

Regulatory DNA and stem cells in complex genetic diseases

Claartje Meddens

Beoordelingscommissie:

Prof. dr. J.M. Beekman
Prof. dr. E.P.J. Cuppen
Prof. dr. J.C. Escher
Prof. dr. N. Geijssen
Dr. S. Fuchs

Promotiecommissie:

Prof. dr. J.C. Escher
Prof. dr. J.M. Beekman
Prof. dr. E.P.J. Cuppen
Prof. dr. N. Geijssen
Prof. dr. S.B. Snapper
Prof. dr. D.P. McGovern
Dr. S. Fuchs

©2019 by Claartje Meddens, Utrecht, the Netherlands

ISBN: 978-90-393-7187-9

Cover design: Claartje Meddens

Lay-out: Marjolein Meddens

Print: Ridderprint BV, the Netherlands

The research described in this thesis was performed at the Wilhelmina Children's Hospital and the Regenerative Medicine Center Utrecht and was supported by the Alexandre Suerman Stipend (UMCU).

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged. Printing of this thesis was kindly supported by: ChipSoft, Pfizer, Formex Medical and Novuqare.

Regulatory DNA and stem cells in complex genetic diseases

Niet-coderend DNA en stamcellen in
complex genetische ziekten
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit
Utrecht op gezag van de rector magnificus, prof.dr.
H.R.B.M. Kummeling, ingevolge het besluit van het college
voor promoties in het openbaar te verdedigen op

donderdag 3 oktober 2019 des middags te 12.45 uur

door

Claartje Aleid Meddens

geboren op 24 december 1986 te Delft

Promotor:

Prof. dr. E.E.S. Nieuwenhuis

Copromotor:

Dr. M. Mokry

Table of contents

Chapter 1	Introduction	7
Chapter 2	Non-coding DNA in inflammatory bowel disease: From sequence variation in DNA regulatory elements to novel therapeutic potential	19
Chapter 3	Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium	47
Chapter 4	Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease	59
Chapter 5	Additional candidate genes for human atherosclerotic disease identified through annotation based on chromatin organization	79
Chapter 6	Chromatin conformation links distal target genes to chronic kidney disease loci	95
Chapter 7	Stem cells are the principal intestinal epithelial responders to bacterial antigens	115
Chapter 8	Discussion	135
Chapter 9	Supplementary material	145
Chapter 10	Samenvatting	169
	Dankwoord	172
	Curriculum vitae	176
	List of publications	177



Introduction

1

Developments over the past two decades have greatly impacted the execution and outcome of biomedical research. High throughput sequencing techniques provide a basis where information about the content and functioning of genomes is constantly growing and this newly gained information directly leads to the development of new applications of nucleotide sequencing which in turn feeds the continuous unravelling of novel molecular paradigms in cell biology. The development of *ex vivo* model systems provides us with the possibility to study each tissue in each individual human being under any condition we can think of. The combination of rapidly increasing possibilities to study and manipulate nucleotide composition in the plethora of available model systems lays the ground for a new era in biomedical research.

As a scientist and medical doctor, I grew up in this world full of challenges, mysteries and opportunities. An exciting world that is fun to be part of and demands us to critically follow what happens around us, find clever ways to connect and intersect the large amount of available data and carefully choose the research paths we want to pursue. This thesis is the result of my fascination with complex genetic diseases. Diseases in which genetic variants play an important role in the susceptibility and pathogenesis of the disease. Diseases in which the growing knowledge of the complex genetic basis, from time to time, confuses our understanding of the pathogenesis.

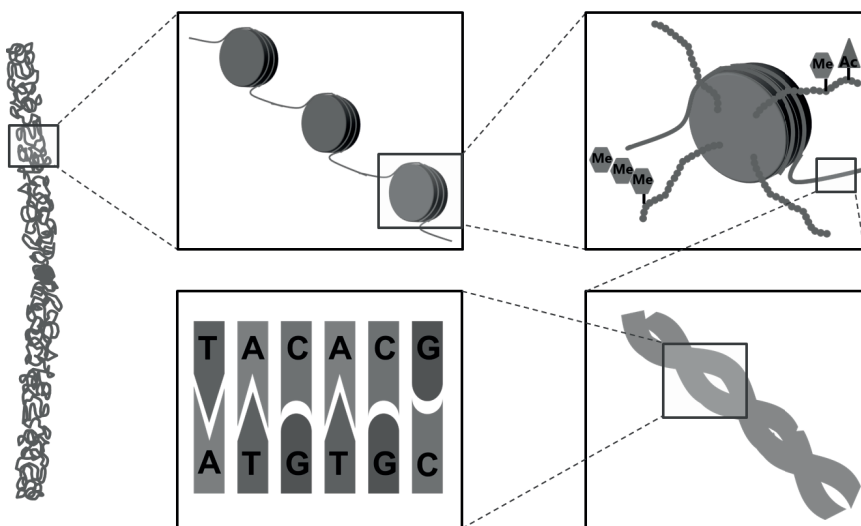
The human genome

DNA is a large polymer that is made of four nucleotides: guanine, cytosine, thymidine and adenosine.¹ The sequence in which these nucleotides are chained together defines the genetic code that is inherited from cell to daughter-cell and from parent to sibling. Each polymer is bound to another polymer with a complementary sequence that together form a double helix.² These double stranded DNA molecules are wrapped around protein complexes called histones and together form a chromatin structure.³ All genetic information of the human genome is arranged and stored into 46 chromosomes (23 from each parent) that are made up from a condensed organization of chromatin. This way, over 12 billion nucleotides that make-up a diploid human genome, are packed into the nucleus of every cell.⁴

The classical function of a genome is to carry genes. Genes can be divided into protein-coding and non-protein coding genes. Protein-coding genes are sequences that encode for mRNA molecules that subsequently determine the amino acid sequence of a protein. These genes contain intronic sequences that are involved in creating a variety of proteins from a single gene (i.e. splice-variants).⁵ In contrast, the functional molecule that is synthesized from a non-protein coding gene is the RNA molecule itself. These RNA molecules are involved in a plethora of processes, many of which influence post-transcriptional regulation of expression.⁶

The human genome harbors a little more than 22,000 genes, that together make up approximately 2% of the total sequence.⁴ Whereas in the 20th century the non-coding DNA was considered to be ‘junk’, nowadays biochemical functions are assigned to more than 80%.⁷ Since the first publication of the sequence of the human genome^{8,9} many types of DNA regulatory elements (DRE) elements have been identified. DRE are elements that are involved in transcription regulation. Transcription of a gene is dependent on the presence of a promoter element that is located directly upstream to the gene body. Promoters have multiple transcription factor binding sites (TFBSs) and provide a platform where the transcriptional machinery can be assembled. The expression level of a gene can be modified through enhancer elements. These elements are found distal to a gene and can be located up to 1 Mbp up- or downstream.¹⁰ When enhancers are active they bind transcription factors and form a DNA-loop through which they physically interact with the gene they regulate. This interaction results in the upregulation of transcriptional activity and therefore in higher expression of the regulated gene.

The activity of genomic elements does not fully rely on the DNA sequence. Transcriptional activity is a dynamic process that varies between cell types, cell states and developmental stages. Histone complexes can carry a variety of post-translational modifications that influence the accessibility of the DNA that is wrapped around the histones and influence the affinity for transcription factors and polymerases.¹¹ The distribution and deposition of histone modifications provides a cell type specific layer of expression regulation, especially the modifications that are found at enhancer elements. Another level of regulation is provided by the 3D organization of chromatin into topologically associating domains, that causes parts of chromosomes to be physically close to each other inside the nucleus.¹²



Genetic variation

Many genes and non-coding elements are conserved between different species. However, genetic variation that evolved throughout evolution resulted in different phenotypes and eventually the branching of species. Although we share a common ancestor, genetic variation within the humans species is ubiquitously present and is scattered throughout the 3 billion basepairs, thereby providing each individual with a unique genome. Estimations by the 1000 genomes project suggest that, on average, a human genome differs from the reference genome at 20 million basepairs.¹³ Pinpointing differences between human genomes is not difficult, but defining genetic variation is less straightforward. Different fields have their own definitions in which the effect of a variant on the outcome of interest plays a major role. In evolution biology for example, deleterious mutations are defined as variants that reduce reproductive success as compared to a hypothetical genome that does not carry that variant.¹⁴ In biomedical research the outcome of interest is usually a diagnostic, therapeutic or prognostic effect. However, when it comes to genetic variation, the effect is usually the relative risk or predisposition of an individual that carries a variant to develop a certain disease.

Genetic variation can be categorized based on the nature of the variant. This ranges from the loss or gain of a full chromosome to the change of a single nucleotide and covers duplications, insertions, deletions, substitutions and translocations.¹⁵ Furthermore, the frequency of a variant in a population (from common to rare) is taken into account. In this thesis I will focus on Single Nucleotide Polymorphisms (SNPs). This is a group of common variants that encompasses the substitution of one nucleotide compared to a reference genome. A single nucleotide variant is called a SNP when the minor allele frequency (MAF, i.e. the frequency at which the lowest abundant allele is present in a population) is higher than 1%. Every individual carries approximately 3.5 to 4.3 million SNPs.¹³

Variation and complex genetic diseases

The protein composition of a cell greatly influences the way a cell responds to the intra cellular and extra cellular environment. Genetic variants can alter this interplay, by affecting the constitution of a protein or its expression level. In some cases, a single variant results in a non-functional protein and subsequently a severe phenotype. This is seen in monogenic diseases like cystic fibrosis and Beta thalassemia.^{16,17} The variants that cause monogenic diseases are rare and in most cases a mutation in the allele of both parents is needed to result in the disease phenotype. Furthermore, individuals that carry such variants in both copies of a gene have a chance that is close to 100% to develop the disease.

Complex genetic diseases are caused by common genetic variants that have a less severe effect on protein function or expression. As a stand-alone variant, these would not result

in a clinically relevant phenotype. However, every individual carries many SNPs that all slightly influence the response to internal and environmental factors and thereby determine how susceptible we are to develop a certain disease.¹⁸

Studying nucleic acids

In this thesis, the sequence, composition and quantification of many RNA and DNA samples is studied. We make use of publically available datasets and data that we generated from cell lines, primary tissue and ex vivo cultures. Together, the data have been generated through the application of dozens of techniques and platforms. The purpose of all of these techniques is to determine which nucleotide polymers are present in a certain sample. They differ in the way through which the polymer is identified, which can be primer based (Polymerase Chain Reaction, PCR), probe-based (arrays) and sequence-based (sequencing platforms). The nature of all input molecules is DNA, independent of an expression or genomic DNA-related question. This is achieved through the reverse transcription of RNA molecule into cDNA (copy DNA, the reverse complement of the RNA molecule).

In PCR, DNA polymerases are used to amplify DNA molecules in cycles that result in doubling of the molecule of interest. The specificity of the technique is based on the need of the polymerase for a double stranded piece of molecule to start the amplification. Therefore, short DNA molecules that are complementary to the sequence of interest (primers) are added to the reaction.¹⁹ This results in the amplification of specific pre-determined sequences. The read-out of PCR can be semi-quantitative or qualitative. In the latter, the presence and length of the amplified molecule is visualized on a DNA-agarose gel. Semi-quantitative PCR (qPCR) is mainly used for expression analysis. The abundance of a transcript is measured relative to an internal reference gene that is not influenced by the study conditions (i.e. the housekeeping gene).²⁰ The fluorescence that is generated by the binding of a dye to double stranded DNA is used as a read-out. Probe-based read-out of qPCR was not used.

DNA microarrays provide the possibility to simultaneously measure the abundance of thousands of molecules. The association of SNPs to complex genetic diseases relies on genome wide profiling of variants through SNP-arrays.²¹ Therefore, DNA microarrays have been inevitable in genetic research since the 1980's. Profiling through arrays is done with probes that are immobilized in spots on a substrate, each spot representing a different known sequence. The choice of platform therefore determines the genes, variants or other sequences that will be profiled. Hybridization of the fluorescently labeled target DNA results in a signal from all spots that bound DNA molecules that were present in the sample. The intensity of the signal in each spot is compared to the intensity of that spot under a different condition, which provides the basis for relative-quantitation. It is an inexpensive method compared to sequencing, however variations in

hybridization and normalization cause some people to prefer RNA-seq over expression arrays.

In contrast to the other techniques, sequencing platforms are not dependent on pre-selected known sequences. Instead, they determine the sequence of the molecules that are present in a sample at a single nucleotide resolution. Quantification can be done on the absolute number of identified sequences. There are many platforms available, accounting for differences in template amplification strategies, sequencing, read length, error rates and costs.²¹ Sequencing can be used to determine novel or patient-specific sequences in genomic DNA, but also to quantify sequences of interest. In this thesis, sequencing is used for transcription profiling and as a targeted approach to gain insight of functional aspects of the genome. In targeted sequencing, pre-treatment of a sample captures genomic sequences with specific characteristics. We make use Chromatin Immunoprecipitation-sequencing (ChIP-seq), Chromatin Conformation Capture-on-chip-sequencing (4C-seq), Self-Transcribing Active Regulatory Region-sequencing (STARR-seq) and Single Cell RNA-sequencing (scRNA-seq).

ChIP-seq

The ChIP-seq technique enables the identification of DNA-sequences to which a given protein is bound.²² First, the DNA and proteins are crosslinked to fix their localization within the nucleus. Next, DNA is digested to create a solution of DNA-protein molecules. Antibodies are then used to precipitate the protein of interest together with the crosslinked DNA. Finally, the proteins are de-crosslinked and the released DNA molecules are sequenced and mapped to a reference genome, to determine their origin. We applied ChIP-seq to profile the chromatin landscape in multiple cell types and conditions. By using antibodies against histone modifications (H3K27Ac, H3K4me3, H3K4me1) we identified active regulatory elements and genes.

4C-seq

Genomic DNA is often interpreted as a linear stretch of nucleotides. However, the 3D organization of the nucleus results in functional interactions between loci that are not close to each other on the linear scale. To address this, chromatin conformation capture (3C) was developed, followed by multiple high-throughput variants of the 3C-technique.²³ We used 4C-seq in which all DNA interactions with a single locus of interest can be identified. 4C-seq was applied to identify physical interactions between genes and enhancers.²⁴ To achieve this, the interactions between DNA loci inside living cells were fixed through crosslinking. Through 2 cycles of digestion and ligation with specific restriction enzymes, circular DNA is generated, that contains two DNA molecules that physically interacted *in vivo*. To selectively amplify all molecules that interacted with the enhancer of interest, outward facing primers are used. Finally the interacting sequences are identified and mapped.

STARR-seq

The activity of regulatory elements is usually profiled based on the co-localization of active histone modifications. STARR-seq was developed as a high-throughput method that directly measures enhancer activity.²⁵ This enabled us to quantify the effect of SNPs on enhancer activity. The technique is based on a library in which sequences of putative enhancers are cloned downstream of a minimal promoter in a plasmid. The library of STARR-plasmids with all putative enhancers is transfected into the cells of interest. If an enhancer is active in a cell type, it activates transcription from its own promoter which results in transcription of its own sequence. Sequencing of the transcriptome from these cells reveals the sequences of active enhancers and abundances reflect the level of activity.

Single Cell RNA-seq

Classical RNA-seq approaches are used to profile the transcriptome of a pool cells and therefore often called bulk RNA-seq. Recently, it has become possible to profile RNA of individual cells. Single cell RNA-sequencing (SCS) involves the isolation of single cells, reverse transcription and amplification of cDNA to enable sequencing on a regular sequencing platform.²⁶ These steps were initially carried out after sorting individual cells into separate wells. New techniques enable the capture of single cells in a droplet in which RNA can be reverse transcribed and subsequently labeled to provide all molecules from one cell with the same barcode for further processing.²⁷ A major challenge in SCS is the identification of rare transcripts as the resolution currently results in the sequencing of a limited percentage of molecules per cell. We use scRNA-seq to identify differences in responsiveness of the multitude of intestinal epithelial cell types.

Studying complex genetics

With the development of high-throughput-genomics, novel approaches to delineate the genetics behind complex genetic diseases have arisen. These studies started from the assumption that complex genetic disease are (in part) caused by the presence of an unfortunate combination of common genetic variants that lead to the disease phenotype when the carrier encounters certain environmental factors (ranging from nutrients to stress and from infections to exposure with chemicals). Identification of the genetic variants that underlie this predisposition would provide insight into the pathological mechanisms. Over the past two decades, many genome wide association studies (GWASs) have successfully associated thousands of SNPs to a great variety of diseases. Although several associations have lead to the identification of pathogenic mechanisms, the pathogenesis of complex genetic diseases is generally not well understood. This is due to the large number of genetic variants that are associated to single diseases and to the problematic translation of associated variants to the pathogenic nature of the variants.

The association of a SNP to a disease does not necessarily mean that this SNP is the disease causing variant. It is rather seen as a marker for a genetic locus that segregates in a population (i.e. a haploblock). Whereas the associated SNP is the marker of this locus, another variant within the locus likely causes the disease phenotype. Therefore, the association of a SNP is the lead of departure for the in depth study of the genetic locus. Classical approaches to unravel the link between the locus and the disease were focused on the genes that are localized within the locus. When there is functional or genetic (presence of mutations) evidence for the relation between a gene and a disease, the gene is called a candidate gene. However, many associated variants cannot be linked to missense mutations within gene bodies.

With the growing knowledge on DNA regulatory elements, it is hypothesized that the association of a SNP can also be caused by variants that alter DNA regulatory elements. This implies that the disease phenotype can be due to altered gene regulation and therefore altered expression levels. Since DRE's can regulate genes at great distance, the candidate genes in this model are not limited to the genes within the associated locus. In this thesis, we make use of high throughput genomics to study the relation between GWAS-associations and DRE in three complex genetic diseases: inflammatory bowel disease (IBD), chronic kidney disease and cardiovascular disease (CVD). We use ChIP-seq to identify active DRE in relevant cell types, STARR-seq to study how associated variants influence DRE-activity and 4C-seq to study the 3D interactions between associated loci and the genes they regulate. By intersecting these data with a vast amount of publically available datasets, we try to broaden the view on the functional genomics of complex genetic diseases in search for pathological mechanisms that underlie these diseases and for novel putative therapeutic approaches.

Complex disease, complex tissue, complex environment

The interpretation of disease-associated genetic variants depends on our knowledge of the function of genes at the molecular, cellular, tissue, organ and entire body-level. On the one hand genetic findings provide novel insight in pathogenic mechanisms and the involvement of specific cell types, while on the other hand, knowledge on cellular processes helps to interpret and prioritize results from genetic studies.

Studies on the genetics of inflammatory bowel disease by us and others, point out an important role for the intestinal epithelium. Genetic defects in epithelial-specific genes have been identified^{28,29} and IBD-associated variants are found to co-localize with epithelial DRE³⁰. We therefore have to understand the processes that take place at the epithelium to be able to delineate how genetic predisposition leads to IBD. In the last part of this thesis we focus on the influence of bacterial antigens on the homeostasis of the intestinal epithelium. By studying how differentiation, proliferation and responsiveness

are regulated in the ever changing environment of the intestinal epithelium we aim to unravel mechanisms that prevent the development of intestinal inflammatory conditions like IBD.

Intestinal epithelium

The intestinal tract is lined with a monolayer of intestinal epithelial cells (IECs). The epithelium consist of multiple specialized cell types that are organized in crypt and villus structures. The proximal small intestine is made up of long villi and moving down the GI tract, the villi become shorter where in the colon multiple crypts dip down from a flat luminal surface.³¹ Crypts and villi harbor different cell types, with stem cells residing at the bottom of the crypt, while differentiated cells, mainly enterocytes, can be found in the villus.³² The intestinal epithelium forms a physical barrier that separates the luminal content of the intestine from the body. Besides the barrier function, IECs play an important role in nutrient absorption and transport. As IECs are the first cells to encounter and sense luminal antigens, they may also be regarded as crucial cells in orchestrating local immune responses. Eliciting the appropriate epithelial response is crucial to maintain intestinal homeostasis. If an imbalance occurs in the responsiveness to bacterial antigens the epithelial homeostasis will be challenged by ongoing inflammation and susceptibility to infection as is seen in inflammatory bowel disease (IBD).

Organoids as model for the intestinal epithelium

We make use of an *in vitro* model to study the intestinal epithelium. We isolate stem cells from human intestinal biopsies and provide them with niche factors (WNT3A, p38 inhibitor, Alk inhibitor) to enable the growth of epithelial structures called intestinal organoids.^{32,33} Organoids can be expanded over long periods and by altering the culture conditions they can form many of the native specialized epithelial cell types.³⁴ Intestinal organoids grow as spherical structures with the luminal membrane facing the inside and the basolateral membrane attached to a 3D matrix on the outside.

As organoids have the same genetic background as their donors they are often used as a disease model. For example, they can be used to predict a patient's prognosis or response to drugs.^{35,36} Next to the genetic background, the intestinal stem cells have been shown to be programmed with their location specific profile and therefore resemble the epithelium of the intestinal location of the biopsy (colon stem cells form colon organoids etc.).³⁷ These characteristics enable us to study the intestinal epithelium at a location, cell type and membrane polarity specific level.

References

1. Levene, Phoebe, A. T. The structure of yeast nucleic acid. IV. Ammonia hydrolysis. J. Biol. Chem. 40, 415–424 (1919).
2. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid.

- Nature 171, 737–738 (1953).
3. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389, 251–260 (1997).
 4. Ensembl. Human assembly and gene annotation. Available at: https://www.ensembl.org/Homo_sapiens/Info/Annotation. (Accessed: 3rd September 2018)
 5. Sharp, P. A. Split genes and RNA splicing. *Science* 77, 805–815 (1994).
 6. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution - Trashing old rules to forge new ones. *Cell* 157, 77–94 (2014).
 7. Dunham, I. Et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
 8. Lander, E. S. Et al. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).
 9. Venter, J. C. Et al. The Sequence of the Human Genome. *Science* (80-). 291, 1304–1351 (2001).
 10. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–86 (2014).
 11. Heintzman, N. D. Et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318 (2007).
 12. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* 62, 668–680 (2016).
 13. Auton, A. Et al. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
 14. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* 16, 333–343 (2015).
 15. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* 10, 241–251 (2009).
 16. Schwank, G. Et al. Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell Stem Cell* 13, (2013).
 17. Thein, S. L. The molecular basis of β -thalassemia. *Cold Spring Harb. Perspect. Med.* 3, 1–24 (2013).
 18. Meddens, C. A., van der List, A. C. J., Nieuwenhuis, E. E. S. & Mokry, M. Non-coding DNA in IBD: from sequence variation in DNA regulatory elements to novel therapeutic potential. *Gut* *gutjnl-2018-317516* (2019). Doi:10.1136/gutjnl-2018-317516
 19. Mullis, K. B. & Faloona, F. A. B. T.-M. In E. In *Recombinant DNA Part F* 155, 335–350 (Academic Press, 1987).
 20. Overbergh, L. Et al. The use of real-time reverse transcriptase PCR for the quantification of cytokine gene expression. *J. Biomol. Tech.* 14, 33–43 (2003).
 21. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351 (2016).
 22. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* (80-). 316, 1497–1502 (2007).
 23. Wit, E. De & Laats, W. De. A decade of 3C technologies-insights into nuclear organization. *Genes Dev.* 11–24 (2012). Doi:10.1101/gad.179804.111.GENES
 24. Van de Werken, H. J. G. Et al. 4C technology: protocols and data analysis. *Methods in enzymology* 513, (Elsevier Inc., 2012).
 25. Arnold, C. C. D. C. Et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* (80-). 339, 1074–7 (2013).
 26. Barbacioru, C. Et al. Mrna-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382 (2009).
 27. Macosko, E. Z. Et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214 (2015).
 28. Kaser, A. Et al. XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease. *Cell* 134, 743–756 (2008).
 29. Wehkamp, J. Et al. NOD2 (CARD15) mutations in Crohn’s disease are associated with diminished mucosal alpha-defensin expression. *Gut* 53, 1658–64 (2004).
 30. Mokry, M. Et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology* 146, 1040–1047 (2014).
 31. Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nat. Rev. Mol. Cell Biol.* 15, 19–33 (2014).
 32. Sato, T. Et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* (2009). Doi:10.1038/nature07935
 33. Sato, T. Et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett’s epithelium. *Gastroenterology* 141, 1762–72 (2011).
 34. Fujii, M. Et al. Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. *Cell Stem Cell* 23, 787–793.e6 (2018).

35. Weeber, F. Et al. Preserved genetic diversity in organoids cultured from biopsies of human colorectal cancer metastases. *Proc. Natl. Acad. Sci.* 112, 13308–13311 (2015).
36. Dekkers, J. F. Et al. A functional CFTR assay using primary cystic fibrosis intestinal organoids. *Nat. Med.* 19, 939–45 (2013).
37. Middendorp, S. Et al. Adult stem cells in the small intestine are intrinsically programmed with their location-specific function. *Stem Cells* (2014). Doi:10.1002/stem.1655



Non-coding DNA in inflammatory
bowel disease: From sequence variation
in DNA regulatory elements to novel
therapeutic potential

2

Claartje A. Meddens, Amy C.J. van der List, Edward E.S.
Nieuwenhuis, Michal Mokry

Based on: *Gut* 68(5):928–41 2019

Inflammatory bowel disease

Inflammatory Bowel Disease (IBD) is a group of disorders of the gastro-intestinal (GI) tract that are characterized by intermittent, chronic or progressive inflammation. There are two main groups of IBD: Crohn's disease (CD) in which transmural inflammation of the intestinal wall occurs in patches throughout the whole intestine, and ulcerative colitis (UC) in which inflammation is limited to the mucosa of the colon and rectum.¹ The pathogenesis of IBD is multifactorial and includes genetic susceptibility as a major contributor. Alongside the growing number of identified genetic risk variants for IBD, there is growing knowledge on the role and functions of many different elements that often reside outside of the protein-coding regions of the human genome. This impacts the interpretation of the role that these variants play in IBD pathogenesis. In this review, we will introduce the concept of DNA regulatory regions, elucidate the possible roles of non-coding regulatory DNA in IBD pathogenesis and discuss how this creates novel therapeutic possibilities.

Genetic susceptibility in IBD pathogenesis

It has been known for decades that both Crohn's disease and ulcerative colitis show familial clustering, albeit in the absence of a clear Mendelian inheritance pattern. Therefore, both diseases are generally considered polygenic disorders with variable phenotypic penetrance. Reported risk ratios for siblings of CD and UC patients compared to the general population vary between 15-42 and 7-17, respectively.² Three nationwide twin studies performed in Sweden, Denmark and the UK revealed increased concordance of IBD in monozygotic twins when compared to dizygotic twins, thereby elucidating that familial clustering is not solely based on shared environmental factors but rather on a shared genetic background.³⁻⁵ In CD, the difference in concordance is more pronounced than it is in UC, implying a greater genetic influence on the pathogenesis of CD than UC. The Swedish study further revealed significantly higher co-occurrence rates of phenotypic characteristics in monozygotic twins when compared to dizygotic twins, especially for age of onset and location of disease, suggesting that not only disease but also phenotypic manifestations of IBD are heritable.³³ Both disease phenotype and clinical course have been shown to be influenced by IBD-associated genetic variants.^{6,7}

Based on these observations, a substantial effort was made to identify genetic elements involved in IBD pathogenesis. In this respect, multiple Genome Wide Association Studies (GWASs) were performed over the past years.⁸⁻¹³ These studies assayed common genetic variants (Single Nucleotide Polymorphisms - SNPs) spanning the whole genome in search of SNPs that are significantly overrepresented in patients when compared to healthy controls. SNPs that occurred more frequently in patients are thus called disease-associated variants.

Many genetic loci and variants that are located side by side are inherited together during meiosis (i.e. there is a very small change of the occurrence of recombination sites between

them). Due to this phenomenon, associated SNPs are generally considered to be markers for other variants located in the coding region of nearby genes. This approach has led to the identification of numerous crucial genes and pathways involved in IBD pathogenesis and has allowed for the development of novel therapeutic approaches.¹⁴ GWAS meta-analyses have identified over 200 loci that are associated with IBD and account for the increased risk of development of the disease.^{15,8,16} Because the search for candidate genes at the associated loci has proven difficult, most loci are not yet functionally linked to IBD. There are indications that pathogenic processes between CD and UC are convergent and share the majority of their associated loci (e.g. the IL23R locus).^{8,17}

Overall, Inflammatory Bowel Disease consists of a range of inflammatory disorders in the GI tract with distinctive onsets, severities, localizations and complications. Although a clear role for genetic susceptibility in IBD pathogenesis has been identified through multiple GWASs, the translation of these findings into patient benefits has been limited. The majority of SNPs that are associated with IBD were found to be located in non-coding DNA. Therefore, these SNPs cannot be causative in the sense that they directly lead to amino acid changes at the protein level.^{8,18,9-12,19-21} Notably, knowledge on the functional non-coding elements in the human genome has tremendously increased. This knowledge can now be used to better interpret GWAS-findings. Here we review new approaches in which the focus shifts from coding sequences to other functional parts of the human genome and how this further enhances our knowledge of causative IBD genetic mutations and the awaited therapeutic perspectives.

DNA regulatory elements

Genomic DNA functions as a carrier of information that dictates the primary sequences of genes. In addition to this important function, non-coding DNA contains elements that are involved in transcriptional regulation.²² According to the 2013 UCSC genome build (GRCh38.p12, Dec 2013, last updated Jan 2018) the human genome consists of 3,609,003,417 base pairs and 20,376 coding genes.²³ The coding sequences of genes make up only 2% of the whole genome, which designates 98% of the 3 billion base pairs as non-coding DNA.²⁴ Over the last years, some functions of non-coding DNA have been discovered and the role of regulatory sequences in transcriptional regulation, development, disease and determination of cell type specificity is now widely appreciated.²⁵⁻²⁸ Multiple distinct elements have been identified, each of which displays specific characteristics and plays different roles in transcriptional regulation (**Figure 1**).

Enhancers

Enhancers are located distal from the genes they regulate and can be found up to several megabases away from the transcription start site.²⁹ They are involved in enhancing the transcription of a regulated gene (**Figure 1A,B**). These distal regulatory elements can regulate multiple genes and one gene can be regulated by multiple enhancers. The

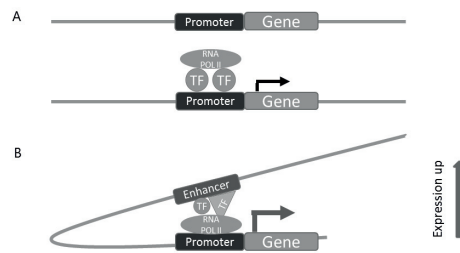


Figure 1. DNA Regulating Elements.

A) Promoter elements are located directly upstream of genes. By binding transcription factors and recruiting RNA Polymerase II, promoters mediate transcription. B) Enhancer elements are located distally from the gene(s) they regulate(s). Active enhancer elements increase transcription levels, which is mediated through transcription factor binding. Abbreviations: TF: transcription factor, RNA Pol II: RNA polymerase II.

mechanisms of enhancer function have been studied at both the molecular (enhancement of transcription) and cellular levels (cell differentiation and development). The mechanism through which enhancers regulate the transcription from promoters has been widely discussed.³⁰ Multiple models have been proposed, and the ‘looping’ model received support after the development of a method known as chromosome conformation capture (3C). The looping model states that the action of enhancers relies on a physical interaction between the enhancer and the promoter of a reporter gene. Various 3C-based methods have been developed to enable the identification of genomic regions that are in physical contact with each other. By using these techniques, long-range interactions between enhancers and promoters have been discovered.³¹ In studies of the β -GLOBIN locus and its distant locus control region, specific interactions between enhancer-bound and promoter-bound transcription factors (TFs) have been discovered.^{32,33} Through these and other studies it is now generally accepted that enhancer function is established by TFs that bind to transcription factor binding sites (TFBSs) and subsequently by the interaction between the TFs that are bound to promoters at one site and the TFs bound to enhancers at other sites.³¹

Enhancers play a key role in the establishment of complex expression patterns that determine the diversity of spatial and temporal gene expression within an organism.²² Enhancer activity is highly cell type specific and can be correlated with gene expression patterns. Some enhancers can be found in clusters that drive the expression of cell type specific genes. These clusters of enhancers are called super-enhancers and have been shown to be enriched in disease-associated variants.^{34,35} The combination of the histone modifications histone 3 lysine 4 mono-methylation (H3K4me1) and histone 3 lysine 27 acetylation (H3K27ac) predominantly mark active enhancers, whereas H3K27 acetylation or H3K4me1 alone are predominantly found at ‘poised’ or inactive genes.^{36–38} Interestingly, many regulatory elements can have both enhancer and promoter characteristics and can therefore show different histone marks in time, in cell types and cell states.^{39–43} By investigating H3K4me1 distribution, between 24,000 and 36,000 enhancers were identified per cell line. Surprisingly, only a minority of these locations (approximately

5,000) overlapped between the two cell types examined, showing that enhancers exhibit cell type-specific patterns. The contribution of epigenetic signatures on development is further established by the finding that fetal gut shows distinct DNA methylation patterns and dynamics from paediatric and adult intestinal tissue.⁴⁴ By identifying the divergent patterns of enhancer activity in different cell types within one organism, enhancer activity was shown to play an important role during differentiation and development.⁴⁵⁻⁴⁶

Promoters

Promoters are regulatory elements that can be found at the 5'-end of a gene and are involved in transcription initiation. Promoter sequences contain TFBSs that recruit (TFs) and assemble the transcription pre-initiation complex that guides RNA polymerase II to the transcription start site.⁴⁷ The location of promoters relative to genes, usually located just upstream of a gene, has facilitated their identification and annotation (**Figure 1A**).⁴⁸ Although active promoters are classically marked by histone 3 lysine 4 tri-methylation (H3K4me3) histone modifications,³⁷ recent studies have shown that some elements can exhibit promoter activity for one gene and enhancer activity for another gene.^{39-41,43} As described above, one DRE can be marked by different histone modifications, which reflects the potential of elements to execute multiple regulatory functions.⁴⁰ Promoters and enhancers might therefore no longer be seen as individual entities, but rather as a spectrum of DRE-activity.⁴¹

Next to enhancers and promoters, multiple other non-coding elements are involved in the regulation of gene expression. For example silencer- and insulator elements are mainly involved in negative regulation of gene expression. As such, expression levels are subject to the composition of all sorts of elements along the chromosomes. In contrast to the cell type-specific activity of enhancer elements, the activity of most other elements is consistent throughout different cell types, as is reflected by their stable chromatin states.⁴⁹

Mechanisms through which variants in DRE cause human genetic disease

Since the knowledge on DNA regulatory elements (DREs) has increased, it has become clear that DREs are involved in determining cell types and differentiation. Sequence variants can be located in the whole genome, including DNA regulatory elements, and causative variants in both coding and non-coding elements are important players in the pathogenesis of complex genetic diseases such as IBD.

There are multiple ways through which sequence variants in DNA regulatory elements can contribute to the development of a genetic disease. One of the proposed mechanisms is through the alteration of TFBSs. In this model, a sequence variant in the TFBS changes the affinity of a TF for this specific sequence. This can subsequently change the regulatory activity of DREs. **Figure 2** shows an example of an altered NF- κ B-binding site in an

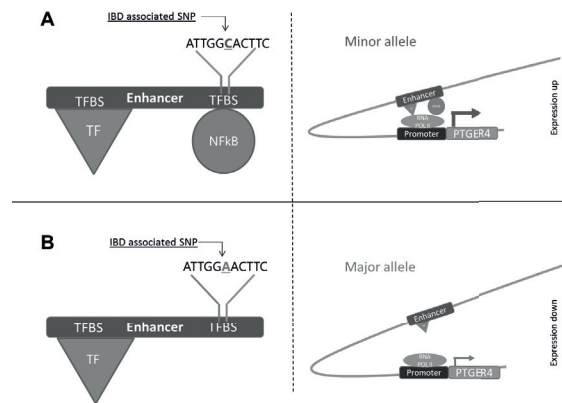


Figure 2. Influence of sequence variants on protein expression levels.

The presence of a SNP in the TFBS of an active enhancer enables the enhancer to increase or decrease the transcription of genes. A) Example of a SNP (rs4495224) that is associated with CD and is located in a TFBS in an enhancer element that regulated the expression of PTGER4. In silico analysis demonstrated that the C-allele of the rs4495224 polymorphism enables the binding of NF-κB.⁵⁰ B) The A-allele of the rs4495224 polymorphism is thought to alter the TFBS in an enhancer, which will cause decreased affinity of NFκB. This could result in less enhancer activity and a subsequent decrease in PTGER4 expression levels.

enhancer that regulates the expression of the PTGER4-gene. The disruption of this binding site results in less binding of NF-κB and subsequently a decreased expression of an IBD-associated gene (PTGER4).⁵⁰ Another mechanism through which variants can contribute to pathogenic processes is by their influence on DRE activity and corresponding histone modifications and DNA methylation profiles.^{51–53} Recently, these theoretical models were supported by the finding that allelic differences lead to allele-specific activity of many regulatory layers. Sequence variants were shown to affect transcription levels, binding of transcription factors, DNA methylation, histone modifications and histone positioning at the locus of the variant.^{54–56} Furthermore, this sequence-based variation in regulatory modules can be transmitted from parent to child, indicating that altered transcription regulation might be a basis for the heritable pathogenic processes.^{57,58}

Besides variants in DNA regulatory elements, there are variants in non-protein-coding DNA that result in a phenotype through different mechanisms. Variants that cause alternative splicing and thereby affect the protein structure have been described.⁵⁹ Furthermore, variants in genes that code for miRNAs or lncRNAs can result in affect the function of these non-coding RNAs and can thereby contribute to IBD pathogenesis.^{60,61}

Several pathogenic mechanisms through which non-coding DNA can result in disease phenotypes have been delineated in monogenic diseases. For example, pre-axial polydactyly was found to be caused by a single mutation in a regulatory element of the Sonic hedgehog-gene that is found 1 Mb upstream of the gene.²⁹ This study and other studies⁶² demonstrate the possibility that even a single deleterious sequence variant in a regulatory region can cause disease. The mechanisms through which less common

genetic variants (i.e., deletions, translocations, etc.) alter enhancer activity and contribute to genetic disease has been extensively reviewed.⁶³

Taken together, these data establish that sequence variants in non-coding DNA can influence the activity of regulatory elements and that such sequence variants are often associated with human genetic diseases.

