Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos

Elo Madissoon1,*, Eeva-Mari Jouhilahti1,*, Liselotte Vesterlund1, Virpi Töhönen1, Kaarel Krjutškov1,2, Sophie Petropoulos3, Elisabet Einarsdóttir1,4, Sten Linnarsson5, Fredrik Lanner3, Robert Månsson6, Outi Hovatta3, Thomas R. Bürglin7, Shintaro Katayama1 & Juha Kere1,4

PAIRED (PRD)-like homeobox genes belong to a class of predicted transcription factor genes. Several of these PRD-like homeobox genes have been predicted in silico from genomic sequence but until recently had no evidence of transcript expression. We found recently that nine PRD-like homeobox genes, ARGFX, CPHX1, CPHX2, DPRX, DUXA, DUXB, NOBOX, TPRX1 and TPRX2, were expressed in human preimplantation embryos. In the current study we characterized these PRD-like homeobox genes in depth and studied their functions as transcription factors. We cloned multiple transcript variants from human embryos and showed that the expression of these genes is specific to embryos and pluripotent stem cells. Overexpression of the genes in human embryonic stem cells confirmed their roles as transcription factors as either activators (CPHX1, CPHX2, ARGFX) or repressors (DPRX, DUXA, TPRX2) with distinct targets that could be explained by the amino acid sequence in homeodomain. Some PRD-like homeodomain transcription factors had high concordance of target genes and showed enrichment for both developmentally important gene sets and a 36 bp DNA recognition motif implicated in Embryo Genome Activation (EGA). Our data implicate a role for these previously uncharacterized PRD-like homeodomain proteins in the regulation of human embryo genome activation and preimplantation embryo development.

Human preimplantation development, starting with the fusion of egg and sperm and continuing to blastocyst implantation1 is characterized by drastic changes in gene expression, especially a massive degradation of oocyte transcripts and a gradual, cascade-like initiation of transcription from the embryo genome2,3. The cellular factors responsible for the Embryo Genome Activation (EGA) have not been fully characterized. However, recently the PRD-like homeodomain transcription factors were implicated in human EGA and preimplantation development3.

The PRD-like homeobox genes form one of several classes of homeobox genes1. Homeobox genes are evolutionarily conserved and the majority of these genes are critical for the regulation of development2. Examples of homeobox genes include the HOX cluster genes5 that are known to play key roles in cell fate determination and cell differentiation. Homeobox genes are relatively well characterized in several model organisms, especially in Drosophila melanogaster6, and several of the PRD-like homeobox genes have previously been suggested to be expressed in the human germline7. The PRD-like homeobox genes share a similar homeodomain, but they lack...
the actual PRD domain. The homeodomains consists of approximately 60 amino acid (aa) residues and folds into a N-terminal arm and three alpha-helices. The N-terminal arm and the third helix have been implicated in DNA-binding specificity and DNA-sequence recognition.

Our recent study on human preimplantation development revealed the expression of 14 PRD-like homeodomain transcription factors during EGA, and we validated the presence of seven factors by cDNA cloning. Most of these transcription factors had been predicted, but only a few had been validated on transcript level.

Here we describe the molecular cloning of several new splice variants for altogether nine human PRD-like homeobox genes, \textit{ARGFX} (arginine-fifty homeobox; altogether six splice variants), \textit{CPHX1} (cytoplasmic polyadenylated homeobox 1), \textit{CPHX2} (cytoplasmic polyadenylated homeobox 2; three splice variants), \textit{DPRX} (divergent paired-related homeobox; two splice variants), \textit{DUXA} (double homeobox A), \textit{DUXB} (double homeobox B; three splice variants), \textit{TPRX1} (tetra-peptide repeat homeobox 1), \textit{TPRX2} (tetra-peptide repeat homeobox 2; three splice variants) and \textit{NOBOX} (NOBOX oogenesis homeobox). We show the first evidence of \textit{TPRX2P} expression and propose that it should be renamed \textit{TPRX2} as previously suggested. Many of these PRD-like homeobox genes appeared primate specific, expressed at very low levels in most tissues and restricted to preimplantation development. Experimental overexpression of the PRD-like homeodomain transcription factors in human embryonic stem cells (hESCs) reveals target genes that intersect with a significant number of the genes first activated in human embryos. The target gene promoters are enriched for the 36 bp motif previously predicted to play a major role in human EGA.

Results
Cloning of PRD-like homeobox genes from human 8-cell stage embryos. Our previous transcription start site (TSS) enriched data on human preimplantation embryos suggested unannotated transcripts for \textit{ARGFX}, \textit{CPHX1}, \textit{CPHX2}, \textit{DPRX}, \textit{DUXA} and \textit{DUXB}. Furthermore, two highly similar Tprx family genes \textit{TPRX1} and \textit{TPRX2} were predicted. We described previously the cDNA cloning of a major splice form for \textit{ARGFX}, \textit{CPHX1}, \textit{CPHX2}, \textit{DPRX}, \textit{DUXA} and \textit{DUXB}. In the present study, we cloned novel splice variants of \textit{NOBOX}, \textit{TPRX1} and \textit{TPRX2} together with additional splice variants for the \textit{ARGFX}, \textit{CPHX2}, \textit{DPRX} and \textit{DUXB} genes from human 8-cell embryos.

We amplified and cloned transcripts for each gene by PCR from non-sheared full-length cDNA libraries of three whole 8-cell embryos prepared according to the sequencing library preparation protocol by Tang et al. The validated transcripts are visualized in Fig. 1 and in UCSC Genome Browser views in Supplemental Figure S1.
Expression of PRD-like homeobox genes in early embryos and pluripotent stem cells. To investigate the developmental specificity of PRD-like homeobox gene expression, we used single cell RNA sequencing (RNAseq) on single blastomeres from two 8-cell stage embryos and two different hESC lines and compared the expression patterns (Fig. 2).

