low-quality artifacts.

To formally quantify read-to-transcript mapping ambiguity, for each sample across both cohorts, we calculated transcript-level inferential relative variance (InfRV). As InfRV is roughly stabilized with respect to the mean expression level, higher InfRV values indicate greater uncertainty in estimating transcript abundance (Methods)[54]. Mean transcript-level InfRV was consistently reduced when quantified with our placenta-specific annotations compared to GENCODE v45 (38.1% reduced in GUSTO and 35.9% reduced in the Gen3G) (Figure S4C-D), indicating improved isoform-level quantification precision with our tissue-specific reference. When stratified by transcript structural category, ISM transcripts exhibited considerably lower mean InfRV than FSM, NIC and NNC, likely because ISM transcripts share fewer exons with other transcripts of the same gene, reducing read assignment ambiguity (Figure 1I; Figure S4E). Similarly, transcripts in "Other" structural categories showed remarkably low mean InfRV but demonstrated the weakest between-sample correlation coefficients, suggesting highly sample-specific expression (Figure S4F-G).

Novel transcripts show distinct structural characteristics compared to annotated transcripts

Compared to transcripts of known genes, transcripts of novel genes were significantly less abundant in the long-reads (Mann-Whitney-Wilcoxon test: $W = 1.07 \times 10^{7}$, $p < 2.2 \times 10^{-16}$) (Figure S5A), shorter in length ($W = 1.31 \times 10^{7}$, $p < 2.2 \times 10^{-16}$) (Figure S5B), and contained fewer exons ($W = 1.43 \times 10^{7}$, $p < 2.2 \times 10^{-16}$) (Figure S5C). Similarly, novel transcript isoforms of annotated genes were significantly less abundant than FSM transcripts ($W = 1.25 \times 10^{8}$, $p < 2.2 \times 10^{-16}$) and shorter in length ($W = 1.58 \times 10^{8}$, $p < 2.2 \times 10^{-16}$), although they contained comparable numbers of exons ($W = 1.15 \times 10^{8}$, $p = 0.99$). These characteristics suggest novel placenta-specific genes are recently evolved considering gene length and transcript complexity follow clear directional trends toward increasing structural elaboration over evolutionary time[55].

Pregnancy-related genes exhibit extraordinary isoform diversity

We identified substantial diversity in the number of multi-exonic RNA isoforms per gene (range: 1–108) with over half of all detected genes (n = 7,343 [59.69%]) expressing more than one isoform, and a smaller subset characterized by exceptional isoform diversity with more than ten distinct isoforms (n = 418 [3.39%]) (Figure 2A). The number of detected isoforms significantly correlated with the number of detected exons per isoform ($r = 0.87$, $p < 2.2 \times 10^{-16}$) (Figure S5D), and the strength of this correlation was stronger among highly expressed genes ($r = 0.93$, $p < 2.2 \times 10^{-16}$), likely reflecting greater sensitivity for detecting splicing events from abundantly expressed genes.

Notably, genes with critical functions in pregnancy and placental development exhibited the highest isoform diversity (Figure 2B). Chorionic somatomammotropin hormone 1 (*CSH1*), which regulates human placental lactogen production in pregnancy[56], displayed the greatest isoform diversity with 108 distinct transcripts. This was followed by pregnancy-specific glycoproteins *PSG5* (56 isoforms) and *PSG6* (48 isoforms), and the CSH-like

gene *CSHL1* (50 isoforms). Similarly, growth hormone 2 (*GH2*), which encodes placental growth hormone, also showed expanded isoform diversity with 22 distinct transcripts. Both *CSH1* and *GH2* belong to the growth hormone/chorionic somatomammotropin gene family clustered on chromosome 17, and their protein products play coordinated roles in maternal metabolic adaptation to pregnancy and fetal growth regulation. The extensive isoform diversity across this gene family suggests complex transcriptional regulation of placental endocrine functions, supported by significant enrichment for "hormone receptor binding" (odds ratio = 79.3, adjusted p = 1.29×10^{-3}) for the 50 most isoform-rich genes through Gene Ontology (GO) analysis (Table S8).

Novel transcript isoforms show varied coding potential and encode potentially functional proteins

We evaluated the coding potential and protein-level validation of detected transcripts. Open reading frames (ORFs) were predicted for 19,670 (52.2%) isoforms (Figure S6A), with similar coding sequence (CDS) length across structural categories (median = 1.11 kb) (Figure S6B). The coding potential of these predicted ORFs, as measured by CPC2 coding probability scores[57], varied by structural category: FSM, NIC and NNC transcripts showed similarly high coding probability (median = 85.39%) while ISM transcripts (median = 66.02%) and other structural categories (median = 57.19%) displayed lower coding potential (Figure S6C).

Using BlastP[58] to assess homology to human peptide sequences identified by tandem mass spectrometry (MS/MS) of placental tissue samples, including from GDM-affected pregnancies[59,60], novel isoforms of known genes (NIC: 56.71%; NNC: 57.98%) showed the highest peptide validation rates, surprisingly exceeding FSM transcripts (51.21%) (Figure 2C; Table S9). ISM transcripts showed a striking tissue-specificity: 32.29% homology to placental peptides versus only 13% to the full human proteome, indicating 2.5-fold enrichment for placenta-specialized protein isoforms (Figure S6D). The higher peptide validation rate for novel, compared to annotated, isoforms demonstrates that our assembly captures functionally important but previously uncharacterized proteins in the placenta, and incomplete models may still encode functional proteins with placenta-specific roles, potentially including variants relevant to GDM pathophysiology.

We identified 80 novel gene transcripts with predicted ORFs (17.62% of novel genes), with 44 showing MS/MS validation and encoding functionally relevant proteins including fibronectin (a potential GDM biomarker)[61], E-cadherin (regulating trophoblastic differentiation)[62], alpha-enolase (5 distinct isoforms, differentially expressed in GDM)[63], occludin (placental barrier integrity)[64], and CCL14 (trophoblast migration and maternal-fetal communication)[65]. The protein-level validation demonstrates that our assembly captures functionally important but previously uncharacterized proteins relevant to placental adaptation in GDM.

Placenta-specific annotations reveal more extensive GDM-associated transcriptional changes

We next examined placental transcriptional differences associated with GDM status in GUSTO and Gen3G, comparing analyses using our tissue-specific assembly versus GENCODE v45 for differential transcript and gene expression, adjusting for gestational age, fetal sex, and residual variation estimated with RUVr[66] (Figure 3A; Figure S7A-B; Methods). Using FDR < 10% thresholds, our placenta-specific annotations identified an average of 138 differentially expressed transcripts between cohorts (SD = 159.8) and 792.5 differentially expressed genes (SD = 955.3), while GENCODE v45 annotations identified 333.5 differentially expressed transcripts (SD = 99.7) and 934.5 differentially expressed genes (SD = 839.3). Despite identifying fewer total features, analyses using our placenta-specific reference resulted in identification of a higher percentage of annotated features with GDM-associated differences compared to analyses using GENCODE v45 (Tables S10-11). This enrichment for differential expression, combined with the reduced inferential variance we observed across transcripts, suggests that our tissue-specific annotation approach concentrates the analysis on biologically relevant transcripts while filtering out noise from irrelevant or poorly quantified features.

Placenta-specific annotations redistribute short-read assignments to dominant and novel isoforms

To better understand how differences in transcript annotations affect differential expression results, we tracked changes in effect sizes in GUSTO between quantifications on GENCODE v45 and our placenta-specific annotations. We first examined *CSH1*, which exhibited the most dramatic expansion in isoform number in our

placenta-specific annotation. For the FSM isoforms of *CSH1*, one isoform (ENST00000558284.1) showed significant differential expression associated with GDM when quantified against GENCODE but lost this association when quantified against our placenta-specific annotations (Figure 3B). While this isoform showed moderate expression in both annotations, it had notably higher expression when quantified against GENCODE (Figure 3C), owing to a redistribution of reads to ENST00000558284.1 or to novel isoforms when the data is quantified with our annotation (Figure 3D). Conversely, *GNAS*—a known imprinted gene in placenta[67]—which experienced one of the most dramatic contractions in isoform number (77 in GENCODE reduced to 17 in our placenta-specific annotation), showed the opposite pattern. Isoform ENST00000476196.5 showed a significant GDM association when quantified against our placenta-specific annotations but not against GENCODE (Figure 3E), with increased expression in our annotations (Figure 3F-G). Together, these findings illustrate how adding unannotated transcripts and removing irrelevant transcripts from a transcriptome annotation results in the redistribution of reads among isoforms of a gene and to other genes. This redistribution is likely beneficial, as it reduces noise by reassigning reads from spurious annotations, improving both quantification accuracy and the detection of biologically relevant differential expression patterns.

Placental transcription mediates significant portion of GDM effects on birth weight

We performed pathway enrichment analysis on differentially expressed transcripts (FDR < 10%) in GUSTO, recognizing that different annotation approaches may capture distinct but potentially valid biological signals (Table S12). Our placenta-specific annotations identified pathways enriched for endocrine and signaling functions: "hormone receptor binding" (odds ratio = 101.89, adjusted $p = 3.20 \times 10^{-3}$), "growth factor activity" (odds ratio = 48.42, adjusted $p = 2.49 \times 10^{-3}$), and "receptor ligand activity" (odds ratio = 18.81, adjusted $p = 2.49 \times 10^{-3}$) (Figure S7C). GENCODE annotations revealed pathways related to structural processes: "PI3K-Akt signaling pathway" (odds ratio = 8.19, adjusted $p = 7.42 \times 10^{-7}$) and "focal adhesion" (odds ratio = 10.16, adjusted $p = 4.23 \times 10^{-6}$) (Figure S7D). To determine which approach better captures functionally relevant changes that mechanistically link GDM to birth weight outcomes, we performed mediation analysis.

