nature communications

Article

https://doi.org/10.1038/s41467-026-71303-4

Long-read assembly reveals vast transcriptional complexity in the placenta associated with metabolic and endocrine function

Received: 2 September 2025

Accepted: 19 March 2026

Published online: 02 April 2026

Check for updates

Sean T. Bresnahan[1] ✉, Hannah E. J. Yong[2], Aryun Nemani[3], 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[2,4,6,13] & Arjun Bhattacharya[1,12,13]

The placenta is critical for fetal development and mediates effects of pregnancy complications on offspring metabolic health yet remains poorly characterized in genomic studies. Existing transcriptomic analyses rely on adult tissue reference annotations, overlooking developmentally important splicing diversity. Using largest-in-class long-read RNA-seq (n = 72), we create a comprehensive placental transcriptome reference identifying 37,661 high-confidence isoforms (14,985 previously unannotated) across 12,302 genes (2,759 previously unannotated). Contrary to characterizations of the placenta as a transcriptomic void, we find transcriptional breadth and complexity comparable to adult tissues, with high splicing diversity of obesity- and growth-related gene transcripts, including 108 distinct *CSH1* (placental lactogen) isoforms. Applying this reference to short-read RNA-seq from diverse populations (n = 352) reduced inferential uncertainty in isoform quantification by approximately 30%. We find placental transcription mediated 36% of gestational diabetes mellitus effects on birth weight, with ancestry-specific effects including previously unannotated *CSH1* isoforms mediating larger effects in European (24.4%) than Asian (13.4%) populations. These findings illustrate the importance of tissue-matched, long-read annotations for isoform-resolved transcriptomics.

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 metabolic programming of later-in-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 characterizes physiological pregnancy, GDM involves insulin insufficiency relative to the exaggerated decline in maternal insulin sensitivity and glucose metabolism, potentially through altered placenta-mediated endocrine sig-

A full list of affiliations appears at the end of the paper. ✉e-mail: stbresnahan@mdanderson.org

naling such as the pro-diabetogenic human placental lactogen (hPL) encoded by *CSH1*[10]. GDM-exposed fetuses experience placenta-mediated alterations in intrauterine signals promoting macrosomia, altered body composition with increased adiposity, and heightened risk for metabolic disorders later in life[11,12]. These associations 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].

Despite the placenta's central role in prenatal development, it has been notably underrepresented in large-scale multi-tissue 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 transcriptional breadth and complexity relevant to fetal growth and metabolic programming.

Developing tissues, including the placenta, exhibit distinct patterns of alternative splicing and isoform usage compared to adult tissues[20–23]. Using 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, which faces 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, probabilistic read assignment introduces significant uncertainty in estimating isoform-level expression[26], a problem particularly acute in developing tissues exhibiting extensive alternative splicing[22].

Analyses of existing placental RNA-sequencing datasets have focused overwhelmingly on aggregated gene-level measurements[5,14,21,23,27–34], missing critical biological information by obscuring 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 are mediated through changes in specific isoforms that could be masked or compensated by expression of other, potentially non-functional isoforms, 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.

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 transcriptional breadth and complexity[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 current understanding of transcriptome regulation. Previous work, including from our group, 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 the discovery of transcriptome-mediated genetic associations with complex traits by approximately 60%[35,37,51].

Here, we develop a comprehensive transcriptome reference for placenta derived from $n = 72$ long-read RNA-sequencing libraries and orthogonal sequencing data. Our analysis identified 37,661 high-confidence transcripts representing 12,302 genes, including isoforms of both known and potentially novel genes, with pregnancy-related genes exhibiting extraordinary isoform diversity compared to other human tissues. We show that this assembly offers two complementary advantages: (1) improved specificity and accuracy of transcript quantification in short-read datasets ($n = 352$) by reducing mapping ambiguity and (2) the ability to identify isoform-level regulatory mechanisms that may be critical for understanding GDM-associated modulation of fetal development. We leveraged this annotation to study GDM and fetal growth, demonstrating how improved isoform resolution reveals molecular mechanisms underlying complex pregnancy outcomes. Our approach uncovered previously uncharacterized placental isoforms of genes involved in glucose metabolism, placental lactogen production and fetal growth regulation that mediate the effects of maternal hyperglycemia on birth weight. These findings demonstrate that long-read sequencing-derived placental transcript annotations are essential for understanding pregnancy pathophysiology and detecting mechanistic signals in observational studies of environmental and intrauterine exposures. Given the improved quantification precision and novel mechanistic insights demonstrated here, our placenta reference transcriptome provides a valuable resource for reanalyzing existing short-read datasets to uncover previously missed isoform-level regulatory mechanisms.

Results

Long-read sequencing reveals placental transcriptional diversity

Using samples collected from a subset of two prospective prebirth cohorts[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 GDM (Fig. 1A). Using the Oxford Nanopore platform, we sequenced cDNA libraries, obtaining approximately 310.8 million reads. After stringent quality control and alignment to 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) (Supplementary Data 1). These reads mapped to 68,002 distinct transcripts with a mean length of 1.36 kb (SD = 1.67 kb, range = 0.06–205 kb) (Supplementary Fig. 1A) and an average of 4.5 exons per transcript (Supplementary Fig. 1B). These transcripts represented 23,069 genes, with an average of 2.95 isoforms per gene (SD = 3.84) (Supplementary Fig. 1C). Our analysis captured 20,803 of 63,187 genes annotated in GENCODE v45 and identified 2266 previously unannotated genes including 102 fusion products of multiple genes (Supplementary Data 2).

One-quarter of placental transcripts are unannotated

Among transcripts annotated to known genes ($n = 65,386$ [96.15%]), we identified three major categories by comparison to GENCODE v45: transcripts matching known annotations, previously unannotated isoforms of known genes, and transcripts of previously unannotated genes (Fig. 1B, C). Most transcripts of known genes matched existing annotations, classified as full splice match (FSM: $n = 45,082$ [68.95%]), where all splice junctions exactly matched a reference transcript, or incomplete splice match (ISM: $n = 5176$ [7.91%]), where a subset of splice junctions matched a reference.

We identified 14,957 previously unannotated transcripts (22.87% of transcripts expressed in placenta) associated with 5630 known genes (26.93% of genes expressed in placenta). These previously unannotated 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 previously unannotated transcripts are categorized as: "novel not in catalog" (NNC, $n = 9578$ [63.9%]), with at least one previously unannotated splice donor or acceptor site; "novel in catalog" (NIC, $n = 5190$, [34.6%]), combining known splice sites in previously unobserved patterns; or "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 2759 transcripts representing previously unannotated genes, categorized as: antisense ($n = 1421$, [51.5%]), intergenic ($n = 1006$ [36.5%]), originating within

https://doi.org/10.1038/s41467-026-71303-4

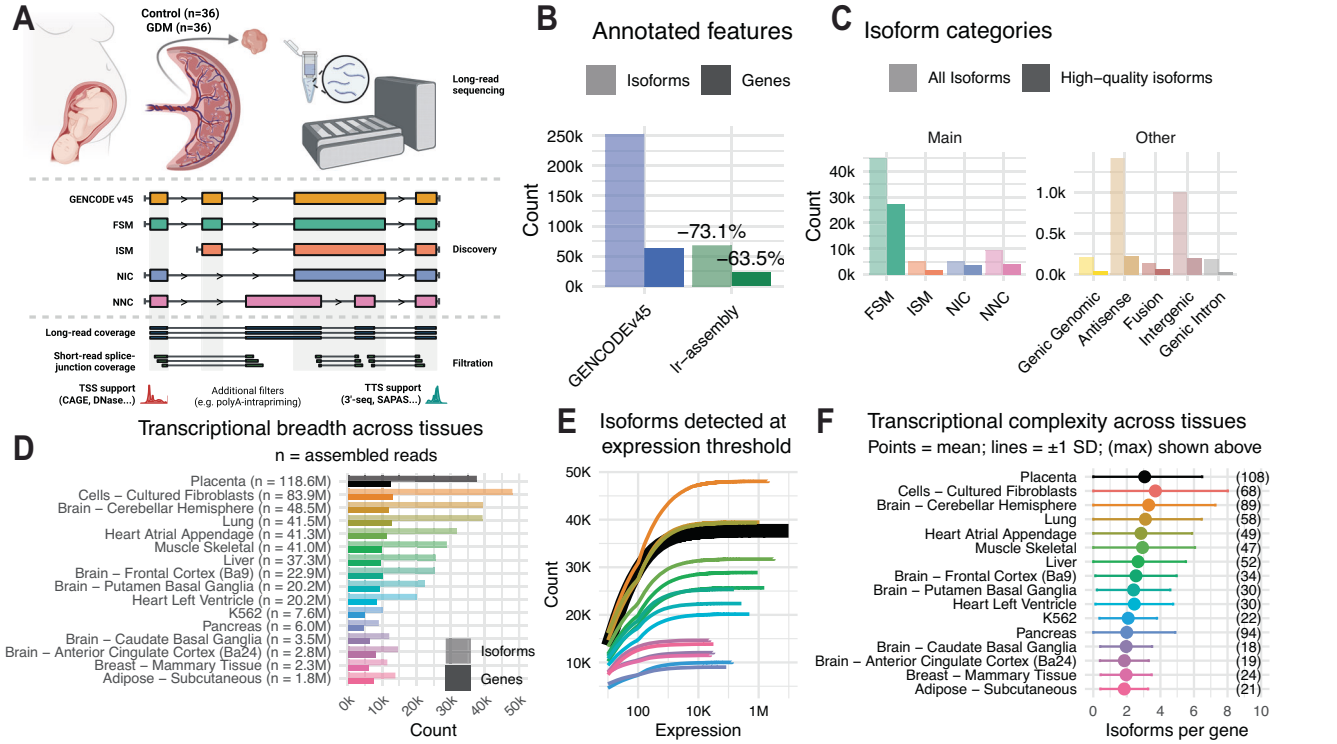


Fig. 1 | The placenta exhibits transcriptional breadth and complexity comparable to other adult tissues, challenging its characterization as a transcriptomic void. A Study design and transcript discovery pipeline. Nanopore cDNA libraries were prepared from villous placenta tissue of $n = 72$ term, live births from independent pregnancies ($n = 36$ controls, $n = 36$ GDM-affected). Discovery phase: transcripts were assembled and compared to 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, intergenic, genic, intronic, and antisense). Filtration phase: transcripts were retained if they had long-read coverage support, short-read splice junction coverage, transcription start site (TSS) support from CAGE and DNase-seq, transcription termination site (TTS) support from 3′-seq and SAPAS (Sequencing Alternative PolyAdenylation sites), and passed quality control filters (no poly(A)-intrapriming, RT-switching, or NMD signatures). Created in BioRender. Bhattacharya, A. (2026) https://BioRender.com/8fkghxk. **B** Comparison of annotated features between GENCODE v45 reference and unfiltered placenta transcriptome (lr-assembly). **C** Transcript distribution by structural category. Light bars: all isoforms ($n = 68{,}002$). Dark bars: high-confidence filtered isoforms ($n = 37{,}661$). **D** Transcriptional breadth across 15 tissues/cell lines from GTEx v9 long-read data. Bars show counts of unique genes (dark) and isoforms (light). Sample sizes shown as total assembled reads per tissue. **E** Isoforms detected at increasing expression thresholds. Placenta shown as a thick black line. Cumulative counts across samples per tissue. **F** Transcriptional complexity as isoforms per gene. Points: mean. Lines: ±1 SD. Parentheses: maximum isoforms per gene (placenta max = 108). Source data are provided as a source data file.

introns of known genes ($n = 189$ [6.8%]), or fusion products spanning multiple genes ($n = 143$ [5.2%]).

