medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Title: Transcriptomic and Metabolomic analyses in Monozygotic and Dizygotic twins.

Running Title: Transcriptomics and metabolomics in MZ and DZ twins

Nikki Hubers[1,2,3], Gabin Drouard[4], Rick Jansen[3,5], René Pool[1, 3], Jouke Jan Hottenga[1], Miina Ollikainen[4,6], Xiaoling Wang[7], Gonneke Willemsen[8], Jaakko Kaprio[4], Dorret I. Boomsma[2,3,9] & Jenny van Dongen[1,2,3]

[1]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[2]Amsterdam Reproduction & Development (AR&D) research institute, Amsterdam, the Netherlands

[3]Amsterdam Public Health research institute, Amsterdam, the Netherlands

[4]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

[5] Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Psychiatry & Amsterdam Neuroscience - Complex Trait Genetics (VUmc) and Mood, Anxiety, Psychosis, Stress & Sleep

[6] Minerva Foundation Institute for Medical Research, Helsinki, Finland.

[7]Georgia Prevention Institute, Medical College of Georgia, Augusta University, Augusta, GA, USA

[8]Faculty of Health, Sport and Wellbeing, Inholland University of Applied Sciences, Haarlem, the Netherlands

[9]Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam, Vrije Universiteit Amsterdam

Corresponding Author:

Nikki Hubers

PhD student, Department of Biological Psychology, Vrije Universiteit Amsterdam

Van der Boechorststraat 7-9, Room MF-H-557, 1081 BT, Amsterdam, The Netherlands

Phone: +31625508825

Email: n.hubers@vu.nl

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Monozygotic (MZ) and dizygotic (DZ) twins are often studied to determine genetic and environmental influences on complex traits, however, the biological mechanisms behind MZ and DZ twinning are not completely understood. Genomic and epigenomic studies have identified SNPs associated with DZ twinning and DNA methylation sites associated with MZ twinning. To enhance the discovery of molecular biomarkers of twinning, we compare transcriptomics and metabolomics data from MZ with those from DZ twins. We compared 42,663 RNA transcripts in 1,453 MZ twins and 1,294 DZ twins from the Netherlands Twin Register (NTR), followed by sex-stratified analyses. The 5% transcripts with lowest p-values were selected for replication analysis in 217 MZ and 158 DZ twins from the older Finnish Twin cohort (FTC). In the NTR sample, we observed upregulations of the protein coding *PURG* gene in MZ twins. The female-only analyses confirmed the *PURG* gene and identified four other genes, while the male-only analyses indicated three other genes associated with either MZ and DZ twinning. Replication results in the FTC, did not confirm *PURG*, but revealed seven differentially expressed genes with nominal significant p-values in both cohorts, none of which have been implicated for twinning before. Pathway enrichment showed differences in expression in the WNT-pathway and cell adhesion processes, which were previously indicated with MZ twinning though an epigenetic study, and the TGF-B pathway known to be associated with DZ twinning through genetic studies. We additionally meta-analyzed 169 serum metabolites from a NMR platform in 2,797 MZ and 2,040 DZ twins from the NTR, FTC and the FinnTwin12 (FT12), and show no metabolomic differences between the MZ and DZ twins. Overall, we identified novel transcriptomics biomarkers of twinning in peripheral blood and provide partial converging evidence for multiple pathways previously identified in the GWAS of DZ and EWAS of MZ twinning.

Keywords: Twinning, Twins, Transcriptomics, Metabolomics, Multi-omics

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Twinning is defined as the process that gives rise to either dizygotic (DZ) or monozygotic (MZ) twins, and in rarer occasions to triplets and other higher-order multiples (van Dongen et al., 2023). MZ and DZ twins are often studied to provide insight into the genetic and environmental influences on complex traits through the classical twin design or other twin-family designs (D. Boomsma et al., 2002; Hagenbeek et al., 2023). However, much remains unclear about the etiology of MZ and DZ twinning. DZ twinning is the result of a spontaneous double ovulation, runs in families and is often studied as a model for super fertility (Beck et al., 2021; D. I. Boomsma, 2020). MZ twins arise after a fertilized egg cell splits, but the mechanisms behind this are still largely a mystery (van Dongen et al., 2021, 2023).

Genome-wide association studies (GWAS) of DZ twinning highlighted multiple genes such as *FSHB*, *FSHR* and *GNRH*, with obvious roles in female reproduction (Mbarek et al., 2016, 2024). An epigenome-wide association study (EWAS) highlighted DNA methylation differences at over 800 sites in blood from MZ twins when compared to DZ twins and singletons (van Dongen et al., 2021). The methylation differences in blood also replicated in buccal cells. Still, mechanisms leading to the formation of twins, in particular MZ twins, are not fully understood. Furthermore, it is unknown whether molecular markers of twinning might be found in other omics layers that have not yet been studied in connection to twinning, including the transcriptome and metabolome. Studying these layers may provide additional insights into the etiology and can lead to the identification of biomarkers for MZ and DZ twinning (Johnson et al., 2016; Soininen et al., 2015; Whipp et al., 2022).

Transcriptomics is the study of all RNA molecules from protein coding to noncoding RNA (Thompson et al., 2016). Data from protein coding RNAs are sometimes integrated with GWAS results to identify novel gene-trait associations (Yin et al., 2022). These so-called transcriptome-wide association studies (TWASs) are usually performed as post-GWAS analyses using databases such as GTEx and tools like SMR/HEIDI (Gamazon et al., 2015; Zhu et al., 2016). In the largest DZ twinning GWAS, eight genes were identified using a TWAS analysis in several female fertility related tissues: *ARL14EP*, *CAPRIN2*, *ZFPM1*, *SMAD3*, *MPPED2-AS*, *GOLGA8T*, *PCBP2*, and *FAM66D*. Twins have also been used to study the heritability of RNA sequence levels (Ouwens et al., 2019), but RNA transcripts have not been used further to investigate the twinning etiologies.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Metabolomics is the study of the small molecules involved in cellular metabolism (Patti et al., 2012). Unlike other omics, metabolomics provides tools to measure biochemical activity directly, by monitoring the substrates and products involved in cellular metabolism. Studies focusing on metabolomics are widely performed in search of biomarkers for physical and mental health conditions, but have never been performed before for twinning (Guijas et al., 2018).

In this study, we compare serum metabolomics and blood transcriptomics profiles between MZ and DZ twins to enhance the discovery of molecular biomarkers of twinning. We include adult participants from the Netherlands Twin Register (NTR) and from two Finnish twin cohorts, FinnTwin12 (FT12) and the older Finnish Twin Cohort (FTC) of whom many have been involved in one of the two previous omics studies into twinning (Kaprio et al., 2019; Ligthart et al., 2019; Mbarek et al., 2024; Rose et al., 2019; van Dongen et al., 2021; Willemsen et al., 2013). For the transcriptomics, we analyzed gene expression levels of 42,663 RNA transcripts in 2,747 participants of the NTR and repeated our analysis in the 5% transcripts with the lowest p-values in 375 participants of the FTC. For the metabolomics, the three cohorts had measurements from the same platform allowing us to meta-analyze 169 metabolite levels from 4,837 DZ and MZ twins.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Participants

NTR

The NTR is a population based cohort that has been collecting data from twins and their families since the 1980's (Ligthart et al., 2019). In 2004 the NTR started a large-scale biological sample collection of nearly 10,000 participants to create a resource for future omics and biomarker studies (Willemsen et al., 2010). During a home visit in the morning, eight tubes of fasting blood and a morning urine sample were collected along with phenotypic information on health, medication use, body composition and smoking. For fertile women samples were obtained, as much as possible, on days 3–5 of their menstrual cycle or in the pill-free week if on oral contraception. RNA data were generated with the Affymetrix U219 array for 3,362 participants of the NTR, of whom 2,828 were twins. From these twins we excluded 81 individuals who were pregnant or did not have complete covariate data leading to a total of 2,747 twins who we included in our study (Table 1). Metabolomics data were collected for 4,227 participants of the NTR of whom 3,638 were twins with complete phenotypic information that could be included in this study. Zygosity of the participants was determined using genotype information, blood group information or multiple survey items (Ligthart et al., 2019).

FTC

One subset of the Finnish Twin Cohort is the older Finnish Twin Cohort (FTC), which includes twins born before 1958 (Wright et al., 2014). Four waves of questionnaires have been sent to unselected members of the cohort in 1975, 1981, 1990 and 2011, respectively (Kaprio et al., 2019). In 2015, a selected subset of these twins came in for measurement of their blood pressure, completed interviews and questionnaires and provided a fasting blood sample for biochemical measures, and samples for omics (Huang et al., 2018). Metabolomic data for 435 participants were included in the study. Out of the 402 peripheral blood samples, high quality RNA was obtained for 391 subjects and 375 twins of them had complete phenotypic data and could be included in the current study (Table 1). Zygosity of the twins was determined by the first questionnaires for twins and confirmed by genotyping.