Table 1. Cloning of PRD-like homeobox genes. *TFE, Transcript far 5’ end, TFE ID refers to TSS data by Töhönen & Katayama, et al.33.

| Gene name | Full name & first reference | Gene ID | TFE ID | TFE Region | Strand | Forward/Reverse Primer | Clone origin |
|-----------|-----------------------------|---------|--------|------------|--------|------------------------|-------------|
| ARGFX     | arginine-fifty homeobox*    | 503582  | FE367949 | chr1:121286750-121286860 | +     | GAGAGACACACACAGTGAGACTCGAGAAATCCCAAGTCACCC | 8-cell embryo |
| CPHX1     | cytoplasmic polyadenylated homeobox 1* | FE200101 | chr1:675706315-75760435 | - | TCTCAGTACTTGCCTGCTCTCCTGAGACCTGCT | 8-cell embryo |
| CPHX2     | cytoplasmic polyadenylated homeobox 2* | FE200054 | chr1:675710885-75711165 | - | GAGGGTCAGACTTGCCTGCTGAGTTCAGGCTC | 8-cell embryo |
| DPRX      | divergent paired-related homeobox* | 503834  | FE227374 | chr1:945135194-54135433 | +     | TATCCCTGGACAGAACCCCAACATATACATTCACATCATCAGAC | 8-cell embryo |
| DUXA      | double homeobox A* | 503835  | FE262918 | chr1:57678787-57678897 | +     | GCTCAGGCTTCCAGGACTCTCTTCGAGACAGATTTGGGGTCCA | 8-cell embryo |
| DUXB      | double homeobox B* | 10033411 | FE200082 | chr1:675703289-75734520 | -     | TTATCGCCGTATCTCCCTGTCAGCGAGATTTGGGGTCCA | 8-cell embryo |
| NOBOX     | NOBOX oogenesis homebox* | 135935  | FE518599 | chr7:14410745-14410865 | -     | ATGGAGACGATCGAGATGAGACCCCTCACACAGG | Ovary |
| TPRX1     | Tetra-peptide repeat homeobox 1* | 284355  | chr1:48307780-48307850 | -     | TCAGGACTCGAGATGCAAGACGTCCACAGAGAAACGCCTC | 8-cell embryo |
| TPRX2     | Tetra-peptide repeat homeobox 2* | 503627  | chr1:48362000-48363000 | +     | TCAGGACTCGAGATGCAAGACGTCCACAGAGAAACGCCTC | 8-cell embryo |
In concordance with our previous results, we detected the expression of ARGFX, CPHX1, CPHX2, DPRX, DUXA, DUXB, NOBOX, TPRX2 and OTX2 in single blastomeres from 8-cell stage embryos. OTX2 was included as a positive control gene representing a previously characterized PRD-like homeobox gene. CPHX1, DPRX, DUXA, DUXB, and OTX2 expression was detected in all blastomeres with highest expression levels for DUXA and OTX2. ARGFX and CPHX2 expression was detected in the majority of blastomeres. NOBOX and TPRX2 expression was detected in one and four single blastomeres, respectively. The largest variation in expression levels among the different blastomeres was detected for ARGFX and DPRX, whereas DUXA and DUXB expression was at similar levels for all blastomeres.

In contrast to the expression of PRD-like homeobox genes in blastomeres, the expression in hESCs was much lower. Thus, CPHX1, CPHX2, TPRX2 and OTX2 expression was not detected in any of the two different hESC lines (HS980 and HS983a, 15 single cells each) at single cell resolution. ARGFX and DPRX expression was detected in only three or one single cell, respectively. Only DUXA was readily detectable in the hESCs, albeit only in a subset of the cells. DUXA was detected in seven out of 15 HS983a cells and in two out of 15 HS980 cells. The positive control OTX2 was detected...
in 12 out of 15 cells from both hESC lines. The expression levels detected in the hESCs were lower than that for human single blastomeres for all PRD-like homeobox genes detected. In order to independently confirm the developmental pattern of PRD-like homeobox gene expression, we applied PCR on an additional set of cDNA samples, including H9 and whole 8-cell stage embryo libraries. In this manner, we detected only low levels of expression of ARGFX, CPHX1 and DUXA in three hESC lines, including H9 (Supplementary Figure S6).

**The majority of cell types and tissues lack PRD-like homeobox gene expression.** To investigate mRNA expression of the PRD-like homeobox genes in different cell types and tissues, we used the FANTOM5 database\(^\text{16,17}\). Only OTX2 was shown to be expressed in the FANTOM5 database, whereas all the new genes were barely detectable in very few of the 1829 samples (Supplementary Table S1). Specifically, none of the new transcripts had an annotated transcriptional start site (TSS) in FANTOM5 data. Only one unannotated promoter corresponded to one of our genes, namely the DUXB TSS, but it lacked correct annotation due to the fact that there was no annotation for DUXB itself in FANTOM5. The lack of expression in FANTOM5 data suggested possible silencing of the genes in later stages of development. To investigate methylation as a possible silencing mechanism of the PRD-like homeobox genes, we considered the DNA methylation status of their promoters in human sperm, preimplantation embryos and embryonic tissues\(^\text{18}\). Percentage of methylated CpG-s from around 1000 bp from the homeobox genes TSS was plotted for all the samples and replicates with sequencing coverage of at least 5x (Supplementary Figure S7). The results show hypomethylation for all the detected homeobox genes in preimplantation embryos compared to various ESC-s and embryonic tissues. This suggests a rapid epigenetic silencing mechanism after the preimplantation development. Additionally we assessed the methylation status of these genes in blood cells\(^\text{19}\). The results show that in blood cells these genes were highly methylated, supporting the mechanism of epigenetic silencing in adult tissues (Supplementary Figure S8).

