LETTER

https://doi.org/10.1038/s41586-019-1312-2

Noncoding deletions reveal a gene that is critical for intestinal function

Danit Oz-Levi[1], Tsviya Olender[1,25], Ifat Bar-Joseph[2,3,25], Yiwen Zhu[4,25], Dina Marek-Yagel[2,3,5], Iros Barozzi[4,23], Marco Osterwalder[4], Anna Alkelai[6], Elizabeth K. Ruzzo[7], Yujun Han[8], Erica S. M. Vos[9], Haike Reznik-Wolf[2,3], Corina Hartman[3,10], Raanan Shamir[3,10], Batia Weiss[3,5], Rivka Shapiro[3,10], Ben Pode-Shakked[3,5], Pavlo Tatarskyy[1], Roni Milgrom[1], Michael Schvimer[11], Iris Barshack[3,11], Denise M. Imai[12], Devin Coleman-Derr[13], Diane E. Dickel[4], Alex S. Nord[4,24], Veena Afzal[4], Kelly Lammerts van Bueren[14], Ralston M. Barnes[14], Brian L. Black[14], Christopher N. Mayhew[15], Matthew F. Kuhar[15], Amy Pitstick[15], Mehmet Tekman[16], Horia C. Stanescu[16], James M. Wells[15,17,18], Robert Kleta[16], Wouter de Laat[9], David B. Goldstein[6], Elon Pras[2,3], Axel Visel[4,19,20], Doron Lancet[1,25,26]*, Yair Anikster[3,5,21,25,26]* & Len A. Pennacchio[4,20,22,25,26]*

Large-scale genome sequencing is poised to provide a substantial increase in the rate of discovery of disease-associated mutations, but the functional interpretation of such mutations remains challenging. Here we show that deletions of a sequence on human chromosome 16 that we term the intestine-critical region (ICR) cause intractable congenital diarrhoea in infants[1,2]. Reporter assays in transgenic mice show that the ICR contains a regulatory sequence that activates transcription during the development of the gastrointestinal system. Targeted deletion of the ICR in mice caused symptoms that recapitulated the human condition. Transcriptome analysis revealed that an unannotated open reading frame (*Percc1*) flanks the regulatory sequence, and the expression of this gene was lost in the developing gut of mice that lacked the ICR. *Percc1*-knockout mice displayed phenotypes similar to those observed upon ICR deletion in mice and patients, whereas an ICR-driven *Percc1* transgene was sufficient to rescue the phenotypes found in mice that lacked the ICR. Together, our results identify a gene that is critical for intestinal function and underscore the need for targeted in vivo studies to interpret the growing number of clinical genetic findings that do not affect known protein-coding genes.

In contrast to whole-exome sequencing (WES)[3], whole-genome sequencing (WGS) can in principle identify mutations in noncoding sequences, as well as in genes that are not annotated in the reference genome. However, sequence variation that affects poorly annotated sequences outside of known genes is challenging to interpret because of the lack of structural and functional annotation of these regions. In this study, we demonstrate how the identification of noncoding deletions in a small number of patients, together with purpose-built mouse models, can be used to elucidate the regulatory and genetic basis of an inherited severe disease (Fig. 1).

Congenital diarrhoeal disorders are a heterogeneous group of inherited diseases of the digestive system and are frequently life-threatening if untreated[1,2,4] (see Supplementary Information for additional clinical background). We studied eight patients from seven unrelated families of common ethnogeographic origin, each of whom had intractable diarrhoea of infancy syndrome (IDIS)[2]—an autosomal recessive pattern of severe congenital malabsorptive diarrhoea (Fig. 1a, b, Extended Data Fig. 1, Supplementary Information). Initial WES analysis revealed no rare exonic sequence variants with the appropriate patient segregation. However, whole-genome linkage analysis and haplotype reconstruction detected a single significant telomeric linkage interval on chromosome 16 (log of the odds ratio (LOD) = 4.26; Extended Data Fig. 2a, Supplementary Information). We examined WES and WGS data from selected patients and observed a 7,013-bp deletion, which we term ΔL, in the absence of other structural changes or coding mutations at the affected locus (Fig. 1c, Extended Data Figs. 1, 2b, Supplementary Information). Two of the patients were compound heterozygous for ΔL and a second variant, ΔS, which contains a 3,101-bp deletion that partially overlaps with ΔL. We were therefore able to define a minimal sequence of 1,528 bp as the intestine-critical region (ICR) (Fig. 1c). All eight patients in this study showed $ICR^{\Delta S/\Delta S}$, $ICR^{\Delta L/\Delta S}$ or $ICR^{\Delta L/\Delta L}$ genotypes, which resulted in a homozygous deletion of the ICR. By contrast, this deletion was not detected in any of the control groups we examined (Extended Data Fig. 1, Supplementary Information). These data suggest that the deletion of the ICR causes the congenital diarrhoea phenotype.

To investigate possible noncoding functions of the ICR, we examined data from the Encyclopedia of DNA Elements (ENCODE)[5]. The ICR contains a 400-bp region that is highly conserved across vertebrates; it also includes CpG island and DNase hypersensitivity signatures, and encompasses a cluster of multiple binding sites for transcription factors including FOXA1 and FOXA2[6,7] (Fig. 1c). To test the hypothesis that the ICR contains a regulatory sequence, we examined its in vivo activity in a reporter assay in transgenic mice[8]. In transgenic embryos ranging from embryonic day (E)11.5 to E14.5, we observed robust and reproducible reporter activity in the stomach, pancreas and duodenum (Fig. 2a, b). These results support the notion that the ICR, which is absent in patients with congenital diarrhoea, contains a gene-regulatory sequence that is active in the developing digestive system.

To assess whether loss of the ICR sequence is sufficient to cause the phenotypes observed in humans, we deleted a 1,512-bp interval from the mouse genome that included the mouse orthologues of the human

[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. [2]The Danek Gertner Institute of Human Genetics, Sheba Medical Center, Ramat Gan, Israel. [3]The Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel. [4]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Ramat Gan, Israel. [6]Institute for Genomic Medicine, Columbia University Medical Center, New York, NY, USA. [7]Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA. [8]Center for Human Genome Variation, Duke University School of Medicine, Durham, NC, USA. [9]Oncode Institute, Hubrecht Institute-KNAW and University Medical Center Utrecht, Utrecht, the Netherlands. [10]Schneider Children's Medical Center, Petach Tikva, Israel. [11]Department of Pathology, Sheba Medical Center, Ramat Gan, Israel. [12]Comparative Pathology Laboratory, University of California Davis, Davis, CA, USA. [13]Plant Gene Expression Center, USDA ARS, Albany, CA, USA. [14]Cardiovascular Research Institute, University of California San Francisco, San Francisco, CA, USA. [15]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [16]Centre for Nephrology, University College London, London, UK. [17]Division of Endocrinology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [18]Center for Stem Cell and Organoid Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [19]School of Natural Sciences, University of California, Merced, CA, USA. [20]US Department of Energy Joint Genome Institute, Walnut Creek, CA, USA. [21]Wohl Institute for Translational Medicine, Sheba Medical Center, Ramat Gan, Israel. [22]Comparative Biochemistry Program, University of California Berkeley, Berkeley, CA, USA. [23]Present address: Department of Surgery and Cancer, Imperial College London, London, UK. [24]Present address: Center for Neuroscience, University of California Davis, Davis, CA, USA. [25]These authors contributed equally: Tsviya Olender, Ifat Bar-Joseph, Yiwen Zhu. [26]These authors jointly supervised this work: Doron Lancet, Yair Anikster, Len A. Pennacchio. *e-mail: Doron.Lancet@weizmann.ac.il; Yair.Anikster@sheba.health.gov.il; LAPennacchio@lbl.gov

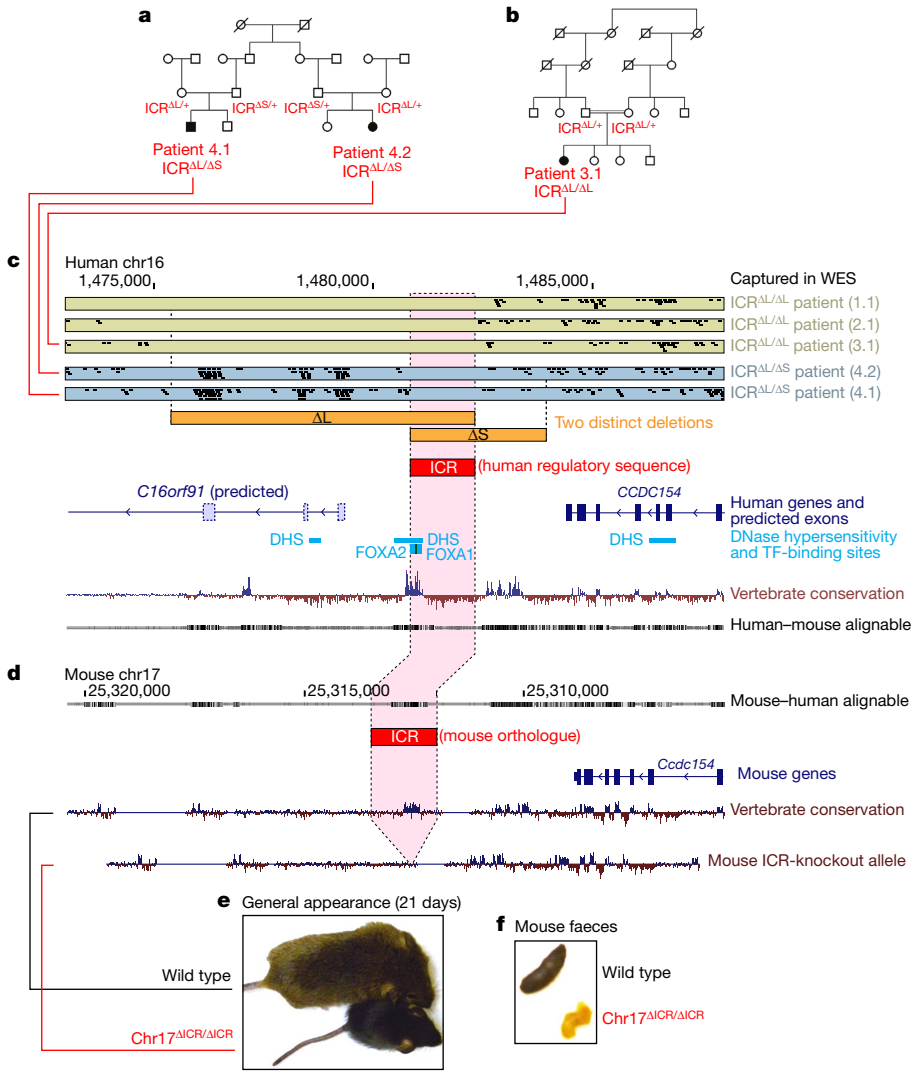


Fig. 1 | Overview of the human and mouse locus and key findings. **a**, **b**, Selected family pedigrees and genotyping results for patients who are compound heterozygous for the two deletion alleles ($ICR^{\Delta L/\Delta S}$) (**a**) or homozygous for one of the deletion alleles ($ICR^{\Delta L/\Delta L}$) (**b**). **c**, **d**, Genomic map of the deletion alleles in human (chromosome 16; genome build GRCh37) (**c**); and mouse (chromosome 17) (**d**), indicating the location of ΔL and ΔS as well as their minimal overlapping region (that is, the ICR). Exome-sequencing data are capped at up to five overlapping tags for visualization; vertebrate conservation is measured by 100-vertebrate PhyloP scores; and only selected transcription-factor-binding sites and DNase hypersensitivity clusters that displayed a signal in more than 20 out of 125 ENCODE cell types are shown. TF, transcription factor. **e**, General appearance of wild-type ($n = 50$) and $chr17^{\Delta ICR/\Delta ICR}$ ($n = 46$) at P21, showing the overall significantly reduced size of $chr17^{\Delta ICR/\Delta ICR}$ compared with wild-type mice (see Fig. 2c, d). **g**, Abnormal appearance of faecal pellets from $chr17^{\Delta ICR/\Delta ICR}$ mice ($n = 46$).

DNase hypersensitivity signature and the predicted binding sites of FOXA1 and FOXA2 (Fig. 1d, Extended Data Fig. 3). Mouse pups that were homozygous for the ICR deletion on mouse chromosome 17 ($chr17^{\Delta ICR/\Delta ICR}$) were born at the expected Mendelian frequency and showed no gross phenotypes at birth. However, starting within the first few days of life, $chr17^{\Delta ICR/\Delta ICR}$ mice displayed overall reduced size (Figs. 1e, 2c), low body weight (Fig. 2d) and substantially decreased survival (Fig. 2e), compared to wild-type littermates. Examination of faecal pellets and internal organs revealed abnormal digestive-tract function in $chr17^{\Delta ICR/\Delta ICR}$ mice (Fig. 1f, Extended Data Fig. 4a), as well as changes in the composition of the intestinal microbiome (Extended Data Fig. 4b, Supplementary Information). Our combined results indicate that deletion of the ICR in mice causes disruption of intestinal function—recapitulating the congenital diarrhoea phenotype that is observed in patients with homozygous ICR deletions.

Next, we sought to elucidate the molecular mechanisms through which deletion of the ICR causes deficiencies in gastrointestinal function. Targeted expression analysis of tissue samples from mouse stomach and intestine, over a range of developmental stages (E14.5 to postnatal day (P)20), revealed an unannotated region that flanked the ICR and that showed very low levels of expression in wild-type prenatal stomach tissue (Fig. 3a). Expression of this sequence was completely lost in matched tissues from $chr17^{\Delta ICR/\Delta ICR}$ littermates, which suggested the presence of a gene that has not previously been annotated in the human or the mouse genome (Extended Data Fig. 3c). Sequence analysis identified an 897-bp open reading frame that is predicted to encode a protein that we name PERCC1 on the basis of its properties (proline (P)- and glutamate (E)-rich with a coiled-coil domain) (Fig. 3b). Although PERCC1 shows strong evolutionary conservation across vertebrates (Fig. 3c), searches based on structure and homology failed to identify any similarities with known proteins. However, a comparison of the rate of human-to-mouse non-synonymous and synonymous substitutions (dN and dS, respectively) showed a dN/dS ratio of 0.17—which suggests that *PERCC1* represents a bona fide protein-coding gene (Fig. 3d).

To compare the spatiotemporal expression of mouse *Percc1* with the in vivo regulatory activity of the ICR, we performed mRNA in situ hybridization for *Percc1*. At E14.5, we observed a pattern of punctate *Percc1* expression in the stomach, pancreas and intestine that strongly resembled the pattern of ICR activity (Fig. 2a, b, Extended Data Fig. 5).