Given the known impacts of GDM on birth weight, in GUSTO, we examined whether some of these effects are mediated by placental transcription using multi-mediator analysis (Figure 4). To address the high-dimensional nature of transcript expression data while maintaining statistical power, we performed principal component analysis on differentially expressed transcripts (FDR < 10%) and selected the number of principal components that cumulatively explained at least 75% of the total variance in transcript expression as mediators[68,69]. We selected this threshold based on the balance between capturing sufficient variance and maintaining interpretability. To address potential confounding by population structure, we included maternal ethnicity as a covariate in addition to adjusting for fetal sex and gestational age. We confirmed that our component selection was optimal through sensitivity analysis (Figure 4B; Figure S8; Table S13).

In GUSTO, GDM was significantly associated with increased birth weight (total effect = 0.430, 95% CI [0.131, 0.729], p = 0.005). Importantly, the indirect effect of GDM on birth weight through placental transcript expression when quantifying short-reads against our placenta-specific transcript annotations was statistically significant (0.155, 95% CI [0.016, 0.293], p = 0.029), with approximately 35.8% of the total effect being mediated through placental transcriptomic changes (Figure 4A). Consistent with the pathway enrichment results, our tissue-specific approach identified mediators with clear placental endocrine functions (*CSH1*, *CDC25C*), aligning with the hormone and growth factor pathways detected in the enrichment analysis. In contrast, when analyzing the same samples using GENCODE v45 annotations, the association between GDM and birth weight was not statistically significant (0.095, 95% CI [-0.057, 0.247], p = 0.219), and GENCODE-based analysis produced more heterogeneous mediators including several transcripts of uncharacterized genes. This finding demonstrates that our focused annotation approach not only detects biologically relevant pathway enrichments but also better captures mechanistically relevant transcript changes that mediate clinically important outcomes, highlighting how the intrauterine environment affects fetal development through altered placental transcription.

Novel CSH1 isoforms emerge as key mediators of GDM effects on fetal growth

We further identified specific transcript isoforms that mediate the effect of GDM on birth weight (Table S14). In GUSTO, 11 transcripts showed significant mediation effects ($p < 0.05$), including *ESPRESSO:chr17:12985:70* (Figure 4E), a novel isoform of *CSH1* not annotated in GENCODE (0.021, 95% CI [0.001, 0.048], $p < 2 \times 10^{-16}$), accounting for approximately 13.4% of the effect of placental transcription (Figure 4C). Similarly, in Gen3G, 19 transcripts showed a significant effect, including *ESPRESSO:chr17:12985:62* (Figure 4E), a different novel isoform of *CSH1* (0.216, 95% CI [0.071, 0.389], $p < 2 \times 10^{-16}$), accounting for approximately 24.37% of the effect of placental transcription (total effect of GDM on birth weight = 0.132, indirect effect of placental transcription = 0.117) (Figure 4D). The identification of different *CSH1* isoforms mediating GDM effects across two independent cohorts illustrates the plasticity of alternative splicing in response to genetic and environmental factors. This population-specific variation was unique to *CSH1*: GENCODE-annotated transcripts showed no significant cross-cohort differences in expression patterns, indicating that the observed isoform differences reflect genuine alternative splicing rather than systematic artifacts. This highlights the potential for developing population-specific biomarkers for pregnancy complications like GDM.

Discussion

Current placental studies involving transcriptome evaluations rely on gene-level expression measurements, obscuring critical transcript-level regulatory mechanisms and missing key placental functions. This approach fails to capture scenarios where adverse outcomes arising from pathological exposures alter functionally distinct transcript isoforms—effects that could be masked by compensatory transcriptional responses that maintain total gene expression levels. Our findings provide empirical evidence for the benefits of long-read RNA-seq-based transcript-resolved placental analysis by identifying specific placental transcript isoforms—particularly novel *CSH1* transcripts—that mediate the relationship between maternal gestational diabetes mellitus (GDM) and newborn birth weight, effects that would be undetectable in conventional gene-level analyses. This transcript-level precision provides novel molecular evidence by demonstrating how the placenta actively shapes fetal development through alternative splicing that fine-tunes endocrine functions and placental signaling in response to the intrauterine environment. By integrating long- and short-read RNA sequencing data across three diverse birth cohorts, our work shifts the field from descriptive placental transcriptomics toward more effective mechanistic understanding, setting the foundation for developing isoform-resolved biomarkers and therapeutic targets to monitor and modulate placental contributions to early-life disease risk.

Despite the placenta's critical role in fetal programming and later-life health outcomes[1–7], it has been notably absent from large-scale genomic initiatives[18–20]. Comparing to other long-read defined transcriptomes across adult tissues in GTEx[47] v9, the placenta demonstrates substantially higher isoform complexity and a nearly 2 times more transcript diversity. This finding fundamentally challenges previous short-read based characterizations of the placenta as having a "transcriptomic void" with the seventh least complex transcriptome among GTEx tissues[21,23,70]. Gong *et al*. concluded that the placental transcriptome was skewed toward a small number of highly abundant transcripts, suggesting this reflected the placenta's nature as a rapidly growing and transient organ where many transcripts are "dispensable." However, our long-read assembly reveals that this apparent simplicity likely reflects the technical limitations of short-read sequencing in resolving complex alternative splicing patterns rather than true biological simplicity. The substantially higher isoform diversity we observe may be facilitated by the placenta's unique DNA methylation landscape, characterized by global hypomethylation with organized partially methylated domains (PMDs) interspersed with highly methylated regions and relatively higher gene body methylation compared to other tissues[71,72] – a pattern that is conserved across placental mammals[73]. This bimodal distribution may create spatially distinct regulatory environments where genes with PMDs experience reduced methylation-mediated splicing repression while genes in highly methylated domains undergo enhanced exon inclusion[74]. Our findings suggest that the placenta employs sophisticated transcript-level regulation that has been systematically underestimated, calling into question conclusions about placental transcriptional diversity and highlighting the necessity of tissue-specific long-read references for accurately characterizing placental transcriptome complexity.

A key biological insight enabled by our placenta-specific transcriptome reference is the improved resolution with which we can detect transcriptomic alterations associated with GDM. First, the remarkable complexity concentrated in key regulators of placental endocrine function, including *CSH1*, *CSH1L*, *PSG5*, *PSG6*, and *GH2*, suggests that isoform-level modulation provides flexibility in hormone signaling to both the mother and fetus. Second, by reducing inferential uncertainty[25,26] in short-read RNA-seq data—particularly for full splice match transcripts—we uncovered a more complete and concordant landscape of differential transcript expression and usage. This enhanced precision revealed GDM-linked regulatory changes that would have been missed using conventional annotations. In the context of placental biology, where transcriptional responses are tightly constrained, the ability to detect subtle but functionally significant isoform shifts may be critical for identifying early biomarkers of maternal-fetal maladaptation and in designing mitigating interventions. These findings underscore the importance of using tissue-specific transcriptomic references in developmental contexts[20–22] and illustrate how isoform-resolved analyses can uncover biological insights missed by gene-level approaches[35–41].

Pathway enrichment analyses using our placenta-specific annotations revealed GDM-associated transcripts enriched for receptor-ligand activity, hormone receptor binding and growth factor signaling, consistent with key endocrine and paracrine roles of the placenta in orchestrating maternal metabolic adaptations and fetal growth trajectories[8,9]. In contrast, analyses using standard GENCODE annotations highlighted cellular processes like RNA processing and chromatin regulation. To empirically test which pathway findings better capture functionally relevant biology, we performed mediation analysis and found that our placenta-specific annotations identified statistically significant mediation with 35.8% of the GDM-birth weight effect mediated through placental transcription, while GENCODE-based analysis failed to detect significant mediation. This empirical validation demonstrates that tissue-specific annotations not only improve pathway interpretability[36,37] but also better capture mechanistically relevant transcript changes that mediate clinically important outcomes.

Notably, our approach revealed population-specific variation in novel *CSH1* isoforms mediating GDM effects across cohorts. *CSH1* encodes human placental lactogen (hPL), a key endocrine regulator of maternal insulin resistance and nutrient partitioning[10]. The novel CSH1 isoforms contain intron retention events (Figure 4E), a form of alternative splicing increasingly recognized as an important regulatory mechanism of tissue-specific gene expression[75,76], particularly during cellular differentiation[77] and in response to cellular stress[78] including aging-induced transposable element expression[79,80,81]. These findings reinforce the DOHaD hypothesis by identifying transcript-level mediators of maternal-fetal metabolic coupling that reveal population-specific molecular mechanisms linking environmental exposures to gene regulation. Additionally, they provide a potential resolution to previous findings that circulating hPL levels do not correlate with insulin sensitivity changes during pregnancy[42]—studies which measured aggregate hormone levels without accounting for functionally distinct isoforms that may have differential metabolic effects. These transcript-level regulatory mechanisms that gene-level approaches miss may enable population-tailored biomarkers for placental dysfunction[82,83].