Orthogonal data validate high-confidence transcript models

For downstream analyses, we filtered out low-confidence isoforms using criteria from previous long-read transcriptome assembly studies (Methods)[36,45–49]. We used orthogonal short-read RNA-seq, DNase-seq, CAGE-seq, and SAPAS data to validate splice junction coverage and transcription start and termination sites (TSS and TTS) (Supplementary Data 16). Transcript models showed robust support (Table 1; Supplementary Data 2) with an average of 554.23 long-reads (Nanopore; SD = 3.89×10^4) (Supplementary Fig. 2A) and an average minimum splice junction support of 2.33×10^4 short-reads (Illumina; SD = 2.47×10^5) (Supplementary Fig. 2B). We observed strong enrichment of TSS near CAGE-seq or DNase-seq peaks (median distance = 90 bp; 23,740 [34.9%] of transcripts within 50 bp) (Supplementary Fig. 2C), and TTS near alternative polyadenylation sites or PolyA motifs (median distance = 15 bp; 26,549 [40.5%] of transcripts within 5 bp) (Supplementary Fig. 2D, E).

Quality filtering identified a few problematic transcripts: $n = 2243$ isoforms (3.3%) showed evidence of nonsense-mediated decay (NMD), $n = 483$ (0.7%) exhibited RT-switching artifacts, and $n = 4$ (0.005%) showed evidence of PolyA intrapriming (Supplementary Fig. 2F). While NMD transcripts can be biologically meaningful, incomplete assemblies or artifacts can also produce premature termination codons. Therefore, for NMD-predicted transcripts, we required additional evidence of substantial 5′ expression to distinguish genuine NMD substrates from assembly artifacts; of 2243 NMD-predicted transcripts, 892 were filtered exclusively due to insufficient 5′ expression evidence, and the remaining 1351 were filtered for other quality control reasons. After removing 30,341 low-confidence isoforms (44.62%; Supplementary Data 3), our placenta transcript assembly contained 37,661 high-confidence isoforms (Fig. 1C) with mean length of 1.60 kb (SD = 1.52 kb, range = 0.08–34.63 kb; Supplementary Fig. 3A) and an average of 5.62 exons per transcript (Supplementary Fig. 3B), representing 12,302 genes (mean = 3.06, SD = 3.43 isoforms per gene), including 11,854 GENCODE v45 genes and 448 previously unannotated genes (Supplementary Fig. 3C). Among the 500 most abundant placental long-read transcripts, we observed a 4.3-fold enrichment for placenta-specific genes (adjusted $p = 7.85 \times 10^{-7}$; Supplementary Data 4; Supplementary Fig. 3D).

Unannotated transcripts show distinct structural features

Compared to transcripts of known genes, transcripts of previously unannotated genes were significantly less abundant in the long-reads (Mann-Whitney-Wilcoxon test: $W = 1.07 \times 10^7$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. 4A), shorter in length ($W = 1.31 \times 10^7$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. 4B), and contained fewer exons ($W = 1.43 \times 10^7$,

https://doi.org/10.1038/s41467-026-71303-4

Table 1 | Summary of orthogonal support for placenta transcripts assembled from long-read sequencing data

Structural Category	Long-read coverage	Splice-junction coverage	TSS support	TTS support	No poly(A)-intrapriming	No NMD	No RTS
FSM	32722	36111	45082	45082	45082	45082	45082
ISM	4263	5100	3549	3266	5175	5176	5176
NIC	4909	4966	4666	4581	5189	5190	5190
NNC	8632	7298	8076	8212	9578	7821	9285
Genic Genomic	173	99	169	150	215	186	201
Antisense	1094	551	1049	812	1421	1193	1309
Fusion	129	120	116	120	143	124	137
Intergenic	790	478	797	765	1006	812	974
Genic Intron	118	67	143	117	189	175	165

Stratified by transcript structural category relative to GENCODE v45, the count of isoforms passing thresholds for long-read coverage, short-read coverage of splice junctions, with TSS and TTS support, and without evidence of poly(A)-intrapriming, nonsense mediated decay (NMD) or reverse template switching (RTS) are indicated.

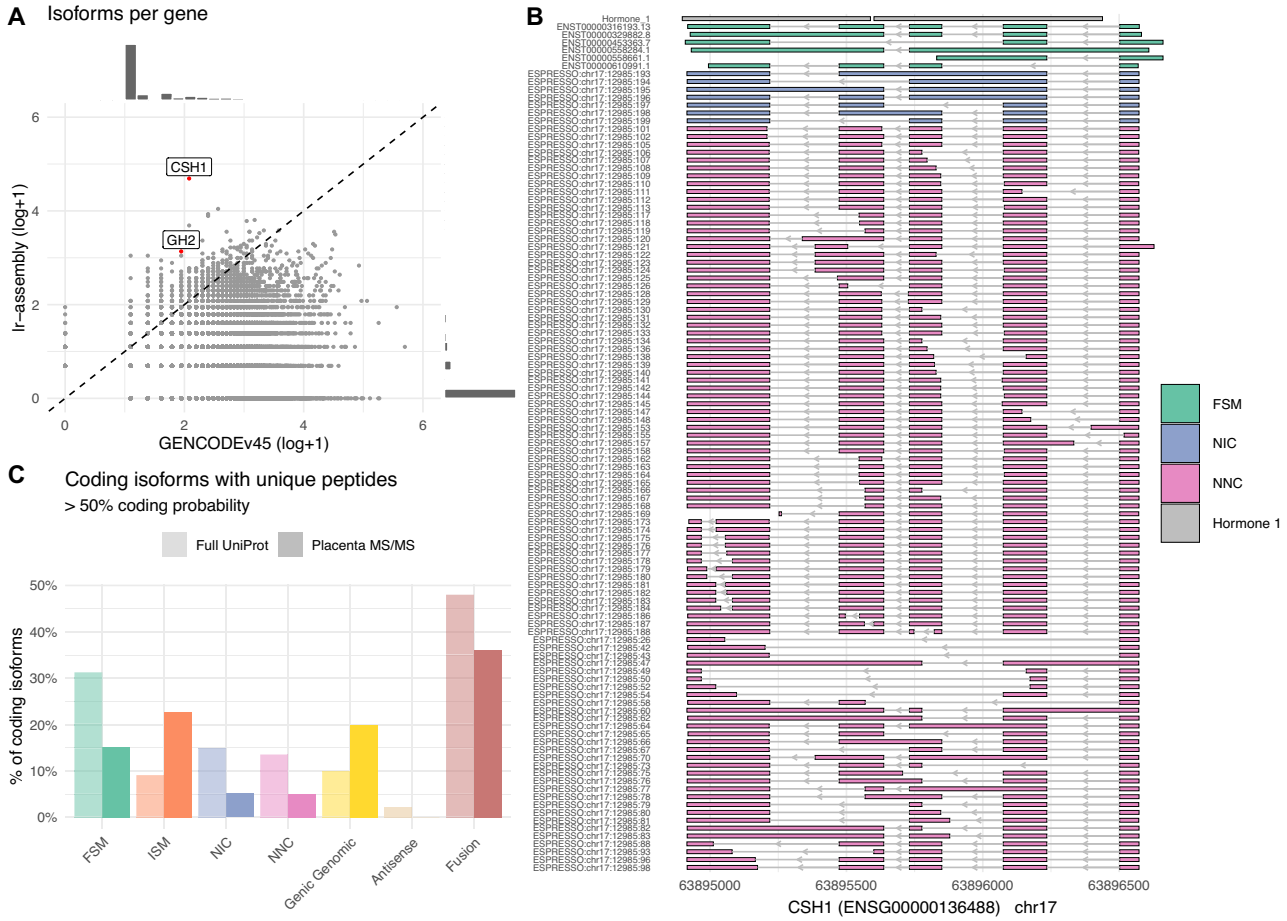


Fig. 2 | The placenta exhibits relatively high splicing diversity of pregnancy-related genes. A Comparison of isoforms per gene between GENCODE v45 reference and high-confidence placenta transcriptome (lr-assembly). Each point represents a gene. Total genes in lr-assembly: $n = 12{,}302$. Total genes in GENCODE v45: $n = 61{,}572$. Red points highlight *CSH1* ($n = 6$ isoforms in GENCODE, $n = 108$ in lr-assembly) and *GH2* ($n = 6$ in GENCODE, $n = 22$ in lr-assembly). **B** Transcript structure visualization of *CSH1* gene (chr17:63,895,000-63,896,500) showing high-confidence isoforms expressed in placenta. Isoforms colored by structural category: FSM (full splice match, green; $n = 6$ isoforms), NIC (novel in catalog, blue; $n = 7$ isoforms), NNC (novel not in catalog, pink; $n = 95$ isoforms). Boxes represent exons; lines represent introns with arrows indicating strand direction. The top track shows the Hormone 1 protein domain (Pfam) in gray. **C** Validation of coding potential for isoforms across the structural categories. Bars show percentage of isoforms with predicted open reading frames (ORFs) and CPC2 coding probability scores >50%, that have coding sequence (CDS) support from the full human UniProt database (light bars) or placental tissue tandem mass spectrometry (MS/MS, dark bars). Source data are provided as a source data file.

$p < 2.2 \times 10^{-16}$) (Supplementary Fig. 4C). Similarly, previously unannotated transcripts 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 ($W = 1.15 \times 10^8$, $p = 0.99$). We identified substantial isoform diversity in multi-exonic RNA isoforms (range: 1–108 isoforms per gene) with over half of all detected genes ($n = 7343$ [59.69%]) expressing more than one isoform (Fig. 2A). The number of

detected isoforms was correlated with the number of detected exons per isoform ($r = 0.87$, $p < 2.2 \times 10^{-16}$) 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 (Supplementary Fig. 4D).