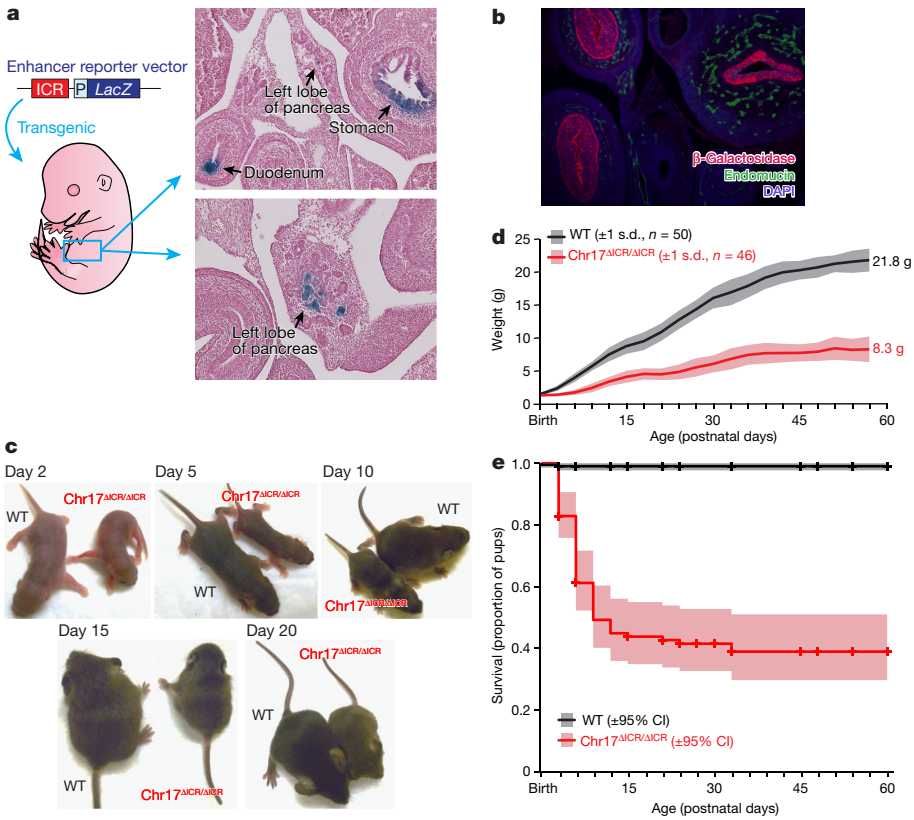


Fig. 2 | Enhancer activity of the ICR, and phenotypes of chr17$^{\Delta ICR/\Delta ICR}$ mice. **a**, **b**, Enhancer reporter activity in transgenic mouse embryos. **a**, Cross-sections of mouse embryos at E13.5, stained with X-gal to show the activity of β-galactosidase in the stomach, pancreas and duodenum. **b**, E14.5 cross-section showing immunofluorescence with anti-β-galactosidase (marking ICR activity; red), anti-endomucin (endothelial cells; green) and DAPI (DNA; blue). Two embryos for each experiment and each condition were collected and a minimum of three sections from each embryo were examined; representative sections are shown (**a**, **b**). **c**, Chr17$^{\Delta ICR/\Delta ICR}$ mice ($n = 46$) are viable but show a reduction in size and weight compared with wild-type (WT) littermates ($n = 50$). **d**, **e**, Reduction in body weight among surviving offspring (**d**) and increased mortality of chr17$^{\Delta ICR/\Delta ICR}$ (**e**) compared with wild-type mice. The body weights of female mice are shown in **d**; male wild-type and chr17$^{\Delta ICR/\Delta ICR}$ mice had similar genotype-dependent differences in weight. Shaded regions represent s.d. (**d**) or 95% confidence interval (CI) (**e**).

To further establish a functional connection between the ICR regulatory sequence and the predicted open reading frame, we used genome editing to disrupt the *Percc1* open reading frame in mice. *Percc1*$^{-/-}$ mice mimicked the key phenotypes of chr17$^{\Delta ICR/\Delta ICR}$ mice, including low body weight (Extended Data Fig. 6a) and abnormal appearance of faeces ($n = 11/12$ (92%) in *Percc1*$^{-/-}$; $n = 20/21$ (95%) in chr17$^{\Delta ICR/\Delta ICR}$; $P = 0.7$ by two-tailed *t*-test). Finally, to establish that PERCC1 expression is sufficient to rescue the phenotypes that result from deletion of the ICR regulatory sequence, we performed a complementation experiment in which we generated chr17$^{\Delta ICR/\Delta ICR}$ mice with an ICR-driven *Percc1* transgene. On complementation, we observed a reversal of the reduced body weight, high lethality and intestinal dysfunction found in chr17$^{\Delta ICR/\Delta ICR}$ mice (Extended Data Fig. 6b). These results demonstrate that a lack of gastrointestinal expression of PERCC1, which is normally controlled by the ICR, causes the phenotypes observed in chr17$^{\Delta ICR/\Delta ICR}$ mice—and probably those of the patients with IDIS who were examined in this study.

To investigate the cell-type specificity and function of PERCC1, we generated a transgenic mouse line that expresses a PERCC1–mCherry fusion protein (Fig. 4a, Extended Data Fig. 7). At P8, some PERCC1–mCherry-positive (PERCC1^{+}) cells were detected in the epithelium of the intestinal villi (Extended Data Fig. 7c), whereas distal stomach compartments displayed a high density of strongly positive cells, in particular in the epithelial layers of the pylorus, antrum and corpus (Fig. 4a, Extended Data Fig. 7b). Co-localization with the pan-endocrine marker synaptophysin (SYP) revealed that a major fraction of the PERCC1^{+} cells in these compartments are endocrine cells (Fig. 4a, Extended Data Fig. 7c). The vast majority of PERCC1^{+} cells detected in the antral epithelium were gastrin-expressing G cells (Fig. 4a), which are required for the secretion of gastric acid and promote growth of the gastrointestinal tract[9]. By contrast, the glandular epithelium at the entrance to the pyloric canal also contained clusters of PERCC1^{+} cells, but these cells did not display endocrine signatures and were located at the base of the gland—a region known to contain gastric stem cells[10] (Fig. 4a). As the majority of PERCC1^{+} cells in the distal stomach were G cells, we next investigated whether loss of *Percc1* in mice affected the development of these cells. We observed that the number of gastrin-expressing cells was reduced in the absence of *Percc1* (Extended Data Fig. 7d). Finally, analysis of 11,665 single-cell transcriptomes from mouse intestinal epithelium[11] showed that *Percc1* is expressed in a small proportion of cells that is strongly enriched for enteroendocrine cells (EECs) ($P = 2.9 \times 10^{-20}$ by chi-squared test; Extended Data Fig. 8a). *Percc1*-positive cells express *Sox4* and *Neurog3*, and their expression profiles are most consistent with an identity as enteroendocrine progenitors (Extended Data Fig. 8b, c). These data indicate that expression of PERCC1 in gastrointestinal tissue is restricted primarily to gastric G cells and duodenal EECs, and suggest that disrupted development of these cells causes the observed phenotype.

To characterize the molecular consequences of loss of the ICR or *Percc1* in more detail, we examined RNA sequencing (RNA-seq) data from relevant stages of mouse development. Among the 100 genes that showed the greatest reduction in expression, seven encode gastrointestinal peptide hormones that are secreted by EECs[12] (Fig. 4b, Table 1, Supplementary Table 1). Notably, one of the most robust changes was a reduction in the expression of gastrin (*Gast*), along with changes in the expression of somatostatin (*Sst*) and ghrelin (*Ghrl*) (Fig. 4c). We observed similar changes in gene expression in duodenal and stomach biopsies that were obtained from an ICR$^{\Delta L/\Delta L}$ patient, as compared with an unaffected ICR$^{+/+}$ sibling (Table 1, Supplementary Table 2, Extended Data Fig. 9). Together, these results are consistent with major disruptions of normal gastrointestinal physiology in chr17$^{\Delta ICR/\Delta ICR}$ mice and patients with IDIS, and highlight the close resemblance between the human disease condition and our mouse knockout models.

We also generated induced pluripotent stem (iPS) cells from an ICR$^{\Delta L/\Delta L}$ patient and an unaffected ICR$^{+/+}$ sibling and differentiated them into human intestinal organoids (HIOs)[13] (Extended Data Fig. 10). The gross morphology of ICR$^{\Delta L/\Delta L}$ and control ICR$^{+/+}$ HIOs was similar, and at early stages (21 days) the ICR$^{\Delta L/\Delta L}$ HIOs showed an abundant presence of EECs (Extended Data Fig. 7e). However, by day 42 the number of EECs was severely reduced in ICR$^{\Delta L/\Delta L}$ HIOs compared with control HIOs, and expression of the EEC markers SYP and chromogranin A (CHGA) was also reduced in ICR$^{\Delta L/\Delta L}$ HIOs

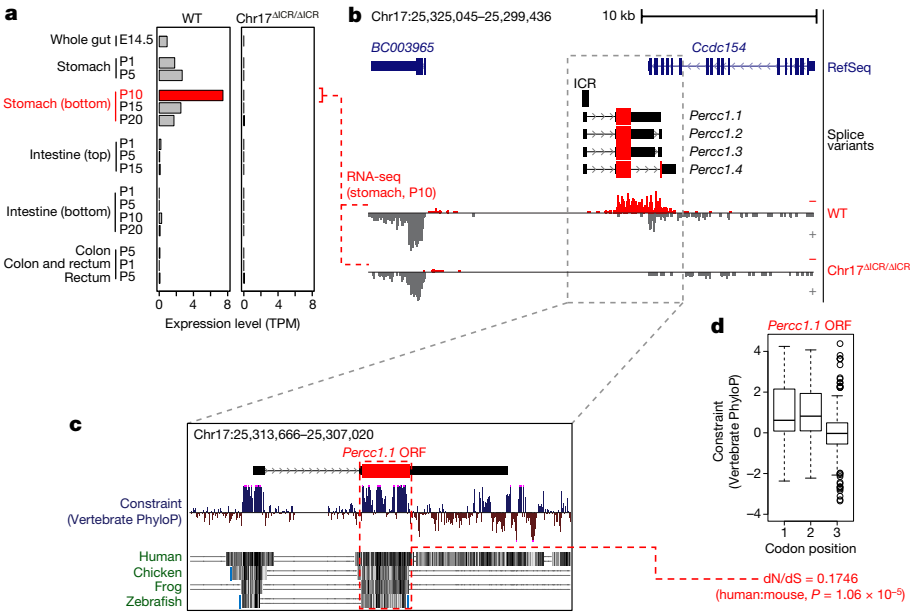


Fig. 3 | Discovery of a gene, *Percc1*, that flanks the ICR. a, Levels of *Percc1* expression in gastrointestinal tissues from wild-type and chr17$^{\Delta ICR/\Delta ICR}$ mice. The highest levels of expression were detected in the stomach at P10. TPM, transcripts per million mapped reads. **b**, Top, localization and structure of the *Percc1* gene within the mouse genome. Bottom, stranded RNA-seq data, indicating loss of expression of *Percc1* in chr17$^{\Delta ICR/\Delta ICR}$ mice compared with wild-type controls. **c**, Detailed view of the *Percc1* gene, and its evolutionary conservation. ORF, open reading frame. **d**, *Percc1* codon-position analysis, illustrating the relaxation of constraint in the third codon position of the predicted PERCC1 protein ($n = 274$). The dN/dS ratio and corresponding *P* value were calculated by phylogenetic analysis using maximum likelihood (chi-squared distribution). Box plots indicate median (centre line), interquartile values (box limits) and range (whiskers) of PhyloP scores, and PhyloP scores at individual positions (dots).

(Fig. 4c, Extended Data Fig. 7e). These results suggest that disruption of the ICR does not affect the initial formation of EECs, but interferes with their subsequent development.

Limited understanding of the in vivo functions of human protein-coding genes and noncoding sequences continues to be a challenge that hinders the systematic interpretation of disease-related data from WES and WGS studies. Here, we have established that IDIS—a severe, recessively inherited gastrointestinal disease—is caused by microdeletions that disrupt a regulatory sequence that is required for intestinal expression of the previously unannotated *PERCC1* gene. The molecular,

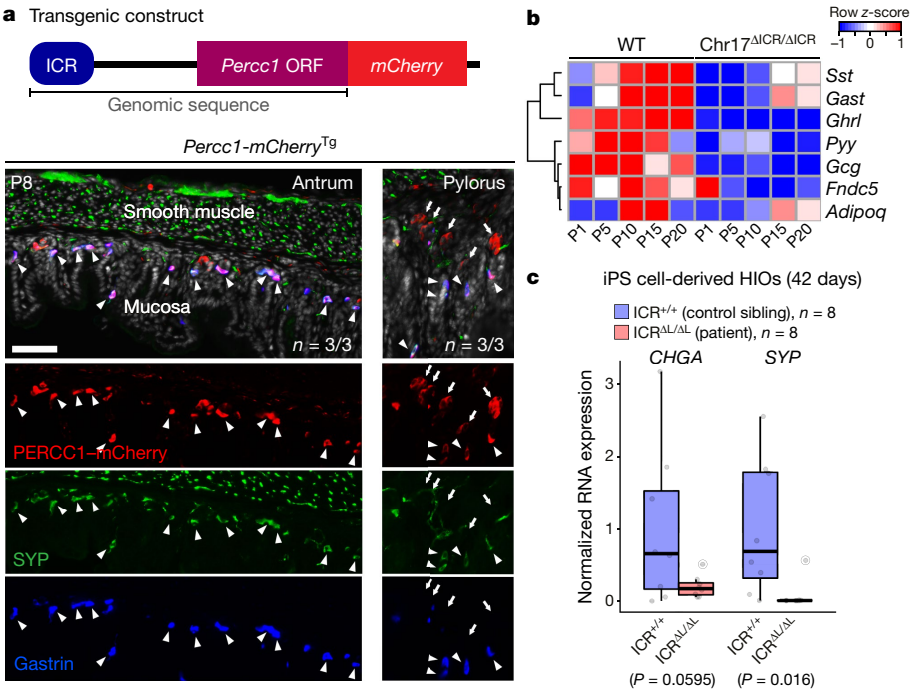


Fig. 4 | PERCC1 is abundant in G cells and its genetic disruption impairs the expression of gastrointestinal peptide hormones and the development of EECs. a, Top, generation of a reporter fusion transgene to track the localization of PERCC1 in mouse gastrointestinal tissues. The genomic sequence spanning the ICR and the *Percc1* open reading frame was fused to *mCherry*. Bottom left, PERCC1^{+} cells (red) in the pyloric antrum at P8 show endocrine identity (SYP; green) and extensive overlap with the endocrine subset of gastrin-expressing G cells (blue). Arrowheads indicate triple-positive cells. Nuclei are shown in grey. Bottom right, PERCC1^{+} cells in the pyloric canal either show endocrine G cell identity (arrowheads) or appear clustered at the base of the gland (arrows). *Percc1-mCherry*^{Tg} indicates transgenic mice for the *Percc1-mCherry* transgene. *n* represents independent biological replicates with similar results. Scale bar, 50 μm. **b**, RNA-seq of stomach samples from chr17$^{\Delta ICR/\Delta ICR}$ mice across different time points ($n = 1$ biological replicate), showing reduced transcript levels of different gastric peptide hormones (somatostatin (*Sst*), gastrin (*Gast*), ghrelin (*Ghrl*), peptide YY (*Pyy*), glucagon (*Gcg*), fibronectin type III domain-containing protein 5 (*Fndc5*) and adiponectin precursor (*Adipoq*)). **c**, Quantitative polymerase chain reaction with reverse transcription (RT–qPCR) analysis of iPS cell-derived HIOs from patients with disrupted PERCC1 (ICR$^{\Delta L/\Delta L}$), compared with control siblings (ICR$^{+/+}$). *SYP* is significantly downregulated in patient-derived HIOs ($P < 0.05$) (two-tailed, unpaired *t*-test), in contrast with *CHGA*. Box plots indicate median (centre line), interquartile values (box limits), range (whiskers), outliers (circled dots) and individual technical replicates (from independent organoid preparations) (dots).

