HAL
open science

A Core Response to the CDX2 Homeoprotein During Development and in Pathologies

Victor Gourain, Isabelle Duluc, Claire Domon-Dell, Jean-Noël Freund

To cite this version:

Victor Gourain, Isabelle Duluc, Claire Domon-Dell, Jean-Noël Freund. A Core Response to the CDX2 Homeoprotein During Development and in Pathologies. Frontiers in Genetics, 2021, 12, pp.744165. 10.3389/fgene.2021.744165 . hal-03788711

HAL Id: hal-03788711

https://hal.science/hal-03788711v1

Submitted on 26 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 25 October 2021
doi: 10.3389/fgene.2021.744165

A Core Response to the CDX2 Homeoprotein During Development and in Pathologies

Victor Gourain[1†], Isabelle Duluc[2], Claire Domon-Dell[2] and Jean-Noël Freund[2]*

[1]Karlsruhe Institute of Technology, Institute of Biological and Chemical Systems, Karlsruhe, Germany, [2]Université de Strasbourg, Inserm, IRFAC / UMR-S1113, FHU ARRIMAGE, FMTS, Strasbourg, France

Whether a gene involved in distinct tissue or cell functions exerts a core of common molecular activities is a relevant topic in evolutionary, developmental, and pathological perspectives. Here, we addressed this question by focusing on the transcription factor and regulator of chromatin accessibility encoded by the Cdx2 homeobox gene that plays important functions during embryonic development and in adult diseases. By integrating RNAseq data in mouse embryogenesis, we unveiled a core set of common genes whose expression is responsive to the CDX2 homeoprotein during trophectoderm formation, posterior body elongation and intestinal specification. ChIPseq data analysis also identified a set of common chromosomal regions targeted by CDX2 at these three developmental steps. The transcriptional core set of genes was then validated with transgenic mouse models of loss or gain of function of Cdx2. Finally, based on human cancer data, we highlight the relevance of these results by displaying a significant number of human orthologous genes to the core set of mouse CDX2-responsive genes exhibiting an altered expression along with CDX2 in human malignancies.

Keywords: homeobox gene, embryo, cancer, gene expression, chromatin targets

OPEN ACCESS

Edited by:
Hauke Busch,
University of Lübeck, Germany

Reviewed by:
Claudia Pommerenke,
German Collection of Microorganisms and Cell Cultures GmbH (DSMZ), Germany
Christopher Fields,
University of Illinois at Urbana-Champaign, United States

****Correspondence:***
Jean-Noël Freund
jean-noel.freund@inserm.fr

†Present Address:
Victor Gourain,
Centre de recherche en transplantation et immunologie, UMR 1064, Nantes, France

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *20 July 2021*
Accepted: *07 October 2021*
Published: *25 October 2021*

Citation:
Gourain V, Duluc I, Domon-Dell C and Freund J-N (2021) A Core Response to the CDX2 Homeoprotein During Development and in Pathologies. Front. Genet. 12:744165. doi: 10.3389/fgene.2021.744165

INTRODUCTION

That evolution makes new out of old suggests the existence of shared properties between the functions of a given gene at its different times or sites of action. The homeobox gene encoding the CDX2 transcription factor allows addressing this assumption since it drives three major developmental processes in mammals. At the blastula stage, Cdx2 is pivotal during the segregation of pluripotent cells into the first two lineages by acting downstream of the lineage allocation process between trophectodermal and inner mass cells to repress Oct4 and Nanog in the trophectoderm (Niwa et al., 2005; Strumpf et al., 2005; Ralston and Rossant, 2008). Then, Cdx2 actively participates in axial posterior body growth at gastrulation through a convergent effect with T-Brachyury to maintain stemness properties of neuro-mesodermal axial progenitors and to sustain Fgf and Wnt signaling (van Rooijen et al., 2012; Amin et al., 2016). Finally, Cdx2 determines intestinal identity of the mid-/hindgut endoderm in embryos and allows identity maintenance of the adult gut epithelium by regulating the proliferation of stem/progenitor cells and the differentiation of mature enterocytes (Gao et al., 2009; Verzi et al., 2010; Stringer et al., 2012). Molecularly, the CDX2 protein has been shown to bind the proximal promoter of a number of target genes, as first uncovered with the intestinal sucrase-isomaltase gene (Suh et al., 1994). In addition, it also binds distant chromatin regions to prevent epigenetic silencing and keep chromatin domains open and active (Saxena et al., 2017).

