



Analysis of DNA variants in miRNAs and miRNA 3'UTR binding sites in female infertility patients

Katarzyna M. Tyc^{1,5} · Anthony Wong¹ · Richard T. Scott Jr² · Xin Tao³ · Karen Schindler^{1,4} · Jinchuan Xing^{1,4}

Received: 15 June 2020 / Revised: 24 September 2020 / Accepted: 25 September 2020 / Published online: 17 October 2020

© The Author(s), under exclusive licence to United States and Canadian Academy of Pathology 2020

Abstract

Early human embryogenesis relies on maternal gene products accumulated during oocyte growth and maturation, until around day-3 post-fertilization when human zygotic genome activation occurs. The maternal-to-zygotic transition (MZT) is a tightly coordinated process of selective maternal transcript clearance and new zygotic transcript production. If MZT is disrupted, it will lead to developmental arrest and pregnancy loss. It is well established that microRNA (miRNA) mutations disrupt regulation of their target transcripts. We hypothesize that some cases of embryonic arrest and pregnancy loss could be explained by the mutations in the maternal genome that affect miRNA-target transcript pairs. To this end, we examined mutations within miRNAs or miRNA binding sites in the 3' untranslated regions (3'UTR) of target transcripts. Using whole-exome sequencing data from 178 women undergoing in vitro fertilization (IVF) procedures, we identified 1197 variants in miRNA genes, including 93 single nucleotide variants (SNVs) and 19 small insertions/deletions (INDELs) within the seed region of 100 miRNAs. Eight miRNA seed-region variants were significantly enriched among our patients when compared to a normal population. Within predicted 3'UTR miRNA binding sites, we identified 7393 SNVs and 1488 INDELs. Between our patients and a normal population, 52 SNVs and 30 INDELs showed significant association in the single-variant testing, whereas 51 genes showed significant association in the gene-burden analysis for genes that are expressed in preimplantation embryos. Interestingly, we found that many genes with disrupted 3'UTR miRNA binding sites follow gene expression patterns resembling MZT. In addition, some of these variants showed dramatic allele frequency difference between the patient and the normal group, offering potential utility as biomarkers for screening patients prior to IVF procedures.

These authors contributed equally: Katarzyna M. Tyc, Anthony Wong

Supplementary information The online version of this article (<https://doi.org/10.1038/s41374-020-00498-x>) contains supplementary material, which is available to authorized users.

✉ Jinchuan Xing
xing@biology.rutgers.edu

¹ Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

² Reproductive Medicine Associates of New Jersey, Basking Ridge, NJ, USA

³ Foundation for Embryonic Competence, Basking Ridge, NJ, USA

⁴ Human Genetics Institute of New Jersey, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

⁵ Present address: VCU Massey Cancer Center, Richmond, VA, USA

Introduction

Infertility is a major reproductive health issue that affects nearly 8.8% of all women in the United States [1]. Despite our advances in studying the underlying genetic factors that affect fertility rates, most of the studies of genetic variants focus on protein-coding variants (reviewed in [2]). Although regulatory non-protein-coding genes and variants also likely contribute to embryo viability variation, few studies focus on non-coding variants.

One of the important stages during embryonic development where non-coding genes play an important role is the maternal-to-zygotic transition (MZT). A properly functioning mechanism of MZT is essential to ensure successful early development, subsequent implantation, and development into a healthy offspring. MZT is a coordinated process of maternal transcript clearance and embryonic genome activation, which occurs around the

8-cell stage in humans [3]. The existence of both maternally- and zygotically-driven transcript clearance mechanisms are well documented to involve microRNAs (miRNAs), which are recognized as important regulators in these processes in many species (e.g., reviewed in [3–5]). For instance, selected miRNAs regulate their target DNA repair genes in human oocytes and blastocysts, supporting a role for miRNAs in maternal transcript clearance in human early embryogenesis [6]. It is widely accepted that dysregulation of the MZT process will lead to abnormal early embryogenesis, with a likely failed implantation and/or high early embryonic mortality.

In this study, we explore whether non-coding variants, such as variants in miRNA genes and miRNA binding sites in the 3' untranslated region (3'UTR) of target genes, could be used as biomarkers for infertile patients undergoing in vitro fertilization (IVF) procedures. Specifically, because a large proportion of failed IVF cycles can be attributed to embryonic aneuploidy (i.e., having a wrong number of chromosomes), we sought to understand whether there is a class of mutated miRNAs that is enriched among the patients producing an unusually high proportion of aneuploid embryos. Of note, a large number of miRNAs are expressed in human blastocysts [7]. Among miRNAs that were differentially expressed between euploid and aneuploid blastocysts, all but one were depleted in compromised embryos [7]. We thus specifically focused on variants predicted to abolish miRNA function.

To this end, we examined whole-exome sequencing data from 178 women undergoing IVF. Production of aneuploid embryos greatly reduces a patient's chances of implantation (fewer euploid embryos available for transfer) and increases the chances of miscarriage. Thus, we grouped our patients as either producing low or high proportions of aneuploid blastocysts (here, referred to as low rate group and high rate group, LRG, and HRG, respectively). We then identified DNA variants within miRNA genes and predicted miRNA binding sites in the 3'UTRs of the miRNA targeting genes. To further determine if our patient cohort harbors variants that can be used as biomarkers to predict infertility, we compared the allele frequency of the identified variants between our samples and a general population. Our analysis revealed that several candidate genes involved in crucial steps during embryo development harbor non-coding variants within their 3'UTR miRNA binding sites. In addition, these genes largely followed the expected MZT expression dynamics during the preimplantation embryonic development. Therefore, variants within these genes are likely to result in mis-regulation of their transcript levels during MZT, which could lead to infertility and thus deserve further investigation.