The majority of IBD-associated variants map to DNA regulatory elements

To understand the pathogenesis of IBD, the major contributors should be identified on the level of proteins, RNA transcripts and consequently genes. The genetic background of IBD has been successfully studied through GWASs, but unfortunately extracting the major players and causative genes is not straightforward. Initially, the translation of GWAS data into candidate genes focused on variants that could be linked to genes via their localization in close proximity to the associated variant (**Figure 3A,B**).^{11,19,20} However, the association of many variants that lie outside gene bodies cannot be linked to missense mutations.

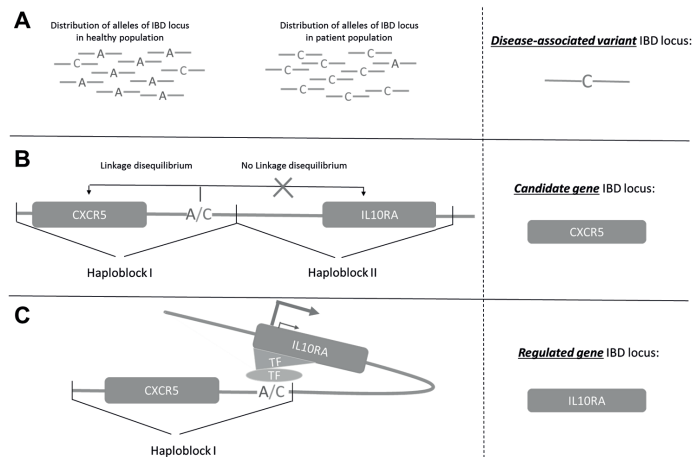


Figure 3. Candidate gene approaches for disease-associated variants.

A) In Genome Wide Association Studies, sequence variants at multiple genomic loci are studied. In this example, studying sequence variants at an IBD-associated locus reveals a C-allele was found more frequently in the patient population, therefore this allele is determined to be a disease-associated variant. B) Model of classical candidate gene approaches. A disease-associated variant is considered to be a marker for a causative coding variant in a gene that is transmitted to offspring on the same stretch of DNA, i.e. is inherited on the same haplotype. This gene is considered to be a possible candidate gene. For example, the IBD-associated SNP rs630923 (A/C) is located in the vicinity of CXCR5. CXCR5 is therefore considered to be an IBD candidate gene.⁸ C) Novel model of candidate gene approaches. In this model, the disease associated variant lies within a DNA regulating element. The SNP in this element results in changes in the expression of the target gene. The target gene can be located outside the haplotype and will therefore not be found by using the classical model. For example, the IBD-associated SNP rs630923 is located in an enhancer that was found to regulate the IL10RA gene that is found in another haplotype. Therefore, IL10RA is considered as a candidate gene as well.⁸⁵

As over 80% of the human genome consists of functional regulatory elements, it was hypothesized that the association of non-coding SNPs to IBD could be due to their effect on regulatory elements. Maurano et al. determined the large overlap between GWAS-associated SNPs for multiple diseases and open chromatin (defined by DNase Hyper Sensitivity sites, DHSs). DHSs are a proxy for chromatin that is accessible for proteins such as transcription factors and therefore, active regulatory elements are generally found to map to DHSs. This study revealed a significant enrichment of disease associated SNPs to DHSs and showed that these putative regulatory elements can interact with promoters over long genomic distances *in vivo*.⁵¹ **Figure 3C** shows how novel approaches can result in the identification of novel IBD candidate genes (for example *IL10RA*) at previously associated loci. Disease-associated SNPs were also found to be enriched in a combination of DNase footprints, TBFSS and histone modifications.⁶⁴ The role of the chromatin landscape in IBD was further established by the finding that colon tissue of CD patients can be subdivided into two clinically relevant subtypes based on chromatin accessibility profiles.⁶⁵

To address whether loci specific for IBD localize to DRE, we profiled active regulatory elements in cell types that are relevant for IBD. We found that 56% of IBD-associated SNPs can be linked to either immune or intestinal epithelial cell-specific regulatory regions (as determined based on the presence of histone modifications).¹⁸ This co-localization occurred approximately three times more frequently in IBD-associated SNPs when compared to randomized sets of SNPs. Furthermore, the enrichment of the localization of IBD-associated SNPs in enhancers and promoters that are activated upon active UC and CD has been revealed.⁶⁶ These data suggest that a large part of the IBD-associated loci can be explained by sequence variants within DNA regulatory elements.

Individual non-coding IBD-variants alter DNA regulatory elements

As the vast majority of associated SNPs lies in non-coding DNA, much effort has gone into identifying candidate genes by linking specific variants to genomic regulatory functions. Delineating individual loci has led to the identification of pathogenic regulatory mechanisms and the genes that are affected by non-coding variants.

IRGM-locus

The association of the *IRGM*-gene locus to IBD has been established through multiple GWASs.⁶⁷⁻⁶⁹ *IRGM* is of great interest as it is involved in the early phases of autophagy, a process implicated in IBD pathogenesis.^{67,70} In depth sequencing of the coding sequence of this gene could not reveal any nonsynonymous mutations.⁶⁷ This implies that the causal variant must be located in non-coding sequences. Indeed, a common 20-kb deletion polymorphism and two small insertions were found upstream of the *IRGM*-gene body and turned out to be in strong linkage disequilibrium (LD) with the most strongly CD-

associated SNP. Subsequently, the effect of the deletion and insertions was studied.^{67,69} The risk alleles were found to perturb *IRGM* expression levels and, subsequently manipulated *IRGM* expression affected cellular autophagy.⁶⁹ Finally, a family of miRNAs are involved in the regulation of *IRGM*-expression and were found to downregulate the protective allele, but not the risk allele.⁷¹

PTGER4-locus

A GWAS by Libioulle *et al.* identified a novel IBD-risk locus on chromosome 5p13.1. The associated risk variants map to a 1.25 Mb gene desert, which complicated candidate gene identification. To test the regulatory potential of the locus, the effect of the SNPs on expression levels of the neighboring genes was profiled. This approach revealed two SNPs that significantly increase the expression of *PTGER4*.¹² *In silico* analysis showed that this is likely caused by the alteration of two transcription factor binding sites (*NFKB* and *XBP1* respectively). The increased affinity of these TFs for the associated locus could explain the increased expression of *PTGER4*.⁵⁰ The role of this locus is supported by the increased susceptibility of *Ptger4* mutant mice to dextran sulfate sodium (DSS)-induced colitis and the established roles of *NFKB* and *XBP1* in IBD pathogenesis.⁷² Furthermore, variants at the associated locus were systematically screened for aberrant regulatory capacity in lymphoblastoid cells. This revealed a single SNP in a distal enhancer that affects the expression of *PTGER4* *in vitro* and could be rescued by genome editing with CRISPR/Cas9.⁷³

TRIB1-locus

There are two SNPs associated with Crohn's disease that are located upstream of the coding sequence and locate to open and therefore likely active chromatin.⁵¹ These two SNPs were found to alter T-Bet binding motifs and result in reduced T-bet binding *in vivo*. This is relevant, because T-Bet is a transcription factor that is involved in T cell differentiation and plays a key role in T cell-mediated colitis.⁷⁴ Furthermore, a direct link between T-bet and *TRIB1* was established by showing that *T-Bet* *-/-* mice show decrease *TRIB1* expression.⁷⁵ These data show that these CD-associated SNPs alter transcription regulatory processes that correlate with *TRIB1* expression.

IL18RAP-locus

In a GWAS from 2010 an IBD-associated variant on chromosome 2q12 was identified. This variant was studied through mRNA profiling in whole blood samples. This revealed that the minor allele specifically correlated with altered expression of the *IL18RAP* gene.⁷⁶ Other associated SNPs at the same locus were found to alter the T-bet TFBS that results in differential expression of *IL18RAP*.

TPL2-locus

TPL2 (*MAP3K8*) maps to a locus that is associated to Crohn's disease.⁸ Although no variants that caused amino acid changes were found, *TPL2* was an obvious IBD candidate gene because it is an important player in T cell and innate responses and it induces

cytokine production in a Pattern Recognition Receptors (PRR)-signaling cascade.⁷⁷ In mice, *TPL2* was shown to contribute to the development of colitis.^{78,79} Hedl *et al.* revealed the putative role of associated variants on DNA regulation, by showing that the disease-associated allele has an increased expression of *TPL2* in monocyte-derived macrophages.⁸⁰ Furthermore, this study elucidated that carriers of the risk allele demonstrate increased *NOD2*- and *NFKB*-signaling that subsequently results in increased cytokine production.

The role of regulatory elements in IBD studied through systematic approaches

The studies described above concern examples of the association of single IBD-associated loci with a putative DNA regulatory function. However, there is a strong need for novel approaches to systematically identify the genes that are regulated through elements that carry IBD variants (**Figure 4A**).

IBD associated variants overlap eQTLs and influence transcription levels

A common feature of regulatory elements is their influence on RNA expression. Therefore, SNPs that contribute to a disease phenotype are expected to affect allele-specific RNA-expression levels. This mechanism is used to profile eQTLs (expression quantitative trait loci) i.e. loci that contain sequence variants that influence the expression of specific genes, thereby identifying SNPs that are linked to transcription regulation (**Figure 4B**). As regulatory activity is cell type specific, the influence of sequence variants is limited to

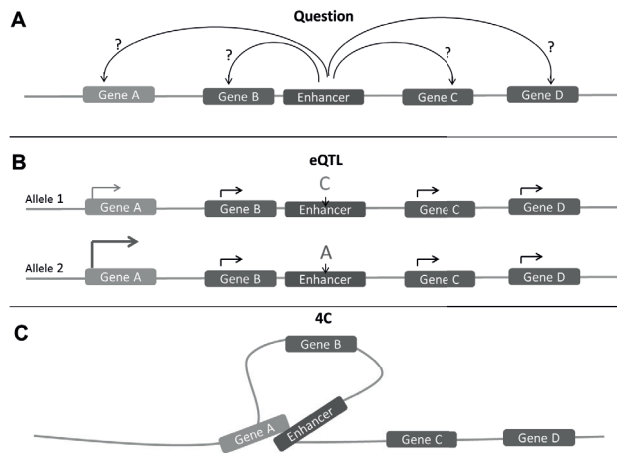


Figure 4. Candidate gene identification through eQTLs and 4C.

A) Identification of candidate genes based on the proximity of a gene to the enhancer that carries the IBD-variant is a biased approach. From the linear composition of DNA it is not possible to identify the genes that are regulated by associated-enhancers. B) eQTLs (expression Quantitative Trait Loci) are used to identify differential expression between alleles that carry different IBD-associated variants. The differentially expressed genes are considered to be candidate genes. C) Circular Chromatin Conformation Capture (4C) is used to study the 3D conformation of the chromatin and identify candidate genes based on their physical interaction with IBD-associated-enhancers.

cells in which the overlapping DRE is active. Therefore, eQTL databases of many cell types have been developed.^{76,81–83} eQTL profiles of cell types that are likely involved in the IBD pathogenesis are used to assign candidate genes to IBD-associated SNPs that influence expression levels. This approach is now commonly used to complement candidate gene approaches that are based on genomic distances between SNPs and genes.^{15,8,9,13} Using eQTLs and regulatory information enables the identification of candidate genes for many loci that could not have been done previously. This is exemplified by a GWAS in which only 3 of 38 novel SNPs were in LD with known missense mutations, whereas 14 SNPs showed eQTL-effects.¹³

Besides allele-specific influence on RNA expression, SNPs can also influence chromatin landscapes and transcription factor binding. IBD associated SNPs were found to influence the accessibility of the DNA and thereby likely influence the extent to which DREs can execute their function in individuals that carry risk variants.⁵¹ Furthermore, IBD associated SNPs were found to be enriched for localization to TFBSs^{16,18,66,51} and to result in allele-specific affinity of transcription factors.^{75,84} Finally, DNA methylation that influences DNA accessibility, was confirmed to show allele-specific patterns.⁵² As such, various layers and mechanisms have been discovered that enable a better understanding of the complex role that genetic variants can play, ultimately through (defected) regulation of DNA transcription of relevant IBD related genes.

Candidate genes physically interact with DNA regulatory regions

Many studies specifically analyze whether associated SNPs affect the expression of genes that are located in the vicinity of the SNP. To overcome this biased approach for the identification of candidate genes, we developed a novel, unbiased method that complements classical approaches to candidate gene identification (*Figure 4C*).⁸⁵ This approach relies on the physical interaction between enhancer elements and the genes they regulate. Using 4C-seq (Circular Chromatin Conformation Capture-sequencing) we studied each gene that interacts with one of the 92 active enhancer elements that carries an IBD-associated SNP. We applied this technique to intestinal epithelial cells, monocytes and lymphocytes and identified 902 putative candidate genes with an average distance of 300 kbp (kilo base pairs) per interaction. These results emphasize that genes that contribute to the IBD pathogenesis can be located further from the associated SNP than is assumed in classical candidate gene approaches and has recently been applied to other complex genetic diseases.^{86,87} This approach identified noteworthy genes including *ATG9A*⁸⁸, a gene involved in autophagy that has been implicated in IBD-pathogenesis, and *IL10RA*, a gene that has been shown to play a role in monogenic forms of IBD.⁸⁹

Gene prioritization

With increasing knowledge on the regulatory functions of the majority of IBD-associated loci, the complexity of the genetic background of IBD seems to be ever expanding. Therefore, *in silico* methods are being developed that integrate all regulatory, chromatin,

coding and non-coding information to identify the key regulators in IBD.^{90,91} Peters et al. applied a predictive model to identify gene networks that play a role in IBD and subsequently validated 12 key drivers of IBD pathogenesis. Our approach, through systematic analysis of chromatin interactions, helped identify common upstream regulators of the IBD candidate genes.⁸⁵ As such, our results suggest an important role for *HNF4A* (a transcription factor that belongs to the nuclear hormone receptor superfamily) in both intestinal epithelium and immune cells. These in silico approaches are crucial to translate the extending number of IBD associated genes to key regulators of the main pathogenic pathways and finally to identify novel therapeutic targets.

Regulatory information creates novel insight in cell types and disease states

On the one hand, the cell type specificity of DRE activity can complicate the studies of genetic variants in regulatory elements, as the functional impact of each associated SNP may be present only in specific cells. Reciprocally, this phenomenon can help us to pinpoint cell types that are involved in IBD pathogenesis as the active regulatory elements from these cells are enriched for IBD associated variants. Therefore, a characteristic of DRE that seemed complicating at first, has paved the way for the identification of cell types that play a role in the IBD pathogenesis. The first study to apply this strategy showed that T helper 17 and T helper 1 cells have the highest accumulation of SNPs in accessible chromatin.⁵¹ Studying the enrichment of IBD-associated SNPs in cell type specific active DRE revealed marked differences between UC and CD.

We and others have shown that in UC both intestinal epithelial cells (IECs) and immune cells seem to be important players, whereas in CD IECs are found to play a less important role than immune cells.^{16,18,66,92} This implies that there are distinct pathological processes underlying CD and UC and that these are limited to immune cells in CD, though in UC the intestinal tissue itself is a major player.

DREs are not only differentially active among cell types, but also among cell states. Studying the activation of monocytes through stimulation with Lipopolysaccharides (LPS) or IFN γ , revealed many context-specific eQTLs. This means that some SNPs will only affect pathogenicity upon, for example, inflammation or during infection. These context-specific eQTLs are now used in search of causative genes in GWASs.¹⁵ A recent study has identified eQTLs in ileal biopsies from patients in different stages of CD (healthy, complication-free disease and disease progression with stricturing or penetrating disease). Through this approach, disease stage-specific eQTLs have been identified. The IBD-associated SNPs that form these eQTLs are therefore likely stage-specific and may be used to predict disease progression.

DNA regulatory elements as therapeutic targets

The majority of IBD-associated SNPs have been shown to affect transcription regulation by altering the sequence of DRE. This creates possibilities for novel therapeutic strategies. Based on the common features of the SNPs that are involved in DNA regulation, druggable targets may be defined.

The presence of a SNP in a disease associated locus can result in changes in TFBSs and in the chromatin landscape. Although there is an interplay between the deposition of histone modifications and the affinity of transcription factors for TFBSs, they can be targeted through different mechanisms. Here we will discuss how the chromatin landscape and key IBD-transcription factors can be targeted and we will review the progress that has been made on this score.

Targeting the chromatin landscape

The activity of DNA regulatory elements strongly correlates with the co-occurrence of histone modifications, DNA methylation, and transcription factor binding.⁹³ Histone modifications are covalent post-translational modifications of one of the four histone tails including acetylation and methylation (*Figure 5*). The tails of histones H3 and H4 are important for transcriptional regulation of numerous genes. Modifications of histones by acetylation are known to weaken the chemical attractions between nucleosome components, enabling the DNA to uncoil from nucleosomes and allowing access to proteins important for transcription such as RNA Polymerase II and TFs. Acetylation of histones by ‘writer’ enzymes called acetyltransferases (HATs) is known to increase the expression of the genes that are regulated by the acetylated enhancers.⁹⁴ The addition

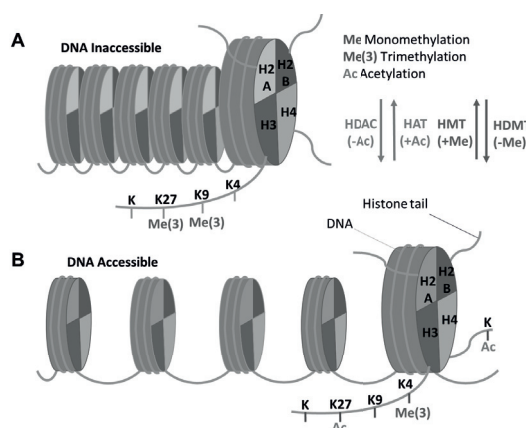


Figure 5. Histone modifications.

Histone tails that belong to one of the 4 subgroups of histones, can be post-translationally modified by covalent attachment of (amongst others) Acetyl and Methyl groups. These modifications are deposited and removed by specific enzymes. Each modification has a different effect on transcriptional activity. A Repressive histone modifications cause the DNA to be densely packed and be inaccessible for transcription factors. B Activating histone modifications are found at accessible DNA, which can result in transcription activation of genes found in this genomic region.

of a methyl group is mediated by the writer enzyme called histone methyltransferases (HMT).⁹⁵ This enzyme can either activate or further repress transcription, depending on the histone tail and subsequently the amino acid that is being methylated and the presence of other methyl or acetyl groups in the vicinity.⁹⁴ Acetyl and methyl groups can be removed by ‘eraser’ enzymes called histone deacetylases (HDACs) and histone demethylases (HDMTs), respectively.⁹⁵ Moreover, ‘reader’ enzymes such as bromodomains, can recognize histone modifications and direct a specific transcriptional outcome by modifying chromatin structure or recruiting machinery involved in gene expression.⁹⁶ Due to their influence on chromatin structure and transcription, drugs targeting enzymes capable of adding (writers), removing (erasers) and recognizing (readers) major enhancer-associated histone modifications could be a promising therapy for reversing aberrant DRE activity seen in the context of disease. One such histone modification which has shown great promise in the treatment of IBD is histone acetylation. *Table 1* presents the available compounds with clinical potential.

HDAC-inhibitors

The first potential targets are HDACs. Studies in both murine and human immune cells and colonic mucosa have shown that a lack of DRE-acetylation of immune genes is involved in the pathogenesis of IBD.⁹⁷ Inhibitors of HDACs increase the levels of acetylation and administration to mice resulted in the amelioration of colitis, a reduction of pro-inflammatory cytokines and a decrease of migratory inflammatory cells in colonic mucosa.⁹⁸ There are 11 isoforms of HDACs, each of which can be inhibited by specific compounds (Table 1).⁹⁹ HDAC-inhibitors have shown to have multiple substrates that are not limited to HDACs. The effect of HDAC-inhibitors can therefore not be automatically and solely ascribed to their effect on the chromatin landscape.^{100,101} Here, we will review HDAC-inhibitors that affect the chromatin landscape and subsequently result in a putative beneficial effect in IBD. Many of these inhibitors are already being used in the clinic, mainly as anticancer treatment.¹⁰² One HDAC-inhibitor, the short chain fatty acid butyrate, is already used as treatment for IBD.¹⁰³ The putative mechanism is inhibition of HDAC9, which enhances histone H3 acetylation in the promoter region of *NOD2*.^{104,105} In both an *in vitro* study on human intestinal epithelial cells and in *in vivo* murine models of experimental colitis, this drug increased *Nod2* expression that was associated with reduced nuclear transcription factor *Nfkb* signaling, reduction of inflammation and improved integrity of the intestinal epithelium.^{104,105} The effect of butyrate on intestinal epithelial cell proliferation was shown to be due to its effect on HDAC-activity and independent of its potential to target G-protein coupled receptors.¹⁰⁶

Table 1. Therapeutics that target chromatin landscape and key regulators.

Overview of putative and novel therapeutics for IBD based on their potential to target chromatin modifiers or key regulators. Clinicaltrials.gov was consulted for data on clinical trails and phases that are executed for each listed compound. We define IBD-related pre-clinical data as studies performed on IBD related cells or tissues. Abbreviations: MM: Multiple myeloma, CTCL: cutaneous T-cell lymphoma, AML: Acute Myeloid Leukemia, MDS: Myelodysplastic Syndrome, SMA: Spinal Muscular Atrophy, ALPS: Autoimmune Lymphoproliferative Syndrome.

	Putative effect in IBD	Compound	Target	IBD-related pre-clinical studies	Clinical trials (phase)
HDAC-inhibitors	Promote acetylation in DREs leading to increased transcription of genes in murine and human immune cells and colonic mucosa associated with reduction of disease severity and inflammation	Sodium butyrate (NaB)	isoform 1-5, 7-9 ¹³⁸	Human ¹³⁹ , mouse ^{105,140,141}	Ulcerative Colitis, Shigellosis (II)
		Valproic acid (VPA)	isoform 1-3, 8 ¹¹¹	Human ^{103,109} , mouse ^{98,107}	Solid tumors (II), Alzheimer's Disease (I), ALPS (II), Schizophrenia (III), Glycogen Storage Disease Type V (II), SMA (II), Supranuclear Palsy (II), Bipolar disorder (III), HIV (I), Leukemia (II)
		Tacedinaline	isoform 1-3 ¹⁴²	Human ¹⁴²	MM(II), Solid tumors (III), MDS (II)
		Quisinostat	isoform 1-11 ¹¹¹	Human ¹⁴³	Leukemia (I), MM (I), Lymphoma (I), Solid tumors (II)
		trichostatin A (TSA)	isoform 1,3, 4, 6, 10	Mouse ^{108,144,145}	-
		Vorinostat/suberoyl-anilide hydroxamic acid (SAHA)	isoform 1-11 ¹¹¹	Human, mouse ^{98,107}	CTCL (II), advanced clinical stages for anticancer treatment (Ali, 2018)
		Givinostat	isoform 1-10	Mouse ^{108,144-146}	Leukemia (II), MM (II), CTCL (III), systemic juvenile idiopathic arthritis (EMA), polycythaemia vera (EMA)
		Entinostat	isoform 1-3, 9 ¹¹¹	Human ^{103,109} , mouse ¹⁴⁶	Leukemia (I), solid tumors (II)
		Panobinostat	isoform 1-11 ¹¹¹	Human ¹⁴⁷ , mouse ^{108,144,145}	HIV(II), AML(I), Lymphoma(III), MM(II), Leukemia(I), CTCL(III), Solid tumors(II)
		Resminostat	isoform 1, 3, 6, 8 ¹¹¹	-	Lymphoma(II), Solid tumors(II)
		Mocetinostat	isoform 1-5, 9-11 ¹¹¹	Guinea pig ¹⁴⁸ , human ¹⁴⁹	MDS(II), Leukemia(II), Lymphoma(II), Solid tumors (II)
		Abexinostat	isoform 1-3, 6, 10 ¹¹¹	Human ¹⁵⁰	Lymphoma(II), Solid tumors(II)
		Pracinostat	isoform 1-4, 7-11 ¹¹¹	Human ¹⁵¹	MDS(II), Myeloproliferative Disease(II), Leukemia(I), Solid tumors(I)
		Belinostat	isoform 1-4, 6-9 ¹¹¹	Human ¹⁵²	Lymphoma(II), Leukemia(II), MDS(II), MM(II), Solid tumors(II)
		Tubastatin A	isoform 6 ¹⁵³	Human ¹⁵⁴ , mouse ¹⁵⁵	-
		Tubacin	isoform 6 ¹⁵⁶	Mouse ¹⁵⁵	-
Santacruzamate A	isoform 2, 4, 6 ¹⁴²	Human ¹⁴²	-		
Romidepsin	isoform 1,2 ¹⁴²	Human ¹⁴²	Lymphoma(II), Solid tumors(II)		
Abexinostat	isoform 2, 3, 6, 10	Human ¹⁵⁰	Lymphoma(II), MM(II) Leukemia(II) Solid tumors(II)		
CUDC-101	isoform 1-10	-	Solid tumors(I)		
BET-inhibitors	Preferential inhibition of inflammation-associated gene expression involved in proinflammatory activity of murine and mouse monocyte, macrophages, and T lymphocytes	JQ1	BET	Human ^{114,157} , mouse ¹¹⁴	None
		I-BET762	BET	Mouse ¹¹⁴	Solid tumors(II)
		I-BET151	BET	Mouse ¹¹⁴	-
		ZEN3694	BET	-	Solid tumors (II)
		INCB054329	BET	-	Solid tumors and Hematologic malignancies(II)
		BMS-986158	BET	-	Solid tumors(II)
		FT-1101	BET	-	Leukemia(I), MDS(I) Lymphoma(I)
		RO6870810/TEN-010	BET	-	MM(I), Leukemia(I), MDS(I), Solid tumors(I)
		RVX000222	BET	Human ¹⁵⁸	Diabetes Mellitus Type 2(III), Cardiovascular Disease(III), Fabry Disease(I), Chronic Kidney Failure(I),
		CPI-0610	BET	-	MM(I), Lymphoma(I), Myelofibrosis(II), Solid tumors(II)
OTX015/MK-8628	BET	-	MDS(I), Lymphoma(I), Solid tumors(I)		
Methyltransferase inhibitors	Upregulation of FOXP3 to increase EZH2 expression	Decitabine (5-aza-2'-deoxycytidine or 5-aza-dC)	DNA Methyltransferase	Human ¹⁵⁹	-
		Tazemetostat	EZH2	Mouse ¹¹⁶	Lymphoma(II), Solid tumors(II)
		SHR2554	EZH2	-	Lymphoma(I)
		CPI-1205	EZH2	Human ¹⁶⁰	Lymphoma(I), Solid tumors(II)
Key regulator modulators	NF-κB inhibition	Dehydroxymethylepoxyquinomicin (DHMEQ)	NF-κB (blocks nuclear translocation)	Mouse ¹⁶¹	-
		Curcumin	NF-κB	Mouse ¹⁶²	IBD(III) (NCT00779493) ¹⁶³
	NF-κB decoy oligonucleotide	NF-κB	Rat ¹²⁶	Atopic Dermatitis(II)	
	HNF4a agonists show protective effect	C14-C18 Fatty acids	HNF4a	Crystal structure ^{123,124}	-
Inhibition of the JAK/STAT pathway	Tofacitinib	JAK1, JAK2, JAK3, TYK2	Human ¹⁶⁴	Crohn's Disease(II) ¹³¹ and Ulcerative Colitis(III) ¹⁶⁵ , Immune related diseases(III)	
	Peficitinib	JAK1, JAK3	Human ¹⁶⁴	Ulcerative Colitis(II) ¹⁶⁶ , Rheumatoid arthritis(II)	
	Upadacitinib	JAK1	Human ¹⁶⁴	Crohn's Disease(NCT02365649) (II)	
	Filgotinib	JAK1	Human ¹⁶⁴	Crohn's Disease(II) ¹⁶⁷ , Immune-related diseases(II)	

SAHA and VPA are HDAC-inhibitors that have passed to advanced clinical stages for anticancer treatment.¹⁰⁷ These inhibitors cause a dose-dependent increase in H₃ acetylation at the site of inflammation and are associated with macroscopic and histological reduction of disease severity as well as marked suppression in pro-inflammatory cytokine expression in the colon.^{98,107} This anti-inflammatory effect could be explained by HDAC inhibition in dendritic cells that results in decreased expression of inflammatory cytokines.^{108,109} Furthermore, the HDAC inhibitor givinostat is being studied in clinical trials for systemic-onset juvenile idiopathic arthritis and has obtained a good safety profile.^{110,111}

KAT2B is a lysine acetyltransferase that is down-regulated in inflamed colonic tissue of CD and UC patients. Inhibition of KAT2B by anacardic acid demonstrated reduced levels of histone H₄ lysine 5 acetylation (H₄K₅ac) in the interleukin-10 (*IL-10*) promoter region which was associated with a dose-dependent decrease in expression of *IL-10*.⁹⁷ HDAC1-selective inhibitors such as tacedinaline and quisinostat promote H₄K₅ acetylation and restore *IL-10* transcription.⁹⁷ The reduction of *IL-10* has been linked to IBD as patients with deleterious mutations in *IL10*, *IL10RA*, and *IL10RB* suffer from severe infantile-onset inflammatory bowel disease.¹¹²

Overall, many studies have demonstrated that these small molecules can target important pathways in the pathogenesis of IBD and their efficacy is currently being studied in clinical trials.

BET-inhibitors

BET (bromodomain and extra-terminal) proteins are 'reader' proteins which recognize acetylated lysine residues on regulatory elements and influence expression by recruiting transcription factors and chromatin remodeling factors such as the SWI/SNF complex.¹¹³ BET proteins can be specifically found at super-enhancers that are characterized by extensive acetylation of histone H₃ at lysine 27 and increased binding transcription factors. Super enhancers are highly cell type and cell state specific and might therefore be able to target disease-specific processes. After all, both healthy controls and IBD-patients may contain the same SNPs in an enhancer, but the enhancer may only be active due to tissue or cell context like inflammation and the presence of cytokines. Therefore, the enhancers that only become active in certain contexts are more likely to have an influence on pathogenesis of a disease and thus it is beneficial that BET-inhibitors are studied in the context of inflammatory disease. There is evidence that inhibiting BET proteins leads to a reduction in inflammation-associated gene expression and can modulate the pro-inflammatory activity of adaptive immune cells.¹¹⁴ For example, the BET-inhibitors JQ1, I-BET762, and I-BET151 were shown to reduce the production of pro-inflammatory cytokines in monocytes and macrophages stimulated by LPS *in vitro* and in mice.¹¹⁴ Studies on human inflamed joint synovial fluid show that treatment with JQ1 on CD4⁺ memory/effector T cells preferentially inhibited the expression of genes involved

in pro-inflammatory and cytokine-related processes regulated by super enhancers.¹¹⁵ Although BET-inhibitors have not been studied in the context of IBD, their immunomodulating potential points to promising effects for IBD treatment.

MT-inhibitors

The methyltransferase *EZH2* (Enhancer of Zeste homolog 2) is an important player in IBD. Patients show significantly reduced expression and it has been shown in mice that inhibition of *EZH2* leads to increased immune responsiveness that is associated with an increased sensitivity to DSS- and 2,4,6-trinitrobenzene sulfonic acid (TNBS) induced experimental colitis.¹¹⁶ In addition, mice that lack *EZH2* specifically in regulatory T cells (Tregs) develop spontaneous IBD. *EZH2* functions as a cofactor of FOXP3 for the regulation of Treg-specific gene networks. As dysregulation of *EZH2* plays a role in the development of IBD in both mice and humans, treatment that restores histone methyltransferase activity in Tregs could be beneficial for treating IBD.¹¹⁶

Targeting key regulators

Inflammatory pathways that play a role in the pathogenesis of IBD can be targeted through cytokine inhibitors. This way, the cascade that results from cytokine-receptor binding can be blocked after which inflammation is dampened. Several monoclonal antibodies have been developed and have become common therapeutics used in IBD. The majority of clinically approved antibodies target TNF α (Tumor Necrosis Factor α), a cytokine that can be produced by many cell types and is found to be upregulated in IBD.¹¹⁷ Other cytokines including IFN γ , IL-6, TGF β and IL-12/p40 are being studied as therapeutic targets in IBD with varying degrees of success.^{118,119} However, these targets may not completely represent the common pathways that have been identified through integrating genetic and epigenetic data.

As described above, recently, many attempts have been done to identify key regulators in the pathogenesis of IBD. The goal of these approaches is to identify single players preferably affecting multiple IBD genes at the same time.^{85,90} Many of these key regulators turn out to be transcription factors. This can be explained by both the many associated loci that contain the same binding motifs for a limited number of TFs and by the involvement of pathways that are regulated by a limited number of TFs. Alterations in transcription factor activity therefore can play a central role in the IBD pathogenesis and targeting these factors seems a valid approach. However, TFs have proved to be challenging targets and have often been termed ‘undruggable’.⁹⁵ Nevertheless, several strategies for TF-targeting are being developed (**Table 1**).

HNF4A

We and others have identified *HNF4A* as an important key regulator in IBD that interferes with pathways in multiple relevant cell types.^{85,90} *HNF4A* provides an interesting target, as there are multiple levels in which *HNF4A* contributes to IBD: enrichment of IBD-SNPs in *HNF4A* binding sites, altered DNA binding¹²⁰, regulation of expression of IBD

candidate genes⁸⁵ and differential expression of *HNF4A* itself. Notably, *Hfn4a* knock-out mice are prone to develop DSS-induced colitis.¹²¹ There are currently no HNF4A agonists ready to be tested in clinical studies. Conversely, multiple HNF4A inhibitors are used *in vitro*.¹²² However, HNF4A seems to have a protective effect in IBD, therefore agonists rather than antagonists are of interest as putative therapeutic compounds. Identification of HNF4A agonists has proven difficult and there are currently no compounds available that upregulate *HNF4A*. Nevertheless, studies on the *in vivo* ligands of HNF4A reveal that medium and long chain fatty acids are the natural ligands and activators of HNF4A.^{123–125}

NFKB

As for agonists of HNF4A, finding specific antagonists for another IBD-key regulator, NFKB, has been a challenge. For NFKB, an alternative TF targeting approach has shown success in this context. By using decoy oligonucleotides, NFKB-DNA binding is decreased through competitive inhibition. Decoy oligonucleotides are short, double-stranded DNA molecules containing the binding motif of a specific transcription factor. When TFs bind to decoy oligonucleotides their availability for binding to DNA regulatory elements decreases. A major limitation of this technique is that decoy oligonucleotides are easily degraded by nucleases. However, recent developments have increased their stability and made them more nuclease-resistant. Administration of a NFKB decoy oligonucleotide in rats limited the expression of pro-inflammatory pathways and showed increased survival rate upon inducing DSS-induced colitis.¹²⁶ Another NFKB inhibitor, dehydroxymethylepoxyquinomicin (DHMEQ), blocks the nuclear translocation of NFKB and ameliorated experimental TNBS and DSS-induced colitis in mice.¹²⁷ Finally, the natural NFKB inhibitor curcumin has shown positive effects on the treatment of IBD in phase I, II and III trials.¹²⁸

STAT3

STAT3 is a key transcription factor in the pathogenesis of IBD. The nuclear translocation of STAT3 is mediated by Janus kinase (JAK).¹²⁹ Therefore, inhibition of JAKs subsequently leads to the inhibition of STAT3 and thereby of the downstream JAK/STAT pathway. Upon the discovery of a role for STAT3 in IBD pathogenesis, JAK-inhibitors have been developed and tested in clinical trials. To date, such drugs have not been proven efficacious in CD.^{130,131} However, JAK-inhibitors are currently used for the treatment of UC.^{132,133}

Conclusion and discussion

Although genome-wide association studies of IBD have revealed many associated sequence variants and some involved genes, these findings only explain a minor portion of the genetic background of IBD.^{8,19} Therefore, new insights into the genetic makeup of these complex genetic diseases is needed. We have reviewed the accumulating evidence regarding the contribution of sequence variants in DNA regulatory elements

to the pathogenesis of inflammatory bowel disease. This contribution is supported by many studies and steps are being taken to translate these findings into new diagnostic, preventive and therapeutic measures.

Studying genetic variants in regulatory elements and their effect on the pathogenesis of complex genetic diseases remains challenging. The context specific activity of DRE makes their identification and annotation of DRE difficult. The limited effect of SNPs in DRE compared to the effect of variants that alter gene coding sequences, further complicates the identification of pathogenic regulatory variants. Over the last years, multiple high-throughput techniques have been developed that enable the efficient annotation of DRE and the evaluation of allele-specific effects. These techniques are built on epigenetic signatures, chromosome conformation, genome editing and self-transcribing activity of regulatory elements and have been thoroughly reviewed by Elkon and Agami.¹³⁴ To assay the effects of genetic variants on the activity of regulatory region, the cell type specific activity of enhancers needs to be taken into account as the phenotype that is caused by a SNP might only be present in a limited number of cell types. Assaying a SNP in the wrong cell type or developmental stage will not reveal a deleterious effect. This is further complicated by the finding that the phenotype of some genetic variants is dependent on the cell state (for example activated cells vs non-activated cells).⁸² Similar as in eQTL-studies, the cell type, developmental stage and cell state dependent effect of SNPs on DRE causes a high chance of false negative detection. To increase detection of disease-associated SNPs assays should be performed in a multitude of cell types and conditions. To address this, eQTL-databases that contain a plethora of cell types and developmental stages are now being developed.¹³⁵

New therapeutic measures for IBD include targeting histone writers, readers and erasers, as well as key regulators of IBD networks. A possible drawback of targeting histone modifications and key regulators is that the effect of the compounds will not be limited to the tissues that are affected by the disease. However, the predictive value of IBD-associated SNPs on the pathogenic cell types can be a lead to develop therapeutics that can be delivered to specific cells. This may be associated with adverse effects as is seen for many therapeutics that target general processes, including immune modulators and chemotherapeutics. The extent of the adverse effects and the efficacy of these putative new compounds is currently being studied in clinical trials. Although most trials concern malignant diseases, the outcome of the trials can be highly relevant for IBD (as for safety, dose ranging and side effects). In *Table 1* we reviewed multiple compounds that have shown to be effective *in vitro* and *in vivo* and are already being used in human diseases.