While our long-read sequencing approach provides unprecedented resolution of placental transcript diversity, the relatively modest sequencing depth limits detection of very low-abundance isoforms that may have important regulatory roles. Our long-read reference cohort (N = 72) represents the largest placental long-read RNA sequencing dataset to date and successfully enabled transcript discovery across two independent validation cohorts with different population ancestries (East/Southeast Asian and White). However, our findings of population-specific differential transcript expression and distinct *CSH1* isoforms mediating GDM effects in each cohort suggest that transcript diversity may vary across populations. Thus, expanding long-read sequencing to additional populations may reveal novel, population-specific transcript isoforms not captured in our current reference. Additionally, experimental validation is necessary to confirm that the novel protein-coding genes identified here are actively translated and to determine their contributions to placental function. Furthermore, functional follow-up studies of genes like CSH1 require precise isoform-level resolution, as experimental interventions like transcriptional knockdown or overexpression depend on targeting the exact isoform mediating the biological effect.

This work opens new research avenues across multiple domains. Most importantly, it enables more precise investigation of environmental exposure effects by identifying specific mechanistic pathways that mediate exposure-outcome relationships. Traditional gene-level approaches suffer from confounding by multiple,

potentially opposing regulatory mechanisms within the same gene, obscuring causal inference about environmental effects. By resolving transcript-level mediators, our approach enables identification of less confounded mechanistic pathways that can support causal inference methods—complementing Mendelian randomization approaches by focusing on intermediate biological mechanisms rather than genetic instruments. This mechanistic precision is particularly valuable for studying environmental exposures like maternal diabetes, where effects may operate through distinct transcript isoforms with opposing functions. Importantly, our tissue-specific assembly reduces the overall transcript annotation set while enriching for biologically relevant isoforms, facilitating fine-mapping of quantitative trait loci (QTLs) and enabling more precise identification of causal "effect isoforms" at specific loci for targeted functional validation studies. Furthermore, our demonstration of population-specific transcript mediators establishes a framework for investigating differential environmental exposure effects by ancestry and context—enabling identification of population-tailored intervention targets and biomarkers. Beyond exposure studies, this resource provides a foundation for isoform-resolved association studies of developmental traits and birth outcomes, enabling precise mapping of regulatory variants influencing fetal growth and early-life disease risk[5,36–38,84,85]. It will enhance comparative evolutionary transcriptomics by improving identification of lineage-specific innovations—such as transposon exonization[86,87] and transposon-driven alternative promoters[88,89]—that have shaped placental evolution[33,80,81,90]. Additionally, it will support detailed investigations of imprinting and parent-of-origin effects[91], which are enriched in placental tissues and contribute substantially to pregnancy outcome variation[67,92–94].

This long-read defined placental transcriptome assembly fills a major gap in understanding the human placenta by enabling isoform-resolved quantification and revealing novel transcripts that mediate maternal GDM effects on fetal growth. Our findings advance the mechanistic basis for how in utero exposures shape long-term metabolic health, providing a foundation for discovering early biomarkers and therapeutic targets for pregnancy complications.

Materials and Methods

Biological samples and sequencing

We included a subset of two mother-offspring cohort studies[28,52] of term, live births (N = 72) without known placentation pathologies—such as intrauterine growth restriction and pre-eclampsia—and 50% from pregnancies affected by gestational diabetes mellitus (GDM; categorized by a 75 g two-timepoint oral glucose tolerance test using the WHO 1999 criteria which was in clinical use at that time). Placenta was processed within 1 hr of delivery. Five villous biopsies of the placenta were rinsed with phosphate buffer saline (PBS), snap-frozen with liquid nitrogen and stored at −80 °C before DNA extraction. Approximately 75 mg of crushed placenta tissue was combined with 500 µL Buffer ATL (Qiagen, Cat #: 939011) and 30 µL Proteinase K (from Qiagen QIASymphony DSP DNA Midi Kit) and incubated overnight in a shaking incubator. Extracted samples were sequenced via 3x 24-plex on PromethION (release 23.04.6, Oxford Nanopore, ONT) with manufacturer's PCR-cDNA library prep kit (#SQK-PCB111.24) at the Genome Institute of Singapore using manufacturer's protocols. Median read lengths were approximately 1,000 bp (vs. 150 bp via NovoSeq) and all QC and RNA quality metrics were met.

Paired-end Illumina libraries of villous placenta from term, live births from the Growing Up in Singapore Toward healthy Outcomes[28,95] (GUSTO, N = 200;) and Genetics of Glucose Regulation in Gestation and Growth[29,96] (Gen3G, N = 152; phs003151) birth cohorts were sequenced previously. Briefly, placental villous tissue samples were collected after delivery and stored at -80 °C. In total RNA extracted using phenol-chloroform followed by purification with the NucleoSpin RNA kit (Macherey-Nagel). In GUSTO, RNA quality was assessed by Nanodrop spectrophotometry and Agilent TapeStation (RIN > 6 required), with ribosomal RNA depletion performed using Illumina Ribo-Zero kits. Sequencing libraries were prepared using NEB Next Ultra Directional RNA Library Prep Kit and sequenced on Illumina HiSeq 4000 platform to a minimum of 50 M paired-end 150 bp reads per library. In Gen3G, RNA quality was assessed using an Agilent Bioanalyzer (RIN > 5 required) and sent to the Broad Institute for sequencing. Libraries were prepared from 250 ng of RNA using Illumina TruSeq Stranded mRNA Sample Preparation Kits and sequenced on an Illumina HiSeq 4000 platform to a minimum of 14 M paired-end 100-bp reads per library.

Transcriptome assembly

FAST5 files of the raw ONT signals for each sample were processed with Guppy v.6.5.7 for base calling on a GPU with barcode adapter trimming enabled. FASTQ files labeled "pass" were concatenated together for each sample. Pychopper2 was used to identify, orient and trim full-length cDNA reads containing primer sequences used in library construction with the PCS111 sequencing kit (ONT). Full length and rescued cDNA reads were concatenated for each sample, and microindels and sequencing errors were corrected with FMLRC2[97] using short-read RNA-seq samples of villous placenta from GUSTO (N = 200)[28]. The corrected reads were aligned to the human genome assembly analysis set (GenBank assembly 000001405.15, no ALT contigs, hg38) using Minimap2[98] with a kmer length of 14, minimizer window size of 4, and in splice-aware mode with --splice-flank=no to relax assumptions about bases adjacent to splice junction donors and acceptors. Secondary alignments were filtered, and primary alignments were coordinate sorted with samtools[99]. The filtered and sorted alignments to chromosomes 1-22, X, Y and M for all samples were then concatenated together to generate a single BAM file used for de novo transcript assembly with ESPRESSO, which has shown favorable balance in maximizing discovery while limiting false discoveries[49].

Transcriptome annotation and filtering

The assembled transcript models were filtered using a custom R[100] script to remove isoforms with un-stranded exons (retaining only "+" or "-" strand isoforms), isoforms with an exon less than 3 nucleotides long, and isoforms with exons positions outside the transcript start and end positions. We then used SQANTI3[101] to characterize the *de novo* assembly and classify isoforms into structural categories relative to GENCODE v45 transcripts: full splice matches (FSM), incomplete splice matches (ISM), novel in catalog (NIC), novel not in catalog (NNC), and other classes (e.g. fusion products and antisense transcripts). Open reading frames and coding sequences were predicted with TransDecoder[102]. Coding potential was calculated with CPC2[57]. For coding sequences of novel genes, we used BlastP[58] to search for homology to peptide sequences identified by tandem mass spectrometry (MS/MS) of placental tissue samples, including from GDM-affected pregnancies[59,60]. Tissue-specific gene enrichment was assessed with TissueEnrich[53].

For quality control of the isoform calls, we considered long-read coverage of the assembled transcripts quantified with ESPRESSO, in addition to splice junction coverage in the short-read RNA-seq samples, proximity of transcription start sites (TSS) from Cap Analysis of Gene Expression sequencing (CAGE-seq) peaks, genome-wide sequencing of regions sensitive to cleavage by DNase I (DNase-seq) peaks or reference TSS, and proximity of transcription termination sites (TTS) from sequencing alternative polyadenylation (polyA) sites (SAPAS), polyA motifs or reference TTS. We defined high-quality isoforms across all structural categories as those with a minimum of 3 long (Nanopore) reads covering the transcript and at least 3 short (Illumina) reads covering each splice junction. We validated transcription start sites (TSS) by requiring them to be within 5 kb of a CAGE-seq peak in the FANTOM5 placenta dataset[103], within 5 kb of a DNase-seq peak in the ENCODE placenta dataset[19], or within 100 bp of a reference TSS. Similarly, transcription termination sites (TTS) were required to be within 100 bp of a SAPAS site in the APASdb placenta dataset[104], within 100 bp of a reference TTS, or within 40 bp of a PolyA motif in the PolyA_DB dataset[105]. For all categories except FSM, we required transcripts to have less than 80% adenosine content in the 20 nucleotides at the 3' end to eliminate artifacts of PolyA intrapriming[101]. For novel isoform categories (excluding FSM, ISM and NIC), we applied additional filters: removing transcripts with evidence of reverse transcriptase (RT) switching at splice junctions, and excluding transcripts predicted to undergo nonsense-mediated decay (NMD) unless they showed substantial 5' expression evidence (defined as a ratio of > 0.5 between short-reads mapping to the first and last splice junctions).