Materials and methods

Ethical approval and patient selection

Patient DNA samples were obtained from Reproductive Medicine Associates of New Jersey (RMANJ) DNA Bank. Analysis was approved by the IRB #RMA1-09-165 at Copernicus Group IRB and IRB #Pro2018000106 at Rutgers University.

Each patient was diagnosed with infertility but had normal hormone profiles and no diagnosed structural abnormalities. The IVF-derived embryos were tested for aneuploidy on day 5. Aneuploidy rate was calculated as previously described [8]. Briefly, patients with self-reported “White, non Hispanic” ancestry and with a minimum four embryos tested were used for aneuploidy rate calculation with the formula:

$$\text{(no. of aneuploid embryos)} / \text{(total no. of embryos tested)}$$

Robust nonlinear regression analysis was applied to determine the extreme samples on both ends of the age spectrum. For regression analysis, R package robustbase (v. 0.92.7) and its implemented function nlrob() was used as follows:

```
nlrob(y ~ (a * age^2 + b * age + c), data = list(x = age, y = aneuploidy), start = list(a = 1, b = 1, c = 1), psi = psi.bisquare)
```

Patients with corresponding “rweight” < 0.9 were selected for whole-exome sequencing (WES). Each patient's age, aneuploidy rate, and group assignment are listed in Table S1.

Sequencing and read mapping

WES was performed on blood DNA samples by Novogene (Sacramento, CA, USA). Sequencing libraries were prepared by Novogene from one microgram of DNA using Agilent SureSelect Human All Exon V6 kit (Agilent Technologies, CA, USA), according to specifications provided by the manufacturer. Barcoded and pulled samples were subjected to 2×150 base-pair (bp) pair-end sequencing using Illumina sequencing platform (Illumina, San Diego, CA, USA).

Raw reads from each lane were then mapped separately to the human reference genome (version hg19) with bwa-mem (v 0.7.15, [9]) and converted to the BAM format using samtools (v 1.4.1, [10]):

```
bwa mem -K 100000000 -Y -t 14 -R
  '@RG\tID:'$id'\tPU:'$pu'\tSM:'$sample'\tPL:ILLUMINA\tLB:'$library'\tCN:
  Novogene' refgenome.fasta file_R1.fastq
  file_R2.fastq | samblaster -a -addMateTags |
  samtools view -b -S /dev/stdin > file.bam
```

Picard Tools [11] was used to sort the mapped reads and mark duplicate reads:

```
java -Xmx16g -jar picard-2.9.2.jar Sort-Sam I=file.bam
O=file.qname.bam SORT_ORDER=queryname
java -Xmx16g -jar picard-2.9.2.jar
MarkDuplicates
I=file.qname.bam O=file.mdups.bam M=
file.mdups_metrics.txt
ASSUME_SORT_ORDER=queryname VALIDA-
TION_STRINGENCY=LENIENT
```

Finally, the alignment files for samples sequenced on multiple lanes were merged with `samtools merge` [10].

Variant calling

Target regions were extended by 50 bp on each end for variant calling. The GATK v3.8 pipeline (GenomeAnalysisTK-3.8-0-ge9d806836) was applied following GATK best practices [12, 13]. In short, reads were recalibrated with `BaseRecalibrator`, followed by `PrintReads`. `HaplotypeCaller` was called on 0.5 Mbp intervals, and the resulting genome variant call format (gVCF) files were concatenated with `CatVariants`. `CombineGVCF` was used to merge all the samples on a per chromosome level, followed by a joint-genotype call with `GenotypeGVCFs`. `CatVariants` was used again to concatenate the chromosome-based multi-sample variant call format (VCF) files. Single nucleotide variants (SNVs) and small insertions/deletions (INDELs) were then sequentially recalibrated with `VariantRecalibrator` followed by `ApplyRecalibration` with default settings.

Variant and sample filtering

Only variants with a PASS flag were kept in the analysis. Quality of the multi-sample VCF file was assessed with `bcftools-1.9 stats` function (available as part of SAMtools software package, [10]). One sample showed excess heterozygous calls and was removed with `VCFtools` (v 0.1.17, [14]). Following principal component analysis (PCA, as described in [8]), three samples were four standard deviations away from the mean among the patients on the first two principal components and were removed with `VCFtools`.

miRNA annotation and structure prediction

ADmiRE, a comprehensive miRNA database, was used to extract miRNA genomic locations [15]. miRVaS was used to predict structural effects of a given miRNA variant [16]. The location of the variant was assigned to specific

functional regions within the miRNA: flank, which denotes the region 100 nucleotides upstream and downstream of a miRNA gene; loop, indicating the terminal loop of the precursor miRNA (pre-miRNA); arm, which forms the hairpin of the pre-miRNA; mature, indicating mature miRNA; and seed, indicating the seed sequence of the mature miRNA [16]. Genomic positions of annotated miRNA genes in hg19 coordinates were downloaded for ADmiRE (ADmiRE.tab, <https://github.com/nroak/ADmiRE>) and miRVaS (mirna_hg19_mirbase20.tsv). Mutations within miRNA genes in ADmiRE database were extracted from the VCF file using BEDtools [17]:

```
bedtools intersect -a input.vcf -b
ADmiRE.bed
```

Mutations within miRNA genes in miRVaS database were extracted using the miRVaS command:

```
mirvas <VCF_file> <output_file> <genome_
hg19.ifas>
<miRNA_Database>
```

miRVaS also performs pre-miRNA secondary structure prediction using RNAFold (version 2.1.5 [18]) and denotes the most important variant with a structural impact. This is performed with three different prediction algorithms: Minimal Free Energy (MFE), Maximal Expected Accuracy (MEA), and centroid (CEN).