We believe that recent insights into the roles of DRE in IBD will result in the development of novel therapeutic strategies. With novel techniques like CRISPR/Cas9, that are now widely used for scientific purposes, genetic diseases may eventually be treated by restoring genetic defects. Although we reviewed the contribution of genetic variants to the pathogenesis of IBD, we have not discussed the therapeutic potential of genome

editing. As patients carry multiple SNPs that contribute to the pathogenesis of IBD, treatment of patients by targeting these common variants through genome editing does not seem an efficient approach. Even more so, because the loci, but not the disease-causing variants have been identified. However, profiling the disease-associated SNPs in individual patients could be useful in predicting response to therapeutics and thereby in defining a personalized therapeutic strategy. We therefore suggest including SNP-profiles in the analyses of clinical trials, to establish the variants that are predictive for treatment outcome.^{136,137}

In this review, we have shown that sequence variants in DNA regulatory elements are involved in the pathogenesis of inflammatory bowel disease. The majority of IBD-associated sequence variants co-localize with active DNA regulatory elements and the mechanisms through which these SNPs lead to pathogenic processes have been intensively studied over the last decade. This resulted in the identification of novel therapeutic targets including regulatory layers such as the chromatin landscape and key regulatory transcription factors.

References

1. Ruuska, T., Vaajalahti, P., Arajärvi, P. & Mäki, M. Prospective evaluation of upper gastrointestinal mucosal lesions in children with ulcerative colitis and crohn's disease. *J. Pediatr. Gastroenterol. Nutr.* 19, 181–186 (1994).
2. Halme, L. *et al.* Family and twin studies in inflammatory bowel disease. *World J. Gastroenterol.* 12, 3668–3672 (2006).
3. Halfvarson, J., Bodin, L., Tysk, C., Lindberg, E. & Järnerot, G. Inflammatory bowel disease in a Swedish twin cohort: A long- term follow-up of concordance and clinical characteristics. *Gastroenterology* 124, 1767–1773 (2003).
4. Orholm, M., Binder, V., Sorensen, T. I. A., Rasmussen, L. P. & Kyvik, K. O. Concordance of inflammatory bowel disease among Danish twins: Results of a nationwide study. *Scand. J. Gastroenterol.* 35, 1075–1081 (2000).
5. Thompson, N. P., Driscoll, R., Pounder, R. E., Wakefield, A. J. & Bowel, I. Genetics versus environment in inflammatory bowel disease: Results of a British twin study. *Br. Med. J.* 312, 95–96 (1996).
6. Cleynen, I. *et al.* Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: Results from the IBDchip European project. *Gut* 62, 1556–1565 (2013).
7. Tyler, A. D. *et al.* The NOD2zinc polymorphism is associated with worse outcome following ileal pouch-anal anastomosis for ulcerative colitis. *Gut* 62, 1433–1439 (2013).
8. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–24 (2012).
9. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42, 1118–1125 (2010).
10. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* 43, 246–52 (2011).
11. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* 39, 596–604 (2007).
12. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 3, e58 (2007).
13. Liu, J. Z. J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–989 (2015).
14. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* 43, 1066–73 (2011).
15. De Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 49, 256–261 (2017).
16. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* 547, 173–178 (2017).
17. Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics

- of immune-related diseases. *Nat. Rev. Genet.* 10, 43–55 (2009).
18. Mokry, M. *et al.* Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology* 146, 1040–1047 (2014).
 19. Barrett, J. C. J. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40, 955–62 (2008).
 20. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314, 1461–3 (2006).
 21. Hong, S. N. *et al.* Deep resequencing of 131 Crohn's disease associated genes in pooled DNA confirmed three reported variants and identified eight novel variants. *Gut* 65, 788–796 (2016).
 22. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144, 327–339 (2011).
 23. Ensembl. Human assembly and gene annotation. Available at: https://www.ensembl.org/Homo_sapiens/Info/Annotation. (Accessed: 3rd September 2018)
 24. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* 24, 344–352 (2008).
 25. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
 26. Peterson, T. A. *et al.* Regulatory Single-Nucleotide Variant Predictor Increases Predictive Performance of Functional Regulatory Variants. *Hum. Mutat.* 37, 1137–1143 (2016).
 27. Pazin, M. J. Using the encode resource for functional annotation of genetic variants. *Cold Spring Harb. Protoc.* 2015, 522–536 (2015).
 28. Rickels, R. & Shilatifard, A. Enhancer Logic and Mechanics in Development and Disease. *Trends Cell Biol.* 28, 608–630 (2018).
 29. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12, 1725–1735 (2003).
 30. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: Five essential questions. *Nat. Rev. Genet.* 14, 288–295 (2013).
 31. de Laat, W. *et al.* Three-Dimensional Organization of Gene Expression in Erythroid Cells. *Curr. Top. Dev. Biol.* 82, 117–139 (2008).
 32. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233–1244 (2012).
 33. Vakoc, C. R. *et al.* Proximity among distant regulatory elements at the β -globin locus requires GATA-1 and FOG-1. *Mol. Cell* 17, 453–462 (2005).
 34. Khan, A. & Zhang, X. DbSUPER: A database of Super-enhancers in mouse and human genome. *Nucleic Acids Res.* 44, D164–D171 (2016).
 35. Hnisz, D. *et al.* XSuper-enhancers in the control of cell identity and disease. *Cell* 155, (2013).
 36. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21931–21936 (2010).
 37. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318 (2007).
 38. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.* 20, 1147–1155 (2013).
 39. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539, 452–455 (2016).
 40. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* 30, 4198–4210 (2011).
 41. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46, 1311–1320 (2014).
 42. Marsh, M. N. Gluten, Major Histocompatibility and the Small Intestine Complex., 330–354 (1992).
 43. Kowalczyk, M. S. *et al.* Intragenic Enhancers Act as Alternative Promoters. *Mol. Cell* 45, 447–458 (2012).
 44. Kraiczy, J. *et al.* DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. *Gut* 1–13 (2017). doi:10.1136/gutjnl-2017-314817
 45. Vahedi, G. *et al.* STATs shape the active enhancer landscape of T cell populations. *Cell* 151, 981–993 (2012).
 46. Sakabe, N. J., Savic, D. & Nobrega, M. A. Transcriptional enhancers in development and disease. *Genome Biol.* 13, (2012).
 47. Spedale, G. *et al.* Tight cooperation between Motif and NC2 β in regulating genome-wide transcription, repression of transcription following heat shock induction and genetic interaction with SAGA. *Nucleic Acids Res.* 40, 996–1008 (2012).
 48. Bajic, V. B., Sin, L. T., Suzuki, Y. & Sugano, S. Promoter prediction analysis on the whole human genome.

- Nat. Biotechnol.* 22, 1467–1473 (2004).
49. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112 (2009).
 50. Glas, J. *et al.* PTGER4 Expression-Modulating Polymorphisms in the 5p13.1 Region Predispose to Crohn's Disease and Affect NF- κ B and XBP1 Binding Sites. *PLoS One* 7, (2012).
 51. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associate Variation in Regulatory DNA. *Science (80-.)*. 337, 1190–1195 (2012).
 52. Chiba, H. *et al.* Allele-specific DNA methylation of disease susceptibility genes in Japanese patients with inflammatory bowel disease. *PLoS One* 13, (2018).
 53. Ventham, N. T. *et al.* Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat. Commun.* 7, (2016).
 54. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* 50, 424–431 (2018).
 55. Izzi, B. *et al.* Allele-specific DNA methylation reinforces PEAR1 enhancer activity. *Blood* 128, 1003–1012 (2016).
 56. Zhang, P. *et al.* High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat. Commun.* 9, (2018).
 57. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342, 744–7 (2013).
 58. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–9 (2013).
 59. Li, Y. I. *et al.* Between Genetic Variation and Disease. *Science (80-.)*. 352, 600–4 (2016).
 60. Schaefer, J. S. MicroRNAs: How many in inflammatory bowel disease? *Curr. Opin. Gastroenterol.* 32, 258–266 (2016).
 61. Zacharopoulou, E., Gazouli, M., Tzouvala, M., Vezakis, A. & Karamanolis, G. The contribution of long non-coding RNAs in Inflammatory Bowel Diseases. *Dig. Liver Dis.* 49, 1067–1072 (2017).
 62. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* 46, 61–64 (2014).
 63. Kleinjan, D. J. & Coutinho, P. Cis-rupture mechanisms: Disruption of cis-regulatory control as a cause of human genetic disease. *Briefings Funct. Genomics Proteomics* 8, 317–332 (2009).
 64. Schaub, M. a, Boyle, A. P., Kundaje, A. & Frazer, K. a. Linking disease associations with regulatory information in the human genome Toward mapping the biology of the genome. 1748–1759 (2012). doi:10.1101/gr.136127.111
 65. Weiser, M. *et al.* Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut* 67, 36–42 (2018).
 66. Boyd, M. *et al.* Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nat. Commun.* 9, (2018).
 67. Parkes, M. *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* 39, 830–2 (2007).
 68. Prescott, N. J. *et al.* Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum. Mol. Genet.* 19, 1828–1839 (2010).
 69. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* 40, 1107–1112 (2008).
 70. Singh, S. B., Davis, A. S., Taylor, G. A. & Deretic, V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science (80-.)*. 313, 1438–1442 (2006).
 71. Brest, P. *et al.* A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat. Genet.* 43, 242–245 (2011).
 72. Kaser, A. *et al.* XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease. *Cell* 134, 743–756 (2008).
 73. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* 165, 1519–1529 (2016).
 74. Neurath, M. F. *et al.* The Transcription Factor T-bet Regulates Mucosal T Cell Activation in Experimental Colitis and Crohn's Disease. *J. Exp. Med.* 195, 1129–1143 (2002).
 75. Soderquest, K. *et al.* Genetic variants alter T-bet binding and gene expression in mucosal inflammatory disease. *PLoS Genet.* 13, (2017).
 76. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* 21, 48–54 (2013).
 77. Dumitru, C. D. *et al.* TNF- α induction by LPS is regulated posttranscriptionally via a Tpl2/ERK-dependent pathway. *Cell* 103, 1071–1083 (2000).

78. Kontoyiannis, D. *et al.* Genetic Dissection of the Cellular Pathways and Signaling Mechanisms in Modeled Tumor Necrosis Factor–induced Crohn’s-like Inflammatory Bowel Disease. *J. Exp. Med.* 196, 1563–1574 (2002).
79. Lawrenz, M. *et al.* Genetic and pharmacological targeting of TPL-2 kinase ameliorates experimental colitis: A potential target for the treatment of Crohn’s disease. *Mucosal Immunol.* 5, 129–139 (2012).
80. Hedl, M. & Abraham, C. A TPL2 (MAP3K8) disease-risk polymorphism increases TPL2 expression thereby leading to increased pattern recognition receptor-initiated caspase-1 and caspase-8 activation, signalling and cytokine secretion. *Gut* 65, 1799–1811 (2016).
81. Marigorta, U. M. *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn’s disease. *Nat. Genet.* 49, 1517–1521 (2017).
82. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* (80-.). 343, 1246949 (2014).
83. Momozawa, Y. *et al.* IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* 9, (2018).
84. Glas, J. *et al.* PTPN2 gene variants are associated with susceptibility to both Crohn’s disease and ulcerative colitis supporting a common genetic disease background. *PLoS One* 7, (2012).
85. Meddens, C. A. *et al.* Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease. *Genome Biol.* (2016). doi:10.1186/s13059-016-1100-3
86. Haitjema, S. *et al.* Additional Candidate Genes for Human Atherosclerotic Disease Identified Through Annotation Based on Chromatin Organization. *Circ. Cardiovasc. Genet.* 10, (2017).
87. Brandt, M. M. *et al.* Chromatin Conformation Links Distal Target Genes to CKD Loci. *J. Am. Soc. Nephrol.* 29, 462–476 (2018).
88. Stappenbeck, T. S. *et al.* Crohn disease: A current perspective on genetics, autophagy and immunity. *Autophagy* 7, 355–374 (2011).
89. Glocker, E.-O. *et al.* Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N. Engl. J. Med.* 361, 2033–45 (2009).
90. Peters, L. A. *et al.* A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nat. Genet.* 49, 1437–1449 (2017).
91. Chahar, S. *et al.* Chromatin profiling reveals regulatory network shifts and a protective role for hepatocyte nuclear factor 4 α during colitis. *Mol. Cell. Biol.* 34, 3291–304 (2014).
92. Raine, T., Liu, J. Z., Anderson, C. A., Parkes, M. & Kaser, A. Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease. *Gut* 64, 250–259 (2015).
93. Pradeepa, M. M. Causal role of histone acetylations in enhancer function. *Transcription* 8, 40–47 (2017).
94. Peeters, J. G. C., Vastert, S. J., van Wijk, F. & van Loosdregt, J. Review: Enhancers in Autoimmune Arthritis: Implications and Therapeutic Potential. *Arthritis Rheumatol. (Hoboken, N.J.)* 69, 1925–1936 (2017).
95. Johnston, S. J. & Carroll, J. S. Transcription factors and chromatin proteins as therapeutic targets in cancer. *BBA - Rev. Cancer* 1855, 183–192 (2015).
96. Gillette, T. G. & Hill, J. A. Readers, writers, and erasers: chromatin as the whiteboard of heart disease. *Circ. Res.* 116, 1245–53 (2015).
97. Bai, A. H. C. *et al.* Dysregulated Lysine Acetyltransferase 2B Promotes Inflammatory Bowel Disease Pathogenesis Through Transcriptional Repression of Interleukin-10. *J. Crohn’s Colitis* 10, 726–734 (2016).
98. Glauben, R. *et al.* Histone hyperacetylation is associated with amelioration of experimental colitis in mice. *J. Immunol.* 176, 5015–22 (2006).
99. Ventham, N. T., Kennedy, N. A., Nimmo, E. R. & Satsangi, J. Beyond Gene Discovery in Inflammatory Bowel Disease: The Emerging Role of Epigenetics. *Gastroenterology* 145, 293–308 (2013).
100. Markozashvili, D. *et al.* Histone deacetylase inhibitor abexinostat affects chromatin organization and gene transcription in normal B cells and in mantle cell lymphoma. *Gene* 580, 134–143 (2016).
101. Solomon, J. M. *et al.* Inhibition of SIRT1 Catalytic Activity Increases p53 Acetylation but Does Not Alter Cell Survival following DNA Damage Inhibition of SIRT1 Catalytic Activity Increases p53 Acetylation but Does Not Alter Cell Survival following DNA Damage. *Mol. Cell. Biol.* 26, 28–38 (2006).
102. Holtzman, L. & Gersbach, C. A. Editing the Epigenome: Reshaping the Genomic Landscape. *Annu. Rev. Genom. Hum. Genet* 1918, (2018).
103. Frikeche, J. *et al.* Impact of HDAC inhibitors on dendritic cell functions. *Exp. Hematol.* 40, 783–91 (2012).
104. Lee, C. *et al.* Sodium butyrate inhibits the NF-kappa B signaling pathway and histone deacetylation, and attenuates experimental colitis in an IL-10 independent manner. *Int. Immunopharmacol.* 51, 47–56 (2017).
105. Simeoli, R. *et al.* An orally administered butyrate-releasing derivative reduces neutrophil recruitment and inflammation in dextran sulphate sodium-induced murine colitis. *Br. J. Pharmacol.* 174, 1484–1496 (2017).
106. Kaiko, G. E. *et al.* The Colonic Crypt Protects Stem Cells from Microbiota-Derived Metabolites. *Cell* 167, 1137 (2016).

107. Ali, M. N. *et al.* The HDAC Inhibitor, SAHA, Prevents Colonic Inflammation by Suppressing Pro-inflammatory Cytokines and Chemokines in DSS-induced Colitis. *ACTA Histochem. Cytochem.* 51, 33–40 (2018).
108. Reddy, P. *et al.* Histone deacetylase inhibition modulates indoleamine 2,3-dioxygenase-dependent DC functions and regulates experimental graft-versus-host disease in mice. *J. Clin. Invest.* 118, 2562–73 (2008).
109. Nencioni, A. *et al.* Histone deacetylase inhibitors affect dendritic cell differentiation and immunogenicity. *Clin. Cancer Res.* 13, 3933–41 (2007).
110. Vojinovic, J. *et al.* Safety and efficacy of an oral histone deacetylase inhibitor in systemic-onset juvenile idiopathic arthritis. *Arthritis Rheum.* 63, 1452–1458 (2011).
111. Felice, C., Lewis, A., Armuzzi, A., Lindsay, J. O. & Silver, A. Review article: selective histone deacetylase isoforms as potential therapeutic targets in inflammatory bowel diseases. *Aliment. Pharmacol. Ther.* 41, 26–38 (2015).
112. Shouval, D. S. *et al.* Enhanced TH17 Responses in Patients with IL10 Receptor Deficiency and Infantile-onset IBD. *Inflamm. Bowel Dis.* 23, 1950–1961 (2017).
113. Brown, J. D. *et al.* NF- κ B directs dynamic super enhancer formation in inflammation and atherogenesis. *Mol. Cell* 56, 219–231 (2014).
114. Tough, D. F. & Prinjha, R. K. Immune disease-associated variants in gene enhancers point to BET epigenetic mechanisms for therapeutic intervention. *Epigenomics* 9, 573–584 (2017).
115. Peeters, J. G. C. *et al.* Inhibition of Super-Enhancer Activity in Autoinflammatory Site-Derived T Cells Reduces Disease-Associated Gene Expression. *Cell Rep.* 12, 1986–1996 (2015).
116. Sarmento, O. F. *et al.* The Role of the Histone Methyltransferase Enhancer of Zeste Homolog 2 (EZH2) in the Pathobiological Mechanisms Underlying Inflammatory Bowel Disease (IBD). *J. Biol. Chem.* 292, 706–722 (2017).
117. Rutgeerts, P., Van Assche, G. & Vermeire, S. Optimizing Anti-TNF treatment in inflammatory bowel disease. *Gastroenterology* 126, 1593–1610 (2004).
118. Caprioli, F., Caruso, R., Sarra, M., Pallone, F. & Monteleone, G. Disruption of inflammatory signals by cytokine-targeted therapies for inflammatory bowel diseases. *Br. J. Pharmacol.* 165, 820–828 (2012).
119. Sedda, S., Marafini, I., Dinallo, V., Di Fusco, D. & Monteleone, G. The TGF- β /Smad System in IBD Pathogenesis. *Inflamm. Bowel Dis.* 21, 2921–2925 (2015).
120. Chahar, S. *et al.* Chromatin profiling reveals regulatory network shifts and a protective role for hepatocyte nuclear factor 4 α during colitis. *Mol. Cell Biol.* 34, 3291–304 (2014).
121. Ahn, S.-H. *et al.* Hepatocyte nuclear factor 4 α in the intestinal epithelial cells protects against inflammatory bowel disease. *Inflamm. Bowel Dis.* 14, 908–920 (2008).
122. Kiselyuk, A. *et al.* HNF4 α antagonists discovered by a high-throughput screen for modulators of the human insulin promoter. *Chem. Biol.* 19, 806–818 (2012).
123. Dhe-Paganon, S., Duda, K., Iwamoto, M., Chi, Y.-I. & Shoelson, S. E. Crystal structure of the HNF4 α ligand binding domain in complex with endogenous fatty acid ligand. *J. Biol. Chem.* 277, 37973–37976 (2002).
124. Wisely, G. B. *et al.* Hepatocyte nuclear factor 4 is a transcription factor that constitutively binds fatty acids. *Structure* 10, 1225–1234 (2002).
125. McIntosh, A. L., Petrescu, A. D., Hostetler, H. A., Kier, A. B. & Schroeder, F. Liver-type fatty acid binding protein interacts with hepatocyte nuclear factor 4 α . *FEBS Lett.* 587, 3787–3791 (2013).
126. Ozaki, K. *et al.* Therapeutic effect of ribbon-type nuclear factor- κ B decoy oligonucleotides in a rat model of inflammatory bowel disease. *Curr. Gene Ther.* 12, 484–92 (2012).
127. Funakoshi, T. *et al.* A novel NF- κ B inhibitor, dehydroxymethylepoxyquinomicin, ameliorates inflammatory colonic injury in mice. *J. Crohn's Colitis* 6, 215–225 (2012).
128. Brumatti, L. *et al.* Curcumin and Inflammatory Bowel Disease: Potential and Limits of Innovative Treatments. *Molecules* 19, 21127–21153 (2014).
129. Shuai, K. & Liu, B. Regulation of JAK-STAT signalling in the immune system. *Nat. Rev. Immunol.* 3, 900–911 (2003).
130. Sandborn, W. J. *et al.* A phase 2 study of Tofacitinib, an oral janus kinase inhibitor, inpatients with crohn's disease. *Clin. Gastroenterol. Hepatol.* 12, (2014).
131. Panés, J. *et al.* Tofacitinib for induction and maintenance therapy of Crohn's disease: Results of two phase IIb randomised placebo-controlled trials. *Gut* 66, 1049–1059 (2017).
132. Sandborn, W. J. *et al.* Tofacitinib, an oral Janus kinase inhibitor, in active ulcerative colitis. *N. Engl. J. Med.* 367, 616–624 (2012).
133. Sandborn, W. J. *et al.* Tofacitinib as induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* 376, 1723–1736 (2017).
134. Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. *Nat. Biotechnol.* 35, 732–746 (2017).

135. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–5 (2013).
136. López-Hernández, R. *et al.* Genetic polymorphisms of tumour necrosis factor alpha (TNF- α) promoter gene and response to TNF- α inhibitors in Spanish patients with inflammatory bowel disease. *Int. J. Immunogenet.* 41, 63–68 (2014).
137. Lacruz-Guzmán, D. *et al.* Influence of polymorphisms and TNF and IL1 β serum concentration on the infliximab response in Crohn's disease and ulcerative colitis. *Eur. J. Clin. Pharmacol.* 69, 431–438 (2013).
138. Davie, J. R. Inhibition of Histone Deacetylase Activity by Butyrate. *J. Nutr.* 133, 2485S–2493S (2003).
139. Brogdon, J. L. *et al.* Histone deacetylase activities are required for innate immune cell control of Th1 but not Th2 effector cell function. *Blood* 109, 1123–30 (2007).
140. Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* 504, 446–450 (2013).
141. Arpaia, N. *et al.* Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature* 504, 451–455 (2013).
142. Zhou, H. *et al.* Pharmacological or transcriptional inhibition of both HDAC1 and 2 leads to cell cycle blockage and apoptosis via p21^{Waf1/Cip1} and p19^{INK4d} upregulation in hepatocellular carcinoma. *Cell Prolif.* 51, e12447 (2018).
143. Méhul, B. *et al.* Mass spectrometry and DigiWest technology emphasize protein acetylation profile from Quisinosat-treated HuT78 CTCL cell line. *J. Proteomics* (2018). doi:10.1016/j.jpro.2018.07.003
144. Jung, I. D. *et al.* Apicidin, the histone deacetylase inhibitor, suppresses Th1 polarization of murine bone marrow-derived dendritic cells. *Int. J. Immunopathol. Pharmacol.* 22, 501–15 (2009).
145. Bode, K. A. *et al.* Histone deacetylase inhibitors decrease Toll-like receptor-mediated activation of proinflammatory gene expression by impairing transcription factor recruitment. *Immunology* 122, 596–606
146. Glauben, R., Sonnenberg, E., Wetzel, M., Mascagni, P. & Siegmund, B. Histone deacetylase inhibitors modulate interleukin 6-dependent CD4⁺ T cell polarization in vitro and in vivo. *J. Biol. Chem.* 289, 6142–51 (2014).
147. Song, W. *et al.* HDAC inhibition by LBH589 affects the phenotype and function of human myeloid dendritic cells. *Leukemia* 25, 161–8 (2011).
148. Assem, E.-S. K. *et al.* Effects of a selection of histone deacetylase inhibitors on mast cell activation and airway and colonic smooth muscle contraction. *Int. Immunopharmacol.* 8, 1793–1801 (2008).
149. Sikandar, S. *et al.* The Class I Hdac Inhibitor Mgcdo103 Induces Cell Cycle Arrest and Apoptosis in Colon Cancer Initiating Cells by Upregulating Dickkopf-1 and Non-Canonical Wnt Signaling. *Oncotarget* 1, 596–605 (2010).
150. Park, J. M., Huang, S., Tougeron, D. & Sinicrope, F. A. MSH3 Mismatch Repair Protein Regulates Sensitivity to Cytotoxic Drugs and a Histone Deacetylase Inhibitor in Human Colon Carcinoma Cells. *PLoS One* 8, e65369 (2013).
151. Yong, W. P. *et al.* Phase I and pharmacodynamic study of an orally administered novel inhibitor of histone deacetylases, SB939, in patients with refractory solid malignancies. *Ann. Oncol.* 22, 2516–2522 (2011).
152. Campbell, G. R., Bruckman, R. S., Chu, Y.-L. & Spector, S. A. Autophagy induction by histone deacetylase inhibitors inhibits HIV type 1. *J. Biol. Chem.* 290, 5028–40 (2015).
153. Butler, K. V. *et al.* Rational Design and Simple Chemistry Yield a Superior, Neuroprotective HDAC6 Inhibitor, Tubastatin A. *J. Am. Chem. Soc.* 132, 10842–10846 (2010).
154. Vishwakarma, S. *et al.* Tubastatin, a selective histone deacetylase 6 inhibitor shows anti-inflammatory and anti-rheumatic effects. *Int. Immunopharmacol.* 16, 72–78 (2013).
155. de Zoeten, E. F. *et al.* Histone deacetylase 6 and heat shock protein 90 control the functions of Foxp3(+) T-regulatory cells. *Mol. Cell. Biol.* 31, 2066–78 (2011).
156. Haggarty, S. J., Koeller, K. M., Wong, J. C., Grozinger, C. M. & Schreiber, S. L. Domain-selective small-molecule inhibitor of histone deacetylase 6 (HDAC6)-mediated tubulin deacetylation. *Proc. Natl. Acad. Sci.* 100, 4389–4394 (2003).
157. Peeters, J. G. C. *et al.* Autoimmune disease-associated gene expression is reduced by BET-inhibition. *Genomics Data* 7, 14–17 (2016).
158. Lu, P. *et al.* BET inhibitors RVX-208 and PFI-1 reactivate HIV-1 from latency. *Sci. Rep.* 7, (2017).
159. Kim, S. W. *et al.* Genetic polymorphisms of IL-23R and IL-17A and novel insights into their associations with inflammatory bowel disease. *Gut* 60, 1527–36 (2011).
160. Goswami, S. *et al.* Modulation of EZH2 expression in T cells improves efficacy of anti-CTLA-4 therapy. *J. Clin. Invest.* 128, (2018).
161. Funakoshi, T. *et al.* A novel NF- κ B inhibitor, dehydroxymethylepoxyquinomicin, ameliorates inflammatory colonic injury in mice. *J. Crohn's Colitis* 6, 215–225 (2012).
162. Tambuwala, M. M. Natural Nuclear Factor Kappa Beta Inhibitors. *Inflamm. Bowel Dis.* 22, 719–723 (2016).
163. Suskind, D. L. *et al.* Tolerability of curcumin in pediatric inflammatory bowel disease: A forced-dose

- titration study. *J. Pediatr. Gastroenterol. Nutr.* 56, 277–279 (2013).
164. De Vries, L. C. S., Wildenberg, M. E., De Jonge, W. J. & D’Haens, G. R. The Future of Janus Kinase Inhibitors in Inflammatory Bowel Disease. *J. Crohn’s Colitis* 11, 885–893 (2017).
165. Panés, J. *et al.* Tofacitinib in patients with ulcerative colitis: Health-related quality of life in phase 3 randomised controlled induction and maintenance studies. *J. Crohn’s Colitis* 12, 145–156 (2018).
166. Sands, B. E. *et al.* Peficitinib, an Oral Janus Kinase Inhibitor, in Moderate-to-severe Ulcerative Colitis: Results From a Randomised, Phase 2 Study. 1–12 (2018). doi:10.1093/ecco-jcc/jjy085
167. Vermeire, S. *et al.* Clinical remission in patients with moderate-to-severe Crohn’s disease treated with filgotinib (the FITZROY study): results from a phase 2, double-blind, randomised, placebo-controlled trial. *Lancet* 389, 266–275 (2017).



Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium

3

Michal Mokry, Sabine Middendorp, Caroline L. Wiegerinck, Merlijn Witte, Hans Teunissen, Claartje A. Meddens, Edwin Cuppen, Hans Clevers, and Edward E. S. Nieuwenhuis

Based on: *Gastroenterology* 146(4):1040–47 2014

Introduction

IBD is a multifactorial disorder resulting from aberrant mucosal immune responses to environmental stimuli.¹ Population-based studies, ethnic differences and twin and family studies have established a role for genetics in IBD pathology.² In the search for genetic components of IBD, several risk alleles localized in protein coding genes have been identified, implicating altered functions of immune-cells and intestinal epithelial cells, in particular Paneth cells. As such, SNPs in the cytoplasmic pathogen recognition receptor NOD2 gene are associated with ileal Crohn's disease and Paneth cell dysfunction.³ In addition, deleterious mutations in several other genes including IL10R⁴, ADAM17⁵ and XIAP⁶ were shown to cause IBD. However, the majority of genetic variants involved in IBD are still unknown.

In search for causal variants for IBD, GWASs have identified numerous susceptibility regions that were marked by single nucleotide polymorphisms (SNPs).⁷⁻⁹ As nearly all of these SNPs are located in non-coding regions, they are generally considered as markers for the causal variants that are located in nearby genes. Therefore, downstream studies mainly focused on finding the disease-associated variants in the protein coding genes located within these susceptibility loci.¹⁰ It has been shown that up to 80% of non-coding DNA possesses functional genomic elements that can regulate expression of coding genes.¹¹ In addition, recent studies in various traits showed significant overlap of GWAS SNPs with gene promoters¹² and open chromatin^{13,14} that contain regions with potential regulatory function in a cell type-specific manner. This suggests that SNPs identified by GWASs for IBD that are not located within protein coding regions may contribute to the disease pathogenesis by affecting the function of non-coding DNA regulatory elements (DRE), such as distal enhancers (DE) or promoters. As activity of DRE has been shown to be highly cell specific¹⁵, we hypothesized that IBD-associated SNPs located in non-coding DNA regions are co-localizing with active regulatory elements in either immune cells or intestinal epithelial cells as these are the main cell types involved in initiating and maintaining the disease.

Materials and Methods

Human material

The procedures for obtaining human samples were approved by the Ethics Committee of the University Medical Centre Utrecht. Biopsies were obtained by ileo-colonoscopy performed as part of the standard diagnostics procedures in suspected IBD patients. Colon and terminal ileum samples were obtained from colectomy material.

Crypt isolation

Intestinal crypts were isolated as described previously.¹⁶ Details are given in the Supplementary materials and methods section.

Organoid cultures

Human organoids were maintained as described previously.¹⁶ Details are given in the Supplementary materials and methods section.

ChIP-seq

Chromatin IP was performed using MAGnify ChIP kit (Invitrogen). One μ l anti H3K27ac (ab4729, Abcam) was used per IP. Sequencing libraries from immunoprecipitated chromatin were prepared as described previously¹⁷ and sequenced on SOLiD 5500 or WildFire sequencer in a multiplexed way to produce 50-bp long reads. Cisgenome v.2¹⁸ software package was used for peak-calling from the ChIP-seq against the common input sample. H3K27 peaks from immune cells were called in the same way from publically available datasets^{19,20}. Datasets were submitted to GEO database with accession number: GSE51425. Details are given in the Supplementary materials and methods section.

Overlapping of SNPs with active DRE and DE

A SNP falling within DRE coordinates was considered as overlapping SNP. Random matched SNP sets (500 unique random SNPs for each GWAS SNP) were generated from variants present on Human Omni1S genotyping chip (Illumina). Random SNP selection was matched for similar minor allele frequency and distance to the closest gene. Linkage Disequilibrium (LD) information, specific to European population, was accessed for both GWAS SNPs and random matched SNPs from HapMap.²¹ To depict the significance of overlap of the regulatory elements with GWAS SNPs, we calculated the p-value using binominal cumulative distribution function $b(x; n, p)$ using R²² `pbinom()` function as described previously¹³. Details are given in the Supplementary materials and methods section.

Overlapping of DRE and DE with susceptibility loci

A susceptibility locus was defined as a 500,000 base pair long genomic region with the susceptibility SNP in the middle. DREs falling within susceptibility locus coordinates were considered as overlapping with the susceptibility locus. Details are given in the Supplementary materials and methods section.

Computational Data analysis

A combination of custom PERL and R scripts and Cisgenome functions were used for computational data analysis. Mapping of 143 JASPAR non-redundant motif matrices²³ was performed using Cisgenome v2.0 utilities (`-r 500`, Supplementary Fig 3DE). Permutation tests and Wilcoxon signed rank test with continuity correction were used to calculate p-values.

Animal husbandry

Zebrafish strains were maintained at the Hubrecht Institute using standard husbandry conditions. Experiments were performed in accordance with the animal research guidelines of The Royal Netherlands Academy of Arts and Sciences (KNAW).

In vivo zebrafish enhancer assay

Seven intestine-specific regulatory sequences were amplified by Phusion TAQ PCR (New England Biolabs) and cloned into the ZED vector²⁴ by using Gateway technology (Life Technologies BV). Constructs were injected in WT zebrafish embryos at the 1-cell stage with 1 nl of a final concentration of 25 ng/ μ l in the presence of 25 ng/ μ l TOL2 transposase RNA.²⁵ Embryos were kept at 28.5°C in E3 medium with phenylthiocarbamide (PTU) and scored for muscle-specific red fluorescent protein, and subsequently intestinal-specific GFP fluorescence at 5 days post fertilization (dpf) on a Leica M165FC fluorescence stereomicroscope (Leica Microsystems GmbH). In vivo imaging was carried out on a Leica AF7000 High Speed Fluorescence microscope mounted with a Leica DFC360 FX camera.

Reporter assays

Genomic regions of 1 – 1.4 kb wide, containing regulatory elements and susceptibility SNPs, were PCR amplified from human genomic DNA and cloned in front of a luciferase gene in pGL4.10 plasmid with a minimal TATA box promoter. Regulatory regions overlapping with promoters were cloned into pGL4.10 plasmid that lacks the minimal promoter. The constructs were transfected using pPEI or Fugene HD (Sigma) into HeLa, 293T, K562, DLD-1, Caco-2 and Jurkat cells with Renilla luciferase as a transfection control. The activity was measured using dual-luciferase reporter assay system (Promega) with a pGL4.10TATA vector serving as a negative control.

Results

Identification of active regulatory regions

First, we identified DRE in intestinal epithelium, as DRE in various immune cells were already publically available.^{19,20} Both active promoters and active enhancers are specifically marked by acetylated histone 3 lysine 27 (H3K27ac).^{15,26} Therefore, to identify epithelial DRE we performed H3K27ac chromatin immunoprecipitation and sequencing (ChIP-seq) on four intestinal crypt samples and three intestinal organoid^{16,27} cultures. The samples were derived from four different individuals, including individuals without diagnosed IBD. To increase the chance of identifying not only general intestinal regulatory elements, but also elements that are specifically active in the inflammatory state or at a specific intestinal location, we assayed representative samples derived from colon and ileum, inflamed and non-inflamed mucosa from freshly isolated crypts or primary organoid cultures (**Suppl. Table 1**). In total, we identified 25-43 thousand H3K27ac peaks in intestinal epithelial cells that were considered as active DRE (**Suppl. Table 1**). We show that H3K27ac profiles from intestine and immune cells cluster separately and contain different binding motifs for transcription factors (**Suppl. Fig 3.1AB**). Furthermore, only a small proportion (mainly promoter regions) of the identified intestinal DRE overlaps with DRE identified in immune cells (**Suppl. Fig 3.1C**); thereby confirming tissue specificity. From the active DRE, we regarded peaks that were located distal (> 2.5 kilobase) to annotated transcriptional start sites (TSS) as active distal enhancers (DE) and peaks <2.5 kb from TSS as active promoters (AP) (**Suppl. Fig 3.1D**).

Active regulatory regions overlap with IBD-associated SNPs

In order to determine the involvement of active chromatin regions in IBD, we mapped IBD-associated SNPs from a recent GWAS⁷ to the identified active regulatory elements in intestinal epithelium and immune cells (**Figure 1A, Suppl. Table 2**). The low variability in activity of DRE between the various intestinal and lymphoid samples indicates that only a minor part of DRE is influenced by disease type or inflammatory status of the cells and the vast majority of overlapping IBD-associated SNPs is located in DREs that are shared by multiple samples.

We found large overlap of DRE with numerous SNPs that can often be located within known binding motifs²⁸ (**Suppl. Table 2**). In this way, specific SNPs potentially affect the binding affinity of transcriptional regulators and thereby alter the expression of regulated

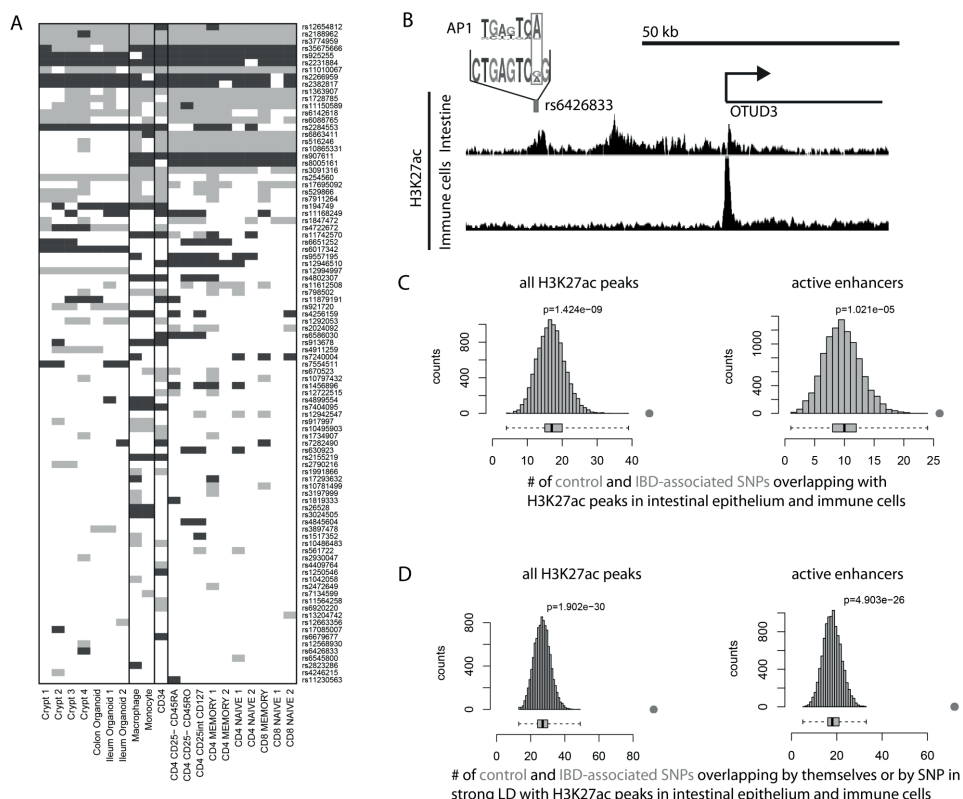


Figure 1. Overlap of IBD-associated SNPs with the regulatory elements in intestinal epithelium and immune cells.

