Supplemental Material

Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape

J. Gray Camp, Christopher L. Frank, Colin R. Lickwar, Harendra Guturu, Tomas Rube, Aaron M. Wenger, Jenny Chen, Gill Bejerano, Gregory E. Crawford, and John F. Rawls

Supplemental Figure Summary

FIGURE S1: RNA-seq biological replicate data are highly correlated.

FIGURE S2: Differential ileal and colonic IEC transcriptomes reveal significant physiological differences between these distinct tissues.

FIGURE S3: DNase-seq biological replicate data are highly correlated.

FIGURE S4: IEC open chromatin exhibits hallmarks of *cis*-regulatory regions.

FIGURE S5: DNase-seq in IECs is sensitive to open chromatin at biomarker genes associated with rare and abundant cell types.

FIGURE S6: Location of differential DHS near genes differentially expressed between ileal and colonic IECs.

FIGURE S7: Differential DNase hypersensitivity in the duodenum compared to ileum and colon.

FIGURE S8: Loosening the FDR threshold identifies DHS differential between microbial states with minimal qualitative certainty.

FIGURE S9: Identification of transcription factors differential between CR and CV conditions in the ileal and colonic epithelium.

FIGURE S10: Clustering reveals mouse transcripts that are robustly differentially expressed in GF vs. colonized ileum or colon across multiple studies.

FIGURE S11: Schematic summary model.

Supplemental Table Summary with Description

TABLE S1: DNase-seq and RNA-seq sample and sequencing summary.

- This table contains metadata describing the samples (ID, age, diet, strain, sex, etc.) and summarizes the sequencing data for each sample.

TABLE S2: Differentially expressed genes and GO annotations between CR ileum and colon.

- This is a multi-tab spreadsheet providing (i) the FPKM values and differential statistics calculated from RNA-seq data for conventionally-raised (CR) ileal and colonic intestinal epithelial cells (IECs), and (ii) Gene Ontology enrichments for genes differentially expressed between ileal and colonic IECs.

TABLE S3: Genes differentially expressed in the presence and absence of microbiota.

- This is a multi-tab spreadsheet providing the FPKM values, gene names, gene coordinates (mm9), and differential statistics calculated from pairwise comparisons between conventionally-raised (CR), germ-free (GF), and conventionalized (CV) ileal or colonic intestinal epithelial cell RNA-seq data.

TABLE S4: Hierarchical clustering of genes differentially regulated in the presence or absence of microbiota.

- This is a multi-tab spreadsheet summarizing Gene Ontology enrichments, statistics, and gene names for gene clusters (Figure 1) differentially expressed in the presence or absence of microbiota based on transcriptomes from our study.

TABLE S5: Differential DHS along the length of the intestine.

- This is a multi-tab spreadsheet providing genome coordinates (mm9), signal value, genetic feature, and nearest gene for each DNase hypersensitive site (DHS) that is significantly different between segments of the conventionally-raised (CR) intestinal tract.

TABLE S6: GREAT summary for segment-specific DHS.

- This is a multi-tab spreadsheet providing functional enrichments and associated statistics calculated using GREAT (version 2.0.2) for DNase hypersensitivity sites (DHS) that are significantly different between segments of the conventionally-raised (CR) intestinal tract.

TABLE S7: DHS putatively differential in the presence and absence of microbiota.

- This is a multi-tab spreadsheet providing the genome coordinates (mm9), genetic feature, and nearest gene for DNase hypersensitive sites (DHS) that are putatively differential (at two false discovery rate (FDR) thresholds) in pairwise comparisons between conventionally-raised, conventionalized, and germ-free conditions.

TABLE S8: DHS putatively differential that are within the regulatory domain of genes that are also differentially expressed between GF and CV ileum.

- These are the genome coordinates (mm9) and nearest gene for DNase hypersensitive sites from ileal intestinal epithelial cells that are putatively differential between germ-free and conventionalized conditions and are within the regulatory domain of differentially expressed genes.

TABLE S9: Genome coordinates of DHS within the regulatory domain of genes with differential expression between microbiota colonization states.

- This is a multi-tab spreadsheet providing the genome coordinates (mm9) for all DNase hypersensitive sites that are within the regulatory domain of genes that are up- or down-regulated in the intestinal epithelium when comparing germ-free versus conventionalized transcriptomes from our study.

TABLE S10: Motif prediction in DHS near genes with differential expression between microbiota colonization states.

- This table provides transcription factor binding motif enrichment analyses using Homer v4.5 for DNase hypersensitive sites within the regulatory domain of genes that are up- or down-regulated in the intestinal epithelium when comparing germ-free versus conventionalized transcriptomes from our study.

TABLE S11: Transcripts that are consistently up- or down-regulated in the ileum or colon of colonized vs. germ-free mice and their GO enrichments.

- This is a multi-tab spreadsheet providing (i) gene symbols and (ii) Gene Ontology enrichments and statistics for transcripts that are consistently up- or down-regulated in the ileum or colon of C57BL/6 mice reared in the presence of microbiota (compared to germ-free) across multiple independent studies.

TABLE S12: Genome coordinates of DHS within the regulatory domain of genes consistently up- or down-regulated in the ileum or colon of colonized vs. germ-free mice.

- This is a multi-tab spreadsheet providing the genome coordinates (mm9) for DNase hypersensitive sites that are within the regulatory domain of genes that are consistently up- or down-regulated in the ileum or colon of colonized vs. germ-free mice.

TABLE S13: Motif prediction in DHS near genes consistently up or down-regulated in colonized versus germ-free ileum and colon.

- This is a multi-tab spreadsheet providing transcription factor binding motif enrichment analyses using Homer v4.5 for DNase hypersensitive sites within the regulatory domain of genes consistently up- or down-regulated in the ileum or colon of colonized vs. germ-free mice.

TABLE S14: Overlap of DHS near microbiota regulated genes with TF ChIP data sets.

- This is a multi-tab spreadsheet containing (i) statistics describing the overlap of DNase hypersensitive sites that are near microbiota-regulated genes with ChIP-seq data from multiple transcription factors and (ii) metadata (TF name, cell type, cell line, method, antibody) and source (reference lead author, PubMed ID) of ChIP-seq data used for the overlap analyses.

Supplemental Scripts Used for Data Analysis

SCRIPT S1: Hierarchical clustering and visualization of RNA-seq data.

- *This is the base R script used to cluster and visualize the intestinal epithelial cell (IEC) transcriptome data from multiple microbiota conditions presented in Figure 1F.*

SCRIPT S2: Overlap enrichment testing given two sets of genomic intervals.

- *This is the base shell script used for calculating the fold of the overlap count statistic given two sets of genomic intervals, which was used as the basis for Figure 6E, 7E, and S4D-F.*

Supplemental Discussion

The initial goal of our study was to use DNase-seq as a quantitative method to identify *cis*-regulatory regions used by host epithelial cells to respond to microbiota activity. By loosening the statistical thresholds that identified differential DHS between ileal and colonic IECs, we discovered sets of DHS putatively differential in the presence or absence of microbiota. However, we were not convinced that these results represent a biologically meaningful response for several reasons. First, visual inspection of these DHS were often not convincing across replicates. Second, most of the differential DHS were not linked to differential gene expression. Third, there were no robust functional enrichments of genes nearby the differential DHS. Taken together, these results suggested that an FDR threshold of 0.05 was not strict enough to filter spurious results. Importantly, these results do not exclude the possibility that other chromatin-based methods such as ChIP-seq using antibodies targeting TFs and post-translationally

modified histones may disclose dynamic chromatin changes in the intestinal epithelium in response to the commensal microbiota. In particular, the gut microbiota has a profound effect on cellular metabolite concentrations, such as butyrate, which may influence the status of histone acetylation (Donohoe et al. 2011). Consistent with this notion, a recent study showed that histone deacetylase 3 is required in IECs for maintaining microbiota-dependent intestinal homeostasis (Alenghat et al. 2013). It is also possible that averaging across the ensemble of cell types in our IEC preparations obscured any subtle microbiota-induced alterations in chromatin accessibility in a subpopulation of epithelial cells. Future work purifying these cell types will be necessary to explore this possibility. Moreover, gene expression changes induced by the microbiota in the intestinal epithelium may be regulated by post-transcriptional processes, such as RNA or protein stability.