To assess pairwise sample correlations of the transcript abundances, we used Oarfish[106] to quantify the transcripts for each long-read RNA-seq library including 50 bootstrap replicates. Transcript quantifications and bootstrap replicates were loaded into R, and overdispersion arising from read-to-transcript ambiguity (RTA) was estimated and divided from the counts[25] using custom R scripts. Pearson correlation coefficients were calculated for each pair of libraries on the transcripts-per-million normalized RTA-corrected counts.

Preparation of orthogonal datasets for assembly filtering

Short-read RNA-seq reads from term, villous placenta from GUSTO (N = 200) were aligned to hg38 using STAR[107] with settings recommended by ENCODE. Splice junctions with < 10 supporting reads were filtered. CAGE-seq reads from placenta tissue (N = 1) and cells (pericyte, N = 2; epithelial, N = 4, trophoblast, N = 1) were retrieved from the FANTOM5 Project[103], trimmed with cutadapt v4.1[108] to remove linker adapters, EcoP15 and 5'poly-G sequences, and filtered reads were aligned to h38 using STAR with settings recommended by ENCODE. CAGE tags were counted and clustered using CAGEr[109] in R. DNAse-seq peaks from placenta tissue (N = 22) mapped to h38 were retrieved in BED format from the ENCODE consortium and intersected with bedtools[110]. SAPAS peaks in placenta tissue mapped to hg38 were retrieved in BED format from APASdb[104], and human polyA motifs annotated in PolyA_DB[105] were retrieved from PolyA-miner[111].

Short-read RNA-seq analyses

Paired-end Illumina libraries of villous placenta from term, live births of the GUSTO (N = 200) and Gen3G were sequenced previously. We used fastp[112] to trim sequencing adapters and filter low-quality reads. Decoy-aware transcriptome indices for Salmon[113] for both the GENCODE v45 annotations and our placenta-specific annotations were constructed with MashMap[114]. Transcriptome quantification was then performed with Salmon with settings --libType A --validateMappings --seqBias and including 50 bootstrap replicates to estimate:

$$E_{i,j}^{(k)}$$

the expression E of transcript T_j in sample S_i for B_k bootstrap replicates, where j = 1,…,n transcripts in the given assembly, i = 1,…,n short-read samples, and k = 1,…,50 replicates. Transcript quantifications were loaded into R with tximport[115], and overdispersion arising from read-to-transcript ambiguity was estimated with catchSalmon() from the edgeR package[25] and divided from the counts[25] prior to downstream analyses. To ensure reliable detection and reduce noise in downstream analyses, we filtered the dataset to retain only transcripts with counts of at least 5 in a minimum of 3 samples.

To test for differences in total transcript expression (TPMs) between assemblies, we used linear mixed-effects models with assembly and cohort as fixed effects and sample as a random effect. To examine differences in mean and variance properties between assemblies and transcript sets, we performed nested ANOVAs using log-transformed mean or variance in expression as the outcome variable, with cohort-by-transcript-set and cohort-by-assembly-set interaction terms. Post-hoc pairwise comparisons between transcript sets within each cohort were conducted using estimated marginal means with Tukey adjustment for multiple comparisons.

Differential transcript and gene expression analyses of GDM status were performed with DESeq2[116]. Models for all analyses adjusted for gestational age (scaled), fetal sex and 2 dimensions of residual variation estimated with RUVr[66]. Isoform-level mediation analyses[117] of the GDM-associated differentially expressed transcripts at FDR <10% were performed with the mediation package[118] using variance stabilizing transformation (VST) normalized counts to estimate the indirect effect of GDM on birth weight (scaled). Multi-mediator analyses were performed with the lavaan package[119] with the number of principal components[68] that cumulatively explained at least 75% of the total variance in the VST-normalized counts of the GDM-associated transcripts (DTEs), adjusting for gestational age and fetal sex to estimate the total effect of GDM and the indirect effect of placental transcription on birth weight. To assess potential confounding by population structure, we used one-way ANOVA to test for associations between each PC and maternal ethnicity, calculating both p-values and eta-squared effect sizes using the effectsize package[120] to quantify the strength of associations.

Transcript-level Inferential relative variance (InfRV) across median-ratio scaled bootstrap replicates[54] was calculated as

$$\mathrm{InfRV} = 0.01 + \frac{\max\left(s^2 - \mu, 0\right)}{\mu + 5}$$

where s^2 is the sample variance of scaled counts over bootstrap replicates, and μ is the mean scaled count over bootstrap replicates, with a pseudocount of 5 added to the denominator and 0.01 added overall to stabilize the statistic and make it strictly positive for log transformation. To visualize read mapping transitions between assemblies for N = 10 randomly selected samples from GUSTO (with 50% from GDM affected pregnancies), we aligned the short-reads with BWA-MEM2[121] to the transcript sequences in each assembly with settings "-O 12,12 -E 4,4" to prevent gapped alignments. End-to-end alignments to the transcript sequences were then filtered with a custom Perl script to retain only the first mate in the read pair and to remove secondary and supplementary alignments.

Ethics Approval

The Singaporean studies follow guidelines set forth in the Declaration of Helsinki and all study activities were approved by the National Healthcare Group Domain Specific Review Board and the SingHealth Centralized Institutional Review Board (GUSTO only). Informed written consent was obtained from all mothers at recruitment. The ethical committee from the Centre Hospitalier Universitaire de Sherbrooke (CHUS) reviewed and approved Gen3G study protocols.

Acknowledgements

We thank members of both the Bhattacharya and Huang labs for critical reading of our manuscript, and members of the Love lab for feedback on methods.

Data Availability

We have made our high-confidence placenta-specific transcript annotations publicly available at https://github.com/sbresnahan/lr-placenta-transcriptome.

Funding

The GUSTO study is supported by the National Research Foundation (NRF) under the Open Fund-Large Collaborative Grant (OF-LCG; MOH-000504) administered by the Singapore Ministry of Health's National Medical Research Council (NMRC) and the Agency for Science, Technology and Research (A*STAR). In RIE2025, GUSTO is supported by funding from the NRF's Human Health and Potential (HHP) Domain, under the Human Potential Programme. Placental RNA sequencing was supported by an NMRC Open Fund-Young Individual Research Grant awarded to JYH (MOH-000550-00). JYH was also supported by an A*STAR Human Health and Potential–Prenatal/Early Childhood Grant (H24P2M0002) and Pilot Projects Program funding from the NIMHD RCMI-CC (U24MD015970). The Gen3G study was initially supported by a Fonds de recherche du Québec—Santé operating grant (to M-.F.H., grant no. 20697), Canadian Institute of Health Research (CIHR) operating grants (to M.-F.H., grant no. MOP 115071) and a Diabète Québec grant.

References

1. McKay, R. Remarkable role for the placenta. *Nature* **472**, 298–299 (2011).
2. Baron-Cohen, S. *et al.* Foetal oestrogens and autism. *Mol. Psychiatry* **25**, 2970–2978 (2020).
3. Thornburg, K. L., O'Tierney, P. F. & Louey, S. Review: The Placenta is a Programming Agent for Cardiovascular Disease. *Placenta* **31**, S54–S59 (2010).
4. Ursini, G. *et al.* Convergence of placenta biology and genetic risk for schizophrenia. *Nat. Med.* **24**, 792–801 (2018).
5. Peng, S. *et al.* Genetic regulation of the placental transcriptome underlies birth weight and risk of childhood obesity. *PLOS Genet.* **14**, e1007799 (2018).
6. Tedner, S. G., Örtqvist, A. K. & Almqvist, C. Fetal growth and risk of childhood asthma and allergic disease. *Clin. Exp. Allergy* **42**, 1430–1447 (2012).
7. Bronson, S. L. & Bale, T. L. The Placenta as a Mediator of Stress Effects on Neurodevelopmental Reprogramming. *Neuropsychopharmacology* **41**, 207–218 (2016).
8. Stern, C. *et al.* Placental Endocrine Activity: Adaptation and Disruption of Maternal Glucose Metabolism in Pregnancy and the Influence of Fetal Sex. *Int. J. Mol. Sci.* **22**, 12722 (2021).
9. Olmos-Ortiz, A. *et al.* Immunoendocrine Dysregulation during Gestational Diabetes Mellitus: The Central Role of the Placenta. *Int. J. Mol. Sci.* **22**, 8087 (2021).
10. Jin, Y., McNicol, I. & Cattini, P. A. A locus control region generates distinct active placental lactogen and inactive growth hormone gene domains in term placenta that are disrupted with obesity. *Placenta* **164**, 64–72 (2025).
11. Lowe, L. P. *et al.* Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study. *Diabetes Care* **35**, 574–580 (2012).
12. Page, K. A. *et al.* Children Exposed to Maternal Obesity or Gestational Diabetes Mellitus During Early Fetal Development Have Hypothalamic Alterations That Predict Future Weight Gain. *Diabetes Care* **42**, 1473–1480 (2019).
13. Lacagnina, S. The Developmental Origins of Health and Disease (DOHaD). *Am. J. Lifestyle Med.* **14**, 47–50 (2020).
14. Liu, Z. *et al.* Large-scale chromatin reorganization reactivates placenta-specific genes that drive cellular aging. *Dev. Cell* **57**, 1347-1368.e12 (2022).
15. Rowan, J. A. *et al.* Metformin in gestational diabetes: the offspring follow-up (MiG TOFU): body composition and metabolic outcomes at 7–9 years of age. *BMJ Open Diabetes Res. Care* **6**, e000456 (2018).
16. Crume, T. L. *et al.* Association of exposure to diabetes in utero with adiposity and fat distribution in a multiethnic population of youth: the Exploring Perinatal Outcomes among Children (EPOCH) Study. *Diabetologia* **54**, 87–92 (2011).
17. Pinney, S. E. *et al.* Exposure to Gestational Diabetes Enriches Immune-Related Pathways in the Transcriptome and Methylome of Human Amniocytes. *J. Clin. Endocrinol. Metab.* **105**, 3250–3264 (2020).
18. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
19. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
20. The GTEx Consortium. Atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
21. Gong, S. *et al.* The human placenta exhibits a unique transcriptomic void. *Cell Rep.* **42**, 112800 (2023).
22. Ruano, C. S. M. *et al.* Alternative splicing in normal and pathological human placentas is correlated to genetic variants. *Hum. Genet.* **140**, 827–848 (2021).
23. Gong, S. *et al.* The RNA landscape of the human placenta in health and disease. *Nat. Commun.* **12**, 2639 (2021).
24. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
25. Baldoni, P. L. *et al.* Dividing out quantification uncertainty allows efficient assessment of differential transcript expression with edgeR. *Nucleic Acids Res.* **52**, e13–e13 (2024).