FT12

FT12 is a population-based longitudinal cohort including twins born in Finland between 1983 and 1987 (Rose et al., 2019). FT12 was designed for baseline assessments at an early age

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

preceding onset of regular exposure to alcohol, tobacco or other substances (Rose et al., 2019). Subsequent waves of follow-up in FT12 at ages 11-12, 14, 17, and 22 years, have created a rich dataset of behavioral assessments from multiple sources across three stages of adolescence and into early adulthood. The "age 22" assessment wave involved 1,347 twin individuals, with 779 individuals attending in-person assessments and thus venous blood plasma samples could be collected (Table 1; (Whipp et al., 2022)). We removed 15 participants from the data due to pregnancy or lipid lowering medication, resulting in a total sample of 764 participants from FT12. Zygosity was determined using questionnaires at baseline and confirmed later by genotyping. No transcriptomics data were available in the FT12.

Omics data

Transcriptomics

NTR

The generation of the gene expression arrays and data in the NTR was described before (Jansen et al., 2014; Wright et al., 2014). Briefly, peripheral venous blood samples were drawn in the morning (7:00—11:00 a.m.) after an overnight fast. Within 20 minutes of sampling, heparinized whole blood was transferred into PAXgene Blood RNA tubes (Qiagen) and stored at -20°C. The frozen PAXgene tubes were shipped to the Rutgers University and DNA Repository (RUCDR, http://www.rucdr.org). RNA was extracted by the Qiagen Universal liquid handling system, as per the manufacturer's protocol. RNA quality and quantity was assessed by Caliper AMS90 with HT DNA5K/RNA LabChips. Samples were hybridized to Affymetrix U219 array plates (GeneTitan) which contains 530,467 probes for 49,293 transcripts. Array hybridization, washing, staining, and scanning were carried out in an Affymetrix GeneTitan System per the manufacturer's protocol. Twin pairs were randomized over the sample plates. Expression data were required to pass standard Affymetrix Expression Console quality metrics before further undergoing quality control. Samples were excluded if they showed sex inconsistency. Log2 transformation and quantile normalization were applied to the gene expression data. In the main analyses we excluded 1,578 transcripts located on the X and Y chromosome, leading to a total of 42,663 RNA transcripts included in main analysis, 44,182 in the female-only analysis and 44,241 in the male-only analysis.

FTC

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

An in-depth description of the transcriptomics measurements of the transcriptomics data collection can be found in Huang et al (2018) (Huang et al., 2018). RNA samples were extracted from the peripheral leukocytes stored in the RNA cell protection reagents (QIAGEN Inc. Valencia, CA, USA) using the miRNeasy Mini Kit (QIAGEN Inc. Valencia, CA, USA). Transcriptome-wide gene expression data were obtained using the Illumina HumanHT-12 v4 Expression BeadChip (Illumina Inc) which targets more than 48,000 transcripts that provide transcriptome-wide coverage of well-characterized genes, gene candidates and splice variants. A block design was used to keep the distributions of sex and zygosity similar across chips (12 samples/chip) with the co-twins assigned to the same chip. The twin pairs were randomly assigned as pairs to the 12 positions on each chip. Quantification was performed employing the Genome-Studio Gene Expression Module and the lumi R package for data preprocessing and quality control. The quality controls consisted of two key steps: 1; Transcripts with detection P-value <0.05 in more than 50% of the samples were defined as "present"; 2; Log2 transformation and quantile normalization were applied to the gene expression data. There were 19,530 transcripts that passed the quality control steps and they were used as indices of gene expression levels in further analyses. For the replication analyses, we selected the top 2,133 (5% lowest p-value fraction) RNA transcripts from the analyses in NTR and selected the same transcript in the FTC data. We selected the probes based on genetic location and/or gene name. A total of 1,325 transcripts shared overlap between the two platforms and were included in the replication analyses.

The analysis of the transcriptomics data of the FTC cohort showed a high inflation (lambda = 1.3) of the p-values (Supplementary figure 1). This is potentially due to the decision to retain MZ and DZ twin pairs on the same RNA chips, as we observe a high correlation of 0.75 between the RNA chips and age (Supplement Figure 2), which can introduce multi-collinearity (*10.4 - Multicollinearity | STAT 462*, n.d.). To address this, we repeated our analyses of zygosity in 1,325 randomly selected RNA (excluding the initial 1,325 probes) transcripts in the FTC data. When repeating the analyses to random RNA transcriptomics data, we do observe several transcripts surpassing our significance threshold after multiple testing correction burden, but when looking at the enrichment analyses we do not observe any pathways associated with MZ or DZ twinning before (Supplementary Figures 3-4).

Metabolomics

In all three cohorts, metabolites were measured on the metabolomics platform from Nightingale (Soininen et al., 2015). The panel included amino acids (alanine, glutamine, histidine,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

isoleucine, leucine, phenylalanine, tyrosine, and valine) and ketone bodies (acetate, acetoacetate, and 3-hydroxybutyrate), lipids and fatty acids. All metabolite data were available in units mmol/L, nm, g/L or µmol/L (Supplementary Table 1). The metabolomics data were pre-processed for each of the three cohorts similarly, but separately.

In the NTR, the metabolomics data were collected in different batches, which were all pre-processed separately (Hagenbeek et al., 2020; Pool et al., 2020). Metabolites were excluded from the analysis when the mean coefficient of variation exceeded 25% and the missing rate exceeded 5%. Metabolite measurements were set to missing if they were below the lower limit of detection or quantification or could be classified as an outlier (five standard deviations greater or smaller than the mean). If missing, the metabolite was imputed with half of the value of this limit, or when this limit was unknown with half of the lowest observed level for this metabolite.

In the FT12 and FTC, the metabolomics data were measured in one batch each. Metabolite values below the limit of detection were reported as missing, as were metabolite values that deviated from the mean by more than 5 standard deviations (SD). Metabolites with more than 10% missing values were discarded. Missing values were imputed in the FT12 and using the sample minimum of each metabolite for which imputation was to be performed. Outlier presence was assessed by verifying that no participant had at least one of its first three principal components greater than or less than 5 SD from the mean.

In all three cohorts, metabolites were normalized by inverse normal rank transformation so that the distribution of metabolites are of mean 0 and variance 1 (Demirkan et al., 2015; Kettunen et al., 2016). In FT12, 119 metabolites could be included and in NTR and FTC all 169 metabolites were included (Supplementary Table 1 & 2).

Statistical analyses

Transcriptomics

We performed logistic regression for each of the 42,663 RNA transcripts in the NTR twin status (MZ twins coded as= 1 and DZ twins coded as= 2) as a predictor. All analyses were performed in R version 4.3.0. The models were run by using Generalized estimating equations (GEE) to account for family structure (package = *geepack*; function = *geelm).* We included sex, biological age at blood sampling, body mass index (BMI) at time of blood sampling, plate, well and cell counts of blood cells including neutrophiles, lymphocytes, monocytes, eosinophils and

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

basophiles, as covariates. Given that transcriptomics data are correlated (Jansen et al., 2014), adjusting p-values to correct for multiple testing was based on the number of independent dimensions in the data. We performed principal component analyses of the transcripts (Jollife & Cadima, 2016), and corrected the transcriptomic analyses of twins for the number of effective tests performed. 2,814 principal components captured 95% of the total variation in the RNA transcript levels (Supplementary Table 3), leading to a correction where significance is reached if a p-value is below 0.05/2,814 = 1.78E-05. In addition, we ran sex stratified analyses without the sex covariate and we performed sensitivity analyses of the female-only analyses in 669 females after excluding all participants taking contraceptives and post-menopausal women (N=1,128). In the analyses of the females we additionally included the 1,519 transcripts on the X chromosome and in the analyses of the males we included the X chromosome transcripts and 59 transcripts of the Y chromosome.

We took forward the 5% fraction of transcripts with the lowest p-value in NTR in the main analyses (2,213 transcripts in 1,876 genes) and searched for matching transcripts in the FTC data. We identified 1,325 transcripts from the Illumina chip for replication, based on the location of the transcripts and the gene in which the transcript is located and included in both platforms. Given that these transcripts have already been preselected, a more stringent Bonferroni correction was used for replication in the FTC, leading to a p-value to be lower than 0.05/1,325=3.77E-05 to be significant.

Metabolomics

We performed logistic regression for each metabolite with twin zygosity as a predictor (MZ and DZ twin, coded 1 and 2), for each cohort separately. The covariates for each cohort comprise age, sex, smoking status, BMI, batch, and additionally only for NTR fasting status (yes/no) and use of lipid lowering medication. All variables were measured at time of blood sampling. All analyses were performed in R version 4.3.0. The models were run by using GEE to account for family structure (package = *geepack*; function = *geelm*). For the meta-analyses of NTR, FT12 and FTC, we applied fixed effect meta-analyses (package = *metafor*; function = *rma*). We estimated the number of effective tests using the method of Matrix Spectral Decomposition (Chen et al., 2022). In total, we included 169 metabolites which resulted in a total of 27 effective tests, leading to a p-value to be lower than 0.05/27 = 1.85E-03 to be significant, based on Bonferroni correction.

Enrichments analyses

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Enrichr

We performed enrichment analysis using Enrichr on the 5% most significant transcripts in NTR and the 19 genes of the 21 significant transcripts in FTC. Enrichr is a gene set search engine that enables the querying of hundreds of thousands of annotated gene sets (Xie et al., 2021). Enrichr uniquely integrates knowledge from many high-profile projects to provide synthesized information about mammalian genes and gene sets. We employed the *enrichR* R package as an R interface to the Enrichr database (Kuleshov et al., 2016). We used the databases Gene Ontology 2023 molecular functions database (GO2023) and KEGG pathway database (KEGG) to look at the enrichment (Consortium et al., 2023; Kanehisa et al., 2023).