Supplemental Methods

Mouse IEC isolation

Eight-week old mice were terminally anesthetized with 0.5 ml isofluorane in an airtight container and euthanized via cervical dislocation. Duodenum (anterior 5 centimeters of midgut), ileum (posterior 6 centimeters of midgut), and colon (6 centimeters of terminal hindgut) were harvested and placed into three separately labeled 50ml conical tubes containing ice-cold PBS. Using a dissecting scope, intestine-associated mesentery, adipocytes, and blood vessels were removed from each segment. Using scissors and starting with colon, each segment was splayed, vigorously washed quickly five to ten times with 50 ml ice-cold PBS, and transferred into dissociation reagent 1 (DR1; 30 mM

EDTA, 1.5 mM DTT, 0.5x Complete protease inhibitors (Roche), in 1x PBS) for 15 minutes on ice. Segments were transferred to Dissociation Reagent 2 (DR2; 30 mM EDTA, 0.5x Complete protease inhibitors (Roche), in PBS) and moderately shaken by hand for 5-20 minutes (depending on tissue) until most epithelial cells are sloughed. Note that isolation of intestinal epithelial cells were staggered at 5 minute intervals because the colonic epithelium takes approximately 15 minutes to slough, the ileal epithelium takes 10-15 minutes, and the duodenal epithelium takes 5-10 minutes. Intestinal lamina propria was removed and 8 ml of cold PBS was added to the cells on ice. Care was taken to minimize contact of cells with polypropylene pipette tips and polystyrene pipettes due to cell stickiness and loss. Cells were pelleted at 400 x G at 4°C and washed twice with 13 ml cold PBS. During each wash cells were resuspended by flicking in 1 ml PBS before adding the remaining 12 ml. After the 2nd wash, cells were resuspended in 0.5 ml cold PBS and 0.1 ml was reserved for RNA extraction.

Flow cytometry and fluorescence activated cell sorting

Isolated intestinal epithelial cells were dissociated into single cell suspensions, immunostained with EpCAM or CD31 antibodies, and analyzed by flow cytometry exactly as described (Gracz et al. 2013).

DNase hypersensitivity assay on mouse IECs

Cells were gently lysed by adding 10 ml 0.1% Igepal in Resuspension Buffer (10 mM Tris-Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) containing 1x Complete Protease Inhibitors, inverted three times, and three gentle shakes. Nuclei were pelleted at 600 x G for 10

min at 4°C. Nuclei were resuspended by flicking in 0.73 ml RSB. Nuclei aliquots (0.12 ml) were transferred to labeled 1.5 ml eppendorf tubes on ice using a wide bore pipette tip (cut with razor). Nuclei were incubated at 37°C for 30 seconds, 1 minute, 2 minutes, 4 minutes, or 8 minutes and reactions stopped by addition of 0.33 ml cold 50mM EDTA. Agarose plugs were made by pipetting 0.45 ml of 55°C 1% low-melting point agarose (in sterile 50 mM EDTA, pH 8.0; InCert, Lonza, 50121) directly to the reactions on ice and quickly distributing approximately 80 µL per plug yielding about 10 plugs per time point. Plugs were solidified at 4°C and transferred to 50 ml of LIDS Buffer (10 mM Tris-Cl, pH 8.0; 1% lauryl sulfate lithium salt (Sigma); 100 mM EDTA) and shaken for 1 hour at room temperature at 60 rpm. LIDS Buffer was changed and plugs were incubated overnight at 37°C. Plugs were washed five times with 50 ml of 50 mM EDTA for 1 hour each wash. Plugs were then stored at 4°C in 50 mM EDTA. Half of one plug at each condition was used to determine the appropriate amount of digestion to be used for constructing sequencing libraries.

DNase-seq library sequencing

DNase-seq libraries were sequenced at the Duke Sequencing and Analysis Core Resource using Illumina HiSeq 50 bp single-end sequencing. Data on the number of sequences and mapped reads are described in Table S1. Note that the DNase-seq protocol generates essentially fixed-length 20 bp insert fragments due to MmeI cleavage of DNA 20 bp from the MmeI recognition site included in the first ligated linker (Song and Crawford 2010). Sequences were aligned to the mouse genome (NCBI37/mm9) using BWA (Li and Durbin 2009) (seed length of 20 bp, allowing up to 2

mismatches, multi-reads mapping up to 4 locations). We allocated reads at random when mapping to 2-4 genomic locations equally well. When mapping to multiple locations but one location had a superior alignment score, the read was placed at the position with the best score. CR artifacts were filtered and raw aligned sequencing reads were smoothened using a kernel density estimation function called Parzen windowing (Parzen 1962) for visualization. DNase hypersensitive sites (or peaks) used for analysis were called using F-Seq v1.84 (Boyle et al. 2008).

Bioinformatic analysis of RNA-seq datasets extended

To assess the global association of differential DHS and nearby gene expression differences between ileum and colon, we linked DHS found within 2 kb upstream or downstream of a gene body (as defined by RNA-seq; only genes with detectable expression, >10 reads/transcript in at least one tissue type, were used) with that gene for high-confidence putative regulation. Less than 10 reads per transcript could result in erroneously high fold-change values that would bias subsequent analyses. This was done for all ileum-specific, all colon-specific, the top 1000 ileum-specific, and the top 1000 colon-specific DHS. The distributions of fold-change FPKM values between tissues were compared to the distribution of all 13,256 expressed genes by a two-sided Kolmogorov-Smirnov test.

Bioinformatic analysis of DNase-seq datasets extended

Sequencing depth normalization, variance fitting, and pairwise differential analyses were performed via DESeq (Anders and Huber 2010).. To identify DHS specific to one

out of the three tissue types profiled, pairs of duodenum, ileum, and colon were used to compare to the third tissue in pairwise fashion. In cases where a window showed significant variation between a tissue pair and third tissue, but also between the two tissues of the pair individually, the window was discarded as insignificant (to eliminate DH site significance driven solely by one tissue type). The overlap of DHS from duodenum, ileum, and colon was determined using the top 2% of non-genic mm9 phastCons elements downloaded from UCSC table browser. To generate the heatmap for visualization of DHS, all differential peak calls found in CR and GF ileum and colon were merged (31,745 unique sites), and sequencing-depth normalized DNase signal was summed in each region for each sample. Standard z-score normalization and hierarchical clustering were performed on the DHS for visualization.

Quantifying DHS overlap with ChIP datasets

TF binding regions from various cell types and tissues from mouse and human were manually compiled and curated from the literature (Table S14). Human sets were lifted over to the mouse genome (mm9) using a minimum match threshold of 0.25 bases remapping. Most datasets were from non-intestinal tissues and experiments or replicates using the same TF were merged into a single set keeping all peaks. Putative tissue-specific enhancer marked regions were sourced from the Mouse ENCODE Project and padded 250 bases to either side of the peak maximum (Shen et al. 2012). To quantify the enrichment for DHS overlap with putative enhancers or TF binding regions, DHS that overlap at least one region were counted and compared to the corresponding distribution in a null model where the DHS are randomly distributed

across the genome with a uniform probability. For this null model, the expectation value and the standard deviation of the overlap count can be calculated analytically. The size and significance of the enrichment were, respectively, characterized by the fold, i.e. the observed count divided by the expected count, and the z-score, i.e. the number of standard deviations between the observed and expected counts. Custom scripts used for analysis and the curated TF binding regions are available upon request.