**Overexpression of the PRD-like homeobox genes in hESC identifies target genes.** We overexpressed the PRD-like homeobox genes in hESC to identify their target genes. The pFastBac vector used contained both the gene of interest and an eGFP fluorescent marker, enabling separation of transfected cells from non-transfected cells by Fluorescence Activated Cell Sorting (FACS). We could then verify the co-expression of the two proteins in the same cells by the simultaneous expression of eGFP and the red fluorescence gene mCherry control both by microscopy (Supplementary Figure S9a) and by FACS analysis (Supplementary Figure S9b). Successful sorting of the transfected versus non-transfected cells was further confirmed by mapping the sequencing reads to the original pFastBac vector backbone. The vast majority of reads from the GFP positive samples originated from the overexpression vector, while there were very few reads detected in the GFP negative samples (Supplementary Figure S9c).

After quality control and the exclusion of one disqualified GFP negative control sample, we performed the differential expression analysis as previously described\(^\text{20}\). We listed the differentially expressed TSSs from each comparison at \(p < 0.05\), and for robustness, we considered for further analyses the intersection of lists against three control conditions. The analysis yielded 18 gene sets in total: upregulated and downregulated gene sets for all the PRD-like homeobox genes and the pre-mRNA processing genes SRSF11 and KRT18, the pluripotency associated gene PRRC2C, the cell-cycle gene NASP and the homeodomain gene OTX2 itself (Fig. 3a).

Among the regulated target genes we identified 12 common target genes that were either upregulated by at least four or downregulated by at least five of the different PRD-like homeodomain transcription factors. These 12 common target genes included, among others, the cell-cycle gene NASP, the pluripotency associated gene DPPA3, the trophoblast determining gene KRT18, the pre-mRNA processing genes PRRC2C and SRSF11 and the homeodomain gene OTX2 itself (Fig. 3b).

Furthermore, there were some targets that were specific for a subset of the homeodomain genes. For example, three out of the four Yamanaka factors (KL14, MYC, POU5F1, SOX2) were differentially regulated by TPRX2, OTX2, DUXA and DPRX: SOX2 was downregulated by TPRX2 and OTX2, MYC was downregulated by OTX2, DUXA and DPRX, and POU5F1 was upregulated by DPRX and downregulated by OTX2. Interestingly, CPHX1 upregulated BCA31 and USP9X, two genes important for gamete generation\(^\text{21}\).

**The PRD-like homeodomain transcription factors may act both as activators and repressors.** The number of differentially expressed genes varied between the different PRD-like homeodomain transcription factors, with DPRX regulating the largest number of genes. The majority of DPRX target genes were downregulated and showed overlap with TPRX2 and ARGFX downregulated target genes (Fig. 3c). The smallest number of significantly regulated genes was found for DUXB. We found distinct roles for the PRD-like homeodomain genes as either mainly activators, such as ARGFX, CPHX1 and CPHX2 (Fig. 3a,c), or as mainly repressors, such as DPRX, DUXA and TPRX2 (Fig. 3a,c). For testing the application of our results to the context of human embryos, we looked at the expression of up-regulated genes by the three activators (CPHX1, CPHX2 and ARGFX) in human embryos\(^\text{22}\). An independent RNA sequencing dataset from human embryos and ESC-s showed up-regulation of the target genes in the course of embryo development for all three activators (Supplementary Figure S10). Full lists of the target genes, their corresponding TSSs and genomic annotations are given in Supplementary Table S2.

To further investigate the target gene profiles, we performed cross-comparisons between all upregulated or downregulated genes. The significance of the overlap was evaluated by Chi-squared test statistic or Fisher's exact test. A major overlap was found between the upregulated targets of CPHX1 and CPHX2 (\(p < 2 \times 10^{-14}\)) (Fig. 3d), and genes upregulated by DUXB and CPHX1 (\(p < 2 \times 10^{-8}\)), ARGFX and OTX2 (\(p < 2 \times 10^{-5}\)), and CPHX1 and NOBOX (\(p < 2 \times 10^{-5}\)).
In the same manner, the comparison between downregulated target genes yielded a number of significant overlaps; the target genes downregulated by DPRX, TPRX2, DUXB and NOBOX all have significant overlap between each other (Fig. 3c). Furthermore, DPRX and ARGFX share a significant number of downregulated target genes with all other PRD-like homeodomain genes except for CPHX1, CPHX2 (both) or DUXB (only ARGFX). We also tested the overlaps of the target genes upregulated by one factor and downregulated by another (Supplementary Figure S11). A significant overlap was observed between the downregulated targets of DPRX and upregulated targets of both CPHX1 (p < 2 × 10⁻⁸) and CPHX2 (p < 2 × 10⁻¹⁴). These data suggest tentative regulatory networks controlling human EGA.

**Target gene profiles can be influenced by the amino acid sequence of the homeodomain.** Since the members of the PRD-like homeobox gene family share similar amino acid sequences in the homeodomains (Fig. 3e), we hypothesized that they might target the same genes through similar DNA-binding properties. Thus,
we correlated the observed overlaps of the target genes with sequence similarities of the homeobox domains (Fig. 3c,d).

*CPHX1* and *CPHX2* shared a large number of upregulated target genes (Fig. 3c), and they had an almost identical homeodomain, with only 4 aa difference (Fig. 3e). *DUXA* and *DUXB* differed from the others in having two homeodomains. Only *DUXB* had the residues V-KN-A, but both *DUXA* and *DUXB* shared a number of target genes with the others (Fig. 3d).

*ARGFX* stood out with a different homeodomain sequence and significant overlap between its downregulated target genes and the target genes of *DPRX*, *TPRX2*, *OTX2*, *DUXA* and *NOBOX*.

Comparing the overall homeodomain sequence, *ARGFX* shared 35% and 33% aa identity with *DUXA*, whereas the homeodomains in *DUXB* shared 33% and 28% aa identity with *ARGFX*. The same degree of aa identity in the homeodomain (35%) was seen between *ARGFX* and *TPRX2*. This shows that although the similarities in the DNA binding domain indicate similar target gene profiles, the homeodomain similarity was not alone determining the target gene expression profiles.