BIOS consortium database

We performed a look-up of the 5% most significant transcripts in the NTR data and the significantly different expressed transcripts in FTC in the methylation–expression association (expression quantitative trait methylation; eQTM) database of the BIOS consortium (Bonder et al., 2017; Zhernakova et al., 2016). The eQTM database did not include participants from NTR. The database contains information on associations between DNA methylation levels and gene expression, based on RNA sequencing data from ~2000 whole blood samples. We checked whether the identified RNA transcripts identified in the current study were located in genes that showed overlap with the eQTMs of the CpGs associated with MZ twinning from our earlier EWAS study (van Dongen et al., 2021).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Transcriptomics

We compared 42,663 RNA transcripts between 1,453 MZ and 1,291 DZ twins of the NTR (Table 1). One RNA transcript, located at 11727318_at (chr8:30853321:30854213), was significantly differentially expressed between MZ and DZ twins (higher expression in MZ twins; beta = 1.78E-02; p value = 6.19E-06). The transcript maps to chromosome 8 located in the *PURG* gene (Figure 1; Supplement Figure 5; Supplementary table 4). We performed enrichment analyses in the GO2023 and the KEGG dataset on the genes in which 5% transcripts with the lowest p-values are located (2,133 transcripts in 1,876; Figure 2; Supplementary table 5-6). In the GO2023 database, 56 pathways were significantly enriched in the genes and the most strongly enriched pathway was ”molecular function of the estrogen response element binding”. The KEGG database highlighted 24 significant enrichment of pathways involved in breast cancer and cell adhesion elements, as well as the WNT-signaling pathway. We performed a look-up of the 5% most significant transcripts in the BIOS consortium eQTM database and identified that two transcripts (located in the genes *HTT* and *FANC)* that were significantly associated with DNA methylation of CpG sites in the genes previously found to be differentially methylated in MZ twins (P<3e-08; Supplementary table 7). Three methylation sites in the *HTT* gene (cg01807241, cg02566259 and cg06421590) showed lower methylation in MZ twins while in our result the *HTT* is higher expressed in the MZ twins. In line with this, in the BIOS dataset, lower methylation correlated with higher expression. The methylation site in *FANCC*, cg14127626, also showed less methylation in the BIOS dataset and more gene expression in the MZ twins.

Sex-stratified analyses

We reran the analyses in females and males separately, including the sex chromosomes. In the female-only analyses we included 1,008 MZ twins and 789 DZ twins and 44,182 RNA transcripts. In these analyses the same transcript in *PURG*, 11727318_at (chr8:30853321:30854213), showed significantly higher expression in MZ twins (beta = 2.50E-02; p value = 2.65E-07). In addition, we found four more transcripts, 11745032_a_at (ch4:114821440:114821440), 11736249_x_at (chr15:64657193:64658274), 11732371_at (chr19:51479729:51480947) and 11759996_at (chr5:35960858:35963053) which significantly associated with zygosity in the female-only analyses (Supplementary table 8). These transcripts are of the genes *ARSJ* (beta = -2.30E-02; p value = 1.04E-05), *KIAAA0101* (beta = -3.19E-02;

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

p value = 1.26E-05), *KLK7* (beta = -2.43E-02; p value = 1.42E-05) and *UGT3A1* (beta = -3.81E-02; p value = 1.73E-05), respectively. We further performed enrichment analyses in the GO2023 and the KEGG dataset on the genes in which the 5% transcripts with the lowest p-value are located (2,009 transcripts; Figure 3; Supplementary tables 9-10). In total 44 pathways were enriched in the KEGG dataset and 58 function in GO2023. The strongest enriched pathways were the DNA binding Transcription Activator Activity, RNA Polymerase II-specific and pathways in cancer.

We performed a sensitivity analysis in the NTR female-only sample after removing all females that took contraceptives or indicated to be past menopause (N=1,128) at the time of the biological sampling (new N=669) In this analysis, six genes in the SMAD pathway (*SMAD2*, *SMAD4*, *TGFB1*, *SMAD5*, *TGFBR1*, *SMAD7*), previously indicated for DZ twinning, were downregulated in the female DZ twins based on a nominal significance ($p<0.05$). In addition, the R-SMAD pathway was the most significantly enriched (Supplementary tables 16-19), which was also among the enriched pathways in the female-only analyses (Supplementary table 10).

In the male-only analyses we included 445 MZ twins and 505 DZ twins and 44,241 RNA transcripts, including both sex chromosomes. In this analysis the transcript in *PURG* was not significantly different in MZ and DZ twins (beta = 7.00E-03; p value = 0.260). However, we observed three other transcripts, 11739295_a_at (chr21:45651163: 45655445), 11738638_a_at (chr20:2397878:2413399) and 11726545_x_at (chr12: 100550135: 100550836), that were significantly associated with zygosity in the male-only analyses (Supplementary table 11). These transcripts are located respectively in the *ICOSLG* (beta = 3.93E-02; p value = 4.08E-07), *TGM6* (beta = 3.22E-02; p value = 7.29E-06) and *GOLGA2P5* (beta = 3.43E-02; p value = 1.50E-05) genes. We further performed enrichment analyses in the GO2023 and the KEGG dataset on the genes in which the 5% most significant transcripts are located (2,112 transcripts; Figure 4; Supplementary Tables 12-13). In total 21 pathways were enriched in the KEGG dataset and 55 functions in GO2023. The most strongly enriched pathways were the Ankyrin Binding pathway and the Thyroid hormone signaling pathway.

Replication analyses

Replication analyses were carried out for 1,325 transcripts of the 2,133 (5%) most significant ones in the results in NTR in 375 twins of the FTC cohort (Table 1). We observed that 21 transcripts, from 19 genes, were differentially expressed in the FTC (Supplementary table 14).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Eight of these genes showed lower expression levels in the MZ twins while the other 11 showed higher expression in the MZ twins. We performed enrichment analyses in the GO2023 and KEGG databases on the 19 genes (Figure 5). The GO2023 indicated 10 significant enrichments including the cAMP regulatory function and the R-SMAD pathway (Supplementary table 15). The TGF-B pathway was also enriched in the KEGG database, together with the cell cycle, the WNT-pathway and 7 other pathways (Supplementary table 16). A look-up of the 19 genes of the 21 significant transcripts in the BIOS consortium eQTM database did not reveal associations with CpGs that were previously associated with MZ twinning. The NTR findings for *PURG* were not replicated. For the FTC cohort the most significant transcript (ILMN_7746) lies in the *JUN* gene (beta = -1.165; $p < 1E-10$). Of the 19 genes, 7 (37%) showed a replication of the same direction of effect in the NTR (Table 2).

Metabolomics

We meta-analyzed 169 metabolites in 2,797 MZ twins and 2,040 DZ from the NTR, FT12 and the FTC (Table 1) with a fixed effect model. The results of the individual analyses of each cohort and the meta-analyses are summarized in supplementary table 2. The results show that none of the metabolite levels differed significantly between the MZ and DZ twins after multiple testing correction (alpha=2.96E-04). Five metabolites showed differences between the MZ and DZ twins, on the nominal p-value of 0.05. The levels of glycoprotein acetyls, linoleic acid, omega-6 fatty acids, albumin and phospholipids in medium LDL were lower in the MZ twins. In three out of these five metabolites the effects of the NTR and the FT12 show the same direction, while the FTC shows the opposite effect (Supplementary figure 6). We found that the overall correlations in the betas between the three cohorts were weak with a range from -0.29 to 0.24 (Supplementary Figure 6).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

In this study we compared transcriptomic and metabolomic data between MZ and DZ twins of the NTR, FT12 and the FTC. We show differences in the RNA transcript levels between MZ and DZ twins in genes whose relations to twinning have partially been established through other, previous, omics studies (GWAS, and EWAS) and show that the differences between the transcriptome in MZ and DZ twins depend on sex. We find some overlap in the differentially expressed gene transcripts with the monozygotic twinning epigenetic signature and the enrichment analyses point to pathways, such as cell adhesion, the WNT-pathways and the TGF-B pathway, that have been indicated before with either MZ or DZ twinning (Mbarek et al., 2024; van Dongen et al., 2021).

We found one transcript located in the *PURG* gene to be significantly associated with zygosity in the total NTR sample. When dividing into males and females, we identified an additional seven transcripts (three in males and four in females). None of these eight genes were replicated in the Finnish samples (which contained both men and women), but eight other transcripts located in these seven genes showed nominal significant association with zygosity in both cohorts. These seven genes could potentially serve as biomarkers of DZ or MZ twinning. In the females-only analyses, but not in the males-only analyses, we confirm the finding of *PURG* indicating a sex-specific twinning maker. *PURG* is a gene that codes for Purine Rich Element Binding Protein G, and the direct function of the gene is unknown. It is however highly similar to purine-rich element binding protein A, which is a DNA-binding protein which has been implicated in the control of both DNA replication and transcription (Liu & Johnson, 2002). *PURG* is highly expressed in the brain, endometrium and testis and, in all analyses, showed higher expression in the MZ twins (PURG; Gene - NCBI, n.d.). In the FinnGen GWAS study, *PURG* was also associated with hypertension, BMI and pregnancy related conditions including poor fetal growth (Kurki et al., 2023). Of note, previous studies have shown that mothers of dizygotic twins have a higher BMI compared to mothers of monozygotic twins (Hubers et al TRHG in press).