Transcription factor binding site prediction

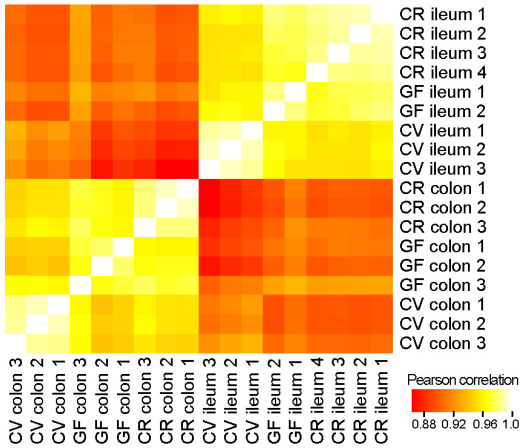
We combined gene expression data sets from several studies to identify mRNA transcripts that are differentially regulated in the intestines of mice reared germ free (GF) versus those colonized with a conventional mouse microbiota from birth (conventionally-raised, CR) or from a specified postnatal stage (conventionalized, CV), robust to different microarray or sequencing platforms, mouse age, length of colonization, housing locations, and microbiota composition. Referenced RNA-seq and microarray datasets included those comparing ileal or colonic IECs from CR vs. GF C57BL/6 mice at 8-12.5 weeks of age (this study), ileal IECs from CR vs. GF C57BL/6 mice at 4 weeks of age (Pott et al. 2012), colonic IECs from CR vs. GF C57BL/6 mice at 8-14 weeks of age (Donohoe et al. 2011), whole ileum or colon from CR vs. GF C57BL/6J mice at 12 weeks of age (Larsson et al. 2012), whole ileum from CV (colonized for 14 days) vs. GF NMRI mice at 9-13 weeks of age (Rawls et al. 2006), whole ileum or colon from CV (colonized for 30 days) vs. GF C57BL/6J mice at 8-10 weeks of age(Aidy et al. 2012). For all included data sets, normalized log2 microarray or RNA-seq data comparing GF and colonized conditions was z-scored. Experiments were

grouped by their tissue source (i.e. ileum or colon) and each group was filtered to include only those genes that were expressed (FPKM > 0) in IECs for the respective tissue type based on the RNA-seq data described in this study. Only genes with 1 or 0 missing values among the respective data sets were included in subsequent analysis. Ileum and colon data sets were then analyzed separately with k-means clustering (k=5, Euclidean distance, 1000 iterations; Cluster 3.0) to identify genes that were consistently up or down regulated across different experiments in either tissue. This provided 4 total gene lists that showed either a consistent large increase or decrease of mRNA levels following the colonization by microbes in either ileum or colon. We used the Animal Transcription Factor Database to filter lists of microbiota-responsive genes to identify transcription factors (i.e., DNA-binding TFs, chromatin remodeling factors, and transcription co-factors) (Zhang et al. 2011). For each gene consistently regulated by the microbiota, we assigned a basal regulatory domain of a minimum distance (5 kb) upstream and downstream (1 kb) of the TSS (regardless of other nearby genes). The gene regulatory domain was extended in both directions to the nearest gene's basal domain but no more than the maximum extension (1,000 kb) in one direction. The gene regulatory domain assignment scheme was found to out-perform several other approaches (McLean et al. 2010). Motif enrichment was performed on repeat masked, exon masked, and evolutionarily conserved DHS within the regulatory domain of genes in each list using Homer v4.5 (Heinz et al. 2010). Briefly, differential motif enrichment from a library of 245 vertebrate TF position weight matrix models was calculated by comparing DHS sequences within the regulatory domains of genes that were up-regulated by the microbiota to DHS sites of those that were down- regulated in the

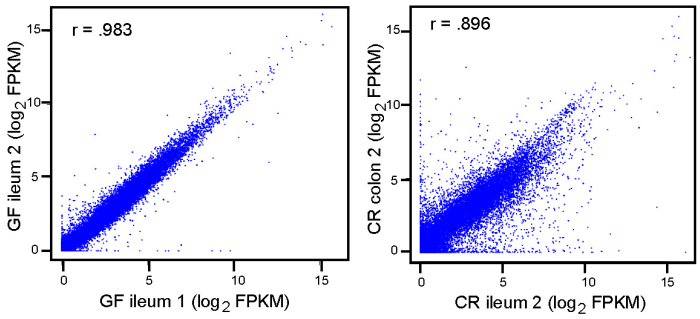
same tissue (e.g. Ileum-Up DHS vs Ileum-Down DHS). Motif enrichment was also separately calculated using a GC-matched background (10x as many sequences as the foreground) (Guturu et al. 2013). Ingenuity pathway analysis (IPA, http://www.ingenuity.com/) was used to identify upstream regulators whose predicted activation or inhibition control response to microbiota colonization. Briefly, IPA is a commercially available informatics package that queries a database of molecular interactions and functional gene annotations curated from the scientific literature and publicly available knowledge repositories. The “Upstream Regulators” analysis looks for upstream factors (TFs, signaling molecules) that have been shown experimentally to regulate genes within an input set and reports a p-value using a Fisher’s exact test and z-scores derived from multiple probability distributions (binomial, Gaussian) (Krämer et al. 2014).

SUPPLEMENTAL FIGURES

A



B



C

	CR ileum 1	CR ileum 2	CR ileum 3	CR ileum 4	GF ileum 1	GF ileum 2	CV ileum 1	CV ileum 2	CV ileum 3	CR colon 1	CR colon 2	CR colon 3	GF colon 1	GF colon 2	GF colon 3	CV colon 1	CV colon 2	CV colon 3
CR ileum 1	1.000	0.985	0.981	0.983	0.968	0.971	0.913	0.894	0.896	0.926	0.897	0.908	0.955	0.953	0.954	0.905	0.898	0.897
CR ileum 2	0.985	1.000	0.989	0.988	0.978	0.982	0.908	0.896	0.898	0.929	0.902	0.911	0.964	0.960	0.959	0.907	0.897	0.897
CR ileum 3	0.981	0.989	1.000	0.987	0.976	0.978	0.912	0.898	0.900	0.928	0.903	0.911	0.955	0.954	0.953	0.903	0.895	0.895
CR ileum 4	0.983	0.988	0.987	1.000	0.975	0.973	0.913	0.898	0.901	0.928	0.902	0.913	0.957	0.958	0.954	0.906	0.899	0.898
GF ileum 1	0.968	0.978	0.976	0.975	1.000	0.983	0.922	0.910	0.914	0.936	0.914	0.919	0.962	0.962	0.954	0.916	0.908	0.908
GF ileum 2	0.971	0.982	0.978	0.973	0.983	1.000	0.903	0.893	0.898	0.928	0.901	0.910	0.968	0.962	0.965	0.905	0.896	0.894
CV ileum 1	0.913	0.908	0.912	0.913	0.922	0.903	1.000	0.982	0.982	0.978	0.967	0.963	0.887	0.893	0.878	0.956	0.961	0.965
CV ileum 2	0.894	0.896	0.898	0.898	0.910	0.893	0.982	1.000	0.991	0.971	0.969	0.961	0.874	0.884	0.863	0.948	0.955	0.955
CV ileum 3	0.896	0.898	0.900	0.901	0.914	0.898	0.982	0.991	1.000	0.974	0.969	0.962	0.876	0.886	0.864	0.948	0.954	0.953
CR colon 1	0.926	0.929	0.928	0.928	0.936	0.928	0.978	0.971	0.974	1.000	0.968	0.973	0.912	0.914	0.902	0.963	0.966	0.966
CR colon 2	0.897	0.902	0.903	0.902	0.914	0.901	0.967	0.969	0.969	0.968	1.000	0.980	0.880	0.886	0.871	0.941	0.941	0.946
CR colon 3	0.908	0.911	0.911	0.913	0.919	0.910	0.963	0.961	0.962	0.973	0.980	1.000	0.891	0.897	0.883	0.946	0.948	0.948
GF colon 1	0.955	0.964	0.955	0.957	0.962	0.968	0.887	0.874	0.876	0.912	0.880	0.891	1.000	0.990	0.992	0.930	0.918	0.913
GF colon 2	0.953	0.960	0.954	0.958	0.962	0.962	0.893	0.884	0.886	0.914	0.886	0.897	0.990	1.000	0.987	0.936	0.927	0.920
GF colon 3	0.954	0.959	0.953	0.954	0.954	0.965	0.878	0.863	0.864	0.902	0.871	0.883	0.992	0.987	1.000	0.921	0.911	0.907
CV colon 1	0.905	0.907	0.903	0.906	0.916	0.905	0.956	0.948	0.948	0.963	0.941	0.946	0.930	0.936	0.921	1.000	0.985	0.986
CV colon 2	0.898	0.897	0.895	0.899	0.908	0.896	0.961	0.955	0.954	0.966	0.941	0.948	0.918	0.927	0.911	0.985	1.000	0.991
CV colon 3	0.897	0.897	0.895	0.898	0.908	0.894	0.965	0.955	0.953	0.966	0.946	0.948	0.913	0.920	0.907	0.986	0.991	1.000

FIGURE S1: RNA-seq biological replicate data are highly correlated. (A) Heatmap of pairwise Pearson correlation coefficients between 26,900 transcripts profiled for all RNA-seq samples in this study. (B) Scatterplots show two example correlations between individual samples. Note that biological replicates have a higher correlation than between tissue comparisons.