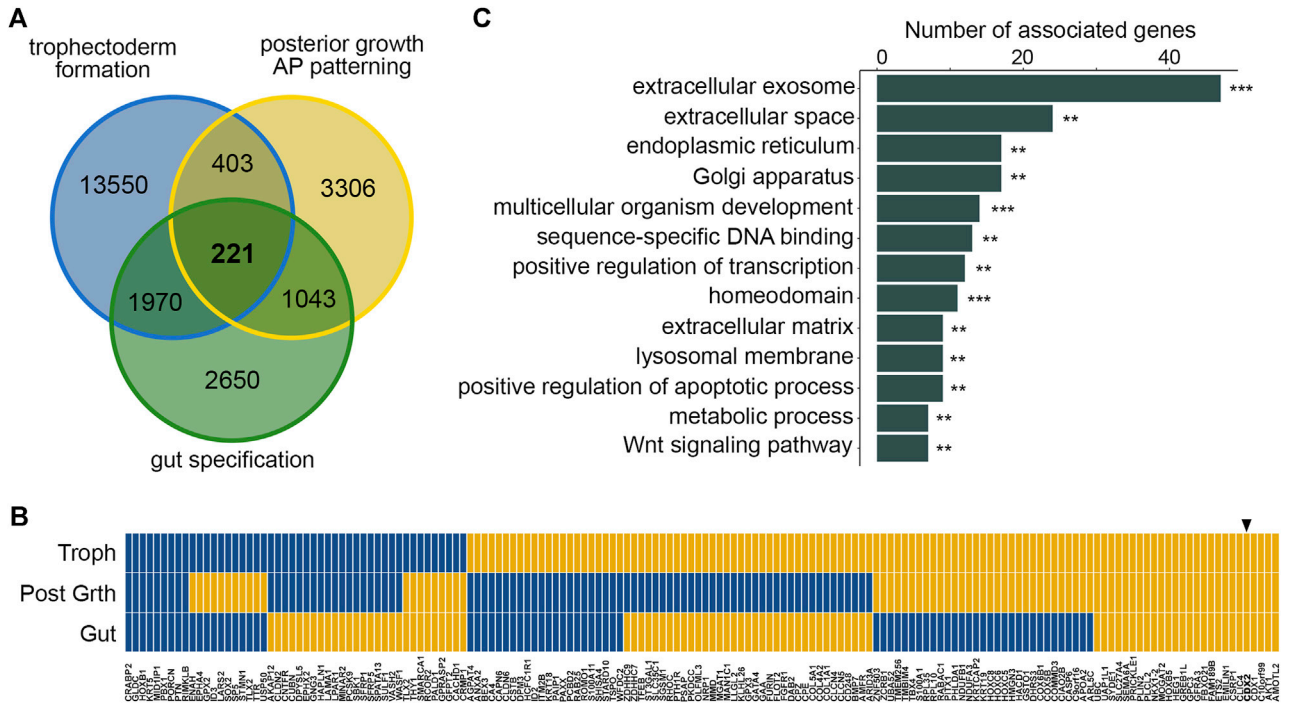

FIGURE 1 | A core set of DEGs responsive to CDX2 in mouse development. (**A**) Comparison of differentially expressed genes between Cdx2-overexpressing vs wild type ES cells (trophectoderm formation; blue), E8 Cdx-null vs wild type embryos (posterior growth, AP patterning; yellow) and intestinal epithelial cells of E16 Cdx2$^{-/-}$ vs wild type embryos (gut specification; green) showing the core set of 221 mouse DEGs. (**B**) Correlation map of the expression changes of the mouse core set of 162/221 DEGs having human orthologues, at the three developmental steps (yellow: up-regulation; blue: down-regulation): trophectoderm formation (Troph; Cdx2-overexpressing vs wild type ES cells), posterior growth (Post Grth; E8 wild-type vs Cdx-null embryos) and gut specification (Gut; E16 wild-type vs Cdx2$^{-/-}$ embryos). At each developmental stage, the map illustrates the comparison of the Cdx2-expressing samples vs the Cdx2 non-expressing counterparts. The arrowhead shows CDX2. (**C**) Ontology analysis of the set of 162 human orthologues to the mouse core set of DEGs. $^{**}p < 0.01$, $^{***}p < 0.001$.

While physiologically restricted to the gut epithelium in adults, CDX2 expression becomes reduced and heterogeneous in human colorectal cancer, particularly in tumors with the worst prognosis (Baba et al., 2009; Dalerba et al., 2016; Balbinot et al., 2018). This reduction facilitates tumor progression, as shown in mouse models of intestinal cancer, indicating a tumor suppressor role in the gut (Bonhomme et al., 2003; Sakamoto et al., 2017; Balbinot et al., 2018). Inversely, CDX2 is ectopically turned on outside the gut in precancerous intestine-type metaplasia and associated adenocarcinoma of foregut-derived organs including stomach and esophagus (Moskaluk et al., 2003), even though patients survival correlates with the CDX2 level in gastric cancers (Seno et al., 2002). Beside the upper digestive tract, CDX2 is also ectopically expressed in 80% of acute myeloid leukemia (AML) irrespective of the cytogenetic group but correlating with disease burden (Scholl et al., 2007). Thus, unlike the gut, CDX2 has on oncogenic effect in the hematopoietic lineage, as recently demonstrated in mice (Vu et al., 2020; Galland et al., 2021).

On this basis, the present work interrogates whether some elements of the response to CDX2 are shared during the successive steps of embryonic development in mice and subsequently whether these elements are altered in human pathologies along with CDX2.

RESULTS

A Core Set of Genes Responsive To CDX2 During Mouse Development

To address if there is a common set of genes responsive to the CDX2 transcription factor during its successive functions in mouse embryogenesis, we analyzed publicly available RNAseq data related to trophectoderm formation, posterior growth, and intestinal fate determination (see **Supplementary Table S1.1**). For this purpose, we compared the consequences of Cdx2 overexpression in embryonic stem (ES) cells (Cambuli et al., 2014; Rhee et al., 2017), of Cdx loss of function in E8 growing embryos (Amin et al., 2016), and of Cdx2 deficiency in the intestinal endoderm of E16 embryos (Banerjee et al., 2018). With $|\log_2\text{fold-change}|>2$ and $p < 0.05$, a core set of 221 differentially expressed murine genes (DEGs), corresponding to 162 human orthologues, was identified in common between these three conditions (**Figure 1A**; **Supplementary Table S1.2**). Interestingly, the up or down expression changes of the DEGs were not always consistent at the three developmental steps, indicating a context-dependent response to CDX2 (**Figure 1B**; **Supplementary Table S1.3**). Ontology enrichment analysis of the 162 human orthologues revealed a significant association with "extracellular exosome", "extracellular matrix", "multicellular

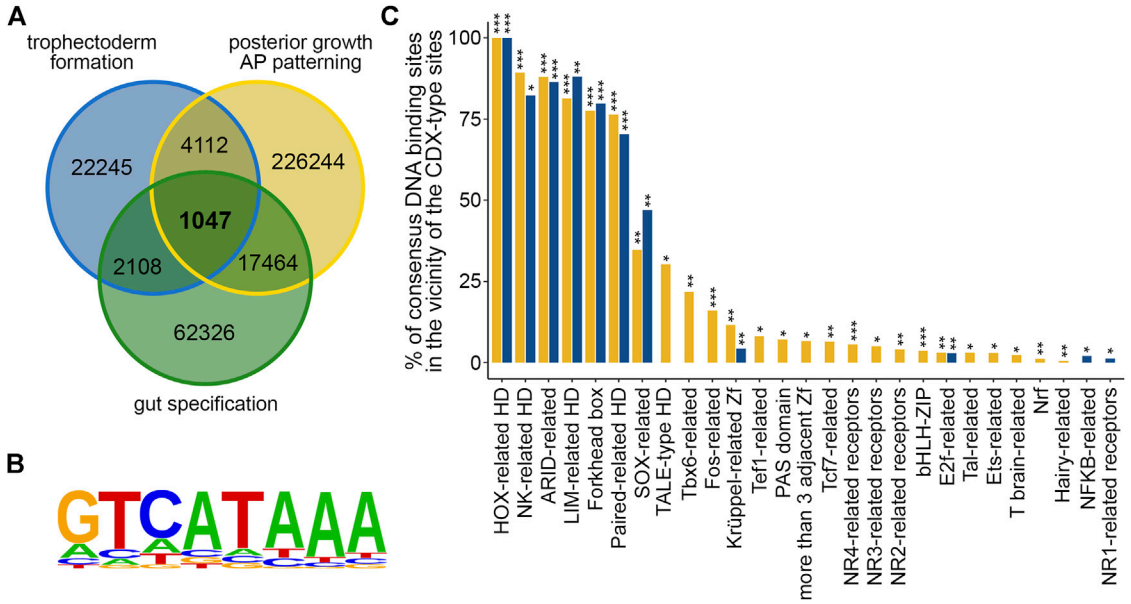

FIGURE 2 | ChIPseq targets of CDX2 during mouse development. (**A**) Comparison of the CDX2-associated chromosomal regions during trophectoderm formation (trophectoderm formation; blue), embryonic posterior elongation (posterior growth, AP patterning; yellow) and gut formation (gut specification; green) showing the core of 1,047 common regions. (**B**) Consensus sequence of the 1801 CDX-type motifs present in the 1047 CDX2-bound chromosomal regions. (**C**) Enrichment in consensus binding motifs for the indicated transcription factors in the vicinity (+/50 bp) of the 1801 CDX-type sites present in the common ChIPseq regions (yellow) and in the vicinity of the 1340 CDX-type sites present in the promoters of the 221 common DEGs (blue). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.00001$. HD: homeodomain; Zf: zinc finger; h: hormone.