Association testing of miRNA seed region mutation among IVF patients and between IVF patients and a general population

Allele frequencies (AFs) among LRG and HRG patients were calculated for miRNA variants using the `vcftools --freq` function [14]. The AFs were then compared between the two groups using the Fisher's Exact test to identify variants that are significantly associated with either LRG or HRG (R, version 3.5.1). Multiple-testing correction on the Fisher's Exact test *p* values was performed with the Benjamini–Hochberg (BH) method [19].

The variant calls based on the whole-genome sequencing of non-Finnish European (NFE) population from the Genome Aggregation Database (gnomAD v.2.0.1; [ftp.ensembl.org/pub/data_files/homo_sapiens/GRCh37/variation_genotype/gnomad.genomes.r2.0.1.sites.noVEP.vcf.gz](ftp://ftp.ensembl.org/pub/data_files/homo_sapiens/GRCh37/variation_genotype/gnomad.genomes.r2.0.1.sites.noVEP.vcf.gz)) was used as the reference in the population comparison analysis. gnomAD aggregates whole-genome sequencing information from case-control studies and provides expected AF of genetic variants among human populations [20]. The AFs were compared between the infertile patients (LRG + HRG) and NFE population using the Fisher's Exact test (R, version 3.5.1), followed by multiple-testing correction of the *p* values using the BH method. Variants that are not found within the gnomAD database were excluded from analysis.

Identification of miRNA target genes

DIANA mirPath (<http://snf-515788.vm.okeanos.grnet.gr/>) and its microT-CDS prediction algorithm (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index) was used to generate a list of microRNA-to-gene associations [21]. Genes that were predicted to be regulated by mutated miRNAs were uploaded to DIANA mirPath as a gene list to be functionally characterized. Statistically significant target pathways, along with the associated miRNAs and genes found in each pathway, were identified by employing the mirPath's empirical sampling algorithm.

Selecting functionally relevant 3'UTR variants for association testing

Variants were annotated with ANNOVAR [22] to identify 3'UTR SNVs and INDELS:

```
table_annoar.pl vcf file.vcf annovar/
humandb/ -out myanno -buildver hg19 -proto-
col refGene,gnomad_genome,snp135,avsnp150
-operation g,f,f,f -argument "-separate",,,
-nastring.
```

3'UTR miRNA binding sites were predicted by TargetScanHuman 7.2 (http://www.targetscan.org/vert_72/). TargetScanHuman predicts miRNA binding sites by searching for 6–8mer 3'UTR sites matching seed regions of miRNAs [23]. 3'UTR variants within predicted miRNA binding sites were selected using BEDtools intersection [17].

Gene expression measurements collected on zygote, 4-cell, 8-cell, compacted morula, inner cell mass (ICM) from early blastocyst as well as epiblast and primitive endoderm from late blastocyst stages [24] were used to filter out the non-expressed genes prior to association testing. All genes with a maximum Fragments Per Kilobase of transcript per Million mapped reads (FPKM) < 1 across these stages were deemed non-expressed.

The Fisher's Exact test and the BH correction method were applied to identify 3'UTR SNVs and INDELS that were significantly associated with either LRG or HRG. To calculate the association of the variants in our patient cohort compared to gnomAD NFE population, the Fisher's Exact test and the BH correction method were performed using gnomAD NFE AFs as controls and AFs of all patients (LRG + HRG) as cases.

Functional enrichment analysis of genes harboring variants in 3'UTR miRNA binding sites

The list of genes harboring significantly enriched 3'UTR variants that overlap miRNA binding sites was uploaded to

the functional enrichment analysis tool WebGestalt [25] and DAVID 6.8 [26]. WebGestalt parameters of analysis were set to “over-representation analysis (ORA)”, with “gene-ontology:BP” as the functional database, and “genome” as the reference gene list. For DAVID 6.8, the default genome-wide list of genes for *Homo sapiens* was used as the background gene list [26].

Early embryogenesis gene expression analysis

Gene expression profiles for early preimplantation embryonic stages were obtained from a previous study [24]. Seven expression samples were provided by the study from six developmental stages: zygote, 4 cell, 8 cell, compacted morula, early blastocyst (ICM), and late blastocyst (epiblast and primitive endoderm). Pseudo count of one was added to FPKM values for each gene in the seven samples, and the FPKM values were \log_2 transformed (i.e., $\log_2(\text{FPKM} + 1)$) and mean centered. Hierarchical clustering of the expression data of each candidate gene list was performed with Pearson's correlation coefficient as the distance metric using Cluster 3.0 [27] and visualized using JavaTreeView [28].

To calculate the expression level change of late embryonic stages compared to early embryonic stages for each gene, the mean FPKM value of the 8 cell, compacted morula, ICM, epiblast, and primitive endoderm samples was divided by the mean FPKM value of the zygote and 4 cell stage. All genes with the \log_2 fold-change of the two means ($\log_2\text{FC}^{\#1}$) < -1 were defined as following a “high-to-low” expression pattern. Similarly, the mean FPKM value of compacted morula, ICM, epiblast, and primitive endoderm samples was divided by the mean FPKM value of the zygote, 4 cell, and 8 cell samples for each gene. Genes with the \log_2 fold-change of these two means ($\log_2\text{FC}^{\#2}$) > 1 were defined as following a “low-to-high” expression pattern. The calculated means and fold-changes are listed in Table S12. The number of genes that are found in each category in the whole transcriptome data was used as the expected values in the Fisher's Exact test when calculating the significance of the enrichment of gene expression patterns in the analysis of genes with 3'UTR variants.