The four extra genes identified in the females-only analyses all showed increased expression in the DZ twins and all show some relation to female reproduction. *ARSJ and UGT3A1* have a direct link to female fertility as both are involved in the synthesis of hormones (Mackenzie et al., 2008). *ARSJ* is also associated with height in the FinnGen study, possible reflecting the increased body composition of mothers of DZ twins (Kurki et al., 2023; van Dongen et al., 2023). *KIAAA0101* has been associated before with ovarian cancer and *KLK7* has been

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

implicated with litter size in pigs and encodes for a member of the kallikrein subfamily of serine proteases, which have diverse physiological functions and many kallikrein genes are biomarkers for cancer. (Dan & Yonggang, 2020; Jin et al., 2018; Kurki et al., 2023). DZ twinning is often seen as super fertility and these findings further highlight the relation between DZ twinning and fertility. It would be interesting to investigate if these genes are also more upregulated in mothers of DZ twins and if the differences in gene expression are also present around ovulation.

When comparing the results with the TWAS results from the latest GWAS we find *PCBP2* to be higher expressed in MZ twins in the 5% transcripts with lowest p-values of the NTR results (Mbarek et al., 2024). The TWAS analyses were performed in eight tissues mostly relevant to female fertility; breast, hypothalamus, ovary, pituitary gland, testis, uterus, vagina, and whole blood tissue, while we are only focusing on RNA transcript levels in blood (Mbarek et al., 2024). In the TWAS, the whole blood data showed the lowest amount of genes with an association with DZ twinning, hinting that blood probably is not the most relevant tissue to study. Obtaining RNA transcript data from any other tissue at such a large scale is quite challenging and the previously detected DNA methylation signature was observed in both blood and buccal samples from MZ twins (van Dongen et al., 2021). In the current study, we show that differential DNA methylation in monozygotic twins is accompanied by a difference in transcript levels at two genes in blood. Still, investigating the RNA transcript levels of other tissues has potential in understanding the twinning traits.

HTT and *FANCC*, are indicated both in the 5% transcripts with lowest p-values RNA transcripts and in the epigenetic MZ twinning signature (Bonder et al., 2017; van Dongen et al., 2021; Zhernakova et al., 2016). The former gene has been associated with Huntington's diseases and the latter with a process that is activated when DNA replication is blocked called the Fanconi anemia pathway (Gordon & Buchwald, 2013; Jain & Roy, 2023). The directions of effects appear reversed in the two studies, implying that the increased methylation of these genes found by van Dongen et al. (2021) correlated with a lower expression of these genes in the MZ twins as shown in our study (van Dongen et al., 2021).

Additionally, we observe enrichment of the associated RNA transcripts in cell adhesion processes and the WNT-signaling pathways. Both were also enriched in the epigenetic signature and our results further underline the hypothesis that these processes may therefore be important in the etiology of MZ twinning (van Dongen et al., 2021) The WNT-signaling pathway may also be important for DZ twinning as SMAD3 has been hypothesized to influence the WNT-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

signaling through an interaction with β-catenin (Funa et al., 2015; Hernandez Gifford, 2015). *SMAD3* is the second top hit of the DZ twinning GWAS and also plays a central role in the TGF-B pathway as it is needed to translocate transcription factors to the cell nucleus (Hata & Chen, 2016; Mbarek et al., 2024). In our transcriptomics analysis, two genes in the TGF-B pathways pathway, *CDKN2B* and *SKP1*, were respectively lower and higher expressed in DZ twins of the FTC. These genes have not been previously indicated for any type of twinning, but the TGF-B pathway has been implicated for DZ twinning before (Mbarek et al., 2024).

As for metabolites, we do not observe differences between MZ and DZ twins after correcting for multiple testing. We show that the three cohorts did not show strong correlations between coefficients, and even showed a negative correlation between the NTR and FTC, based on analyses in the total sample of males and females combined. The negative correlation could have been caused by menopause or contraceptive effects, which are known to influence metabolite levels (Bot et al., 2020; Hagenbeek et al., 2020) or the small sample of the FTC cohort. Furthermore, since dizygotic twinning depends on a double ovulation, a molecular signature of dizygotic twinning may disappear when ovulation stops beyond the female reproductive age or through the use of contraceptives. In the transcriptomics data, we performed a sensitivity analysis in the NTR female-only sample by removing all females that took contraceptives or indicated to be in menopause at the time of the biological sampling, and saw that the signal became stronger for genes in the SMAD pathway implicated in dizygotic twinning.

A final consideration of our study is that we do not have a non-twin control group. Therefore, for the transcriptome differences we observe between MZ and DZ twins, it is not necessarily evident whether they are connected to MZ or DZ twinning. However, some pathways clearly show overlap with previous DZ twinning GWAS results and thus are more likely driven by differential expression in DZ twins, whereas other transcriptomic results converge with the DNA methylation signature for MZ twinning (Mbarek et al., 2024; van Dongen et al., 2021).

To conclude, our study identifies novel transcriptomic biomarkers for twinning and provides converging evidence for multiple pathways previously identified in the GWAS of DZ and EWAS of MZ twinning. More functional studies on twinning are to be encouraged, and pursuing efforts to include other omics (e.g. proteomics) and other tissues (e.g. ovary, and pre- and perinatal tissues) may pave the way to a better understanding of how these biological pathways are involved in twinning.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Declarations

Ethics approval

The study protocol was approved for the NTR by the Ethical Review Board of the VU University Medical Center and informed consent was obtained from all participants.

Ethical approval for all data collection waves from FT12 was obtained from the ethical committee of the Helsinki and Uusimaa University Hospital District and the Institutional Review Board of Indiana University. All data collection and sampling protocols were performed in compliance with the ethical guidelines. Parents provided consent for the twins aged 12 and 14 years old, while twins aged 17 and 22 years old provided written consent themselves for sample collection.

The study protocol for the FTC was approved by the Institutional Ethics Board of the Hospital District of Helsinki and Uusimaa, Finland (ID 154/13/03/00/11) and the Institutional Review Board of Augusta University.

Availability of data and materials

The data of the Netherlands Twin Register (NTR) may be requested through the NTR data access committee (https://tweelingenregister.vu.nl/information_for_researchers/working-with ntr-data).

The Finnish Twin Cohort data used in the analysis is deposited in the Biobank of the Finnish Institute for Health and Welfare (https://thl.fi/en/web/thl-biobank/forresearchers). It is available to researchers after written application and following the relevant Finnish legislation.

FinnTwin12 data analyzed in this study is not publicly available due to the restrictions of informed consent. Requests to access these datasets should be directed to the Institute for Molecular Medicine Finland (FIMM) Data Access Committee (DAC) (fimmdac@helsinki.fi) for authorized researchers who have IRB/ethics approval and an institutionally approved study plan. To ensure the protection of privacy and compliance with national data protection legislation, a data use/transfer agreement is needed, the content and specific clauses of which will depend on the nature of the requested data.

Competing interest

None of the authors have competing interests.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Funding

NH is supported by the Royal Netherlands Academy of Science Professor Award (PAH/6635) to DIB and the 2023 talent travel grand award by the faculty of behaviour and movement sciences at the Vrije Universiteit Amsterdam. JvD is supported by NWO Large Scale infrastructures, X-omics (184.034.019). The NTR is supported by multiple grants from the Netherlands Organizations for Scientific Research (NOW; 480-15-001/674; 480-04-004; 400-05-717); and Medical Research (ZonMW; 912-10-020); the European Science Council (ERC) Genetics of Mental Illness (ERC Advanced, 230374); Developmental trajectories of psychopathology (NIMH 1RC2 MH089995). The metabolomics data of the NTR was collected through BBMRI: BBMRI.1 (184.021.007) and BBMRI.2 (184.033.111).

The FTC has been supported by the Academy of Finland (Grants 265240, 263278, 308248, 312073, 336832 to Jaakko Kaprio and 297908 to Miina Ollikainen) and the Sigrid Juselius Foundation (to Miina Ollikainen). The transcriptome study in FTC was supported by NIH/NHLBI grant HL104125. Phenotype and genotype data collection in FinnTwin12 cohort has been supported by, ENGAGE – European Network for Genetic and Genomic Epidemiology, FP7-HEALTH-F4-2007, grant agreement number 201413, National Institute of Alcohol Abuse and Alcoholism (grants AA-12502, AA-00145, and AA-09203 to R J Rose; AA15416 and K02AA018755 to D M Dick; R01AA015416 to Jessica Salvatore) and the Academy of Finland (grants 100,499, 205,585, 118,555, 141,054, 264,146, 308,248 to JK, and the Centre of Excellence in Complex Disease Genetics (grants 312,073, 336,823, and 352,792 to JKaprio).