**Target genes of PRD-like transcription factors are activated during human preimplantation development.** We suggested earlier that the new PRD-like homeobox genes might act as key regulators of human early development3. In order to further test this hypothesis, we analyzed the overlap between the differentially expressed target genes in our overexpression experiment and the genes upregulated at various stages of preimplantation development: EGA2,3, trophectoderm (TE), epiblast (EPI), primitive endoderm (PE) or inner cell mass (ICM)22 (Fig. 4). The similarity of targets was tested by Chi-squared test after matching the dataset-specific TSSs for EGA genes from Töhönen, *et al.* The lists of upregulated genes from Yan *et al.* and Blakeley *et al.* were calculated as described in materials and methods, and the Chi-squared test was performed by matching the gene names. The number of observed intersecting genes between the experimentally regulated gene sets and the preimplantation gene sets was calculated and compared with the expected number of intersecting genes occurring by chance. We report multiple testing corrected p-values.

Most of the comparisons with PE, ICM and EPI gene sets yielded no statistically significant overlaps with the experimental gene sets, except for OTX2. Eight target genes of OTX2 overlapped with ICM genes, while overlap of only 2 genes was expected (p = 0.012). Larger overlap than expected was noted for DPRX (110/72, p = 3.7 × 10−5) and TPRX2 (32/18, p = 0.045) with the TE genes. The most significant overlaps were observed for DPRX with both EGA datasets (Töhönen *et al.*3; 64/20, p = 5.5 × 10−25 and Yan *et al.*2; 64/24, p = 3 × 10−17).

**Activity of the homeobox genes via DNA motif implicated in preimplantation development.** The PRD-like homeobox genes were suggested to act on a motif found enriched in the promoters of EGA genes3. In order to test further the hypothesis, we studied the enrichment of the 36 bp embryo motif in the promoters of the experimental target genes (2000 bp upstream and 500bp downstream from TSS) by the MAST software23. The results confirmed the enrichment of the 36 bp motif upstream of the TSS for genes regulated by most of the PRD-like homeobox genes (Fig. 5a). This result further strengthened the notion of the functionality of the
predicted motif for PRD-like genes, and supported the 36 bp motif as a key regulatory element in human preimplantation development.

To further investigate the suggested main activators CPHX1, CPHX2 and ARGFX as transcription activators, we performed luciferase reporter assays after co-transfecting human embryonic kidney cells (HEK-293) with a transcription factor at a time and a reporter construct with ZSCAN4 promoter upstream of the luciferase gene. The ZSCAN4 promoter contained multiple repeats of the 36 bp motif, and ZSCAN4 was upregulated in EGA3. The ZSCAN4 promoter yielded up to 27 fold increase in luciferase expression when co-transfected with CPHX1, but not with CPHX2 or ARGFX, suggesting further specificity of function (Fig. 5b).

Discussion
The homeodomain proteins make up approximately 15–30% of all animal transcription factors24. A number of homeodomain proteins have been shown to be important for early development in vertebrate species25,26, and some homeodomain proteins are restricted in expression to specific cells at specific developmental stage. In our previous study we detected the expression of PRD-like homeobox genes expressed in oocytes, zygotes or single blastomeres from 4- and 8-cell embryos3. The transcript variants of the PRD-like homeobox genes ARGFX, DPRX, DUXA, DUXB, TPRX1 and TPRX2 cloned in this study may be expressed only during a restricted time period, explaining their recent discovery. A thorough investigation of human homeobox loci revealed that ARGFX, DPRX, TPRX1 and DUXA were initially predicted and annotated based on a few existing EST, retrotransposed pseudogene and genomic sequences7,10. Of the genes presented here only NOBOX was previously well characterized and found expressed at high levels in the ovary, specifically in primordial and growing oocytes27. Interestingly, our sequencing data indicated a new TSS yielding a novel splice variant of NOBOX that we confirmed by cDNA cloning. This novel NOBOX TSS had barely detectable expression in 4- or 8-cell embryos and embryonic stem cells.

In our database searches the PRD-like homeobox genes were undetectable in most cell types and tissues, including the comprehensive FANTOM5 database. In agreement with their absent expression, these genes were rapidly methylated in directly after preimplantation stage in human embryos18, suggesting an epigenetic silencing mechanism (Supplementary Figure S7). This observation was also consistent with the commonly observed silencing of Alu elements enriched in the promoters of the new genes3,28. During human preimplantation development the DNA methylation levels are reset, which is critical for the normal development29,30.

Figure 5. Promoter properties of the homeodomain genes. (a) Enrichment of a 36 bp DNA motif in the promoter region (~−2,000 ~ +500 bp distance around the center of the TFE) of the PRD-like homeodomain transcription factor target genes. The motif was enriched upstream of the promoters of genes during human pre-implantation development (red line). The motif is over-represented upstream of the up- and down-regulated genes by homeobox genes and under-represented from about 0 to 500 bp downstream from the TFE position. The motif is not enriched in any specific regions among random start-sites from the FANTOM database (blue dotted line). (b) Luciferase expression Fold Change between ZSCAN4 promoter-containing vector and corresponding empty vector (pGL4.25) with transfection by activators CPHX1, CPHX2 and ARGFX in HEK293 cell lines. The average values from three biological replicas are shown, error bars represent standard deviation. The values are normalized to corresponding vector without transcription factor overexpression.
Due to ethical reasons, functional studies in human embryos were not possible for these genes. However, we chose the closest and biologically most relevant cell line hESC-s for overexpression experiment. The PRD-like homeobox genes could induce both up- and downregulation of target genes based on the hESC overexpression experiment (Fig. 3a), and the target genes showed significant overlaps both between the different factors themselves (Fig. 3c,d) and between previously published human preimplantation data sets (Fig. 4). While the target genes show expected overlaps with published datasets from embryos, we have to acknowledge that some of the identified target genes might not be the same in human embryos due to the usage of cell model. Since the expression of PRD-like homeodomain transcription factors was very low or not at all detected in hESC-s (Supplementary Figure 6), some of the target genes in human embryos might not be overexpressed due to the cellular context such as due to methylation of the gene region.