These structural features are consistent with patterns observed in evolutionarily younger genes across multiple species, where gene length and transcript complexity show directional trends toward increasing structural elaboration over evolutionary time[53]. Approximately half of the previously unannotated transcripts in our assembly are non-coding, raising the possibility that some may represent transcriptional noise rather than functional innovations under selection. However, definitive characterization of evolutionary origin and functional significance would require synteny analysis, phylogenetic reconstruction, or experimental validation to distinguish between lineage-specific functional genes, neutrally-evolving transcripts, and transcriptional noise[54–59].

The placenta is not a transcriptomic void

To evaluate placental transcriptional breadth and complexity with reference to other human tissues, we reprocessed 15 GTEx[47] tissues and cell lines using an identical bioinformatic pipeline to our placental samples (Tables S16, S17; "Methods"). The placenta exhibits transcriptional breadth (total number of isoforms and genes expressed) comparable to other human tissues (Fig. 1D), with saturation analysis confirming sufficient sequencing depth for robust comparison (Fig. 1E). Transcriptional complexity (the mean number of isoforms per gene) ranged from 1.8 to 2.8 across tissues (Fig. 1F). These findings challenge previous characterizations of the placenta as a transcriptomic void expressing few, high-abundance transcripts[21,23,60]. Instead, the placenta exhibits transcriptional breadth and complexity comparable to other tissues and utilizes complex splicing regulation—particularly for genes involved in maternal-fetal signaling.

Pregnancy-related genes show high transcriptional complexity

While transcriptional breadth and complexity of the placenta are comparable to other tissues, it exhibits relatively high transcriptional complexity in specific genes with critical functions in pregnancy and placental development. Chorionic somatomammotropin hormone 1 (*CSH1*), which regulates human placental lactogen production in pregnancy[61], displayed the greatest splicing diversity with 108 distinct isoforms (Fig. 1F, maximum values shown in brackets; Fig. 2A, B). 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 splicing diversity with 22 distinct isoforms. 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 transcriptional complexity across this gene family suggests complex regulation of placental endocrine functions, supported by significant enrichment for hormone receptor binding (odds ratio = 79.3, adjusted $p = 1.29 \times 10^{-3}$) for the 50 most isoform-rich genes through Gene Ontology (GO) analysis (Supplementary Data 5).

Unannotated transcripts encode potentially functional proteins

We evaluated the coding potential and peptide validation of detected transcripts. Open reading frames (ORFs) were predicted for 19,670 (52.2%) isoforms (Supplementary Fig. 5A), with similar coding sequencing (CDS) length across structural categories (median = 1.11 kb) (Supplementary Fig. 5B). Coding potential, measured by CPC2 coding probability scores[62], varied by structural category: FSM, NIC, and NNC transcripts showed similarly high coding potential (median = 85.39%) while ISM (median = 66.02%) and other categories (median = 57.19%) displayed lower potential (Supplementary Fig. 5C).

We used BlastP[63] to identify isoform-unique peptides from tandem mass spectrometry (MS/MS) of human placental tissue, including GDM-affected pregnancies[64,65], ensuring validation of isoform-specific features. We found that 15.2% of reference transcripts (FSM) and 5.0–5.2% of novel isoforms (NIC, NNC) contained unique peptides detected in placental MS/MS (Fig. 2C; Supplementary Data 6). Notably, ISM (22.7%) and genic genomic transcripts (20.0%) showed higher validation rates in placental MS/MS compared to the full UniProt database (9.1% and 10.0%, respectively). In contrast, the full UniProt database revealed higher validation rates for reference and novel junction transcripts (FSM: 31.1%, NIC: 15.0%, NNC: 13.6%), with lower validation in the tissue-specific MS/MS reflecting the limited dynamic range and condition-specific nature of proteomic detection.

We identified 80 transcripts of potentially novel genes (17.62%) with predicted ORFs—44 of which show MS/MS validation—encoding functionally relevant proteins including fibronectin (a potential GDM biomarker)[66], E-cadherin (regulating trophoblastic differentiation)[67], alpha-enolase (5 distinct isoforms, differentially expressed in GDM)[68], occludin (placental barrier integrity)[69], and CCL14 (trophoblast migration and maternal-fetal communication)[70]. Transcripts without peptide validation may include both non-coding RNAs, which can have important regulatory functions, as well as proteins that may be below MS/MS detection limits due to low abundance, post-translational modifications affecting peptide recovery, condition-specific expression not represented in available proteomic datasets, or cell-type-specific expression diluted in bulk tissue samples.

Placenta annotations reduce short-read quantification uncertainty

To evaluate the impact of transcript discovery and filtration on short-read transcript quantification, we compared quantifications against our placenta reference transcriptome, GENCODE v45, and their union (GENCODE+). We analyzed Illumina cDNA libraries from villous placental tissue collected in two prospective prebirth cohorts: GUSTO[28] ($n = 200$; "Methods") and the Genetics of Glucose regulation in Gestation and Growth (Gen3G[29]; $n = 152$; "Methods"), both with approximately 20% of samples from GDM-affected pregnancies (Fig. 3A; Supplementary Data 7 and Supplementary Fig. 6A).

Our placenta reference transcriptome demonstrated population-specific differences in total quantified expression that varied by cohort ancestry. In the GUSTO cohort (Singaporean), which shared ancestry with the samples used to generate the placenta reference transcriptome, lr-assembly achieved the highest total quantified expression with an increase of 7351 log(TPM) over GENCODE v45 (17.9% increase, $p < 2 \times 10^{-16}$) and 4622 log(TPM) over GENCODE+ (10.6%, $p < 2 \times 10^{-16}$) (Supplementary Data 8). In contrast, the Gen3G cohort (white European) showed the highest total quantified expression with GENCODE + , with an increase of 13,044 log(TPM) over GENCODE v45 (31.2%, $p < 2 \times 10^{-16}$). The lr-assembly alone provided a smaller but significant 3,896 log(TPM) increase over GENCODE v45 in Gen3G (9.3%, $p = 1.45 \times 10^{-10}$). These differences in total quantified expression across cohorts (interaction $p < 2 \times 10^{-16}$) likely reflect the ancestry composition of our long-read samples, with GENCODE+ detecting additional isoforms in Gen3G not captured in the Singaporean-derived lr-assembly (Fig. 3B).

Despite these ancestry-specific differences in total quantified expression, transcripts that failed to meet our stringent quality criteria showed significantly lower mean expression and variance in the short-read quantifications compared to FSM and previously unannotated transcripts in both cohorts (Fig. 3C; Supplementary Fig. 6B). Statistical analysis revealed a clear hierarchy in transcript expression levels and variance (Supplementary Data 9): shared (FSM) transcripts showed the