Authors contributions

NH, GD, RP and JvD have contributed to the conceptualization of the study. NH, GD, RJ and RP prepared the data for inclusion. NH has performed the analyses and was responsible for the first draft of the manuscript. RP, JvD, JJH, JK, GW and DIB provided supervision of the analyses and writing. GW and DIB were involved in the initial data collection of the NTR and MO and XW in the data collection of the FTC and FT12. All authors provided feedback and revision on the manuscript and supported the submitted version of the manuscript.

Acknowledgement

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

This work largely builds upon the biobanking efforts performed by the BIOS and the BBMRI consortia and we thank all participants and staff for their contributions. Additionally we thank Dr. Fiona Hagenbeek for providing expertise and feedback throughout the project.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

References

10.4 - Multicollinearity | STAT 462. (n.d.). Retrieved January 29, 2024, from https://online.stat.psu.edu/stat462/node/177/

Beck, J. J., Bruins, S., Mbarek, H., Davies, G. E., & Boomsma, D. I. (2021). Biology and Genetics of Dizygotic and Monozygotic Twinning. *Twin and Higher-Order Pregnancies*, 31–50. https://doi.org/10.1007/978-3-030-47652-6_3/FIGURES/6

Bonder, M. J., Luijk, R., Zhernakova, D. V., Moed, M., Deelen, P., Vermaat, M., Van Iterson, M., Van Dijk, F., Van Galen, M., Bot, J., Slieker, R. C., Jhamai, P. M., Verbiest, M., Suchiman, H. E. D., Verkerk, M., Van Der Breggen, R., Van Rooij, J., Lakenberg, N., Arindrarto, W., … Heijmans, B. T. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, *49*(1), 131–138. https://doi.org/10.1038/NG.3721

Boomsma, D., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews. Genetics*, *3*(11), 872–882. https://doi.org/10.1038/NRG932

Boomsma, D. I. (2020). The Genetics of Human DZ Twinning. *Twin Research and Human Genetics*, *23*(2), 74–76. https://doi.org/10.1017/THG.2020.15

Bot, M., Milaneschi, Y., Al-Shehri, T., Amin, N., Garmaeva, S., Onderwater, G. L. J., Pool, R., Thesing, C. S., Vijfhuizen, L. S., Vogelzangs, N., Arts, I. C. W., Demirkan, A., van Duijn, C., van Greevenbroek, M., van der Kallen, C. J. H., Köhler, S., Ligthart, L., van den Maagdenberg, A. M. J. M., Mook-Kanamori, D. O., … Sattar, N. (2020). Metabolomics Profile in Depression: A Pooled Analysis of 230 Metabolic Markers in 5283 Cases With Depression and 10,145 Controls. *Biological Psychiatry*, *87*(5), 409–418. https://doi.org/10.1016/J.BIOPSYCH.2019.08.016

Chen, Y., Li, E. M., & Xu, L. Y. (2022). Guide to Metabolomics Analysis: A Bioinformatics Workflow. *Metabolites*, *12*(4). https://doi.org/10.3390/METABO12040357

Consortium, T. G. O., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., … Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1). https://doi.org/10.1093/GENETICS/IYAD031

Dan, L., & Yonggang, L. (2020). Molecular characteristics and association analysis with litter size trait for porcine KLK7 gene. *Animal Biotechnology*, *31*(5), 377–381. https://doi.org/10.1080/10495398.2019.1604379

Demirkan, A., Henneman, P., Verhoeven, A., Dharuri, H., Amin, N., van Klinken, J. B., Karssen, L. C., de Vries, B., Meissner, A., Göraler, S., van den Maagdenberg, A. M. J. M., Deelder, A. M., C 't Hoen, P. A., van Duijn, C. M., & van Dijk, K. W. (2015). Insight in Genome-Wide Association of Metabolite Quantitative Traits by Exome Sequence Analyses. *PLoS Genetics*, *11*(1), 1004835. https://doi.org/10.1371/JOURNAL.PGEN.1004835

Funa, N. S., Schachter, K. A., Lerdrup, M., Ekberg, J., Hess, K., Dietrich, N., Honoré, C., Hansen, K., & Semb, H. (2015). β-Catenin Regulates Primitive Streak Induction through Collaborative Interactions with SMAD2/SMAD3 and OCT4. *Cell Stem Cell*, *16*(6), 639–652. https://doi.org/10.1016/J.STEM.2015.03.008

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. https://doi.org/10.1038/NG.3367

Gordon, S. M., & Buchwald, M. (2013). *The FANCC Gene and Its Products*. https://www.ncbi.nlm.nih.gov/books/NBK6419/

Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E., & Siuzdak, G. (2018). Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nature Biotechnology*, *36*(4), 316. https://doi.org/10.1038/NBT.4101

Hagenbeek, F. A., Hirzinger, J. S., Breunig, S., Bruins, S., Kuznetsov, D. V., Schut, K., Odintsova, V. V., & Boomsma, D. I. (2023). Maximizing the value of twin studies in health and behaviour. *Nature Human Behaviour*, *7*(6), 849–860. https://doi.org/10.1038/S41562-023-01609-6

Hagenbeek, F. A., Pool, R., van Dongen, J., Draisma, H. H. M., Jan Hottenga, J., Willemsen, G., Abdellaoui, A., Fedko, I. O., den Braber, A., Visser, P. J., de Geus, E. J. C. N., Willems van Dijk, K., Verhoeven, A., Suchiman, H. E., Beekman, M., Slagboom, P. E., van Duijn, C. M., Barkey Wolf, J. J. H., Cats, D., … Boomsma, D. I. (2020). Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. *Nature Communications*, *11*(1). https://doi.org/10.1038/S41467-019-13770-6

Hata, A., & Chen, Y. G. (2016). TGF-β Signaling from Receptors to Smads. *Cold Spring Harbor Perspectives in Biology*, *8*(9). https://doi.org/10.1101/CSHPERSPECT.A022061

Hernandez Gifford, J. A. (2015). The role of WNT signaling in adult ovarian folliculogenesis. *Reproduction (Cambridge, England)*, *150*(4), R137. https://doi.org/10.1530/REP-14-0685

Huang, Y., Ollikainen, M., Sipil, P., Mustelin, L., Wang, X., Su, S., Huan, T., Levy, D., Wilson, J., Snieder, H., Kaprio, J., & Wang, X. (2018). Genetic and environmental effects on gene expression signatures of blood pressure: a transcriptome-wide twin study. *Hypertension (Dallas, Tex. : 1979)*, *71*(3), 457. https://doi.org/10.1161/HYPERTENSIONAHA.117.10527

Jain, S., & Roy, I. (2023). Aptamer Reduces Aggregation of Mutant Huntingtin and Rescues Proteostasis Network in Non-Neuronal and Neuronal Cells. *ACS Chemical Neuroscience*, *14*(12), 2385–2395. https://doi.org/10.1021/ACSCHEMNEURO.3C00226

Jansen, R., Batista, S., Brooks, A. I., Tischfield, J. A., Willemsen, G., Van Grootheest, G., Hottenga, J. J., Milaneschi, Y., Mbarek, H., Madar, V., Peyrot, W., Vink, J. M., Verweij, C. L., de Geus, E. J. C., Smit, J. H., Wright, F. A., Sullivan, P. F., Boomsma, D. I., & Penninx, B. W. J. H. (2014). Sex differences in the human peripheral blood transcriptome. *BMC Genomics*, *15*(1), 1–12. https://doi.org/10.1186/1471-2164-15-33/FIGURES/3

Jin, C., Liu, Z., Li, Y., Bu, H., Wang, Y., Xu, Y., Qiu, C., Yan, S., Yuan, C., Li, R., Diao, N., Zhang, Z., Wang, X., Liu, L., & Kong, B. (2018). PCNA-associated factor P15PAF, targeted by FOXM1, predicts poor prognosis in high-grade serous ovarian cancer patients. *International Journal of Cancer*, *143*(11), 2973–2984. https://doi.org/10.1002/IJC.31800

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews. Molecular Cell Biology*, *17*(7), 451. https://doi.org/10.1038/NRM.2016.25

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065). https://doi.org/10.1098/RSTA.2015.0202

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, *51*(D1), D587–D592. https://doi.org/10.1093/NAR/GKAC963

Kaprio, J., Bollepalli, S., Buchwald, J., Iso-Markku, P., Korhonen, T., Kovanen, V., Kujala, U., Laakkonen, E. K., Latvala, A., Leskinen, T., Lindgren, N., Ollikainen, M., Piirtola, M., Rantanen, T., Rinne, J., Rose, R. J., Sillanpää, E., Silventoinen, K., Sipilä, S., … Waller, K. (2019). The Older Finnish Twin Cohort - 45 Years of Follow-up. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *22*(4), 240–254. https://doi.org/10.1017/THG.2019.54

Kettunen, J., Demirkan, A., Würtz, P., Draisma, H. H. M., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A. J., Lyytikäinen, L. P., Pirinen, M., Pool, R., Sarin, A. P., Soininen, P., Tukiainen, T., Wang, Q., Tiainen, M., Tynkkynen, T., Amin, N., Zeller, T., … Ala-Korpela, M. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications*, *7*. https://doi.org/10.1038/NCOMMS11122

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Maayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. https://doi.org/10.1093/NAR/GKW377

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., … Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature 2023 613:7944*, *613*(7944), 508–518. https://doi.org/10.1038/s41586-022-05473-8