organism development", "sequence-specific DNA binding", "gene regulation", "metabolic process" and "Wnt signaling" (**Figure 1C**; **Supplementary Table S1.4**). Twenty-eight genes of the DEGs core encoded nuclear proteins involved in chromatin conformation, DNA transcription and repair (Arid3a, Bmyc, Cdx1, Cdx2, Commd3, Ets2, Gata4, Hmgn3, Hoxb1, Hoxb5, Hoxc5, Hoxc6, Hoxc8, Id2, Id3, Nkx1.2, Pbx1, Prickle1, Prr13, Pitx1, Rcor2, Smarca1, Sox2, Sp5, Tbx4, Tfeb, Tlx2, Znf503), of which 11 homeobox genes known to play important roles in morphogenesis (underlined). Taken together, these results demonstrate the existence of a core set of genes responsive to CDX2 during its successive functions in embryonic development.

A Core Set of Chromatin Sites Bound by CDX2 During Mouse Development

Next, publicly available ChIPseq data (Amin et al., 2016; Rhee et al., 2017; Banerjee et al., 2018) were used to compare the location of the CDX2 protein on chromatin at the three developmental stages analyzed above by RNAseq. It gave a core set of 1,047 chromosomal regions sharing overlapping peaks in the three conditions (**Figure 2A**; **Supplementary Table S2.1**). 265 and 466 of these peaks respectively fell into protein coding genes and their promoters (defined as the 2-kb segment upstream of the transcription start site), 52 into non-protein coding genes and their 2-kb promoters, and 264 into intergenic regions. Among the 1,047 regions, 835 (77.75%) exhibited at least one conserved motif analogous to the mouse CDX2 binding site reported in the JASPAR database (#PH0013.1), based on the functional characterization of CDX-binding sites by SELEX (T/C-A-T-A-A-A-T/G, Margalit et al., 1993). This gave a total of 1,801 CDX-type sites (enrichment p-value = 10^{-152}) (**Figure 2B**; **Supplementary Table S2.2**). Interestingly, the ± 50 bp segments around these CDX-type sites were enriched in DNA-binding motifs for 149 transcription factors ($p < 0.05$) grouped into 25 families (**Figure 2C**; **Supplementary Table S2.3**). Moreover, 71 of these transcription factor binding motifs ($p < 0.05$), belonging to nine families, were also enriched within the ± 50 bp segments centered on the 1,314 CDX-type sites present in the promoters of the 221 DEGs (**Figure 2C**; **Supplementary Table S2.2** and **Supplementary Table S2.4**). The presence of enriched binding motifs for these transcription factors nearby the CDX binding sites suggests possible direct or indirect interactions. Among the CDX2 ChIPseq peaks located in gene promoters, 8 were associated with genes of the core set of DEGs (Arid3a, Epha4, Hoxc6, Man1c1, Mgat1, Mid1ip1, Sgsm1, Tfeb), whereas 75 out of the 264 intergenic peaks (28.41%) fell into Super-Enhancer domains (**Supplementary Table S2.5**).

Validation of the Core Set of Differentially Expressed Murine Genes in Independent Transgenic Mouse Models

Five transgenic mouse models targeting the Cdx2 gene have been reported together with corresponding RNAseq data: 1) the ectopic expression of human CDX2 in the anterior epiblast at gastrulation (RsCDX2:Sox2Cre^{ERT2} embryos)