https://doi.org/10.1038/s41467-026-71303-4

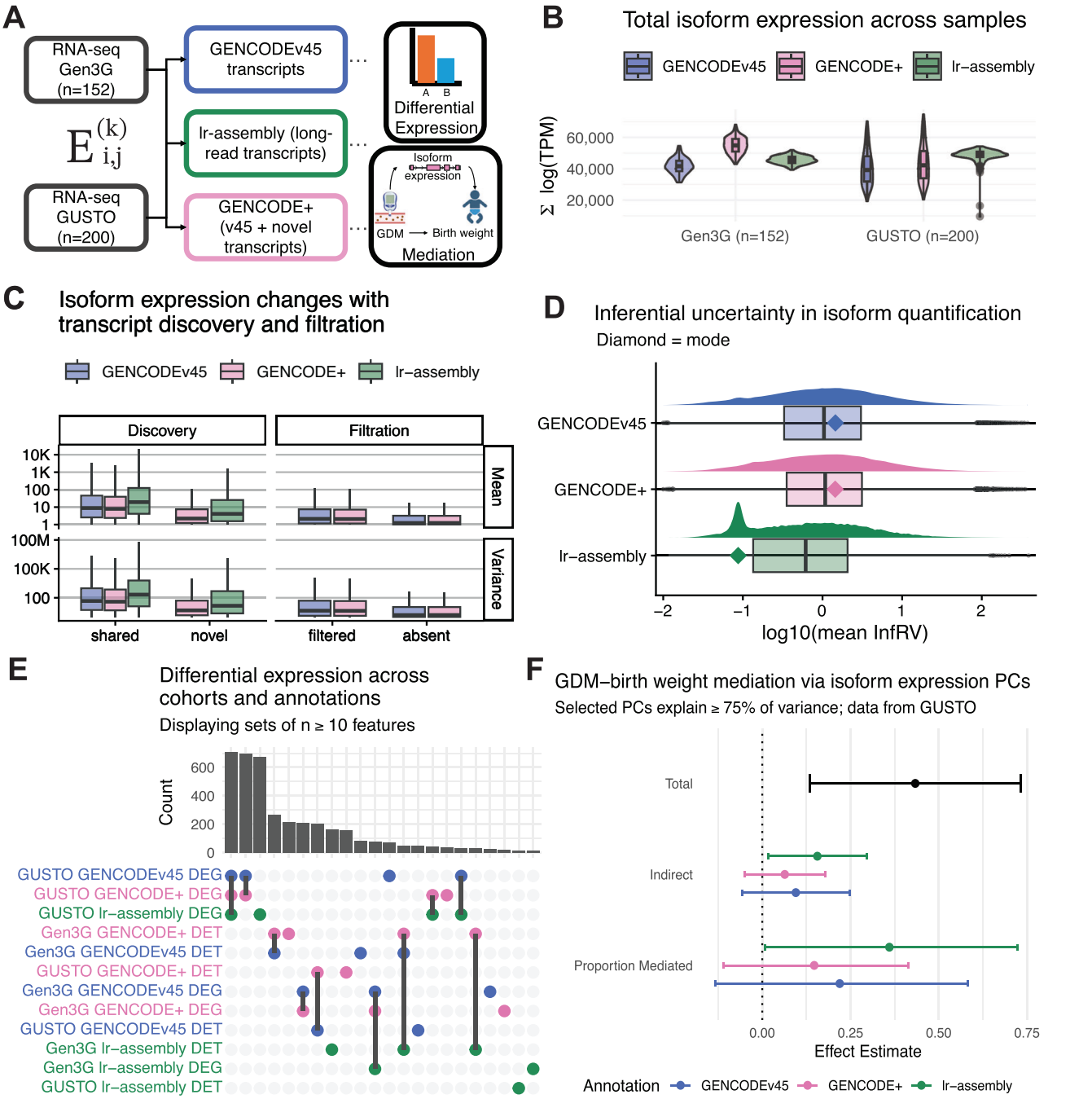


Fig. 3 | Isoform-aware quantification of placental short-read RNA-seq reduces inferential uncertainty and reveals transcriptional mediation of GDM effects on birth weight. A Overview of the multi-cohort placental RNA-seq analysis framework for differential expression and mediation analyses. Short-read RNA-seq of villous placental tissue from two prospective prebirth cohorts (~20% GDM-affected) was quantified against GENCODE v45, high-confidence placenta transcripts (lr-assembly), and their union (GENCODE+). Transcript expression was estimated across 50 bootstrap replicates using Salmon to estimate overdispersion from read-to-transcript mapping ambiguity and inferential relative variance (InfRV), a measure of quantification uncertainty. Created in BioRender. Bhattacharya, A. (2026) https://BioRender.com/3udkvk4. **B** Total transcript expression (TPM) per short-read sample in GUSTO ($n = 200$ independent placenta samples, Singaporean cohort) and Gen3G ($n = 152$ independent placenta samples, white European cohort) per annotation. **C** Mean and variance in GUSTO short-read samples across annotations for transcripts shared between annotations, previously unannotated (novel), removed during filtering (filtered), and not assembled (absent). **D** InfRV in isoform quantification in GUSTO samples per annotation; diamond = mode. **B**–**D** Box plots show median (center line), interquartile range (IQR, box bounds), and 1.5× IQR from quartile bounds (whiskers); points beyond whiskers are outliers. **E** Count of DEGs and DETs per cohort and annotation; set labels colored by annotation (green, lr-assembly; blue, GENCODE v45; pink, GENCODE+) in GUSTO ($n = 44$ GDM-affected, $n = 145$ unaffected) and Gen3G ($n = 17$ GDM-affected, $n = 135$ unaffected). Differential expression tested using DESeq2 with negative binomial GLMs, technical variation estimated with RUVr ($k = 2$), two-sided Wald test, Benjamini-Hochberg FDR < 0.1, adjusting for gestational age and infant sex. **F** Effect estimates of the GDM–birth weight relationship mediated through placental isoform expression in GUSTO. PCs derived from DETs per annotation; mediators were the first PCs explaining ≥75% variance (lr-assembly: 5 PCs = 77.1%; GENCODE v45: 8 PCs = 76.7%; GENCODE + : 6 PCs = 75.0%). Mediation models fit using structural equation modeling (lavaan), adjusting for infant sex, gestational age, and maternal ethnicity; 95% CIs and p-values from the asymptotic covariance matrix. Data presented as effect estimate ± 95% CI. Source data are provided as a source data file.

highest expression, followed by previously unannotated transcripts, then filtered transcripts, and finally absent transcripts. In GUSTO, shared transcripts had 1.14–1.67 log-fold higher expression than other categories (all $p < 1 \times 10^{-4}$), with expression variance showing similar patterns (1.88–2.78 log-fold differences, all $p < 1 \times 10^{-4}$). Importantly, previously unannotated transcripts showed expression levels only slightly lower than shared transcripts, both of which had higher expression levels than filtered or absent transcripts, confirming that our assembly captures functionally relevant isoforms rather than low-quality artifacts. Notably, transcripts absent from our lr-assembly but present in GENCODE (which contributed to Gen3G's higher expression with GENCODE+) also showed significantly lower expression and variance, suggesting many represent low-confidence annotations.

To formally quantify read-to-transcript mapping ambiguity, we calculated transcript-level inferential relative variance (InfRV) for each sample across both cohorts. As InfRV is roughly stabilized with respect to the mean expression level, higher InfRV values indicate greater uncertainty in estimating transcript abundance[71] ("Methods"). Mean transcript-level InfRV was consistently reduced when short reads were quantified against our placenta reference transcriptome compared to both GENCODE v45 (30.4% reduced in GUSTO and 30.3% in Gen3G) and GENCODE+ (31.5% reduced in GUSTO and 30.4% in Gen3G) (Fig. 3D; Supplementary Fig. 6C), indicating improved isoform-level quantification precision. Critically, this improvement was observed in both cohorts despite ancestry-specific differences in total expression, indicating that the lr-assembly reduces mapping ambiguity even for the Gen3G cohort, where GENCODE+ achieved higher total quantified expression. When stratified by structural category, transcripts in "Other" categories (antisense, intergenic, fusion, and genic genomic) displayed low mean InfRV and the weakest between-sample correlations (Supplementary Fig. 6D–G), suggesting unambiguous read mapping due to simpler transcript structures with high biological or technical variance reflecting potentially stochastic transcription or lower abundance that amplifies technical noise.

As reduced InfRV is positively correlated with reduced exon sharing between transcripts (Supplementary Fig. 6H; $r = 0.687$, $p = 3.85 \times 10^{-279}$), smaller annotations would be expected to show lower InfRV values. To confirm that our observed reduction reflected tissue-matched isoform content rather than annotation size alone, we quantified placental short-read RNA-seq against long-read transcript assemblies for two GTEx tissues: subcutaneous adipose (one of the smallest of our assemblies with 13,829 transcripts) and cultured fibroblasts (the largest; 47,850 transcripts). Both GTEx annotations reduced InfRV relative to GENCODE v45 or GENCODE + , consistent with reduced exon sharing from smaller annotations. However, the placenta reference had the lowest mode InfRV (mode $\log_{10}$ of InfRV estimated by kernel density: placenta = −1.05, adipose = −0.86, fibroblasts = −0.86; Supplementary Fig. 6I). This demonstrates that non-tissue-matched assemblies contain tissue-specific isoforms not expressed in the tissue of interest, introducing irrelevant exon sharing. In contrast, a tissue-matched annotation process both *excludes* irrelevant isoforms through assembly and filtering and *includes* relevant, previously unannotated isoforms through discovery. These data confirm that tissue-matched long-read annotations improve quantification precision beyond what would be expected from annotation size reduction alone and that the benefits of our placenta reference transcriptome extend across ancestry groups, even when not all ancestry-specific isoforms are captured.

Placenta annotations reveal extensive GDM-associated changes

We next examined differences at gene and transcript levels associated with GDM status in GUSTO and Gen3G. In the following analyses, we excluded eleven GUSTO libraries as GDM status was not available for eight samples, and three were identified as outliers (Supplementary Data 7; Supplementary Fig. 6A). We compared analyses using our placenta reference transcriptome versus GENCODE v45 and their union (GENCODE+), adjusting for gestational age, infant sex, and technical variation estimated with RUVr[72] (Fig. 3E; Supplementary Fig. 7A, B). Using FDR < 10% thresholds, our placenta reference transcriptome revealed an average of 200 differentially expressed transcripts (DETs) between cohorts (SD = 244.6) and 1492 differentially expressed genes (DEGs; SD = 18.4), while GENCODE v45 yielded 333.5 DETs (SD = 99.7) and 914.5 DEGs (SD = 867.6), and GENCODE+ yielded 402.5 DETs (SD = 41.7) and 923 DEGs (SD = 823.1). Analyses using our placenta reference transcriptome resulted in the identification of fewer, but a higher annotated percentage of, GDM-associated features (Tables S10, 11). This enrichment, combined with the reduced inferential uncertainty 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 annotated features.

Placental transcription mediates 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 (Supplementary Data 12). Our placenta reference transcriptome 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 $p = 2.49 \times 10^{-3}$) (Supplementary Fig. 7C). In contrast, 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 $p = 4.23 \times 10^{-6}$) (Supplementary Fig. 7D). To determine which approach better captures functionally relevant changes that mechanistically link GDM to birth weight outcomes, we performed causal mediation analyses.

Given the known impacts of GDM on birth weight, we examined whether some of these effects are mediated by placental transcription in GUSTO using multi-mediator analysis. 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, for equal comparison, selected the first *PC* principal components which cumulatively explained at least 75% of the total variance in transcript expression as mediators[73,74] (lr-assembly, *PC* = 5 [77.09% total variance]; GENCODE v45, *PC* = 8 [76.72%]; GENCODE+, *PC* = 6 [75.00%]). We addressed GDM-transcriptome (exposure-mediator) and transcriptome-birthweight (mediator-outcome) confounding by adjusting for maternal ethnicity, infant sex, and gestational age. We confirmed these principal components met our 75% variance threshold through sensitivity analysis (Supplementary Fig. 8A, B; Supplementary Data 13).

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 transcription when quantifying short-reads against our placenta reference transcriptome was statistically significant (0.155 [0.016, 0.293], $p = 0.029$), with approximately 35.8% of the total effect being mediated through placental transcript expression (Fig. 3F). Consistent with the pathway enrichment results, our tissue-matched 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 either GENCODE v45 or GENCODE+, the proportion mediated by placental transcript expression was smaller and not statistically significant (GENCODE v45: 0.095 [−0.057, 0.247], $p = 0.219$) (GENCODE+: 0.064 [−0.050, 0.178], $p = 0.273$) and these annotations produced more heterogeneous mediators including several transcripts of uncharacterized genes.