26. Singh, N. P., Wu, E. Y., Fan, J., Love, M. I. & Patro, R. Tree-based differential testing using inferential uncertainty for RNA-Seq. Preprint at https://doi.org/10.1101/2023.12.25.573288 (2023).
27. Santos Jr, H. P. *et al.* Evidence for the placenta-brain axis: multi-omic kernel aggregation predicts intellectual and social impairment in children born extremely preterm. *Mol. Autism* **11**, 97 (2020).
28. Soh, S.-E. *et al.* Insights from the Growing Up in Singapore Towards Healthy Outcomes (GUSTO) Cohort Study. *Ann. Nutr. Metab.* **64**, 218–225 (2014).
29. Guillemette, L. *et al.* Genetics of Glucose regulation in Gestation and Growth (Gen3G): a prospective prebirth cohort of mother–child pairs in Sherbrooke, Canada. *BMJ Open* **6**, e010031 (2016).
30. Paquette, A. *et al.* A genome scale transcriptional regulatory model of the human placenta. *Sci. Adv.* (2024).
31. Tekola-Ayele, F. *et al.* Placental multi-omics integration identifies candidate functional genes for birthweight. *Nat. Commun.* **13**, 2384 (2022).
32. Saben, J. *et al.* A comprehensive analysis of the human placenta transcriptome. *Placenta* **35**, 125–131 (2014).
33. Armstrong, D. L. *et al.* The core transcriptome of mammalian placentas and the divergence of expression with placental shape. *Placenta* **57**, 71–78 (2017).
34. Baker, B. H. *et al.* Placental transcriptomic signatures of prenatal and preconceptional maternal stress. *Mol. Psychiatry* **29**, 1179–1191 (2024).
35. Bhattacharya, A. *et al.* Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Nat. Genet.* **55**, 2117–2128 (2023).
36. Aguzzoli Heberle, B. *et al.* Mapping medically relevant RNA isoform diversity in the aged human frontal cortex with deep long-read RNA-seq. *Nat. Biotechnol.* **43**, 635–646 (2025).
37. Wen, C. *et al.* Cross-ancestry atlas of gene, isoform, and splicing regulation in the developing human brain. *Science* **384**, eadh0829 (2024).
38. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018).
39. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).
40. Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.* **54**, 1355–1363 (2022).
41. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
42. Kirwan, J. P. *et al.* TNF-alpha Is a Predictor of Insulin Resistance in Human Pregnancy.
43. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
44. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
45. Leung, S. K. *et al.* Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep.* **37**, 110022 (2021).
46. Paniagua, A. *et al.* Evaluation of strategies for evidence-driven genome annotation using long-read RNA-seq. *Genome Res.* **35**, 1053–1064 (2025).
47. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. (2023).
48. Veiga, D. F. T. *et al.* A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci. Adv.* **8**, eabg6711 (2022).
49. Gao, Y. *et al.* ESPRESSO: Robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. *Sci. Adv.* **9**, eabq5072 (2023).
50. Han, S. W., Jewell, S., Thomas-Tikhonenko, A. & Barash, Y. Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads. *Genome Res.* **34**, 1624–1635 (2024).
51. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
52. Yong, H. E. J. *et al.* Increasing maternal age associates with lower placental CPT1B mRNA expression and acylcarnitines, particularly in overweight women. *Front. Physiol.* **14**, 1166827 (2023).
53. Jain, A. & Tuteja, G. TissueEnrich: Tissue-specific gene enrichment analysis. *Bioinformatics* **35**, 1966–1967 (2019).
54. Zhu, A., Srivastava, A., Ibrahim, J. G., Patro, R. & Love, M. I. Nonparametric expression analysis using inferential replicate counts. *Nucleic Acids Res.* **47**, e105–e105 (2019).

55. Gorlova, O., Fedorov, A., Logothetis, C., Amos, C. & Gorlov, I. Genes with a large intronic burden show greater evolutionary conservation on the protein level. *BMC Evol. Biol.* **14**, 50 (2014).
56. Männik, J. *et al.* Differential Expression Profile of *Growth Hormone/Chorionic Somatomammotropin* Genes in Placenta of Small- and Large-for-Gestational-Age Newborns. *J. Clin. Endocrinol. Metab.* **95**, 2433–2442 (2010).
57. Kang, Y.-J. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).
58. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
59. Lin, Z. *et al.* Alternative Strategy To Explore Missing Proteins with Low Molecular Weight. *J. Proteome Res.* **18**, 4180–4188 (2019).
60. Wei, Y., He, A., Huang, Z., Liu, J. & Li, R. Placental and plasma early predictive biomarkers for gestational diabetes mellitus. *PROTEOMICS – Clin. Appl.* **16**, 2200001 (2022).
61. Rasanen, J. P. *et al.* Glycosylated Fibronectin as a First-Trimester Biomarker for Prediction of Gestational Diabetes: *Obstet. Gynecol.* **122**, 586–594 (2013).
62. Li, H. W., Cheung, A. N. Y., Tsao, S. W. & Cheung, A. L. M. Expression of E-Cadherin and Beta-Catenin in Trophoblastic Tissue in Normal and Pathological Pregnancies. *Int J Gynecol Pathol* **22**, (2003).
63. Lin, J. *et al.* Proteomic analysis of plasma total exosomes and placenta-derived exosomes in patients with gestational diabetes mellitus in the first and second trimesters. *BMC Pregnancy Childbirth* **24**, 713 (2024).
64. Villota, S. D., Toledo-Rodriguez, M. & Leach, L. Compromised barrier integrity of human feto-placental vessels from gestational diabetic pregnancies is related to downregulation of occludin expression. *Diabetologia* **64**, 195–210 (2021).
65. Hannan, N. J., Jones, R. L., White, C. A. & Salamonsen, L. A. The Chemokines, CX3CL1, CCL14, and CCL4, Promote Human Trophoblast Migration at the Feto-Maternal Interface1. *Biol. Reprod.* **74**, 896–904 (2006).
66. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
67. Hanna, C. W. Placental imprinting: Emerging mechanisms and functions. *PLOS Genet.* **16**, e1008709 (2020).
68. Zeng, P., Shao, Z. & Zhou, X. Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. *Comput. Struct. Biotechnol. J.* **19**, 3209–3224 (2021).
69. Zhao, Y., Lindquist, M. A. & Caffo, B. S. Sparse principal component based high-dimensional mediation analysis. *Comput. Stat. Data Anal.* **142**, 106835 (2020).
70. Gonzalez, T. L. *et al.* High-throughput mRNA-seq atlas of human placenta shows vast transcriptome remodeling from first to third trimester. *Biol. Reprod.* **110**, 936–949 (2024).
71. Yuan, V. *et al.* Cell-specific characterization of the placental methylome. *BMC Genomics* **22**, 6 (2021).
72. Schroeder, D. I. *et al.* The human placenta methylome. *Proc. Natl. Acad. Sci.* **110**, 6037–6042 (2013).
73. Schroeder, D. I. *et al.* Early Developmental and Evolutionary Origins of Gene Body DNA Methylation Patterns in Mammalian Placentas. *PLOS Genet.* **11**, e1005442 (2015).
74. Lev Maor, G., Yearim, A. & Ast, G. The alternative role of DNA methylation in splicing regulation. *Trends Genet.* **31**, 274–280 (2015).
75. Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene expression programs. *Hum. Genet.* **136**, 1043–1057 (2017).
76. Alharbi, A. B. *et al. Ctcf* haploinsufficiency mediates intron retention in a tissue-specific manner. *RNA Biol.* **18**, 93–103 (2021).
77. Ullrich, S. & Guigó, R. Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. *Nucleic Acids Res.* **48**, 1327–1340 (2020).
78. Hadar, S., Meller, A., Saida, N. & Shalgi, R. Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention. *iScience* **25**, 105543 (2022).
79. Pabis, K. *et al.* A concerted increase in readthrough and intron retention drives transposon expression during aging and senescence. *eLife* **12**, RP87811 (2024).
80. Keighley, L. M., Lynch-Sutherland, C. F., Almomani, S. N., Eccles, M. R. & Macaulay, E. C. Unveiling the hidden players: The crucial role of transposable elements in the placenta and their potential contribution to pre-eclampsia. *Placenta* **141**, 57–64 (2023).
81. Frost, J. M. *et al.* Regulation of human trophoblast gene expression by endogenous retroviruses. *Nat. Struct. Mol. Biol.* **30**, 527–538 (2023).