Table 1 | Significant changes in gene expression in chr17$^{\Delta ICR/\Delta ICR}$ mice and corresponding changes in human gastrointestinal tissues

Gene	Gene description	Functional annotation	Stomach		Intestine	
			Mouse	Human	Mouse	Human
Downregulated						
Adipoq	Adiponectin	Hormone	6.46	NS	>100	NS
Sst	Somatostatin	Hormone	2.39	>100	NS	NS
Gast	Gastrin	Hormone	2.95	>100	NS	NS
Ghrl	Ghrelin	Hormone	26.82	5.49	4.65	>100
Fndc5	Fibronectin type III domain-containing protein 5	Hormone	3.22	NS	9.96	NS
Gcg	Glucagon	Hormone	4.96	NS	NS	NS
Pyy	Peptide YY	Hormone	2.90	NS	NS	NS
Abca13	ABC sub-family A, member 13	ABC transporter	NS	NS	9.86	NS
Abcc2	ABC sub-family C, member 2	ABC transporter	(+3.68)	NS	1.57	1.64
Isl1	ISL LIM homeobox 1	Transcription factor	1.92	7.75	NS	3.94
Upregulated						
Il6	Interleukin 6	Inflammation	NS	NS	>100	NS
Ccl3	C–C motif chemokine ligand 3	Inflammation	NS	NS	>100	NS
Cxcl5	C–X–C motif chemokine ligand 5	Inflammation	NS	NS	11.18	NS
Cxcl1	C–X–C motif chemokine ligand 1	Inflammation	NS	NS	8.82	2.19
C3	Complement component 3	Inflammation	2.30	(−4.88)	1.67	2.30
Spp1	Secreted phosphoprotein 1	Inflammation	NS	NS	9.78	NS
Il1rn	Interleukin 1 receptor antagonist	Inflammation	NS	NS	10.39	NS
Ggt1	γ-glutamyltransferase 1	Metabolism	2.36	5.47	7.85	NS
B4galnt2	β-1,4-*N*-acetyl-galactosaminyltransferase 2	Metabolism	2.71	5.66	1.56	NS
Nox1	NADPH oxidase 1	Metabolism	NS	NS	>100	NS
Duox2	Dual oxidase 2	IBD	NS	(−3.16)	14.21	50.66

A selection of genes representing major functional classes that are deregulated either in the stomach or in the intestine of P10 mice. Fold changes are shown when a significant difference compared with wild-type control was identified (fold changes of significant deregulation with opposite sign are shown in parentheses). The categories of genes were selected based on Gene Ontology and pathway enrichment of the complete list of deregulated genes in the mouse stomach (Supplementary Table 3). The indicated genes were all found among the top 100 up- or downregulated genes in the P10 mouse samples. Among the 100 genes that showed the largest increase in expression, 7 are related to inflammatory response, including interleukin 6 (*Il6*), complement 3 (*C3*) and 3 chemokine ligands (*Ccl3*, *Cxcl5* and *Cxcl1*) (Supplementary Table 1). ABC, ATP-binding cassette; IBD, inflammatory bowel disease; NS, not significant.

cellular and physiological phenotypes observed in patients and in engineered mice indicate that *PERCC1* is required for normal development of EECs and, thereby, for normal enteroendocrine hormone secretion. The phenotype of chr17$^{\Delta ICR/\Delta ICR}$ mice resembles that of mice with an intestine-specific deletion of *Neurog3* (a pro-endocrine transcription factor required for development of EECs[14]), which provides further evidence to suggest that abnormal development of EECs causes IDIS. The different aetiologies of chronic diarrhoea of infancy and enteropathies have recently been reviewed and classified into distinct categories[15]. Among these categories, disorders of EEC function that are caused by mutations in *NEUROG3*, *ARX*, *PCSK1* and *RFX6* are described as a unique and separate entity that manifests with malabsorptive diarrhoea. The phenotypes observed in patients and engineered mice indicate that IDIS is most similar to this specific class of chronic diarrhoeal disorders. Beyond congenital diarrhoea, our results serve as a reminder that, despite extensive annotation efforts, protein-coding genes associated with disease phenotypes remain to be discovered. As WGS is increasingly used for studies of rare diseases, our work underscores the importance of detailed experimental follow-up of such findings through in vivo models.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1312-2.

Received: 19 September 2014; Accepted: 16 May 2019;
Published online: 19 June 2019

1. Avery, G. B., Villavicencio, O., Lilly, J. R. & Randolph, J. G. Intractable diarrhea in early infancy. *Pediatrics* **41**, 712–722 (1968).
2. Straussberg, R. et al. Congenital intractable diarrhea of infancy in Iraqi Jews. *Clin. Genet.* **51**, 98–101 (1997).
3. Bamshad, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
4. Canani, R. B. & Terrin, G. Recent progress in congenital diarrheal disorders. *Curr. Gastroenterol. Rep.* **13**, 257–264 (2011).
5. Qu, H. & Fang, X. A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics* **11**, 135–141 (2013).
6. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
7. Eeckhoute, J. et al. Cell-type selective chromatin remodeling defines the active subset of FOXA1-bound enhancers. *Genome Res.* **19**, 372–380 (2009).
8. Pennacchio, L. A. et al. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
9. Dimaline, R. & Varro, A. Novel roles of gastrin. *J. Physiol.* **592**, 2951–2958 (2014).
10. Barker, N. et al. Lgr5^{+ve} stem cells drive self-renewal in the stomach and build long-lived gastric units *in vitro*. *Cell Stem Cell* **6**, 25–36 (2010).
11. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
12. Helander, H. F. & Fändriks, L. The enteroendocrine "letter cells" – time for a new nomenclature? *Scand. J. Gastroenterol.* **47**, 3–12 (2012).
13. Spence, J. R. et al. Directed differentiation of human pluripotent stem cells into intestinal tissue *in vitro*. *Nature* **470**, 105–109 (2011).
14. Mellitzer, G. et al. Loss of enteroendocrine cells in mice alters lipid absorption and glucose homeostasis and impairs postnatal survival. *J. Clin. Invest.* **120**, 1708–1721 (2010).
15. Thiagarajah, J. R. et al. Advances in evaluation of chronic diarrhea in infants. *Gastroenterology* **154**, 2045–2059 (2018).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2019

METHODS

Experimental design. All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory (LBNL) Animal Welfare and Research Committee. All mice used in this study were housed at the Animal Care Facility (ACF) at LBNL. Mice were monitored daily for food and water intake, and inspected weekly by the Chair of the Animal Welfare and Research Committee and the head of the animal facility in consultation with the veterinary staff. The LBNL ACF is accredited by the American Association for the Accreditation of Laboratory Animal Care (AAALAC). Generation of transgenic mice and manipulation of genomic sequence were performed in *Mus musculus* FVB strain mice, mice with C57/129S6 mixed background or W4/129S6 embryonic stem cells (see section 'Generation of ICR-knockout mice (chr17$^{\Delta ICR/\Delta ICR}$)' for details). Mice of both sexes were used in the analysis. Strategies for sample-size selection and randomization were conducted as described in the relevant paragraphs below. Cell lines were not tested for mycoplasma contamination.

Transgenic mouse assays. Sample sizes were selected empirically based on our previous experience of performing transgenic mouse assays for >2,000 total putative enhancers (VISTA Enhancer Browser, https://enhancer.lbl.gov/). Mouse embryos or postnatal mice were excluded from further analysis if they did not contain the reporter transgene or if the stage was not correct. All transgenic mice were treated with identical experimental conditions. Randomization and blinding of the investigators were unnecessary and not performed.

Genomic knockouts. Sample sizes were selected empirically on the basis of our previous studies[16]. All phenotypic characterization of knockout mice used a matched littermate selection strategy. All phenotyped homozygous knockout mice described in the paper resulted from crossing heterozygous knockout mice together to allow for the comparison of matched littermates of different genotypes. Embryonic samples used for RNA-seq and immunofluorescence were dissected blind to genotype. RNA-seq libraries were prepared and sequenced in mixed batches (including both knockout and wild-type samples).

Subjects. Patients with IDIS were recruited at Schneider and Sheba medical centres in Israel. Clinical details of the subjects are provided in Supplementary Table 4. All procedures performed in this study that involved human participants were in accordance with the ethical standards of the Sheba Medical Center institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all the individual participants (and/or their legal guardian) who were involved in the study.

Exome sequencing and identification of variants. Exome sequencing was performed using Agilent SureSelect Human All Exon technology (Agilent Technologies). The captured regions were sequenced using Genome Analyzer IIx (Illumina). The resulting reads were aligned to the reference genome (build GRCh37) using the Burrows–Wheeler Alignment tool[17]. We obtained a coverage of 70×, in which a base was considered covered if 5 or more reads spanned the nucleotide. Genetic differences relative to the reference genome were identified by SAMtools v.0.1.7, a variant-calling program[18] that identifies both single-nucleotide variants and small insertion–deletions (indels). Finally, Sequence Variant Analyzer software (SVA v.1.10)[19] was used to annotate all identified variants. For comparison to controls, we used 1,000 samples that were subjected to exome or whole-genome sequencing at the Center for Human Genome Variation (CHGV, Duke University), dbSNP, 1000 Genomes and the NHLBI Grand Opportunity Exome Sequencing Project.

WGS. WGS of individual 2.1 was performed at CHGV, using the Illumina HiSeq platform (Illumina), and analysed as described for exome data. A total of 275 CHGV whole-genome-sequenced, unrelated samples were used as controls. To detect copy number variants from WGS we used the Estimation by Read Depth with Single-nucleotide variants (ERDS) tool[20].

Collection of biopsies. Subjects underwent gastroduodenoscopy following Institutional Review Board approval (9881-12-SMC) at Sheba Medical Center and written informed consent of the patients and family members.

RNA extraction from biopsies. RNA isolation from frozen biopsies was performed using the TRI Reagent Protocol (Sigma-Aldrich) according to the manufacturer's instructions, or using the Qiagen RNeasy Mini Kit (Qiagen). The integrity of the samples was tested by measuring concentration and purity using a NanoDrop Spectrophotometer (Nanodrop Technologies).

RNA-seq of human samples. Total RNA was prepared according to the Illumina RNA-seq protocol. In brief, globin reduction, enrichment of poly (A) mRNA (polyA RNA), chemical fragmentation of the polyA RNA, cDNA synthesis and size selection of 200-bp cDNA fragments were performed. Next, the size-selected libraries were used for cluster generation on the flow cell, and prepared flow cells were run on the Illumina HiSeq2000 (Illumina). We obtained a total of 74.18 million paired-end reads of 100 bp for the affected sample and 72.53 million reads for the healthy sample. Reads were aligned to the human genome (build GRCh37) using TopHat v.2.0.4[21] with the default parameters. Quantification of gene expression was performed with Cuffdiff v.2.0.2[21], using the Illumina iGenome project UCSC annotation file as a reference. Differentially expressed genes were defined using the following thresholds: fold change of 1.5, $P \leq 0.05$.

RT–qPCR. RNA that was extracted from biopsies was used for RT–PCR expression analyses. RT–PCR experiments were performed using TaqMan Gene Expression Assays (Applied Biosystems), with the Applied Biosystems StepOnePlus (Applied Biosystems). From 1 μg of biopsy RNA, cDNA was synthesized using the SuperScript First-Strand Synthesis System for RT–PCR (Invitrogen), according to the manufacturer's instructions. A total of 20 μl cDNA was added with 30 μl water to 50 μl TaqMan Universal PCR Master Mix (Applied Biosystems) and the resulting 100-μl reaction mixtures were loaded onto a 96-well PCR plate. We used 14 different TaqMan Gene Expression Assays including three housekeeping genes with the following assay IDs: Hs00757713_m1 (*MLN*), Hs01074053_m1 (*GHRL*), Hs00175048_m1 (*NTS*), Hs00356144_m1 (*SST*), Hs00174945_m1 (*PYY*), Hs01062283_m1 (*GAST*), Hs00292465_m1 (*ARX*), Hs00174937_m1 (*CCK*), Hs00175030_m1 (*GIP*), Hs00219734_m1 (*GKN1*) and Hs00699389_m1 (*GKN2*). The housekeeping genes we used were *HMBS* (Hs00609297_m1), *ACTB* (Hs99999903_m1) and *GAPDH* (Hs99999905_m1). Reference cDNA samples were synthesized using 200 ng RNA from RNA extracted from stomach and duodenal tissues of two healthy controls (BioCat) for use in the normalization calculations. qPCR for expression analysis on the missing exons in *C16orf91* was done using cDNA extracted from the Human Digestive System MTC Panel (Clontech Laboratories).

Collection of serum. Whole blood was withdrawn into a Vacutainer serum tube without anticoagulant. The blood was immediately treated with 1 μM AEBSF (a protease inhibitor) and left at room temperature for 30 min to clot before centrifugation (15 min at 2,500 r.p.m. at 4 °C).

Enzyme-linked immunosorbent assay. Serum hormone levels were determined using the sandwich enzyme-linked immunosorbent assay (ELISA) technique. We used the following commercial kits according to the manufacturer's instructions: Human Ghrelin (Total) ELISA (Millipore), Human PYY (Total) ELISA (Millipore) and Human GIP (Gastric Inhibitory Polypeptide) ELISA (ENCO).

Linkage analysis and homozygosity mapping. Genome-wide single-nucleotide polymorphism (SNP) genotyping from the DNA of 6 children with IDIS and 22 relatives from families 1–5 was performed using the Illumina HumanCytoSNP 12 v.2.1_H, according to the manufacturer's recommendations (Illumina), in conjunction with SNP genotypes retrieved from whole-exome data. For linkage studies 35,845 informative equally spaced SNP markers were chosen after filtering for Mendelian errors and unlikely genotypes. Genotypes were examined with the use of a multipoint parametric linkage analysis and haplotype reconstruction for an autosomal recessive model with complete penetrance and a disease-allele frequency of 0.001, as previously described[22]. Homozygosity mapping was performed using PLINK v.1.07[23] with the default parameters (length 1,000 kb, SNP(N) 100, SNP density 50 kb/SNP, largest gap 1,000 kb).