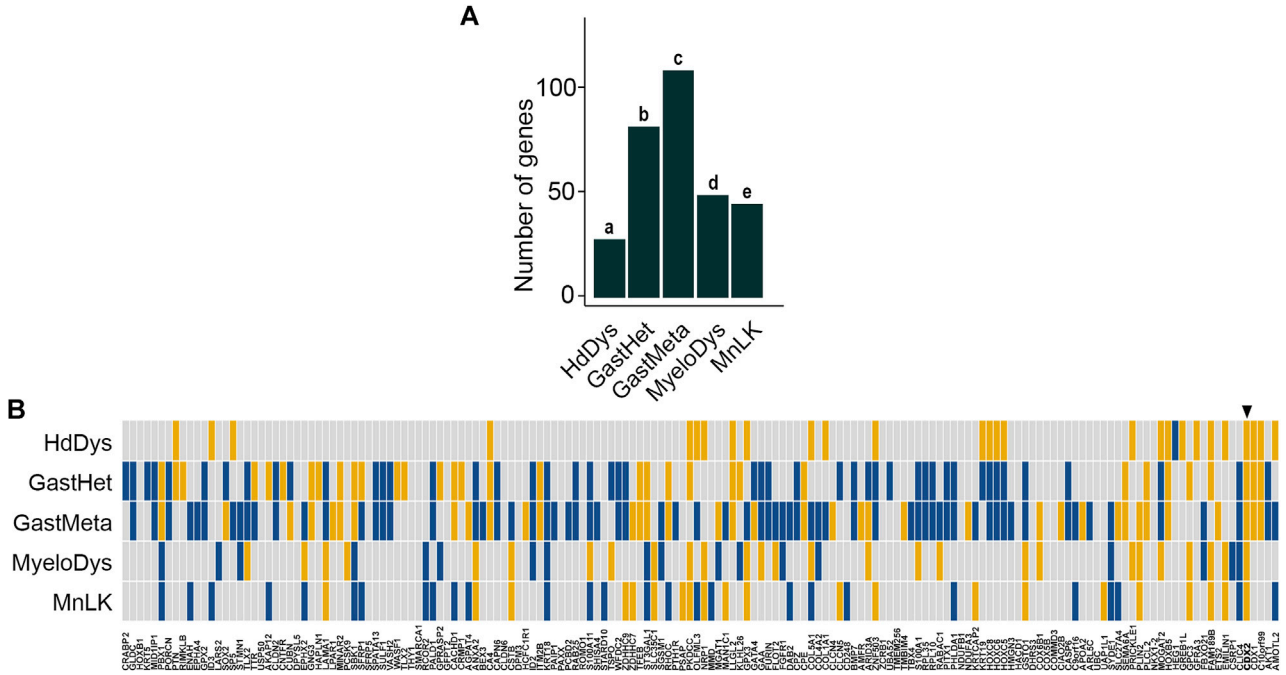

FIGURE 3 | Validation of the core set of DEGs in transgenic mouse models. (**A**) Enrichment analysis of 162/221 genes of the mouse core set of DEGs having human orthologues in Cdx2-dependent transgenic mouse models of head dysgenesis (HdDys), gastric-type heteroplasia in the cecum (GastHet), gastric-type metaplasia in the pericecal region (GastMeta), myelodysplasia (MyeloDys) and monoblastic leukemia (MnLK). *p*-values are a = 5.030 E^{-15}; b = 3.020 E^{-12}; c = 1.578 E^{-18}; d = 4.324 E^{-5}; e = 1.546 E^{-4}. (**B**) Correlation map of the above 162 genes of the mouse core set of DEGs in the validation models (yellow: up-regulation; blue: down-regulation). In each case, the map illustrates the comparison of the Cdx2-expressing samples vs the Cdx2 non-expressing counterparts. The arrowhead shows CDX2. The order of the genes is the same as in **Figure 1B**.

resulting in severe head dysgenesis (HdDys) (Grall et al., 2019); 2) the sporadic silencing of the single wild type Cdx2 allele in heterozygous $Cdx2^{+/-}$ embryos leading to congenital gastric-type heteroplasia in the cecum (GastHet) (Beck et al., 1999; Balbinot et al., 2018); 3) the mosaic invalidation of Cdx2 in the adult intestinal epithelium ($AhCre^{ERT}$:$Cdx2^{f/f}$ mice) inducing pericecal gastric-type metaplasia (GastMeta) (Balbinot et al., 2018); 4) the ectopic expression of mouse Cdx2 in hematopoietic stem cells ($SclCre^{ERT}$:Rosa-LSL-Cdx2 mice) leading to myelodysplasia (MyeloDys) (Vu et al., 2020), and 5) the ectopic induction of human CDX2 in bone marrow stem/progenitor cells (Mx1Cre:RsCDX2 mice) inducing monoblastic leukemia (MnLK) (Galland et al., 2021). Testing the deregulated genes in these five murine models against the 162 human orthologues to the mouse core set of DEGs revealed a significant number of genes in common, namely 28 genes in HdDys (enrichment *p*-value = 5.030 E^{-15}), 82 genes in GastHet (enrichment *p*-value = 3.020 E^{-12}), 109 genes in GastMeta (enrichment *p*-value = 3.020 E^{-12}), 45 genes in MyeloDys (enrichment *p*-value = 4.324 E^{-5}) and 49 genes in MnLK (enrichment *p*-value = 1.546 E^{-4}) (**Figures 3A,B**; **Supplementary Tables S3.1–5**). These results validate the core set of DEGs responsive to CDX2 in mice. In addition, they reinforce the notion of context-dependent effect.

Pattern of the Core Set of Differentially Expressed Murine Genes in Human Pathologies

Having established and validated the core set of DEGs in mice, we addressed the pattern of the 162 orthologues in human diseases exhibiting alterations in CDX2 levels (**Figure 4A**; **Supplementary Tables S4.1–2**). Several pathological conditions were considered. First, given that the physiological expression of CDX2 is limited to the gut epithelium in adults and that it is reduced in colon cancers with bad prognosis (Balbinot et al., 2018), we compared the transcriptomes in the deciles of tumors exhibiting the lowest vs highest CDX2 levels (n = 44 each) among The Cancer Genome Atlas (TCGA) collection of 436 colon adenocarcinomas (COAD). Overall, a total of 46 genes among the 162 human orthologues of the core set of DEGs were differentially expressed between both groups (enrichment *p*-value = 0.044) (**Figures 4B,C**; **Supplementary Table S4.3**). Second, we considered pathological situations exhibiting abnormal ectopic expression of CDX2 outside the gut in the upper digestive tract, namely the esophagus and stomach, where ectopic CDX2 associates with precancerous metaplasia and adenocarcinoma (Moskaluk et al., 2003). In the esophagus, retrieving the list of differentially expressed genes between healthy CDX2-free mucosa (n = 17)