82. Li, J. *et al.* Isoform usage as a distinct regulatory layer driving nutrient-responsive metabolic adaptation. *Cell Metab.* **37**, 772-787.e6 (2025).
83. Verta, J.-P. & Jacobs, A. The role of alternative splicing in adaptation and evolution. *Trends Ecol. Evol.* **37**, 299–308 (2022).
84. Bhattacharya, A. *et al.* Placental genomics mediates genetic associations with complex health traits and disease. *Nat. Commun.* **13**, 706 (2022).
85. Peng, S. *et al.* Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases. *Hum. Mol. Genet.* **26**, 3432–3441 (2017).
86. Pasquesi, G. I. M. *et al.* Regulation of human interferon signaling by transposon exonization. *Cell* **187**, 7621-7636.e19 (2024).
87. Arribas, Y. A. *et al.* Transposable element exonization generates a reservoir of evolving and functional protein isoforms. *Cell* **187**, 7603-7620.e22 (2024).
88. Miao, B. *et al.* Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol.* **21**, 255 (2020).
89. Lee, H. J. *et al.* Epigenomic analysis reveals prevalent contribution of transposable elements to *cis* -regulatory elements, tissue-specific expression, and alternative promoters in zebrafish. *Genome Res.* **32**, 1424–1436 (2022).
90. Lynch, V. J. *et al.* Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy. *Cell Rep.* **10**, 551–561 (2015).
91. Gardner, A. & Úbeda, F. The meaning of intragenomic conflict. *Nat. Ecol. Evol.* **1**, 1807–1815 (2017).
92. Soliman, H. K. & Coughlan, J. M. United by conflict: Convergent signatures of parental conflict in angiosperms and placental mammals. *J. Hered.* **115**, 625–642 (2024).
93. Kobayashi, E. H. *et al.* Genomic imprinting in human placentation. *Reprod. Med. Biol.* **21**, e12490 (2022).
94. Vincenz, C., Lovett, J. L., Wu, W., Shedden, K. & Strassmann, B. I. Loss of Imprinting in Human Placentas Is Widespread, Coordinated, and Predicts Birth Phenotypes. *Mol. Biol. Evol.* **37**, 429–441 (2020).
95. Fitzgerald, E. *et al.* Hofbauer cell function in the term placenta associates with adult cardiovascular and depressive outcomes. *Nat. Commun.* **14**, 7120 (2023).
96. Hivert, M.-F. *et al.* Placental IGFBP1 levels during early pregnancy and the risk of insulin resistance and gestational diabetes. *Nat. Med.* **30**, 1689–1695 (2024).
97. Mak, Q. X. C. Polishing De Novo Nanopore Assemblies of Bacteria and Eukaryotes With FMLRC2.
98. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
99. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
100. R Core Team. R: A language and environment for statisticla computing. (2024).
101. Pardo-Palacios, F. J. *et al.* SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat. Methods* **21**, 793–797 (2024).
102. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
103. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
104. You, L. *et al.* APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* **43**, D59–D67 (2015).
105. Zhang, H. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33**, D116–D120 (2004).
106. Jousheghani, Z. Z. & Patro, R. Oarfish: Enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. Preprint at https://doi.org/10.1101/2024.02.28.582591 (2024).
107. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
108. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12.
109. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51–e51 (2015).
110. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

111. Yalamanchili, H. K. *et al.* PolyA-miner: accurate assessment of differential alternative poly-adenylation from 3′Seq data using vector projections and non-negative matrix factorization. *Nucleic Acids Res.* **48**, e69–e69 (2020).
112. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
113. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
114. Kille, B., Garrison, E., Treangen, T. J. & Phillippy, A. M. Minmers are a generalization of minimizers that enable unbiased local Jaccard estimation. *Bioinformatics* **39**, btad512 (2023).
115. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016).
116. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
117. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **15**, 309–334 (2010).
118. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. **mediation** : *R* Package for Causal Mediation Analysis. *J. Stat. Softw.* **59**, (2014).
119. Rosseel, Y. **lavaan** : An *R* Package for Structural Equation Modeling. *J. Stat. Softw.* **48**, (2012).
120. Ben-Shachar, M., Lüdecke, D. & Makowski, D. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *J. Open Source Softw.* **5**, 2815 (2020).
121. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, Rio de Janeiro, Brazil, 2019). doi:10.1109/IPDPS.2019.00041.

Figures

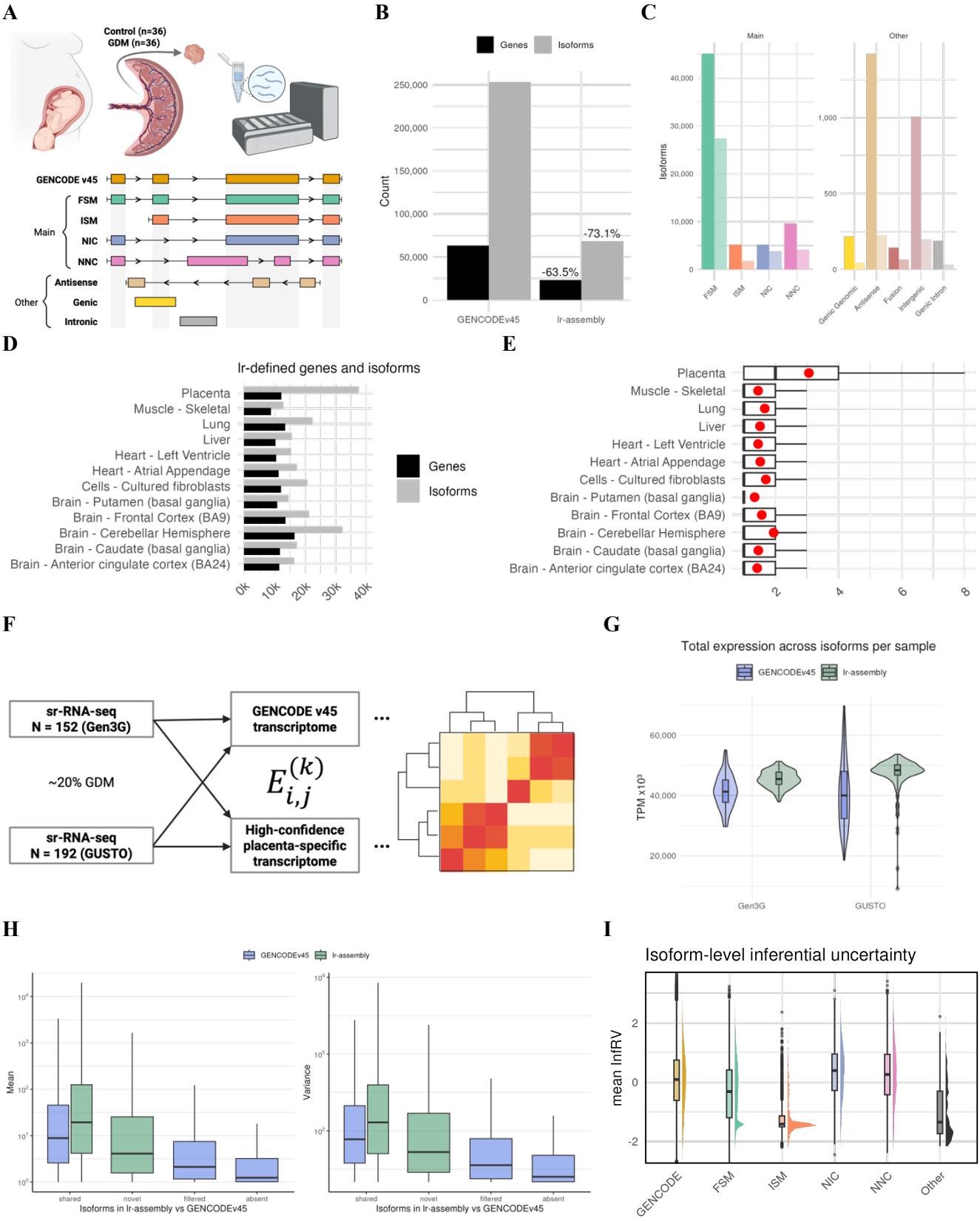

A
Control (n=36)
GDM (n=36)
GENCODE v45
FSM
ISM
NIC
NNC
Main
Antisense
Genic
Intronic
Other
B
Genes
Isoforms
Count
250,000
200,000
150,000
100,000
50,000
0
-63.5%
-73.1%
GENCODEv45
Ir-assembly
C
Main
Other
Isoforms
40,000
30,000
20,000
10,000
0
1,000
500
FSM
ISM
NIC
NNC
Genic Genomic
Antisense
Fusion
Intergenic
Genic Intron
D
Ir-defined genes and isoforms
Placenta
Muscle - Skeletal
Lung
Liver
Heart - Left Ventricle
Heart - Atrial Appendage
Cells - Cultured fibroblasts
Brain - Putamen (basal ganglia)
Brain - Frontal Cortex (BA9)
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Brain - Anterior cingulate cortex (BA24)
Genes
Isoforms
0k
10k
20k
30k
40k
E
Placenta
Muscle - Skeletal
Lung
Liver
Heart - Left Ventricle
Heart - Atrial Appendage
Cells - Cultured fibroblasts
Brain - Putamen (basal ganglia)
Brain - Frontal Cortex (BA9)
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Brain - Anterior cingulate cortex (BA24)
2
4
6
8
F
sr-RNA-seq
N = 152 (Gen3G)
~20% GDM
sr-RNA-seq
N = 192 (GUSTO)
GENCODE v45
transcriptome
High-confidence
placenta-specific
transcriptome
$E_{i,j}^{(k)}$
G
Total expression across isoforms per sample
GENCODEv45
Ir-assembly
TPM x10³
60,000
40,000
20,000
Gen3G
GUSTO
H
GENCODEv45
Ir-assembly
Mean
Variance
shared
novel
filtered
absent
Isoforms in Ir-assembly vs GENCODEv45
I
Isoform-level inferential uncertainty
mean InfRV
2
0
-2
GENCODE
FSM
ISM
NIC
NNC
Other