A) Each rectangle depicts overlap between IBD-associated tag SNP (dark grey) or SNP in the strong LD ($r_2 > 0.8$) with the tag SNP (light grey) and DRE identified in the separate samples (in columns) B) Browser view of H3K27ac signal in intestinal epithelium and immune cells depicting the intestine-specific active regulatory element overlapping with a SNP involved in the IBD7. The rs6426833 variant alters the AP1 binding motif that has possible functional effect on this DRE28. C) Number of IBD-associated SNPs (dot) overlapping with the regulatory regions compared to 10,000 random SNP sets (grey bars) D) Number of IBD-associated SNPs (dot) overlapping by themselves or by SNP in strong LD ($r_2 > 0.8$) with the regulatory regions compared to 10,000 random SNP sets (grey bars). P-values in C) and D) were calculated with binominal cumulative distribution function.

genes as can be predicted by the RegulomeDB²⁸ (Suppl. Table 2). As an example, we show that one of the intestine-specific active DRE peaks, an active distal enhancer near the OTUD3 gene, is overlapping with the IBD-associated risk variant rs6426833 (Figure 1B).⁷ This specific SNP alters an AP1 binding motif that potentially affects transcription of one or more genes that are under control of this regulatory element. In addition, many other known IBD genes involved in intestinal inflammation, such as HNF4²⁹, harbor IBD-associated risk variants overlapping with DREs in their vicinity (Suppl. Fig 3.2).

Next, to prove the significance of this overlap, we show that active DRE, and more interestingly, numerous active DE in both intestinal epithelium and immune cells co-localized with recently identified IBD-associated SNPs⁷, more frequent than predicted

from random sampling (**Figure 1C**). The significance of this co-localization was further confirmed by including two other independent GWAS data sets^{8,9} and by using DRE from immune cells and intestine separately (**Suppl. Fig 3.3A**). Due to extensive linkage disequilibrium (LD) at each genomic locus in the human genome, many of the IBD-associated SNPs, presented on genotyping arrays (tag SNPs), are linked to other SNPs in their vicinity. In that case, the GWAS signal from tag SNPs can be caused by biological activity of other linked SNPs. To reflect this, we repeated our analysis by taking into account known SNPs from the 1000 Genomes³⁰ project which were in strong LD ($r_2 > 0.8$) with the tag SNP (**Figure 1D, Supplementary Figure 3.3A**). By this means, the significance of the co-localization even increased in comparison to the situation where SNPs in LD are ignored.

In total, we show that 92 out of 163 SNPs co-localized with active regulatory elements identified in intestinal epithelium or immune cells by the tag SNP itself or by a SNP in LD. This is 2.5 to 3.5 times more frequent than expected from random sampling (**Figure 1CD**). We also confirmed significance of this enrichment independently in all 20 separate samples (**Suppl. Fig 3.3BC**). In addition, we show that these DRE-overlapping IBD-associated loci, were more frequently co-localized with known transcription factor binding motifs (**Suppl Fig 3.3DE, Suppl Table 2**).

Since regulatory elements with the highest activity are responsible for the major transcriptional output in the cell^{31,32}, we also explored the activity status (measured by levels of H3K27ac signal) of overlapping DRE. Interestingly, DRE that overlap with IBD-associated loci contain higher levels of the H3K27ac mark compared to non-overlapping ones (**Figure 2, Suppl. Fig 3.4**). Altogether, the enrichment of IBD-associated SNPs in DRE with highest activity and tendency to affect binding sites suggest their functional involvement.

Besides H3K27ac, a variety of other histone marks and/or transcription factors can be used to identify regulatory elements. As an example: mono-methyl histone 3 lysine 4 (H3K4me1) is present on both active and poised enhancers^{15,26}, while histone acetyltransferases p300 and CREB binding protein (CBP), responsible for acetylation of H3K27, are present on a subset of active regulatory regions. Therefore, we examined the overlap of IBD-associated SNPs with numerous publically available ChIP-seq datasets for H3K4me1^{11,19,33}, p300^{11,34} and CBP³⁴⁻³⁶ from relevant cell types and tissues (**Suppl Fig. 3.5**). The overlap was highly significant in the majority of the cases, further suggesting the involvement of these specific regulatory sites in IBD.

Validation of identified DREs

To evaluate the activity of the tissue-specific DRE, we selected overlapping regions that were active in either intestinal epithelium or immune cells or in both and performed luciferase reporter assays. DRE activity was defined by an induction of at least 1.5 fold in Caco-2, DLD1 (intestinal epithelium), Jurkat (T lymphocytes) or K562 (erythroleukemia)

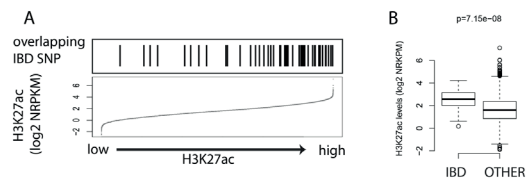


Figure 2 Activity of overlapping DRE

A) DRE are ranked according to the maximal H₃K₂₇ac signal. DRE overlapping with IBD-associated SNPs are indicated by black vertical lines. B) Boxplot depicts normalized maximal H₃K₂₇ac signal of DRE overlapping with IBD-associated SNPs compared to all other DRE. P-value was calculated using Wilcoxon signed rank test with continuity correction.

cells when compared to an empty vector. In concordance to their tissue specificity, seven out of eleven elements tested, showed activity in immune cells and/or epithelial cells (**Suppl. Fig 3.6**). In addition, for one out of seven enhancers, we were able to induce expression of a reporter gene in zebrafish intestine, supporting the functional significance and the evolutionary conservation of the identified regulatory sequences (**Figure 3**). Furthermore, the intestine-specific DRE were largely less active in HeLa or 293T cells, demonstrating the tissue-specific activity of the identified DRE.

Regions marked by IBD-associated SNPs are epigenetically more active in relevant cells

Next, we established that the studied IBD-associated SNPs were predominantly enriched in epigenetically more active regions (estimated by the number of active DRE within a 500kb region surrounding the SNP) compared to control-matched SNPs (**Figure 4A**), which was also confirmed in all separate DRE sets (**Suppl. Fig 3.7**). This enrichment was lower or not present in the susceptibility loci for other unrelated traits³⁷, such as cardiac arrest and clefting, where the involvement of the immune system or intestinal epithelium is not expected (**Suppl. Fig 3.8**). The 500,000 base pair size was selected based on the cutoff for the locus size used in the GWAS that identified 163 IBD loci.⁷ Interestingly, a substantial part of distal chromatin interaction of promoters with regulatory elements is within this distance.³⁸ This finding suggests that next to a direct effect on the activity of

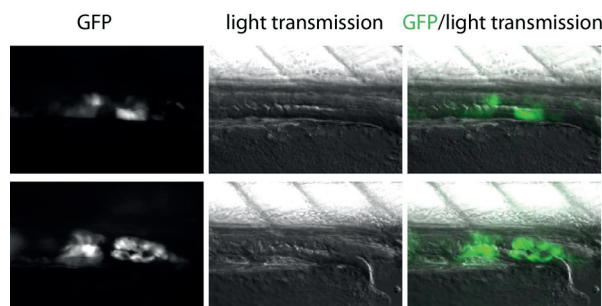


Figure 3 GFP expression driven by rs17085007-marked region in 5dpf zebrafish embryo.

Terminal part of gastrointestinal tract of two representative individual fishes is shown. GFP expression is shown in green.

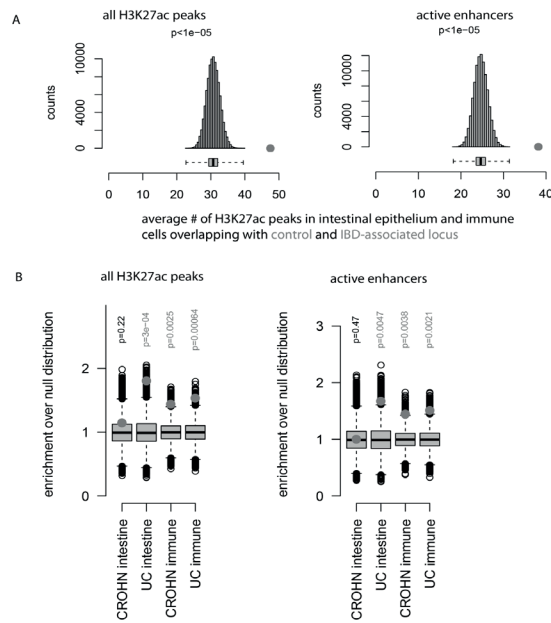


Figure 4. Accumulation of regulatory elements from intestinal epithelium and immune cells with IBD-associated loci.

A) Average number of regulatory elements overlapping with IBD-associated locus (grey dot) compared to 100,000 matched control sets (grey bars). B) Enrichment of regulatory elements from intestinal epithelium and immune cells in UC- and CD-specific loci (grey dots) normalized to the average counts in 100,000 matched control sets (grey bars). p-Values were calculated with permutation test. Statistically significant co-localizations are depicted in red.

regulatory elements, some of the IBD-associated SNPs can be markers for other unknown variants in nearby regulatory regions.

To investigate the possible differential involvement of DRE in different subtypes of IBD, we examined the enrichment of DRE in loci marked by SNPs that were found to be specific for Crohn's disease (CD) or ulcerative colitis (UC) separately.⁷ Notably, while DRE identified in immune cells show similar enrichments in both CD- and UC-specific loci, enrichment of DRE found in the intestinal epithelium was present only in UC loci and was highly diminished in CD loci (**Figure 4B**), which was reproducible over majority of separate DRE sets (**Suppl. Fig 3.9**). These data suggest that even though CD and UC have many overlapping pathophysiological elements, certain active tissue-specific DRE may also be associated with a subtype of IBD. As such, CD-associated loci overlap with DRE in both immune and intestinal epithelial cells, whereas UC-associated loci tend to be more restricted to DRE in intestinal epithelial cells.

Discussion

In this study we demonstrate overlap of IBD-associated SNPs, previously identified by GWAS, and active regulatory elements assayed in involved cell types – intestinal

epithelium and immune cells. More than half (54%) of the IBD-associated SNPs showed association with regulatory elements in at least one of the 20 samples from intestinal epithelium and immune cells, which is ~3.5 times more than expected from matched random variants.

Among the SNPs that did not show overlap with our newly identified DRE, a substantial part can still be functional by overlapping with DRE we did not annotate. Even though our setup covered the major cell types involved in IBD, some of the relevant genomic interactions could still be missed since they might be present in other cell types, e.g. fibroblasts, other leukocyte subsets or even unknown cell types with their own set of unique regulatory elements. Moreover, certain regulatory elements involved in the IBD pathogenesis could be active only under special circumstances, such as active inflammation, stress, hypoxia, regeneration or differentiation.

Taking into account the large overlap between DRE and GWAS SNPs, novel approaches to link tag SNPs and candidate genes might be considered in the future. Currently, GWASs report candidate genes mainly based on the functional relationships between genes found in the vicinity of the tag SNPs.³⁹⁻⁴² Even though these approaches have led to the identification of novel and valid candidate genes and pathways, they are mainly limited and biased by prior knowledge like functional gene or pathway annotations. Possible novel approaches based on chromosome conformation capture technologies⁴³⁻⁴⁵ and 3D organization of chromatin (4C technology) might prioritize candidate genes based on physical proximity of their promoters to a regulatory region. This allows identifying the gene that is directly regulated by DRE overlapping with a certain GWAS SNP.

Next, besides common GWAS SNPs, rare variants in DRE, which are missed by GWASs may also play a role in IBD pathogenesis. Therefore, besides exon-centric re-sequencing projects aiming to find causal deleterious sequence variants, which are directly affecting protein-coding genes, re-sequencing of DRE might also reveal causal variants. This proposed approach might particularly apply to cases with a clear Mendelian inheritance pattern in the absence of causative mutations in protein coding genes. In general, in 75-81% of such cases^{46,47}, exome-centric sequencing does not lead to identification of a causal variant⁴⁷, suggesting the existence of causal damaging variants in non-coding DNA.

Overall, we show that numerous SNPs in non-coding regions, which were identified in GWASs, co-localize with active DRE in immune and intestinal epithelial cells. These findings support the possible involvement of active DRE in the IBD pathogenesis. Further studies are necessary to determine which protein coding genes are regulated by these DRE. We suggest that genetic studies in IBD and other complex diseases may consider the cell type-specific involvement of functional genomic elements other than protein coding regions.

References

1. Nieuwenhuis EE, Blumberg RS. The role of the epithelial barrier in inflammatory bowel disease. *Adv Exp Med Biol* 2006;579:108-16.
2. Cho JH, Brant SR. Recent Insights Into the Genetics of Inflammatory Bowel Disease. *Gastroenterology* 2011;140:1704-U21.
3. Hamm CM, Reimers MA, McCullough CK, et al. NOD2 status and human ileal gene expression. *Inflamm Bowel Dis* 2010;16:1649-57.
4. Glocker EO, Kotlarz D, Boztug K, et al. Inflammatory Bowel Disease and Mutations Affecting the Interleukin-10 Receptor. *New England Journal of Medicine* 2009;361:2033-2045.
5. Blaydon DC, Biancheri P, Di WL, et al. Inflammatory Skin and Bowel Disease Linked to ADAM17 Deletion. *New England Journal of Medicine* 2011;365:1502-1508.
6. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;13:255-62.
7. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;491:119-24.
8. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010;42:1118-+.
9. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47 (vol 43, pg 246, 2011). *Nat Genet* 2011;43:919-919.
10. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011;43:1066-73.
11. Consortium EP, Dunham I, Kundaje A, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57-74.
12. Trynka G, Sandor C, Han B, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 2013;45:124-30.
13. Maurano MT, Humbert R, Rynes E, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 2012;337:1190-1195.
14. Schaub MA, Boyle AP, Kundaje A, et al. Linking disease associations with regulatory information in the human genome. *Genome Res* 2012;22:1748-1759.
15. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;473:43-9.
16. Sato T, Stange DE, Ferrante M, et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* 2011;141:1762-72.
17. Mokry M, Hatzis P, Schuijers J, et al. Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res* 2012;40:148-58.
18. Jiang H, Wang F, Dyer NP, et al. CisGenome Browser: a flexible tool for genomic data visualization. *Bioinformatics* 2010;26:1781-2.
19. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010;28:1045-8.
20. Pham TH, Benner C, Lichtinger M, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* 2012;119:e161-71.
21. International HapMap C, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52-8.
22. R_Development_Core_Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.
23. Sandelin A, Alkema W, Engstrom P, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;32:D91-4.
24. Bessa J, Tena JJ, de la Calle-Mustienes E, et al. Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dyn* 2009;238:2409-17.
25. Urasaki A, Morvan G, Kawakami K. Functional dissection of the Tol2 transposable element identified the minimal cis-sequence and a highly repetitive sequence in the subterminal region essential for transposition. *Genetics* 2006;174:639-49.
26. Creighton MP, Cheng AW, Welstead GG, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* 2010;107:21931-6.
27. Sato T, Clevers H. Growing Self-Organizing Mini-Guts from a Single Intestinal Stem Cell: Mechanism and Applications. *Science* 2013;340:1190-1194.

28. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 2012;22:1790-7.
29. Ahn SH, Shah YM, Inoue J, et al. Hepatocyte nuclear factor 4 alpha in the intestinal epithelial cells protects against inflammatory bowel disease. *Inflamm Bowel Dis* 2008;14:908-920.
30. Genomes Project C. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061-73.
31. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;153:307-19.
32. Loven J, Hoke HA, Lin CY, et al. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* 2013;153:320-334.
33. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science* 2012;336:736-9.
34. Wang Z, Zang C, Cui K, et al. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* 2009;138:1019-31.
35. Hollenhorst PC, Chandler KJ, Poulsen RL, et al. DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet* 2009;5:e1000778.
36. Ram O, Goren A, Amit I, et al. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* 2011;147:1628-39.
37. Hindorf LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;106:9362-7.
38. Sanyal A, Lajoie BR, Jain G, et al. The long-range interaction landscape of gene promoters. *Nature* 2012;489:109-U127.
39. Raychaudhuri S, Plenge RM, Rossin EJ, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* 2009;5:e1000534.
40. Krauthammer M, Kaufmann CA, Gilliam TC, et al. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc Natl Acad Sci U S A* 2004;101:15148-15153.
41. Wang K, Li MY, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007;81:1278-1283.
42. Franke L, van Bakel H, Fokkens L, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;78:1011-1025.
43. Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 2006;38:1348-54.
44. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem* 2009;107:30-9.
45. Simonis M, Kooren J, de Laat W. An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* 2007;4:895-901.
46. de Ligt J, Willemsen MH, van Bon BW, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 2012;367:1921-9.
47. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;369:1502-11.



Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease.

4

Claartje A. Meddens, Magdalena Harakalova, Noortje A. M. van den Dungen, Hassan Foroughi Asl, Hemme J. Hijma, Edwin P. J. G. Cuppen, Johan L. M. Björkegren, Folkert W. Asselbergs, Edward E. S. Nieuwenhuis, and Michal Mokry

Based on: *Genome Biology* 17(1):247 2016

Background

Inflammatory Bowel Disease (IBD) is an inflammatory disorder of the gastro-intestinal tract with an intermittent, chronic or progressive character. Studies on the pathogenesis of IBD have elucidated the involvement of a broad range of processes that mainly regulate the interaction between the intestinal mucosa, the immune system and microbiota.¹ A role for genetics in the pathogenesis of IBD has been established through twin-, family- and population based studies.¹ Subsequently, a substantial effort to identify genetic elements involved in the IBD pathogenesis followed. In this respect, multiple Genome Wide Association Studies (GWASs) have been performed over the past years.²⁻⁵ In these studies, common genetic variants (Single Nucleotide Polymorphisms, SNPs) are assayed across the whole genome in search of variants that are significantly over- or underrepresented in patients compared to healthy controls. Although GWASs have revealed many IBD associated loci, for most loci the causal genes that led to the associations have not been identified. Furthermore, the majority of IBD-associated SNPs are located in non-coding DNA and therefore cannot be causal in the sense that they directly lead to amino acid changes at the protein level.^{2-4,6-9} Therefore, these SNPs are generally thought to be markers for disease-causing variants in nearby genes. This model is used in classical approaches for candidate gene identification. These approaches are mainly based on the selection of genes that have shared functional relationships and are localized in the vicinity of the identified loci.^{10,11} This has led to the identification of crucial genes and pathways involved in IBD pathogenesis.¹² However, over the past decade it has been established that besides genes, the human genome consists of many other functional elements in the non-protein coding regions. These regions of the genome can play a role in the pathogenesis of complex diseases. As such, many types of DNA Regulatory Elements (DRE), especially enhancer elements, are involved in establishing spatiotemporal gene expression patterns in a cell type specific manner.¹³ These elements are crucial in the regulation of developmental processes and in maintaining cell type specific functionality. It is therefore now widely appreciated that part of the GWAS associations is due to sequence variation in DNA regulatory elements, but this information has largely been ignored in candidate gene identification.^{9,14-18}

We have recently shown that 92 of 163 IBD GWAS susceptibility loci localize to DNA regulatory elements (DRE, identified through the presence of H3K27Ac in relevant cell types).⁹ DRE are involved in transcription regulation and establishing cell type specific expression patterns.¹⁹ The genes that are regulated by the IBD-associated elements are likely to play a role in IBD and can therefore be considered as IBD candidate genes. This information has not been used in previous candidate gene approaches, because the identification of these genes comes with several hurdles. Since regulatory elements can regulate genes via chromatin-chromatin interactions that comprise up to 1 megabase^{20,21}, these genes cannot be identified based on their linear distance from the regulatory regions. Classical methods for candidate gene identification, that take regulatory mechanisms

into account, have mainly been restricted to computational approaches.^{14,16,22,23} So far, a limited number of studies have shown the value of using physical interactions between regulatory elements and the genes they regulate through studying the 3D nuclear conformation chromatin interactions in GWAS interpretation. These studies analyzed either single interactions (3C) or many-vs-many interactions (Hi-C) and were performed in colorectal cancer, auto-immune diseases and multiple other diseases.²⁴⁻²⁷ In contrast to these approaches we make use of 4C-seq, thereby increasing the number of analyzed interactions compared to 3C and increasing the resolution compared to Hi-C. Our study provides the first systematic analysis of chromatin interactions between disease-associated DRE and candidate genes in IBD. We have identified 902 novel IBD candidate genes, consisting of many noteworthy genes, for example *IL10RA*, *SMAD5* and *ATG9A*.

Methods

Cell culture

DLD-1 cells were cultured in RPMI-1640 with 10% FCS and standard supplements. Cells were harvested for 4C template preparation by trypsinization at 60-80% confluence.

Monocyte and Peripheral Blood Lymphocyte (PBL) isolation

Peripheral blood was collected from two healthy donors (one for monocyte isolation, one for PBL isolation) in sodium-heparin tubes. Peripheral Blood Mononuclear Cells (PBMCs) were isolated by Ficoll-Paque gradient centrifugation. PBMCs were incubated with magnetic CD14+ microbeads (Milteny, order nr. 130-050-201) according to manufacturer's manual. Thereafter cells were magnetically separated by the AutoMACS™ Separator, negative fraction consisted of PBLs, positive fraction of monocytes.

Circular Chromosome Conformation Capture - Sequencing

Template preparation

For each cell type one 4C-template was prepared. 4C-chromatin preparation, primer design and library preparation were described previously.²⁸ 10×10^6 cells were used for chromatin preparation per cell type (monocytes, PBLs and DLD-1). Primer sequences are listed in **Supplementary table 1**. The library preparation protocol was adapted to make it compatible with the large number of viewpoints. Details can be found in the **Supplementary methods**.

Sequencing

Libraries were sequenced using the HiSeq2500 platform (Illumina), producing single end reads of 50 bp.

Data analysis

The raw sequencing reads were de-multiplexed based on viewpoint specific primer sequences (the datasets are accessible through GEO Series accession number GSE89441). Reads were then trimmed to 16 bases and mapped to an *in silico* generated library of fragends (fragment ends) neighboring all DpnII sites in human genome (NCBI37/hg19), using the custom Perl scripts. No mismatches were allowed during the mapping and the reads mapping to only one possible fragend were used for further analysis. To create the 4C signal tracks in the UCSC browser, we have generated the *.bed files with information for each mappable fragend on the coordinates and their covered/non-covered (1 or 0) status. Visualization of the tracks in the UCSC browser was done with the following settings: windowing function: mean; smoothing window: 12 pixels.

Identification of the interacting genes

First, we calculated the number of covered fragends within a running window of k fragends throughout the whole chromosome where the viewpoint is located. This binary approach (i.e. a fragend is covered or is not covered in the dataset) was chosen to overcome the influence of PCR-efficiency-based biases, however this approach decreases the dynamic range of the 4C-seq and may overestimate the strength of distal interactions compared to proximal interactions. The k was set separately for every viewpoint so it contains on average 20 covered fragends in the area around the viewpoint (± 100 kbp), e.g. when 100 out of 150 fragends around the viewpoint were covered the window size was set to 30 fragends. Next, we compared the number of covered fragends in each running window to the random distribution. The windows with significantly higher number of covered fragends compared to random distribution ($p < 10^{-8}$ based on binominal cumulative distribution function; R *pbinom*) were considered as significant 4C signal. The following criteria were defined for the identification of the candidate genes; i) the Transcriptional Start Site (TSS) co-localizes with a significant 4C-seq signal ($p < 10^{-8}$) within 5 kbp; ii) the susceptibility variant or other variant in linkage disequilibrium (LD) co-localizes with the H₃K₂₇ac signal (that marks activating regulatory elements) in the cell type from which the 4C signal was obtained (68 loci in monocytes, 73 in lymphocytes and 52 in intestinal epithelial cells)⁹ and iii) the gene is expressed ($\log_2(\text{RPKM}) > -0.5$) in the assayed cell type (Supplementary table 2). Datasets used for expression analysis are listed in **Supplementary table 3**. Quality measures for the 4C library preparation and sequencing can be found in **Supplementary figures 4.1-4.3**. The use of single 4C templates per cell type was validated in a biological duplicate of the lymphocyte 4C template that is derived from a different donor (**Supplementary figure 4.4A**) and the reproducibility in other chromatin interaction datasets was established by intersecting our findings with two Hi-C datasets²⁵ (**Supplementary figure 4.4B, Supplementary table 3**).

TSS occupancy by H₃K₂₇ac and H₃K₄me₃

The publically available datasets of H₃K₂₇ac and H₃K₄me₃ occupancy were accessed from the UCSC/ENCODE browser (<http://genome.ucsc.edu/ENCODE/>). Datasets are listed in **Supplementary table 3**. The occupancy around 2kbp \pm of TSS of was calculated using custom Perl scripts and Cisgenome²⁹ functions.

eQTL analyses

GTEx: A manual look-up was performed for expression Quantitative Trait Loci (eQTL) in the Genotype-Tissue Expression (GTEx) database (accession dates; eQTL-genes: 05-2016; p-values: 09-2016). The presence of eQTL genes for each of the 92 IBD associated SNPs was performed in four different tissues: Colon-Transverse; Colon-Sigmoid; Small Intestine-Terminal Ileum; Whole Blood.³⁰ Next, for each gene for which an IBD associated SNP turned out to be an eQTL, its presence among the 4C-seq identified genes was evaluated (**Supplementary table 4**). All transcripts in the GTEx database that were not included in the gene annotation (UCSC genes 2009) that was used for the analysis of the 4C-seq data were removed from the analysis.

STAGE: eQTLs were analyzed using the Stockholm Atherosclerosis Gene Expression (STAGE)³¹ dataset (**Supplementary methods**). Identified loci from GWAS for IBD were matched with imputed and genotyped SNPs and were selected for eQTL discovery. We compared the amount eQTLs between ‘SNP-candidate gene’-pairs and ‘SNP-control gene’-pairs. Control genes are genes within the same locus that are not interacting with the IBD associated locus. An empirical FDR was estimated for each eQTL-gene by shuffling patient IDs 1000 times on genotype data as described previously³².

Gene Set Enrichment Analysis (GSEA)

GSEA³³ was performed using gene expression datasets³⁴ from intestinal biopsies obtained from Ulcerative Colitis patients (datasets available at GSE11223). The “normal uninfamed sigmoid colon” and “UC inflamed sigmoid colon” were used and the fold change in expression were calculated

using the GEO2R tool³⁵ with default settings. Significance of the enrichment was calculated based on 1000 cycles of permutations.

Signaling pathway analysis

The IL10 signaling pathway components were retrieved from Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City). Genes upregulated upon IL10 signaling (target genes) and genes involved in the bilirubin cascade were removed before further analysis. The interactions between the members of the IL-10 signaling pathway were visualized using the GeneMania tool <<http://www.genemania.org/>>.

The general pathway analysis was performed with the Ingenuity Pathway Analysis software (IPA®, QIAGEN Redwood City), based on the candidate genes from the three cell types, separately.

Upstream regulators

Upstream regulators that are enriched regulators of the candidate genes in our datasets were identified with the Ingenuity Pathway Analysis software (IPA®, QIAGEN Redwood City), based on the candidate genes from the three cell types separately. The Ingenuity's Upstream Regulator Analysis algorithm predicts upstream regulators from gene datasets based on the literature and compiled in the Ingenuity knowledge base.

CTCF tracks

CTCF tracks were accessed from the UCSC/ENCODE browser (<http://genome.ucsc.edu/ENCODE/>). Datasets are listed in **Supplementary table 3**.

Tracks used for rs630923 and rs2382817 (Figure 4, Supplementary figure 4.9)

All tracks were accessed from the UCSC/ENCODE browser (<http://genome.ucsc.edu/ENCODE/>). Datasets are listed in **Supplementary table 3**. Haploblock structures were visualized with Haploview³⁶; Pairwise LD statistics of variants with a distance up to 500 kbp was used in the analyses.

Organoid culture

Colon biopsies were obtained by colonoscopy. The biopsies were macroscopically and pathologically normal. Crypt isolation and culture of human intestinal cells from biopsies have been described previously^{37,38}. In summary, human organoids were cultured in expansion medium (EM) containing RSPO1, noggin, EGF, A83-01, nicotinamide, SB202190, and WNT3A. The medium was changed every 2–3 days and organoids were passaged 1:4 every 9 days.

5–7 Days after passaging, the organoids were exposed to 10 μ L sterilized *E.Coli*-lysate (control organoids were not stimulated). After 6 hours of exposure, the organoids were harvested and RNA was extracted using TRIzol LS (Ambion™). cDNA was synthesized by performing reverse-transcription (iScript, Biorad). mRNA abundances were determined by real-time PCR using primer pairs that target *HNF4 α* *NFKB1* (**Supplementary table 1**) with the SYBR Green method (Bio-Rad). *ACTIN* mRNA abundance was used to normalize the data.

Results

Genes interacting with DRE at IBD associated loci

A meta-analysis on GWASs performed in IBD resulted in the confirmation of 163 susceptibility loci.³ We have recently shown that 92 of these 163 loci overlap with enhancer elements (regulatory elements that enhance transcription) that are active in relevant cell types for IBD (i.e. intestinal epithelial cells and immune cells).⁹ We now use this information to identify novel IBD candidate genes. We do so by identifying the genes

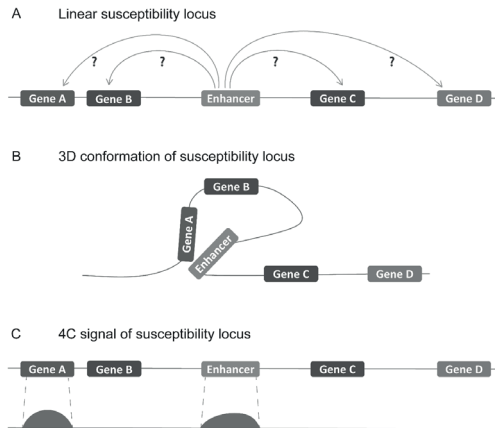


Figure 1. 3D Nuclear organization in candidate gene identification.

A) The linear organization of the genome does not provide sufficient information to predict which gene is regulated by an enhancer of interest. B) Genes that are regulated by an enhancer form a 3D nuclear interaction. C) The 4C-seq technique captures the 3D conformation and results in a signal around the gene that was interacting with the viewpoint (i.e. the SNP). For a detailed explanation of the 4C-seq procedure we refer to the published 4C protocol.²⁸ In this study, the analysis of the 3D conformation of chromatin will reveal which genes interact with an enhancer that is found at an IBD-susceptibility locus. The 4C analysis of a locus will show an interaction signal that can be mapped to the gene with which the interaction was formed. Therefore, 4C-seq can be used as a tool to use information on DNA regulation for candidate gene identification.

that are regulated by these 92 regulatory elements. Since the regulated genes cannot be pinpointed by studying the linear organization of the susceptibility loci, we assayed the 3D conformation of these loci (**Figure 1**). The effect of common variants, especially those in regulatory elements, is relatively mild. Therefore it is very unlikely that a single common variant will ablate or create a whole regulatory region and its 3D interaction.³⁹ By the same reasoning, we do not expect that the 3D interactions in patients will be fundamentally different compared to healthy controls or cell lines. However, regulation of genes can be genotype specific¹⁶, which demands for the identification of genes that are dysregulated in IBD. For these reasons we decided on an experimental setup where we assay chromatin conformation in healthy control cells and a cell line, to identify genes that can be dysregulated in IBD under pathological conditions. Therefore, we have performed 92 high resolution 4C-seq (circular chromosome conformation capture-sequencing) experiments to cover all individual IBD susceptibility loci that overlap DRE in three cell types, thereby creating 276 individual chromatin interaction datasets. This way, we could identify all genes that physically interact with the regulatory elements that are found at IBD associated loci. As the activity of enhancers is known to be cell type specific¹⁹, we assayed chromatin interactions in monocytes (i.e. CD14⁺ fraction of PBMCs), lymphocytes (i.e. CD14⁻ fraction of PBMCs) and in an intestinal epithelial cell line (DLD-1, derived from colorectal adenocarcinoma).

4C-seq identifies different sets of candidate genes in different cell types

The candidate genes we reported here all meet the following criteria: 1) the enhancer element physically interacts with the candidate gene ($p > 10^{-8}$) 2) the enhancer element is

active in the assayed cell type (i.e. the associated variant or a variant in LD colocalizes with the histone mark H₃K₂₇Ac)⁹ and 3)

the candidate gene is expressed in the assayed cell type ($\log_2(\text{RPKM}) > -0.5$). With this approach we identified 1409 candidate genes: 923 genes in monocytes, 1170 in lymphocytes and 596 in DLD-1 cells of which 796 were shared by two or more cell types and 810 were found in only one cell type (**Figure 2A, B**). We identified 902 IBD candidate genes that have not been reported by GWASs before (**Table 1, Supplementary table 2**). 22 of the 92 studied loci are associated to only one of the IBD subtypes (11 to Crohn's disease, 11 to ulcerative colitis). The candidate genes that were identified for these loci might contribute to the mechanisms that lead to the subtype specific phenotypes. Interestingly, for two loci on chromosome 7 that give separate GWAS signals for CD (rs10486483) and UC (rs4722672), the 10 candidate genes that were identified for this CD locus were also found in the UC locus. This implies that in some cases, although the genetic risk factor is different between the subtypes, the mechanisms that underlie the genetic risk can share downstream components. Notably, this UC locus is active in intestinal epithelium, whereas the CD locus is not, which resulted in the identification of additional candidate

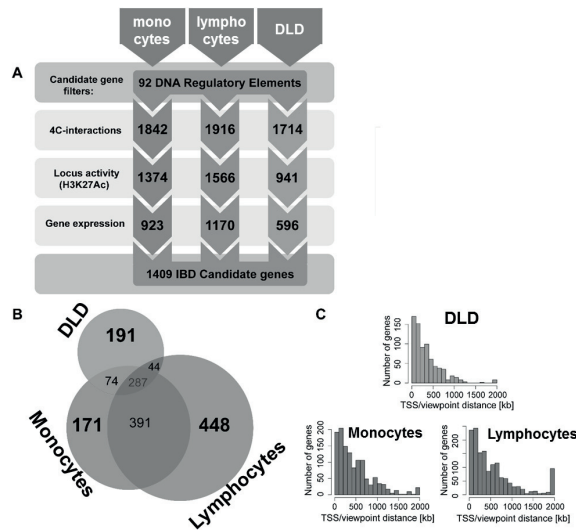


Figure 2. Candidate gene characteristics

A) Flowchart of filtering steps that were performed to identify IBD candidate genes (4C interactions with $p > 10^{-8}$; locus activity based on the colocalization of the associated variant or a variant in LD with H₃K₂₇Ac; gene expression: $\log_2(\text{RPKM}) > -0.5$). The number of remaining genes after each steps is depicted in the corresponding arrow. B) A Venn diagram of the candidate genes (that meet all three criteria) identified in the three separate cell types. The surface of the circles corresponds with the numbers of genes that are unique for one cell type and with the genes that where only two cell types overlap. The number of genes shared by all three cell types is depicted in the center of the diagram. The differences between DLDs and the immune cells is not solely due to shared active enhancers between monocytes and lymphocytes that are inactive in DLDs. To address this, Supplementary figure 5 depicts a Venn diagram of all genes interacting with one of all (92) assayed viewpoints. These results confirm the ability of 4C-seq to detect cell type specific chromatin-chromatin interactions. C) Distribution of the distance between the reported candidate genes and the viewpoints. The majority of the genes is located several hundreds of kilo-bases away from the susceptibility locus.

Table 1. 4C-seq output per locus.

This table shows the 4C output for each associated IBD locus that overlaps an active regulatory element based on the presence of H₃K₂₇Ac and is arranged on the GWAS association of the loci to either Crohn's disease (CD), ulcerative colitis (UC) or both (IBD). The positions are given as in the GWAS in which the association was found³ and are relative to human reference genome GRCh37. Depicted are the name of the SNP and the key novel genes that were identified at that locus. The numbers refer to the number of novel genes that were identified at that locus and, between brackets, the total number of candidate genes identified at that locus.