Results

Overview of the study

Patients undergoing IVF in combination with preimplantation genetic testing for aneuploidy were divided into LRG and HRG based on embryo aneuploidy rates (Table S1, see Methods). Using WES, a total of 449,358

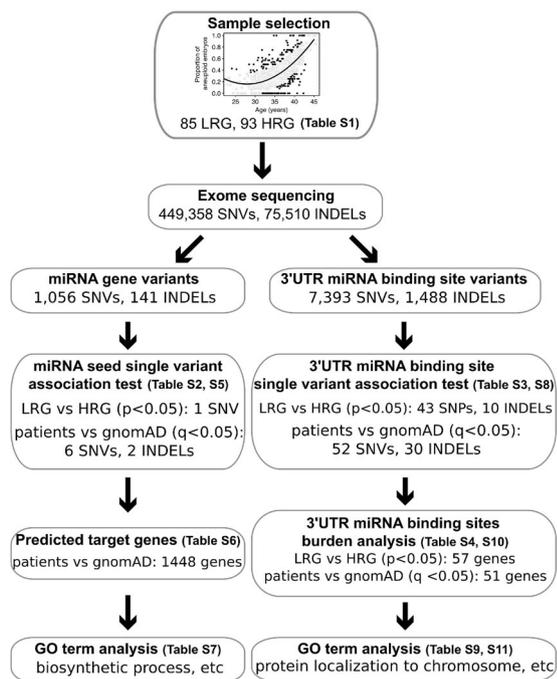


Fig. 1 Analysis workflow to identify candidate non-coding variants implicated in fertility. Infertile patients were divided into LRG and HRG based on their aneuploidy rates. Following WES and variant identification, SNVs and INDELs that overlap miRNA seed-sequences (left) and 3'UTR miRNA binding sites (right) were selected for analyses as outlined in the figure. Gene expression analysis and GO term enrichment analysis were used to identify candidate variants that potentially dysregulate pathways that influence fertility phenotype.

SNVs and 75,510 INDELs were identified in 178 patients (85 LRG and 93 HRG, Fig. 1) and miRNA-related variants were analyzed from two perspectives. First, we examined variants within miRNA genes that have a potential to disrupt the miRNA's binding capacity to its target sites. Second, we searched for variants within 3'UTRs that could reduce miRNA binding affinity and thus modify a miRNA's regulatory effect.

miRNAs and their associated variants

We used ADmiRE and miRVA_S to select variants that are within 100 bp of annotated miRNA genes. ADmiRE contains 1878 miRNAs, whereas miRVA_S contains 1871. Most of the miRNA genes (1850) were annotated in both databases. Next, we determined the number of miRNA genes that were fully covered by the exome sequencing. About 90% of the miRNA genes were fully captured by the WES target regions, according to the annotations provided with ADmiRE (1670, 89.7%) and miRVA_S (1674, 90%; Fig. 2a). When we considered 100 bp flanking regions on each side of a miRNA gene, 604 (spanning 161,906 bp) and 603

(spanning 161,641 bp) miRNAs genes listed in ADmiRE and miRVA_S were fully covered by our data, respectively. For completeness, we included miRNA genes from both databases in our following analyses.

In total, 1056 SNVs and 141 INDELs overlapped miRNA gene regions. Expanding the overlapping regions to include 100 bp upstream and downstream of the miRNA genes, 2942 SNVs and 402 INDELs overlapped miRNA genes (all variants are listed in Table S2). Of the 266 mutated mature miRNAs, 237 (89.1%) had 1 SNV, 19 (7.1%) contained 2 SNVs, and 10 (3.8%) contained 3 or more SNVs. mir-8078-3p was most frequently mutated in our patient population (15 SNVs) with all variants identified in the gnomAD database. mir-8078-3p is located in a highly variable genomic region and this genomic context might contribute to the high number of SNVs in this miRNA. Among the 29 miRNAs that contained INDELs, 27 (93.1%) contained a single INDEL, 1 contained 2 INDELs, and 1 contained 4 INDELs. We then stratified miRNA variants by miRNA regions as depicted in Fig. 2b (adapted from [16]). In total, 93 SNVs and 19 INDELs were within the miRNA seed regions and thus most likely to affect the miRNA function (summarized in Fig. 2c and detailed in Table S2). Besides seed variants, of all SNVs and INDELs in non-seed regions (labeled as *mature (non seed)*, *arm*, *loop*, or *flank*, Fig. 2c), only 152 SNVs and 18 INDELs were predicted by all three miRVA_S prediction algorithms to have an impact on miRNA seed region (the prediction scores were collected in Table S2 in columns "PredictedEffect_Centroid", "PredictedEffect_MEA", and "PredictedEffect_MFE", see Methods for details).

Next, we applied the Fisher's Exact test with the BH multiple-testing correction to determine which of the miRNA variants were significantly enriched among HRG or LRG patients. No variant reached a statistically significant level after adjustment for multiple testing. In the seed region, only one SNV had a raw p value <0.05 (chr1:228284991 C-T, $p = 0.011$). The SNV was enriched in LRG and within the seed region of miR-3620-5p (Table S2). Outside of the seed region, 2 variants in mature regions, 5 in arm regions, 1 in the loop, and 19 within flanking regions of miRNAs also had raw p values <0.05 . None of these variants were predicted to have an impact on miRNA seed region by all three algorithms implemented in miRVA_S (Table S2).