Deletion analysis. Boundaries for the two deletion alleles were determined by PCR using amplified DNA and Sanger sequencing. The specific primers used to amplify across both deletions and inside the overlapping region for the two deletions are reported in Supplementary Table 5. In parallel, we used polymorphic markers that were identified by electronically screening genomic clones located on Chr16 0.86-2.8Mb. Primers were designed with Primer3 software (frodo.wi.mit.edu; from the Whitehead Institute, Massachusetts Institute of Technology). The specific primers used are reported in Supplementary Table 6. Amplification of the polymorphic markers was performed in a 25-μl reaction containing 50 ng DNA, 13.4 ng of each primer and 1.5 mM deoxyribonucleotide triphosphate (dNTP) in1.5 mM $MgCl_2$ PCR buffer with 1.2 U *Taq* polymerase (Bioline). After an initial denaturation of 5 min at 95 °C, 30 cycles were performed (94 °C for 2 min, 56 °C for 3 min and 72 °C for 1 min), followed by a final step of 7 min at 72 °C. PCR products were separated by electrophoresis on an automated genetic analyser (Prism 3100; Applied Biosystems). The breakpoint coordinates were chr16: 1475365-1482378 (for ΔL) and chr16: 1480850-1483951 (for ΔS), with an overlapping region at chr16: 1480850-1482378 (ICR).

Mouse transgenic assays. The candidate gene-regulatory sequence (chr 16: 1479875–480992) was amplified from human genomic DNA using PCR, and was cloned into the hsp68-lacZ transgenic vector containing a minimal *hsp68* promoter coupled to a *lacZ* reporter gene. The purified transgene construct was microinjected into fertilized FVB/N mouse oocytes, which were implanted into pseudopregnant foster females, and embryos were collected at E11.5–E14.5. Transgenic activity was determined by X-gal staining to detect β-galactosidase activity. Only patterns observed in at least three different embryos resulting from independent transgenic events were considered reproducible positive enhancers. All animal work was reviewed and approved by the LBNL Animal Welfare and Research Committee.

Generation of ICR-knockout mice (chr17$^{\Delta ICR/\Delta ICR}$). Homologous arms were generated by PCR (see Supplementary Table 7 for primers) and cloned into the

ploxPN2T vector, which contains a neomycin (G418)-resistant cassette flanked by *loxP* sites for positive selection, and an HSV-tk cassette for negative selection. Constructs were linearized and electroporated (20 μg) into W4/129S6 mouse embryonic stem cells (Taconic). The electroporated cells were selected under G418 (150 μg ml^{-1}) and 0.2 μM FIAU for a week. Surviving colonies were picked and expanded on 96-well plates, screened by PCR and by sequencing with primers that were outside but flanked the homology arm. Clones that were correctly targeted were electroporated with 20 μg of the Cre recombinase-expressing plasmid TURBO-Cre. TURBO-Cre was provided by T. Ley of the Embryonic Stem Cell Core of the Siteman Cancer Center, Washington University Medical School. Clones positive for Neo removal were screened by PCR and checked for G418 sensitivity. PCR products covering the deleted region and part of the homologous arms were gel-purified and sequenced to confirm the deletion of the ICR. Correctly targeted clones were subsequently injected into C57BL/6J embryos at the blastocyst stage. Chimeric mice were then crossed to C57BL/6J mice (Charles River) as well as 129S6/SvEvTac mice (Taconic) to generate heterozygous ICR-null mice, followed by breeding of heterozygous littermates to generate homozygous null mice.

Genotyping of chr17$^{\Delta ICR/\Delta ICR}$ mice. Genomic DNA was extracted from a 0.2–0.3-cm section of tail that was incubated overnight in lysis buffer (containing 100 mM Tris HCl pH 8.5, 5 mM EDTA, 0.2% SDS, 200 mM NaCl and 50 μg proteinase K) at 55 °C. Genotyping was carried out using standard PCR techniques (see Supplementary Table 7 for primers). Approximately 1–2 μl of 50–100-fold diluted tail lysate was used in a 20-μl PCR that contained 200 μM dNTP, 1.5 mM $MgCl_2$, 5 pmol of each forward and reverse primer and 0.5 U Taq polymerase.

RNA-seq of mouse tissues. Total RNA was extracted from different intestinal regions and stomach of mice at E11.5, P1, P5, P10, P15 and P20 using TRIzol Reagent (Invitrogen). RNA-seq libraries were then constructed using the Illumina TruSeq Stranded Total RNA Sample Preparation Kit, following the manufacturer's recommendations. The libraries were sequenced using a 50-bp single-end strategy with four samples per lane on an Illumina HiSeq instrument, and data were analysed using the same protocols as described for human, although with the mm9 mouse reference and Illumina iGenome project mouse genome annotation data. The RNA-seq data can be accessed at the Gene Expression Omnibus (GEO) under accession GSE94245. DAVID v.6.7[24] was run separately on the differentially expressed genes in the P10 stomach (167 downregulated, 187 upregulated, Supplementary Table 1) or in the P10 intestine (326 downregulated, 327 upregulated, Supplementary Table 1). The resulting lists of clustered terms were inspected for highly significant enrichment (q-value ≤ 0.05) for biological process (Gene Ontology), molecular function (MF), or pathway (Kyoto Encyclopedia of Genes and Genomes; KEGG) (see Supplementary Table 3). The most significant term for each cluster was retained. In total, 59 genes showing any of these annotations were also found in the top 100 deregulated genes (either up- or downregulated, separately for each tissue). A selection of genes from these 59 is highlighted in Table 1.

16S rRNA amplicon analysis (iTags) of microbial community diversity. Samples of faeces and gut content were collected from chr17$^{\Delta ICR/\Delta ICR}$ mice and wild-type littermates. DNA was extracted from these samples using the PowerFecal DNA Isolation Kit (MO Bio Laboratories). V4 16S regions were amplified from the DNA samples using barcoded primers and 5 PRIME HotMasterMix (Fisher Scientific) as previously described[25]. Amplicons were pooled in equal amounts, purified with AMPureXP magnetic beads (Beckman Coulter) and sequenced.

***Percc1* RT–PCR, RT–qPCR and 5′ or 3′ rapid amplification of cDNA ends.** Total RNA was extracted from the intestine and stomach of mice using TRIzol Reagent (Thermo Fisher Scientific) and treated with DNase (Promega). First strand cDNA was generated from the DNase-treated total RNA using the SuperScript First-Strand Synthesis System (Thermo Fisher Scientific). RT–qPCR was performed using the KAPA SYBR FAST Roche LightCycler 480 (2X) qPCR Master Mix (KAPA Biosystems). To identify the expression boundaries of the *Percc1* transcript(s), regular RT–PCR, 5′ rapid amplification of cDNA ends (5′-RACE) and 3′-RACE were performed (see Supplementary Table 8 for primers). Standard RT–PCR used Platinum *Taq* DNA Polymerase High Fidelity (Thermo Fisher Scientific); 5′- and 3′- RACE used the SMARTer RACE 5′/3′ Kit (Clontech). PCR products were gel-purified and Sanger-sequenced. The cDNA and predicted protein sequence are available in GenBank (record KY964488). It should be noted that since our original identification of *Percc1*, the Havana gene annotation project manually curated this open reading frame as uncharacterized LOC105371045, also known as AL032819.3 (http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000284395;r=16:1431035-1433397;t=ENST00000640283).

Generation of *Percc1*-knockout mice by CRISPR–Cas9. *Percc1*-knockout mice were generated by CRISPR–Cas9 editing as previously described[26,27]. Single-guide RNAs (sgRNAs) were constructed using 60-mer oligonucleotides and an sgRNA cloning vector (Addgene plasmid 41824)[28] according to the protocol described in http://www.addgene.org/static/cms/files/hCRISPR_gRNA_Synthesis.pdf. The sgRNA target-site sequences are provided in Supplementary Table 9. *Cas9* mRNA was generated using a human codon-optimized *Cas9* gene from plasmid pDD921[29]. T7-promoter-Cas9-polyA and T7 promoter-sgRNA amplicons were amplified by PCR from pDD921 and sgRNA clones, respectively, using Phusion polymerase (New England Biolabs). *Cas9* RNA was generated by in vitro transcription from the T7-promoter-Cas9-polyA amplicon using the mMESSAGE mMACHINE T7 Kit (Thermo Fisher Scientific), following the manufacturer's instructions. sgRNA was generated by in vitro transcription from the T7-promoter-sgRNA amplicon using the MEGAshortscript Kit (Thermo Fisher Scientific), following the manufacturer's instructions. In vitro-transcribed RNA was cleaned using the MEGAclear Kit (Thermo Fisher Scientific), following the manufacturer's instructions. RNA was eluted into RNase-free microinjection buffer (10 mM Tris pH 7.5; 0.1 mM EDTA). The RNA was then assessed by electrophoresis on a 10% TBE urea PAGE gel.

Microinjection and generation of genetically modified mice. A mixture of 100 ng μl^{-1} *Cas9* RNA and 50 ng μl^{-1} total sgRNA in microinjection buffer was injected into the cytoplasm of fertilized mouse strain FVB/N mouse oocytes. The pups that were generated (F_0) were screened by PCR (primers listed in Supplementary Table 9) and the resulting PCR products were sequenced to identify deletion breakpoints.

Complementation of chr17$^{\Delta ICR/\Delta ICR}$ with a *Percc1* mouse transgene. An 8,530-bp region was PCR-amplified from W4/129S6 mouse genomic DNA (see Supplementary Table 10 for primer sequences). The region amplified included all putative *Percc1* exons, a possible promoter and the 3′-untranslated region. The resulting PCR product was cloned into the pCR2.1 vector (Invitrogen TOPO TA Cloning Kit) and verified by sequencing. To generate transgenic mice, a purified, linear transgene fragment was injected into the pronucleus of fertilized eggs from chr17$^{\Delta ICR/+}$ intercrosses. Offspring were genotyped by PCR for existence of the *Percc1* transgene and examined for phenotypes found in chr17$^{\Delta ICR/\Delta ICR}$ chr17 mice.

Generation of transgenic mice that express the PERCC1–mCherry fusion protein. A 3,308-bp fragment that covered the *Percc1* promoter to the last amino acid of the *Percc1* coding region was amplified by PCR and cloned into the pcDNA3 mCherry LIC cloning vector 6B (Addgene, 30125) at the 5′-end of the mCherry cDNA to generate a mouse *Percc1-mCherry* fusion construct (see Supplementary Table 11 for primers). The construct was linearized with NotI and injected into the pronucleus of fertilized FVB eggs to generate transgenic mice.

Western blotting. Tissues (stomach and intestine) were lysed and homogenized in T-PER Tissue Protein Extraction Reagent (Thermo Fisher Scientific, 78510) plus Protease Inhibitor (Sigma, P8340). Protein concentration was determined using the Pierce Coomassie (Bradford) Protein Assay Kit (Thermo Fisher, 23200). A total of 100 μg of lysate per well of each sample was loaded on Bolt 4–12% Bis-Tris Plus Gel (Thermo Fisher, NW04120BOX), along with SeeBlue Plus2 Pre-Stained Standard (Thermo Fisher Scientific, LC5925) and 0.1 μg mCherry protein (VWR 10190-818) as a positive control. After electrophoresis at 200 V for 35 min, the gel was blotted on a PVDF membrane with a pore size of 0.45 μm, using XCell II Blot Module (Thermo Fisher Scientific). The membrane was then blocked with 3% nonfat milk in TBS (Sigma, T8793). The primary antibody, mCherry antibody (Thermo Fisher Scientific, PA5-34974), was added to the blot at 1:1,000 dilution in 3% nonfat milk in TBS, and incubated overnight at 4 °C. Following several washes with TBST (Sigma, T9039-10PAK), horseradish peroxidase (HRP)-conjugated goat anti-rabbit antibody (Thermo Fisher, 31460) was added at 1:3,000 dilution in TBST and incubated at room temperature for 1 h. mCherry protein was detected with 1-Step Ultra TMB-Blotting Solution (Thermo Fisher Scientific, 37574).

Tissue embedding and cryosectioning. The stomach was dissected out from *Percc1-mCherry* transgenic mice at P8 and fixed in 4% PFA for 2–3 h at 4 °C. After several washes with PBS, the tissue was soaked in 10%, 20% and 30% sucrose in PBS, each for 1 h at 4 °C, and then embedded in a 1:1 mixture of Tissue-Tek OCT Compound (VWR, 25608-930) and 30% sucrose in PBS. Embedded tissue was sectioned using a cryostat (10-μm sections).

Immunofluorescence and cell counting. Cryosections were washed 3 times with PBS for 5 min each wash, incubated in 0.2% TritonX-100 in PBS for 20 min and washed again 3 times with PBS for 5 min each wash at room temperature. Sections were blocked with 1% BSA in 0.1% PBT for 1 h at room temperature. Primary antibodies against mCherry (1:500, Thermo Fisher Scientific, PA5-34974), synaptophysin (1:1,000, Synaptic Systems, 101004), gastrin (1:500, C-20, Santa Cruz Biotechnology, sc-7783), anti-E-cadherin (1:500, BD Biosciences, 610181), β-galactosidase (1:500, Abcam, ab9361) and endomucin (1:500, eBiosciences, 14-5851-82) were used and incubated overnight at 4 °C. After washing sections 3 times with PBS for 5 min each wash, sections were incubated for 1 h at room temperature with combinations of Alexa Fluor 488 and 594 (for double fluorescence) or Alexa Fluor 488, 568 and 647 (for triple fluorescence) conjugate secondary antibodies (Thermo Fisher Scientific), each at a 1:1,000 dilution in 1% BSA in PBS. Sections were then washed twice with PBS, treated with Hoechst for counterstaining of nuclei and mounted in Mowiol or VECTASHIELD HardSet Mounting Medium (Vector Laboratories, H-1500). Gastrin-positive cells were counted on consecutive stomach sections of wild-type ($n = 3$) and ICR-knockout ($n = 2$) mice at P8. The

fraction of gastrin-expressing cells was calculated by normalization to the total number of nuclei determined via Cell Profiler (http://cellprofiler.org). Cell counts from each antral side (as shown in the schematic of Extended Data Fig. 7d) were averaged.

Single-cell analysis of mouse intestine. Single-cell RNA-seq profiles[11] were retrieved from the GEO[30] (accession GSE92332). SAM alignment files[18] were downloaded from the Short Read Archive using the sam-dump utility (part of the SRA Toolkit v.2.8.1; http://ncbi.github.io/sra-tools/). Coordinates of the reads that mapped to *Percc1* were extracted using SAMtools v.1.3.1[18], along with the corresponding cell barcodes and unique molecular identifiers (UMIs) (CB:Z and UB:Z, respectively). These allowed the assignment of each unique transcript (via UMI) to the parental cell (via cell barcode). Cell-type classification, as previously described[11], was parsed from GSE92332_Regional_UMIcounts.txt.gz (available at the GEO; GSE92332). All the further data-processing steps, plots and calculations of *P* values were performed in R v.3.4.2 (www.r-project.org).