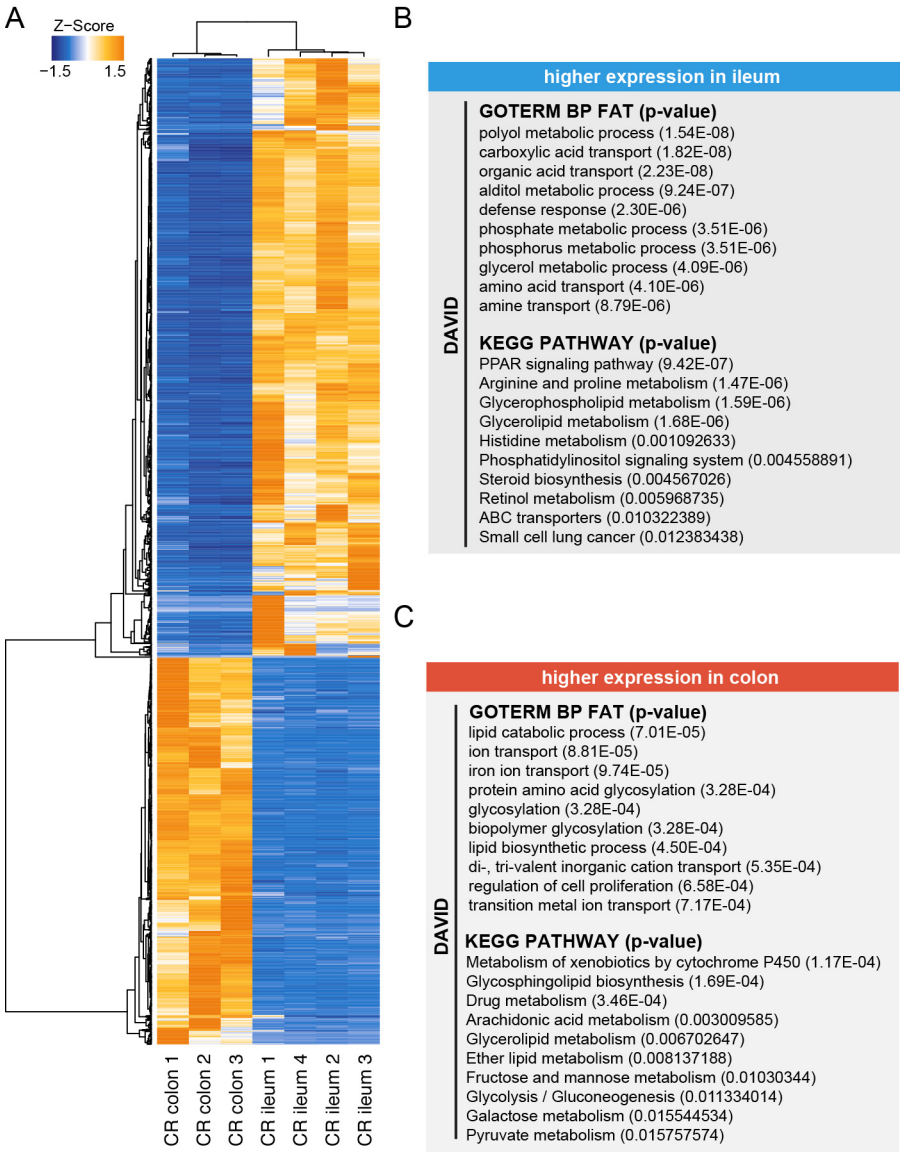


FIGURE S2: Differential ileal and colonic IEC transcriptomes reveal significant physiological differences between these distinct tissues. (A) Hierarchical clustering of FPKM values for all genes that exhibited differential expression in the pairwise comparison between CR ileal and CR colonic IEC (see Figure 3A). (B,C) The top 10 functional enrichments are shown for genes exhibiting higher expression in the ileum (B) or colon (C). See also Table S2.

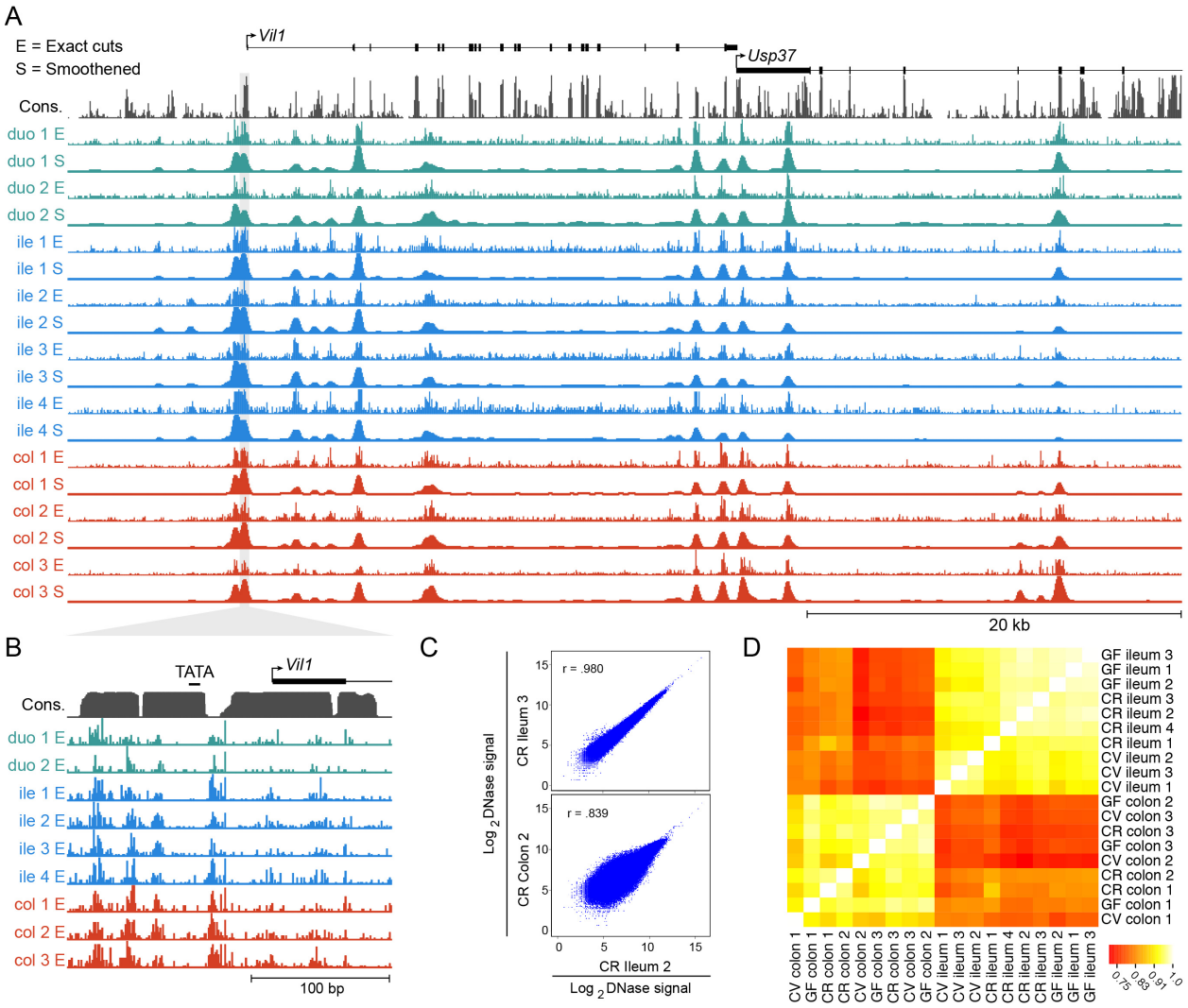


FIGURE S3: DNase-seq biological replicate data are highly correlated. (A) Screenshot of the *Villin 1* locus showing each DNase-seq biological replicate from the duodenal, ileal, and colonic epithelium from conventionally-raised (CR) mice. Shown are the tag counts and smoothened signal intensity tracts. Note the highly reproducible signature of open chromatin using endogenous DNase activity. The gray track on top displays mammalian conservation scores (phyloP). (B) High-resolution view of the exact cut track shows that the TATA box near the villin 1 transcription start site is protected from DNase cleavage indicative of a RNA Polymerase II footprint. Note the highly reproducible signature using endogenous DNase activity. (C) Scatterplots show two

example correlations between individual DNase-seq samples. Note that biological replicates have a higher correlation than between tissue comparisons. (D) Heatmap of pairwise Pearson correlation coefficients between DNase-seq tag counts in the union set of top 100,000 DHS identified in all samples.