IBD				CD				UC						
Chr.	Position (Mb)	SNP	Key Novel Genes	Novel (total)	Chr.	Position (Mb)	SNP	Key Novel Genes	Novel (total)	Chr.	Position (Mb)	SNP	Key Novel Genes	Novel (total)
IBD	1	8.02	rs35675666	CTNNB1P1	12(17)	IBD	12	48.2	rs11168249	12(18)				
IBD	1	151.79	rs4845604	NOTCH2NL	25(36)	IBD	13	99.95	rs9557195	5(10)				
IBD	1	155.67	rs670523	MUC1	34(55)	IBD	13	27.52	rs17085007	14(15)			CDK8	15(19)
IBD	1	200.87	rs7554511		3(7)	IBD	14	69.27	rs194749	15(19)			ATP6V1D	6(13)
IBD	1	206.93	rs3024505	CD55, CD46, IKBKE	9(12)	IBD	14	75.7	rs4899554	13(15)			BATF	13(15)
IBD	1	22.7	rs12568930	CDC42, EPH2B	5(6)	IBD	14	88.47	rs8005161	3(5)			PIAS1	4(8)
IBD	2	25.12	rs6545800		4(8)	IBD	15	67.43	rs17293632	5(15)			LAT	8(15)
IBD	2	219.14	rs2382817	ATG9A	14(25)	IBD	16	23.86	rs7404095	11(17)			TNFRSF1	18(26)
IBD	2	43.81	rs10495903		9(12)	IBD	16	28.6	rs26528	12(26)			TRIM37	0(0)
IBD	2	102.86	rs917997		8(12)	IBD	16	11.54	rs529866	6(8)			DNM2	13(37)
IBD	2	191.92	rs1517352		8(12)	IBD	17	57.96	rs1292053	2(18)			PGLYRP1	2(18)
IBD	2	28.61	rs925255	TANK	4(11)	IBD	17	37.91	rs12946510	1(5)			BCL2L1	2(13)
IBD	3	48.96	rs3197999	HYAL1, HYAL2, HYAL3	43(90)	IBD	17	40.53	rs12942547	11(20)			SUMO3	4(4)
IBD	3	18.76	rs4256159	KAT2B	7(7)	IBD	17	32.59	rs3091316	9(18)			CRKL	17(24)
IBD	4	74.85	rs2472649	AREG	4(10)	IBD	18	46.39	rs7240004	6(9)			NOTCH2NL	4(7)
IBD	5	176.79	rs12654812	FGFR4	15(30)	IBD	19	10.49	rs11879191	14(17)			FBXW11	18(20)
IBD	5	96.24	rs1363907		8(13)	IBD	19	46.85	rs4802307	2(5)			MYC	0(3)
IBD	5	131.19	rs2188962	SKP1	17(32)	IBD	19	1.12	rs2024092	6(10)			TAX1BP1	2(2)
IBD	5	141.51	rs6863411	HDAC3	13(18)	IBD	20	31.37	rs4911259	13(31)			BAX	5(16)
IBD	5	10.69	rs2930047		5(6)	IBD	20	48.95	rs913678	5(16)			MUL1	7(11)
IBD	5	40.38	rs11742570	RICTOR	9(11)	IBD	20	30.75	rs6142618	10(15)			PPP3CA	5(11)
IBD	6	138	rs6920220	IFNGR1	12(13)	IBD	21	45.62	rs7282490	9(15)			SMAD5	9(15)
IBD	6	167.37	rs1819333	DLL1	9(13)	IBD	21	16.81	rs2823286	8(22)			TAX1BP1	3(7)
IBD	6	90.96	rs1847472	MAP3K7	7(8)	IBD	22	21.92	rs2266959	32(50)			MAPK3	15(22)
IBD	7	50.245	rs1456896		3(5)	CD	1	114.3	rs6679677	3(10)			NFATC3	3(10)
IBD	7	100.34	rs1734907		45(61)	CD	1	120.45	rs3897478	10(18)				10(18)
IBD	8	126.53	rs921720	KIAA0196	3(5)	CD	2	62.55	rs10865331					
IBD	8	130.62	rs1991866	MYC	2(2)	CD	2	234.15	rs12994997					
IBD	9	139.32	rs10781499	TRAF2	12(28)	CD	5	173.34	rs17695092					
IBD	10	94.43	rs7911264		3(7)	CD	6	128.24	rs13204742					
IBD	10	82.25	rs6586030		6(9)	CD	6	21.42	rs12663356					
IBD	10	6.08	rs12722515	PRKCC	10(17)	CD	7	26.88	rs10486483					
IBD	10	101.28	rs4409764		1(7)	CD	8	129.56	rs6651252					
IBD	10	35.3	rs11010067	FZD8	3(7)	CD	19	49.2	rs516246					
IBD	10	30.72	rs1042058	ITGB1	7(9)	CD	21	34.77	rs2284553					
IBD	10	81.03	rs1250546	DLG5	4(8)	UC	1	2.5	rs10797432					
IBD	10	59.99	rs2790216		0(4)	UC	1	20.15	rs6426833					
IBD	11	1.87	rs907611	TOLLIP, DUSP8	19(25)	UC	4	103.51	rs3774959					
IBD	11	65.65	rs2231884	PELI3	53(76)	UC	5	134.44	rs254560					
IBD	11	61.56	rs4246215		5(16)	UC	7	2.78	rs798502					
IBD	11	76.29	rs2155219		8(11)	UC	7	27.22	rs4722672					
IBD	11	118.74	rs630923	IL10RA	15(28)	UC	11	114.38	rs561722					
IBD	11	60.77	rs11230563		12(20)	UC	16	30.47	rs11150589					
IBD	12	68.49	rs7134599	MDM2	5(6)	UC	16	68.58	rs1728785					
IBD	12	40.77	rs11564258		2(3)	UC	20	43.06	rs6017342					
IBD	12	12.65	rs11612508		4(12)	UC	20	33.8	rs6088765					

genes for rs4722672 that are UC specific (Table 1). Among the identified candidate genes are many noteworthy genes that have been implicated in the IBD pathogenesis, but that were never identified through GWAS-associations (Table 2⁴⁰⁻⁴⁶). We have now identified these novel candidate genes that have been missed by classical approaches for candidate gene identification.

As expected, based on their common hematopoietic origin, the two immune cell types show larger overlap compared to DLD-1 cells (Figure 2B, Supplementary figure 4.5). With a median enhancer-to-gene distance of 261, 370 and 354 kilobasepairs in DLD-1, lymphocytes and monocytes, respectively, a large proportion of the genes we report are found outside the GWAS susceptibility loci (Figure 2C). Notably, some of the interactions between IBD loci and candidate gene span over 5 megabases. For example,

Table 2. Noteworthy novel candidate genes.

IL10RA, interleukin 10 receptor subunit alpha; IL10RB interleukin 10 receptor subunit beta; Th17 cells, T-helper 17 cells, Th2 cells, T-helper 2 cells; SMAD, named after their homologous genes Mothers Against Decapentaplegic (MAD) and the Small Body Size protein (SMA) in *Drosophila* and *C. Elegans* respectively; CD, complement-decay accelerating factor; MCP, membrane cofactor protein; DAF, decay accelerating factor.

Noteworthy novel candidate genes
<i>ATG9A</i> : <i>ATG9A</i> encodes autophagy related protein 9A. Autophagy plays an important role in host defense by eliminating pathogens. ATG-family member <i>ATG16L1</i> has previously been associated with Crohn's disease ²⁹ .
<i>BATF</i> : Basic leucine zipper transcription factor ATF-like (BATF) belongs to the activator protein 1 family that is involved in transcription regulation in all immune cells. <i>Batf</i> -deficient mice do not develop Th17 cells and do not produce IL17. Furthermore, BATF regulates cell type specific gene expression in Th2 cells, germinal center B-cells, and T-follicular helper cells ³⁰ .
<i>CD46/CD55</i> : <i>CD46</i> (also known as <i>MCP</i>) and <i>CD55</i> (also known as <i>DAF</i>) are regulatory proteins expressed on surface membranes. These proteins protect the host from autologous complement-mediated injury upon activation of the complement cascade. <i>Daf</i> -deficient mice show increased epithelial damage upon induction of colitis, delayed healing, and elevated expression of proinflammatory cytokines ³¹ .
<i>IL10RA</i> : The IL10-receptor consists of the two subunits IL10RA and IL10RB. Sequence variants in genes encoding these two subunits are known to cause severe very early onset IBD in a monogenic fashion ³² . While the association of <i>IL10RB</i> with the complex form of IBD was reported by GWASs, the link with <i>IL10RA</i> was so far missing.
<i>SMAD5</i> : <i>SMAD5</i> is a downstream effector in BMP signaling. <i>SMAD5</i> expression was found to be downregulated in intestinal cells of IBD patients. Furthermore, conditional depletion of <i>Smad5</i> in mice results in increased susceptibility for development of colitis upon DSS-induction (dextran sulfate sodium) ³⁴ .

rs925255 shows a significant ($p=6.068 \times 10^{-9}$) physical interaction with *TANK* (TRAF family member-associated NF- κ B activator), a gene that is localized 30 megabases from this locus (**Supplementary table 2**).

Validation and reproducibility of 4C-seq data

To validate the reproducibility of our data we prepared a 4C template from lymphocytes from a different donor and performed 4C-seq for the 92 regions on this material. **Supplementary figure 4.4A** shows that 91% of the candidate genes that are identified in the replicate dataset were also identified in the dataset that is used throughout this study. This demonstrates the reproducibility of the 4C-technique, not only in technical, but also in biological duplicates. These results are in line with studies that have previously shown that in 3C-based methods, results from biological duplicates are highly reproducible.⁴⁷ Furthermore, we validated the reproducibility of our data by intersection the 4C datasets with Hi-C datasets that were created in CD34+ leukocytes and a lymphoblastoid cell line.²⁵ This confirmed a high reproducibility by showing that 99% (CD34+) and 87% (lymphoblastoid) of the genes that were found by Hi-C were also found in our 4C-data (**Supplementary figure 4.4B**).

Identified candidate genes are actively expressed

We reasoned that genes that are truly regulated by active enhancers *in vivo* would, on average, be more highly expressed than other genes within the region of the 4C-signal. The quantitative examination of expression levels and histone modifications that mark active enhancers and promoters confirmed that the genes that were detected by our

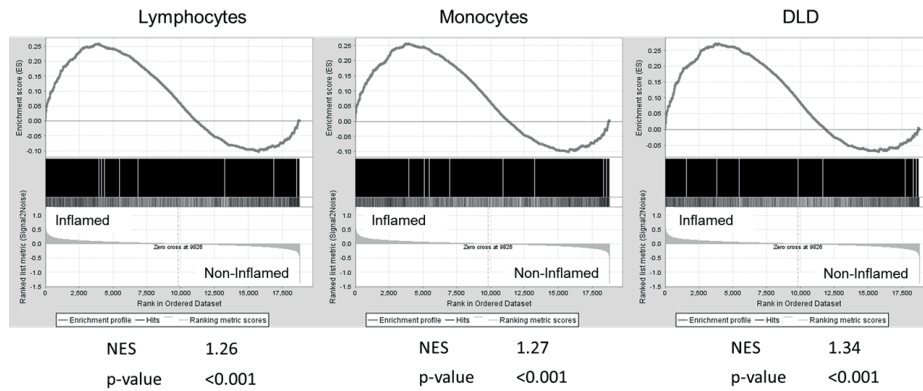


Figure 3. GSEA for candidate genes in intestinal inflammation in IBD.

The figure shows the Gene Set Enrichment Analysis for the candidate genes reported in monocytes, lymphocytes and DLDs. Genes that are upregulated (red) in inflamed compared to non-inflamed biopsies are plotted to the left of the x-axes, downregulated genes (blue) on the right. 4C-seq gene sets are significantly ($p < 0.001$) enriched for genes that are upregulated in the inflamed intestine of IBD patients (reflected by positive normalized enrichment score, NES). Enrichment score (ES) reflects the degree to which the 4C-seq genes sets are overrepresented at the differentially expressed genes in intestinal biopsies. The nominal p-value and the normalized enrichment score (NES, normalized for the size of the gene sets) are shown below each graph.

method indeed are more actively transcribed than all other genes (also than genes that were not detected by 4C and are found in the same genomic region, **Supplementary figure 4.6, 4.7**). These results support the detection of functional interactions by the 4C-seq approach that was executed here. Furthermore, we assessed ‘possible’ insulator elements (i.e. insulators occupied by CTCF protein) between the 92 DRE and the candidate genes. Interestingly, the majority of interactions bypasses several CTCF sites and numerous interactions skip over 50 sites bound by CTCF (**Supplementary figure 4.8**). In addition, genes that do not interact with the 4C viewpoint do not seem to have more CTCF sites between the viewpoint and their promoter compared to the interacting genes (**Supplementary figure 4.8**). This is in line with observations from Hi-C datasets where 82% of long range interactions bypass at least one CTCF site.²⁵

Previously, insulator regions have been shown to prevent enhancer-gene interactions.⁴⁸ We therefore investigated whether assessment of the CTCF-binding can be used as an alternative to the 4C-method by predicting the borders of the regions in which our candidate genes were found. We conclude that CTCF-binding information cannot be used as an alternative for the 4C-based candidate gene approach presented here.

4C-seq candidate genes have SNP dependent expression profiles

We hypothesize that the candidate genes that we identify are contributing to the IBD pathogenesis via impaired transcription regulation caused by variants in DNA regulatory elements. To test this hypothesis, we studied whether 4C-seq candidate genes show different expression profiles in different genetic backgrounds (i.e. in individuals that carry the associated SNP vs individuals that do not) through expression Quantitative Trait Loci

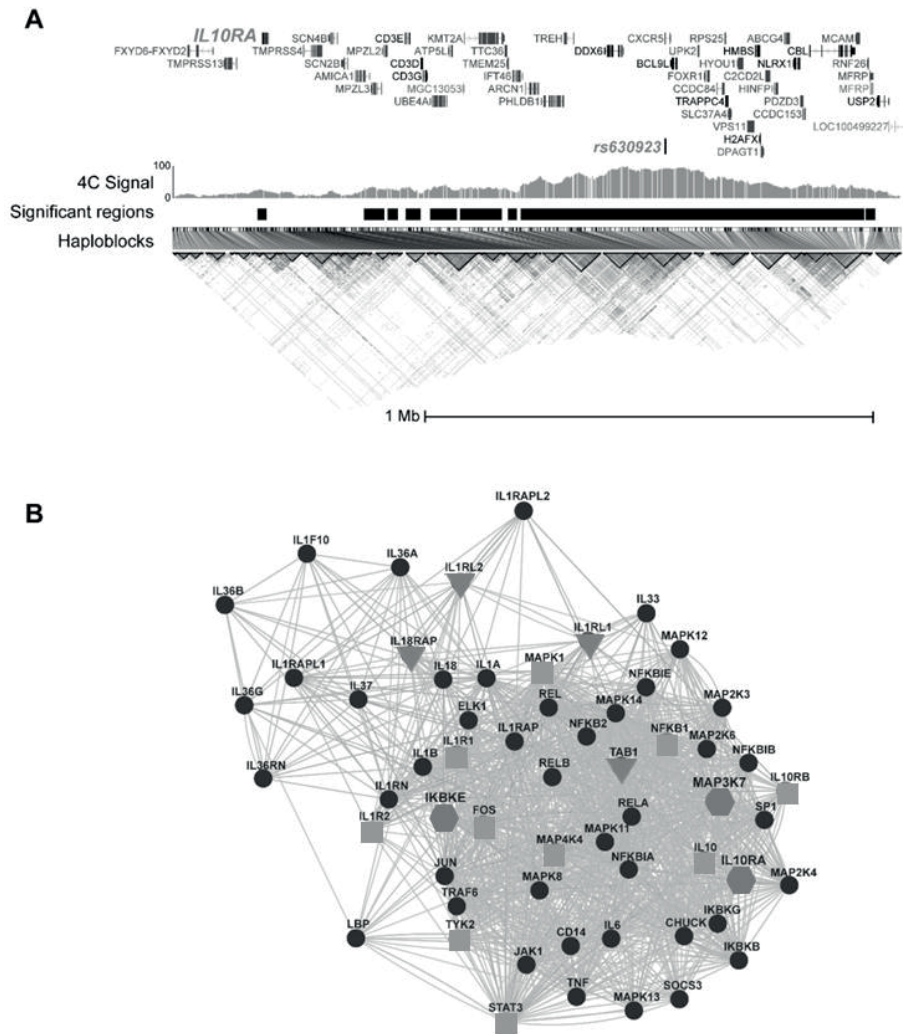


Figure 4. *IL10RA* is a novel IBD candidate gene.

A) The 4C-signal from the rs630923 locus in lymphocytes; signal on the y-axis is depicted as the percentage of fragments covered per pixel. Black bars indicate significant 4C signal ($p < 10^{-8}$); all coding genes located in these regions are shown. The TSS of *IL10RA* co-localizes with a distant significant signal (~1Mb from the viewpoint). Rs630923 and *IL10RA* localize to different haploblocks, meaning these regions do not co-segregate. B) A network that consists of members of the IL10 signaling pathway. Hexagons represent novel IBD candidate genes, squares represent candidate genes that were identified by 4C-seq as well as by GWAS, triangles represent previously reported candidate genes that were not identified in the 4C-seq dataset and black dots represent members of the IL10 pathway that have not been associated to IBD. Although many genes of the IL10 signaling pathway have been reported previously, we complement the network with 3 novel candidate genes including *IL10RA*.

(eQTL) analyses.²³ We performed two different analyses in separate databases. First, we used the GTEx database³⁰ to test whether our approach is able to detect the eQTLs that are present in the intestinal epithelium (colon-sigmoid, colon-transverse, terminal ileum) and whole blood.³⁰ We performed an eQTL look-up of the 92 IBD-associated SNPs in these tissues and found 50 genes with a SNP dependent expression profile. Interestingly,

all of the 50 genes were indeed identified by our 4C-seq approach (**Supplementary table 4**). Second, we made use of another eQTL database (STAGE)³¹ and explored the presence of candidate genes among the genes that were found to have expression levels that are dependent on the interacting SNP genotype in white blood cells. This revealed 10 candidate genes that have an eQTL in the STAGE database. Next, we analyzed all non-interacting genes within 2 megabases from the 4C viewpoint (**Supplementary table 4**). In contrast to the interacting genes, none of the non-interacting genes showed genotype dependent expression in the same database. These findings altogether support the capability of our method to identify the candidate genes which expression regulation is dependent on IBD associated genomic variants.

4C-seq gene set is enriched in genes involved in inflammation in IBD patients

After demonstrating that our method enables the identification of novel IBD candidate genes that are likely subject to SNP dependent expression levels, we examined whether the genes we report here are involved in the major pathogenic process in IBD, namely intestinal inflammation. To address this, we performed a gene set enrichment analyses (GSEA)³³ in which we used RNA expression data of intestinal biopsies from IBD patients³⁴. We compared expression levels in inflamed versus non inflamed intestinal biopsies and tested whether the 4C-seq candidate genes were enriched among the differentially

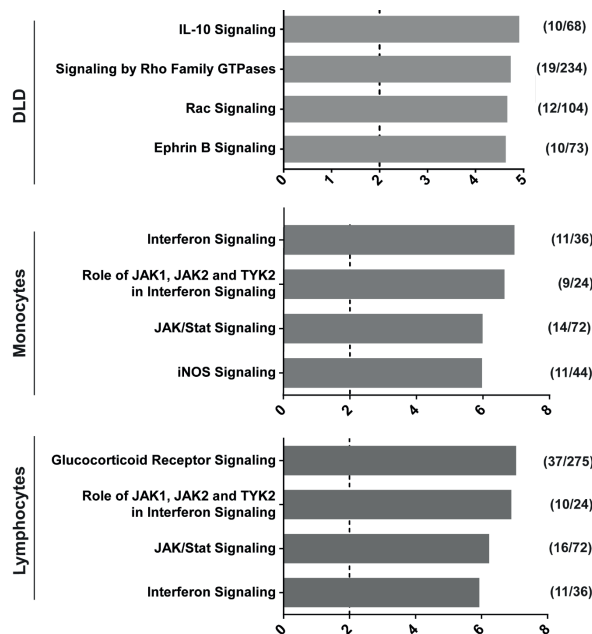


Figure 5. Pathways in IBD.

This figure shows the pathways that are most highly enriched among the identified candidate genes in the three separate cell types. Bars correspond with the $-\log$ of the p-value, the dashed line indicates the threshold for significance. Numbers between the brackets show (amount of pathway members in dataset / total amount of pathway members). Pathway analyses were performed using Ingenuity Pathway Analysis (IPA, see methods). All significantly enriched pathways can be found in **Supplementary table 4**.

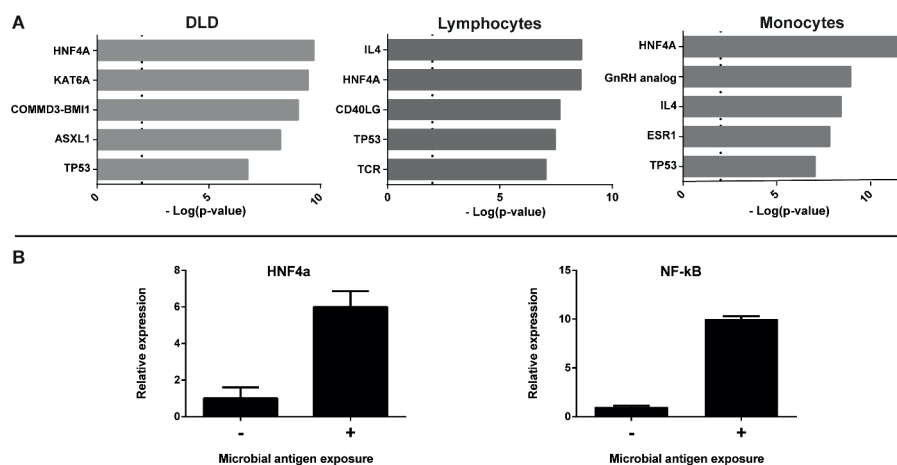


Figure 6. Upstream regulators of IBD candidate genes.

A) The charts show the upstream regulators of the identified IBD candidate genes in the separate cell types. Bars correspond with the $-\text{Log}$ of the p-value, the dashed line indicates the threshold for significance. The analyses were performed using Ingenuity's Upstream Regulator Analysis (see methods for more information). All significantly enriched upstream regulators can be found in Supplementary table 5. B) Relative RNA expression before (-) and upon (+) exposure of human intestinal organoids to microbial antigens. Data were normalized to β -Actin mRNA abundances. HNF4 α and NF- κ B are upregulated upon exposure. HNF4 α , Hepatocyte Nuclear Factor α ; KAT6A, K (lysine) acetyl transferase 6A; COMMD3-BMI1, naturally occurring read-through transcription between the neighboring COMM domain-containing protein 3 and polycomb complex protein BMI-1; ASXL1, Additional Sex Combs Like Transcriptional Regulator 1; TP53, Tumor Protein p53; IL4, Interleukin 4; CD40LG, CD40 ligand; TCR, T-cell receptor, GnRH, Gonadotropin Releasing Hormone; ESR1 Estrogen Receptor 1.

expressed genes. This analysis shows that all three 4C gene sets (monocytes, lymphocytes and intestinal epithelium) are highly enriched ($p < 0.001$) for genes that are upregulated upon intestinal inflammation in IBD patients (Figure 3). These results support the role of the candidate genes reported here in intestinal inflammation in IBD.

Chromatin interactions reveal IL10RA and ATG9A as novel IBD targets

IL10RA is one of the newly identified candidate genes. Previously, sequence variants in genes encoding the two subunits of the interleukin 10 receptor, *IL10RA* and *IL10RB*, were found to cause severe early onset IBD in a Mendelian fashion.⁴³ Our 4C datasets reveal that *IL10RA* interacts with an IBD-associated enhancer element in peripheral blood lymphocytes ($p = 4.1 \times 10^{-10}$). Since *IL10RA* is located -1Mbp upstream of the associated SNP (rs630923) and is separated from the SNP by multiple haploblocks (Figure 4A), this gene has not been identified through classical candidate gene approaches. The enhancer element that co-localizes with rs630923 is active in lymphocytes, but not in monocytes and intestinal epithelial cells (i.e. H3K27Ac marks are present only in lymphocytes). These results imply distinctive and cell type specific regulatory pathways for *IL10RA* expression in immune cells. Besides *IL10RA*, we identified 12 candidate genes that are part of the *IL10*-signaling pathway (Figure 4B), three of which are novel candidate genes

(*IL10RA*, *IKBKE*, *MAP3K7*). Thereby we confirm and further establish the important role of IL10 signaling in IBD.

Furthermore, we identified *ATG9A* (autophagy-related gene 9A) as a novel candidate gene, as its transcriptional start site is physically interacting with an enhancer element in the proximity of rs2382817 in DLDs and monocytes ($p=7.891 \times 10^{-13}$ in monocytes, $p=9.787 \times 10^{-12}$ in DLDs, **Supplementary figure 4.9**). *Atg9a* is known to be involved in the generation of autophagosomes. Furthermore, *Atg9a* has been shown to dampen the innate immune response that occurs in response to microbial dsDNA. *ATG9A* knockout mice show enhanced expression of *IFN- β* , *IL6* and *CXCL10* upon exposure to microbial dsDNA.⁴⁹ This gene is furthermore of interest to IBD, because the association of other autophagy genes to IBD is well established.^{6,50,51} For example, patients that are homozygous for the *ATG16L* risk allele show Paneth cell granule abnormalities.⁵² Based on the role *ATG9A* plays in responding to microbial dsDNA and the role *ATG16L* plays in Paneth cell degranulation, it is possible that *ATG9A* contributes to the IBD pathogenesis in monocytes and intestinal epithelial cells via distinct mechanisms.

Pathway analysis shows cell type specific results

Besides studying the individual associated loci and the genes they regulate, we aimed to elucidate the pathways in which the IBD candidate genes are involved. Since our approach enables us to determine both IBD candidate genes and the cell type in which they are likely dysregulated, we analyzed the pathogenic processes that are possibly involved in monocytes, lymphocytes and intestinal epithelial cells. Therefore, we performed separate pathway analyses on the datasets generated in these three different cell types. This revealed that the enriched pathways in the two immune cell types are mainly similar to each other, whereas the enrichment in epithelial cells shows different pathways (**Figure 5, Supplementary table 5**). Notably, IL10 signaling was found to be highly enriched in the intestinal epithelium dataset. This implies that the members of this pathway are possibly dysregulated in this cell type. As this pathway is also enriched in the immune cells (**Supplementary table 5**), it is likely that the contribution of IL10 signaling to the IBD pathogenesis can be found in the interplay between the intestinal epithelium and immune cells. Furthermore, several JAK/STAT and Interferon signaling pathways were highly enriched in both monocytes and lymphocytes. JAK-STAT is a common signaling pathway used by many cytokines. Dysregulation of the JAK-STAT pathway can lead to a plethora of immune diseases.⁵³ For example, tissue specific disruption of *STAT3* is known to cause an IBD-like phenotype in mice.⁵³ The high enrichment of many pathways that are relevant to IBD in the datasets of the separate cell types, supports the relevance of approaches that take cell type specific role for candidate genes into account.

HNF4 α is a potential key regulator of the IBD candidate genes

The 4C-seq approach reveals candidate genes based on their physical interaction with active regulatory regions. Transcription factors are important mediators in activating

expression from active regulatory regions. Therefore, we aimed to determine which upstream regulators are involved in the regulation of transcriptional activity of the IBD candidate genes. We used an *in silico* analysis that determines which factors regulate expression from the candidate genes and which sets of genes that are regulated by a certain upstream regulator are enriched in our cell type specific datasets. This analysis shows many significantly overrepresented upstream regulators (**Figure 6A, Supplementary table 6**), including numerous transcription factors. Notably, Hepatocyte Nuclear Factor 4 α (*HNF4 α*) is highly enriched in all three cell types. *HNF4 α* is a transcription factor that belongs to the nuclear hormone receptor superfamily.⁵⁴ Recently, the *HNF4 α* -locus was associated to IBD through a GWAS.⁵⁵ Mouse studies revealed that during intestinal inflammation, *HNF4 α* has a reduced ability to bind to active enhancers and that *HNF4 α* knock-out mice spontaneously develop colitis.^{56,57}

Our study confirms that many genes that are likely dysregulated in IBD are regulated by *HNF4 α* . Furthermore, *HNF4 α* was found to be one of our candidate genes that was identified by a distal interaction with rs6017342 in intestinal epithelial cells (**Supplementary table 2**). Upon exposure of intestinal organoids to bacteria lysate, we found that the epithelial response is characterized by a markedly upregulation of both the *NF- κ B* pathway and *HNF4 α* (**Figure 6B**). The kinetics of *HNF4 α* expression upon epithelial responses and the enrichment of *HNF4 α* -regulated genes among the IBD candidate genes, propose *HNF4 α* as a potential key regulator in IBD.

Discussion

Our study shows that using chromatin interactions for GWAS interpretation reveals many novel and relevant candidate genes for IBD. Specifically, we have intersected data on chromatin interactions, mRNA expression and H3K27Ac occupation data (marking active enhancer elements) to identify IBD candidate genes. By applying 4C-seq to cell types involved in IBD, we revealed 902 novel candidate genes, consisting of multiple noteworthy genes like *SMAD5*, *IL10RA* and *ATG9A*. Notably, many novel genes were located outside the associated loci.

There are multiple ways that can be used to identify significant interactions in 4C-seq datasets and none of these methods offer the ideal solution for all interaction ranges (long, short, inter-chromosomal), resolutions and dynamic ranges of signal.^{58,59} In this study we have selected a method that, to our opinion, provides a good balance between the specificity and sensitivity for interactions spanning up to several megabases. In order to reduce the amount of false positive findings we chose to use a stringent cut-off ($p \leq 10^{-8}$).

The identification of functional DRE-gene interactions is further established through the overlap of the candidate gene sets identified in the different cell types. Intestinal epithelial cells are developmentally and functionally very distinct from cells with a

shared hematopoietic origin, in that context monocytes and lymphocytes are more alike. These differences in overlapping background are reflected by the sets of candidate genes identified in the different cell types. Specifically, lymphocytes and monocytes shared a large part of the candidate genes, whereas intestinal epithelial cells showed a more distinct set of genes (for example, monocytes share 42% and 8% of candidate genes with lymphocytes and DLD-1 respectively, **Figure 2A and Supplementary figure 4.5**). Although this approach gives a general overview of the contribution of lymphocytes to the IBD pathogenesis, it does not enable to discriminate between mechanisms in lymphocyte subsets. Analyzing a pool of cell types also decreases the sensitivity of detection candidate genes that are specific to a subset of cells. Therefore, in future approaches, 4C datasets for specific lymphocyte subtypes can provide more insight into the contribution of each of these cell types to the IBD pathogenesis. Furthermore, since UC is limited to the colon and CD can occur throughout the intestine, creating 4C dataset from epithelium derived from different parts the intestine (i.e. duodenum, jejunum, ileum and colon) might help to discriminate between UC and CD specific pathogenic processes.

We examined the presence of eQTLs among the IBD-associated SNPs and the 4C-seq candidate genes. These analyses confirm that our approach is capable to pick up every candidate gene that was found to have SNP dependent expression levels in tissues relevant for IBD. As expected, based on the two eQTL databases that were used, not all 4C-seq candidate genes we found to have a SNP dependent expression pattern. This is (at least in part) due to the highly context specific nature of SNP dependent differential expression of many eQTLs⁶⁰. While eQTLs are usually identified at one specific cell state,⁶⁰ many SNP dependent expression patterns are only present under specific conditions (i.e. developmental stages, presence of activating stimuli etc.), resulting in a high false negative rate of eQTL detection. For example, many 4C-seq candidate genes might be differentially expressed between genotypes in the presence of pro-inflammatory stimuli. Our findings both confirm that our assay enables to detect genes with a SNP dependent expression profile and underlines the need of chromatin-based techniques to identify the genes that are missed by eQTL analyses.

By using GSEA we show that the 4C-seq candidate genes are highly enriched among genes that are upregulated in inflamed intestinal biopsies from IBD patients. Since the GSEA compares inflamed versus non-inflamed intestinal tissue within patients we cannot determine what the baseline difference in expression is between patients and healthy controls. Although the fact that a gene is upregulated upon inflammation does not show a causal relation between the (dys)regulation of that gene and the IBD phenotype, it shows the involvement of the novel 4C-seq candidate genes in IBD.

We have shown that pathway- and upstream regulator- enrichment algorithms can be used to interpret and prioritize this large candidate gene dataset. Interpretation of the 4C-seq data can be further optimized by using this data in a quantitative manner (i.e. correlating peak strength instead of using a cut-off value for peak calling). However,

as with all approaches for candidate gene identification, further validation is needed to identify the causal genes for IBD. The first step towards this confirmation will in this case consist of revealing the dysregulation of the candidate gene expression upon alteration of the enhancer function *in vivo*.

We have profiled the chromatin interactions in primary cells from healthy controls and a cell line, to create a profile of the genes that physically interact with the IBD susceptibility loci under normal conditions in peripheral immune cells derived from healthy individuals and in intestinal epithelium derived cell line. As the effects of common variants in regulatory regions are relatively mild, it is improbable that a single common variant that is present in an IBD patient will ablate or create a whole regulatory region and its 3D interaction.³⁹ We therefore do not expect that the identification of candidate genes in cells derived from patients will reveal a substantial number of additional interactions. On the other hand, these variants are expected to cause dysregulation of the candidate genes and thereby contribute to the disease, possibly under very specific conditions, i.e. during certain stages of development or in presence of specific stimuli.^{16,60}

Our study provides a proof of principle for the usage of chromatin-chromatin interactions for the identification of candidate genes. The approach presented here complements, but does not replace, previously reported approaches for candidate gene identification.¹¹ Candidate gene prioritization models for GWASs currently use multiple types of information, for example protein-protein interactions, expression patterns and gene ontology. We propose that these algorithms should take chromatin interactions into account to optimize gene prioritization. Furthermore, the intensity of the 4C-seq signal combined with the expression levels of 4C-candidate genes can be of added value.

We conclude that 4C-seq and other 3C-derived methods can be applied to candidate gene identification in diseases with a complex genetic background and complement the classical candidate gene identification approaches.

References

1. Kaser, A., Zeissig, S. & Blumberg, R. S. Inflammatory bowel disease. *Annu. Rev. Immunol.* 28, 573–621 (2010).
2. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42, 1118–1125 (2010).
3. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–24 (2012).
4. Anderson, C. A. et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* 43, 246–52 (2011).
5. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–989 (2015).
6. Rioux, J. D. et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* 39, 596–604 (2007).
7. Libioulle, C. et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 3, e58 (2007).
8. Duerr, R. H. et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314, 1461–3 (2006).
9. Mokry, M. et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in

- immune cells and the intestinal epithelium. *Gastroenterology* 146, 1040–7 (2014).
10. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 5, e1000534 (2009).
 11. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81, 1278–83 (2007).
 12. Rivas, M. A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* 43, 1066–73 (2011).
 13. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–86 (2014).
 14. Maurano, M. T. et al. Systematic Localization of Common Disease-Associate Variation in Regulatory DNA. *Science* (80-.). 337, 1190 (2012).
 15. Kleinjan, D. J. & Coutinho, P. Cis-rupture mechanisms: Disruption of cis-regulatory control as a cause of human genetic disease. *Briefings Funct. Genomics Proteomics* 8, 317–332 (2009).
 16. Schaub, M. a, Boyle, A. P., Kundaje, A. & Frazer, K. a. Linking disease associations with regulatory information in the human genome Toward mapping the biology of the genome. 1748–1759 (2012). doi:10.1101/gr.136127.111
 17. McVicker, G. et al. Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–9 (2013).
 18. Kilpinen, H. et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342, 744–7 (2013).
 19. Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112 (2009).
 20. de Laat, W. et al. Three-Dimensional Organization of Gene Expression in Erythroid Cells. *Curr. Top. Dev. Biol.* 82, 117–139 (2008).
 21. Hughes, J. R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* 46, 205–12 (2014).
 22. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* 489, 75–82 (2012).
 23. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368, 20120362 (2013).
 24. Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. 30, 1411–1420 (2010).
 25. Mifsud, B. et al. Sup Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47, 598–606 (2015).
 26. Jäger, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* 6, 6178 (2015).
 27. Martin, P. et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* 6, 10069 (2015).
 28. van de Werken, H. J. G. et al. 4C technology: protocols and data analysis. *Methods in enzymology* 513, (Elsevier Inc., 2012).
 29. Ji, H. et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* 26, 1293–300 (2008).
 30. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–5 (2013).
 31. Hägg, S. et al. Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study. *PLoS Genet.* 5, (2009).
 32. Foroughi Asl, H. et al. Expression quantitative trait Loci acting across multiple tissues are enriched in inherited risk for coronary artery disease. *Circ. Cardiovasc. Genet.* 8, 305–15 (2015).
 33. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–50 (2005).
 34. Noble, C. L. et al. Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut* 57, 1398–405 (2008).
 35. NCBI. GEO2R. at <<http://www.ncbi.nlm.nih.gov/geo/geo2r/>>
 36. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–5 (2005).
 37. Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett’s epithelium. *Gastroenterology* 141, 1762–72 (2011).
 38. Dekkers, J. F. et al. A functional CFTR assay using primary cystic fibrosis intestinal organoids. *Nat. Med.* 19, 939–45 (2013).
 39. Vernot, B. et al. Personal and population genomics of human regulatory variation. *Genome Res.* 22, 1689–97 (2012).

40. Saitoh, T. & Akira, S. Regulation of innate immune responses by autophagy-related proteins. *J. Cell Biol.* 189, 925–35 (2010).
41. Murphy, T. L., Tussiwand, R. & Murphy, K. M. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nat. Rev. Immunol.* 13, 499–509 (2013).
42. Lin, F., Spencer, D., Hatala, D. a, Levine, A. D. & Medof, M. E. Decay-accelerating factor deficiency increases susceptibility to dextran sulfate sodium-induced colitis: role for complement in inflammatory bowel disease. *J. Immunol.* 172, 3836–3841 (2004).
43. Glocker, E.-O. et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N. Engl. J. Med.* 361, 2033–45 (2009).
44. Liu, B., Tahk, S., Yee, K. M., Fan, G. & Shuai, K. The ligase PIAS1 restricts natural regulatory T cell differentiation by epigenetic repression. *Science* 330, 521–525 (2010).
45. Allaire, J. M. et al. Loss of Smad5 leads to the disassembly of the apical junctional complex and increased susceptibility to experimental colitis. *Am. J. Physiol. Gastrointest. Liver Physiol.* 300, G586–G597 (2011).
46. Portillo, J. C., Greene, A., Schwartz, I., Subauste, M. C. & Subauste, C. S. Blockade of CD40 – TRAF2 , 3 or CD40 – TRAF6 is sufficient to inhibit pro-inflammatory responses in non-haematopoietic cells. *Immunology* 21–33 (2014). doi:10.1111/imm.12361
47. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* 489, 109–13 (2012).
48. Bell, A. C., West, A. G. & Felsenfeld, G. The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell* 98, 387–396 (1999).
49. Saitoh, T. et al. Atg9a controls dsDNA-driven dynamic translocation of STING and the innate immune response. *Proc. Natl. Acad. Sci. U. S. A.* 106, 20842–6 (2009).
50. Hampe, J. et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.* 39, 207–11 (2007).
51. Parkes, M. et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn’s disease susceptibility. *Nat. Genet.* 39, 830–2 (2007).
52. Cadwell, K. et al. A key role for autophagy and the autophagy gene Atg16l1 in mouse and human intestinal Paneth cells. *Nature* 456, 259–63 (2008).
53. Shuai, K. & Liu, B. Regulation of JAK-STAT signalling in the immune system. *Nat. Rev. Immunol.* 3, 900–911 (2003).
54. Sladek, F. M., Zhong, W., Lai, E. & Darnell, J. E. Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev.* 4, 2353–2365 (1990).
55. Barrett, J. C. et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* 41, 1330–4 (2009).
56. Chahar, S. et al. Chromatin profiling reveals regulatory network shifts and a protective role for hepatocyte nuclear factor 4 α during colitis. *Mol. Cell. Biol.* 34, 3291–304 (2014).
57. Darsigny, M. et al. Loss of hepatocyte-nuclear-factor-4 α affects colonic ion transport and causes chronic inflammation resembling inflammatory bowel disease in mice. *PLoS One* 4, e7609 (2009).
58. de Wit, E. et al. CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell* 60, 676–684 (2015).
59. Raviram, R. et al. 4C-ker: A Method to Reproducibly Identify Genome-Wide Interactions Captured by 4C-Seq Experiments. *PLoS Comput. Biol.* 12, e1004780 (2016).
60. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 343, 1246949 (2014).