Histological analysis of human biopsies. FFPE blocks were sectioned at a thickness of 4 μm and a positive control was added on the right side of the slides. All immunostainings were fully calibrated on a Benchmark XT staining module (Ventana Medical Systems). In brief, after sections were dewaxed and rehydrated, a CC1 Standard Benchmark XT pre-treatment for antigen retrieval (Ventana Medical Systems) was selected for all immunostainings: CHGA (1:500, Dako), SYP, (1:200, Life Technologies, Invitrogen), GAST and somatostatin (STT). Detection was performed with the iView DAB Detection Kit (Ventana Medical Systems) and counterstained with haematoxylin (Ventana Medical Systems). After the run on the automated stainer was completed, slides were dehydrated in ethanol solutions (70%, 96% and 100%) for one minute each. Sections were then cleared in xylene for two minutes, mounted with Entellan and cover slips were added.

Generation of iPS cells from patient lymphocytes. Whole blood was isolated by routine venipuncture from patient 2.1 and two healthy siblings (2.3 (heterozygous carrier) and 2.4 (unaffected wild type)) at Sheba Medical Center in Israel, in preservative-free 0.9% sodium chloride containing 100 U ml^{-1} heparin. Blood was then shipped overnight to Cincinnati Children's Hospital Medical Center for generation of iPS cells. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood by Ficoll centrifugation as previously described[28] and were used to derive iPS cells. In brief, PBMCs were cultured for 4 d in DMEM containing 10% FCS, 100 ng ml^{-1} SCF, 100 ng ml^{-1} TPO, 100 ng ml^{-1} IL3, 20 ng ml^{-1} IL6, 100 ng ml^{-1} Flt3L, 100 ng ml^{-1} GM-CSF and 50 ng ml^{-1} M-CSF (Peprotech). Transduction using a polycistronic lentivirus that expressed OCT4, SOX2, KLF4, cMYC and dTomato was performed[31] following the second day of culture in this medium. Transduced cells were then cultured for an additional 4 d in DMEM containing 10% FCS, 100 ng ml^{-1} SCF, 100 ng ml^{-1} TPO, 100 ng ml^{-1} IL3, 20 ng ml^{-1} IL6 and 100 ng ml^{-1} FLT3L. The medium was changed every other day. PBMCs were then plated on 0.1% gelatin-coated dishes containing 2×10^4 irradiated mouse embryonic fibroblasts per cm^2 (GlobalStem), and cultured in human embryonic stem cell (hESC) medium containing 20% knockout serum replacement, 1 mM L-glutamine, 0.1 mM β-mercaptoethanol, 1 × non-essential amino acids and 4 ng ml bFGF until formation of colonies of iPS cells. Putative iPS cell colonies were then manually excised and replated in feeder-free culture conditions that consisted of matrigel (BD BioSciences) and mTeSR1 (STEMCELL Technologies). Lines that exhibited robust proliferation and maintenance of stereotypical human pluripotent stem cell morphology were then expanded and cryopreserved before use in experiments. Standard metaphase spreads and G-banded karyotypes were determined by the CCHMC Cytogenetics Laboratory.

Differentiation of iPS cells into intestinal organoids. The differentiation of human iPS cells was performed as previously described[13,32,33] with minor modifications. In brief, two clonal iPS cell lines from each donor were dispase-passaged into a matrigel-coated 24-well tissue-culture plate and cultured for 3 d in mTeSR1. Following definitive endoderm differentiation, the monolayers were treated for 4 d with RPMI medium 1640 (Gibco) containing 2% defined fetal calf serum, 1 × non-essential amino acids, 3 μM CHIR99021 (Stemgent) and 500 ng ml^{-1} rhFGF4 (R&D Systems) to induce hindgut spheroid morphogenesis. After the fourth day, 'day 0' HIOs were collected, embedded in matrigel matrix and cultured in Advanced DMEM/F12 (Gibco) containing 100 U ml^{-1} penicillin/streptomycin (Gibco), 2 mM L-glutamine (Gibco), 15 mM HEPES (Gibco), N2 supplement (Gibco), B27 supplement (Gibco) and 100ng ml^{-1} rhEGF (R&D Systems) for up to 42 days, splitting, passaging and changing the medium periodically.

HIOs that were collected for immunofluorescence analysis were fixed in 4% paraformaldehyde for 1–2 h at room temperature, washed overnight at 4 °C in PBS and embedded in OCT compound (Sakura). Eight-to-ten-micrometre-thick sections were incubated with primary antibodies overnight at 4 °C in 10% normal donkey serum and 0.05% Triton X-100 in PBS solution, and subsequently incubated with secondary antibodies for 1 h at room temperature. The primary antibodies used were: FOXA2 (1:500, Novus), E-cadherin (1:500, R&D Systems), SYP (1:1,000, Synaptic Systems), CDX2 (1:500, Biogenex), PDX1 (1:5,000, Abcam). All secondary antibodies (AlexaFluor; Invitrogen) were used at 1:500 dilution. Confocal microscopy images were captured with a 20× plan apo objective on a Nikon A1Rsi Inverted microscope, using settings of 0.5 pixel dwell time, 1,024 resolution, 2× line averaging and 2.0 × A1 plus scan.

Total RNA was extracted from HIOs using a NucleoSpin RNA II kit (Macherey-Nagel), and cDNA was synthesized with SuperScript VILO (Invitrogen) using 300 ng RNA. qPCR analysis was performed with TaqMan Fast Advanced Master Mix and custom-designed TaqMan Array 96-Well FAST Plates (Applied Biosystems) consisting of the following targets: 18S-Hs99999901_s1; GAPDH-Hs99999905_m1; ARX-Hs00292465_m1; CHGA-Hs00900370_m1; SYP-Hs00300531_m1; NTS-Hs00175048_m1.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All RNA-seq data used in this study have been deposited in the GEO repository (National Center for Biotechnology Information). The files are accessible through the GEO accession number GSE94245. The cDNA and predicted protein sequence for *Percc1* are available in GenBank (record KY964488). All other relevant data are available from the corresponding authors on request.

16. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
18. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
19. Ge, D. et al. SVA: software for annotating and visualizing sequenced human genomes. *Bioinformatics* **27**, 1998–2000 (2011).
20. Zhu, M. et al. Using ERDS to infer copy-number variants in high-coverage genomes. *Am. J. Hum. Genet.* **91**, 408–421 (2012).
21. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
22. Bockenhauer, D. et al. Epilepsy, ataxia, sensorineural deafness, tubulopathy, and *KCNJ10* mutations. *N. Engl. J. Med.* **360**, 1960–1970 (2009).
23. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
24. Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44–57 (2009).
25. Lindemann, S. R. et al. The epsomitic phototrophic microbial mat of Hot Lake, Washington: community structural responses to seasonal cycling. *Front. Microbiol.* **4**, 323 (2013).
26. Wang, H. et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
27. Yang, H. et al. One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**, 1370–1379 (2013).
28. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
29. Kvon, E. Z. et al. Progressive loss of function in a limb enhancer during snake evolution. *Cell* **167**, 633–642 (2016).
30. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
31. Warlich, E. et al. Lentiviral vector design and imaging approaches to visualize the early stages of cellular reprogramming. *Mol. Ther.* **19**, 782–789 (2011).
32. McCracken, K. W., Howell, J. C., Wells, J. M. & Spence, J. R. Generating human intestinal tissue from pluripotent stem cells in vitro. *Nat. Protocols* **6**, 1920–1928 (2011).
33. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L. & Roach, J. C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* **27**, 3216–3217 (2011).

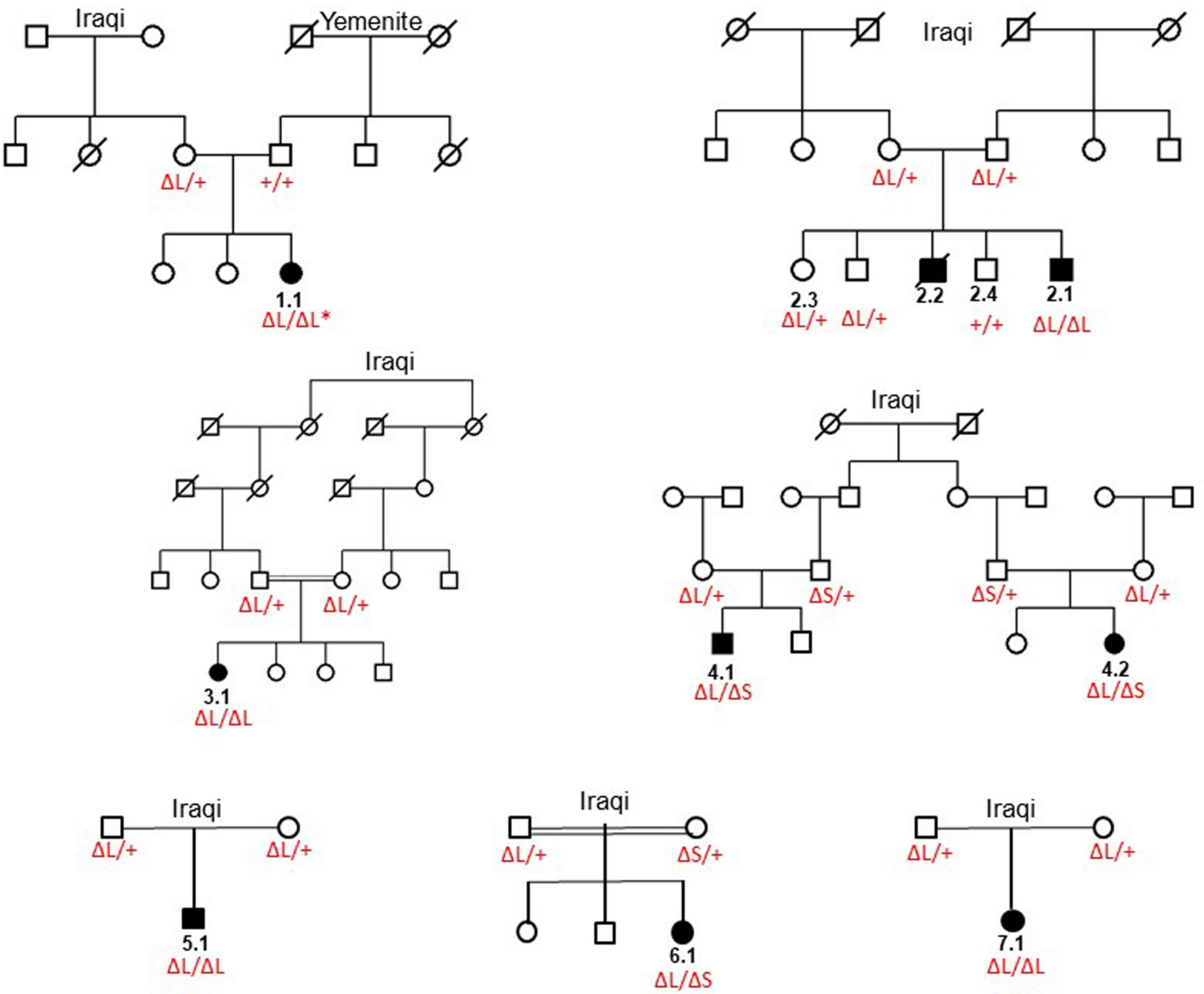
Acknowledgements The authors thank the patients and their families for their cooperation and support. This work was supported by grants to D.L. from the SysKid EU FP7 project (241544), the Wolfson Family Charitable Trust and the Crown Human Genome Center at the Weizmann Institute of Science. A.V. and L.A.P. were supported by NHLBI grant R24HL123879 and NHGRI grants R01HG003988, U54HG006997 and UM1HG009421; J.M.W. and Y.A. were supported by the Cincinnati Children's Hospital and Sheba Medical Center's Joint Research Fund; J.M.W. and M.F.K. were supported by NIH grants 1R01DK092456 and 1U18NS080815 as well as a digestive disease center grant (P30 DK0789392); R.K. was supported by the David and Elaine Potter Charitable Foundation; B.L.B. was supported by NIH grants HL089707, HL064658 and HL136182; and I. Barozzi was funded through an Imperial College Research Fellowship. Research was conducted at the E. O. Lawrence Berkeley National Laboratory and performed under the Department of Energy contract DE-AC02-05CH11231 (University of California). iPS cell lines were generated in collaboration with the Cincinnati Children's Pluripotent Stem Cell Facility. This work was performed in partial fulfilment of the requirements for a PhD degree for D.O.-L. (Weizmann Institute of Science, Rehovot, Israel) and I.B.-J. (The Sackler Faculty of Medicine, Tel Aviv University, Israel).

Author contributions C.H., R. Shamir, R. Shapiro, B.W., B.P.-S., P.T., I. Barshack, E.P. and Y.A. recruited patients, provided patient care and characterized the symptoms. B.W., B.P.-S. and Y.A. obtained biopsies. D.O.-L., D.B.G., E.K.R. and Y.H. managed the sequencing, D.O.-L., T.O., I. Barozzi and A.A. performed the bioinformatic analyses and discovered the mutation. M.T., H.C.S. and R.K. provided SNP genotyping and linkage analysis. I.B.-J., D.M.-Y., H.R.-W., M.O., E.S.M.V., R.M., M.S., I. Barshack and W.d.L. did experimental work, including mutation characterization. Y.Z., M.O., A.S.N., V.A., D.M.I., D.C.-D., D.E.D., K.L.v.B., R.M.B., B.L.B., A.V. and L.A.P. did the transgenic and knockout mouse generation and characterization. J.M.W., M.F.K., A.P. and C.N.M. did the iPS cell generation and human intestinal organoid studies. D.O.-L., A.V., L.A.P. and D.L. wrote the manuscript. R. Shamir, D.B.G., E.P., L.A.P., D.L. and Y.A. provided leadership to the project. All authors contributed to the final manuscript.

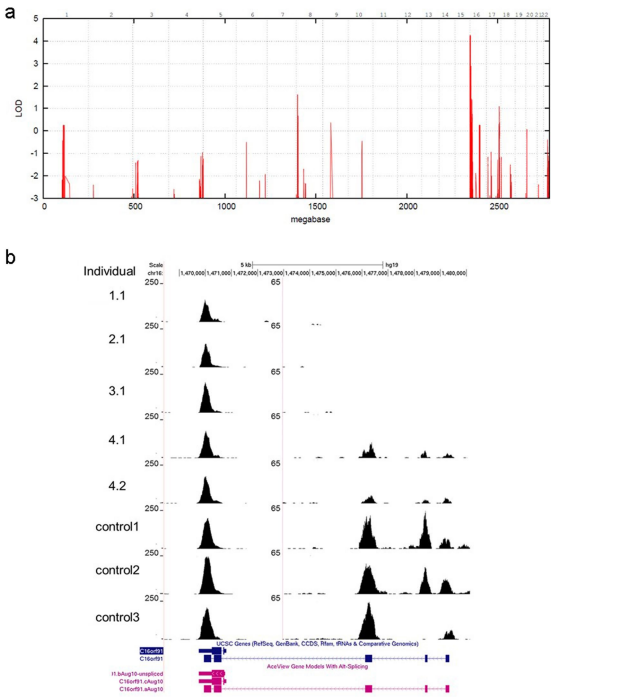
Competing interests The authors declare no competing interests.