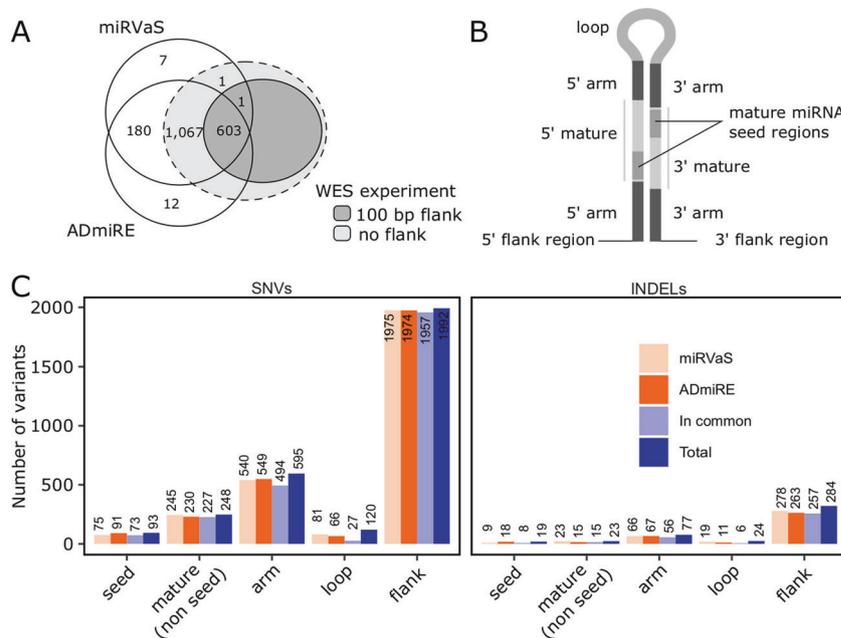
Identification of genes with 3'UTR variants

We annotated the WES-derived variants and identified 10,079 SNVs and 1901 INDELs within 3'UTRs of protein-coding genes. TargetScanHuman [23] predicted that 7393 SNVs and 1488 INDELs were within potential miRNA

binding sites (Fig. 1). Using expression data from [24], variants in non-expressed genes (i.e., genes with a maximum FPKM value less than one) were removed from analysis. Using the Fisher's Exact test, we determined the single-variant association for the two patient groups. In total, 43 SNVs and 10 INDELS in 51 expressed genes showed nominal association between LRG or HRG ($p < 0.05$; Table S3). No variant reached genome-wide significance after the multiple-testing correction. To determine the difference in overall burden of 3'UTR variants affecting miRNA-mRNA interaction in each gene, we aggregated 3'UTR variants across all predicted miRNA binding sites in each gene and performed gene-level association tests between the two patient groups. Among the 4999 genes containing 3'UTR variants that overlapped miRNA binding sites, 57 genes had a potential association with either LRG or HRG (Fisher's Exact test, $p < 0.05$; Table S4).

Fig. 2 Variants stratified by miRNA gene regions.

A Number of miRNA genes from miRVaS and ADmiRE databases that were covered in the WES experiment. **B** Diagram depicting miRNA stem-loop structure with highlighted distinct miRNA gene regions (adapted from [16]). In our analyses 5' and 3' flanks span 100 bps on each end of the miRNA gene. **C** Number of SNVs and INDELS within distinct regions of miRNAs as outlined in (B) and per tool as indicated by the color legend.



Identification of variants with different allele frequencies between the infertility patient group and a general population

Although our patient groups had different outcomes from one another with respect to embryo aneuploidy rates, all the individuals we sequenced were diagnosed with infertility. To determine whether AFs of miRNA-related variants in infertility patients as a whole are different from a general population, we compared their sequences to 7509 NFE individuals from the gnomAD project as controls (Table S5, these variants are the same miRNA variants listed in Table S2, but now contrasted against gnomAD NFE population). Among the 93 seed-SNVs and 19 seed-INDELS in the patient group, 6 SNVs and 2 INDEL in 6 different miRNAs showed significant association with our patient cohort (BH-corrected p value (q) < 0.05 ; Table 1).

Table 1 Significantly enriched miRNA seed variants in the comparison of patient cohort with gnomAD NFE population.

miRNA	Chr	Start	End	Ref	Alt	dbSNP150	AF_Patient ^a	AF_gnomAD NFE ^a	pVal ^b	FDR ^c
hsa-mir-4472-2-5p	12	116866112	116866114	CA	-	rs202127912	0.23	1.60E-02	2.6E-54	1.5E-51
hsa-mir-8078-3p	18	112284	112285	G	C	0	0.17	3.90E-02	2.2E-16	6.1E-15
hsa-mir-8078-3p	18	112286	112287	C	A	rs775697263	0.15	3.50E-02	4.3E-16	1.2E-14
hsa-mir-8078-3p	18	112281	112282	C	A	0	0.04	1.00E-03	5.4E-13	1.2E-11
hsa-mir-1273a-3p	8	101036233	101036234	G	A	rs932668866	0.01	1.34E-04	1.2E-04	1.5E-03
hsa-mir-6859-1-3p	1	17384	17385	G	A	rs201535981	0.25	1.60E-01	6.1E-04	6.8E-03
hsa-mir-1273d-5p	1	10287790	10287791	A	G	0	0.01	6.74E-05	1.6E-03	1.6E-02
hsa-mir-302c-3p	4	113569539	113569543	ACTT	-	rs199971565	0.06	3.00E-02	5.1E-03	4.4E-02

^aAllele frequency of the patient and gnomAD NFE group, respectively.

^bRaw p value from Fisher's Exact test.

^cBenjamini-Hochberg adjusted p value.

Next, we examined if the six miRNAs with mutated seed-sequences can target genes that together constitute one or more pathways. We used DIANA mirPath to predict target genes for the 6 miRNAs and identified 1448 target genes (Table S6). The targeted genes showed statistically significant enrichments in 27 gene ontology (GO) biological processes. Some of the processes are important for cellular division processes (e.g., biosynthetic process, DNA-templated transcription; Table S7).

Within 3'UTR miRNA target binding sites, we again compared the variant AFs of the patient group with the gnomAD NFE population. Out of the 8881 3'UTR variants within miRNA binding sites that we initially described in Table S3, 52 SNVs and 30 INDELS in 64 expressed genes showed significant AF difference between the patient group and the control group after multiple testing correction ($q < 0.05$; Table S8). The 64 expressed genes were not significantly enriched in any specific GO category (Table S9). Sixty-seven of the 82 variants (82%) were more common in our patient group compared to the control group. The most dramatic examples included rs35334288 (patient AF 32.9%, NFE AF 0.7%) and rs146607300 (patient AF 27.2%, NFE AF 0.5%), both of which have AFs more than 40 times higher in our infertility patient group compared to NFE.