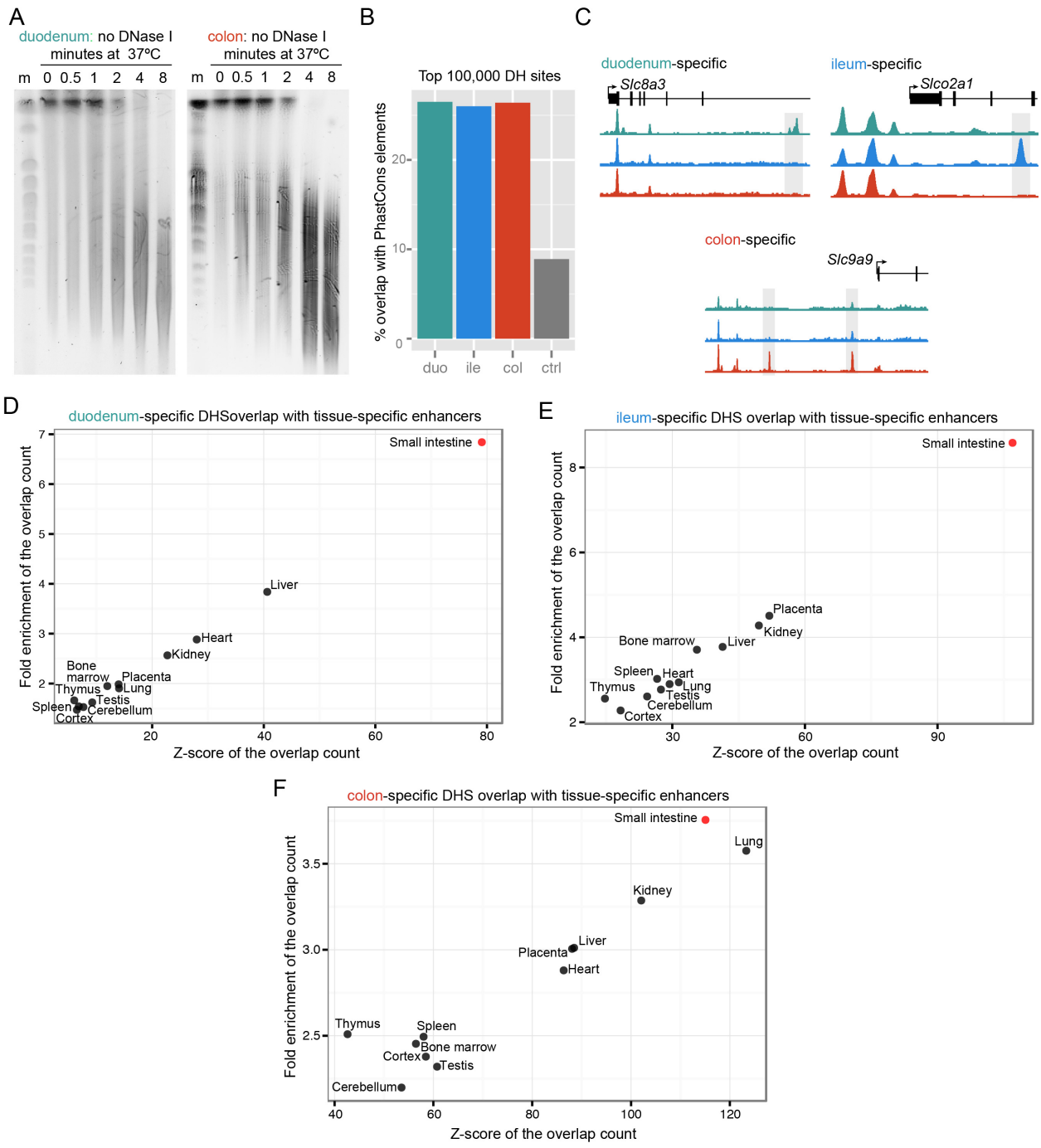


FIGURE S4: IEC open chromatin exhibits hallmarks of *cis*-regulatory regions. (A) Pulse-field gel of duodenal and colonic IECs showing that endogenous DNase activity is detected within 30 seconds after moving nuclei to 37°C and by 8 minutes most HMW

DNA is digested. (B) DHS from duodenum, ileum, and colon overlap conserved DNA (top 2% of non-genic mm9 phastCons elements) more often than control regions. (C) Representative signal track views highlighting duodenum-, ileum-, or colon-specific open chromatin sites (grey shaded areas). solute carrier 8a3, *Slc8a3*; solute carrier 2a1, *Slco2a1*; solute carrier 9a9, *Slc9a9.* (D-F) Plot showing the overlap between putative tissue specific enhancers, determined by H3K4me1 ChIP-seq (Shen et al. 2012), and segment-specific DHS from (D) duodenal, (E) ileal, and (F) colonic IECs. The fold-enrichment and z-score are calculated relative to a uniformly distributed null model.

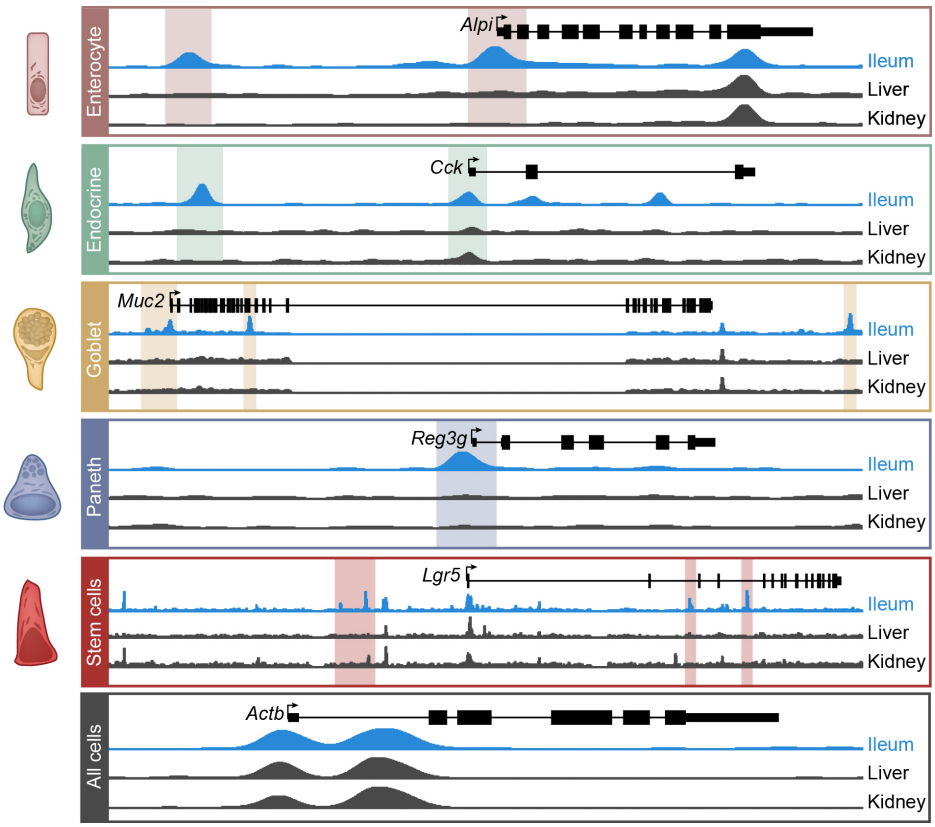


FIGURE S5: DNase-seq in IECs is sensitive to open chromatin at biomarker genes associated with rare and abundant cell types. DNase hypersensitivity at the intestinal alkaline phosphatase (*Alpi*, enterocyte marker), cholecystokinin (*Cck*, enteroendocrine cell marker), mucin 2 (*Muc2*, goblet cell marker), regenerating islet-derived protein 3 gamma (*RegIIIγ*, paneth cell marker), leucine-rich repeat-containing G protein-coupled receptor 5 (*Lgr5*, crypt stem cell marker), and b-actin (*Actb*, ubiquitous in all cells) loci in ileal IECs, liver, and kidney. Note common and distinct peaks. There are no distinct intestine-specific peaks near the *Actb* gene, which is typically expressed in all cell types. Cell-type images sourced from (Clevers and Batlle 2013). Note that the DNase signal intensity scale on the y-axis is fixed for each gene in the cross-tissue comparisons. Similar intestine-specific peaks were observed for the duodenum and colon (data not shown).