Additional candidate genes for human
atherosclerotic disease identified
through annotation based on
chromatin organization.

5

Claartje A. Meddens* ,Saskia Haitjema* , Sander W. van der Laan,
Daniel Kofink, Magdalena Harakalova, Vinicius Tragante, Hassan
Foroughi Asl, Jessica van Setten, Maarten M. Brandt, Joshua C. Bis,
Christopher O'Donnell, Caroline Cheng, Imo E. Hofer, Johannes
Waltenberger, Erik Biessen, J. Wouter Jukema, Pieter A. F. M.
Doevendans, Edward E. S. Nieuwenhuis, Jeanette Erdmann, Johan L. M.
Björkegren, Gerard Pasterkamp, Folkert W. Asselbergs, Hester M. den
Ruijter, and Michal Mokry

Based on: *Circulation: Cardiovascular Genetics* 10(2) 2017

Introduction

Atherosclerosis is a chronic inflammatory disease of the lipid-rich vascular wall that underlies many cardiovascular diseases (CVD).¹ A large part of the disease burden of atherosclerosis can be traced back to coronary artery disease (CAD) and large artery stroke (LAS). Genome-wide association studies (GWAS) have helped to unravel the complex genomic background of these diseases, currently explaining about 10% of heritability.^{2,3} The current approach is to annotate a novel susceptibility locus with the gene at the nearest genomic position. Some alternative strategies also take into account gene expression or protein-protein interactions.^{4,5} A recent effort employing these bioinformatics-based approaches resulted in 98 new candidate genes for CAD.⁶ In the last few years, the evidence that variants identified by GWAS also contribute to the disease pathogenesis by affecting the regulatory DNA sequences they reside in is growing.^{7,9} These genetic variants may affect the activity of the DNA regulatory elements (DRE) and, under specific circumstances, lead to dysregulation of gene expression. This is mediated by long range 3D chromatin-chromatin interactions where the regulated candidate genes can be located up to ~1 MB away¹⁰⁻¹² a distance much larger than is normally

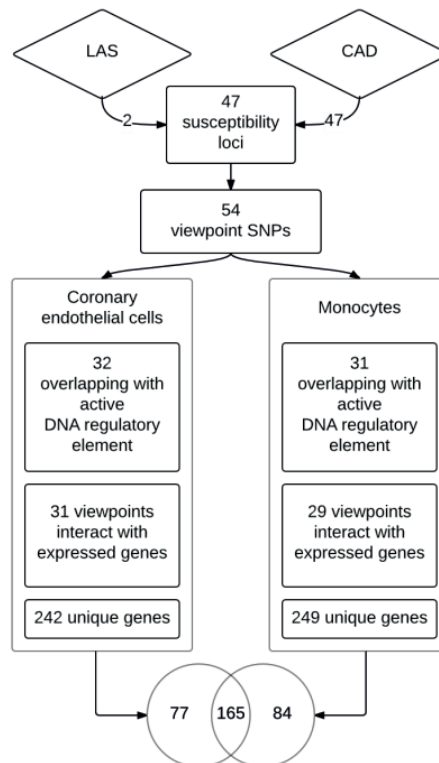


Figure 1 Flowchart of identification of candidate genes.

Susceptibility loci: SNPs associated with risk of disease in METASTROKE and/or CARDIoGRAM. Viewpoint SNPs: SNPs used as the focus point for the primer design of the 4C experiment.

used to annotate candidate genes in GWAS. These candidate genes can be identified by capturing the physical chromatin-chromatin interaction between a known disease susceptibility locus and the promoter of the gene(s) it presumably regulates.¹³ Here we systematically apply this principle (study design is summarized in **Figure 1**) to variants identified by large meta-analyses of GWAS for CAD and LAS; altogether assaying 47 previously identified susceptibility loci.^{2,3} Atherosclerotic disease starts in the endothelial lining of the affected arteries and involves attraction and proliferation of monocytes.¹⁴ Therefore, we studied 37 loci that co-localize with active DRE in human monocytes and/or in cardiac endothelial cells. We used circular chromosome conformation capture sequencing (4C-seq) to identify candidate genes based on their physical interaction with one of the active DRE.

Methods

Cell culture

Primary commercially available human cardiac endothelial cells (CEC) that were isolated by enzymatic detachment (Lonza *Clonetics*[™]) were cultured in RPMI-1640 with 10% FCS and standard supplements. Cells were harvested for 4C template preparation by trypsinisation at 60-80% confluence.

Monocyte isolation

Human peripheral blood was collected from a healthy donor in sodium-heparin tubes. Peripheral Blood Mononuclear Cells (PBMCs) were isolated by Ficoll-Paque gradient centrifugation. PBMCs were incubated with magnetic CD14⁺-microbeads (Milteny, order nr. 130-050-201) according to manufacturer's manual. Thereafter cells were magnetically separated by the AutoMACS[™] Separator, the positive fraction (monocytes) was used for 4C template preparation.

Circular Chromosome Conformation Capture-Template preparation

The 4C template was prepared as previously described.¹³ Summarized, 10x10⁶ cells were used per cell type (monocytes and CEC). Cells were crosslinked in 2% formaldehyde. After chromatin isolation, the chromatin was digested with *DpnII* (NEB, # R0543L). Digestion was stopped through heat inactivation of the restriction enzyme. Samples were diluted and ligated by T4 DNA ligase. The second digestion was carried out with *CviQI* (NEB, #R069S) and inactivated by phenol:chloroform extraction. The chromatin was diluted, for the final T4 ligation and the chromatin was purified. The quality of digestion and ligation was assessed on agarose gels.

Viewpoint selection and primer design

All SNPs from table 1 and 2 and the young-CAD SNP (rs16986953) from the CARDIoGRAMplusC4D paper(2) (n=47) and the two replicated SNPs (rs2383207 and rs2107595) from METASTROKE³ were considered for viewpoint design (**Supplemental table 8**). When the SNPs within the susceptibility locus were less than 15,000 bp apart (e.g. rs12740374, rs602633 and rs599839 in the *SORT1* region), only one SNP was selected as a viewpoint. Susceptibility loci that overlapped with active DRE were identified through FAIRE, the presence of H3K4Me1, H3K4Me3, H3K27Ac, H3K4Me2 or H3K9Ac, EP300 or CTCF binding sites or DNase hypersensitivity sites (**Supplemental table 9**). DRE falling within the susceptibility locus coordinates were considered overlapping with the susceptibility locus. The primers were designed as was described previously.¹³ Primer sequences are listed in **Supplemental table 8**. In summary, primers were designed in a window of 5 kb up- and downstream from the associated SNP. Forward and reverse primers were designed at least 300 bp apart. Forward (reading) primers were designed on top of the first restriction enzyme site. The reverse (non-reading) primer was designed close to (max 100bp away from) the second restriction

enzyme site. In case no primer pair could be designed within the initial window, the window was extended 5 kb up- and downstream (n=8). In the case of rs1561198 this did not result in a suitable primer, so a primer pair that was 299bp apart was selected for this viewpoint.

Circular Chromosome Conformation Capture- Sequencing (4C-seq) library preparation

4C-sequencing library preparation was performed as described previously³³, with minor adaptations in order to make the protocol compatible with the large number of viewpoints: the PCR of 4C template was performed with 600 ng (monocytes) or 1,6 µg (coronary endothelial cells) of 4C template per reaction. 8 to 10 primer pairs were multiplexed in the initial PCR reaction (primer sequences are listed in **Supplemental table 8**). Primers pairs were pooled according to primer efficiency (based on intensity on gel electrophoresis signal after PCR on test template). PCR products were purified after an initial PCR reaction of 6 cycles (reaction volume = 200 µL) and divided among 8-10 PCR reactions containing single primer pairs for another 26 cycles (reaction volume = 25 µL). Thereafter, PCR products derived from the same cells were pooled in equimolar amounts and a final 6 cycle PCR reaction containing 20 ng of pooled PCR product (reaction volume = 100 µL) was performed with primers that contained sequencing adaptor sequences. All fragments >700bp were removed using size selection on a 1% agarose gel followed by gel extraction of the selected products (Qiagen, #28704). Quality measures for the 4C library preparation and sequencing can be found in **Supplemental figure 5.1**.

Sequencing

Libraries were sequenced using the HiSeq2500 platform (Illumina), according to the manufacturer's protocol, producing 50 bp single end reads.

Data analysis

The raw sequencing reads were de-multiplexed based on viewpoint specific primer sequences. Reads were then trimmed to 16 bases and mapped to an *in silico* generated library of fragends (fragment ends) neighboring all *DpnII* sites in human genome (NCBI37/hg19), using the custom Perl scripts. No mismatches were allowed during the mapping and the reads mapping to only one possible fragend were used for further analysis.

Identification of the interacting genes

First, we calculated the number of covered fragends within a running window of *k* fragends throughout the whole chromosome where the viewpoint is located. The *k* was set separately for every viewpoint so it contains on average 20 covered fragends in the area around the viewpoint (+/- 100kb). Next, we compared the number of covered fragends in each running window to the random distribution. The windows with significantly higher number of covered fragends compared to random distribution ($p < 10^{-8}$ based on binominal cumulative distribution function; *R pbinom*) were considered as significant 4C-seq signal. The following criteria were defined for the identification of the candidate genes; i) the Transcriptional Start Site (TSS) co-localizes with a significant 4C-seq signal ($P < 10^{-8}$) within 5 kbp; ii) the susceptibility variant or other variant in linkage disequilibrium (LD) co-localizes with a DNA regulatory element identified through FAIRE, the presence of H3K4Me1, H3K4Me3, H3K27Ac, H3K4Me2 or H3K9Ac, EP300 or CTCF binding sites or DNase hypersensitivity sites (**Supplemental table 9**) in the cell type from which the 4C-seq signal originated and iii) the gene is expressed (RPKM > 0.5) in the assayed cell type.

Identification of gene expression

For monocyte expression, data from the ENCODE database were used (**Supplemental table 10**).¹⁵ For coronary endothelial cell expression, HMVECs (Lonza) were cultured on gelatine coated plates in EGM2-MV (Lonza) supplemented with penicillin and streptomycin. Subsequently, HMVECs were cultured for 20 hours in low serum medium (EBM + 0.5% FCS), followed by cell lysis and RNA isolation using the RNeasy isolation kit (Qiagen). Polyadenylated mRNA was isolated using Poly(A) Beads (NEXTflex). Sequencing libraries were made using the Rapid Directional RNA-Seq Kit (NEXTflex) and sequenced on Illumina NextSeq500 to produce single-end 75 base

long reads (Utrecht Sequencing Facility). Reads were aligned to the human reference genome GRCh37 using STAR version 2.4.2a.¹⁶ Read groups were added to the BAM files with Picard's AddOrReplaceReadGroups (v1.98). The BAM files are sorted with Sambamba v0.4.5 and transcript abundances are quantified with HTSeq-count version 0.6.1p1¹⁷ using the union mode. Subsequently, reads per kilobase of transcript per million reads sequenced (RPKM's) are calculated with edgeR's `rpkm()` function.¹⁸

Pathway analysis

The interacting genes (with and without expressed CARDIoGRAMplusC4D/METASTROKE genes) were analyzed using QIAGEN's Ingenuity Pathway Analysis (IPA, 2015 winter version, QIAGEN Redwood City, www.qiagen.com/ingenuity). We used IPA to identify canonical biological pathways within the Ingenuity Knowledge Base to which the interacting genes were mapped. Limits were set to only direct relationships that were experimentally observed in humans. We performed six rounds of pathway analysis, three in each of the cell types: one with only CARDIoGRAMplusC4D/METASTROKE genes that were expressed in the cell type, one with the CARDIoGRAMplusC4D/METASTROKE genes supplemented by the newly identified genes and one with the novel genes only.

Tracks and plots

All tracks were accessed from the UCSC browser (hg19) (<http://genome.ucsc.edu/>). Regional plots were generated using LocusZoom version 1.3.¹⁹

Gene-based tests

Data for CAD were downloaded from the CARDIoGRAMplusC4D website (<http://www.cardiogramplusc4d.org>). We obtained summary statistics from GWAS on body mass index (BMI), blood lipids including LDL, HDL, total cholesterol and triglycerides, systolic and diastolic blood pressure, coronary calcification, fasting glucose, smoking behavior, and type 2 diabetes from public online resources and data on intima-media thickness and plaque-presence via data request (**Supplemental Table 11**). We used a VEratile Gene-based Association Study (VEGAS) to calculate gene-based association statistics from the summary statistics of each interacting gene for each trait. The details of the methods applied by VEGAS have been described elsewhere.²⁰ In short, SNPs are mapped to the gene (in and around ± 50 kb from 5' and 3' gene borders), and using the GWAS *p*-value a gene-based test statistic is calculated corrected for the underlying population linkage disequilibrium structure. Finally using simulations an empirical gene-based *p*-value of association with the phenotype is calculated per gene. VEGAS results were considered multiple testing significant if they were $P < 6.97 \times 10^{-6}$ ($0.05/22$ phenotypes \times 326 available genes in VEGAS).

eQTL analysis in STAGE

Within the STAGE study, patients undergoing coronary artery bypass grafting (CABG) surgery were sampled for seven different tissues, namely atherosclerotic arterial wall (AAW), internal mammary artery (IMA), liver, skeletal muscle (SM), subcutaneous fat (SF), visceral fat (VF), and fasting whole blood (WB) for RNA and DNA isolation.²¹ Patients that were eligible for CABG and had no other severe systemic diseases (e.g. widespread cancer or active systemic inflammatory disease) were included. For quality control in genotyping, SNPs filtered for minor allele frequency $MAF < 5\%$, Hardy-Weinberg equilibrium (HWE) *p*-value $< 1 \times 10^{-6}$, and call rate of 100%. Imputation was carried out using IMPUTE2 with 1000 Genomes EUR as the reference.²² Quality control for imputed genotypes used additionally an IMPUTE2 INFO score filter < 0.3 . After QC a total of 5,473,585 SNPs remained. The Ethical committee of the Karolinska Hospital approved the study, and all patients gave written informed consent after the nature and possible consequences of the study were explained. An expression trait was tested for association with each genotyped and imputed SNP using Kruskal-Wallis test and false discovery rate to correct for multiple testing as described before. First, all *cis*-pairs of SNPs within 50kb of the transcription start or end site for each gene were identified. Next, *cis* SNP-gene pairs were tested for association in all seven STAGE tissues using `kruX`.²³ The *p*-value for eQTL inclusion in `kruX` was set at 0.05. Finally, an empirical

FDR estimate for each eQTL-gene pair was calculated using ten permutations by shuffling patient IDs on genotype data. As a result, the most significant eQTL-gene association in each tissue was reported.

eQTL analysis in Haploreg

Data on eQTL in healthy individuals were extracted from Haploreg version 4.1 (<http://www.broadinstitute.org/mammals/haploreg/haploreg.php>). The viewpoint SNPs and all SNPs in LD ($r^2 > 0.8$) were used as input. From the output, for each interacting gene, the most significant eQTL within each tissue was extracted.

eQTL analysis in CTMM circulating cells

CTMM circulating cells is a Dutch cohort from four different hospitals comprising of 714 patients undergoing coronary angiography of whom blood was stored. Monocytes were isolated by density centrifugation followed by positive magnetic bead isolation (CD14) and expression was measured using the Illumina humanHT-12 v3 Gene Expression BeadChip Array. After removal of samples with a median intensity of <50 , 370 patients were included in the analysis. The data were quantile normalized and \log_2 transformed using the lumi R package.²⁴

Genotyping was performed using a customized Affymetrix Axiom Tx array containing 767,203 genetic markers. Community standard quality control was performed, filtering out samples with missingness $>5\%$, outlying heterozygosity ($\pm 4SD$ from the cohort mean) or inconsistent sex. Samples of non-European descent or those that were out of Hardy-Weinberg equilibrium ($p < 5 \times 10^{-5}$) were removed. In total, 622 were used in the current analysis. Untyped variants were imputed using a combined reference panel of the 1000 Genomes Project²⁵ and Genome of the Netherlands²⁶ totaling more than 90 million genetic variants across the genome. We used the software packages SHAPEIT²⁷ for phasing and IMPUTE2(22) for imputation. Prior to *cis*-eQTL analysis we filtered the imputed genotype data from CTMM based on $MAF > 0.5\%$, $HWE P > 1 \times 10^{-6}$, $Info\text{-}metric > 0.9$, and only focused on those variants in LD ($r_2 \geq 0.8$) with the CAD associated variants. We then used fastQTL (v2.184)²⁸ to perform I-eQTL analyses using a fixed range (based on the 4C interactions) around each probeID available on the expression array.

Mouse knockout models

Murine gene names were mapped to the genes as follows. First, a custom data file was downloaded from the HUGO Gene Nomenclature Committee (<http://www.genenames.org/cgi-bin/download>) including the Approved gene name and the Mouse Genome Database ID from the Mouse Genome Informatics database and a file containing all available phenotypic information for all knockout mice was downloaded from MGI (<ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno>). Next, for all approved gene names of genes identified through 4C-seq, the mouse phenotypes were looked up by linking the MGI IDs. If no linkage could be made for the MGI ID, this was coded as no available mouse model. If a mouse model was available, but no phenotype was found, this was coded as no available phenotype. If a mouse model was specifically coded as not showing any phenotype upon knockout, this was coded as a gene not resulting in any phenotype. Murine cardiovascular phenotypes were defined as a phenotype resulting in any of the following: impaired blood coagulation or abnormal platelets, abnormal glucose levels or homeostasis, abnormal vascular morphology, vascular remodelling or arterial differentiation, abnormal blood pressure, abnormal vasoconstriction, vasodilatation or vascular permeability, abnormal stress response of the heart, myocardial infarction, abnormal (circulating) lipid levels, abnormal fat morphology or amount, abnormal body weight, abnormal lipid droplet or fat cell size, abnormal macrophage response or inflammation, abnormal wound healing, arteritis, vasculitis, vascular occlusion or atherosclerosis.

Human knockout models

The interacting genes were extracted from the supplementary tables of the studies of Sulem *et al.* and MacArthur *et al.*^{29,30} For each of the interacting genes, all SNPs and indels resulting in human functional knockouts were reported.

Drug targets

For the lookup of existing drugs that target any of the candidate genes, we used a custom built drug pipeline that searches for drug-gene interactions using DGIdb³¹, which merged the most known drug-gene interaction databases, such as DrugBank³² and PharmGKB³³. We removed redundant results using STITCH³⁴ and WHO's INN³⁵. We tested overrepresentation of drug groups according to ATC codes³⁶ using Fisher exact tests.

Results

4C-seq identifies additional candidate genes

We identified 37 active DNA regulatory elements that co-localize with susceptibility loci for CAD or LAS. Twenty-six were active in both monocytes and coronary endothelial cells, 5 were only active in monocytes and 6 were only active in coronary endothelial cells (**Supplemental table 1**). To identify the target genes of these active DRE, we generated 63 4C-seq interaction datasets. We applied the following criteria for the identification of candidate genes: I) the transcriptional start site (TSS) co-localizes with a significant 4C-seq signal ($P < 10^{-8}$) within 5kb; II) the susceptibility variant or any other variant in LD ($r^2 \geq 0.8$) co-localizes with an active DRE signal in the cell type from which the 4C-seq signal was obtained and III) the gene is expressed (RPKM > 0.5) in the studied cell type. With this approach, we identified 326 candidate genes (**Supplemental table 1**), 77 in human male coronary endothelial cells, 84 in human male monocytes and 165 in both cell types (**Figure 1**). In total, we identified 294 candidate genes that were not previously reported by the CAD and LAS GWAS (**Supplemental table 1**). We replicated 235/242 (97.1%) of the chromatin interactions with expressed genes that were identified in male coronary endothelial cells in female coronary endothelial cells (**Supplemental table 1**).

4C-seq identifies candidate genes in novel pathways

We performed cell-type specific pathway analysis of the candidate genes identified by 4C-seq combined with the candidate genes that were previously identified by the GWAS on CAD and LAS (**Supplemental table 2**). Notably, these analyses revealed the *Hypoxia signaling in the cardiovascular system* pathway in monocytes ($P = 0.01$) and the *NRF-mediated oxidative stress response* pathway ($P = 4.68 \times 10^{-4}$ and $P = 0.026$ in coronary endothelial cells and monocytes respectively, **Supplemental table 2**). These pathways are both involved in the cellular response to oxidative stress. Additionally, the 4C-seq approach revealed *PTEN* (a player in the *Hypoxia signaling in the cardiovascular system* pathway) as a novel candidate gene (**Supplemental table 1**). Although this gene was never reported via previous GWAS annotation, *PTEN* (phosphatase and tensin analog) was found to be a likely candidate gene based on dose-dependently upregulation by statins through higher peroxisome proliferator-activated receptor-gamma (PPAR gamma) activity.³⁷ A mutation of *PTEN* led to inflammatory plaque characteristics in human atherosclerotic plaque³⁸ and increased stability of *PTEN* was found to ameliorate atherosclerosis³⁹. Furthermore, *PTEN* shares its upstream transcription regulator *ZEB2* with *CDKN2A* and *CDKN2B* (enrichment P of overlap for *ZEB2*-regulated genes: 3.02×10^{-3}

Chr	Susceptibility Locus	4C-Seq Viewpoint(s)	Gene Identified by 4C-Seq	Cell Type of Identification		eQTL
				Coronary Endothelial Cells	Monocytes	
1	MIA3	rs17464857	<i>AIDA</i>	✓	✓	
			<i>BROX</i>	✓	✓	
			MARC1	✓		
			<i>MIA3</i>	✓	✓	
			<i>TAF1A</i>	✓	✓	
1	SORT1	rs12740374	<i>AMPD2</i>	✓	✓	
			<i>ATXN7L2</i>	✓	✓	
			<i>CYB561D1</i>	✓	✓	
			<i>GNAI3</i>	✓	✓	
			<i>GNAT2</i>	✓		
			<i>GSTM2</i>	✓	✓	
			<i>GSTM4</i>	✓	✓	
			<i>PSMA5</i>	✓	✓	
			PSRC1	✓	✓	Monocytes
			<i>SARS</i>	✓	✓	
			SORT1	✓	✓	
			2	APOB	rs515135	<i>LDHA</i>
2	VAMP5-VAMP8-GGCX	rs1561198	GGCX	✓	✓	CEC
			<i>C2orf68</i>	✓	✓	
			<i>ELMOD3</i>	✓	✓	
			<i>MAT2A</i>	✓	✓	
			<i>RETSAT</i>	✓	✓	
			<i>RNF181</i>	✓	✓	
			<i>TGOLN2</i>	✓	✓	
			<i>TMEM150A</i>	✓	✓	
			<i>USP39</i>	✓	✓	
			VAMP5	✓	✓	
			VAMP8	✓	✓	Monocytes
			<i>CARF</i>		✓	
2	WDR12	rs6725887	<i>FAM117B</i>	✓	✓	CEC/Monocytes
			NBEAL1	✓	✓	CEC
			WDR12	✓	✓	
3	MRAS	rs9818870	MRAS	✓	✓	CEC
6	ANKS1A	rs12205331	<i>C6orf106</i>	✓	✓	
			<i>RPS10</i>	✓	✓	
			<i>SNRPC</i>	✓	✓	
			<i>UHRF1BP1</i>	✓	✓	
6	PHACTR1	rs9369640, rs12526453	<i>MYLIP</i>		✓	
			PHACTR1	✓	✓	CEC
8	LPL	rs264	<i>INTS10</i>	✓	✓	
8	TRIB1	rs2954029	TRIB1	✓	✓	
9	CDKN2BAS	rs1333049, rs3217992, rs2383207	CDKN2A	✓		
			CDKN2B	✓	✓	
10	CYP17A1-CNNM2-NT5C2	rs12413409	<i>ARL3</i>	✓	✓	CEC
			<i>USMG5</i>	✓	✓	Monocytes
10	CNNM2	rs12413409	<i>BORCS7</i>	✓	✓	
			<i>WBP1L</i>	✓	✓	
10	KIAA1462	rs2505083	KIAA1462	✓		CEC
10	LIPA	rs11203042, rs2246833	LIPA	✓	✓	CEC/Monocytes
13	COL4A1-COL4A2	rs4773144	COL4A1	✓		
			COL4A2	✓		
17	RAI1-PEMT-RASD1	rs12936587	PEMT	✓	✓	Monocytes
			RASD1	✓	✓	Monocytes
17	SMG6	rs2281727	<i>SRR</i>	✓	✓	CEC
			SMG6	✓	✓	
17	UBE2Z	rs15563	<i>CALCOCO2</i>	✓	✓	Monocytes
			<i>KPNB1</i>		✓	
			UBE2Z	✓	✓	Monocytes
19	LDLR	rs1122608	<i>C19orf52</i>		✓	
			<i>CARM1</i>		✓	
			LDLR		✓	
			<i>SMARCA4</i>			
			<i>TSPAN16</i>		✓	
			<i>YIPF2</i>		✓	

Table 1. Candidate genes identified by 4C-seq in human coronary endothelial cells and/or human monocytes.

*Only genes that have their interacting viewpoint as eQTL or genes with a significant gene-based P value ($P < 6.97 \times 10^{-6}$ by gene-based test using VEGAS (corrected for multiple-testing $0.05/22$ phenotypes \times 326 available genes) are depicted. The full list of genes identified by 4C-seq can be found in Supplemental table 1. Susceptibility locus: name of the locus as given by CARDIoGRAMplusC4D or METASTROKE; viewpoint: SNP used as the focus point for the primer design of the 4C experiment; Gene: gene physically interacting with viewpoint, determined by 4C-seq; Underlined genes: genes that have previously been reported by CARDIoGRAMplusC4D or METASTROKE; Chr: chromosome; eQTL: expression quantitative trait locus; GWAS: genome-wide association study CAD: coronary artery disease; BMI: body mass index; TC: total cholesterol; LDL: low-density lipoprotein; HDL: high-density lipoprotein; TG: triglycerides

in monocytes and 3.06×10^{-3} in coronary endothelial cells). We reveal multiple novel pathways related to cardiovascular disease and we now show that *PTEN* physically interacts with a DRE at rs2246833 in monocytes (P interaction = 2.36×10^{-10}).

Expression of identified genes is genotype dependent

DRE exert their function through regulation of gene expression. We explored this mechanism by studying expression quantitative trait loci: the GWAS SNPs (or a SNP in LD; $r^2 > 0.8$) that significantly affected the expression of the candidate genes identified by 4C-seq in the studied tissues (**Table 1**). For the candidate genes identified by 4C-seq in coronary endothelial cells, we performed lookups within eQTL data of atherosclerotic artery wall and internal mammary artery in the STAGE cohort of patients undergoing cardiac bypass surgery. We identified two eQTL ($FDR < 0.1$) in atherosclerotic artery wall (rs9818870: *MRAS* and rs2281727: *SRR*, **Supplemental table 3A**). The *SRR* gene, that has not been reported previously, encodes for the serine racemase enzyme that is an endogenous ligand of the glycine site of NDMA receptors in the brain. Blockage of this site was found to prevent stroke damage.⁴⁰ Interestingly, a set of twice the number of genes from the same genetic locus that were not identified by 4C-seq as a target gene resulted in no significant eQTL in STAGE. Using the HaploReg tool, we additionally examined expression in aorta, coronary artery and tibial artery tissue and identified another seven eQTL for genes that we identified in coronary endothelial cells (**Supplemental table 3B**), of which *ARL3* and *FAM117B* were not reported before. Both genes are poorly studied in the context of cardiovascular disease. Within the *VAMP5-VAMP8-GGCX* locus we replicate rs1561198, that was previously reported to be an eQTL for *GGCX* in mammary artery by the CARDIoGRAMplusC4D investigators in the ASAP study⁴¹, as an eQTL for *GGCX* in aorta and tibial artery.

For genes identified by 4C-seq in monocytes, we performed *cis*-eQTL analysis in monocytes from 370 patients undergoing coronary angiography for coronary artery atherosclerosis in the CTMM (Center for Translational Molecular Medicine) Circulating Cells cohort.⁴² We identified four eQTL ($FDR < 0.1$) of which the genes overlap with genes identified by 4C-seq in monocytes of these patients (rs12740374: *PSRC1*, rs1561198: *VAMP8*, rs2246833: *LIPA*, rs12413409: *USMG5*, **Supplemental table 3C**). Previously, the CARDIoGRAMplusC4D investigators also identified rs1561198 as an eQTL for *VAMP8* in

lymphoblastoid cells and skin in the MuTHER study.⁴³ Inclusion of the previously published cardiovascular cohort of Zeller et al. revealed five additional genes (**Supplemental table 3D**).⁴⁴ The SNP that revealed the strongest association with gene expression of *PSRC1* in monocytes of CTMM (rs7528419) is in perfect LD (1) with rs12740374 in the *SORT1* region. Interestingly, whereas the minor allele of the latter SNP is known to increase *SORT1* expression in liver, we found no such association between rs7528419 and *SORT1* expression in monocytes (nominal $P = 0.87$). In addition, we found an association between higher *PSRC1* expression in monocytes with a more severe atherosclerotic phenotype identified by a higher atherosclerotic burden, quantified using the SYNTAX score ($P = 0.003$). This association with high atherosclerosis burden could not solely be explained by LDL levels, the putative mechanism through which *SORT1* expression affects cardiovascular disease phenotypes (P when corrected for circulating LDL levels = 0.01). Expression of *PSRC1* in whole blood has previously been associated with cardiovascular disease in an Asian population.⁴⁵ Largely because the functional significance of the minor allele of rs12740374 as a transcription factor binding site that increases *SORT1* expression directly, no further attention has been given to alternative candidate genes in the *SORT1* region. With our 4C-seq approach in monocytes, we here show first evidence that the expression of *PSRC1*, a candidate gene in the *SORT1* locus, is genotype-dependent expressed in monocytes and related to the severity of atherosclerosis. This example further supports the implication of our additionally identified candidate genes in cardiovascular disease.

Additional genetic annotation

We further explored current genetic knowledge for the candidate genes identified through 4C-seq (**Table 1, Supplemental table 4-7**). First, if the candidate genes are effector genes of the DREs within CVD susceptibility loci, one would expect the genes to be enriched for (common) variants associated with CVD. Using the VEGAS algorithm, we concatenated GWAS p-values of all single-nucleotide polymorphisms (SNPs) in or within 50kb of a gene into a p-value for that particular gene. This way, we studied the genes identified by 4C-seq in published and unpublished GWAS data studying a total of 22 traits, either surrogate markers of atherosclerosis or known risk factors for cardiovascular disease (**Supplemental table 4**). Of all 326 candidate genes, 33 showed a significant association ($P < 0.05/(22 \times 326) = 6.97 \times 10^{-6}$) with coronary artery disease in the CARDIoGRAMplusC4D GWAS and 149 were nominally associated with coronary artery disease in the CARDIoGRAMplusC4D GWAS (significant enrichment, 149/326: binomial $P = 2.9 \times 10^{-102}$). Additionally, we found 7 genes that were significantly associated with BMI in GIANT and 29 genes that were significantly associated with at least one lipid trait in GLGC.

Second, we annotated the prioritized interacting genes from our 4C-seq experiment with phenotypic information of mouse models within the Mouse Genome Informatics database (MGI, www.informatics.jax.org). We found murine phenotypic information on 144 mouse homologues (**Supplemental table 5**). Knockout of 67 of them resulted in a

phenotype related to cardiovascular disease (significant enrichment, 67/144: binomial $P = 1.36 \times 10^{-47}$), such as abnormal blood vessel morphology (*Col4a1*, *Cxcl12*, *Epor*, *Shc1*, *Tcf7l1*), altered circulating fatty acid levels (*Csf2*, *Kdm3a*, *Ldlr*, *Lipa*, *Pten*) and impaired vascular contractility (*Acta2*). Human variants in *ACTA2* are associated with early onset stroke and MI.⁴⁶ Knockout of two candidate gene murine homologues affected development of atherosclerotic lesions, namely *Ldlr* (accelerated development of atherosclerosis) and *Shc1* (resistance to diet-induced atherosclerosis). The p66 isoform of human SHC1 is implicated in reactive oxygen species generation and its knockdown in endothelial cells of obese mice attenuated production of these radicals and of fatty acids oxidation.⁴⁷ Third, we investigated the biological effect of human knockout variation of the candidate genes to study druggability. We queried two datasets of available information on SNPs and insertion/deletion variants that cause human functional knockouts.^{29,30} We found human knockouts, caused by nonsense, splice or frameshift variants, for 89 candidate genes (**Supplemental table 6**).

Fourth, using a custom-built drug discovery pipeline, we found available compounds to target 50 of the candidate genes (**Supplemental table 7A**). These drugs showed a relative overrepresentation for usage as immunomodulating agents ($P = 0.012$ in coronary endothelial cells, $P < 0.001$ in monocytes) (**Supplemental table 7B**).

Together, these findings provide further evidence that by using the 4C-seq method we identified additional candidate target genes for human atherosclerotic disease.

Discussion

Based on 3D chromatin-chromatin interactions with DNA regulatory elements that co-localize with previously identified susceptibility loci, we present 294 additional candidate genes for CAD and LAS that are of potential interest in the pathophysiology of human atherosclerotic disease. This study is the first to systematically study the human chromatin interactions of the CARDIoGRAMplusC4D and METASTROKE loci. Many of the additional genes have not been implicated in atherosclerosis before. Our approach, from a DNA regulatory point of view, complements conventional methods for candidate gene identification of GWAS susceptibility, can help further unravel diseases with a complex genetic background, and pave the way for cell-type specific drug development.

We have highlighted the 4C candidate genes that we could annotate via additional analyses and that therefore have known or foreseeable effects on cardiovascular disease. Based on tissue-specific pathway analyses, we highlighted *PTEN* that is known to be upregulated by statins and to possess effects on atherosclerosis.³⁷⁻³⁹ Furthermore, based on eQTL studies, we identified *SRR*, the effect of which was previously implicated in stroke⁴⁰, and *USMG5*, that was previously associated with white matter hyper-intensities in the brain.⁴⁸ Of special interest is the finding of an alternative mechanism by which the susceptibility locus that contains rs7528419 (*SORT1* region) may exert its effect. Using

4C-sequencing we identified a physical interaction between an active regulatory element that overlaps rs7528419 and *PSRC1* in monocytes. Moreover, we found an association between rs7528419 and the expression of *PSRC1* in monocytes and an association between *PSRC1* expression and atherosclerosis severity. This association was independent of LDL levels, which is the putative mechanism of rs12740374, a SNP in perfect LD (1.0) with rs7528419 that was previously found to increase *SORT1* expression in liver.

Mapping the SNPs that identify susceptibility loci in GWAS to genes that affect a complex disease, such as cardiovascular disease, is a challenging task. By annotating the locus with the linearly closest gene, the 3D conformation of chromatin is inadvertently not taken into account. Many of the additional candidate genes we report are located outside the GWAS susceptibility loci. Using 5C (chromosome conformation capture carbon copy) the importance of studying 3D interactions has been highlighted previously; in a sample of 628 TSS from the ENCODE project only 7% of the over 1000 long-range looping interactions were with the nearest gene.¹⁰ In a previous effort to identify candidate genes based on DNA regulatory mechanisms, 33 enhancers in the 9p21 locus were scrutinized.⁴⁹ Interestingly, the chromatin interaction between the enhancers identified by 3C (chromatin conformation capture) was found to be remodeled upon treatment with interferon- γ in HUVECs. In our 4C-seq experiment, we confirmed the physical chromatin-chromatin interaction between the 9p21 susceptibility locus and several candidate genes, among which interferons, in human coronary endothelial cells and monocytes. However, we found that these genes were not actively expressed in these cell types and therefore did not consider them any further.

There are some limitations to this study. First, there is no consensus about the gold standard approach to analysis of 4C-seq data. For example, we used a conservative cut-off for calling a chromatin-chromatin interaction ($P < 10^{-8}$). Altering this cut-off may result in more candidate genes. However, this likely leads to more false-positive results. We therefore report a quantitative measure for the p-value of the interaction of the DRE with the proposed candidate gene to enable the reader to take these considerations into account when interpreting the data. Second, while 4C-sequencing enables us to look at physical interactions, these interactions do not necessarily mean that the expression in the studied tissue is in fact regulated by the association locus or even expressed. We therefore decided to only report only genes that are actively expressed in the studied tissues. Furthermore, we found no eQTL association between the SNP of interest and any of the genes *in the vicinity* of the genes that were identified by 4C-sequencing, indicating that the resolution of the technique is sufficient to distinguish between candidate genes and less relevant genes within a genomic region. A more accurate cell type-specific mapping of susceptibility loci to candidate genes in humans is of paramount importance for the development of specific compounds in the pharmaceutical industry. The genes we identified display only partial overlap between coronary endothelial cells and monocytes. This finding stresses the importance of cell-specific approaches in order to grasp the

complex biology of atherosclerotic disease. It also highlights the possibility to develop cell-specific compounds to target atherosclerotic disease. Our results therefore underline the need to investigate cell type-specific 3D chromatin conformation in future functional follow-up of GWAS data.