We then aggregated the 3'UTR variants to perform a gene-level association test between our patient cohort against gnomAD NFE. Among the 5161 expressed genes with 3'UTR variants overlapping miRNA binding sites, 51 expressed genes showed significant association with either the gnomAD NFE group or our patient group ($q < 0.05$; Table S10). Although no GO term was significantly enriched among these 51 genes, we observed a nominal enrichment in biological processes such as cellular protein localization, telomere maintenance via semi-conservative replication, and protein localization to chromosome (Table S11).

Genes with 3'UTR variants within miRNA binding sites display two distinct expression patterns during early stages of embryonic development

Successful early embryogenesis relies on accurate transcriptome regulation. During MZT we expect two expression patterns to occur: an initial high level of maternally provided transcripts followed by their rapid degradation, and the appearance of newly transcribed zygotic genes that accompanies the maternal transcript clearance (Fig. 3A). For genes that were significantly associated either with our patient group or the control group, we asked if their expression profiles clustered in these patterns during pre-implantation stage of the embryogenesis. To this end, we used the human embryo gene expression data assessed at the preimplantation stages of the development (see Methods

for more details) and analyzed two sets of genes. The first set included 64 genes that showed significant association in the single-variant analysis (Fig. 3B) and the second set included 51 genes that showed significant association in the burden analysis (Fig. 3C). Because zygotic genome activation in human embryos occurs between 4 and 8 cell stage of development [29], in the following analysis we assumed that clearance of maternal transcripts takes place right after the 4 cell stage, followed by the appearance of newly transcribed zygotic genes around the 8 cell stage. Accordingly, our candidate genes were categorized into two expression patterns: (1) a “high-to-low” expression pattern for all genes with a $\log_2FC^{\#1} < -1$ between the mean expression across the 8 cell, morula, and blastocyst stages and the mean expression across the zygote and 4 cell stage; and (2) a “low-to-high” expression pattern with a $\log_2FC^{\#2} > 1$ between the mean expression across the morula and blastocyst stages and the mean expression across zygote, 4 cell, and 8 cell stages (see Methods for details). Among the 19,641 genes analyzed by Stirparo et al. [24], 5108 and 7713 genes fell into the “high-to-low” and “low-to-high” group, respectively (Table S12). In our single-variant analysis of 64 mutated genes, 25 genes had a “high-to-low” expression and 22 genes had a “low-to-high” expression (Fig. 3B). The “high-to-low” category showed a significant enrichment compared to the whole transcriptome (Fisher's Exact test $p < 0.05$). This analysis suggests that genes harboring significant 3'UTR variants disproportionately fall within the “high-to-low” category. Given their expression pattern, these transcripts are likely to be maternally provided. In our burden analysis, 17 and 18 out of 51 mutated genes displayed a “high-to-low” and a “low-to-high” expression pattern, respectively (Fig. 3C). Neither category showed a significant enrichment compared to the whole transcriptome.

Discussion

In this report, we used whole-exome sequencing from an infertility patient cohort to perform association studies between miRNA gene and miRNA target binding site variants and infertility phenotypes. In particular, we examined whether mutated miRNAs and/or miRNA target binding sites affect genes that function in cellular processes important for early embryogenesis. Identification of genetic variants enriched within the patient group can serve two purposes: (1) these variants can help us understand the genetic risk factors of fertility problems and the underlying molecular mechanisms; (2) variants that are commonly observed in the patient group could serve as potential biomarkers to diagnose infertility or to improve the success rates of assisted reproductive technologies.