Additional information

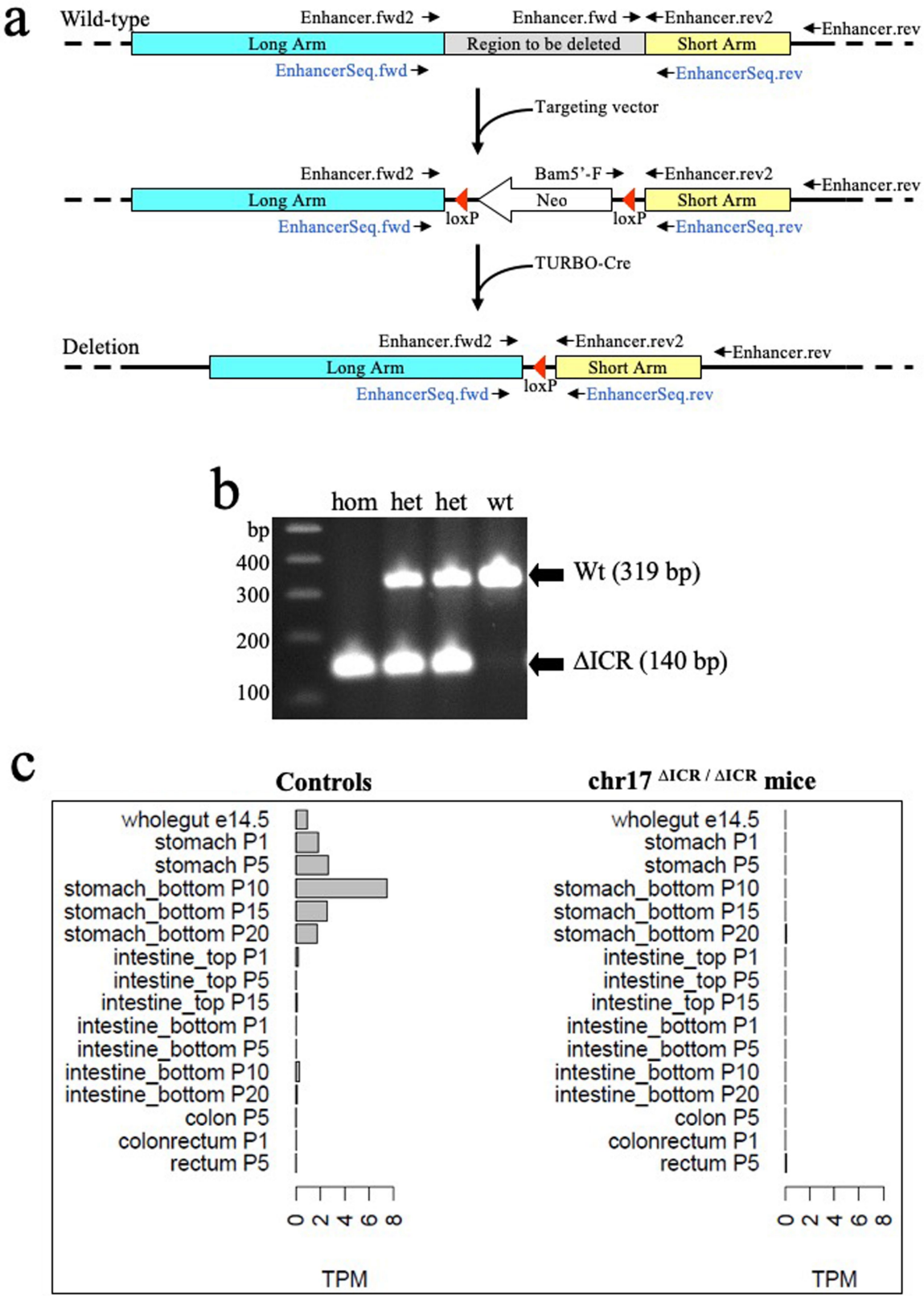
Extended data. is available for this paper at https://doi.org/10.1038/s41586-019-1312-2.
Supplementary information. is available for this paper at https://doi.org/10.1038/s41586-019-1312-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.L., Y.A. or L.A.P.



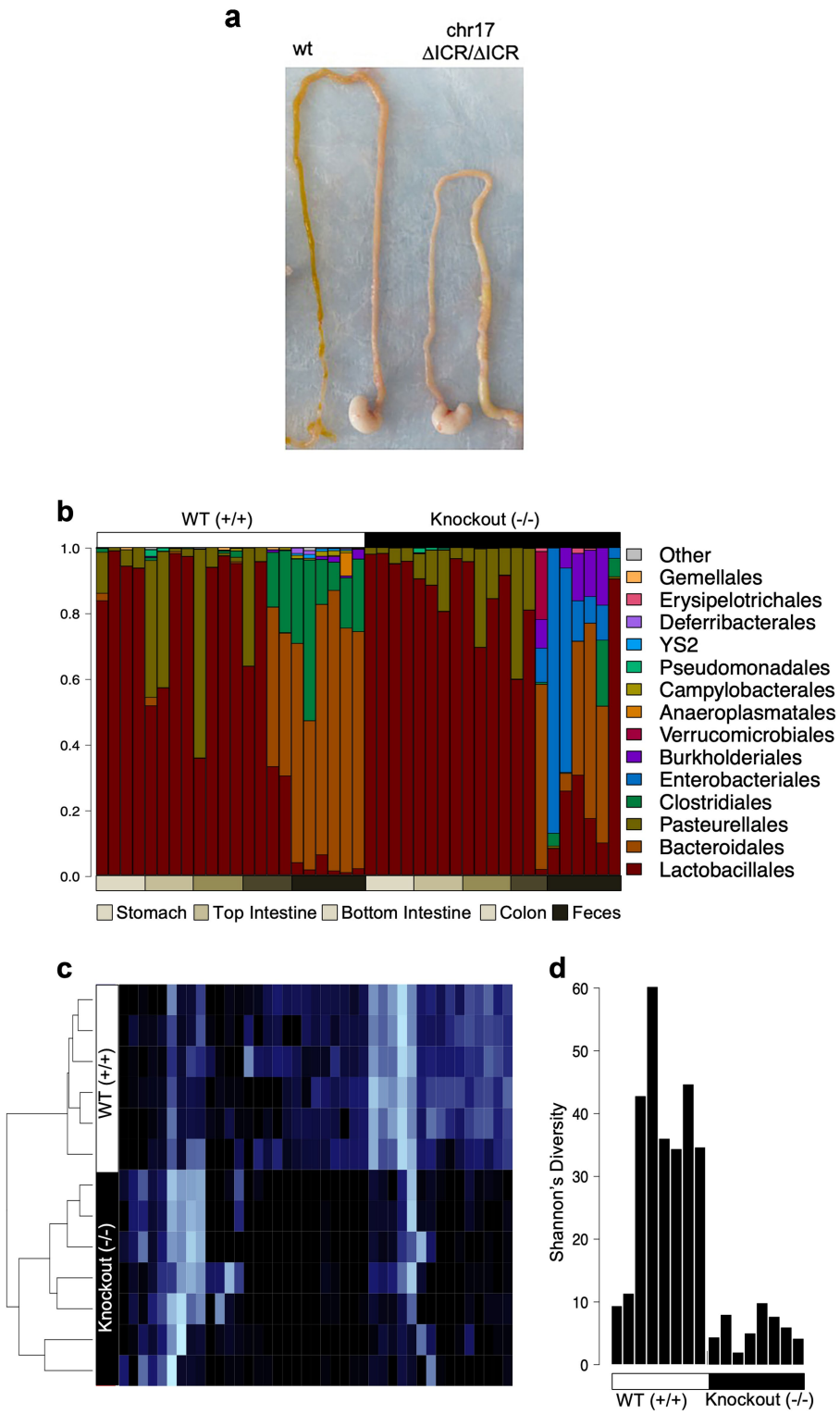
Extended Data Fig. 1 | Family pedigrees. Filled black symbols represent affected individuals, and deletion genotypes are indicated in red. WES was done for individuals 1.1, 2.1, 3.1, 4.1 and 4.2, WGS was done for individual 2.1 and transcriptome analysis was done for individuals 2.1 and 2.4. Patient 1.1 (marked with an asterisk) was found to have uniparental disomy.



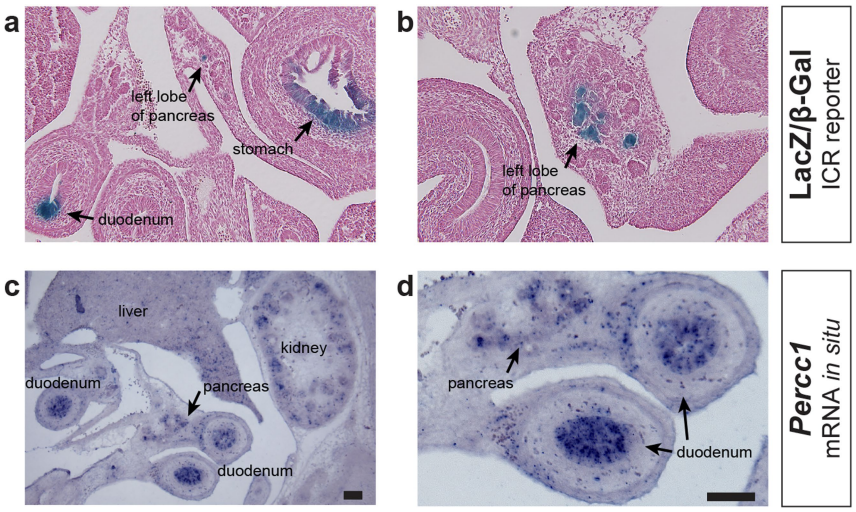
Extended Data Fig. 2 | Genetic analysis in IDIS. a, Analysis of the SNP genotyping that was performed on 6 of the patients in families 1–5 and their 22 relatives detected a single significant telomeric linkage interval on chr16 with a maximum LOD score of 4.26. Haplotype reconstruction confirmed this interval, with flanking marker rs207435 (chr16: 2,984,868), and showed two distinct disease haplotypes either in a homozygous setting (in affected individuals for disease allele 1 (that is, ΔL) in families 2, 3, 5) or in a compound heterozygous setting (in affected individuals for disease alleles 1 and 2 (that is, ΔS) in family 4). All the affected individuals who carried disease allele 1 showed an identical disease haplotype from rs533184 (chr16: 1,155,025) to rs397435 (chr16: 2,010,138). **b**, Schematic of reads covering exons in the *C16orf91* gene, for the five exome-sequenced patients and for three unaffected controls who underwent sequencing under identical conditions. The first three patients (individuals 1.1, 2.1 and 3.1), who had a chr16$^{\Delta L/\Delta L}$ genotype, had zero coverage in the three upstream exons (right). The last two patients (individuals 4.1 and 4.2), who had a chr16$^{\Delta L/\Delta S}$ genotype, had non-zero coverage in these exons, but coverage was lower than in controls. All subjects had high coverage in the downstream exons (left). Numbers indicate the scale in sequencing reads per base.



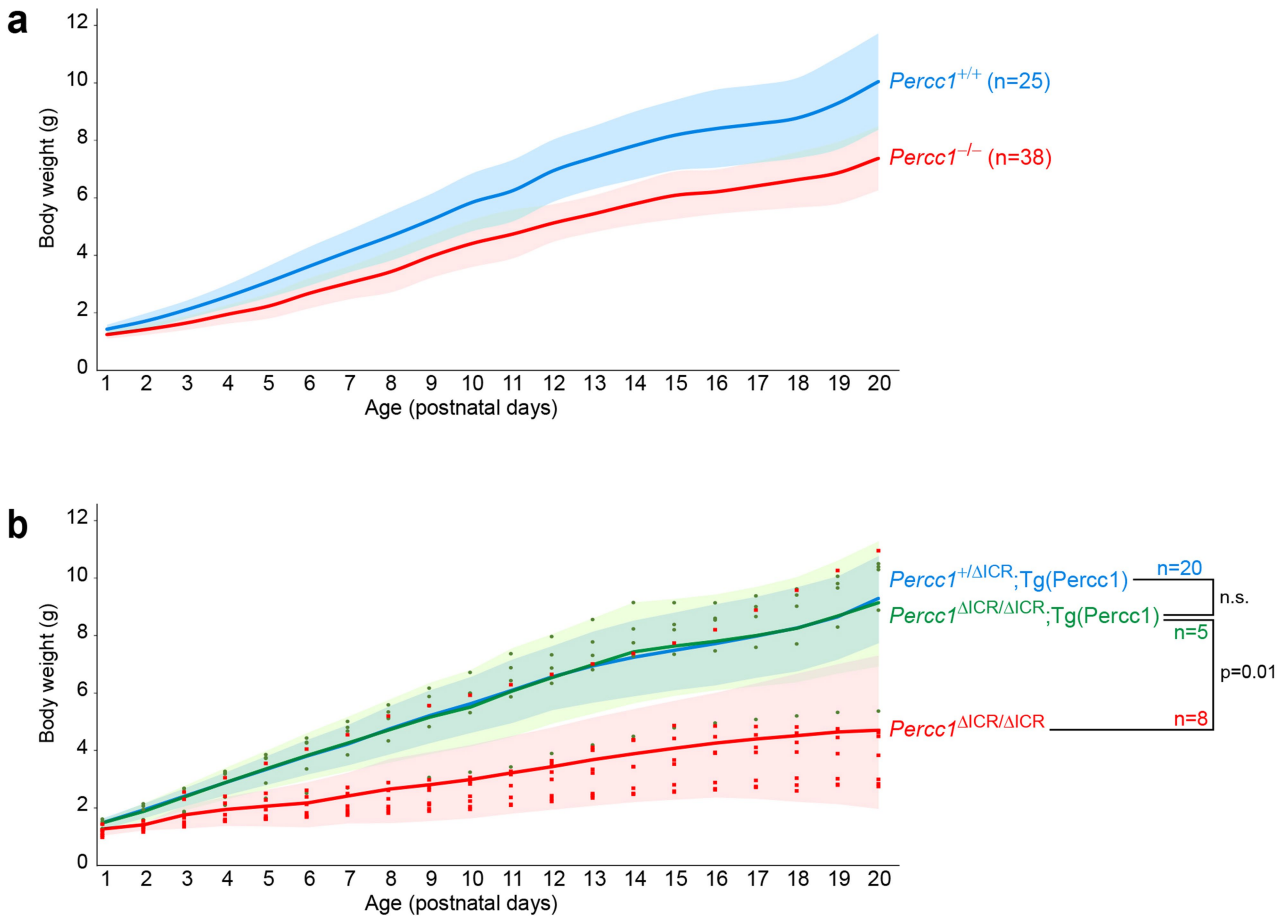
Extended Data Fig. 3 | Targeted deletion of the ICR noncoding sequence in mice. a, Overview of targeting approach. See Methods for details. **b**, Genotyping results obtained from genomic DNA ($n = 554$) isolated from the tails of homozygous and heterozygous ICR-knockout (ΔICR) mice, compared to a wild-type control. See Methods for primers and details. **c**, *Percc1* expression derived from RNA-seq from control littermates (left) and knockout mice (right). Tissues and time points are indicated to the left of each plot.