Figure 1 – A long-read assembly of the placenta transcriptome reveals extensive splicing diversity and improves short-read based transcriptome quantification. (**A**) Nanopore cDNA libraries were prepared from villous placenta tissue of term, live births from mothers with and without gestational diabetes mellitus (GDM). Transcripts were assembled and compared to the GENCODE v45 annotations to classify isoforms into structural categories: full splice matches (FSM), incomplete splice matches (ISM), novel in catalog (NIC), novel not in catalog (NNC), and other classes (fusion products and intergenic, genic, intronic and antisense transcripts). (**B**) Genes and transcripts in the GENCODE v45 reference compared to the full placenta-specific annotation. (**C**) Transcripts in the placenta-specific reference stratified by structural category. Opaque bars represent the full set of transcripts, while transparent bars indicate the high-confidence transcripts. (**D**) Count of long-read defined genes and transcript isoforms in placenta compared to 11 tissues in GTEx[47] v9. (**E**) Transcripts per gene across tissues. Red points indicate the mean. (**F**) Illumina short-read RNA-seq of villous placental tissue—from two independent birth cohort studies with approximately 20% of samples from GDM-affected pregnancies—were quantified against the GENCODE v45 transcript annotations and the high-confidence placenta-specific transcript annotations. Using Salmon[113], for each cohort and annotation, the expression E of transcript T_j in sample S_i was estimated for B_k bootstrap replicates, where $j = 1,\ldots,n$ transcripts in the given annotation, $i = 1,\ldots,n$ short-read samples, and $k = 1,\ldots,50$ replicates. Bootstrap replicates were used to estimate overdispersion arising from read-to-transcript mapping ambiguity, which was divided from the counts prior to downstream analyses[25], and to quantify inferential relative variance (InfRV), a measure of uncertainty in the read-to-transcript mapping[54]. (**G**) The total transcript expression (measured as transcripts per million [TPM]) per sample in each cohort when quantified against the GENCODE v45 annotations or the high-confidence placenta-specific annotations. (**H**) The mean and variance in the GUSTO short-read samples comparing transcripts that were shared between annotations (i.e., FSM transcripts in placenta), transcripts that were novel in placenta, GENCODE v45 transcripts that were initially recovered in the long-readss but removed in the filtering procedure, and GENCODE v45 transcripts that were not assembled from the long-readss. (**I**) Transcript-level InfRV in the GUSTO short-read samples when quantified against the GENCODE v45 annotations compared to the placenta-specific annotations, stratified by transcript structural category.

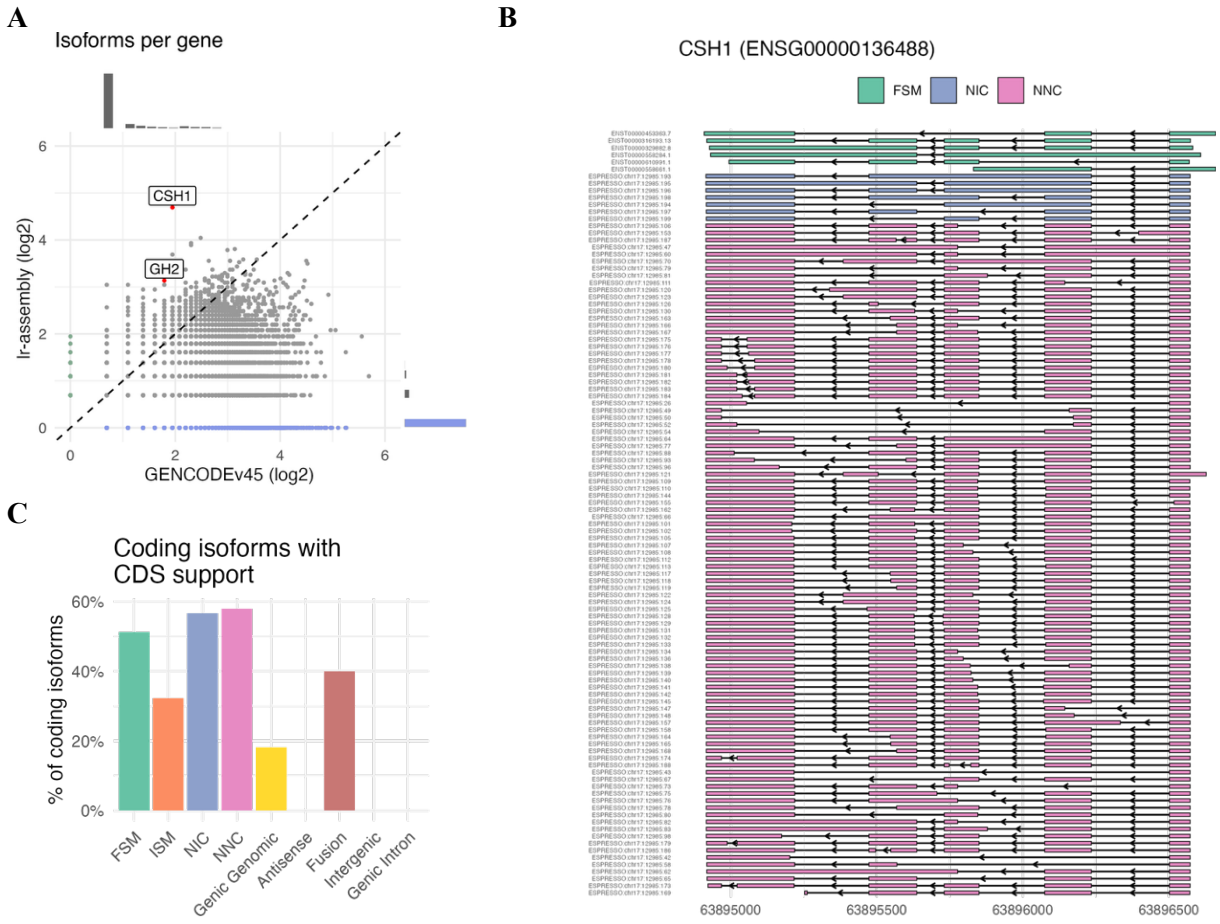
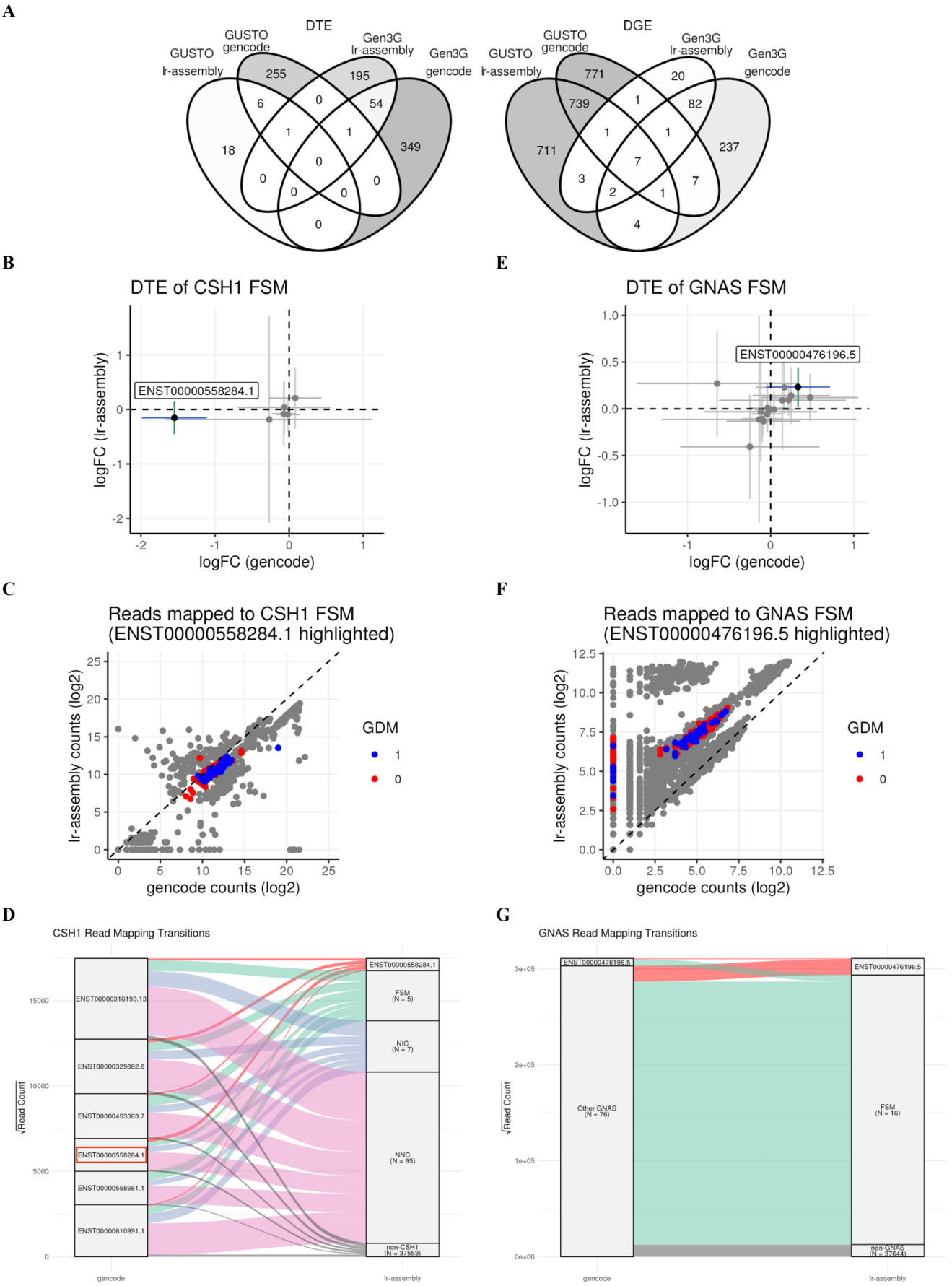