References

1. Hansson GK, Libby P, Tabas I. Inflammation and plaque vulnerability. *Journal of Internal Medicine*. 2015. p. 483–93.
2. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* [Internet]. 2012 Dec 2;45(1):25–33. Available from: <http://www.nature.com/doi/10.1038/ng.2480>
3. Traylor M, Farrall M, Holliday EG, Sudlow C, Hopewell JC, Cheng YC, et al. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE Collaboration): A meta-analysis of genome-wide association studies. *Lancet Neurol*. 2012;11(11):951–62.
4. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* [Internet]. 2015;6:5890. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4420238&tool=pmcentrez&rendertype=abstract>
5. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*. 2011;7(1).
6. Brnne I, Civelek M, Vilne B, Di Narzo A, Johnson AD, Zhao Y, et al. Prediction of causal candidate genes in coronary artery disease loci. *Arterioscler Thromb Vasc Biol*. 2015;35(10):2207–17.
7. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* (80-) [Internet]. 2012;337(6099):1190–5. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3771521&tool=pmcentrez&rendertype=abstract>
8. Mokry M, Middendorp S, Wiegerinck CL, Witte M, Teunissen H, Meddens CA, et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology*. 2014;146(4):1040–7.
9. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* [Internet]. 2012;45(2):124–30. Available from: <http://dx.doi.org/10.1038/ng.2504> <http://www.nature.com/doi/10.1038/ng.2504>
10. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* [Internet]. 2012 Sep 5;489(7414):109–13. Available from: <http://www.nature.com/doi/10.1038/nature11279>
11. De Laat W, Klous P, Kooren J, Noordermeer D, Palstra RJ, Simonis M, et al. Three-Dimensional Organization of Gene Expression in Erythroid Cells. *Current Topics in Developmental Biology*. 2008. p. 117–39.
12. Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* [Internet]. 2014 Jan 12;46(2):205–12. Available from: <http://www.nature.com/doi/10.1038/ng.2871>
13. Van De Werken HJG, De Vree PJP, Splinter E, Holwerda SJB, Klous P, De Wit E, et al. 4C technology: Protocols and data analysis. *Methods Enzymol*. 2012;513:89–112.
14. Hansson GK, Libby P. The immune response in atherosclerosis: a double-edged sword. *Nat Rev Immunol*. 2006;6(7):508–19.
15. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* [Internet]. 2012;489(7414):57–74. Available from: <http://www.nature.com/doi/10.1038/nature11247> <http://www.nature.com/doi/10.1038/nature11247>
16. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
17. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166–9.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139–40.
19. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* [Internet]. 2010 Sep 15;26(18):2336–7. Available from: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq419>
20. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for

- genome-wide association studies. *Am J Hum Genet.* 2010;87(1):139–45.
21. Hägg S, Skogsberg J, Lundström J, Noori P, Nilsson R, Zhong H, et al. Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: The Stockholm Atherosclerosis Gene Expression (STAGE) study. *PLoS Genet.* 2009;5(12).
 22. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Schork NJ, editor. *PLoS Genet* [Internet]. 2009 Jun 19;5(6):e1000529. Available from: <http://dx.plos.org/10.1371/journal.pgen.1000529>
 23. Qi J, Asl H, Björkegren J, Michoel T. kruX: matrix-based non-parametric eQTL discovery. *BMC Bioinformatics* [Internet]. 2014;15(1):11. Available from: <http://www.biomedcentral.com/1471-2105/15/11>
 24. Du P, Kibbe WA, Lin SM. lumi: A pipeline for processing Illumina microarray. *Bioinformatics.* 2008;24(13):1547–8.
 25. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* [Internet]. 2012 Oct 31;491(7422):56–65. Available from: <http://www.nature.com/doi/10.1038/nature11632>
 26. Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, Elbers CC, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* [Internet]. 2014 Jun 29;46(8):818–25. Available from: <http://www.nature.com/doi/10.1038/ng.3021>
 27. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* [Internet]. 2012 Dec 27;10(1):5–6. Available from: <http://www.nature.com/doi/10.1038/nmeth.2307>
 28. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* [Internet]. 2015;1–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26708335>
 29. Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, et al. Identification of a large set of rare complete human knockouts. *Nat Genet* [Internet]. 2015 Mar 25;47(5):448–52. Available from: <http://www.nature.com/doi/10.1038/ng.3243>
 30. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* (80-) [Internet]. 2012 Feb 17;335(6070):823–8. Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1215040>
 31. Griffith M, Griffith OL, Coffman AC, Weible J V, McMichael JF, Spies NC, et al. DGIdb: mining the druggable genome. *Nat Methods* [Internet]. 2013;10(12):1209–10. Available from: <http://dx.doi.org/10.1038/nmeth.2689>
 32. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: A comprehensive resource for “Omics” research on drugs. *Nucleic Acids Res.* 2011;39(SUPPL. 1).
 33. Altman RB. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet* [Internet]. 2007;39:426. Available from: <http://dx.doi.org/10.1038/ng0407-426> <http://www.nature.com/ng/journal/v39/n4/pdf/ng0407-426.pdf>
 34. Kuhn M, Szklarczyk D, Franceschini A, Von Mering C, Jensen LJ, Bork P. STITCH 3: Zooming in on protein-chemical interactions. *Nucleic Acids Res.* 2012;40(D1).
 35. Van Bever E, Wirtz VJ, Azermai M, De Loof G, Christiaens T, Nicolas L, et al. Operational rules for the implementation of INN prescribing. *Int J Med Inform.* 2014;83(1):47–56.
 36. Pahor M, Chrischilles EA, Guralnik JM, Brown SL, Wallace RB, Carbonin P. Drug data coding and analysis in epidemiologic studies. *Eur J Epidemiol.* 1994;10(4):405–11.
 37. Teresi RE, Planchon SM, Waite KA, Eng C. Regulation of the PTEN promoter by statins and SREBP. *Hum Mol Genet.* 2008;17(7):919–28.
 38. Muthalagan E, Ganesh RN, Sai Chandran B V, Verma SK. Phosphatase and tensin analog expression in arterial atherosclerotic lesions. *Indian J Pathol Microbiol.* 2014;57(3):427–30.
 39. Dai XY, Cai Y, Mao DD, Qi YF, Tang C, Xu Q, et al. Increased stability of phosphatase and tensin homolog by intermedin leading to scavenger receptor A inhibition of macrophages reduces atherosclerosis in apolipoprotein E-deficient mice. *J Mol Cell Cardiol.* 2012;53(4):509–20.
 40. De Miranda J, Santoro A, Engelender S, Wolosker H. Human serine racemase: Molecular cloning, genomic organization and functional analysis. *Gene.* 2000;256(1-2):183–8.
 41. Folkersen L, Van't Hooft F, Chernogubova E, Agardh HE, Hansson GK, Hedin U, et al. Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. *Circ Cardiovasc Genet.* 2010;3(4):365–73.
 42. Hoefler IE, Sels JW, Jukema JW, Bergheanu S, Biessen E, McClellan E, et al. Circulating cells as predictors of secondary manifestations of cardiovascular disease: Design of the CIRCULATING CELLS study. *Clin Res Cardiol.* 2013;102(11):847–56.
 43. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory

- effects across multiple tissues in twins. *Nat Genet* [Internet]. 2012;44(10):1084–9. Available from: <http://dx.doi.org/10.1038/ng.2394> \n<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3784328&tool=pmcentrez&rendertype=abstract>
44. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, et al. Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility. *Bochdanovits Z*, editor. *PLoS One* [Internet]. 2010 May 18;5(5):e10693. Available from: <http://dx.plos.org/10.1371/journal.pone.0010693>
 45. Arvind P, Nair J, Jambunathan S, Kakkar V V., Shanker J. CELSR2-PSRC1-SORT1 gene expression and association with coronary artery disease and plasma lipid levels in an Asian Indian cohort. *J Cardiol*. 2014;64(5):339–46.
 46. Guo DC, Papke CL, Tran-Fadulu V, Regalado ES, Avidan N, Johnson RJ, et al. Mutations in Smooth Muscle Alpha-Actin (ACTA2) Cause Coronary Artery Disease, Stroke, and Moyamoya Disease, Along with Thoracic Aortic Disease. *Am J Hum Genet*. 2009;84(5):617–27.
 47. Paneni F, Costantino S, Cosentino F. P66Shc-induced redox changes drive endothelial insulin resistance. *Atherosclerosis*. 2014;236(2):426–9.
 48. Lopez LM, Hill WD, Harris SE, Valdes Hernandez M, Munoz Maniega S, Bastin ME, et al. Genes From a Translational Analysis Support a Multifactorial Nature of White Matter Hyperintensities. *Stroke* [Internet]. 2015 Feb;46(2):341–7. Available from: <http://stroke.ahajournals.org/lookup/doi/10.1161/STROKEAHA.114.007649>
 49. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, et al. 9p21 DNA variants associated with coronary artery disease impair interferon- γ signalling response. *Nature* [Internet]. 2011 Feb 10;470(7333):264–8. Available from: <http://www.nature.com/doi/10.1038/nature09753>



Chromatin conformation links distal
target genes to chronic kidney disease
loci

6

Claartje A. Meddens, Maarten M. Brandt, Laura Louzao-Martinez, Noortje A. M. van den Dungen, Nico R. Lansu, Edward E. S. Nieuwenhuis, Dirk J. Duncker, Marianne C. Verhaar, Jaap A. Joles, Michal Mokry, and Caroline Cheng

Based on: *Journal of the American Society of Nephrology: JASN*
29(2):462–76 2018

Introduction

Chronic kidney disease (CKD) is a condition marked by loss of kidney function, which can lead to end stage renal disease and is associated with a dramatic increase in cardiovascular disease-related morbidity and mortality¹. Based on the latest report of the Center for Disease Control and Prevention (2007-2014), over 15% of the U.S. population is affected by CKD and the numbers are expected to rise. CKD incurs substantial rising medical costs in the U.S., with similar developments observed globally. Over the last decade, the findings of multiple genome wide association studies (GWASs) have established common DNA variants as genetic risk factors for CKD^{2,3}. However, functional annotation and explanation of these loci remains an issue. Currently, the functional annotation of GWAS data is mainly conducted by linking susceptibility loci by spatial proximity to the nearest gene⁴. For example, well-known single nucleotide polymorphisms (SNPs) that are associated with CKD include SNPs annotated with *ALMS1* and *UMOD*. *ALMS1* is required for medullar collecting duct ciliogenesis⁵, whereas *UMOD* is involved in the inhibition of calcium oxalate crystallization in renal fluids⁶ and has an evolutionary role in protection from urinary tract infections⁷. Since these SNPs are located in coding regions of genes with important renal protective functions, it is conceivable that the genetic variation marked by these SNPs affects both genes, contributing to CKD pathogenesis. For many of the CKD associated susceptibility loci that are not directly located in or near protein coding regions, the causal contribution to disease etiology is far less straightforward.

New insights brought by epigenetic research have revealed the prevalence of DNA regulatory elements (DREs), such as enhancers and repressors, located in both coding-

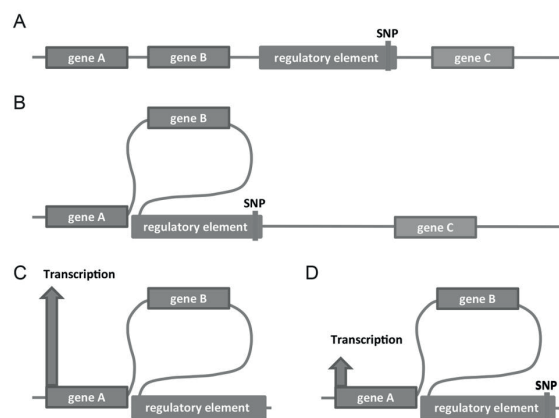


Figure 1: Genetic variation in DREs could be a causative factor in dysregulation of distal target gene expression.

(A) Many of the susceptibility loci that are not located in protein coding regions, overlap with DREs such as enhancers and repressors. (B) DREs play a crucial role in regulating gene expression, in a cell specific manner, by modulating 3D chromatin interactions, and increasing spatial proximity of DREs with transcriptional start sites, thereby regulating transcription of genes on a non-linear DNA scale. (C/D) Distal transcriptional activity of DREs could be compromised by genetic variation (represented by colocalization with disease-associated SNP).

and non-protein-coding DNA regions (**Figure 1A**)⁸. These DREs play a crucial role in regulating gene expression in a cell specific manner. Enhancer elements regulate transcription of their target genes through 3D chromatin interactions with transcriptional start sites (**Figure 1B**). Importantly, DREs can regulate expression levels of gene targets over a distance up to thousands of kilobase pairs (kbp)⁹, far exceeding the current standard distance for GWAS annotation. Common genetic variation in DREs could be a causative factor in dysregulation of target gene expression, leading to disease or other phenotypes (**Figure 1C,D**). This was demonstrated previously for the SNP rs12913832, which was shown to modulate human pigmentation by affecting the enhancer regulation of the *OCA2* promoter¹⁰. Systematic mapping of the target genes of DREs that overlap with known CKD associated SNPs, could greatly improve our understanding of the complex genetics of CKD.

Here we used self-transcribing active regulatory region sequencing (STARR-seq) to evaluate the potential effect of CKD associated genetic variation on transcriptional regulation. In a proof of principle approach we cloned putative DREs located on the same linkage disequilibrium (LD) block as the CKD associated SNP rs11959928 from 20 donors in STARR-seq reporter plasmids. This approach enabled us to study the effect of all variants found on this susceptibility region in the donor pool on enhancer activity in primary human renal proximal tubular epithelial cells (HRPTECs), human renal glomerular endothelial cells (HRGECs), and the human embryonic kidney cell line HEK293a. The findings of this experiment illustrated how regulatory function could be affected by common small variants, thereby highlighting the relevance of studying downstream target genes of DREs overlapping with disease-associated susceptibility regions to add an additional layer to post GWAS analysis. Subsequently, we used circular chromosome conformation capture-sequencing (4C-seq) to identify putative candidate genes for CKD by examining 3D interactions between DREs that colocalize with CKD susceptibility loci and their target genes. This allowed us to study long range regulation of target gene promoters by cross-linking the folded and interacting DRE segments, followed by two restriction-ligation steps of the DNA strands and DNA sequencing. As transcriptional regulation is cell-type specific, and CKD pathogenesis is associated with reduced glomerular filtration rate as a result of tubulo-interstitial fibrosis¹¹ as well as loss of peritubular and glomerular capillaries^{12, 13}, we conducted the 4C-seq in HRPTECs and HRGECs. Chromatin interactions were studied in these primary cells from healthy donors to create an overview of genes interacting with CKD susceptibility loci. We conducted a systematic screen of 39 putative regulatory elements that colocalize with previously reported susceptibility regions for CKD. This led to the identification of 304 target genes that are potentially transcriptionally affected by these CKD-associated SNPs. This study shows for the first time a direct interaction between CKD-associated common variant regions and the promoter regions of CKD-associated target genes. Although additional functional studies are needed to determine the exact mechanism of action, in its current

form our data presents an extensive overview of potential target genes for the previously reported CKD-associated SNPs, providing new gene candidates for hypothesis-driven future studies.

Results

STARR-seq directly demonstrates the potential of genetic variation to affect regulation of gene expression

To illustrate the effect of genetic variation on regulatory activity of DREs as an additional layer to GWAS interpretation, the STARR-seq reporter set-up was used to test the influence of common genetic variants colocalizing with possible DREs positioned on the haploblock marked by the CKD-associated SNP rs11959928. The STARR-seq reporter assay is based on a reporter plasmid containing a minimal promoter, followed by an incorporated candidate enhancer sequence¹⁴. The activity of each enhancer is reflected by its ability to induce the promoter activity leading to RNA transcription of the enhancers sequence (**Figure 2A**). The advantage of this approach over luciferase reporter assays is that STARR-seq allows parallel (and thus “high throughput”) assessment of all genomic variation in the enhancer regions located on this specific haploblock, as the effect of a variant on enhancer strength is reflected by its relative prevalence in transcribed RNA compared to its prevalence in the pool of reporter plasmids. Putative DREs located on the haploblock marked by the CKD-associated SNP rs11959928 (**Figure 2B**), containing at least three potential regulatory regions (I, II, III) as illustrated by H₃K₄Me₁, H₃K₂₇Ac, and DNase clusters in human umbilical vein endothelial cells and human epidermal keratinocytes (adapted from USCS genome browser), were cloned from 20 individual donors into one combined reporter library (**supplemental table 1**). This library was transformed in HRGECs, HRPTECs and HEK293a (the latter cell type was used as an additional control), followed by sequencing of the produced enhancer derived RNA, as well as the library itself, enabling us to compare transcription frequency of each variable allele located in the enhancer sequences on the haploblock with its original frequency in the library (**supplemental table 2**). Via this approach, one particular region containing 5 variants was found to be strongly affected by allele specific activity in all three examined cell types (**Figure 2C**). Four of these variants had a reference allele frequency of 45.2-74.8% in the library input, yet virtually only the reference alleles were transcribed in all three cell types. The other allele had a wild-type penetrance of 31.8-36.5% in the library input, but its frequency was strongly reduced in the transcribed RNA. This example illustrates that disease-associated SNPs may not only affect gene coding sequences, but might also affect the transcriptional regulation of DRE target genes.

4C-seq leads to discovery of new target genes for CKD-associated SNPs

Building on the illustrative STARR-seq findings, 39 CKD associated susceptibility loci that colocalize with DREs were studied in HRGECs and HRPTECs in order to identify the target genes of putative DREs^{2,3}. Activity of these DREs was assessed in renal epithelium

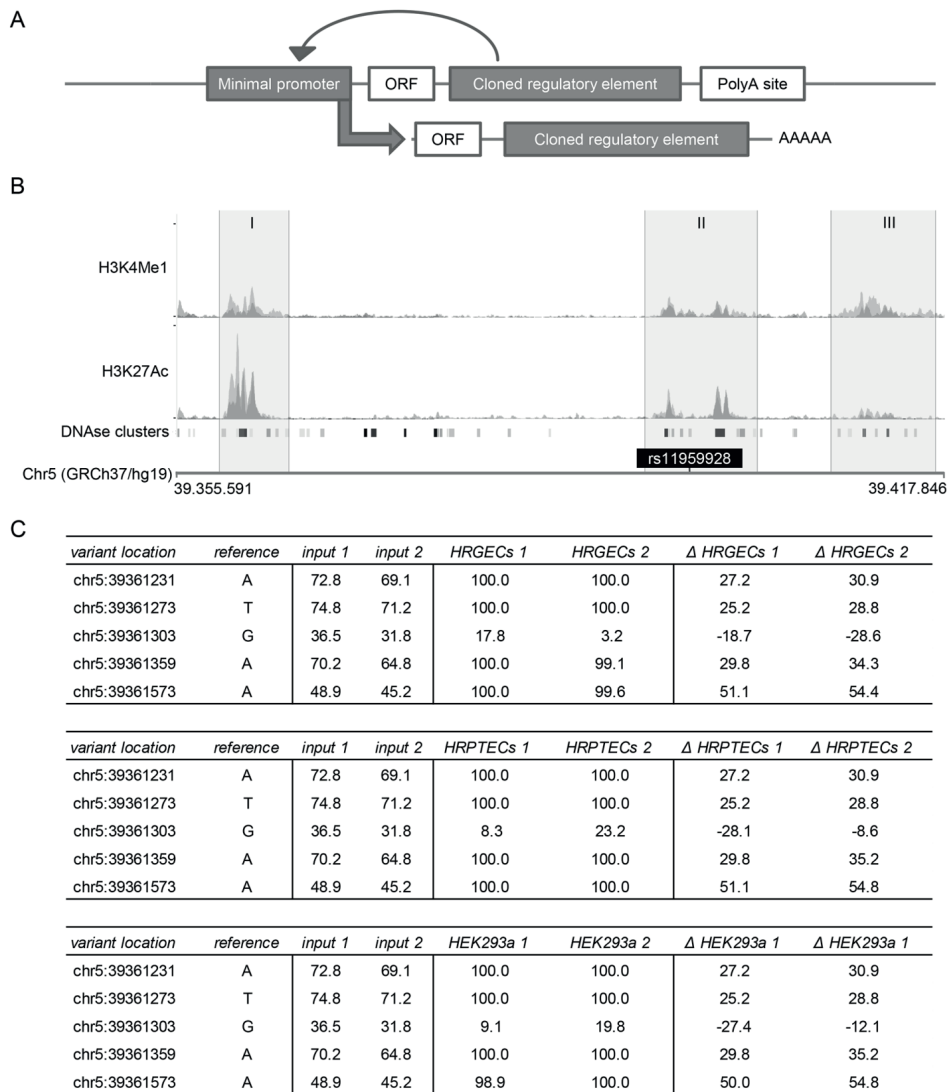


Figure 2: STARR-seq analysis illustrates the effect of CKD-associated genetic variation on transcriptional regulation.

(A) The STARR-seq reporter principle is based on a reporter plasmid containing a minimal promoter, followed by a cloned candidate enhancer sequences. The activity of each enhancer is reflected by its ability to transcribe themselves. (B) Putative DREs, identified by H₃K₄Me₁, H₃K₂₇Ac, and DNase clusters (Human Umbilical Vein Endothelial Cells in blue, Human Epidermal Keratinocytes in pink, overlap in purple - adapted from USCS genome browser), located on the haploblock marked by CKD-associated SNP rs11959928 (I, II, III) were cloned into the STARR-seq plasmid from 20 individual donors. (C) The library of STARR-seq plasmids was transformed in HRGECs, HRPTECs and HEK293a, followed by RNA-seq of the produced enhancer RNA strands. Shown in replicate is the percentage of the reference allele in the input library, the percentage of the reference allele in cellular transcribed RNA, and the delta between the prevalence in the library versus transcribed RNA (found in region I).

and fetal renal tissue for HRPTECs and microvascular endothelium for HRGECs based on DNase hypersensitivity and H₃K₄me₃ chromatin immunoprecipitation data (**supplemental table 3**). Of the 39 studied loci, 6 colocalize only with active DREs in renal epithelium, 5 colocalize only with active DREs in microvascular endothelium, and 28 loci colocalize with active DREs in both renal epithelium and microvascular endothelium. For the discovery of target genes of these regulatory elements, the TSSs that interacted with these loci were examined in HRPTECs and HRGECs using 4C-seq (**Figure 3A-F**). 67 chromatin interaction datasets were generated in twofold, of which only the replicated chromatin interactions were considered as candidate genes. These candidate genes were filtered per cell type for expression in that specific cell type, using in-house and public expression datasets (**Figure 3G**). This led to the discovery of 304 CKD target genes, of which 199 were found in HRGECs (**Figure 4, supplemental table**

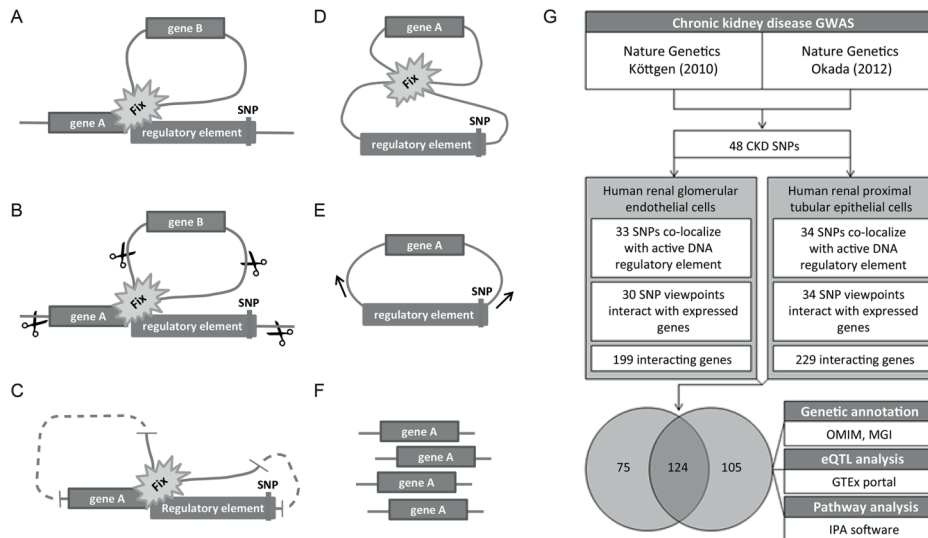


Figure 3: 4C-seq was used to study chromatin interactions leading to the discovery of 304 CKD target genes in total.

(A/B) The 3D chromatin conformation of the DREs was studied in detail based on 4C template, which was generated by fixing the chromatin structure, followed by enzymatic (DPNI) digestion of the fixed chromatin. (C) The digested chromatin was ligated into circular fragments in a diluted environment, (D) after which the chromatin was de-crosslinked. (E) The circular DNA molecules followed another round of enzymatic (CviQI) digestion and ligation, after which the 4C-seq library was prepared with primers that target sequences in close proximity of the CKD susceptibility loci. (F) This library was sequenced to identify genes that were physically interacting with CKD susceptibility loci. (G) The 4C analysis was initially performed on 48 viewpoints based on CKD associated SNPs, of which in total 39 colocalized with active DNA regulatory elements based on mapping with DNase hypersensitivity or H₃K₄me₃ ChIP-seq datasets (33 in HRGECs and 34 in HRPTECs with partial overlap of SNPs). Of these 39 studied viewpoints, only 36 (31 in HRGECs and 34 in HRPTECs with partial overlap of viewpoints) were interacting with a total of 304 target genes with validated expression in the assessed cell types (overlap indicated in Venn graph). These 304 genes were subsequently processed for genetic annotation to renal failure associated traits in the online mendelian inheritance in man (OMIM) database and the mouse genome informatics (MGI) database, in addition to eQTL analysis in the GTEx portal database, and pathway analysis using Ingenuity Pathway Analysis (IPA) software.

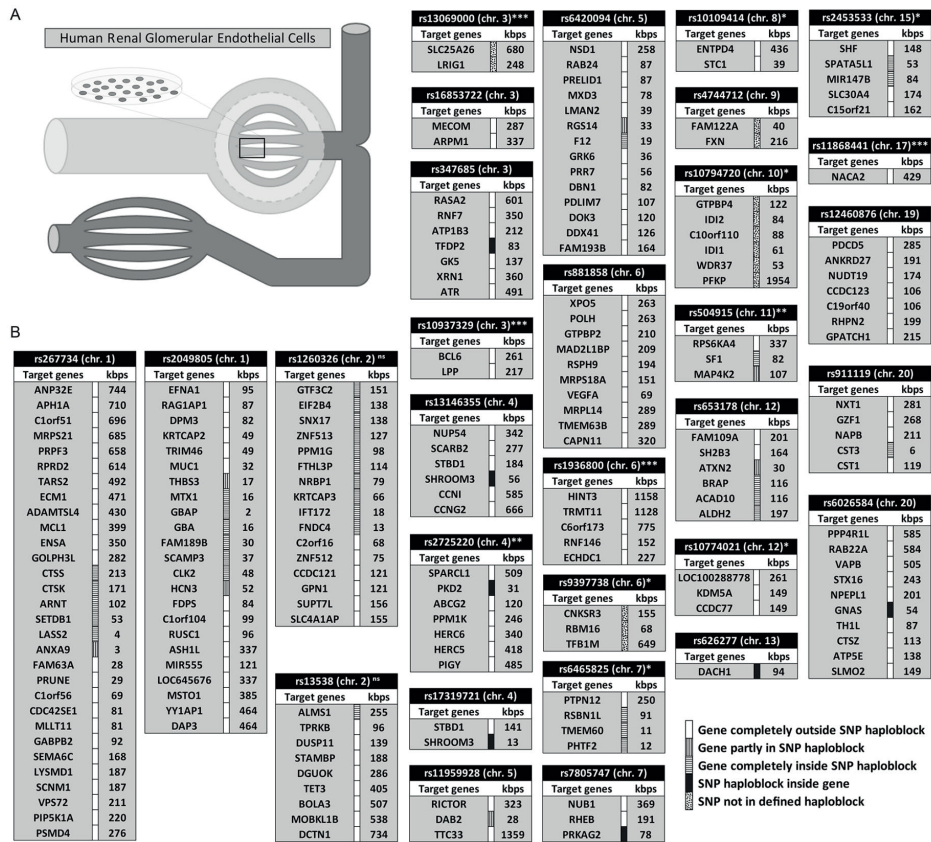


Figure 4: Analysis of chromatin interactions of CKD susceptibility loci that colocalize with regulatory elements using 4C-seq led to the discovery of 199 CKD target genes in glomerular endothelial cells.

(A) Chromatin interactions were studied in cultured HRGECs to define endothelial target genes of CKD susceptibility loci that colocalize with active regulatory elements. (B) Of a total of 33 replicated 4C datasets, based on CKD susceptibility loci that colocalize with active regulatory elements, 30 interacted with at least one target gene that was expressed in endothelium, which led to the identification of 199 CKD target genes in total. Studied SNPs are displayed ordered on position, followed by haplblock information (green; gene completely outside SNP haplblock, blue; gene partly inside SNP haplblock, purple; gene completely inside SNP haplblock, red; SNP haplblock inside gene, white; SNP not in defined haplblock) and the SNP-TSS distance in kbps. *SNP solely associated with serum creatinine (estimated glomerular filtration rate), **SNP solely associated with serum urate, ***SNP solely associated with blood urea nitrogen, nsnon-synonymous SNP.

4) and 229 in HRPTECs (Figure 5, supplemental table 5). Amongst the 199 target genes interacting in HRGECs and 229 target genes interacting in HRPTECs, 124 were identified in both cell types (Figure 3G). These 304 candidate genes all fulfilled the following three criteria: (1) The TSS of the candidate gene colocalizes with a significant 4C-seq signal ($P < 10^{-8}$) within 5kbp. (2) The SNP or any other SNP in LD ($r > 0.8$) colocalizes with active regulatory regions. (3) The candidate gene is expressed in the cell types of interest (reads per kilobase million reads sequenced (RPKM) > 1 and probe intensity > 6 for microvascular endothelium RNA-seq data and HRPTECs microarray data, respectively)¹⁵.

Table 1: Overlapping genes in 4C-seq retrieved genes and CKD associated genes in OMIM.

Gene ID	Phenotype	MIM Number	4C SNP	SNP-TSS (kbp)	HRGECs	HRPTECs
MUC1	Medullary cystic kidney disease 1	158340	rs2049805	32	x	x
PKD2	Polycystic kidney disease 2	173910	rs2725220	31	x	
SCARB2	Epilepsy, progressive myoclonic 4, with or without renal failure	602257	rs13146355	277	x	x
SLC2A9	Hypouricemia	612076	rs3775948	28		x
SLC34A1	Nephrolithiasis/osteoporosis, hypophosphatemic	182309	rs6420094	6		x








 Gene completely outside SNP haploblock
 Gene completely inside SNP haploblock
 SNP haploblock inside gene

Table 2: Overlapping genes in 4C-seq retrieved genes and CKD associated genes in MGI.

Gene ID	Phenotype	4C SNP	SNP-TSS (kbp)	HRGECs	HRPTECs
ALMS1	abnormal kidney morphology, urine-, renal tubules-, and other kidney related abnormalities	rs13538	255	x	x
ASH1L	glomerulus- and other kidney related abnormalities	rs2049805	337	x	x
CCNI	urine-, blood-, glomerulus-, podocyte-, and other kidney related abnormalities	rs13146355	585	x	x
CTSS	abnormal kidney angiogenesis	rs267734	213	x	x
DCTN1	blood abnormalities	rs13538	734	x	
GNAS	urine- and other kidney related abnormalities	rs6026584	54	x	x
GRK6	glomerulus- and other kidney related abnormalities	rs6420094	36	x	x
IFT172	abnormal kidney morphology, glomerulus abnormalities	rs1260326	18	x	
MAF	abnormal kidney morphology, renal tubules abnormalities	rs889472	11		x
MECOM	blood abnormalities	rs16853722	287	x	x
MPV17	abnormal kidney morphology and angiogenesis, urine-, blood-, glomerulus-, renal tubules-, and other kidney related abnormalities	rs1260326	185		x
PKD2	abnormal kidney morphology, kidney cysts, blood-, renal tubules-, and other kidney related abnormalities	rs2725220	31	x	
RNF7	abnormal kidney angiogenesis	rs347685	350	x	x
SCARB2	urine abnormalities	rs13146355	277	x	x
SHC1	abnormal kidney angiogenesis	rs2049805	252		x
SLC14A1	urine- and blood abnormalities, abnormal renal filtration	rs7227483	117		x
SLC14A2	urine abnormalities, abnormal renal filtration	rs7227483	8		x
SLC2A9	kidney cysts, urine-, blood-, renal tubules-, and other kidney related abnormalities	rs3775948	28		x
SLC4A5	urine abnormalities, abnormal renal filtration	rs13538	702		x
SLC7A9	renal tubules abnormalities	rs12460876	4		x
SOD2	blood abnormalities	rs2279463	554		x
THBD	abnormal kidney morphology	rs911119	582		x
VEGFA	abnormal kidney morphology and angiogenesis, blood-, glomerulus-, renal tubules-, and podocyte abnormalities	rs881858	69	x	x

 Gene completely outside SNP haploblock
 Gene partly in SNP haploblock
 Gene completely inside SNP haploblock
 SNP haploblock inside gene

6

Genetic annotation of candidates picked up by 4C-seq in OMIM and MGI demonstrates link with CKD

We evaluated if the identified candidates were associated with CKD, using the Online Mendelian Inheritance in Man (OMIM) and the Mouse Genome Informatics (MGI) database. The OMIM database is a catalogue of human genetic disorders that connects rare gene variants with phenotype. We established the overlap of our HRPTECs and HRGECs gene lists with the genes retrieved from the OMIM Morbid Map by searching for the keywords “kidney”, “renal” and “nepbro”. Monogenetic defects in 5 CKD candidate

genes were directly correlated with a renal disease phenotype, of which 2 are completely located on a different haploblock than the 4C viewpoint (**Table 1**, green mark). The MGI database contains murine phenotypic information of mutant alleles. Analysis revealed that monogenetic silencing of 23 of the CKD candidate genes in mice, caused direct renal failure related traits, including albuminuria (*ALMS1*, *MPV17*, *SCARB2*), abnormal renal filtration rate (*SLC14A1*), and glomerular sclerosis (*CCNI*, *MPV17*, *VEGFA*) (**Table 2**). Of the 23 renal failure traits associated target genes found with MGI, 13 are located entirely on a different haploblock than the 4C viewpoint (green mark).

eQTL analysis reveals genotype dependent expression of CKD candidate genes

A candidate gene with an expression level that is significantly correlated with co-occurrence of a SNP is likely to be transcriptionally regulated by a DRE affected by the SNP. These loci that contribute to variation in gene expression levels, called expression

Table 3: Overlapping genes in 4C-seq retrieved genes and eQTL genes derived from GTEx portal.

eQTL	Gene ID	SNP-TSS (kbp)	eQTL in (tissue):	HRGECs	HRPTECs
rs10794720	ID12	84	Muscle	x	
	NUDT19	174	Artery, Nerve	x	
rs12460876	SLC7A9	4	Adipose, Skin, Thyroid, Nerve, Lung		x
	TDRD12	146	Testis		x
	SNX17	138	Muscle	x	x
rs1260326	FNDCC4	13	Thyroid	x	x
	NRBP1	79	Adipose, Testis	x	x
	KRTCAP3	66	Adrenal gland	x	
	ALMS1	255	Pancreas	x	x
rs13538	TPRKB	96	Artery	x	x
	NAT8	1	Adipose, Thyroid, Skin, Esophagus, Pancreas, Artery, Brain		x
	THBS3	17	Whole Blood, Thyroid, Esophagus, Colon, Lung, Stomach, Testis, Spleen	x	x
	GBA	16	Esophagus, Nerve, Thyroid	x	x
rs2049805	MUC1	32	Esophagus	x	x
	FAM189B	30	Thyroid	x	x
	EFNA1	95	Skin	x	x
	HCN3	52	Nerve	x	
	SPATA5L1	53	Nerve, Adipose, Esophagus, Thyroid, Whole Blood, Artery, Muscle, Heart	x	x
rs2453533	SLC30A4	174	Adipose	x	
	GATM	30	Thyroid, Muscle, Esophagus, Lung, Skin, Whole blood		x
	SLC28A2	97	Nerve, Thyroid, Adrenal gland, Colon, Ovary		x
	ANXA9	3	Skin, Testis	x	x
rs267734	CTSS	213	Adipose, Skin, Artery, Muscle, Thyroid, Nerve, Esophagus	x	x
	CTSK	171	Whole blood	x	x
	ARNT	102	Whole blood	x	
rs2725220	PKD2	31	Testis, Esophagus	x	
rs4744712	FAM122A	40	Artery	x	
rs504915	MEN1	114	Testis		x
	NRXN2	53	Lung, Artery, Esophagus, Nerve, Adipose, Muscle, Skin		x
	RGS14	33	Skin, Artery, Testis	x	x
rs6420094	FGFR4	304	Nerve		x
	SLC34A1	6	Esophagus		x
rs6465825	RSBN1L	91	Brain	x	
	TMEM60	11	Artery, Testis	x	
rs653178	ALDH2	197	Skin	x	x
rs911119	CST3	6	Testis, Lung, Whole blood, Nerve	x	x
rs9895661	TBX2	238	Artery		x

Proefschrift_Layout_CM_Final.indd 104

2019-08-26 19:37:49

quantitative trait loci (eQTL), are identified using GWAS and RNA-seq data of the target organ. To date, no large genome wide eQTL data of the human kidney has been published that allows adequate analysis of all CKD associated SNPs¹⁶. To evaluate if the expression levels of the CKD candidate genes are affected by CKD-associated SNPs, we used the Genotype-Tissue Expression (GTEx) database. Although GTEx so far only contains kidney specific expression data of 26 genotyped donors, which does not reach the GTEx threshold for eQTL analysis (>70), the database does include genotype- and expression-matched data of 449 donors for which eQTL analysis was conducted in 44 non-renal tissues, which is large enough to stratify most of the individual CKD SNPs and wild-type alleles assessed in our study. Of the 39 CKD associated susceptibility loci, 25 were annotated in GTEx. These 25 SNPs were significantly correlated with the expression of 54 genes, of which 48 physically interacted with the 4C-seq input loci (data not shown). Of these 48 captured target genes, 38 were actually expressed in HRPTECs and/or HRGECs and included 10 genes located on a completely different haploblock than the CKD-associated SNP (**Table 3**, green mark). Thus, although the lack of genotype-kidney expression datasets prohibited us to study these eQTLs in renal tissue, this demonstrates the ability of the studied elements to establish a SNP dependent expression pattern of captured target genes in the 4C-seq approach. Interesting eQTL target genes also

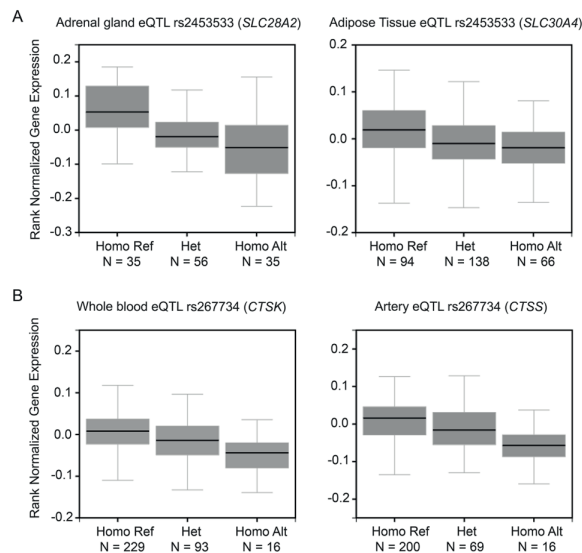


Figure 6: Expression levels of 4C-seq captured genes are correlated with the associated CKD SNP. (A) The expression level of solute carriers SLC28A2 and SLC30A4 is lower in the presence of the heterozygous (Het) and homozygous alternative (Homo Alt) allele (rs2453533), compared to the homozygous reference (Homo Ref) “wildtype” allele, p-value 6.8E-10 and 1.9E-05, respectively (adjusted from GTEx portal). (B) Similarly, the expression levels of the secreted proteases CTSS and CTSK are lower in the presence of the heterozygous and homozygous alternative allele (rs267734), compared to the homozygous reference allele, p-value 1.4E-07 and 0.5E-06, respectively (adjusted from GTEx portal). SNP-target gene pair p-values were based on matrix eQTL analysis in linear regression mode, as described by the GTEx consortium³⁹.