Ligthart, L., Van Beijsterveldt, C. E. M., Kevenaar, S. T., De Zeeuw, E., Van Bergen, E., Bruins, S., Pool, R., Helmer, Q., Van Dongen, J., Hottenga, J. J., Van'T Ent, D., Dolan, C. V., Davies, G. E., Ehli, E. A., Bartels, M., Willemsen, G., De Geus, E. J. C., & Boomsma, D. I. (2019). The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *22*(6), 623–636. https://doi.org/10.1017/THG.2019.93

Liu, H., & Johnson, E. M. (2002). Distinct proteins encoded by alternative transcripts of the PURG gene, located contrapodal to WRN on chromosome 8, determined by differential termination/polyadenylation. *Nucleic Acids Research*, *30*(11), 2417–2426. https://doi.org/10.1093/NAR/30.11.2417

Mackenzie, P. I., Rogers, A., Treloar, J., Jorgensen, B. R., Miners, J. O., & Meech, R. (2008). Identification of UDP Glycosyltransferase 3A1 as a UDP N-Acetylglucosaminyltransferase. *Journal of Biological Chemistry*, *283*(52), 36205–36210. https://doi.org/10.1074/JBC.M807961200

Mbarek, H., Gordon, S. D., Duffy, D. L., Hubers, N., Mortlock, S., Beck, J. J., Hottenga, J.-J., Pool, R., Dolan, C. V, Actkins, K. V, Gerring, Z. F., Van Dongen, J., Ehli, E. A., Iacono, W. G., Mcgue, M.,

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Chasman, D. I., Gallagher, C. S., Schilit, S. L. P., Morton, C. C., … Martin, N. G. (2024). Genome-wide association study meta-analysis of dizygotic twinning illuminates genetic regulation of female fecundity. *Human Reproduction*, *39*(1), 240–257. https://doi.org/10.1093/HUMREP/DEAD247

Mbarek, H., Steinberg, S., Nyholt, D. R., Gordon, S. D., Miller, M. B., McRae, A. F., Hottenga, J. J., Day, F. R., Willemsen, G., De Geus, E. J., Davies, G. E., Martin, H. C., Penninx, B. W., Jansen, R., McAloney, K., Vink, J. M., Kaprio, J., Plomin, R., Spector, T. D., … Boomsma, D. I. (2016). Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. *American Journal of Human Genetics*, *98*(5), 898–908. https://doi.org/10.1016/J.AJHG.2016.03.008

Ouwens, K. G., Jansen, R., Nivard, M. G., van Dongen, J., Frieser, M. J., Hottenga, J. J., Arindrarto, W., Claringbould, A., van Iterson, M., Mei, H., Franke, L., Heijmans, B. T., A. C. ’t Hoen, P., van Meurs, J., Brooks, A. I., Heijmans, B. T., A. C. ’t Hoen, P., van Meurs, J., Isaacs, A., … Penninx, B. W. J. H. (2019). A characterization of cis- and trans-heritability of RNA-Seq-based gene expression. *European Journal of Human Genetics 2019 28:2*, *28*(2), 253–263. https://doi.org/10.1038/s41431-019-0511-5

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Metabolomics: the apogee of the omic triology. *Nature Reviews. Molecular Cell Biology*, *13*(4), 263. https://doi.org/10.1038/NRM3314

Pool, R., Hagenbeek, F. A., Hendriks, A. M., van Dongen, J., Willemsen, G., de Geus, E., Metabolomics Consortium, B., Willems van Dijk, K., Verhoeven, A., Eka Suchiman, H., Beekman, M., Eline Slagboom, P., Harms, A. C., Hankemeier, T., Boomsma, D. I., D Suchiman, H. E., Amin, N., Beulens, J. W., van der Bom, J. A., … Slagboom, P. (2020). Genetics and Not Shared Environment Explains Familial Resemblance in Adult Metabolomics Data. *Twin Research and Human Genetics*, *23*(3), 145–155. https://doi.org/10.1017/THG.2020.53

PURG purine rich element binding protein G [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved May 15, 2024, from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=29942

Rose, R. J., Salvatore, J. E., Aaltonen, S., Barr, P. B., Bogl, L. H., Byers, H. A., Heikkilä, K., Korhonen, T., Latvala, A., Palviainen, T., Ranjit, A., Whipp, A. M., Pulkkinen, L., Dick, D. M., & Kaprio, J. (2019). FinnTwin12 Cohort: An Updated Review. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *22*(5), 302. https://doi.org/10.1017/THG.2019.83

Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation. Cardiovascular Genetics*, *8*(1), 192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216

Thompson, S. D., Prahalad, S., & Colbert, R. A. (2016). Integrative Genomics. *Textbook of Pediatric Rheumatology*, 43-53.e3. https://doi.org/10.1016/B978-0-323-24145-8.00005-3

van Dongen, J., Gordon, S. D., McRae, A. F., Odintsova, V. V., Mbarek, H., Breeze, C. E., Sugden, K., Lundgren, S., Castillo-Fernandez, J. E., Hannon, E., Moffitt, T. E., Hagenbeek, F. A., van Beijsterveldt, C. E. M., Jan Hottenga, J., Tsai, P. C., van Dongen, J., Hottenga, J. J., McRae, A. F., Sugden, K., … Boomsma, D. I. (2021). Identical twins carry a persistent epigenetic signature of early genome programming. *Nature Communications 2021 12:1*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-25583-7

medRxiv preprint doi: https://doi.org/10.1101/2024.06.25.24309452; this version posted June 26, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

van Dongen, J., Hubers, N., & Boomsma, D. I. (2023). New insights into the (epi)genetics of twinning. *Human Reproduction*, *0*, 1–8. https://doi.org/10.1093/HUMREP/DEAD131

Whipp, A. M., Heinonen-Guzejev, M., Pietiläinen, K. H., van Kamp, I., & Kaprio, J. (2022). Branched-chain amino acids linked to depression in young adults. *Frontiers in Neuroscience*, *16*, 935858. https://doi.org/10.3389/FNINS.2022.935858/BIBTEX

Willemsen, G., De Geus, E. J. C., Bartels, M., Van Beijsterveldt, C. E. M. T., Brooks, A. I., Estourgie-van Burk, G. F., Fugman, D. A., Hoekstra, C., Hottenga, J. J., Kluft, K., Meijer, P., Montgomery, G. W., Rizzu, P., Sondervan, D., Smit, A. B., Spijker, S., Suchiman, H. E. D., Tischfield, J. A., Lehner, T., … Boomsma, D. I. (2010). The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *13*(3), 231–245. https://doi.org/10.1375/TWIN.13.3.231

Willemsen, G., Vink, J. M., Abdellaoui, A., Den Braber, A., Van Beek, J. H. D. A., Draisma, H. H. M., Van Dongen, J., Van 'T Ent, D., Geels, L. M., Van Lien, R., Ligthart, L., Kattenberg, M., Mbarek, H., De Moor, M. H. M., Neijts, M., Pool, R., Stroo, N., Kluft, C., Suchiman, H. E. D., … Boomsma, D. I. (2013). The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *16*(1), 271. https://doi.org/10.1017/THG.2012.140

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y. H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T. H., D'Ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., … Boomsma, D. I. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics 2014 46:5*, *46*(5), 430–437. https://doi.org/10.1038/ng.2951

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3), e90. https://doi.org/10.1002/CPZ1.90

Yin, X., Bose, D., Kwon, A., Hanks, S. C., Jackson, A. U., Stringham, H. M., Welch, R., Oravilahti, A., Fernandes Silva, L., Locke, A. E., Fuchsberger, C., Service, S. K., Erdos, M. R., Bonnycastle, L. L., Kuusisto, J., Stitziel, N. O., Hall, I. M., Morrison, J., Ripatti, S., … Wen, X. (2022). Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. *American Journal of Human Genetics*, *109*(10), 1727. https://doi.org/10.1016/J.AJHG.2022.08.007

Zhernakova, D. V., Deelen, P., Vermaat, M., Van Iterson, M., Van Galen, M., Arindrarto, W., Van't Hof, P., Mei, H., Van Dijk, F., Westra, H. J., Bonder, M. J., Van Rooij, J., Verkerk, M., Jhamai, P. M., Moed, M., Kielbasa, S. M., Bot, J., Nooren, I., Pool, R., … Franke, L. (2016). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics 2016 49:1*, *49*(1), 139–145. https://doi.org/10.1038/ng.3737

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics 2016 48:5*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538

Figure 1: Manhattan plot showing the results of the transcriptomics analyses in the Netherlands Twin Register (NTR) comparing monozygotic and dizygotic twins.