https://doi.org/10.1038/s41467-026-71303-4

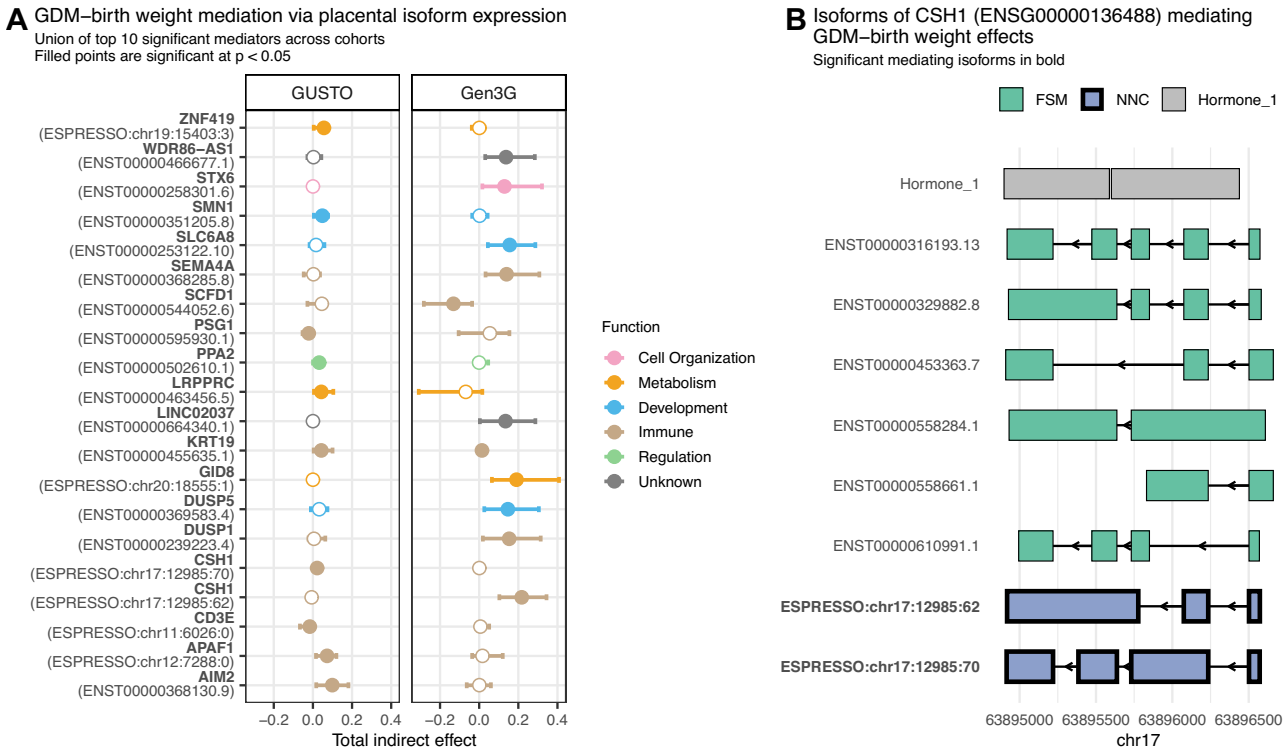


Fig. 4 | GDM-induced shifts in placental transcript expression are associated with variation in birth weight. A Indirect effect estimates for transcripts showing significant GDM-birth weight mediation effects in GUSTO ($n = 189$) and Gen3G ($n = 152$) when short-read samples were quantified against the placenta reference annotations (lr-assembly). Each data point represents an individual transcript from the union of the top 10 significant mediators per cohort ($n = 20$ total transcripts: 10 from GUSTO, 10 from Gen3G). Points colored by inferred biological function based on Gene Ontology Biological Process annotations: Cell Organization, Metabolism, Development, Immune, Regulation, Unknown. Filled points indicate significance at $p < 0.05$ and 95% confidence intervals (CIs) excluding 0. Mediation analysis was performed using the mediation R package with bootstrap resampling. Linear regression models were fit for: (1) mediator model: transcript expression ~ GDM + sex + gestational age (+ maternal ethnicity for GUSTO); (2) outcome model: birth weight ~ GDM + transcript expression + sex + gestational age (+ maternal ethnicity for GUSTO). *p*-values are two-sided from quasi-Bayesian Monte Carlo simulation based on bootstrap samples, and CIs (95%) were calculated from bootstrap resampling. Average causal mediation effects (indirect effects) represent the change in birth weight mediated through changes in transcript expression. Data presented as total indirect effect estimate ± 95% CIs. Exact *p*-values and CIs for all 20 transcripts are provided in Supplementary Data 14. **B** Transcript structures of *CSH1* gene (ENSG00000136488; chr17:63,895,000-63,896,500) showing isoforms with significant GDM-birth weight mediation effects. NNC (novel not in catalog) isoforms in pink (ESPRESSO:chr17:12985:62, ESPRESSO:chr17:12985:70) contain previously unannotated intron retention events (bold labels). FSM (full splice match) reference isoforms in green. Protein domain track (Hormone 1, Pfam) shown in gray. Source data are provided as a source data file.

To confirm that these observations were not merely driven by differences in annotation size, we quantified the GUSTO short-read samples against GTEx transcript assemblies for subcutaneous adipose tissue and cultured fibroblast cells, performing identical differential transcript expression (Supplementary Fig. 8C) and multimediator analyses (Supplementary Fig. 8D, E; Supplementary Data 13) as for the other annotations. Consistent with our observation that annotation size alone does not drive differences in quantification uncertainty, quantification against annotations of different sizes yielded insignificant GDM-birth weight mediation by transcript expression (subcutaneous adipose: $PC = 1$, 0.045 [−0.017, 0.106], $p = 0.154$) (fibroblasts: $PC = 2$, 0.049 [−0.031, 0.128], $p = 0.230$). These findings demonstrate that both transcript discovery and filtration improve detection of biologically relevant pathway enrichments and estimation of mechanistically relevant transcriptional changes that mediate clinically important outcomes.

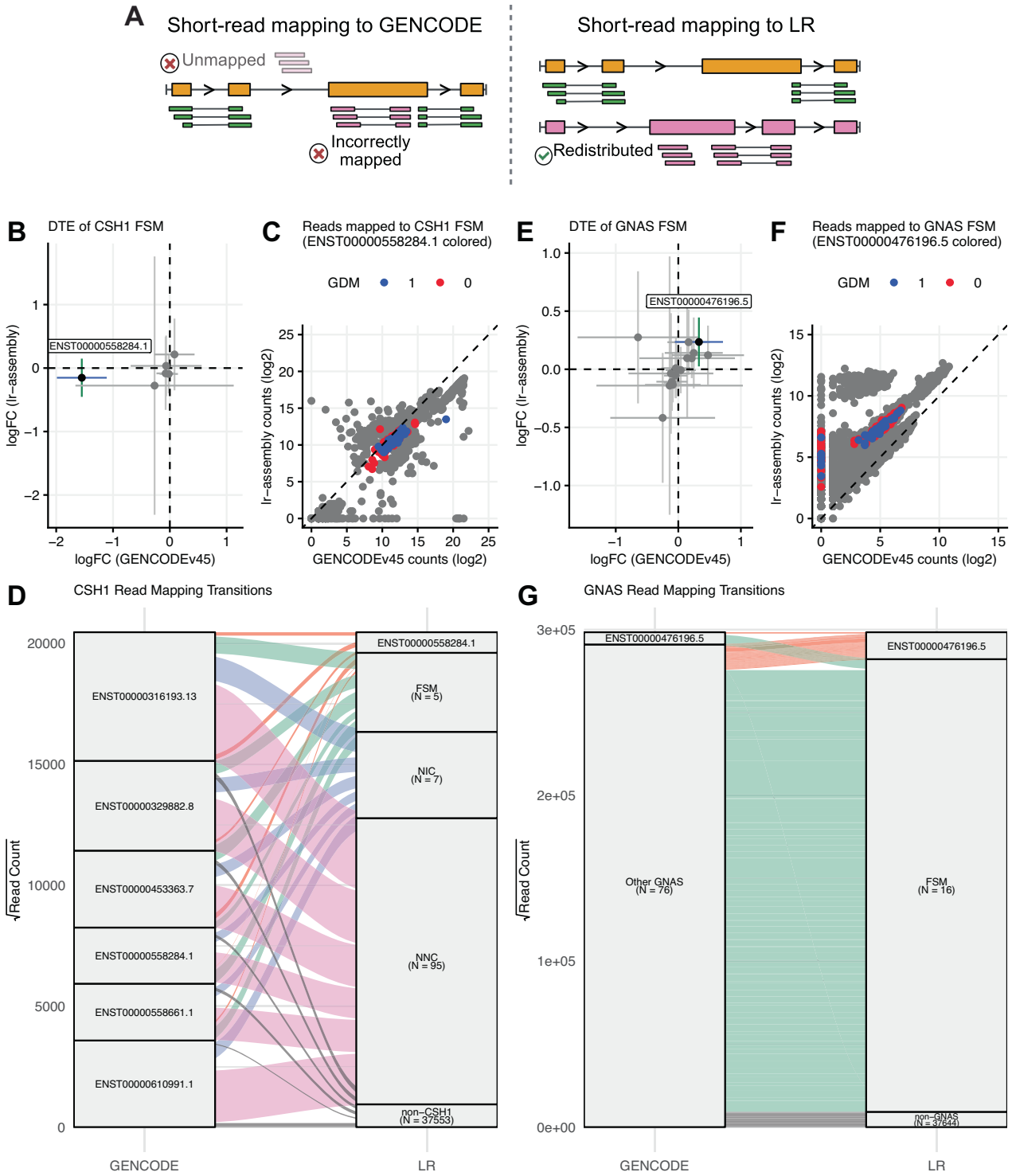
Unannotated *CSH1* isoforms mediate GDM effects on growth

We further identified specific transcripts that mediate the effect of GDM on birth weight (Fig. 4A; Supplementary Data 14). In GUSTO, 11 transcripts showed significant mediation effects ($p < 0.05$), including ESPRESSO:chr17:12985:70 (0.021, 95% CI [0.001, 0.048], $p < 2 \times 10^{-16}$), an isoform of *CSH1* not annotated in GENCODE, accounting for approximately 13.4% of the effect of placental transcription. Similarly, in Gen3G, 19 transcripts showed a significant effect, including ESPRESSO:chr17:12985:62 (0.216 [0.071, 0.389], $p < 2 \times 10^{-16}$), another previously unannotated isoform of *CSH1*, 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). 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 (Fig. 4B). This population-specific variation was unique to *CSH1* in our placenta reference transcriptome: GENCODE-annotated transcripts showed no significant cross-cohort differences in expression patterns.

Long-read annotations redistribute short-read assignments

To understand how differences in transcript annotations affect differential expression (and by extension, causal mediation) analyses, we tracked changes in effect sizes in GUSTO between quantifications on GENCODE v45 and our placenta reference transcriptome (Fig. 5A). For *CSH1*, which exhibited the most dramatic expansion in splicing diversity in our reference, one FSM isoform (ENST00000558284.1) only showed significant differential expression associated with GDM when quantified against GENCODE (Fig. 5B). While this isoform showed moderate expression in both references, it had notably higher expression when quantified against GENCODE (Fig. 5C), owing to a

https://doi.org/10.1038/s41467-026-71303-4



redistribution of reads to ENST00000558284.1 or to previously unannotated isoforms when the data is quantified with our reference (Fig. 5D). In contrast, *GNAS*—a known imprinted gene in placenta[75] that exhibited one of the most dramatic contractions in splicing diversity relative to GENCODE (77 isoforms reduced to 17 in our reference)—showed the opposite pattern. Isoform ENST00000476196.5 showed a significant GDM association when quantified against our placenta reference transcriptome but not against GENCODE (Fig. 5E), and increased expression with our reference (Fig. 5F, G). These findings illustrate how both adding unannotated transcripts and removing irrelevant transcripts from a transcriptome reference 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 signals.