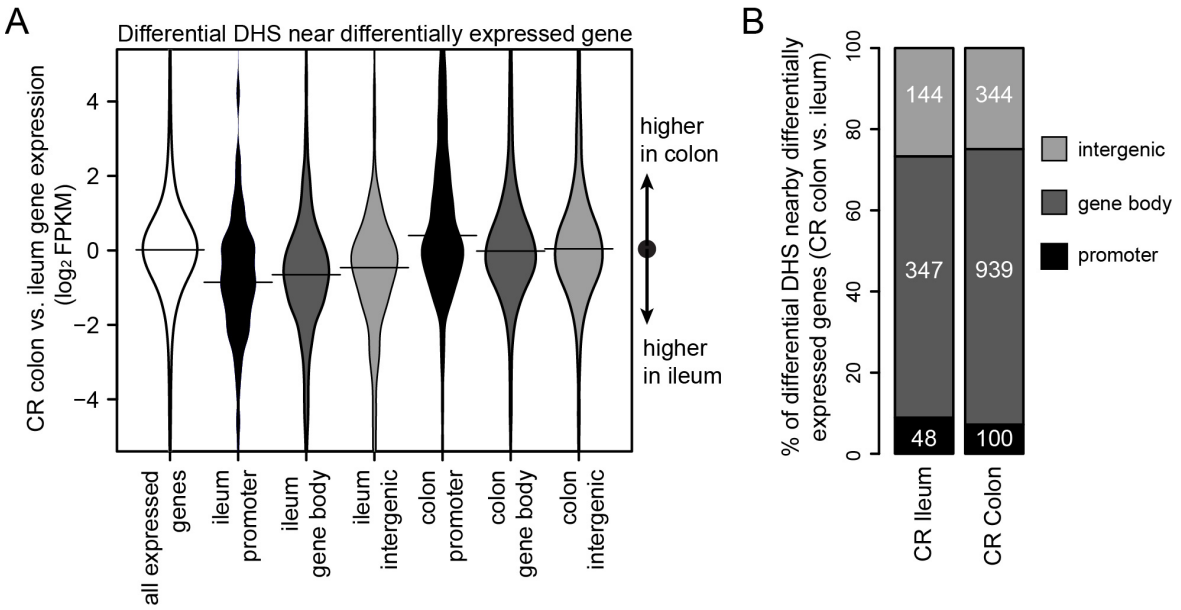


FIGURE S6: Location of differential DHS near genes differentially expressed between ileal and colonic IECs. (A) Beanplot of fold-change (colon over ileum) gene expression levels for all expressed genes (> 10 reads per transcript), and genes nearest ileum-specific and colon-specific DHS, separated by localization to proximal promoters (< 2kb from TSS), gene bodies (intronic and exonic), and intergenic regions. Note promoter-localized differential DHS best predict elevated gene expression in a tissue-specific manner. All distributions exhibited highly significant shifts from all expressed genes (Mann-Whitney U test, $p < 1x10^{-10}$), except for colon-specific gene body and intergenic located DHS. (B) Distribution of genic features for all ileum-specific and colon-specific DHS located nearest to differentially expressed genes shows that the majority are located outside proximal promoter regions.

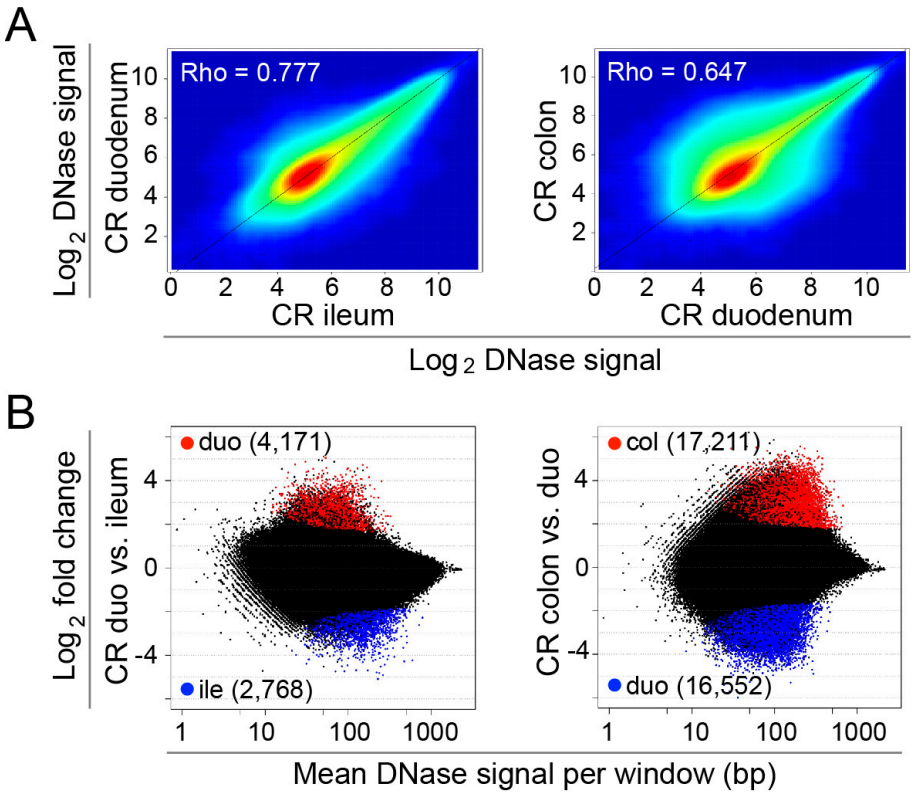


FIGURE S7: Differential DNase hypersensitivity in the duodenum compared to ileum and colon. (A) Density scatterplot showing the correlation of DNase-seq signal intensity for the union top 100,000 DHS for CR duodenum and CR ileum (left) or CR colon and CR duodenum (right). (B) The fold difference in DNase signal intensity for CR duodenum vs. CR ileum (left) and CR colon vs. CR duodenum (right) plotted against the average DNase signal observed in 250 bp windows. Significant differential windows highlighted in red and blue (FDR < .0001).