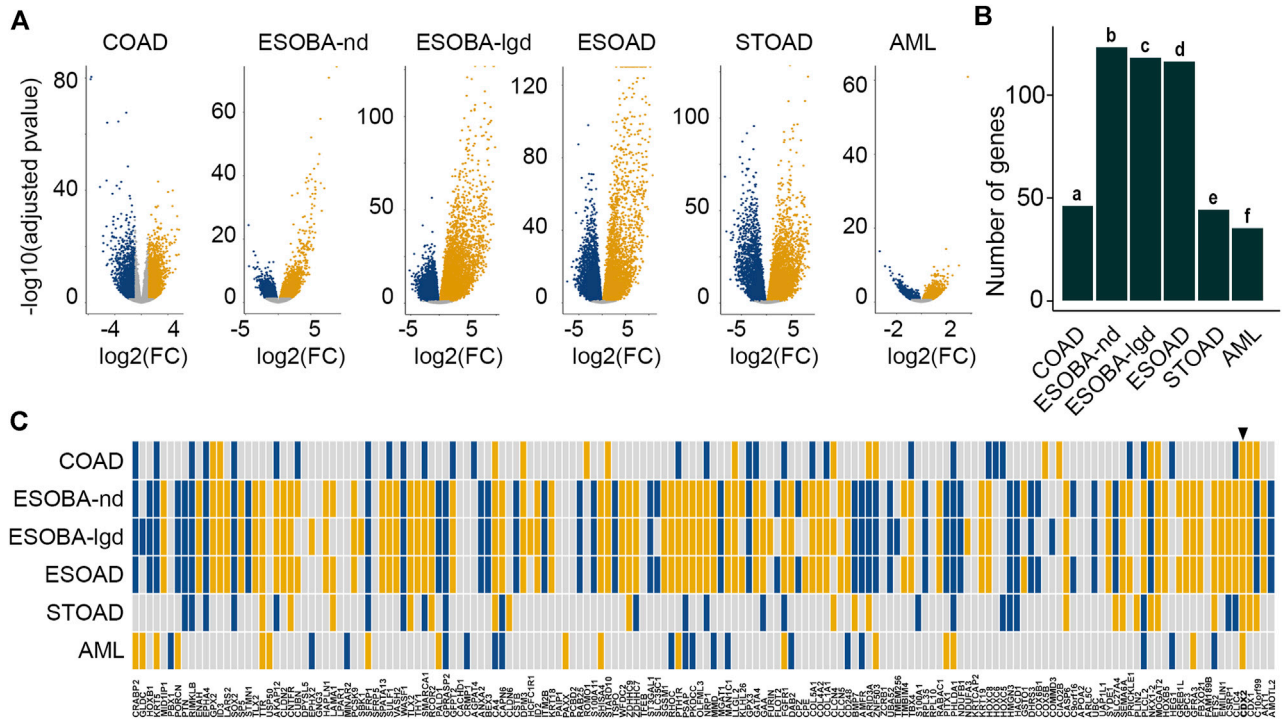

FIGURE 4 | Human orthologues to the core set of mouse DEGs deregulated in diseases. (**A**) Volcano plots of the differentially expressed genes in colon adenocarcinoma (COAD), in esophageal non-dysplastic Barrett metaplasia (ESOBA-nd), low-grade dysplastic Barrett metaplasia (ESOBA-lgd) and adenocarcinoma (ESOAD), in gastric adenocarcinoma (STOAD) and in acute myeloid leukemia (AML); up-regulated (yellow), down-regulated (blue), not significant (grey). (**B**) Enrichment analysis of the set of 162 human orthologues to the mouse core set of DEGs among the differentially expressed genes in human diseases. *p*-values are a = 0.044; b = 0.16 E^{-73}; c = 0.21 E^{-64}; d = 0.96 E^{-44}; e = 0.0028; f = 0.14 E^{-4}. **(C)** Correlation map of the expression changes of the human orthologues to the mouse core set of DEGs in human diseases (yellow: up-regulation; blue: down-regulation). In each case, the map illustrates the comparison of the CDX2-high expressing samples vs CDX2 non-expressing or low-expressing counterparts. The arrowhead shows CDX2. The order of the genes is the same as in **Figure 1B**.

and CDX2-expressing non-dysplastic Barrett metaplasia (ESOBA-nd) (n = 14), low-grade dysplastic Barrett metaplasia (ESOBA-lgd) (n = 8) and adenocarcinoma (ESOAD) (n = 12) (Maag et al., 2017) revealed respectively 123, 118 and 116 orthologues of the core set of DEGs (respective enrichment p-values are 0.16 E^{-73}, 0.21 E^{-64} and 0.96 E^{-44}) (**Figures 4B,C**; **Supplementary Table S4.4–6**). In the stomach, the list of differentially expressed genes in the quartiles of tumors presenting the highest vs lowest levels of CDX2 (n = 35 each) within the series of 272 STOAD samples of the TCGA comprised 44 DEGs of the core (enrichment p-value = 0.0028) (**Figures 4B,C**; **Supplementary Table S4.7**). Third, we analyzed AML in which abnormal ectopic expression of CDX2 is associated with disease burden (Scholl et al., 2007). We found 35 genes of the core set of DEGs among the genes differentially expressed between the quartiles with the highest vs lowest levels of CDX2 (n = 38 each) in the series of 151 AML of the TCGA (enrichment p-value = 0.14 E^{-4}) (**Figures 4B,C**; **Supplementary Table S4.8**). Taken together, these results indicate that a significant proportion of members of the core set of CDX2-responsive genes defined during mouse development is differentially expressed in human diseases along with CDX2 changes.

DISCUSSION

This study identified in mice a core set of common DEGs responsive to the CDX2 homeoprotein and a core set of common chromatin sites bound to the CDX2 protein at three developmental steps at which this transcription factor plays pivotal roles: trophectoderm specification, posterior growth of the embryonic body and intestinal determination. The core of DEGs was validated in transgenic mouse models targeting Cdx2. Moreover, a significant number of human orthologues to the mouse core set of DEGs was altered in human malignancies along with CDX2. Taken together, these results show that a transcription factor, e.g., the CDX2 homeoprotein, while driving distinct functions at different steps during embryonic development, can exert a common subset of molecular activities, and that some of these activities can be subsequently deregulated in adult pathologies along with this factor.

Although studies in mice have highlighted the importance of the Cdx2 gene at many embryonic stages, developmental defects linked to alterations of this gene are rare in human, likely because its constitutive loss of function is expected to prevent trophectoderm formation and uterine implantation of the

blastula. However, human CDX2 gene variants have recently been associated with sirenomelia (Lecoquierre et al., 2020), in accordance with the function attributed to this gene in posterior body elongation and patterning. Moreover, the aberrant expression of CDX2 reported in various forms of congenital endoderm-derived heteroplasia corroborates its key role in intestinal identity determination (Martin et al., 2010). Beyond embryogenesis, pathological alterations of CDX2 levels occur at its physiological site of expression, the gut, as well as ectopically in the upper digestive tract and in leukemia. The fact that the expression of a significant number of genes of the developmental core set of DEGs changed along with CDX2 in human malignancies strengthens the relevance of this DEGs core.

This study reveals that the direction of the changes of several genes of the DEGs core is not consistent at the three mouse developmental steps analyzed here, as well as in human pathologies. It emphasizes the context-dependent activity of this transcription factor. This property can be seen in view of the number of CDX2 ChIPseq peaks overlapping intergenic Super Enhancers known to control the functional activity of large chromosomal regions, and of the anti-repressing effect exerted by the CDX2 protein to prevent the incursion of inactive marks into chromatin domains and keep them accessible to other transcription partners (Verzi et al., 2013; Saxena et al., 2017). Thus, as shown in the gut, CDX2 can have inductive, permissive and repressive transcriptional effects (Verzi et al., 2013; San Roman et al., 2015; Saxena et al., 2017), indicating that its outcome depends not only on the chromatin domains that are kept open, but also on the specific repertoire of nuclear partners present in the cells and able to interact with open chromatin regions to either stimulate or inhibit transcription. Interestingly, in pathological situations the context-dependent activity of CDX2 could provide hints to explain opposite effects, being a tumor suppressor in its physiological site of expression, the gut, but an oncogene when ectopically expressed in the hematopoietic lineage. Thus, the present study opens ways to investigate novel functional interactions between developmental genes and exploit them in a therapeutic perspective.

MATERIALS AND METHODS

Mouse and Human RNAseq and ChIPseq Data

Mouse RNAseq and ChIPseq data were retrieved from the GEO database (https://www.ncbi.nlm.nih.gov/geo/): GSE62149 (Cambuli et al., 2014) and GSE90752 (Rhee et al., 2017) for ES cells, GSE84899 (Amin et al., 2016) for E8 growing embryos, GSE115541 (Banerjee et al., 2018) for E16 intestinal endoderm, GSE123559 (Grall et al., 2019) for the head dysgenesis model, GSE89992 (Balbinot et al., 2018) for gastric-type intestinal hetero- and metaplasia, GSE133679 (Vu et al., 2020) for myelodysplasia, and GSE120487 (Galland et al., 2021) for monoblastic leukemia. The identifiers of samples used for this study are given in the **Supplementary Table S1.1**. Human RNAseq data from colon adenocarcinoma (COAD), stomach adenocarcinoma (STOAD) and acute myeloid leukemia (AML) were obtained from the database The Cancer Genome Atlas (The TCGA research network: https://www.cancer.gov/tcga) with the identifiers given in **Supplementary Table S4.1**. Esophageal metaplasia and adenocarcinoma data were from Maag et al. (2017).