In the hESCs, DPRX was found to downregulate the largest number of target genes (2067), followed by TPRX2 (464) and DUXA (362). We propose that DPRX might be needed to restrict the expression of a number of activated genes after the EGA, allowing later lineage specification.

The luciferase reporter assay using a putative ZSCAN4 promoter placed upstream of the luciferase gene further confirmed that CPHX1, but not the other activators CPHX2 and ARGFX acts on ZSCAN4 promoter (Fig. 5b). Thus the activity of these factors was not determined only by their homeodomain, but rather a combination of homeodomain specificity, possible transactivating domains in the protein or their cellular context and interaction with other proteins. For example, OTX2 has been shown to interact on protein level with LHX1 and FOXA2. Thus, the cellular context may affect the transcriptional activity of the PRD-like homeodomain proteins.

We studied the possibility that the overlap of target genes for the PRD-like homeobox genes might result from the similarities in the DNA-binding regions of their homeodomains. We observed a large number of genes upregulated by CPHX1 and CPHX2 and they differed in six positions only with a high overall protein sequence similarity (Supplementary Figure S4 a). ARGFX, on the other hand, only shared 17 aa identity with the homeodomains of CPHX1 and CPHX2, possibly explaining the low overlap of target genes for these transcription factors even though all three seemed to function mainly as activators (Fig. 3c,d). As an example, ZSCAN4, a known inducer of iPS cells was significantly upregulated by CPHX1 only. Further experimental evidence is needed in order to define the specificity of the target gene profiles of the homeodomain proteins.

It is known that an aa change at position 52 in the DNA-recognition part of the homeodomain in NOBOX may cause premature ovarian failure due to disruption of the transcription factor binding to the NOBOX-binding consensus sequence. When aligning the different homeodomains for our 9 transcription factors we see that at position 52 the aa varies between the different factors. For example, NOBOX, ARGFX, OTX2, TPRX2 and DUXA have an arginine residue at position 52, just as NOBOX. The overall highly similar CPHX1 and CPHX2 differ at this position (lysine or asparagine).

PRD-like transcription factors have been implicated in early development in other organisms such as Drosophila melanogaster. For example, the Drosophila prd-like homeobox gene Odysseus is involved in germ cell formation. Interestingly, we found that CPHX1 upregulates the expression of BCAP31 and USP9X, whereas CPHX2 upregulates SPIN3, all three genes involved in human gamete generation according to DAVID annotation. In addition, BCAP31 is downregulated by DUXB and TPRX2. Also other germ cell related genes, such as ZFP42, TEX14 and TFFAP2 were regulated in our overexpression experiment. Targets of the PRD-like homeodomain transcription factors were enriched in genes upregulated during early development (Fig. 4). OTX2 targets were enriched in ICM, TPRX2 targets in TE and DPRX targets in TE and EGA.

In conclusion, this study presents further evidence of expression for the nine PRD-like genes that we recently identified from human preimplantation embryos. Our data support the hypothesis that the previously uncharacterized PRD-like homeobox genes are specifically expressed during early development, have a key role in the transcriptional dynamics during EGA, and are subsequently silenced as the development progresses from a pluripotent stage towards more highly differentiated stage.

Materials and Methods

Human embryos. Human embryos used in this study were collected in Sweden and donated by informed consent by couples who underwent infertility treatment by in vitro fertilization (IVF). Cryopreserved cells that were not needed for IVF treatment were donated as an alternative to being destroyed as they had exceeded the maximum legal storage time. Cleavage stage embryos were frozen at the 4-cell stage on day 2 after fertilization. After thawing, embryos were allowed to develop until the 8-cell stage in a sequential culture system until reaching the maximum legal storage time. Cleavage stage embryos were frozen at the 4-cell stage on day 2 after fertilization.

Library preparation. Human embryo libraries were prepared after thawing the embryos at the 4-cell stage and incubation allowing them to develop until the 8-cell stage. Each single embryo was put into a 0.5 mL PCR tube containing 4.45μl of freshly prepared cell lysis buffer and the libraries were processed according to the protocol by Tang et al.12. In total, three 8-cell libraries were prepared.

Human ovary poly-A+ RNA was obtained from Clontech (cat# 636152). Human ovary cDNA was prepared using a first strand cDNA synthesis kit (Invitrogen) according to the manufacturer’s instructions.

PCR primer design. The putative novel transcripts were predicted based on our previous RNAseq data3 and these predictions were used for designing PCR cloning primers.

NOBOX primers were designed based on an RNAseq peak at FE518599 (chr7:144100745-144100865,-) located in an intron of the human RefSeq sequence NM_001080413.3 (GeneID135935), that was connected to the 3’ UTR of the predicted sequence. The sequences containing ORFs for CPHX1, CPHX2 and DUXB were predicted.
First-Strand Synthesis Supermix for qRT-PCR (Invitrogen) according to the manufacturer’s instructions. For PCR lines (H9, HS401 and HS980) in three biological replicates were used for cDNA synthesis using SuperScript III (GeneID 503834) and NM_001012659.1 (GeneID 503582) respectively (Table 1).