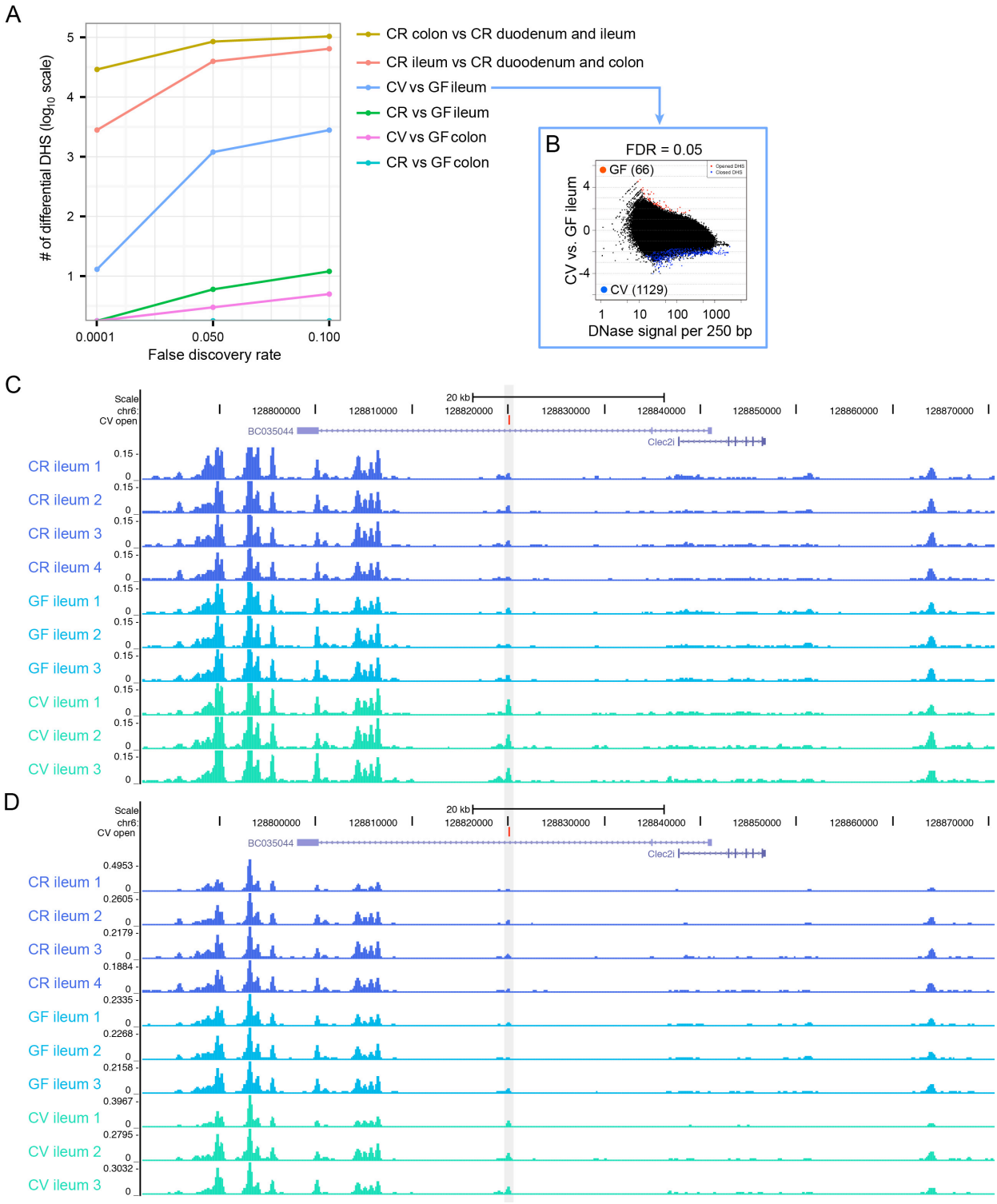


FIGURE S8: Loosening the FDR threshold identifies DHS differential between microbial states with minimal qualitative certainty. (A) Plot showing the relationship between false discovery rate (FDR) thresholds and number of DNase hypersensitive

sites (DHS) determined differential between tissue and microbial state comparisons. Loosening the thresholds to FDR < 0.05 identifies putative DHS differential between microbial states. Conventionally-raised (CR), Conventionalized (CV), Germ-free (GF). (B) Representative signal track view of a top-scoring DHS differential between CV and GF states in the ileal epithelium determined significant at FDR 0.05. Note the y-axis is on a fixed scale. (C) Same view as in panel B, but the y-axis is scaled to the data view. Notice that both the peak height and difference between CV and GF or CR is minimal.

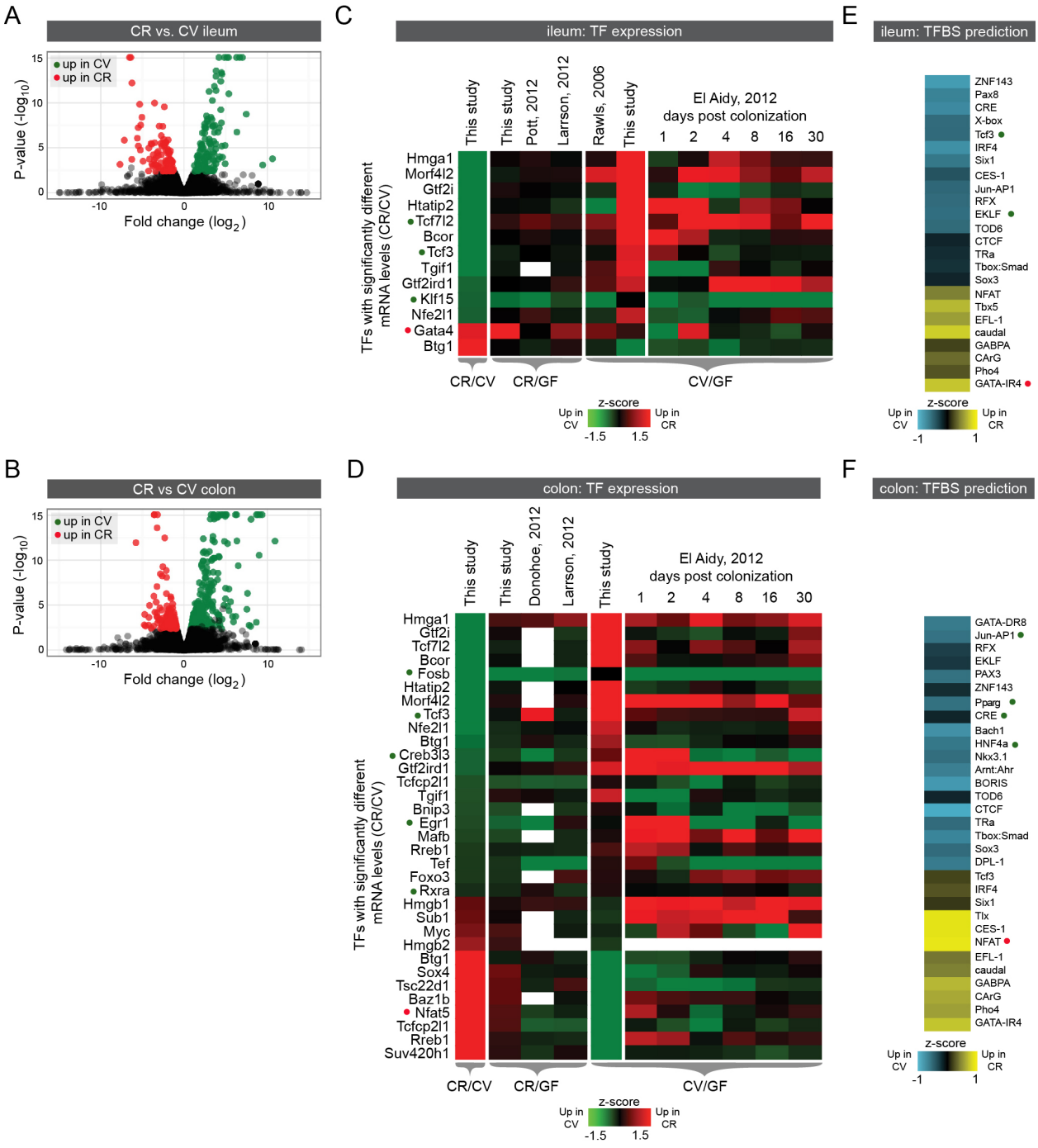


FIGURE S9: Identification of transcription factors differential between CR and CV conditions in the ileal and colonic epithelium. (A,B) Volcano plot showing pairwise comparisons of RNA expression between CR vs. CV conditions for ileal and colonic epithelium. Green and red dots represent genes up- or down-regulated in CR

respectively (FDR < 0.05). (C,D) Heatmap analysis of transcription factors (TF) differentially expressed between CR and CV conditions in our study and corresponding expression patterns in multiple other published datasets. (E,F) Transcription factor binding site (TFBS) prediction in DHS within the regulatory domain of genes in our study that were differentially regulated between CR vs. CV conditions. Red and green dots highlight TFs that were both differentially expressed and had binding sites enriched near genes higher expressed in CR or CV, respectively.