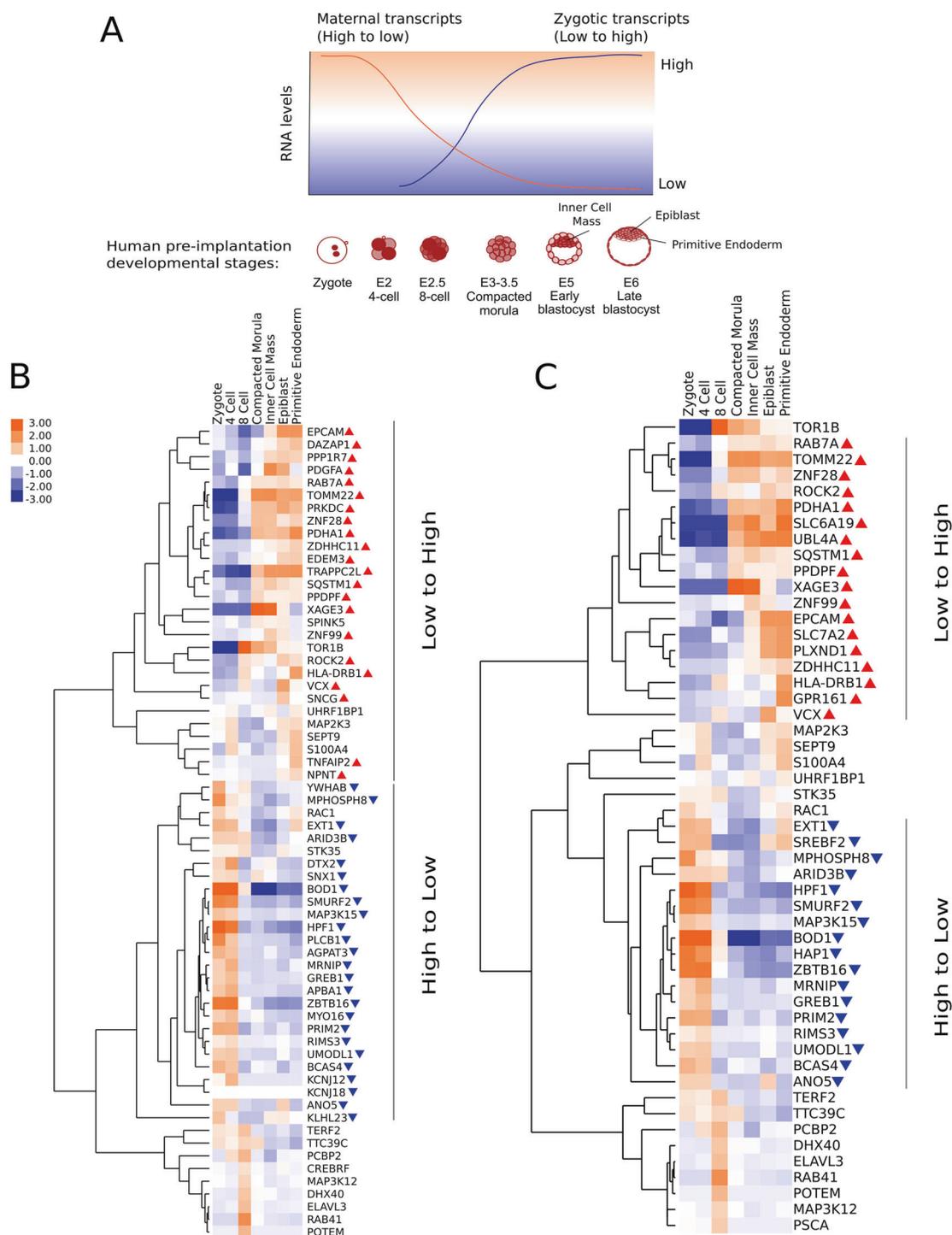


Fig. 3 Genes enriched for 3'UTR variants within miRNA binding sites follow expression resembling maternal transcript clearance pattern. **A** A schematic diagram of anticipated mRNA changes during early stages of human embryogenesis. The human preimplantation developmental stages are labeled on the x-axis, and the presumed RNA time courses for maternal (orange) and zygotic (blue) transcripts are indicated with solid lines. **B** Clustering of 64 genes with significant 3'UTR variants in the single-variant analysis of patient vs control groups. **C** Clustering of 51 Genes with significant 3'UTR variants from the burden analysis of patient versus control groups. Gene

expressions were $\log_2(\text{FPKM} + 1)$ transformed and mean-centered row-wise. Color scales are the same for **B** and **C**. Genes following a “low-to-high” expression pattern are indicated with red upwards triangles. Genes following a “high-to-low” expression pattern are indicated with blue downwards triangles (see Methods). Dendrograms were generated by hierarchical clustering of the transformed expression data. Developmental stages: zygote, 4 cell, 8 cell, compacted morula, early blastocyst (Inner Cell Mass), and late blastocyst (Epiblast and Primitive Endoderm).

Our comparison between patients with high and low rates of aneuploidy did not identify statistically significant associations between non-coding variants and aneuploidy rate. Given the limited sample size of our groups, it is likely that the lack of association is due to the lack of statistical power, rather than miRNAs not engaging in biologically relevant processes. When considering all infertility patients together and comparing the allele frequency differences of miRNA variants between the patient group and a NFE population from the gnomAD database, we identified eight miRNA seed variants within six miRNAs that were significantly enriched among infertility patients. A possible effect of these variants could be that affected miRNAs no longer control turnover of maternal genes that are not compatible with the zygotic transcriptome. This effect could cause a failure in embryonic development, implantation or cause early miscarriage, resulting in an infertility diagnosis. Subsequent GO enrichment analyses of the 1448 potential target genes of these six miRNAs showed that the disruption of the target genes could impact biological processes such as “biosynthetic process” and “DNA-templated transcription” that are important for cell-cycle (Table S7).

In addition to variants in miRNA genes, we also examined variants within regions that are predicted to constitute potential miRNA binding sites within 3'UTRs of protein-coding genes. Among genes that are expressed during early human embryonic development, 64 genes that contained 3'UTR variants within miRNA binding sites showed significant differences in our patient group compared to the general population. Although GO analysis on these genes did not identify any significant term, several genes from the list perform functions important for early embryogenesis. One such gene is *SEPT9* (Septin 9), a gene involved in cytoskeletal and cell-cycle control. *Sept9* knockout mouse embryos demonstrate embryonic lethality around day 10 of gestation [30]. In addition, variants such as rs35334288 in *ELAVL3* and rs146607300 in *MAP3K12* are 40 times more common in the infertility patient cohort compared to the control group. Such common variants in the patient population have the potential to serve as biomarkers for making better-informed decisions regarding reproductive choices.

Burden analysis of the 3'UTR variants identified a total of 51 significant genes. We observed interesting gene expression patterns that were shared among these genes across early stages of embryogenesis. Using gene expression data collected from various preimplantation stages of the developing embryo, we found that genes that harbor 3'UTR miRNA binding site variants displayed two main expression changes patterns; a “low-to-high” pattern with the inflection point occurring between the 8 cell and compacted morula stage; and a “high-to-low” pattern with the expression change between the 4 and 8 cell stages (Fig. 3B, C). Our single-variant analysis revealed that our patient

cohort has a disproportionately high number of genes with 3'UTR mutations that follow the “high-to-low” pattern of expression. Such disrupted maternal transcript clearance may not be compatible with zygotic genome activation. Functional studies are needed to determine whether the presence of these variants affects the clearance dynamics of maternal transcripts in early stages of embryo development and/or how these variants affect embryo development past the 4-cell stage. As such, these variants have a potential to serve as infertility biomarkers.