Mouse mRNAseq Read Mapping and Quantification of Expression

Quality controls of raw RNAseq reads were carried out with the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to assess base quality, nucleotide ratio and sequence duplication rate. RNAseq reads were then mapped with STAR (Dobin et al., 2013) against the mouse reference genome GRCm38. Alignments were filtered in normal mode and multi-mapped reads were discarded. For every splicing junction reconstructed from the first round of mapping, a second mapping was carried out to improve alignment. Metrics on alignment were computed with Samtools Flagstat and Samtools Stat (Danecek et al., 2021) to ensure quality of mapping. Raw gene expression, i.e. the number of mapped reads per annotated gene, were computed with HTSeq, in union mode and with the annotation of the reference genome provided as a GTF file.

Mouse CELseq Read Mapping and Quantification of Expression

For the CELseq data of growing mouse embryos (Amin et al., 2016), sequencing adapters were trimmed with Cutadapt (Martin, 2011). Reads were mapped on the reference genome GRCm38 with BWA.aln (Li et al., 2009) and genomic coordinates were converted to alignment with BWA Samse and Samtools view. Raw read numbers were computed as described above for mRNAseq data.

Differential Expression Analysis

For mouse ES cells data, as no replicate was available, differential expression was assessed by computing the delta of the gene expression values between control and experimental condition in each of the two datasets (Rhee et al., 2017 and Cambuli et al., 2014). Then, common differentially expressed genes between both datasets were selected with a threshold of 2 on delta. For the other mouse embryos data, namely the growing embryo (Amin et al., 2016) and the intestinal endoderm (Banerjee et al., 2018), DESeq2 (Love et al., 2014) was used. Gene expression was normalized with a regression model and differential expression was tested with the Wald test corrected by Bonferonni. False positives were identified with the Cook distance and flagged. Samples segregation was assessed by Principal Component Analysis (PCA, **Supplementary Figure S1**). Genes with significant variations in transcript levels were selected applying a threshold of 2 on |log2 (fold-change)| and a threshold of 0.05 on adjusted p-value. These genes were then compared between the datasets of the three developmental stages, i.e., ES cells, growing embryo and gut endoderm, to create the

core set of common differentially expressed genes (DEGs). The enrichment in genes of the core was tested with the exact Fisher test. Orthology between the mouse DEGs and the human genome was evaluated with Ensembl Compara information based on the annotation of the mouse reference genome GRCm38, with a confidence score of 1 (high) or a minimal sequence homology of 30%, using a custom-made R-script as previously published (Mayrhofer et al., 2017). Enriched biological functions (Gene Ontology Resource, http://geneontology.org/), signaling pathways (Kyoto Encyclopedia of Genes and Genomes, https://www.genome.jp/kegg/) and protein domains (InterPro, http://www.ebi.ac.uk/interpro/) were tested on the core set of genes with DAVID (Huang et al., 2009). Further annotation of genes including symbol and description were collected with a custom-made R script.

Mouse ChIPseq Data Processing

ChIPseq reads were mapped with BWA (Li and Durbin, 2009) against the reference genome GRCm38 as described above for the CELseq data. Unmapped reads, reads with low mapping quality, i.e., a Phred score below 30 for each base, and multi-mapped reads were filtered out. Duplicated reads were removed with GATK MarkDuplicates (Van der Auwera and O'Connor, 2020). Metrics on alignments were collected with Samtools Stats and Samtools Flagstat to ensure a good quality of read mapping (Danecek et al., 2021). Peaks were detected with MACS2 (Zhang et al., 2008). A cutoff of 10^{-05} was set on *p*-values to output peaks and significance of peaks compared to background noise was evaluated with regard to the input control. For each peak the signal was normalized computing fragment pileup per million reads. ChIPseq peaks were then selected applying a threshold of 0.05 on *p*-values and visually controlled in the genome browser IGV (Robinson et al., 2011). A core was created with ChIPseq peaks of the compared datasets overlapping with at least 10 bp in the three conditions: trophectoderm formation, antero-posterior patterning and gut specification. ChIPseq peaks were annotated with an in-house developed R script based on genes present in the annotation of the reference genome GRCm38. Both upstream and downstream genes were annotated. Intergenic ChIPseq peaks were further compared to Super-Enhancers from the database dbSUPER (Khan and Zhang, 2016).

DNA Binding Sites Analysis

All known binding motifs of vertebrate transcription factors present in the core of ChIPseq peaks and in the gene promoters of the core of DEGs (defined as the 2-kb segment upstream of the canonical transcription start site(s) of each gene) were retrieved from the database JASPAR (Khan et al., 2018), classified with TFclass relying on "class" and "family" subdivisions (Wingender et al., 2018), and their position weight matrixes were reformatted. Enrichment for transcription factor binding motifs was tested with HOMER (Heinz et al., 2010) in the direct vicinity (+/− 50 bp) of mapped CDX-type homeobox motifs identified in the promoters of the DEGs and in the overlapping ChIPseq peaks. To test transcription factor binding motif enrichment, background sets of DNA sequences were created. These sets were composed of the same number of tested regions, i.e., promoters or overlapping ChIPseq peaks. The DNA sequences were of the same size as the tested regions and were randomly extracted from the mouse reference genome GRCm38.

Analysis of Mouse Validation Samples and Human Pathological Samples

For samples obtained from mouse models of embryonic head dysgenesis (Grall et al., 2019), gastric-type heteroplasia and metaplasia (GastHet and GastMeta, Balbinot et al., 2018), myelodysplasia (MyeloDys, Vu et al., 2020) and monoblastic leukemia (MnLK, Galland et al., 2021), the log2 (fold-change) and *p*-value were retrieved from the literature.