Discussion

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]. Using matched bioinformatic processing of placental and GTEx[47] tissues, our results demonstrate that the placenta

Fig. 5 | Transcriptome annotation of placenta refines short-read mapping to individual transcript isoforms and uncovers distinct patterns of differential expression. A Tissue-specific transcriptome annotation improves short-read mapping accuracy. Short reads (green and pink boxes) that are unmapped or incorrectly mapped against GENCODE transcripts (gold) are redistributed to the correct transcripts upon quantification against the tissue-specific reference transcriptome (lr-assembly, gold and pink transcripts). Created in BioRender. Bhattacharya, A. (2026) https://BioRender.com/z66on6s. **B** Differential expression of *CSH1* full splice matches (FSM, $n = 6$ transcripts) in GUSTO ($n = 189$). Each point represents one transcript. Vertical error bars: 95% CI of $\log_2$ fold change quantified against lr-assembly (gray = other transcripts; green = ENST00000558284.1). Horizontal error bars: 95% CI quantified against GENCODE v45 (gray = other transcripts; blue = ENST00000558284.1). Differential expression tested using DESeq2 with negative binomial GLMs, technical variation estimated with RUVr ($k = 2$), two-sided Wald test, Benjamini-Hochberg FDR correction, adjusting for gestational age and infant sex. Data presented as $\log_2$ fold change ± standard error. **C** Log_2-transformed read counts for *CSH1* FSM transcripts per sample in GUSTO. ENST00000558284.1 highlighted by GDM status (blue = GDM, red = control); dashed line indicates perfect correlation. **D** Alluvial diagram of *CSH1* read mapping transitions from GENCODE v45 (left) to lr-assembly (right) in GUSTO ($n = 10$ randomly selected samples: 5 GDM, 5 control). Y-axis: square root of read count. Flow colors indicate lr-assembly destination: red = ENST00000558284.1, teal = FSM, purple = NIC, pink = NNC, dark gray = non-*CSH1*. **E** Differential expression of *GNAS* FSM ($n = 17$ transcripts) in GUSTO; plot structure as in (**B**). ENST00000476196.5 highlighted (green = lr-assembly axis, blue = GENCODE v45 axis); statistical testing as in (**B**). **F** Log_2-transformed read counts for *GNAS* FSM transcripts in GUSTO. ENST00000476196.5 highlighted by GDM status (blue = GDM, red = control). **G** *GNAS* read mapping transitions from GENCODE v45 to lr-assembly; flow colors: red = ENST00000476196.5, teal = FSM, dark gray = non-*GNAS*. Source data are provided as a source data file.

exhibits transcriptional breadth and complexity comparable to other adult tissues while displaying relatively high transcriptional complexity of specific pregnancy-related genes, challenging previous characterizations of the placenta as a transcriptomic void[21,23,60]. 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 relatively high transcriptional complexity we observe may be facilitated by the placenta's unique DNA methylation landscape[76–78]. Importantly, transcriptional complexity and breadth metrics are sensitive to sequencing depth, underscoring the importance of adequate sequencing depth and matched processing pipelines for robust cross-tissue comparisons.

Current placental studies involving transcriptome evaluations rely on gene-level expression measurements, obscuring critical isoform-level regulatory mechanisms and missing key placental functions. This gene-level approach is particularly problematic when adverse outcomes arising from pathological exposures alter functionally distinct gene isoforms—effects that could be masked by compensatory transcriptional responses maintaining total gene expression levels. Critically, our results demonstrate that the choice of reference annotation fundamentally alters differential expression outcomes, with tissue-specific placental annotations revealing biologically meaningful associations that remain hidden when using generic adult tissue references[79]. Our findings provide empirical support for the benefits of long-read RNA-seq-based isoform-resolved placental analysis by identifying specific placental transcripts, particularly previously unannotated *CSH1* isoforms, that mediate the relationship between maternal gestational diabetes mellitus (GDM) and newborn birth weight. By demonstrating that tissue-matched annotations, achieved through both transcript discovery and removal of low-confidence transcripts, substantially reduce inferential uncertainty while increasing the interpretability and biological relevance of transcriptomic associations, our work establishes annotation choice as a critical determinant of scientific conclusions in placental genomics and, likely, other tissues.

Our placenta reference transcriptome enables higher-resolution detection of GDM-associated transcriptomic changes. Key placental hormone regulators (*CSH1*, *CSH1L*, *PSG5*, *PSG6*, *GH2*) exhibit high transcriptional complexity, providing flexibility in maternal-fetal hormone signaling. By reducing inferential uncertainty in RNA-seq data[25,26], we revealed GDM-linked regulatory changes missed by conventional gene-level approaches[35–41]. Pathway enrichment analyses using our placenta reference transcriptome revealed GDM-associated transcripts enriched for receptor-ligand activity, hormone receptor binding and growth factor signaling, consistent with endocrine and paracrine roles of the placenta in orchestrating maternal metabolic adaptations and fetal growth trajectories[8,9]. This isoform-resolved precision is critical for identifying early biomarkers of maternal-fetal maladaptation and designing interventions, underscoring the importance of tissue-specific transcriptomic references in developmental contexts[20–22].

Notably, our approach revealed 35.8% of the GDM-birth weight effect was mediated through placental transcription, with population-specific variation in previously unannotated *CSH1* isoforms mediating GDM effects across cohorts. *CSH1* encodes human placental lactogen (hPL), a central endocrine regulator of maternal insulin resistance and nutrient partitioning[10]. The previously unannotated *CSH1* isoforms we identified contain intron retention events (Fig. 4B), a form of alternative splicing increasingly recognized as an important regulatory mechanism of tissue-specific gene expression[80,81], particularly during cellular differentiation[82] and in response to cellular stress[83] including aging-induced transposable element expression[84–86].

These findings reinforce the DOHaD hypothesis by identifying isoform-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—studies which measured aggregate hormone levels without accounting for functionally distinct isoforms that may have differential metabolic effects[42,87,88]. We note that causal mediation analysis assumes no unmeasured exposure-mediator or mediator-outcome confounding, conditional on measured covariates. While this assumption cannot be empirically verified in our study, we minimized potential confounding by adjusting for maternal ethnicity, infant sex, and gestational age. Future applications could employ molecular negative control approaches, such as using maternal epigenomic data as proxies for unmeasured confounders[89], though such methods require additional structural assumptions.

Our long-read cohort ($n = 72$) of term placenta 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/South Asian and white European). However, several limitations warrant consideration. The composition of our long-read samples was of Singaporean ancestry, which may limit the capture of ancestry-specific isoforms expressed in other populations. Expanding long-read sequencing to encompass diverse population ancestries will be essential to comprehensively capture population-specific isoform variation and ensure equitable representation in reference transcriptomes. Similarly, future work should expand upon this reference by incorporating placental tissue across stages of gestation and by drawing from larger, more diverse study cohorts with adequate

sequencing depth, which will be essential for capturing the dynamic regulation of placental transcription throughout development.

An inherent limitation of bulk tissue analysis is that placental tissue comprises multiple cell types (trophoblasts, endothelial cells, stromal cells, and immune cells), potentially obscuring cell-type-specific isoform expression. The isoform diversity we observe likely represents aggregate expression across these cell types, and cell-type-specific isoform switching or differential isoform usage cannot be resolved in our current dataset. Additionally, while long-read sequencing excels at transcript structure discovery, it has lower throughput than short-read sequencing, which we addressed by using our long-read-defined transcriptome as a reference for isoform-level quantification in larger cohorts with short-read sequencing data. Expanding long-read sequencing to additional populations, trimesters, and cell types, including single-cell or spatial transcriptomics approaches, will reveal additional population-, pregnancy stage-, and cell-type-specific isoforms not captured in our current reference.

Experimental validation is also necessary to confirm that the previously unannotated 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. Finally, while our long-read sequencing approach provides high-confidence transcript structures, future validation studies could employ PCR-free methods such as Oxford Nanopore direct RNA sequencing or spatial transcriptomics approaches to provide independent technical confirmation and tissue localization of previously unannotated isoforms.

We also acknowledge that RNA integrity in term placental samples collected under clinical conditions is relatively but expectedly lower than standard quality thresholds. Placental samples must be assessed by a pathologist before collection and transport to the core facility or laboratory handling the RNA extractions, a process that can last 3–4 h and result in RIN values below the typical threshold for library preparation[90]. While our samples showed moderate RNA degradation (long-read samples, mean RIN = 6.95 ± 0.77; GUSTO short-read samples, mean = 3.56 ± 1.60; Gen3G short-read samples, mean = 8.3 ± 0.90), RIN values did not differ significantly by GDM status (Supplementary Data 15), ensuring that our differential expression and mediation findings are not confounded by differences in RNA quality. Long-read sequencing is also robust to moderate RNA degradation, and our stringent filtering criteria ensured that only well-supported transcripts were included in downstream analyses.

Our placental transcriptome reference will enable more precise investigation of environmental exposure effects by identifying mechanistic pathways mediating exposure-outcome relationships, complementing Mendelian randomization approaches by focusing on intermediate biological mechanisms. This mechanistic precision is particularly valuable for studying environmental exposures like maternal diabetes, where effects may operate through distinct gene isoforms with opposing functions. Our tissue-specific reference facilitates fine-mapping of quantitative trait loci (QTLs) and enables more precise prioritization of causal isoforms for targeted functional validation[5,36–38,91,92]. Furthermore, our demonstration of population-specific transcript mediators of gestational exposures on birth outcomes establishes a framework for investigating differential environmental exposure effects by ancestry, enabling identification of population-tailored intervention targets and biomarkers. This resource will also support investigations of imprinting and parent-of-origin effects[93], which are enriched in placental tissues and contribute substantially to pregnancy outcome variation[75,94–96], and enhance comparative evolutionary transcriptomics by improving identification of lineage-specific innovations such as transposon exonization[97,98] and transposon-driven alternative promoters[99,100] that have shaped placental evolution[33,85,86,101]. To facilitate exploration of our placental transcriptome resource, we developed an interactive Shiny application (https://github.com/sbresnahan/lr-placenta-transcriptome-viz) enabling visualization of transcript structures, expression patterns, and differential expression and mediation analysis results.

Methods

Ethics approval

This research was approved by the National Healthcare Group Domain Specific Review Board, the SingHealth Centralized Institutional Review Board, and the ethical committee of the Center Hospitalier Universitaire de Sherbrooke (CHUS), and complies with the Declaration of Helsinki. All three cohorts recruited pregnant individuals assigned female sex at birth. While we recognize that not all pregnant individuals identify as women, analyses are based on placental samples from these participants and their newborns who were assigned male or female sex at birth. Informed written consent was obtained from all mothers at recruitment, and all data are published in accordance with these approvals and consent terms. Tissue donors for long-read sequencing (GUSTO and NUH cohorts) and GUSTO donors for short-read sequencing were compensated for their participation, while Gen3G donors for short-read sequencing were not compensated.

Study populations and cohorts

Samples were collected from term, live births without known placentation pathologies, such as intrauterine growth restriction and pre-eclampsia. Samples for long-read sequencing were collected from a subset of two prospective prebirth cohorts ($n = 72$; maternal age at delivery: 33.1 ± 3.5 years; maternal ethnicity: 54% Chinese, 33% Indian, 13% Malay; gestational age at delivery: 38.6 ± 0.6 weeks; infant sex: 40 male and 32 female [determined at birth by clinical assessment]; birth weight: 3.3 ± 0.3 kg) with 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)[28,52,102]. Of these $n = 72$ samples, $n = 6$ were selected from the Growing Up in Singapore Toward Healthy Outcomes (GUSTO[28,102]) cohort and have matched Illumina short-read libraries from the same biological specimens, while the remaining $n = 66$ were selected from an independent National University Hospital of Singapore (NUH[52]) cohort.