Figure 2 – **Pregnancy-related genes show extraordinary splicing diversity.** (**A**) Transcript isoforms per gene in the GENCODE v45 reference compared to the high-confidence placenta-specific reference. Colored bars in the marginal histograms and points indicate GENCODE v45 genes with no isoforms represented in the placenta-specific reference (blue) and novel placenta-specific genes not represented in the GENCODE v45 reference (green). (**B**) High-confidence transcripts of *CSH1* expressed in the placenta. (**C**) Transcripts with predicted open reading frames (ORFs) with orthogonal support for coding sequences (CDS).


A
DTE
GUSTO lr-assembly
GUSTO gencode
Gen3G lr-assembly
Gen3G gencode
255
195
6
0
54
1
1
18
349
0
0
0
0
0
0
DGE
GUSTO lr-assembly
GUSTO gencode
Gen3G lr-assembly
Gen3G gencode
771
20
739
1
82
1
1
711
237
7
3
2
7
1
4
B
DTE of CSH1 FSM
ENST00000558284.1
logFC (lr-assembly)
logFC (gencode)
E
DTE of GNAS FSM
ENST00000476196.5
logFC (lr-assembly)
logFC (gencode)
C
Reads mapped to CSH1 FSM
(ENST00000558284.1 highlighted)
lr-assembly counts (log2)
gencode counts (log2)
GDM
1
0
F
Reads mapped to GNAS FSM
(ENST00000476196.5 highlighted)
lr-assembly counts (log2)
gencode counts (log2)
GDM
1
0
D
CSH1 Read Mapping Transitions
ENST00000316193.13
ENST00000329882.8
ENST00000453363.7
ENST00000558284.1
ENST00000558661.1
ENST00000610991.1
√Read Count
gencode
lr-assembly
ENST00000558284.1
FSM (N = 5)
NIC (N = 7)
NNC (N = 95)
non-CSH1 (N = 37553)
G
GNAS Read Mapping Transitions
ENST00000476196.5
Other GNAS (N = 76)
√Read Count
gencode
lr-assembly
ENST00000476196.5
FSM (N = 16)
non-GNAS (N = 37644)

34

Figure 3 – Tissue-specific transcriptome assembly yields dramatic differences in transcript- and gene-level associations with GDM. (A) Venn diagrams of the overlapping significant associations between cohorts and assemblies. (**B**) Differential transcript expression (DTE) of *CSH1* FSM transcripts in GUSTO. Vertical (green) lines indicate the 95% CI of the $\log_2$ fold change when quantified with GENCODE v45 (gencode), while horizontal (blue) lines indicate the 95% CI when quantified on the placenta-specific reference (lr-assembly). (**C**) Raw read counts for each *CSH1* FSM transcript for each sample in GUSTO. Highlighted points correspond to ENST00000558284.1, with GDM status indicated. (**D**) Reads mapped to *CSH1* isoforms in GENCODE v45 and their respective mapping locations on the placenta-specific annotations. (**E**) DTE of *GNAS* FSM transcripts in GUSTO. (**F**) Raw read counts for each *GNAS* FSM transcript for each sample in GUSTO. Highlighted points correspond to ENST00000476196.5, with GDM status indicated. (**G**) Reads mapped to *GNAS* isoforms in GENCODE v45 and their respective mapping locations on the placenta-specific annotations.

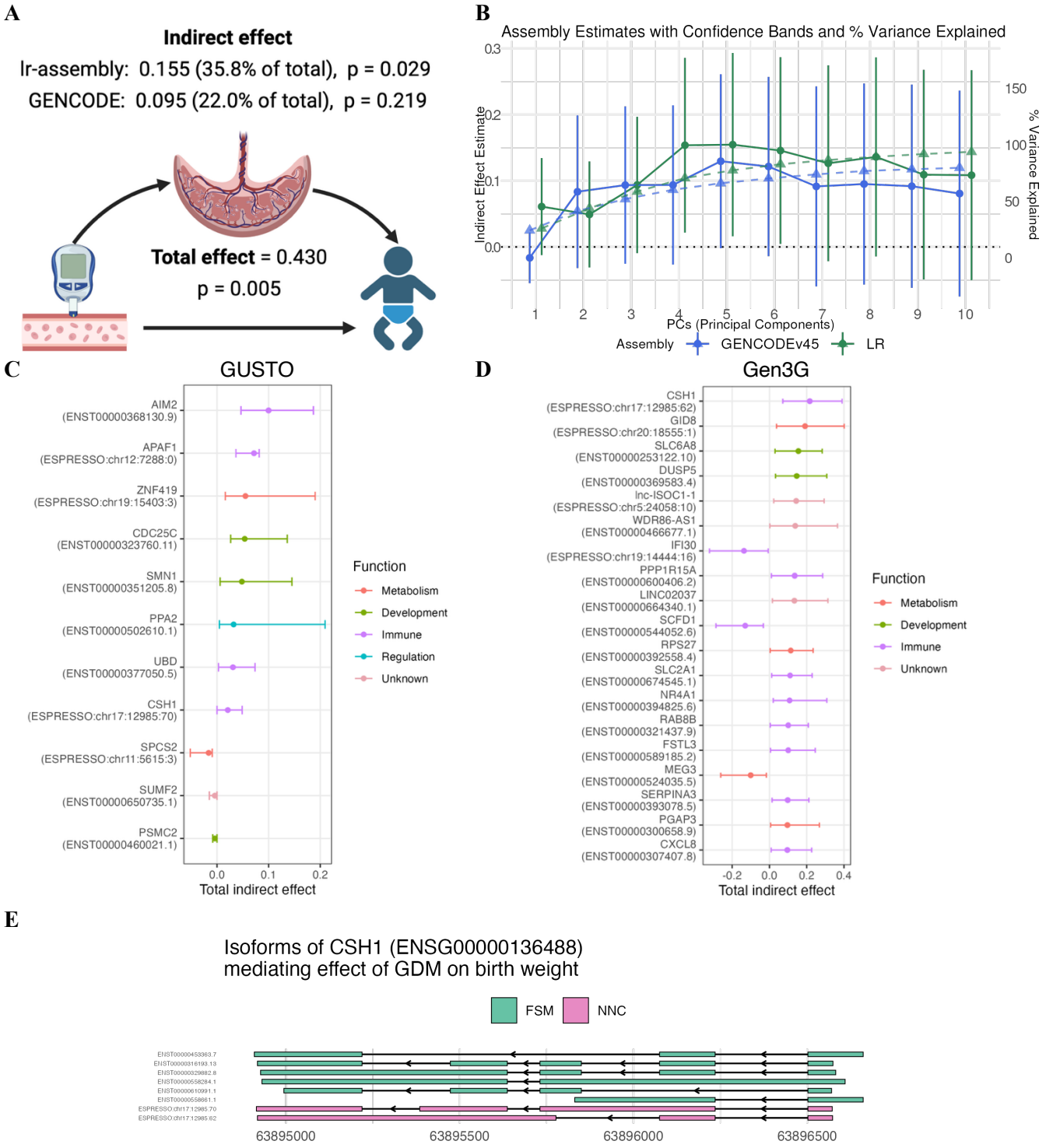


Figure 4 – GDM-induced shifts in placental transcription affect birth weight. (**A**) GDM induces a direct effect on birth weight, partially mediated through the placenta, adjusting for gestational age and fetal sex. Results of a multimediator analysis using the first 5 principal components of expression of GDM-associated DTEs at FDR < 10%. (**B**) Sensitivity analysis of the indirect effect of placental transcription on birth weight across % variance captured by varying numbers of PCs. Points indicate effect estimates while triangles indicate % variance. (**C**) Individual transcripts showing significant ($p < 0.05$) mediation effects in GUSTO when short-reads were quantified against the placenta-specific annotations. (**D**) Similarly, individual transcripts showing significant mediation effects in Gen3G. (**E**) Transcript structures of the NNC isoforms of *CSH1* showing significant mediation effects with reference to the FSM isoforms.