For human pathological samples, raw levels of transcripts were computed with HTSeq (Anders et al., 2015) for colon adenocarcinoma (COAD), stomach adenocarcinoma (STOAD) and acute myeloid leukemia (AML). Each human gene symbol was associated to the corresponding Ensembl gene identifier and the transcript levels were normalized by computing reads per kilobase per million in order to identify groups with high and low levels of CDX2 transcripts. These groups were defined as upper and lower quartiles or deciles with a purpose of comparable size. For pair-wise comparison of groups, raw levels of transcripts were processed with DESeq2 (Love et al., 2014) as described above. Genes with significant variation in transcript levels were selected applying a threshold of 0.05 on adjusted *p*-value (Bonferroni multiple testing method). The significance of the difference in expression level of CDX2 among samples with high versus low expression of CDX2 in COAD, STOAD and AML is shown in the boxplot of the **Supplementary Figure S2** and confirmed with a Wilcoxon test. For Barrett's syndrome and esophagus adenocarcinoma, log2 (fold-change) and *p*-value were retrieved from the literature (Maag et al., 2017). For pathologies and validation datasets, the enrichment in gene of the core was tested with the one-tailed exact Fisher test with gene sets defined by significantly differentially expressed genes and a stringent gene Universe defined as genes confidently associated with a Gene Ontology.

R-Scripts Availability

The code for the analysis of each dataset is available on github (https://github.com/victor-gourain/Gourainetal2021) and Zenodo (https://zenodo.org/badge/latestdoi/407113075).

DATA AVAILABILITY STATEMENT

Publicly available RNAseq and ChIPseq datasets, clearly referred in the article, were used for this study.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Ethical review and approval was not required for the animal study because this study uses only publically available RNAseq and ChIPseq data.

AUTHOR CONTRIBUTIONS

J-NF, VG, ID and CD-D conceived the research, analyzed the data and wrote the manuscript; VG performed the bioinformatics analyses.

FUNDING

This work was supported by the Fondation ARC (PJA #20181208021) and by the Institut National du Cancer (INCa, PLBIO 19–289).

ACKNOWLEDGMENTS

The authors thank the Inserm for the support of the IRFAC laboratory, Uwe Strähle, Karlsruhe Institute of Technology, Germany, for providing computing resources, and Jeremie Poschmann, University of Nantes, France, for his critical review.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.744165/full#supplementary-material

REFERENCES

Amin, S., Neijts, R., Simmini, S., van Rooijen, C., Tan, S. C., Kester, L., et al. (2016). Cdx and T Brachyury Co-Activate Growth Signaling in the Embryonic Axial Progenitor Niche. *Cel Rep.* 17, 3165–3177. doi:10.1016/j.celrep.2016.11.069

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python Framework to Work with High-Throughput Sequencing Data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638

Baba, Y., Nosho, K., Shima, K., Freed, E., Irahara, N., Philips, J., et al. (2009). Relationship of CDX2 Loss with Molecular Features and Prognosis in Colorectal Cancer. *Clin. Cancer Res.* 15, 4665–4673. doi:10.1158/1078-0432.ccr-09-0401

Balbinot, C., Armant, O., Elarouci, N., Marisa, L., Martin, E., De Clara, E., et al. (2018). The Cdx2 Homeobox Gene Suppresses Intestinal Tumorigenesis through Non-Cell-Autonomous Mechanisms. *J. Exp. Med.* 215, 911–926. doi:10.1084/jem.20170934

Banerjee, K. K., Saxena, M., Kumar, N., Chen, L., Cavazza, A., Toke, N. H., et al. (2018). Enhancer, Transcriptional, and Cell Fate Plasticity Precedes Intestinal Determination during Endoderm Development. *Genes Dev.* 32, 1430–1442. doi:10.1101/gad.318832.118

Beck, F., Chawengsaksophak, K., Waring, P., Playford, R. J., and Furness, J. B. (1999). Reprogramming of Intestinal Differentiation and Intercalary Regeneration in Cdx2 Mutant Mice. *Proc. Natl. Acad. Sci.* 96, 7318–7323. doi:10.1073/pnas.96.13.7318

Bonhomme, C., Duluc, I., Martin, E., Chawengsaksophak, K., Chenard, M. P., Kedinger, M., et al. (2003). The Cdx2 Homeobox Gene Has a Tumour Suppressor Function in the Distal Colon in Addition to a Homeotic Role during Gut Development. *Gut* 52, 1465–1471. doi:10.1136/gut.52.10.1465

Cambuli, F., Murray, A., Dean, W., Dudzinska, D., Krueger, F., Andrews, S., et al. (2014). Epigenetic Memory of the First Cell Fate Decision Prevents Complete ES Cell Reprogramming into Trophoblast. *Nat. Commun.* 5, 5538. doi:10.1038/ncomms6538

Dalerba, P., Sahoo, D., Paik, S., Guo, X., Yothers, G., Song, N., et al. (2016). CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. *N. Engl. J. Med.* 374, 211–222. doi:10.1056/nejmoa1506597

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *Gigascience* 10 (2), giab008. doi:10.1093/gigascience/giab008

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Galland, A., Gourain, V., Habbas, K., Güler, Y., Martin, E., Ebel, C., et al. (2021). CDX2 Expression in the Hematopoietic Lineage Promotes Leukemogenesis via TGFβ Inhibition. *Mol. Oncol.* 15, 2318–2329. doi:10.1002/1878-0261.12982

Gao, N., White, P., and Kaestner, K. H. (2009). Establishment of Intestinal Identity and Epithelial-Mesenchymal Signaling by Cdx2. *Dev. Cel* 16, 588–599. doi:10.1016/j.devcel.2009.02.010

Grall, E., Gourain, V., Naïr, A., Martin, E., Birling, M.-C., Freund, J.-N., et al. (2019). Severe Head Dysgenesis Resulting from Imbalance between Anterior and Posterior Ontogenetic Programs. *Cell Death Dis* 10, 812. doi:10.1038/s41419-019-2040-0

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cel* 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2018). JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and its Web Framework. *Nucleic Acids Res.* 46, D260–D266. doi:10.1093/nar/gkx1126

Khan, A., and Zhang, X. (2016). dbSUPER: A Database of Super-Enhancers in Mouse and Human Genome. *Nucleic Acids Res.* 44, D164–D171. doi:10.1093/nar/gkv1002

Lecoquierre, F., Brehin, A. C., Coutant, S., Coursimault, J., Bazin, A., Finck, W., et al. (2020). Exome Sequencing Identifies the First Genetic Determinants of Sirenomelia in Humans. *Hum. Mutat.* 41, 926–933. doi:10.1002/humu.23998

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Maag, J. L. V., Fisher, O. M., Levert-Mignon, A., Kaczorowski, D. C., Thomas, M. L., Hussey, D. J., et al. (2017). Novel Aberrations Uncovered in Barrett's Esophagus and Esophageal Adenocarcinoma Using Whole Transcriptome Sequencing. *Mol. Cancer Res.* 15, 1558–1569. doi:10.1158/1541-7786.MCR-17-0332

Margalit, Y., Yarus, S., Shapira, E., Gruenbaum, Y., and Fainsod, A. (1993). Isolation and Characterization of Target Sequences of the chickenCdxAhomeobox Gene. *Nucl. Acids Res.* 21, 4915–4922. doi:10.1093/nar/21.21.4915