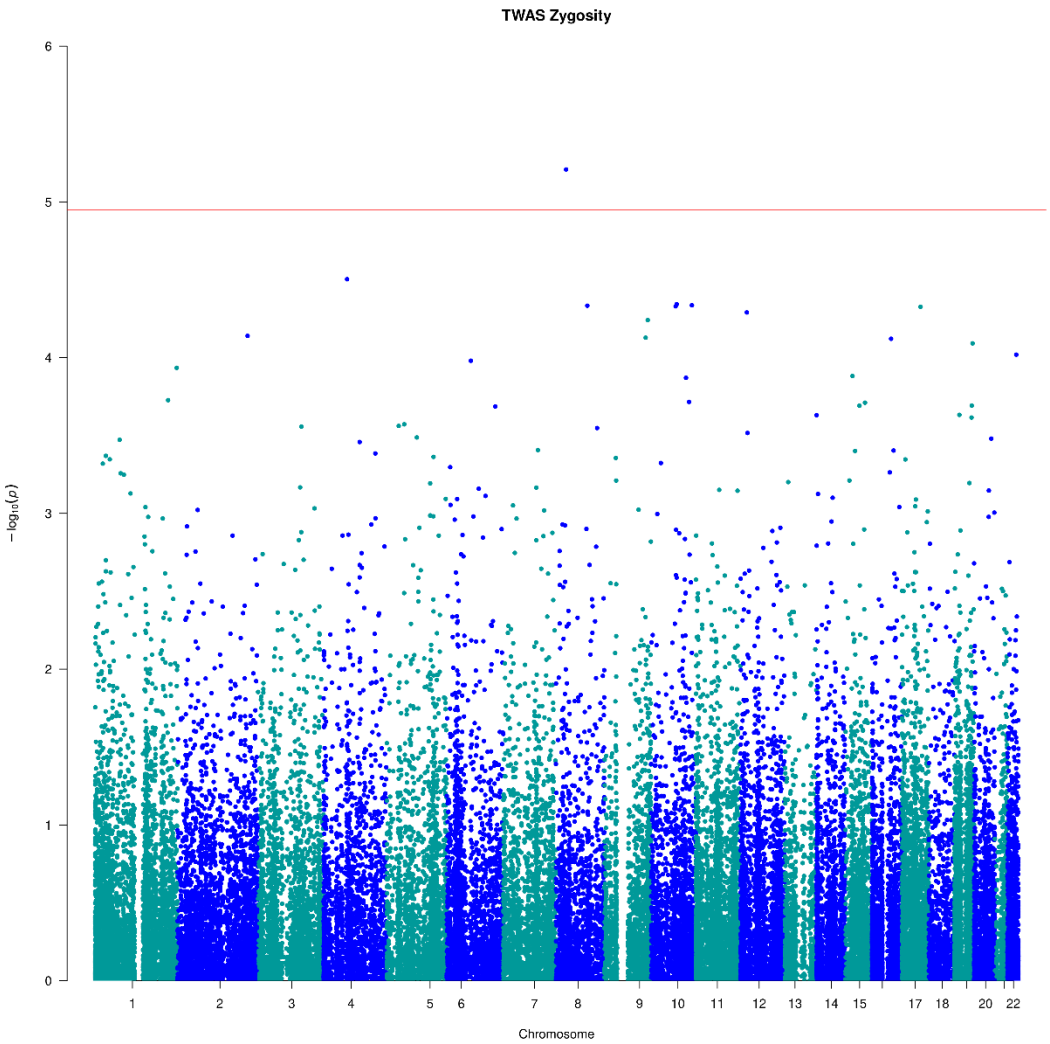


Figure 2: Enrichment analyses of the 5% transcripts with lowest p-values RNA transcript with the lowest p-value in the NTR. The figure only shows the ten most enriched pathways and molecular functions. The left figure shows the enrichment in the Gene Ontology 2023 molecular functions database and the right figure shows the enrichment in the KEGG human pathway database.

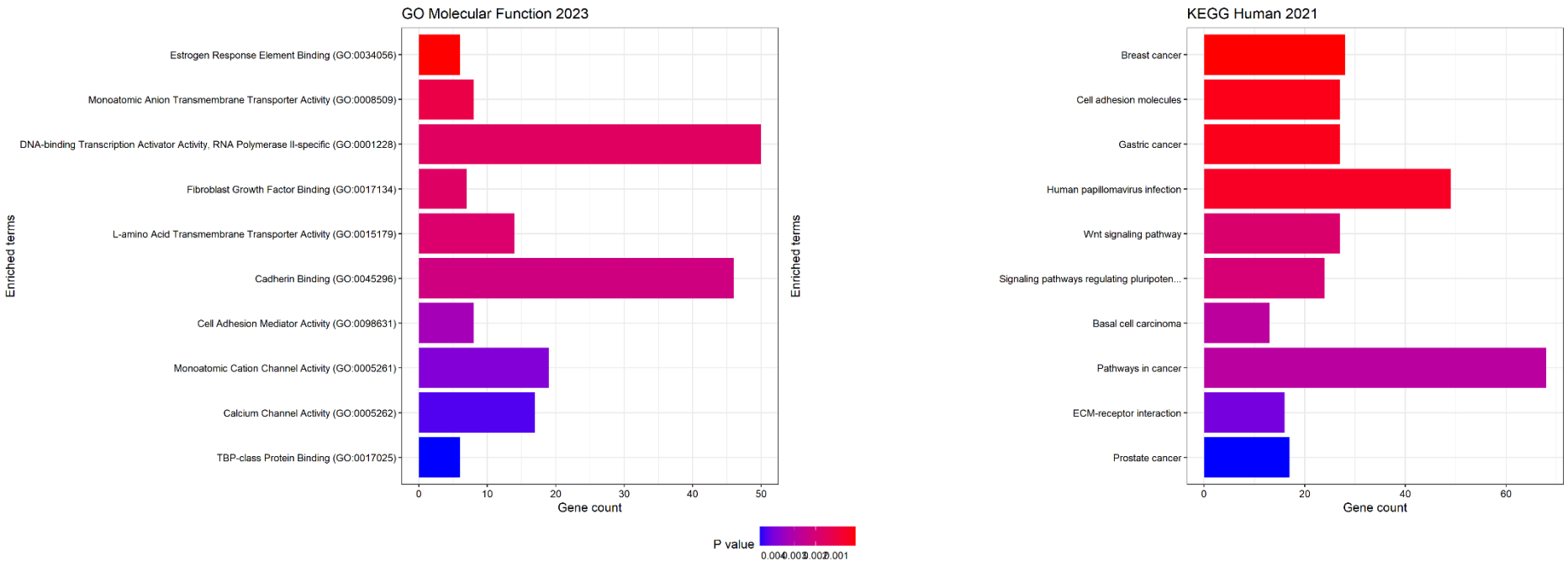


Figure 3: Enrichment analyses of the 5% transcripts with lowest p-values RNA transcript with the lowest p-value in the NTR females. The figure shows the ten most enriched pathways and molecular functions. The left figure shows the enrichment in the Gene Ontology 2023 molecular functions database and the right figure shows the enrichment in the KEGG human pathway database.