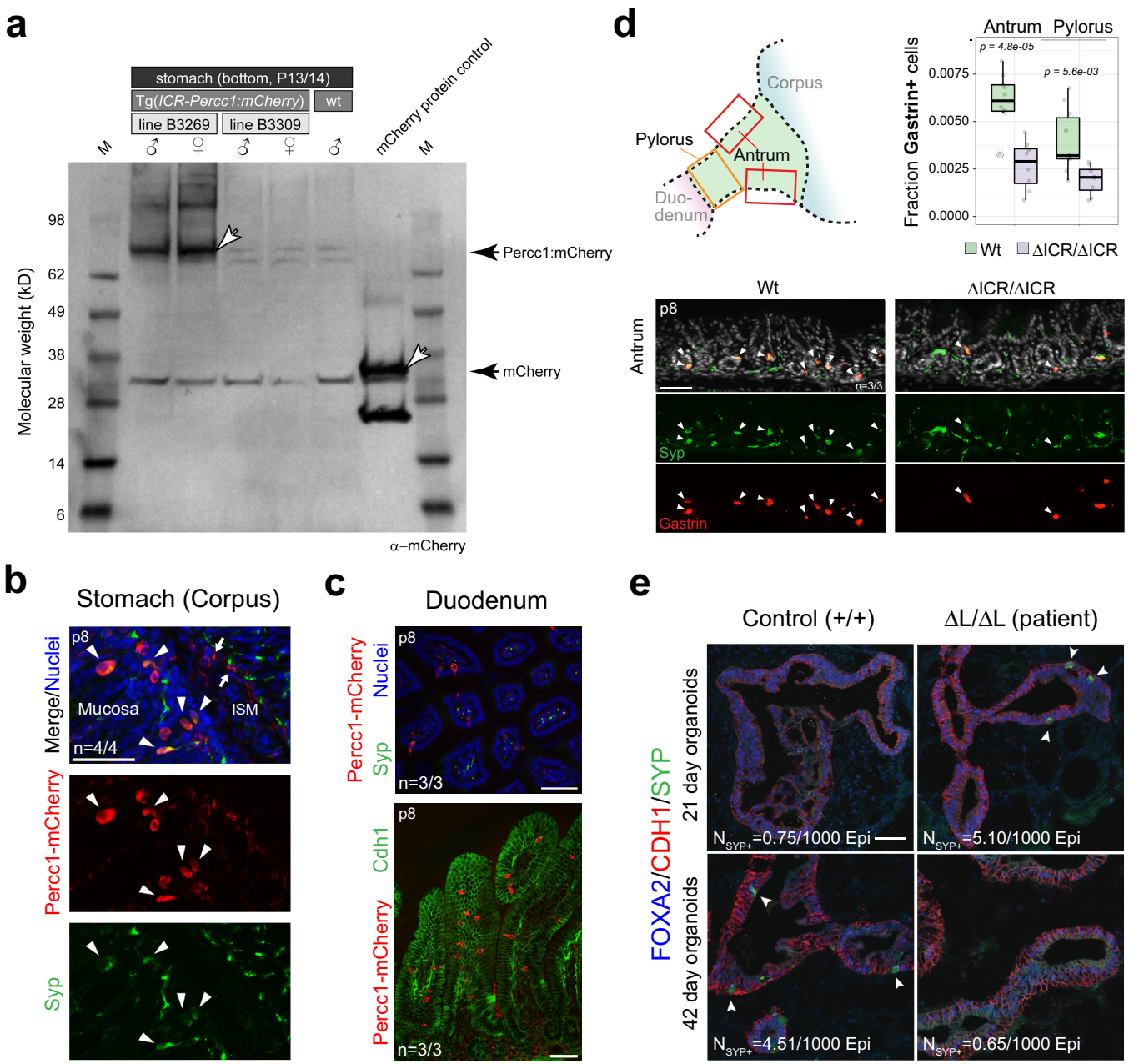
Extended Data Fig. 4 | Gastrointestinal and microbiome analysis in chr17$^{\Delta ICR/\Delta ICR}$ mice. a, Modified intestinal content in wild-type mice (left) compared to chr17$^{\Delta ICR/\Delta ICR}$ mice (right; $n = 45$) at P10. **b**–**d**, ICR deletion causes changes in intestinal and faecal microbiome composition. Microbial communities in different intestinal compartments and faeces were analysed by 16S rRNA-based sequence profiling. **b**, Family-level relative abundance profiles of the top 15 most abundant prokaryotic families for wild-type ($n = 22$) and chr17$^{\Delta ICR/\Delta ICR}$ ($n = 21$) intestinal and faecal samples, organized by sample type. The most pronounced changes were observed in colon and faecal samples. **c**, Heat map of log-transformed read counts for those genera that exhibited the greatest variance (top 60%) across all faecal samples. The abundance profiles exhibit perfect clustering of the faecal samples (rows) into wild-type ($n = 6$) and chr17$^{\Delta ICR/\Delta ICR}$ ($n = 7$) groups. **d**, Bar charts of Shannon's diversity for all faecal samples from **b**, grouped into wild-type and chr17$^{\Delta ICR/\Delta ICR}$ samples.



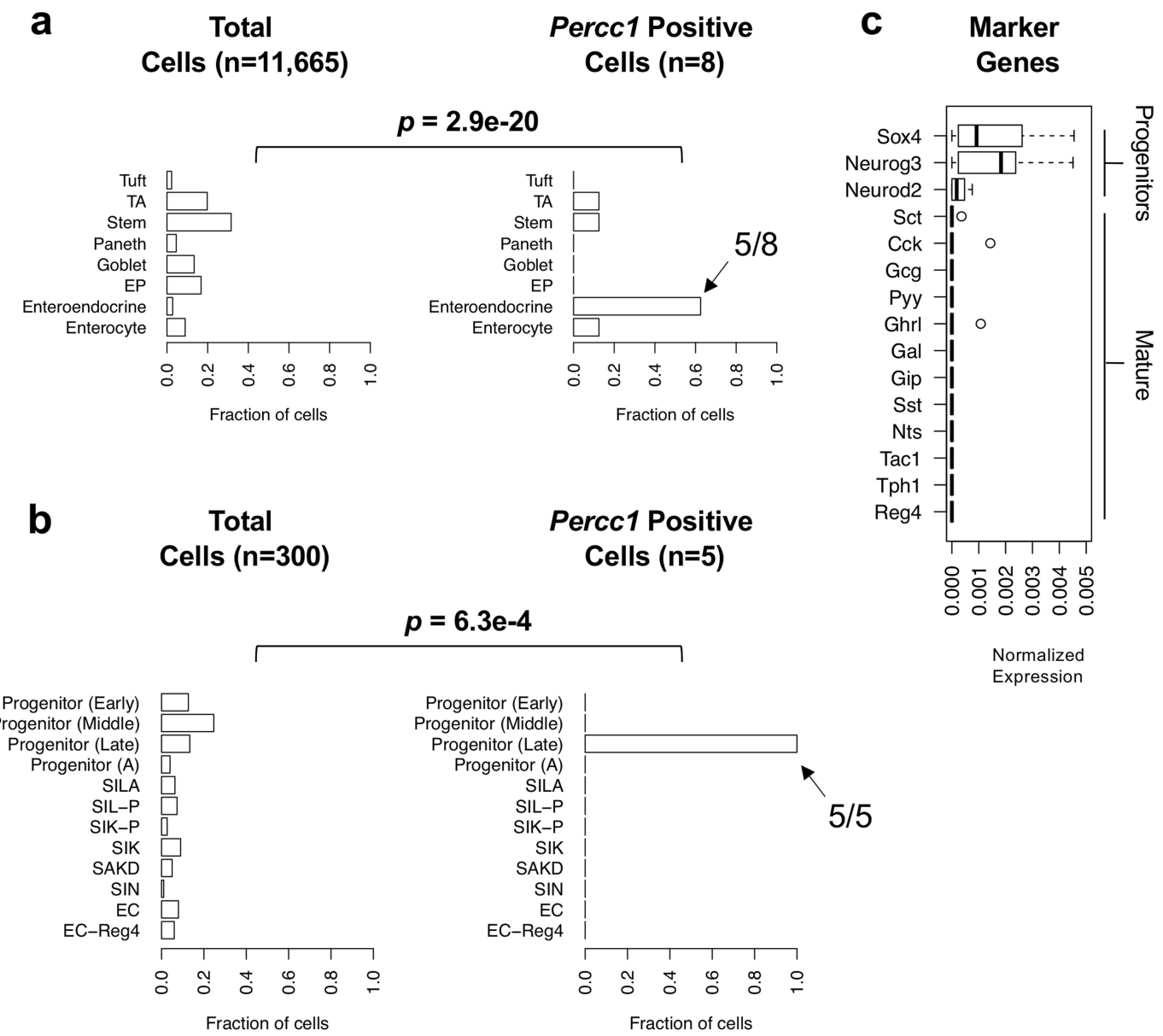
Extended Data Fig. 5 | Gastrointestinal X-gal staining of ICR-reporter transgenic embryos compared to *Percc1* mRNA in situ hybridization. a, **b**, Cross-sections of E14.5 mouse tissues with a β-galactosidase ICR-driven transgene. **c**, **d**, *Percc1* mRNA in situ hybridization analysis on E14.5 wild-type sections. For X-gal staining and in situ hybridization experiments, two embryos for each experiment and each condition were collected at E14.5 and a minimum of three sections from each embryo were examined. Representative sections are shown.



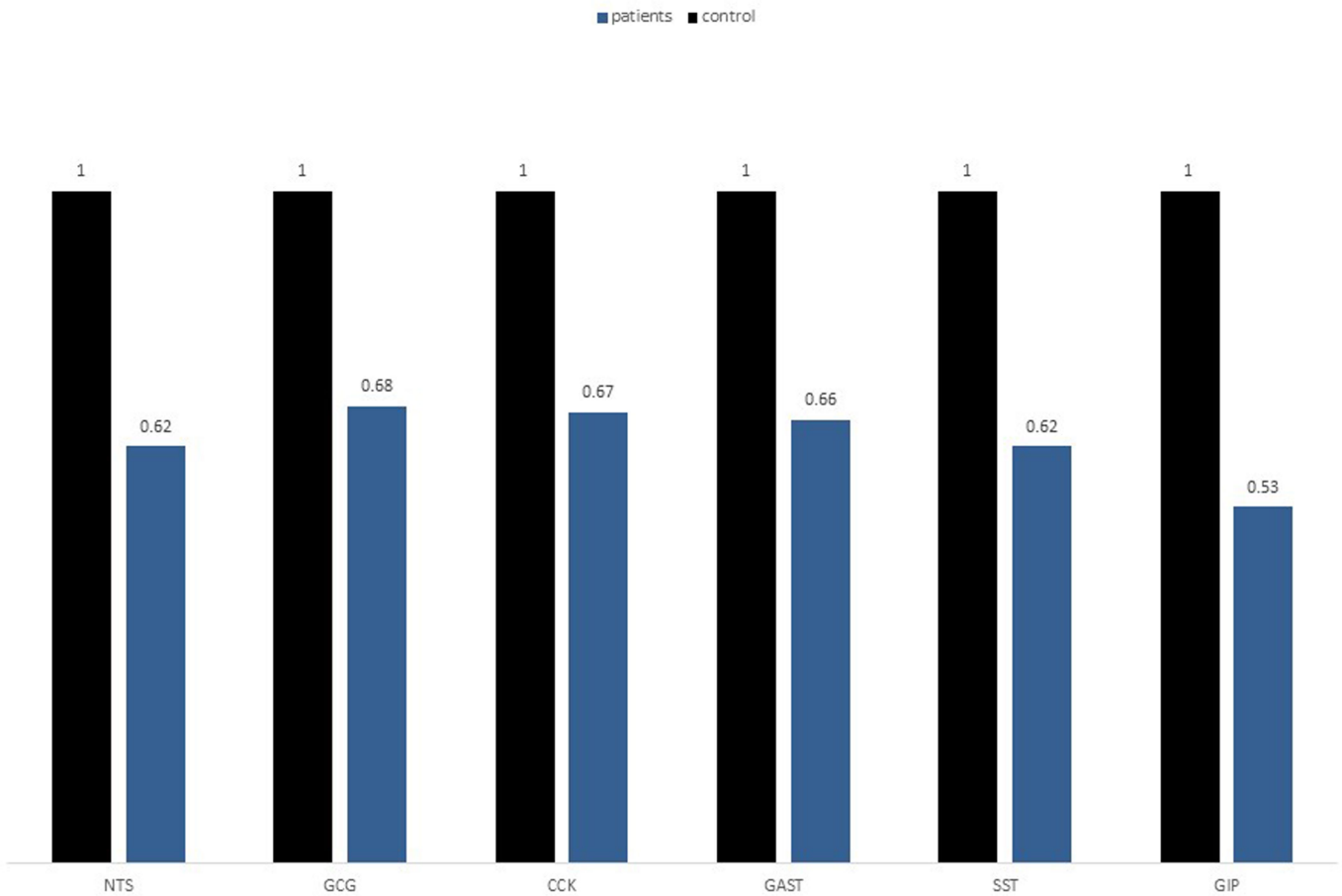
Extended Data Fig. 6 | Analysis of body weight in *Percc1*-knockout and transgenic mice. a, Comparison of weight in *Percc1*-knockout mice ($n = 38$; red) and littermate controls ($n = 25$; blue), showing that *Percc1*-knockout mice have reduced body weight. *Percc1*-knockout mice were generated in an FVB/N genetic background. **b**, *Percc1* transgenic rescue of the body weight phenotype that is found in chr17$^{\Delta ICR/\Delta ICR}$ mice. An 8.5-kb *Percc1* mini gene was constructed (Supplementary Table 10) and used to generate a *Percc1* mouse line that overexpressed PERCC1. When this transgene was introduced into the chr17$^{\Delta ICR/\Delta ICR}$ mouse genetic background, we observed the rescue of all the phenotypes (including the severe reduction in body weight) that were found in chr17$^{\Delta ICR/\Delta ICR}$ mice. Chr17$^{\Delta ICR/\Delta ICR}$ mice were generated in a mixed 129/C57Bl6 background. *P* values were determined using a two-tailed *t*-test; n.s. indicates a *P* value of 0.8–1.0. Lines show the mean and shaded areas represent ±1 s.d.