Cloning of ARGFX, CPHX1, CPHX2, DPRX, DUXA, DUXB and TPRX2. cDNA libraries from three single 8-cell embryos were used for cloning of the putative transcripts. The transcripts were amplified from the cDNA libraries using Phusion High-Fidelity DNA polymerase (New England Biolabs) according to the manufacturer’s instructions. For amplification of DUXA, DPRX, DUXB, TPRX1 and TPRX2, the following PCR program was used: 98 °C for 30 s; 40 cycles of 98 °C for 10 s, 65.9 °C (DUXB)/67.9 °C (DPRX, TPRX1, TPRX2)/71.5 °C (DUXA) for 30 s, 72 °C for 1 min; final extension 72 °C for 10 min. ARGFX, CPHX1 and CPHX2 were amplified using Touchdown PCR: 98 °C for 30 s; 24 cycles of 98 °C for 10 s, annealing for 30 s, temperature decreasing from 63 °C to 56 °C, 1 °C/3 cycles, 72 °C for 30 s; 16 cycles of 98 °C for 10 s, 55 °C for 30 s, 72 °C for 30 s; final extension 72 °C for 10 min. In order to clone the three differently sized TPRX2 isoforms, the PCR products were purified from agarose gel using the QIAquick Gel Extraction Kit (Qiagen), and the obtained amplicons were then reamplified in order to obtain sufficient amount of amplicons for subsequent cloning. PCR products were cloned into the pCRABlunt-TOPO vector using the Zero Blunt TOPO PCR Cloning kit (Invitrogen) and verified by Sanger sequencing (Eurofins Genomics).

Cloning of NOBOX. Human ovary cDNA was used for cloning putative NOBOX transcript. The transcript was amplified from cDNA using HotStarPlus Taq DNA polymerase (Qiagen) according to manufacturer’s instructions. For PCR amplification following steps were applied: 95 °C for 5 min; 40 cycles of 95 °C for 30 s, 60 °C for 30 s, 72 °C for 1 min; final extension 72 °C for 10 min. PCR product was cloned into pCRRII-dual promoter TOPO vector using TOPO TA cloning kit (Invitrogen), and sequence was verified by Sanger sequencing (Eurofins Genomics).

Prediction of ORFs, functional protein domains, and alignment on UCSC genome browser. For open reading frame (ORF) prediction, forward and reverse sequences from each clone were first trimmed for vector sequences using Pregap4 version 1.6-r and then the contig sequences were formed using Gap4 v4.11.2-r (both from the Staden package, http://staden.sourceforge.net/). Mismatching bases were manually checked for quality and edited accordingly. ORFs were predicted and translated using ApE plasmid editor (http://biology.labs.utah.edu/jorgensen/wayne/ape/). Conserved protein domains were predicted applying an NCBI Blastx (http://blast.ncbi.nlm.nih.gov/Blast.cgi) translated nucleotide query on each complete clone sequence against non-redundant protein sequences (nr) and EMBL-EBI InterPro (http://www.ebi.ac.uk/interpro/) on translated sequences. The consensus cDNA and amino acid sequences were aligned by Blat with “psl” as an output (UCSC Genome Browser) and the information was combined to generate a BED file for visualization of custom tracks on a browser (Supplemental Text S2). The alignments were edited manually in order to include short exons.

Gene expression profiling of hESC lines and single 8-cell blastomeres. A total of 48 single cells were manually picked directly to STRT lysis buffer. The cells included 16 human regular ES cells (HS980), 16 human single-8-cell blastomere derived ES cells (HS983a) and 16 human single blastomeres from 8-cell embryos. The embryos were thawed at 4-cells stage (ThawKit™ Cleave, VitroLife) and cultured in G-1™ Plus media (VitroLife) overnight under standard conditions as performed in the IVF Clinic (5% CO2/5%O2). The zona pellucidae were removed with Tyrode’s solution (Sigma-Aldrich) and embryos were dissociated into single cells using enzymatic (TrypLE, Life technologies) and mechanical separation. The STRT library was prepared according to the modified STRT protocol (Krjutškov, et al., accepted) and sequenced on four lanes of Illumina HiSeq 2000 instrument. Pre-processing of STRT reads, quality control and alignments were performed as previously described (Krjutškov, et al., accepted). After quality control, 4 outliers (one regular ES cell, one blastomere-derived ES cell and two 8-cell blastomeres) were removed from further analysis due to low RNA content. The gene expression was quantified per transcript far 5’ end (TFE) units\(^3\). The TFE coordinates for the PRD-like homeobox genes are given in Table 1. The reference file for counting the reads per TFE was defined for PRD-like homeobox genes and spike-in RNAs only. The reads per TFE were counted using SAMtools version 1.1 and htsesq-count version 0.6.1, and the data was normalized using external spike-in RNAs as described previously\(^20\). The raw read counts and normalized expression values for PRD-like homeobox genes and spike-in RNAs used for creating beeswarm plots are given in Supplemental Table S3. The original sequence files as well as aligned BAM-files are accessible from European Nucleotide Archive (ENA) under the accession number PRJEB12467.

Analysis of methylation pattern for homeobox genes. Publicly available dataset for human sperm, preimplantation embryo, ESC and embryonic tissues was downloaded from GEO (accession number GSE51239, file hDev100bpMeFinal.txt.gz)\(^18\). 100 bp methylation value percentages were used for genomic positions in the range of 500 bp up- and downstream from the middle of TFE-s for the homeobox genes.

Publicly available dataset for methylation in human blood cells was downloaded\(^19\). The M-values within the entire homeobox genes’ region was used.

PCR detection of PRD-like homeobox gene expression. Total RNA extracted from three hESC lines (H9, HS401 and HS980) in three biological replicates were used for cDNA synthesis using SuperScript III First-Strand Synthesis Supermix for qRT-PCR (Invitrogen) according to the manufacturer’s instructions. For PCR
we used an input of 22 ng of cDNA and the reactions were done in three biological replicates. As a control, we used 10 ng cDNA from a human 8-cell cDNA library. The PCR reactions were carried out using an ABI PRISM 7500 Fast Real-Time PCR System with FastStart Universal SYBR Green Master mix (Roche) according to the manufacturer’s instructions. The primer sequences are shown in Supplemental Table S4. For validation of the amplicons, the PCR products were cloned into a pCRII-1 dual promoter TOPO vector using the TOPO TA cloning kit (Invitrogen), and verified by Sanger sequencing (Eurofins Genomics).