Samples for short-read sequencing were collected from GUSTO ($n = 200$; maternal age at delivery: 32.8 ± 4.8 years; maternal ethnicity: 70% Chinese, 15% Indian, 15% Malay; gestational age at delivery: 38.7 ± 1.2 weeks; infant sex: 106 male, 94 female [determined at birth by clinical assessment]; birth weight: 3.1 ± 0.4 kg) and Genetics of Glucose Regulation in Gestation and Growth (Gen3G[29,103], $n = 152$; maternal age at delivery: 28.6 ± 4.3 years; maternal ethnicity: white European; gestational age at delivery: 39.3 ± 1.1 weeks; infant sex: 79 male and 73 female [determined at birth by clinical assessment]; birth weight: 3.4 ± 0.4 kg), each with 20% of samples from pregnancies affected by GDM.

Sample collection

Placentae were processed within 1 h of delivery. Five villous biopsies of each placenta were rinsed with phosphate buffer saline (PBS), snap-frozen with liquid nitrogen and stored at −80 °C before RNA extraction. For the $n = 6$ GUSTO samples with long-read sequencing, RNA was extracted from approximately 100 mg of crushed tissue using phenol-chloroform and purified with a Qiagen RNeasy Mini Kit. For the $n = 66$ NUH samples, RNA was extracted following the same protocol as for the $n = 6$ from GUSTO, using approximately 100 mg of crushed tissue using phenol-chloroform and purified with a Qiagen RNeasy Mini Kit. Tissue samples were consumed during these studies and are no longer available.

Paired-end Illumina libraries of villous placenta from term, live births from GUSTO and Gen3G were generated from placental villous tissue samples collected after delivery and stored at −80 °C. For GUSTO Illumina libraries, RNA was extracted using phenol-chloroform followed by purification with the Qiagen RNeasy Mini Kit. RNA quality was assessed by Nanodrop spectrophotometry and Agilent TapeStation, with ribosomal RNA depletion performed using Illumina Ribo-Zero kits. For Gen3G, RNA was extracted using phenol-chloroform followed by purification with the NucleoSpin RNA kit (Macherey-Nagel). RNA quality was assessed using an Agilent Bioanalyzer.

RNA sequencing

Extracted samples from the long-read cohort were sequenced via 3 × 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 1000 bp (vs. 150 bp via NovoSeq). In GUSTO, short-read 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, short-read sequencing libraries were prepared from 250 ng of RNA using Illumina TruSeq Stranded mRNA Sample Preparation Kits and sent to the Broad Institute for sequencing 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. Pychopper v2.7.9 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 micro-indels and sequencing errors were corrected with FMLRC2[104] v0.1.8 using short-read RNA-seq samples of villous placenta from GUSTO[28]. The corrected reads were aligned to the human genome assembly analysis set (GenBank assembly 000001405.15, no ALT contigs, hg38) using Minimap2[105] 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 v1.9[106]. 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 v1.6.0, which shows favorable balance in maximizing discovery while limiting false discoveries[49].

Transcriptome annotation and filtering

The assembled transcript models were filtered using a custom R[107] script to remove isoforms with un-stranded exons (retaining only forward or reverse 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[108] v5.4 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 v5.7.1[109]. Coding potential was calculated with CPC v2.0[62]. For protein validation, we performed BLASTP[63] v2.17.0 searches of predicted peptide sequences against two databases: (1) peptides identified by tandem mass spectrometry (MS/MS) of placental tissue samples, including from GDM-affected pregnancies[64,65], and the full UniProt human proteome. To ensure isoform-specific validation, we extracted the matched peptide sequences from each BLASTP hit and identified peptides unique to individual isoforms within each gene. Only transcripts with one unique peptide match (> 95% sequence identity) and CPC2 coding probability > 0.5 were considered validated. Tissue-specific gene enrichment was assessed with TissueEnrich v1.26.0[110].

For quality control, 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 (Supplementary Data 16). We defined high-confidence 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[111], 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[112], within 100 bp of a reference TTS, or within 40 bp of a PolyA motif in the PolyA_DB dataset[113]. 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[108]. For specific previously unannotated isoforms (excluding FSM, ISM and NIC), we applied additional filters: removing transcripts with evidence of reverse transcriptase (RT) switching at splice junctions (a known technical artifact) and applying additional scrutiny to transcripts predicted to undergo nonsense-mediated decay (NMD) including substantial 5′ expression (defined as a ratio of > 0.5:1.0 between short-reads mapping to the first and last splice junctions) to distinguish genuine NMD substrates from assembly artifacts.

Preparation of orthogonal datasets for assembly filtering

Short-read RNA-seq reads from term, villous placenta from GUSTO were aligned to hg38 using STAR[114] v2.7.11b 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[111], trimmed with cutadapt v4.1[115] to remove linker adapters, EcoP15 and 5′poly-G sequences, and filtered reads were aligned to h38 using STAR with settings recommended by ENCODE (Supplementary Data 16). CAGE tags were counted and clustered using the CAGEr[116] v2.16.0 package 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 v2.31.0[117]. SAPAS peaks in placenta tissue mapped to hg38 were retrieved in BED format from APASdb[112], and human polyA motifs annotated in PolyA_DB[113] were retrieved from PolyA-miner[118].

Comparative analysis of GTEx tissues with matched processing pipeline

To enable direct comparison of placental transcriptional complexity with other human tissues, we reprocessed 15 GTEx[47] v9 tissues and cell lines from raw Oxford Nanopore Technologies (ONT) long-read RNA sequencing fastq files (retrieved from dbGaP, phs000424.v9.p2) using an identical bioinformatic pipeline to our placental samples. Assembled transcripts were annotated using SQANTI3 with the same reference files and parameters as placental samples, including GENCODE v45 comprehensive annotation and the human polyA motifs database. For splice junction validation and transcript filtering, we utilized tissue-matched short-read RNA-seq samples from GTEx (retrieved from dbGaP,

phs000424.v9.p2) and orthogonal datasets from multiple public repositories (FANTOM5, ENCODE, SRA, PolyA-Site) employing diverse sequencing methods (CAGE, RAMPAGE, 3'-seq, SAPAS, PolyA-seq, DRS) (Supplementary Data 16). These orthogonal datasets were processed identically to placental validation data. We applied identical SQANTI3 filtering criteria to all tissues to retain high-confidence isoforms. Long-read sample sizes, read depth metrics, and assembly statistics for all GTEx assemblies are provided in Supplementary Data 17.

Short-read RNA-seq analyses

Paired-end Illumina libraries of villous placenta from term, live births of the GUSTO and Gen3G were sequenced previously. We used fastp[119] v1.0.1 to trim sequencing adapters and filter low-quality reads. Decoy-aware transcriptome indices for Salmon[120] v1.10.2 for GENCODE v45 transcripts, our long-read defined placental transcripts (lr-assembly), and their union (GENCODE+) were constructed with MashMap v3.1.3[121]. Transcriptome quantification was then performed with Salmon with settings --libType A --validateMappings --seqBias and including 50 bootstrap replicates to estimate:

$$E_{i,j}^{(k)} \tag{1}$$

the expression E of transcript T_j in sample S_i for B_k bootstrap replicates, where $j = 1,\ldots,n$ transcripts in the given assembly, $i = 1,\ldots,n$ short-read samples, and $k = 1,\ldots,50$ replicates. Additional quantifications were performed against the long-read transcript assemblies of GTEx[47] v9 subcutaneous adipose tissue and cultured fibroblast cells following the same procedure described above.

Transcript quantifications were loaded into R with tximport v1.34.0[122], and overdispersion arising from read-to-transcript mapping ambiguity was estimated with catchSalmon() from the edgeR v4.4.2 package[25] and divided from the counts[25] prior to downstream analyses. To reduce noise in downstream analyses, we filtered the expression matrices to retain only transcripts with at least 5 read counts in a minimum of 3 samples. Pearson correlation coefficients were calculated for each pair of samples within each cohort on the transcripts-per-million (TPM) normalized counts.

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 v1.45.3[123]. Models for all analyses adjusted for gestational age (scaled), infant sex and 2 dimensions of technical variation estimated with RUVr from the RUVSeq v1.40.0 package[72]. Maternal sex was not treated as a variable as all participants were female. Sex-stratified analyses were not a primary aim of this study, and the cohorts were underpowered for such analyses. Associated sample covariates from the Gen3G study are available via dbGaP (accession phs003151.v1.p1) and from the GUSTO study under restricted access (see Data Availability). Expression count matrices are available in a Zenodo repository (Data Availability), enabling sex-disaggregated analyses by qualified researchers with appropriate data access permissions. Isoform-level mediation analyses[124] of the GDM-associated differentially expressed transcripts at Benjamini-Hochberg adjusted p-value (false discovery rate, FDR) < 10% were performed with the mediation v4.5.1 package[125] 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 v0.6 package[126] selecting as mediators, for equal comparison, the first *PC* principal components[73] that cumulatively explained at least 75% of the total variance in the VST-normalized counts of the GDM-associated transcripts (DTEs). Models estimated the total effect of GDM and the indirect effect of placental transcription on birth weight, adjusting for gestational age and infant sex. To assess potential confounding by population structure, we used one-way ANOVA to test for associations between each PC and maternal ethnicity, calculating p-values and eta-squared effect sizes using the effectsize v1.0.1 package[127] to quantify association strength.

Transcript-level Inferential relative variance (InfRV) across median-ratio scaled bootstrap replicates[71] was calculated as:

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

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[128] v0.7.19 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.

Statistics and reproducibility

No statistical method was used to predetermine sample size; the long-read cohort ($n = 72$) represents the largest available collection of placental long-read RNA-seq samples, and short-read cohorts (GUSTO, $n = 200$; Gen3G, $n = 152$) were previously sequenced. Samples without known placentation pathologies were included. For analyses of GDM effects, we excluded eleven GUSTO samples: GDM status was not available for eight samples, and three were identified as outliers. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Statistical analyses were performed in R. Differential transcript and gene expression analyses were conducted with DESeq2, adjusting for gestational age, infant sex, and 2 dimensions of technical variation estimated with RUVr. Multiple testing correction used the Benjamini-Hochberg procedure (FDR < 10%). Isoform-level mediation analyses were performed with the mediation package and multi-mediator analyses with lavaan. Inferential relative variance (InfRV) was calculated across 50 bootstrap replicates from Salmon quantification using edgeR::catchSalmon() to assess isoform quantification uncertainty. All attempts at replication were successful.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Datasets required to reproduce the analyses conducted in this study are available in the associated Supplementary Data or have been deposited in a Zenodo repository under https://doi.org/10.5281/zenodo.18841087. The repository includes the high-confidence placenta and GTEx v9 reference transcriptome GTFs, SQANTI3 classification tables, and the raw and InfRV-adjusted expression matrices for the GUSTO and Gen3G short-read datasets, quantified against each assembly described in this study. Oxford Nanopore long reads generated in this study are publicly accessible through the NCBI Sequence

Read Archive (BioProject accession PRJNA1373130). All accessions for validation datasets used in transcriptome annotation filtering are available in Supplementary Dataset 16. An interactive Shiny application for visualizing transcript structures and exploring differential expression and GDM–birth weight mediation results is available at https://github.com/sbresnahan/lr-placenta-transcriptome-viz. GTEx v9 Oxford Nanopore long-read and Illumina short-read sequencing data are available under controlled access via dbGaP (accession phs000424.v9.p2 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v9.p2]) and on AnVIL (https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V9_hg38), with access granted to authorized researchers for biomedical research purposes as per dbGaP procedures. Short-read data from GTEx tissues were used for isoform quantification and for splice junction validation of transcript models called from the associated long-read data. Illumina short reads from Gen3G, along with associated sample covariates, are available under controlled access via dbGaP (accession phs003151.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003151.v1.p1]); access requests are reviewed by the Gen3G data access committee, with decisions typically communicated within 4–6 weeks, and approved researchers must sign a data use agreement restricting use to the approved research purpose. Illumina short reads from GUSTO, along with associated sample covariates, are available under restricted access due to privacy and regulatory requirements governing human participant data; access requests should be submitted via https://gustodatavault.sg/about/request-for-data and will be reviewed by the GUSTO data access committee within 4–8 weeks, with approved researchers required to sign a data use agreement limiting use to the specified research purpose and prohibiting participant re-identification. Source data for graphs presented in the main text and supplementary materials are provided with this paper. Source data are provided with this paper.