Although candidate genes harboring 3'UTR mutations that showed the strongest association with our patient group are involved in a broad range of cellular processes (Tables S9 and S11), it is difficult to directly link these processes to fertility issues. Nonetheless, several of these genes display relevant mouse phenotypes upon partial or complete abrogation. Searching MouseMine (<http://www.mousemine.org/mousemine/begin.do>) we found that several of our candidate genes were related to reproductive system phenotype, reduced female fertility, and embryonic or pre-weaning lethality (*Elavl3*, *Pcbp2*, *Map3k12*, *Tomm22*, *Ano5*), suggesting that our single-variant analysis of 3'UTR variants has identified a range of candidate genes that could contribute to human fertility. Future functional studies of these genes and their associated non-coding variants are needed to elucidate their roles in embryogenesis and range of issues underlying infertility (e.g., quality and number of eggs retrieved).

A main limitation of our study is the relatively small patient sample size. The lack of significant variants in our LRG vs HRG comparison after multiple testing correction could suggest that miRNAs do not play an important role in the biological mechanisms of MZT. However, some of our observations argue against this hypothesis. For example, when we compared all of our infertility patients with a normal population, many variants showed significant association, both within miRNAs (Table S5) and within 3'UTR miRNA binding sites (Table S8). In addition, we observed an interesting expression pattern for genes with significant variants in the 3'UTR miRNA binding sites (Fig. 3B, C). Namely, these genes tend to display a “high-to-low” expression pattern during the very first stages of embryonic development ($p < 0.05$, Fisher's Exact test). This expression pattern fits into the widely accepted hypothesis that maternally provided transcripts must be cleared during the MZT and that this likely occurs with miRNA's help. Therefore, we believe that the lack of significant results is more likely due to the small sample size in our study. With regard to the lack of significantly enriched GO terms among the genes with mutated 3'UTR miRNA binding sites (Tables S9 and S11), it is likely that the maternal transcripts involved in MZT are responsible for many aspects of embryonic development and are not enriched in a particular pathway.

In the future, studies with a larger patient cohort could confirm and strengthen the association between certain variants and aneuploidy phenotype and better assess the role of miRNAs in aneuploidy. Such variants could also serve as biomarkers in a clinical setup where they could prove useful in making better-informed decisions regarding reproductive choices. In summary, we present an investigation towards understanding what non-coding genetic variants could be contributing to female infertility.

Acknowledgements We thank the patients who participated in and contributed to this study. This work is supported by a grant from the NIH/NICHD to KS, JX, and XT: R01-HD091331. We gratefully acknowledge access to the HPC facilities and support of the computational STEM and bioinformatics scientists from the Office of Advanced Research Computing at Rutgers University.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Infertility, CDC. 2019. <https://www.cdc.gov/nchs/fastats/infertility.htm>.
- Yatsenko SA, Rajkovic A. Genetics of human female infertility. *Biol Reprod*. 2019;101:549–66.
- Lee MT, Bonneau AR, Giraldez AJ. Zygotic genome activation during the maternal-to-zygotic transition. *Annu Rev Cell Dev Biol*. 2014;30:581–613.
- Walser CB, Lipshitz HD. Transcript clearance during the maternal-to-zygotic transition. *Curr Opin Genet Dev*. 2011;21:431–43.
- Salilew-Wondim D, Gebremedhn S, Hoelker M, Tholen E, Hailay T, Tesfaye D. The role of MicroRNAs in mammalian fertility: from gametogenesis to embryo implantation. *Int J Mol Sci*. 2020;21:585.
- Tulay P, Naja RP, Cascales-Roman O, Doshi A, Serhal P, Sen-Gupta SB. Investigation of microRNA expression and DNA repair gene transcripts in human oocytes and blastocysts. *J Assist Reprod Genet*. 2015;32:1757–64.
- Rosenbluth EM, Shelton DN, Sparks AE, Devor E, Christenson L, Van Voorhis BJ. MicroRNA expression in the human blastocyst. *Fertil Steril*. 2013;99:855–61.e3.
- Tyc KM, El Yakoubi W, Bag A, Landis J, Zhan Y, Treff NR, et al. Exome sequencing links CEP120 mutation to maternally derived aneuploid conception risk. *Hum Reprod*. 2020;35:2134–48.
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 2013;1303:3997.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
- Picard Tools (version 2.9.2), Broad Institute, 2017. <http://broadinstitute.github.io/picard/>.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
- Wang S, Xing J. A primer for disease gene prioritization using next-generation sequencing data. *Genomics Inform*. 2013;11:191–9.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
- Oak N, Ghosh R, Huang KL, Wheeler DA, Ding L, Plon SE. Framework for microRNA variant annotation and prioritization using human population and disease datasets. *Hum Mutat*. 2019;40:73–89.
- Cammaerts S, Strazisar M, Dierckx J, Del Favero J, De, Rijk P. miRVaS: a tool to predict the impact of genetic variants on miRNAs. *Nucleic Acids Res*. 2016;44:e23.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
- Lorenz R, Bernhart SH, Höner zu Siederdisen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*. 2011;6:26.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*. 1995;57:289–300.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434–43.
- Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res*. 2015;43:W460–6.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
- Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 2015;4:e05005.
- Stirparo GG, Boroviak T, Guo G, Nichols J, Smith A, Bertone P. Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. *Development*. 2018;145:dev158501.
- Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis toolkit (WebGestalt): update 2013. *Nucleic Acids Res*. 2013;41:W77–83.
- Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44–57.
- de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. *Bioinformatics*. 2004;20:1453–4.
- Saldanha AJ. Java Treeview—extensible visualization of microarray data. *Bioinformatics*. 2004;20:3246–8.
- Jukam D, Shariati SAM, Skotheim JM. Zygotic genome activation in vertebrates. *Dev Cell*. 2017;42:316–32.
- Fichtbauer A, Lassen LB, Jensen AB, Howard J, Quiroga Ade S, Warming S, et al. Septin9 is involved in septin filament formation and cellular stability. *Biol Chem*. 2011;392:769–77.