**hESC expression vector construction.** In order to overexpress ARGFX, CPHX1, CPHX2, DPRX, DUXA, DUXB, NOBOX, OTX2 and TPRX2 in mammalian cells, the sequences were cloned into a bicistronic pFastBac expression vector. The pFastBac vector was modified as described previously. Briefly, the transcript sequences were amplified from the TOPO vectors using primers containing Ascl and PacI restriction sites, digested using Ascl and PacI (New England Biolabs) and ligated into pFastBac vector. The primer sequences for cloning the genes in this study are shown in Supplemental Table S5. mCherry fluorescent protein was used as a control, and was amplified from the standard injection marker construct elt-2::GFP for C. elegans (kindly provided by Gert Jansen, The Erasmus University Medical Center, Holland). The final structure of the modified expression pFastBac vector was as follows: CMV enhancer – EF1α promoter – Ascl – Gene Of Interest – PacI – IRES – eGFP – WPRE. The IRES (internal ribosomal entry site) element allows for simultaneous expression of two proteins: the inserted gene of interest, and a green fluorescent protein (GFP) marker.

**Overexpression experiments and cell sorting for expression profiling.** The hESC line HS401 was cultured on Laminin–521 (Biolamina) in mTeSR™-1 media (Stemcell Technologies). One μg of expression vector containing the gene of interest was incubated with 3 μl of Lipofectamine2000 in 50 μl of DMEM. Confluent cells (95–100%) were trypsinized, gently washed with DPBS, and re-suspended in 50 μl of transfection solution. The cell suspension was transferred to a new Laminin–521 coated plate with 50 μl of DMEM and the cells were allowed to settle in the transfection solution for 15 minutes. The media was subsequently changed to mTeSR™-1. After 9–11 h of transfection, the cells were washed with DPBS and treated with trypsin to get single-cell solution, washed again with DPBS and suspended in ice-cold DPBS until the analysis. The overexpression time point was chosen by time-course analysis using pFastBac-mCherry vector and further estimation when the first wave of target genes would be expressed. Cell viability was assessed by 1:1000 propidium iodide (Molecular Probes) and cell clusters were removed by FSC and SSC. 75 cells per sample from GFP positive and GFP negative cells were sorted in three replicates into 5 μl of library lysis buffer. The sorting was performed in three replicates from the same transfection well for 75 transfected and 75 non-transfected cells in each replica (altogether 6 samples per gene) for each gene and mCherry control (n = 9), providing two libraries for pooled sequencing of 54 samples by modified STRT protocol (Krjutškov, et al. accepted).

The homeobox genes were distributed between two libraries that were processed simultaneously. The first of these, Library 3, included DUXA, DPRX, ARGFX, OTX2 and mCherry. The second library, Library 4, included CPHX1, CPHX2, DUXB, TPRX1, NOBOX and mCherry.

**STRT RNAseq analysis for target gene detection.** The STRT data was analyzed as described previously. Briefly, the reads were filtered, de-multiplexed by barcodes, joined for the same unique molecular identifiers (UMI), trimmed for the barcodes and UMI-s, and mapped to the human UCSC genome hg19 by TopHat. Outliers were identified by analysis of distance correlation dendrograms, which indicated two outliers for one library (Library 3, Supplementary Figure S12a) and no outliers for the second library (Library 4, Supplementary Figure S12b). The outlier E8 in Library 3 (mCherry, GFP negative) was removed, resulting in one library having one less GFP negative control. The principal components one and two separated two different DPRX wells from the controls, and the components 3 and 4 separated none, therefore DPRX was not removed.

The signal of STRT sequencing forms clusters that were identified as Transcript Far 5′ End (TFE) as described previously (Krjutškov, et al. accepted). Briefly, the aligned BAM files were assembled into putative transcripts by sample types using Cufflinks, followed by extraction of only the 5′-end exons by UCSC tools and followed by merging reads from various sample types by BEDTools. The reads were counted on the identified TFE-s by htseq, in order to confirm overexpression in the GFP positive versus GFP negative cells.

Three sets of controls were used for performing differential analysis: only mCherry control GFP positive wells (n = 3), only GFP-negative wells from the whole library (n = 22 or n = 23), and the combination of both mCherry control and homeobox specific GFP-negative wells (n = 3 + 3). Estimates of differential expression for the TFE-s were obtained using the R package SAMstr with positive samples compared to each of the three control (FDR < 0.05 unless indicated otherwise). Only intersection of the differentially expressed TFE-s from three analyses was used in the study. The identifier TFE-s were mapped to the genome and annotated for genes and genomic regions as described previously. Only the reads mapping to 5′ UTR of coding genes were used in the study.

The original sequence files as well as aligned BAM-files are accessible from ENA by the accession number PRJEB12453.

**Gene expression analysis for public preimplantation datasets.** Three publicly available datasets were used for gaining differentially expressed genes lists. Two independent studies were used for gaining the EGA genes lists. The TFE identifiers up-regulated by either 4- or 8-cell stage were gained from Töhönen et al. and the identifiers were matched with at least 1 bp overlap by BEDTools.

The RPMK values were downloaded from Yan et al. and up-regulated genes by either 4- or 8-cell stage were calculated by students t-test (FDR < 0.05). Only gene names that had a match in our dataset were used for the
The transfections were performed using Lipofectamine 2000 (Invitrogen) 0.5 μg as follows: Luciferase vector 100 ng/well, pFastBac vector 100 ng/well and Renilla luciferase vector 10 ng/well. This was co-transfected with other constructs to enable normalization. The concentrations for single constructs were determined previously.