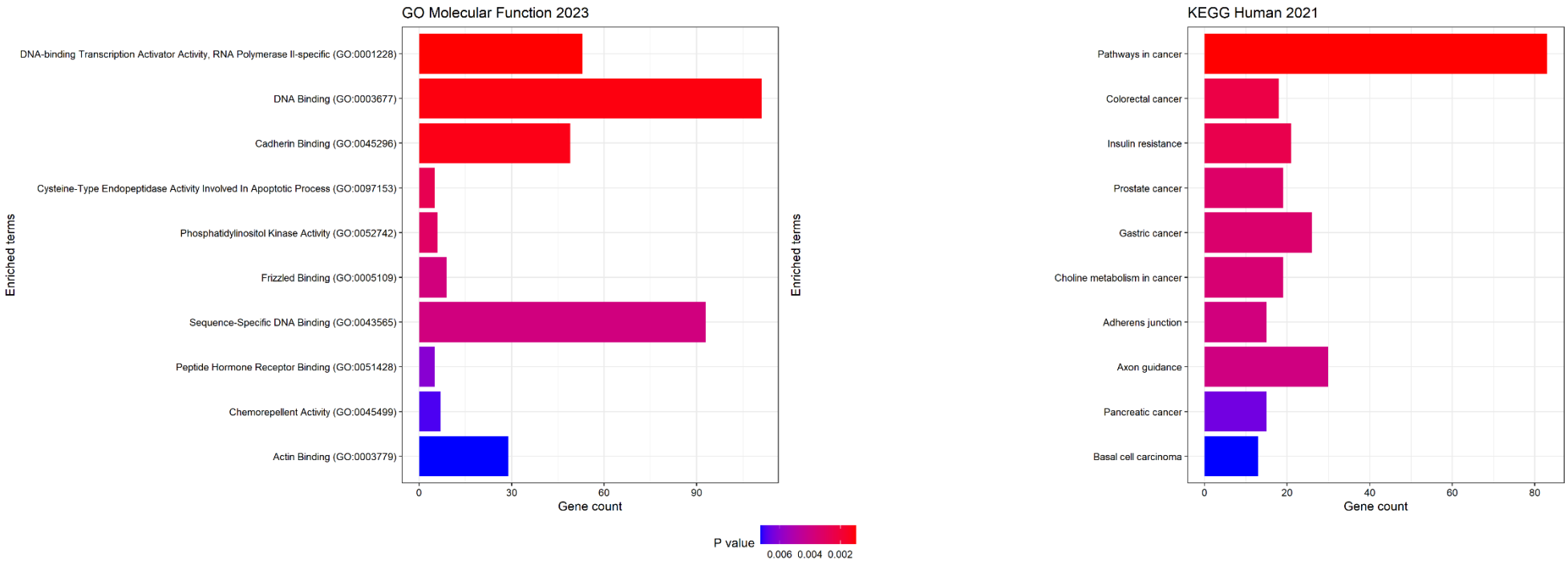
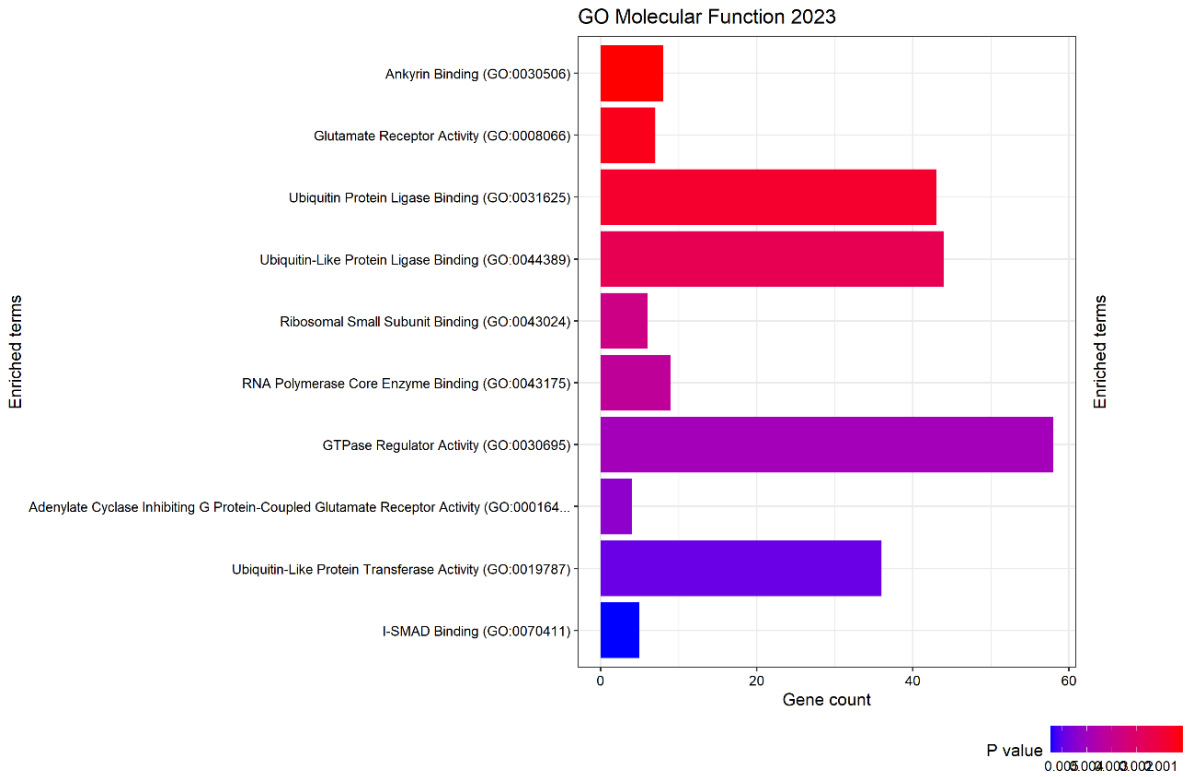


Figure 4: Enrichment analyses of the 5% transcripts with lowest p-values RNA transcript with the lowest p-value in the NTR males. The figure shows the ten most enriched pathways and molecular functions. The left figure shows the enrichment in the Gene Ontology 2023 molecular functions database and the right figure shows the enrichment in the KEGG human pathway database.



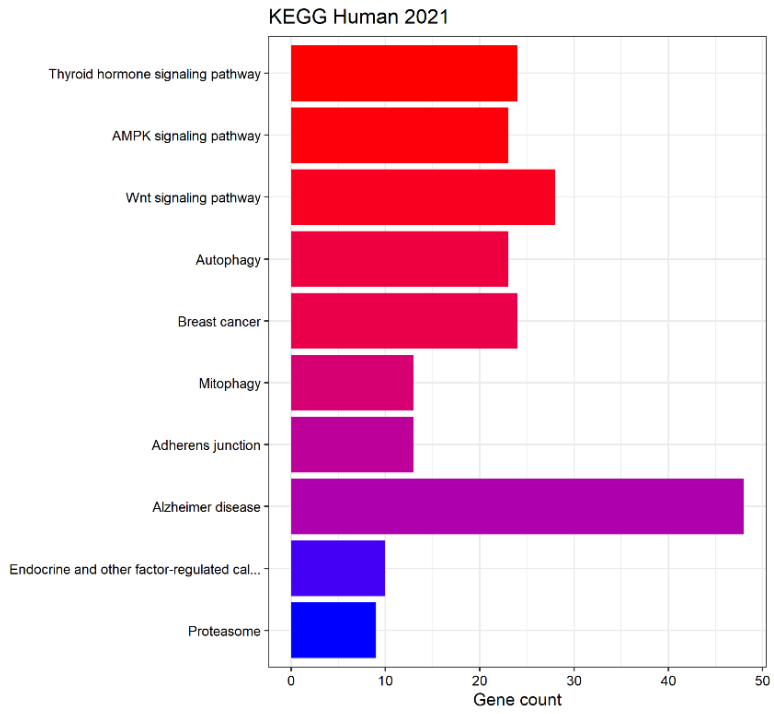
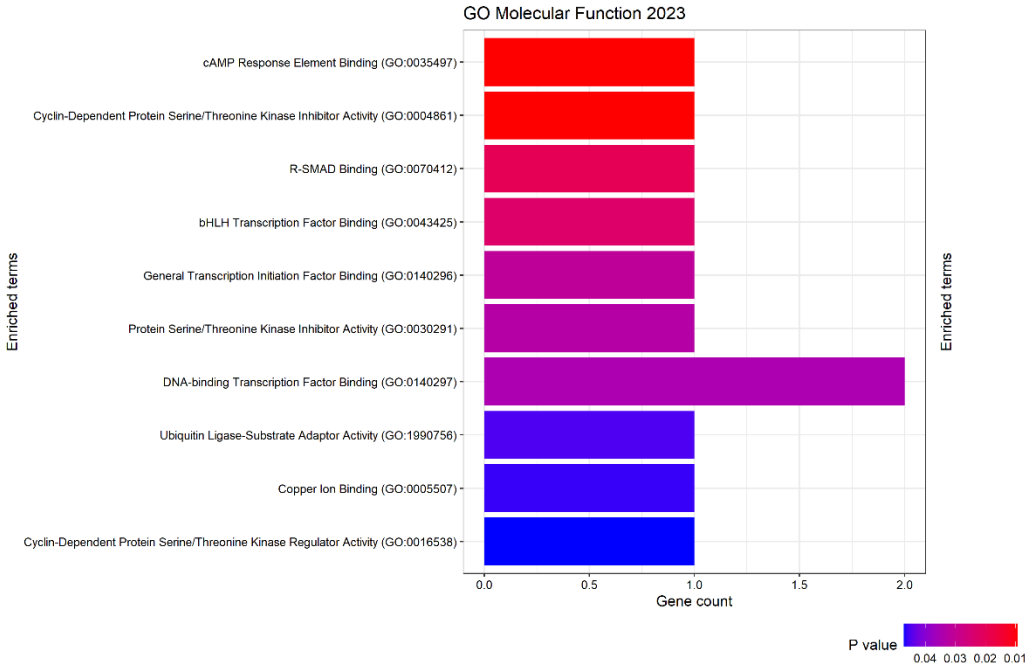
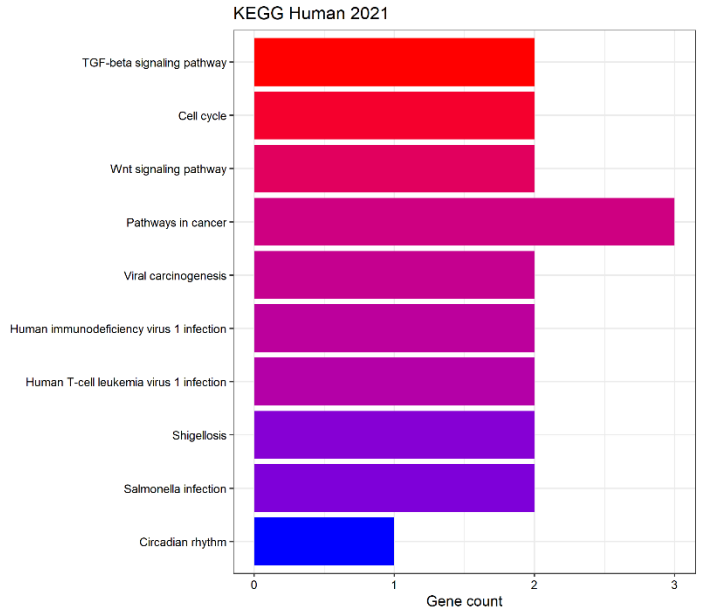



Figure 5: Enrichment analyses of the 19 genes from the 21 significant (p<3.77E-05) RNA transcripts for the transcriptomics analyses in the FTC. The figure shows the ten most enriched pathways and molecular functions. The left figure shows the enrichment in the Gene Ontology 2023 molecular functions database and the right figure shows the enrichment in the KEGG human pathway database.