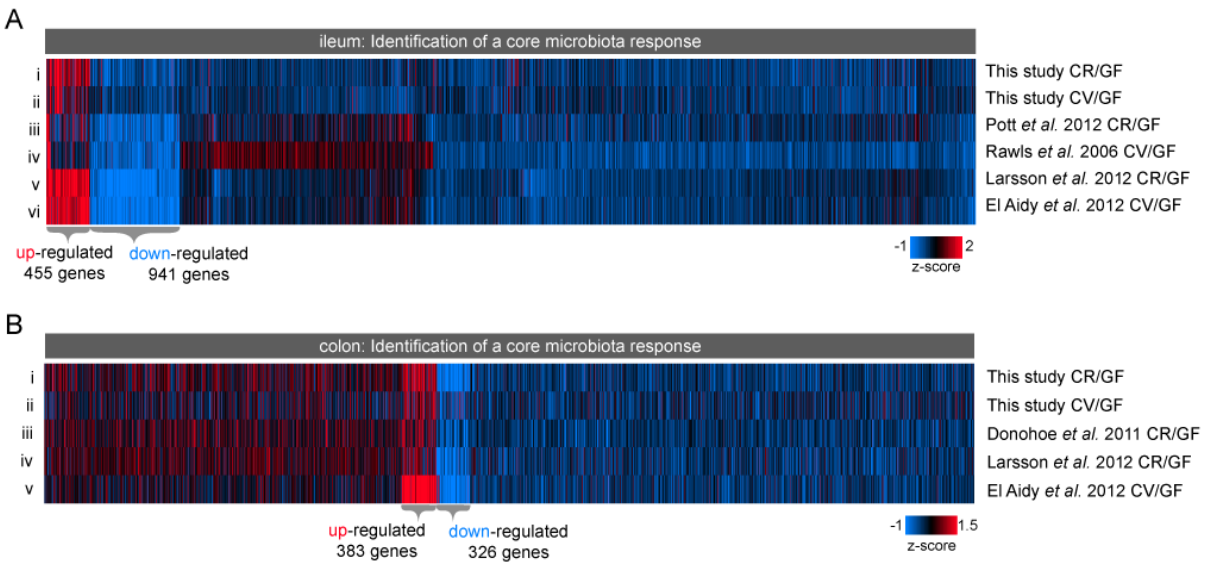


FIGURE S10: Clustering reveals mouse transcripts that are robustly differentially expressed in GF vs. colonized ileum or colon across multiple studies. (A,B) K-means clustering (K=5) of Z-scores representing differences in transcript levels between colonized and GF ileum (A) and colon (B) data sets. Distinct clusters of genes that show consistently strong differential increases (up, red) or decreases (down, blue) in mRNA levels in colonized vs. GF mice are highlighted.

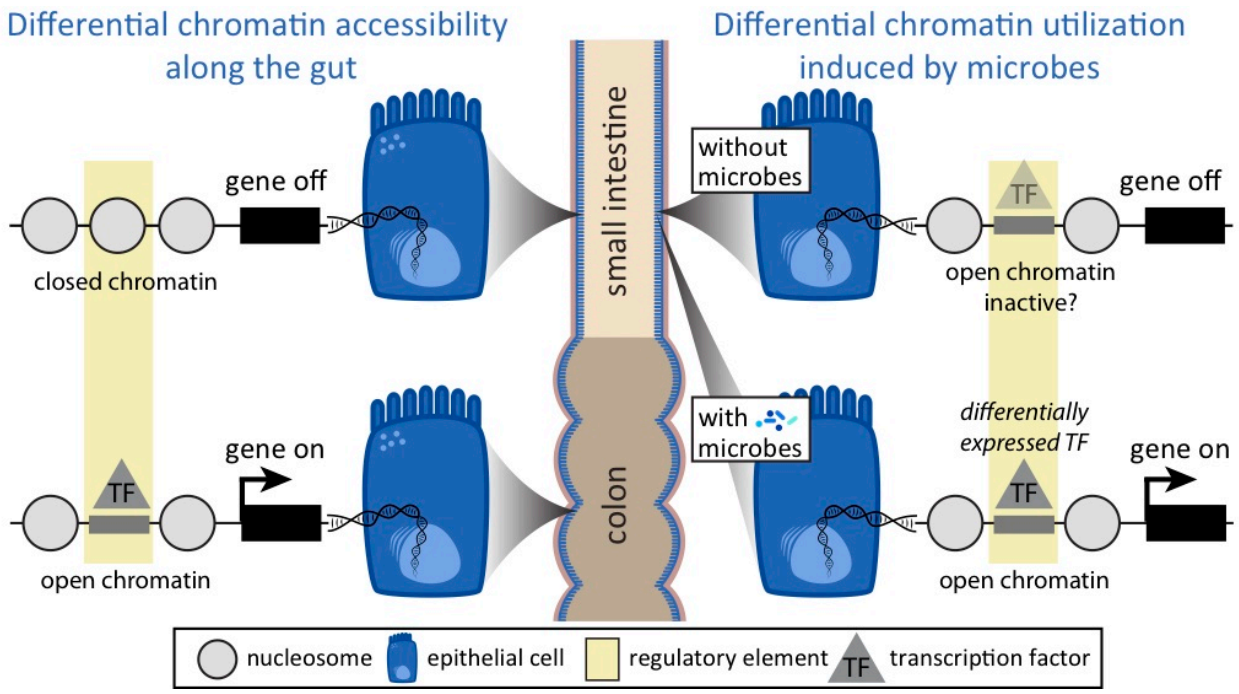


FIGURE S11: Schematic summary model. Our results support a model in which differential gene expression between epithelial cells from small intestine and colon is linked with differential chromatin accessibility (left side of graphic). In contrast, differential gene expression between epithelial cells in the same intestinal region from animals raised without or with microbes is not associated with significant alterations to chromatin accessibility, but is linked to differential utilization of accessible chromatin through differential expression of transcription factors (right side of graphic). This model suggests that the accessible chromatin landscape in intestinal epithelial cells is pre-programmed by the host in a region-specific manner, and poised to permit responses to environmental factors such as microbiota through binding of accessible regulatory elements by specific transcription factors.

REFERENCES

Aidy El S, van Baarlen P, Derrien M, Lindenbergh-Kortleve DJ, Hooiveld G, Levenez F, Doré J, Dekker J, Samsom JN, Nieuwenhuis EES, et al. 2012. Temporal and spatial interplay of microbiota and intestinal mucosa drive establishment of immune homeostasis in conventionalized mice. *Mucosal Immunol* **5**: 567–579.

Alenghat T, Osborne LC, Saenz SA, Kobuley D, Ziegler CGK, Mullican SE, Choi I, Grunberg S, Sinha R, Wynosky-Dolfi M, et al. 2013. Histone deacetylase 3 coordinates commensal-bacteria-dependent intestinal homeostasis. *Nature* **504**: 153–157.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537–2538.

Clevers H, Batlle E. 2013. SnapShot: the intestinal crypt. *Cell* **152**: 1198–1198.e2.

Donohoe DR, Garge N, Zhang X, Sun W, O'Connell TM, Bunger MK, Bultman SJ. 2011. The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metabolism* **13**: 517–526.

Gracz AD, Fuller MK, Wang F, Li L, Stelzner M, Dunn JCY, Martin MG, Magness ST. 2013. Brief report: CD24 and CD44 mark human intestinal epithelial cell populations with characteristics of active and facultative stem cells. *Stem Cells* **31**: 2024–2030.

Guturu H, Doxey AC, Wenger AM, Bejerano G. 2013. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philos Trans R Soc Lond, B, Biol Sci* **368**: 20130029.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**: 576–589.

Krämer A, Green J, Pollard J, Tugendreich S. 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**: 523–530.

Larsson E, Tremaroli V, Lee YS, Koren O, Nookaew I, Fricker A, Nielsen J, Ley RE, Bäckhed F. 2012. Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut* **61**: 1124–1131.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler

transform. *Bioinformatics* **25**: 1754–1760.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**: 495–501.

Parzen E. 1962. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33**: 1065–1076.

Pott J, Stockinger S, Torow N, Smoczek A, Lindner C, McInerney G, Bäckhed F, Baumann U, Pabst O, Bleich A, et al. 2012. Age-dependent TLR3 expression of the intestinal epithelium contributes to rotavirus susceptibility. *PLoS Pathog* **8**: e1002670.

Rawls JF, Mahowald MA, Ley RE, Gordon JI. 2006. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**: 423–433.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**: 116–120.

Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols* **2010**: pdb.prot5384.

Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H, Guo AY. 2011. AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* **40**: D144–D149.