Luciferase reporter assay. HEK-293 cells (ATCC) were seeded on 48-well plates in DMEM containing 1 g/l glucose, L-glutamine, pyruvate and supplemented with 10% FBS and 2 mM L-glutamine (all from Gibco). Cells were grown overnight at 37 °C in 5% CO₂ and transfected with the ZSCAN4 promoter in pGL4.25 reporter vector described previously or pGL4.25 luciferase vector (Promega) in combination with CPHX1, CPHX2 or ARGFX, all cloned into pFastBac vector as described above. Renilla luciferase vector pGL4.74 [hLuc/TK] (Promega) was co-transfected with other constructs to enable normalization. The concentrations for single constructs were determined as follows: Luciferase vector 100 ng/well, pFastBac vector 100 ng/well and Renilla luciferase vector 10 ng/well. The transfections were performed using Lipofectamine 2000 (Invitrogen) 0.5 μg/well according to manufacturer’s instructions. Cells were incubated at 37 °C in 5% CO₂, harvested 24 h after transfection and subjected to Dual luciferase assay (Promega) in three biological replicates with two technical replicates each, according to manufacturer’s protocol. Luciferase signals were measured using a TECAN infinite M200 (Tecan, Männedorf, Switzerland).

Data availability statement. All the novel homeobox genes clone sequences were submitted to EMBL, and their mRNA accession numbers are given in Fig. 1. The short read sequences and aligned BAM-files from the RNA-seq experiments for the overexpression in ES-cells are uploaded in ENA with the accession number PRJEB12453. The RNA-seq dataset from single-cell hESC-s and 8-cell human embryo have the accession number PRJEB12467.

References
1. Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. & Pera, R. A. Human pre-implantation embryo development. Development 139, 829–841, doi: 10.1242/dev.060426 (2012).
2. Yan, L. et al. Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol, doi: 10.1038/nsmb.2660 (2013).
3. Tohonen, V. et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun 6, 8207, doi: 10.1038/ncomms9207 (2015).
4. Burglin, T. R. & Affolter, M. Homeodomain proteins: an update. Chromosoma, doi: 10.1007/s00412-015-0543-8 (2015).
5. Holland, P. W. Evolution of homeobox genes. Wiley Interdiscip Rev Dev Biol 2, 31–45, doi: 10.1002/wdev.78 (2013).
6. Burglin, T. R. A Handbook of Transcription Factors, Chapter 5: Homeodomain Subtypes and Functional Diversity. Vol. 52 (Springer Netherlands, 2011).
7. Booth, H. A. & Holland, P. W. Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. Gene 387, 7–14, doi: 10.1016/j.gene.2006.07.034 (2007).
8. D’Elia, A. V. et al. Missense mutations of human homeoboxes: A review. Hum Mutat 18, 361–374, doi: 10.1002/humu.1207 (2001).
9. Qian, Y. Q. et al. Nuclear magnetic resonance spectroscopy of a DNA complex with the uniformly 13C-labeled Antennapedia homeodomain and structure determination of the DNA-bound homeodomain. J Mol Biol 234, 1070–1083, doi: 10.1006/jmbi.1993.1660 (1993).
10. Holland, P. W., Booth, H. A. & Bruford, E. A. Classification and nomenclature of all human homeobox genes. BMC Biol 5, 47, doi: 10.1186/1741-7007-4-47 (2007).
11. Li, G. & Holland, P. W. The origin and evolution of ARGFX homeobox loci in mammalian radiation. BMC Evol Biol 10, 182, doi: 10.1186/1471-2148-10-182 (2010).
12. Tang, F. et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat Protoc 5, 516–535, doi: 10.1038/nprot.2009.236 (2010).
13. Huntriss, J., Hinkins, M. & Picton, H. M. cDNA cloning and expression of the human NOBOX gene in oocytes and ovarian follicles. Mol Hum Reprod 12, 283–289, doi: 10.1093/molehr/gal035 (2006).
14. Zhong, Y. F. & Holland, P. W. The dynamics of vertebrate homebox gene evolution: gain and loss of genes in mouse and human lineages. BMC Evol Biol 11, 169, doi: 10.1186/1471-2148-11-169 (2011).
15. Rodin, S. et al. Clonal culturing of human embryonic stem cells on laminin-521/E-cadherin matrix in defined and xeno-free environment. J Mol Biol 283–289, doi: 10.1093/molehr/gal035 (2006).
16. Forrest, A. R. et al. A promoter-level mammalian expression database. Nature 507, 462–470, doi: 10.1038/nature13182 (2014).
17. Lizio, M. et al. Gateway to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 16, 22, doi: 10.1186/s13059-015-0560-6 (2015).
18. Smith, Z. D. et al. DNA methylation dynamics of the human preimplantation embryo. Nature 511, 611–615, doi: 10.1038/nature13581 (2014).
19. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One 7, e41361, doi: 10.1371/journal.pone.0041361 (2012).
20. Katayama, S., Tohonen, V., Linnarsson, S. & Kere, J. SAMstr: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics 29, 2943–2945, doi: 10.1093/bioinformatics/bts511 (2013).
21. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25–29, doi: 10.1038/75556 (2000).

Motif analysis. The MAST tool was used to search for significant sequence patterns to all given motifs in a gene set. Sequences – 2000 and +500 bp from the center of each TF for differentially regulated genes (FDR < 0.1) were analyzed. A list of all transcription start sites was gained from FANTOM database, and used as a control set to the THE-s.
Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos.

The authors declare no competing financial interests.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Madissoon, E. et al. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci. Rep. 6, 28995; doi: 10.1038/srep28995 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
Corrigendum: Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos

Elo Madissoon, Eeva-Mari Jouhilahti, Liselotte Vesterlund, Virpi Tööhonen, Kaarel Krjutškov, Sophie Petropoulos, Elisabet Einarsdottir, Sten Linnarsson, Fredrik Lanner, Robert Månsson, Outi Hovatta, Thomas R. Bürglin, Shintaro Katayama & Juha Kere

Scientific Reports 6:28995; doi: 10.1038/srep28995; published online 14 July 2016; updated 02 September 2016

The original version of this Article contained a typographical error in the spelling of the author Sophie Petropoulos, which was incorrectly given as Sophie Petropoulous. This has now been corrected in the PDF and HTML versions of the Article.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016