Code availability

All custom scripts and analysis pipelines used in this study are publicly available at https://github.com/sbresnahan/lr-placenta-transcriptome and archived at https://doi.org/10.5281/zenodo.18841087. This repository includes code for dataset preparation, Oxford Nanopore data processing and transcriptome assembly using ESPRESSO, SQANTI3-based annotation and quality control filtering, and short-read analyses in prospective prebirth cohorts including differential gene and transcript expression, mediation analysis, and GO enrichment analysis.

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. *Diab. 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. *Diab. 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 Diab. 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 diff erential testing using inferential uncertainty for RNA-seq. *Genome Res.* **35**, 2326–2338 (2025).
27. Santos Jr, H. P. et al. Evidence for the placenta-brain axis: multiomic 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.* **10**, eadf3411 (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. *Diabetes* **51**, 2207–2213 (2002).
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. *Nature* **621**, 857–865 (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. 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).
54. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* **41**, 8220–8236 (2013).
55. Haerty, W. & Ponting, C. P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* **14**, R49 (2013).
56. W.-W. Liang, et al. Transcriptome-scale RNA-targeting CRISPR screens reveal essential lncRNAs in human cells. https://doi.org/10.6084/m9.figshare.30370171 (2024)
57. Schüler, A., Ghanbarian, A. T. & Hurst, L. D. Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol. Biol. Evol.* **31**, 3164–3183 (2014).
58. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front. Genet.* **6**, 2 (2015).
59. Mattick, J. S. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).
60. 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).
61. 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).
62. 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).
63. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
64. Lin, Z. et al. Alternative strategy to explore missing proteins with low molecular weight. *J. Proteome Res.* **18**, 4180–4188 (2019).
65. 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).
66. Rasanen, J. P. et al. Glycosylated fibronectin as a first-trimester biomarker for prediction of gestational diabetes. *Obstet. Gynecol.* **122**, 586–594 (2013).
67. 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**, 63–70 (2003).
68. 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).
69. 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).
70. 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).
71. 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).
72. 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).
73. 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).

74. Zhao, Y., Lindquist, M. A. & Caffo, B. S. Sparse principal component based high-dimensional mediation analysis. *Comput. Stat. Data Anal.* **142**, 106835 (2020).
75. Hanna, C. W. Placental imprinting: emerging mechanisms and functions. *PLOS Genet* **16**, e1008709 (2020).
76. Yuan, V. et al. Cell-specific characterization of the placental methylome. *BMC Genomics* **22**, 6 (2021).
77. Schroeder, D. I. et al. The human placenta methylome. *Proc. Natl. Acad. Sci.* **110**, 6037–6042 (2013).
78. Schroeder, D. I. et al. Early developmental and evolutionary origins of gene body DNA methylation patterns in mammalian placentas. *PLOS Genet* **11**, e1005442 (2015).
79. Cole, N., Wu, W., Head, S. T., Bresnahan, S. T. & Bhattacharya, A. Quantification method affects replicability of eQTL analysis, colocalization, and TWAS. Preprint at *bioRxiv* https://doi.org/10.1101/2025.08.20.671303 (2025).
80. Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene expression programs. *Hum. Genet.* **136**, 1043–1057 (2017).
81. Alharbi, A. B. et al. *Ctcf* haploinsufficiency mediates intron retention in a tissue-specific manner. *RNA Biol.* **18**, 93–103 (2021).
82. 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).
83. Hadar, S., Meller, A., Saida, N. & Shalgi, R. Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention. *iScience* **25**, 105543 (2022).
84. Pabis, K. et al. A concerted increase in readthrough and intron retention drives transposon expression during aging and senescence. *eLife* **12**, RP87811 (2024).
85. 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).
86. Frost, J. M. et al. Regulation of human trophoblast gene expression by endogenous retroviruses. *Nat. Struct. Mol. Biol.* **30**, 527–538 (2023).
87. Li, J. et al. Isoform usage as a distinct regulatory layer driving nutrient-responsive metabolic adaptation. *Cell Metab.* **37**, 772–787.e6 (2025).
88. Verta, J.-P. & Jacobs, A. The role of alternative splicing in adaptation and evolution. *Trends Ecol. Evol.* **37**, 299–308 (2022).
89. Huang, J. Y. Leveraging molecular negative controls for effect estimation in non-randomized human health and disease studies: a demonstrative simulation study. *ICML 2021 Neglected Assumptions in Causal Inference Workshop* (accepted preprint). (2021).
90. Sommer, J. et al. Optimal timing for RNA isolation: recommendations for placenta sample collection under clinical conditions. *Placenta* **161**, 52–54 (2025).
91. Bhattacharya, A. et al. Placental genomics mediates genetic associations with complex health traits and disease. *Nat. Commun.* **13**, 706 (2022).
92. 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).
93. Gardner, A. & Úbeda, F. The meaning of intragenomic conflict. *Nat. Ecol. Evol.* **1**, 1807–1815 (2017).
94. 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).
95. Kobayashi, E. H. et al. Genomic imprinting in human placentation. *Reprod. Med. Biol.* **21**, e12490 (2022).
96. 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).
97. Pasquesi, G. I. M. et al. Regulation of human interferon signaling by transposon exonization. *Cell* **187**, 7621–7636.e19 (2024).
98. Arribas, Y. A. et al. Transposable element exonization generates a reservoir of evolving and functional protein isoforms. *Cell* **187**, 7603–7620.e22 (2024).
99. Miao, B. et al. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol.* **21**, 255 (2020).
100. 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).
101. 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).
102. Fitzgerald, E. et al. Hofbauer cell function in the term placenta associates with adult cardiovascular and depressive outcomes. *Nat. Commun.* **14**, 7120 (2023).
103. 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).
104. Mak, Q. X. C. et al. Polishing de novo nanopore assemblies of bacteria and eukaryotes with FMLRC2. *Mol. Biol. Evol.* **40**, msad048 (2023).
105. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
106. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
107. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2024).
108. 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).
109. 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).
110. Jain, A. & Tuteja, G. TissueEnrich: tissue-specific gene enrichment analysis. *Bioinformatics* **35**, 1966–1967 (2019).
111. Noguchi, S. et al. FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
112. 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).
113. Zhang, H. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res* **33**, D116–D120 (2004).
114. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
115. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
116. 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).
117. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
118. 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).

119. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
120. 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).
121. 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).
122. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016).
123. 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).
124. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **15**, 309–334 (2010).
125. Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. mediation: R Package for Causal Mediation Analysis. *J. Stat. Softw.* **59**, 1–38 (2014).
126. Rosseel, Y. lavaan: an R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36 (2012).
127. Ben-Shachar, M., Lüdecke, D. & Makowski, D. effectsize: estimation of effect size indices and standardized parameters. *J. Open Source Softw.* **5**, 2815 (2020).
128. Vasimuddin, M., 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).

Acknowledgments

The authors would like to thank the GUSTO, Gen3G, and NUH prebirth cohort study participants and their families, and the related staff and study teams, without whom this study would not be possible. We also 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. The GUSTO study is supported by the National Research Foundation (NRF) under its Translational and Clinical Research (TCR) Flagship Program (NMRC/TCR/004-NUS/2008 and NMRC/TCR/012-NUHS/2014 to S.-Y.C.) and the Open Fund-Large Collaborative Grant (MOH-000504 to S.-Y.C.) administered by the Singapore Ministry of Health's National Medical Research Council (NMRC) and the Agency for Science, Technology and Research (A*STAR). Additional funding is provided by the Institute for Human Development and Potential (IHDP), A*STAR. The Gen3G study was initially supported by a Fonds de recherche du Québec—Santé operating grant (20697 to M.-F.H.), Canadian Institute of Health Research (CIHR) operating grants (MOP 115071 to M.-F.H.) and a Diabète Québec grant, and placental short-read RNA sequencing was supported by NICHD (R01HD94150 to M.-F.H.). Placental long-read RNA sequencing was supported by an NMRC Open Fund-Young Individual Research Grant (MOH-000550-00 to J.Y.H.). J.Y.H. 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).

Author contributions

S.T.B., J.Y.H., and A.B. designed the study. M.-F.H. managed the Gen3G study and S.-Y.C. managed the GUSTO study. H.E.J.Y., J.K.Y.C., F.W., and P.-E.J. managed data from these studies. S.T.B. directed bioinformatic and statistical analyses and wrote the initial draft of the manuscript. W.H.W. and S.L. contributed to bioinformatic and statistical analyses. A.N. designed the Shiny app. M.I.L. consulted on bioinformatic and statistical analyses. All authors contributed to the writing and editing of the final manuscript.

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/s41467-026-71303-4.

Correspondence and requests for materials should be addressed to Sean T. Bresnahan.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

[1]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]A*STAR Institute for Human Development and Potential, Singapore, Singapore. [3]Department of BioSciences, Rice University, Houston, TX, USA. [4]Department of Public Health Sciences, The University of Hawai'i at Mānoa, Honolulu, HI, USA. [5]KK Women's and Children's Hospital, Singapore, Singapore. [6]Duke-NUS Medical School, Singapore, Singapore. [7]Department of Biology, Université de Sherbrooke, Sherbrooke, QC, Canada. [8]Diabetes Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [9]Department of Medicine, Université de Sherbrooke, Sherbrooke, QC, Canada. [10]Department of Obstetrics & Gynecology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [11]Departments of Genetics & Biostatistics, University of North Carolina—Chapel Hill, Chapel Hill, NC, USA. [12]Institute for Data Science in Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [13]These authors jointly supervised this work: Jonathan Y. Huang, Arjun Bhattacharya. ✉e-mail: stbresnahan@mdanderson.org