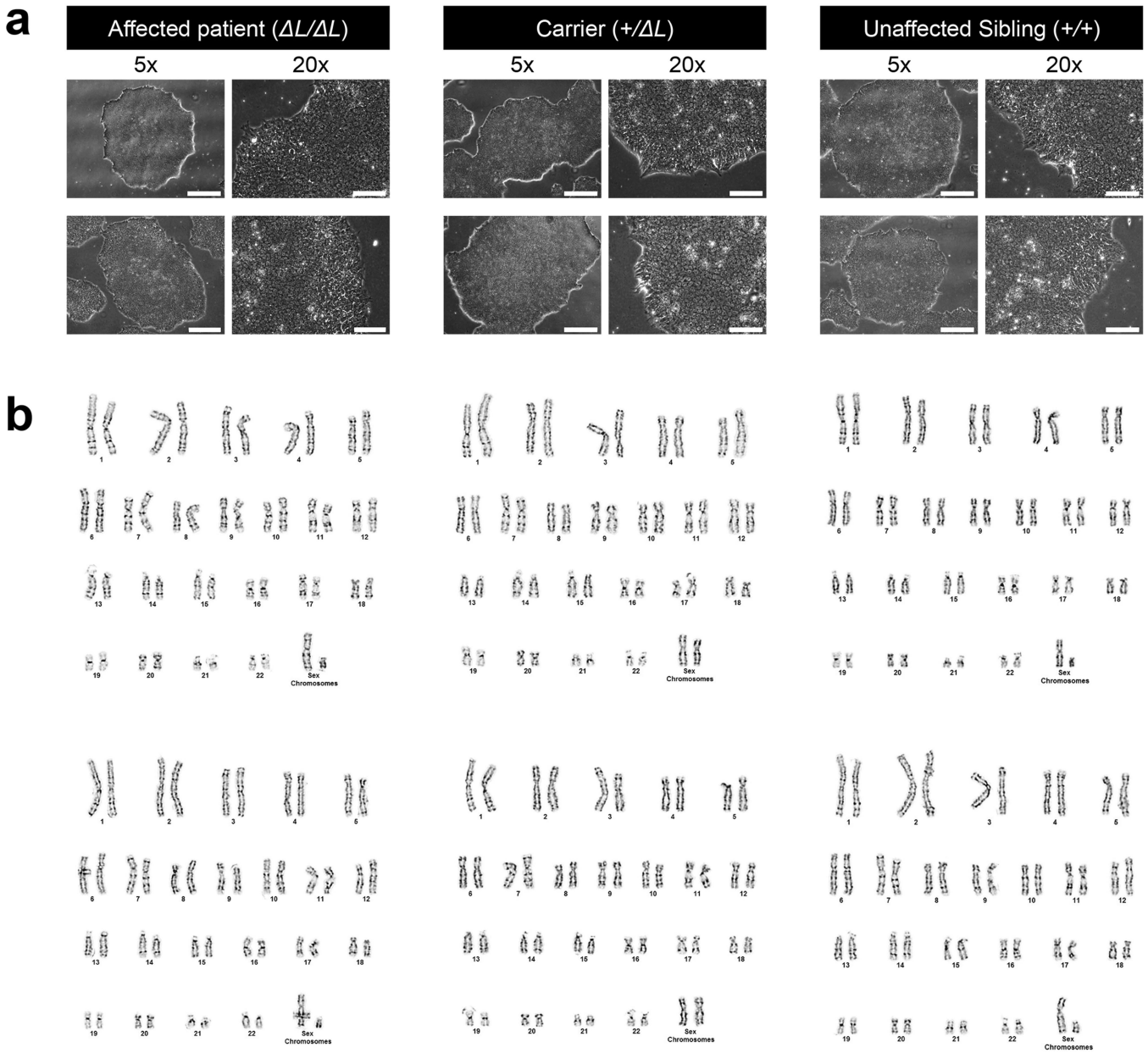
Extended Data Fig. 7 | Characterization of PERCC1 in mice and patients. a, Western blot analysis of PERCC1–mCherry fusion protein. Two stable transgenic lines (B3269 and B3309) were established through standard pronuclear microinjection of fertilized mouse eggs. Protein extracts from juvenile mice (P13–P14) were separated by SDS–PAGE and transferred for western hybridization. Lanes: 1, molecular mass marker (M); 2 and 3, line B3269; 4 and 5, line B3309; 6, wild-type control; 7, mCherry positive control; 8, molecular mass marker. mCherry is predicted to be 28.8 kDa and the PERCC1–mCherry fusion protein is predicted to be 59 kDa, with both proteins running about 5 kDa larger. Line B3309 does not express the fusion protein, in contrast to line B3269 (probably owing to a position effect). These experiments were performed four times. **b**–**e**, Identification of cells with PERCC1$^+$ identity, and the effect of PERCC1 ablation in gastrointestinal tissues. **b**, A subpopulation of PERCC1$^+$ cells (red) in the corpus epithelium (mucosa) expresses SYP (green) at P8. Arrowheads mark double-positive cells. Arrows mark a minor fraction of PERCC1$^+$ cells that were detected in longitudinal smooth muscle (lSM). DAPI-stained nuclei are shown in blue. **c**, Dispersed PERCC1$^+$ cells (red) are observed in the villi of the duodenum at P8. Top, cross-section through villi illustrates the absence of endocrine identity (green) in these cells. Bottom, sagittal section showing the distribution of PERCC1$^+$ cells in the epithelium of villi (CDH1; green). **d**, Top left, schematic depicting the anatomical compartments of the distal stomach and the location of sections used for cell counting. Top right, reduction of the fraction of G cells observed predominantly in the pyloric antrum of *Percc1*-deficient chr17$^{\Delta ICR/\Delta ICR}$) mice at P8. Box plots indicate median (centre line), interquartile values (box limits), range (whiskers), outliers (circled dots) and individual biological replicates (dots). *P* values were determined using an unpaired two-tailed *t*-test. Bottom, comparative immunofluorescence analysis illustrating the reduced number of gastrin-expressing cells (red) in the absence of *Percc1* (chr17$^{\Delta ICR/\Delta ICR}$) in the antrum at P8. SYP-expressing endocrine cells are green and nuclei are grey. **e**, Immunofluorescence from HIOs derived from control (ICR$^{+/+}$) and patient (ICR$^{\Delta L/\Delta L}$) iPS cell lines. Detection of anti-FOXA2 (blue) and anti-CDH1 (red) was used to visualize the HIO epithelium, and EECs were localized at 21 and 42 days on the basis of SYP expression (green) and counted. The average number of SYP$^+$ cells (N_{SYP+}) per 1,000 epithelial (Epi) cells from cell counts in $n = 2$ technical replicates from independent HIO preparations is indicated ($P = 1.75 \times 10^{-18}$ for reduced number of SYP$^+$ cells in ICR$^{\Delta L/\Delta L}$ HIOs; Fisher's exact test). *n* represents independent biological replicates with similar results. Scale bars, 50 μm.



Extended Data Fig. 8 | *Percc1* analysis of single-cell transcriptomes from mouse intestine. a, Left, bar chart showing the fraction of the total cells profiled in a previous study[11] ($n = 11{,}665$) that was assigned to each one of the major cell types identified. Right, the same information, but limited to those cells that express *Percc1* ($n = 8$). **b**, Same as **a** but limited to EECs. *P* values were calculated using a chi-squared test, using data from the corresponding left panel as reference (**a**, **b**). EP, epithelial, TA, transit-amplifying. **c**, Box plots showing the distributions of the normalized gene-expression values for known EEC-associated transcription factors and hormones in the eight *Percc1*-positive cells from **a**. Box plots indicate median (centre line), interquartile range (IQR; box limits) and 1.5 × IQR (whiskers).



Extended Data Fig. 9 | Validation of human RNA-seq data by RT–qPCR in duodenal tissue from two different patients and control tissue. Pairwise comparison of the relative gene-expression levels of six peptide hormones (cholecystokinin (*CCK*), gastrin (*GAST*), glucagon (*GCG*), gastric inhibitory polypeptide (*GIP*), neurotensin (*NTS*) and somatostatin (*SST*)) in duodenal tissue from patients and normal duodenal tissue (control; represented as 1). Relative expression levels for patients represent the average between two patients (patients 1.1 and 5.1).



Extended Data Fig. 10 | Characterization of HOIs and iPS cell lines. **a**, HIOs generated from an affected patient, a carrier and an unaffected sibling all show normal morphology. Differentiation into HIOs was performed in duplicate with qPCR and histological analyses that yielded similar results. **b**, iPS cell lines from an affected patient, a carrier and an unaffected sibling display a normal karyotype. This was a single experiment for each sample, as an assessment of quality control.

nature research

Corresponding author(s): Len Pennacchio

Last updated by author(s): 3/25/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software used.
Data analysis	We have provided all software versions in the main text and they are provided here. Tophat v2.0.4, Cuffdiff v2.0.2, R v.3.4.2, SRA Toolkit v2.8.1, SAMtools v1.3.1 (mouse single cell analysis), SAMtools v0.1.7a (exome analysis), Burrows-Wheeler Alignment (BWA), Sequence Variant Analyzer software (SVA) v1.10, Estimation by read depth with single-nucleotide variants (ERDS), Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi/ from the Whitehead Institute, Massachusetts Institute of Technology, and Cambridge, MA), PLINK v1.07, and DAVID v6.7.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The cDNA and predicted protein sequence will be available upon publication in Genbank record KY964488. The RNA-seq data will be available upon publication with GEO number GSE94245.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Methods, section "Experimental Design." Specific information related to the different experimental approaches can be obtained in the individual methods sections. For samples from human patients no power calculation was performed. Only 7 pedigrees are known to exist for this condition and they were included in this study.
Data exclusions	Methods, section "Experimental Design." Specific information related to the different experimental approaches can be obtained in the individual methods sections.
Replication	Attempts at replication were successful and statistical parameters and reproducibility numbers are indicated in figure panels/text as required.
Randomization	Methods, section "Experimental Design."
Blinding	Methods, section "Experimental Design."

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	All antibodies used are described in the methods section. Primary antibodies used: rabbit polyclonal IgG anti-mCherry (ThermoFisher Scientific, cat# PA5-34974); guinea pig polyclonal anti-Synaptophysin (Synaptic Systems, cat# 101004), goat polyclonal IgG anti-Gastrin (C-20, Santa Cruz, cat# sc-7783); mouse monoclonal IgG2a anti-E-Cadherin (BD Biosciences, cat# 610181); rat monoclonal IgG2a anti-Endomucin (Thermo Fisher/eBiosciences cat# 14-5851-82); chicken polyclonal IgY anti-b-galactosidase (Abcam cat# ab9361); goat polyclonal IgG anti-E-Cadherin (R&D Systems cat# AF648); mouse monoclonal IgG2a anti-FOXA2 (Novus/Abnova cat# H00003170-M01) Secondary antibodies: HRP-conjugated goat anti-rabbit IgG (1:3000, ThermoFisher cat# 31460); donkey anti-rabbit IgG Alexa Fluor 568 (Thermo Fisher cat# A-10042); goat anti-guinea pig IgG Alexa Fluor 488 (Thermo Fisher cat# A-11073); donkey anti-goat Alexa Fluor 647 (Thermo Fisher IgG cat# A-21447); goat anti-mouse IgG Alexa Fluor 488 (Thermo Fisher cat# A-11001); goat anti-mouse IgG Alexa Fluor 594 (Thermo Fisher cat# A-11032); goat anti-rabbit IgG Alexa Fluor 594 (Thermo Fisher cat# A-11012); goat anti-rat IgG Alexa Fluor 488 (Abcam cat# ab150157); goat anti-chicken IgY Alexa Fluor 594 (Thermo Fisher;cat# A-11042);
Validation	All antibodies are commercially available and validated by the vendors. Additional details about primary antibodies can be found here: 1) mCherry Antibody (ThermoFisher Scientific, cat# PA5-34974). Lot# QL2132058 The PA5-34974 immunogen is recombinant fragment contains a sequence corresponding to a region within amino acids 1 and 236 of mCherry. PA5-34974 targets mCherry in ICC, IF, IP, and WB applications and shows reactivity with Human samples. https://www.thermofisher.com/antibody/product/mCherry-Antibody-Polyclonal/PA5-34974 2) Synaptophysin (Synaptic Systems, cat# 101004).

Immunogen is synthetic peptide corresponding to AA 301 to 313 from human Synaptophysin1 (UniProt Id: P08247). Reacts with: human (P08247), rat (P07825), mouse (Q62277), hamster, cow, chicken, frog.
https://www.sysy.com/products/s-physin1/facts-101004.php

3) Gastrin (C-20, Santa Cruz Biotechnology, cat# sc-7783). Lot# F2416
Epitope mapped near the C-terminus of Gastrin of human origin. Recommended for detection of Gastrin 71, Gastrin 17, Gastrin 34 and, to a lesser extent, Cholecystokinin of mouse, rat and human origin by WB, IP, IF and ELISA.
https://www.scbt.com/scbt/product/gastrin-antibody-c-20?requestFrom=search

4) anti-E-Cadherin (BD Biosciences, cat# 610181). Lot# 6168628
Immunogen is human E-Cadherin C-terminal Recombinant Protein. Reacts with: human (QC Testing), mouse, rat, dog (Tested in Development).
http://www.bdbiosciences.com/us/applications/research/stem-cell-research/cancer-research/human/purified-mouse-anti-e-cadherin-36e-cadherin/p/610181

5) anti-Endomucin (Thermo Fisher/eBiosciences cat# 14-5851-82)
The eBioV.7C7 monoclonal antibody reacts with mouse endomucin, which was identified in a search for cell-surface expressed endothelial cell markers. Species reactivity: Mouse. Published applications: IHC, IF, Flow cyt
https://www.thermofisher.com/antibody/product/Endomucin-Antibody-clone-eBioV-7C7-V-7C7-Monoclonal/14-5851-82

6) anti-b-galactosidase (Abcam cat# ab9361)
Immunogen: Full length native protein (purified). The immunogen was purified beta-galactosidase from Escherichia coli. Tested applications: IHC, IF, Flow Cyt, ELISA, WB
https://www.abcam.com/beta-galactosidase-antibody-ab9361.html

7) anti-E-Cadherin (R&D Systems cat# AF648)
Immunogen: Mouse myeloma cell line NS0-derived recombinant human E-Cadherin Asp155-Ile707. Species Reactivity: Human, Mouse. Applications: IHC, WB, Flow Cyt
https://www.rndsystems.com/products/human-mouse-e-cadherin-antibody_af648

8) anti-FOXA2 (Novus/Abnova cat# H00003170-M01)
Immunogen: FOXA2 (NP_068556 363 a.a. - 457 a.a.) partial recombinant protein with GST tag. MW of the GST tag alone is 26 KDa. Species reactivity: Human. Antibody reactivity against cell lysate and recombinant protein for WB. It has also been used for IF, RNAi Validation and ELISA.
https://www.novusbio.com/products/hnf-3-beta-foxa2-antibody-9e12_h00003170-m01

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Strains of mice have been reported.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Clinical details of the subjects are provided in Supp. Table S4 and also in the Supplemental Text which reads "We studied eight patients from seven different families of Jewish Iraqi origin with an apparent autosomal recessive pattern of malabsorptive diarrhea, originally defined as having IDIS8. Identity By Descent (IBD) analysis confirmed the family relations and indicated that the closest cross-family relationship had IBD=0.040."
Recruitment	IDIS patients were recruited at Schneider and Sheba medical centers in Israel. Clinical details of the subjects are provided in Supp. Table S4.
Ethics oversight	Written informed consent to participate in the study was obtained from all individuals or their parents in case of minors. The study was approved by the Institutional Review board of the Sheba Medical Center (Tel Aviv , Israel). All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants and/or their legal guardian involved in the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.