Martin, E., Vanier, M., Tavian, M., Guerin, E., Domon-Dell, C., Duluc, I., et al. (2010). CDX2 in Congenital Gut Gastric-Type Heteroplasia and Intestinal-type Meckel Diverticula. *Pediatrics* 126, e723–e727. doi:10.1542/peds.2009-3512

Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.journal* 17, 10–11. doi:10.14806/ej.17.1.200

Mayrhofer, M., Gourain, V., Reischl, M., Affaticati, P., Jenett, A., Joly, J.-S., et al. (2017). A Novel Brain Tumour Model in Zebrafish Reveals the Role of YAP Activation in MAPK/PI3K Induced Malignant Growth. *Dis. models Mech.* 10, 15–28. doi:10.1242/dmm.026500

Moskaluk, C. A., Zhang, H., Powell, S. M., Cerilli, L. A., Hampton, G. M., and Frierson, H. F., Jr. (2003). Cdx2 Protein Expression in normal and Malignant Human Tissues: An Immunohistochemical Survey Using Tissue Microarrays. *Mod. Pathol.* 16, 913–919. doi:10.1097/01.mp.0000086073.92773.55

Niwa, H., Toyooka, Y., Shimosato, D., Strumpf, D., Takahashi, K., Yagi, R., et al. (2005). Interaction between Oct3/4 and Cdx2 Determines Trophectoderm Differentiation. *Cell* 123, 917–929. doi:10.1016/j.cell.2005.08.040

Ralston, A., and Rossant, J. (2008). Cdx2 Acts Downstream of Cell Polarization to Cell-Autonomously Promote Trophectoderm Fate in the Early Mouse Embryo. *Dev. Biol.* 313, 614–629. doi:10.1016/j.ydbio.2007.10.054

Rhee, C., Lee, B.-K., Beck, S., LeBlanc, L., Tucker, H. O., and Kim, J. (2017). Mechanisms of Transcription Factor-Mediated Direct Reprogramming of Mouse Embryonic Stem Cells to Trophoblast Stem-Like Cells. *Nucleic Acids Res.* 45, 10103–10114. doi:10.1093/nar/gkx692

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative Genomics Viewer. *Nat. Biotechnol.* 29, 24–26. doi:10.1038/nbt.1754

Sakamoto, N., Feng, Y., Stolfi, C., Kurosu, Y., Green, M., Lin, J., et al. (2017). BRAFV600E Cooperates with CDX2 Inactivation to Promote Serrated Colorectal Tumorigenesis. *Elife* 6, e20331. doi:10.7554/eLife.20331

San Roman, A. K., Tovaglieri, A., Breault, D. T., and Shivdasani, R. A. (2015). Distinct Processes and Transcriptional Targets Underlie CDX2 Requirements in Intestinal Stem Cells and Differentiated Villus Cells. *Stem Cel Rep.* 5, 673–681. doi:10.1016/j.stemcr.2015.09.006

Saxena, M., Roman, A. K. S., O'Neill, N. K., Sulahian, R., Jadhav, U., and Shivdasani, R. A. (2017). Transcription Factor-Dependent 'Anti-Repressive' Mammalian Enhancers Exclude H3K27me3 from Extended Genomic Domains. *Genes Dev.* 31, 2391–2404. doi:10.1101/gad.308536.117

Scholl, C., Bansal, D., Döhner, K., Eiwen, K., Huntly, B. J. P., Lee, B. H., et al. (2007). The Homeobox Gene CDX2 Is Aberrantly Expressed in Most Cases of Acute Myeloid Leukemia and Promotes Leukemogenesis. *J. Clin. Invest.* 117, 1037–1048. doi:10.1172/jci30182

Seno, H., Oshima, M., Taniguchi, M.-A., Usami, K., Ishikawa, T.-O., Chiba, T., et al. (2002). CDX2 Expression in the Stomach with Intestinal Metaplasia and Intestinal-Type Cancer: Prognostic Implications. *Int. J. Oncol.* 21, 769–774. doi:10.3892/ijo.21.4.769

Stringer, E. J., Duluc, I., Saandi, T., Davidson, I., Bialecka, M., Sato, T., et al. (2012). Cdx2 Determines the Fate of Postnatal Intestinal Endoderm. *Development* 139, 465–474. doi:10.1242/dev.070722

Strumpf, D., Mao, C.-A., Yamanaka, Y., Ralston, A., Chawengsaksophak, K., Beck, F., et al. (2005). Cdx2 Is Required for Correct Cell Fate Specification and Differentiation of Trophectoderm in the Mouse Blastocyst. *Development* 132, 2093–2102. doi:10.1242/dev.01801

Suh, E., Chen, L., Taylor, J., and Traber, P. G. (1994). A Homeodomain Protein Related to Caudal Regulates Intestine-Specific Gene Transcription. *Mol. Cel Biol* 14, 7340–7351. doi:10.1128/mcb.14.11.7340-7351.1994

Van der Auwera, G. A., and O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. Newton, MA: O'Reilly Media, 1491975148.

van Rooijen, C., Simmini, S., Bialecka, M., Neijts, R., van de Ven, C., Beck, F., et al. (2012). Evolutionarily Conserved Requirement of Cdx for Post-Occipital Tissue Emergence. *Development* 139, 2576–2583. doi:10.1242/dev.079848

Verzi, M. P., Shin, H., He, H. H., Sulahian, R., Meyer, C. A., Montgomery, R. K., et al. (2010). Differentiation-Specific Histone Modifications Reveal Dynamic Chromatin Interactions and Partners for the Intestinal Transcription Factor CDX2. *Dev. Cel* 19, 713–726. doi:10.1016/j.devcel.2010.10.006

Verzi, M. P., Shin, H., San Roman, A. K., Liu, X. S., and Shivdasani, R. A. (2013). Intestinal Master Transcription Factor CDX2 Controls Chromatin Access for Partner Transcription Factor Binding. *Mol. Cell Biol.* 33, 281–292. doi:10.1128/mcb.01185-12

Vu, T., Straube, J., Porter, A. H., Bywater, M., Song, A., Ling, V., et al. (2020). Hematopoietic Stem and Progenitor Cell-Restricted Cdx2 Expression Induces Transformation to Myelodysplasia and Acute Leukemia. *Nat. Commun.* 11, 3021. doi:10.1038/s41467-020-16840-2

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., and Dönitz, J. (2018). TFClass: Expanding the Classification of Human Transcription Factors to Their Mammalian Orthologs. *Nucl. Acids Res.* 46, D343–D347. doi:10.1093/nar/gkx987

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-Based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi:10.1186/gb-2008-9-9-r137

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Gourain, Duluc, Domon-Dell and Freund. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.