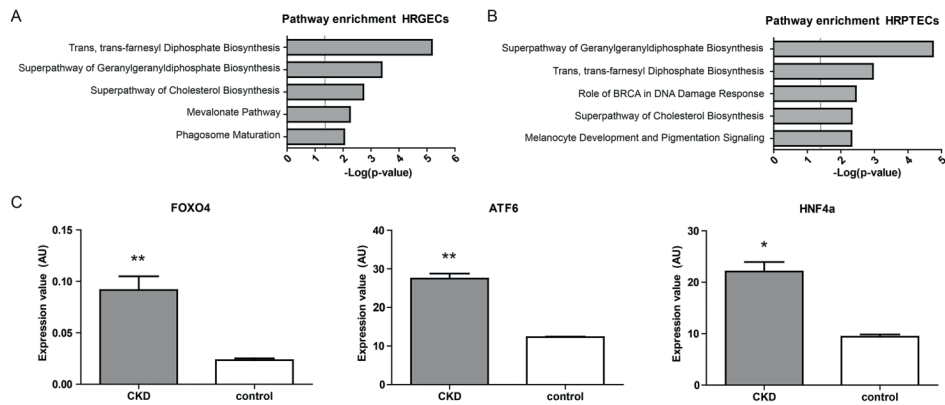


Figure 7: CKD candidate genes are enriched in pathways of biosynthesis, microangiopathy and molecular transport. IPA revealed that CKD candidate genes in both HRGECs (A) and HRPTECs (B) were most enriched in biosynthesis pathways, including the superpathway of geranylgeranyl diphosphate biosynthesis, and the trans, trans-farnesyl diphosphate biosynthesis pathway. These pathways play crucial roles in protein reuptake in HRPTECs. P-values were calculated by a right-tailed Fisher Exact Test. (C) Upstream regulators, identified by IPA based on the enrichment of their target genes in the 4C-seq derived candidates, were significantly higher expressed in renal biopsy specimens from CKD patients (derived from GSE66494). * $p < 0.05$, ** $p < 0.001$, p-values were calculated by a non-parametric t-test.

picked up by 4C-seq include the solute carriers *SLC28A2* and *SLC30A4* (rs2453533), and in relation to renal fibrosis, the genes encoding for the secreted proteases *CTSS* and *CTSK* (rs267734) (Figure 6A,B).

Pathway analysis reveals potentially disrupted mechanisms and regulators in CKD

Besides studying individual loci and interacting genes, we used Ingenuity Pathway Analysis (IPA, QIAGEN) to determine the pathways in which the CKD candidate genes are involved in. In both HRGECs and HRPTECs the 4C candidate genes were most significantly enriched in biosynthesis pathways (supplemental table 6 and 7), including the super-pathway of geranylgeranyl diphosphate biosynthesis ($p = 4.17E-04$, $p = 7.26E-04$), and the trans, trans-farnesyl diphosphate biosynthesis pathway ($p = 6.63E-06$, $p = 1.11E-03$), respectively (Figure 7A-B). Interestingly, both molecular pathways are linked to the mevalonate pathway (supplemental figure 6.2). 4C-seq identified *FDPS* and *PMVK* (rs2049805), and *IDI1* and *IDI2* (rs10794720) as candidate genes in relation to the mevalonate pathway in CKD.

In addition, IPA was used to identify upstream regulators of which the target genes were overrepresented amongst the CKD candidate genes. Targets of the transcription factors ATF6 ($p = 2.00E-03$, $p = 4.50E-04$) and FOXO4 ($p = 5.23E-05$, $p = 1.24E-04$) were significantly enriched in the CKD candidates in both HRGECs and HRPTECs respectively, whereas targets of HNF4 α were only enriched in HRPTECs ($p = 3.76E-02$). It was previously shown that HNF4 α was crucial for establishing and maintaining transcriptional enhancer elements in the renal proximal tubule and that sub-optimal DNA binding properties

amongst others led to transcriptional dysregulation of a variety of solute carriers¹⁷. Interestingly, from publically available micro-array data (NCBI Gene Expression Omnibus accession number GSE66494)¹⁸ all three transcription factors were found to be significantly upregulated in renal biopsies from patients with CKD compared to healthy controls (**Figure 7C**), suggesting these three factors are potentially key regulators in CKD.

Discussion

The main findings of the study are: (1) CKD-associated variation can affect transcriptional regulation, as demonstrated in a proof of principle approach for rs11959928 using STARR-seq. (2) CKD-associated loci interact with promoter regions of target genes via 3D chromatin folding. By taking this DNA regulatory information into account in GWAS annotation, we found many novel CKD candidate genes. (3) Multiple SNP-target genes sets can be distinguished. (4) The identified target genes can be linked to CKD in human and murine disease databases (OMIM and MGI). (5) eQTL analysis reveals that expression of many target genes are genotype dependent. (6) HRGECs and HRPTECs-derived target genes share trans, trans-farnesyl diphosphate biosynthesis pathway as a common molecular mechanism. Combined, our data annotated multiple new genes to previously reported CKD-associated SNPs and provided first time evidence for direct interaction between these common variant regions and their targets. Future studies are now required to pinpoint causal genetic variant(s) at each locus, allowing a deeper understanding of their associated disease mechanisms and their relevance in kidney disease.

Previously reported GWASs for CKD-associated SNPs used classic annotation based on spatial proximity principles to identify affected target genes⁴. This includes annotation of SNPs based on location in coding regions or in close proximity of TSS (taking into account that the average promoter is 100-1000bp long), but also taking into consideration that these SNPs could be markers for less common variants in gene bodies. Using STARR-seq analysis, variants in LD with the CKD-associated SNP rs11959928 were shown to affect activity of a DRE in an allele specific manner, emphasizing that not only protein coding variants, but also variants located in regulatory regions could be at the basis of development or progression of complex diseases. Therefore, we examined the 3D folded state of the chromatin by 4C-seq to list genes that interact with DREs in LD with CKD associated SNPs. This approach led to the identification of 304 CKD candidate genes of which many are not directly located near the associated susceptibility loci. Most enhancers are located several hundreds, sometimes even thousands kbp from their target genes⁹. In our study, the majority of the SNPs are located between 100-500kbp from the target genes' TSS, supporting the idea that the interactions observed in the 4C-seq approach are enhancer-target gene interactions. The effect of common genetic variants in DNA regulatory elements is relatively low¹⁹, therefore it is unlikely that the CKD associated SNPs in these elements will result in creating new or completely ablating

3D DRE-gene interactions. Rather, dysregulation in the expression of a gene profile that is part of the regulation of kidney homeostasis in healthy individuals, is more likely the contributing factor in CKD etiology. The 4C-seq approach helps interpreting genetic variants as a determining factor of the expression levels of interacting targets in the pathogenesis of CKD.

By overlapping our 304 CKD target genes with datasets provided by OMIM and MGI, we confirmed their relevance to kidney disease: Analysis with the MGI database showed that mice deficient for the CKD candidate genes *MPV17*, *CCNI*, *ASH1L* and *SLC4A5* suffer from renal failure related traits²⁰⁻²³. These genes are located 185kbp (rs1260326), 585kbp (rs13246355), 337kbp (rs2049805), and 702kbp (rs13538) from the CKD associated SNP respectively. In addition, validation of the 4C-seq approach was provided by analysis in the OMIM dataset, which demonstrated that multiple target genes were linked to CKD associated traits in human. For example, *MUC1* and *PKD2* were identified by 4C-seq, as a result of the interaction of their promoter regions with regulatory domains that colocalize with rs2049805 and rs2725220, respectively. *MUC1* encodes the protein mucin-1, which is a membrane anchored mucoprotein involved in providing a protective barrier against pathogens. A frameshift mutation in *MUC1*, leading to a novel stop codon, induced medullar cystic kidney disease type 1 (MIM: 158340)²⁴. *PKD2* encodes the polycystin-2 protein, which is involved in renal calcium transport and calcium signaling. Mutations in the gene, leading to loss of function, causes the formation of fluid-filled cysts, eventually leading to progressive destruction of the renal parenchyma (MIM: 173910)²⁵, but was recently also demonstrated to be involved in branching and network formation of lymphatic endothelium which plays a crucial role in renal function^{26, 27}. Such examples illustrate the potential relevance of the candidate genes identified by 4C-seq for renal function, and provide clear evidence for the functional association between the investigated SNP regions and their corresponding 4C-seq captured genes in (human) CKD.

Evidence for the transcription regulatory function of the CKD-associated SNP regions was provided by GTEx database analysis. Many of the CKD susceptibility loci were eQTLs, showing a significant correlation between SNP genotypes and expression level of linked target genes in a variety of tissues. These eQTL target genes include *CTSS* and *CTSK*, coding for cathepsin S and K respectively, both downregulated in the presence of rs267734. Cathepsins are potent proteases and the negative correlation between rs267734 and cathepsin S and K expression might be relevant in relation to renal fibrosis, which is critically involved in CKD progression. In a bleomycin lung fibrosis model it was shown that cathepsin K deficient mice had more severe lung fibrosis than wild-type mice²⁸. Furthermore, it was observed that pharmacological inhibition of cathepsin activity in mice with unilateral ureteral obstruction induced renal fibrosis led to a worse outcome²⁹, indicating that reduced expression of cathepsin S and K in the presence of rs267734 could contribute to CKD.

Pathway analysis demonstrated enrichment of 4C-seq captured genes in multiple biosynthesis pathways involved in the generation of isoprenoid pyrophosphates. Interestingly, this enrichment was observed in HRGECs, and HRPTECs, though with different identified target genes per cell type. Isoprenoid pyrophosphates are indispensable for renal proximal tubular protein reabsorption, as inhibition of 3-hydroxy-3-methylglutaryl CoA reductase in the mevalonate pathway leads to reduced prenylation of GTP binding proteins involved in receptor mediated endocytosis, eventually resulting in proteinuria³⁰. Similarly, altered levels of prenylation of RhoA affects eNOS activity in endothelial cells, resulting in imbalance of ROS levels, contributing to the endothelial dysfunction reported in CKD onset³¹.

In our systemic approach, besides the non-coding variants, we also included 2 non-synonymous SNPs (SNPs in gene coding regions that alter protein sequence: rs1260326 in *GCKR* and rs13538 in *NAT8*) were amongst the studied regions. Presumably, the affected genes are involved in the associated disease phenotype. Especially reports on *NAT8* activity in association with kidney disease seem convincing³²⁻³⁴. However, it was previously demonstrated that DREs can also be located in coding regions³⁵, and it remains of interest that by 4C-seq we found interactions of this locus with TSS of multiple other genes of which the expression levels according to GTEx are significantly associated with the occurrence of the variant. The incorporation of regulatory information provides an additional layer in post-GWAS data to aid in our interpretation of these GWAS datasets, but certainly does not replace the candidate genes identified based on spatial proximity such as *NAT8*. Along the same lines, several SNPs associated with a single trait were included. Several SNPs are solely associated with serum creatinine (estimated glomerular filtration rate). Although these SNPs might be causally associated with CKD, they might also affect creatinine production/secretion independent of renal function. Rs91119 is only associated with serum cystatin C (estimated glomerular filtration rate), and is located directly within the *CST* locus. Again, this SNP does not necessarily have to be causally associated with CKD, but could also be involved in the dynamics of cystatin C production. The same is true for SNPs solely associated with serum urate or blood urea nitrogen, although the authors that identified SNPs associated with the latter group had corrected for non-renal factors³.

In conclusion, taking the 3D structure of chromatin into account, we have identified 304 putative CKD candidate genes of DREs that colocalize with CKD susceptibility loci. In this hypothesis generation-driven approach, we present a new method of GWAS interpretation based on DRE target gene identification by 4C analysis that complements the classic methods of candidate gene identification. In addition, incorporation of the adapted STARR-seq method up or down stream of the 4C pipeline, would further narrow down the identification of causal variants in DNA regulatory function, and help us to greatly expand our understanding of the role that common low risk variants play in the onset of complex diseases such as CKD.

Methods

Cell culture

Primary human renal glomerular endothelial cells (HRGECs; derived from human donor cell biobank Sciencell) and human renal proximal tubular epithelial cells (HRPTECs; Sciencell) were cultured on fibronectin and gelatin coated plates on ECM medium (supplemented with endothelial cell growth kit, and penicillin/streptomycin; Sciencell), and EpiCM medium (supplemented with epithelial cell growth kit, and penicillin/streptomycin; Sciencell) respectively. Human Embryonic Kidney Cells (HEK293a) were cultured on gelatin coated plates on DMEM (Lonza) supplemented with 10% fetal calf serum (Gibco) and 100U/ml penicillin/streptomycin (Lonza). All cells were cultured in 5% CO₂ at 37 °C. The experiments with primary cells were conducted with cells at passage 3.

STARR-seq reporter assay

DNA from 20 individual donors was isolated from whole blood obtained from the Mini Donor Service (positive approval from the medical ethics committee of the UMC Utrecht – protocol number 07/125) via salt precipitation. Regions (~1200 base pairs in size) containing DNase hypersensitivity sites overlapping with candidate variants (minor allele frequency > 0.03) within the susceptibility locus were PCR amplified (primers in supplemental table 1), equimolarly pooled and cloned into pSTARR-seq_human vector (Addgene). The library complexity was verified by dilution series after transformation and was estimated to contain 50.000 individual reporter clones. 40 million cells were placed in 1600µl electroporation buffer (Bio-Rad) supplemented with 120µg (HEK293a) or 240µg (HRPTECs and HRGECs) library, after which electroporation mixture was divided over 16 2mm electroporation cuvettes (Bio-Rad), followed by electroporation with a square wave (110V/25ms for HEK293a, 125V/20ms for HRPTECs and HRGECs) using Gene Pulser Xcell™ (Bio-Rad). After electroporation cells were seeded in normal culture medium for 24h, followed by RNA extraction using the RNeasy isolation kit (Qiagen). The polyadenylated fraction of the total RNA was isolated using Dynabeads Oligo dT 25 (Thermo Fisher). The reporter specific cDNA was synthesized and amplified according to standard STARR-seq protocols³⁴. The amplified cDNA was subsequently fragmented (Covaris S2 ChIP seq programme (power peak 40, duty factor 5 cycle/burst 200)) and cleaned, followed by sequencing library preparation using NEXTflex ChIP-Seq library prep kit for Illumina sequencing. The libraries were sequenced on Illumina NextSeq500 platform to produce 75bp long single end reads.

4C-seq

4C template was prepared as previously described³⁶. In summary, 10 million HRGECs or HRPTECs (both primary cells from healthy donors) were fixed in 2% formaldehyde, after which cells were lysed. The chromatin of the lysed cells was digested with the 4 base cutter DPNII (NEB), followed by ligation in a heavily diluted environment with T₄ ligase (Roche). The ligated samples were de-crosslinked, followed by a second digestion with the 4 base cutter CviQI (NEB). Next, samples were ligated once more in a diluted environment after which the chromatin was purified. The efficiency of each digestion and ligation step was validated on agarose gels. Viewpoints were selected based on the CKD susceptibility loci found in the GWASs of *Y. Okada et al.* (2012) and *A. Köttgen et al.* (2010). If multiple SNPs were found in a genomic region spanning less than 20kbp, only the SNP with the lowest P value was selected as viewpoint. To study the chromatin interactions of CKD associated susceptibility loci with 4C, primers were designed for each viewpoint as described previously³⁶. Briefly, primers were designed within a 5kbp window surrounding the CKD associated SNP. Forward primers were designed in the first restriction site and the reversed primers were designed close to the second restriction site (<100bp), with a minimum distance of 300bp between the forward and the reversed primer. In case no suitable primers could be designed based on these specifications, either the window size surrounding the SNP was increased to 10kbp or the distance between the forward and reversed primer was reduced (supplemental table 8). 4C libraries were sequenced using the NextSeq500 platform (Illumina), producing single end reads of 75bp. The raw

sequencing reads were then de-multiplexed based on viewpoint specific primer sequences. Reads were trimmed to 16 bases and mapped to an *in silico* generated library of fragends (fragment ends) neighboring all DpnII sites in human genome (NCBI37/hg19), using the custom Perl scripts³⁷. No mismatches were allowed during the mapping. The reads mapping to only one possible fragend were used for further analysis.

Target gene identification

First, the number of covered fragends within a running window of k fragends throughout the whole chromosome was calculated (only the viewpoint's chromosome was taken into account). The k was set separately for every viewpoint so it contained on average 20 covered fragends in the area around the viewpoint (± 100 kbp). Next, we compared the number of covered fragends in each running window to the theoretical random distribution. The windows with significantly higher number of covered fragends compared to random distribution ($p < 10^{-8}$ based on binominal cumulative distribution function; R `pbinom`) were considered as a significant 4C signal. The following criteria were defined for the identification of the candidate genes: (1) The Transcriptional Start Site (TSS) colocalizes with a significant 4C-seq signal ($p < 10^{-8}$) within 5kbp. (2) The susceptibility variant or other variant in linkage disequilibrium (LD) colocalizes with at least one of the published datasets that represent candidate regulatory sequences (**supplemental table 3**) in a similar cell type as from which the 4C signal was. (3) The candidate gene has been validated to be expressed by mRNA datasets.

Identification of gene expression

HRPTECs expression data was used from publically available datasets (NCBI Gene Expression Omnibus accession number GSE12792)³⁸. Expression data from microvascular endothelium was generated via RNA extraction from cultured microvascular endothelial cells in low serum medium (EBM-2 medium supplemented with 0.5% fetal calf serum) using the RNeasy isolation kit. Poly(A) Beads (NEXTflex) were used to isolate polyadenylated mRNA, from which sequencing libraries were made using the Rapid Directional RNA-seq kit (NEXTflex). Libraries were sequenced using the Nextseq500 platform (Illumina), producing single end reads of 75bp. Reads were aligned to the human reference genome GRCh37 using STAR version 2.4.2a. Picard's AddOrReplaceReadGroups (v1.98) was used to add read groups to the BAM files, which were sorted with Sambamba v0.4.5 and transcript abundances were quantified with HTSeq-count version 0.6.1p1 using the union mode. Subsequently, RPKMs were calculated with edgeR's RPKM function. Genes were accepted as expressed if probe intensity > 6 or $\log_2(\text{RPKM}) > -1$ in HRPTECs and microvascular endothelium, respectively.

Haploblock localization

Haploview (Broad Institute) was used to download LD plots 500kb up- and downstream from CKD associated SNPs (pairwise comparisons of markers < 2000 kbp apart). From these LD plots, haploblocks, containing CKD associated SNPs, were extracted to evaluate target gene localization in relation to CKD associated susceptibility region.

Genetic annotation with OMIM

The Online Mendelian Inheritance in Man (OMIM) morbid map database was used to find mutant alleles that were associated with CKD. CKD associated traits were mapped based on the phenotype category queries 'renal', 'kidney', and 'nephro'. The gene set found in OMIM was used to identify known CKD associated genes in the list of genes generated via the 4C-seq approach.

Genetic annotation with MGI

The MGI database was used to find monogenic mutant murine alleles that led to CKD related traits. A data file containing the 'approved gene name' and the 'mouse genome database ID' was downloaded from the HUGO Gene Nomenclature Committee to identify the mutated murine genes in the MGI database that led to CKD related traits. CKD related traits were mapped based on the following phenotype categories: Abnormal kidney morphology, abnormal kidney angiogenesis,

urine abnormalities, blood abnormalities, glomerulus abnormalities, renal tubules abnormalities, podocyte abnormalities, kidney cysts, abnormal renal filtration, other kidney related traits. The gene set found in MGI was used to identify known CKD associated genes in the list of genes generated via the 4C-seq approach.

eQTL study in GTEx portal

The GTEx portal database, containing data on eQTL in 449 genotyped donors with expression data in 44 different tissues was used to list genes that significantly correlated in their expression with CKD associated SNPs that colocalized with active DREs. Genes found via this approach were overlapped with the 4C-seq captured gene list to validate whether the 4C-seq approach indeed detected target genes that showed correlations in expression levels with the CKD associated SNPs.

Pathway analysis

Datasets were analyzed using QIAGEN's IPA. IPA was used to study both enrichment of 4C-seq identified genes in canonical pathways and upstream regulators of identified candidate genes, independently for HRGECs and HRPTECs. P-values were calculated based on a right-tailed Fisher Exact Test, calculated by IPA. Expression levels of upstream regulators of which target genes were found enriched in the candidate genes identified by 4C-seq were evaluated in a publically available micro-array dataset which was used to study gene expression in CKD in renal biopsy specimens (NCBI Gene Expression Omnibus accession number GSE66494)¹⁸.

References

1. Levey, AS, Beto, JA, Coronado, et al.: Controlling the epidemic of cardiovascular disease in chronic renal disease: what do we know? What do we need to learn? Where do we go from here? National Kidney Foundation Task Force on Cardiovascular Disease. *Am J Kidney Dis*, 32: 853-906, 1998.
2. Kottgen, A, Pattaro, C, Boger, et al.: New loci associated with kidney function and chronic kidney disease. *Nat Genet*, 42: 376-384, 2010.
3. Okada, Y, Sim, X, Go, et al.: Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat Genet*, 44: 904-909, 2012.
4. Raychaudhuri, S, Plenge, RM, Rossin, EJ, Ng, AC, International Schizophrenia, C, Purcell, SM, Sklar, P, Scolnick, EM, Xavier, RJ, Altshuler, D, Daly, MJ: Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet*, 5: e1000534, 2009.
5. Li, G, Vega, R, Nelms, K, Gekakis, N, Goodnow, C, McNamara, P, Wu, H, Hong, NA, Glynne, R: A role for Alstrom syndrome protein, alms1, in kidney ciliogenesis and cellular quiescence. *PLoS Genet*, 3: e8, 2007.
6. Mo, L, Huang, HY, Zhu, XH, Shapiro, E, Hastly, DL, Wu, XR: Tamm-Horsfall protein is a critical renal defense factor protecting against calcium oxalate crystal formation. *Kidney Int*, 66: 1159-1166, 2004.
7. Ghirotto, S, Tassi, F, Barbujani, G, Pattini, L, Hayward, C, Vollenweider, P, Bochud, M, Rampoldi, L, Devuyst, O: The Uromodulin Gene Locus Shows Evidence of Pathogen Adaptation through Human Evolution. *J Am Soc Nephrol*, 2016.
8. Maurano, MT, Humbert, R, Rynes et al.: Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337: 1190-1195, 2012.
9. Williamson, I, Hill, RE, Bickmore, WA: Enhancers: from developmental genetics to the genetics of common human disease. *Dev Cell*, 21: 17-19, 2011.
10. Visser, M, Kayser, M, Palstra, RJ: HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res*, 22: 446-455, 2012.
11. Bohle, A, Grund, KE, Mackensen, S, Tolon, M: Correlations between renal interstitium and level of serum creatinine. Morphometric investigations of biopsies in perimembranous glomerulonephritis. *Virchows Arch A Pathol Anat Histol*, 373: 15-22, 1977.
12. Nangaku, M: Chronic hypoxia and tubulointerstitial injury: a final common pathway to end-stage renal failure. *J Am Soc Nephrol*, 17: 17-25, 2006.
13. Kozakowski, N, Herkner, H, Bohmig, GA, Regele, H, Kornauth, C, Bond, G, Kikic, Z: The diffuse extent of peritubular capillaritis in renal allograft rejection is an independent risk factor for graft loss. *Kidney Int*, 88: 332-340, 2015.
14. Arnold, CD, Gerlach, D, Stelzer, C, Boryn, LM, Rath, M, Stark, A: Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339: 1074-1077, 2013.

15. Mokry, M, Hatzis, P, Schuijers, J, Lansu, N, Ruzius, FP, Clevers, H, Cuppen, E: Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res*, 40: 148-158, 2012.
16. Keller, BJ, Martini, S, Sedor, JR, Kretzler, M: A systems view of genetics in chronic kidney disease. *Kidney Int*, 81: 14-21, 2012.
17. Martovetsky, G, Tee, JB, Nigam, SK: Hepatocyte nuclear factors 4alpha and 1alpha regulate kidney developmental expression of drug-metabolizing enzymes and drug transporters. *Mol Pharmacol*, 84: 808-823, 2013.
18. Nakagawa, S, Nishihara, K, Miyata, et al.: Molecular Markers of Tubulointerstitial Fibrosis and Tubular Cell Damage in Patients with Chronic Kidney Disease. *PLoS One*, 10: e0136994, 2015.
19. Vernet, B, Stergachis, AB, Maurano, MT, Vierstra, J, Neph, S, Thurman, RE, Stamatoyannopoulos, JA, Akey, JM: Personal and population genomics of human regulatory variation. *Genome Res*, 22: 1689-1697, 2012.
20. Viscomi, C, Spinazzola, A, Maggioni, M, Fernandez-Vizarra, E, Massa, V, Pagano, C, Vettor, R, Mora, M, Zeviani, M: Early-onset liver mtDNA depletion and late-onset proteinuric nephropathy in Mpv17 knockout mice. *Hum Mol Genet*, 18: 12-26, 2009.
21. Griffin, SV, Olivier, JP, Pippin, JW, Roberts, JM, Shankland, SJ: Cyclin I protects podocytes from apoptosis. *J Biol Chem*, 281: 28048-28057, 2006.
22. Xia, M, Liu, J, Wu, X, Liu, S, Li, G, Han, C, Song, L, Li, Z, Wang, Q, Wang, J, Xu, T, Cao, X: Histone methyltransferase Ash1 suppresses interleukin-6 production and inflammatory autoimmune diseases by inducing the ubiquitin-editing enzyme A20. *Immunity*, 39: 470-481, 2013.
23. Groger, N, Vitzthum, H, Frohlich, H, Kruger, M, Ehmke, H, Braun, T, Boettger, T: Targeted mutation of SLC4A5 induces arterial hypertension and renal metabolic acidosis. *Hum Mol Genet*, 21: 1025-1036, 2012.
24. Kirby, A, Gnirke, A, Jaffe, DB et al.: Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat Genet*, 45: 299-303, 2013.
25. Koptides, M, Hadjimichael, C, Koupepidou, P, Pierides, A, Constantinou Deltas, C: Germinal and somatic mutations in the PKD2 gene of renal cysts in autosomal dominant polycystic kidney disease. *Hum Mol Genet*, 8: 509-513, 1999.
26. Stolarczyk, J, Carone, FA: Effects of renal lymphatic occlusion and venous constriction on renal function. *Am J Pathol*, 78: 285-296, 1975.
27. Outeda, P, Huso, DL, Fisher, SA, Halushka, MK, Kim, H, Qian, F, Germino, GG, Watnick, T: Polycystin signaling is required for directed endothelial cell migration and lymphatic development. *Cell Rep*, 7: 634-644, 2014.
28. Buhling, F, Rocken, C, Brasch, F, Hartig, R, Yasuda, Y, Saftig, P, Bromme, D, Welte, T: Pivotal role of cathepsin K in lung fibrosis. *Am J Pathol*, 164: 2203-2216, 2004.
29. Lopez-Guisa, JM, Cai, X, Collins, SJ, Yamaguchi, I, Okamura, DM, Bugge, TH, Isacke, CM, Emson, CL, Turner, SM, Shankland, SJ, Eddy, AA: Mannose receptor 2 attenuates renal fibrosis. *J Am Soc Nephrol*, 23: 236-251, 2012.
30. Sidaway, JE, Davidson, RG, McTaggart, F, Orton, TC, Scott, RC, Smith, GJ, Brunskill, NJ: Inhibitors of 3-hydroxy-3-methylglutaryl-CoA reductase reduce receptor-mediated endocytosis in opossum kidney cells. *J Am Soc Nephrol*, 15: 2258-2265, 2004.
31. Eelen, G, de Zeeuw, P, Simons, M, Carmeliet, P: Endothelial cell metabolism in normal and diseased vasculature. *Circ Res*, 116: 1231-1244, 2015.
32. Chambers, JC, Zhang, W, Lord, GM et al.: Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet*, 42: 373-375, 2010.
33. Suhre, K, Shin, SY, Petersen, AK et al.: Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477: 54-60, 2011.
34. Yu, B, Zheng, Y, Alexander, D, Morrison, AC, Coresh, J, Boerwinkle, E: Genetic determinants influencing human serum metabolome among African Americans. *PLoS Genet*, 10: e1004212, 2014.
35. Birnbaum, RY, Clowney, EJ, Agamy, O et al.: Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res*, 22: 1059-1068, 2012.
36. van de Werken, HJ, de Vree, PJ, Splinter, E, Holwerda, SJ, Klous, P, de Wit, E, de Laat, W: 4C technology: protocols and data analysis. *Methods Enzymol*, 513: 89-112, 2012.
37. Ji, H, Jiang, H, Ma, W, Johnson, DS, Myers, RM, Wong, WH: An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*, 26: 1293-1300, 2008.
38. Beyer, S, Kristensen, MM, Jensen, KS, Johansen, JV, Staller, P: The histone demethylases JMJD1A and JMJD2B are transcriptional targets of hypoxia-inducible factor HIF. *J Biol Chem*, 283: 36542-36552, 2008.
39. Consortium, GT: Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348: 648-660, 2015.



Stem cells are the principal intestinal epithelial responders to bacterial antigens.

7

Claartje A. Meddens, Maaike H. de Vries, Hemme J. Hijma, Anne Claire Berrens, Bas Westendorp, Jet van der Spek, Berend A.P. Kooiman, Renée R.C.E. Schreurs, Miguel Vera, Erik Lijster, Sinisa Prelic, Ninke M. Nieuwenhuis, Evelyn S. Hanemaaijer, Nico Lansu, Noortje van den Dungen, Madeleine J. Bunders, Hans Clevers, Michal Mokry, Edward E.S. Nieuwenhuis

In preparation

Introduction

The intestinal lumen is loaded with a vast number of microorganisms that results in perpetual microbial exposition of intestinal epithelial cells (IECs). Intriguingly, inflammatory responses, including cytokine production by IEC (Damen et al., 2006) are absent under homeostatic conditions. However, upon an encounter with pathogenic bacteria or during epithelial barrier disruption, inflammatory responses must be guaranteed to oppose ongoing invasion and microbial dissemination into the bloodstream. The necessity of two seemingly opposing biological functions dealing with continuous microbial challenge, stresses a high-level of regulation of responsiveness by IECs (Menckeborg et al., 2014; Rakoff-Nahoum et al., 2004). We studied the mechanisms of intestinal epithelial responsiveness at a cellular level, in the context of epithelial cell polarization, the state of epithelial cell differentiation, and the specific location along the proximo-distal axis.

The intestines can be classified in regions that contain different cells that correspond to distinctive functions. Epithelial cells can be studied based on their exact location within the gastrointestinal tract. As such, three axes can be defined. First, the longitudinal axis: the composition of cell types changes from the lining of the duodenum to the colon (Anderle et al., 2005; Haber et al., 2017; Middendorp et al., 2014; Price et al., 2018). Second, the crypt-villus axis: this axis represents the different states of differentiation of stem cells that are located in the crypt bottoms versus fully differentiated IECs at the villus tip or the flat surface epithelium in the colon (van der Flier and Clevers, 2009). Third, the apical-basolateral axis: I.e. the IEC polarity that results from the differential distribution of various molecules including membrane bound receptors that are involved in microbial sensing (Abreu, 2010). Next to the determination of responses based on the exact location along these three axes, epithelial cells can also be studied at a molecular level. In general, microbial responsiveness is regulated through mechanisms that include the alteration of PRR (Pathogen-associated molecular pattern Recognition Receptor), expression and modification of molecules involved in intracellular signal transduction and various transcriptional and post-transcriptional processes (Rescigno and Nieuwenhuis, 2007; Stumpo et al., 2010).

The provision of continuous renewal and repair of the intestinal mucosal epithelium depends on resident intestinal stem cells (ISCs) that constantly undergo the processes of proliferation and differentiation. Along the crypt-villus axis, various transcription factors that include Wnt and Notch and their respective receptors are differentially expressed. These highly regulated signaling pathways result in a stem cell pool that resides in the crypt and can divide and differentiate while moving up towards the tip of the villus (Yin et al., 2014). Previously, it was shown that exposure to microbial metabolites or pathogens, or in case of cellular damage, either induction or inhibition of regenerative

activity by IECs can occur (Jain et al., 2018; Kaiko et al., 2016; Nusse et al., 2018; Schmitt et al., 2018). The fine-tuning of these intricate mechanisms however, remains elusive.

Here we use human intestinal organoids to study epithelial responsiveness to bacterial antigens along the three intestinal-axes as mentioned. We identified major responsive and hypo-responsive epithelial cell types by using single cell RNA-sequencing and we further explored the underlying molecular mechanisms.

Results

Epithelial crypt cells respond to bacterial antigens upon basolateral exposure

To investigate the responsiveness of intestinal epithelial cells (IECs) to microbial antigens, we made use of human intestinal organoids and exposed them to bacterial lysates (**figure 1A**). The bacterial lysates contain a pool of Pathogen-Associated Molecular Patterns (PAMPs) that stimulate IECs through various Pattern Recognition Receptors (PRRs).

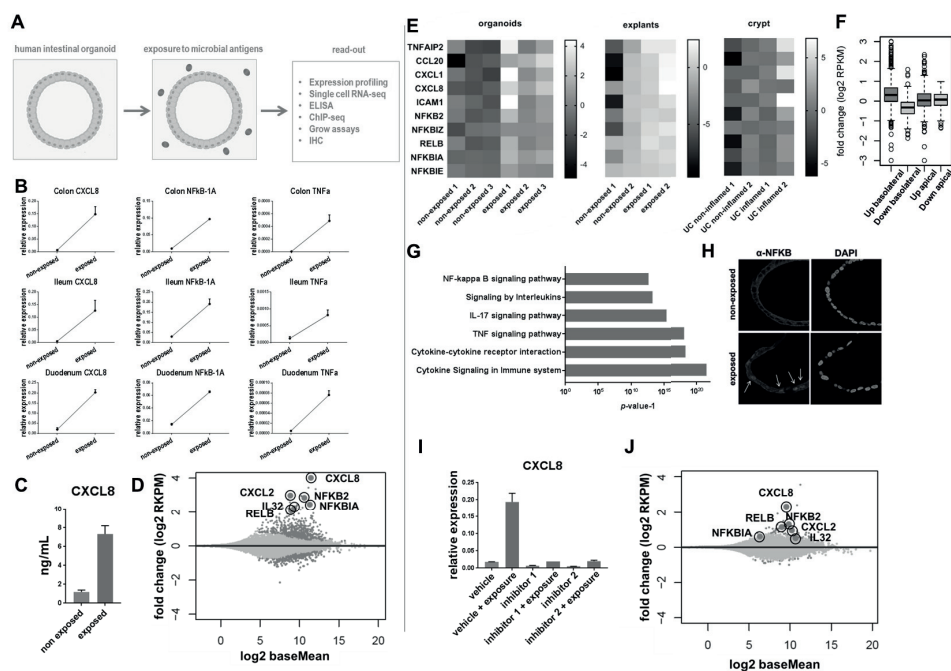


Figure 7: CKD candidate genes are enriched in pathways of biosynthesis, microangiopathy and molecular transport.

IPA revealed that CKD candidate genes in both HRGECs (A) and HRPTECs (B) were most enriched in biosynthesis pathways, including the superpathway of geranylgeranyl diphosphate biosynthesis, and the trans-farnesyl diphosphate biosynthesis pathway. These pathways play crucial roles in protein reuptake in HRPTECs. P-values were calculated by a right-tailed Fisher Exact Test. (C) Upstream regulators, identified by IPA based on the enrichment of their target genes in the 4C-seq derived candidates, were significantly higher expressed in renal biopsy specimens from CKD patients (derived from GSE66494). * $p < 0.05$, ** $p < 0.001$, p-values were calculated by a non-parametric t-test.

The organoids were cultured in expansion medium (EM) containing RSPO1, noggin, EGF, N-acetylcysteine, B27, A83-01, nicotinamide, SB202190, and WNT3A. Under these conditions the organoids mainly consist of undifferentiated cells that *in vivo* reside in the crypt (Sato et al., 2011). All intestinal organoids, regardless of their location of origin along the longitudinal axis, responded to bacterial lysates by upregulating inflammatory genes including *CXCL8*, *TNF* and *NFKBIA* (**figure 1B**, **figure S7.1A**). In line with the typical pro-inflammatory response pattern as shown at the transcriptional level, the organoids released *CXCL8* into the medium upon exposure (**figure 1C**). We have previously shown that intestinal organoids retain their location specific baseline RNA expression profiles when cultured *in vitro* (Middendorp et al., 2014). We conclude that responsiveness to microbial stimulation of IECs is constantly present and is not altered at different locations along the longitudinal axes.

Next, we performed RNA-seq on colon organoids upon microbial challenge, varying the exposure time from 0 to 190 hours (**figure S7.1B**). Exposure for 6 hours showed the highest response and resulted in differential expression of >1000 genes (712 up regulated, 314 down regulated **figure 1D**, **table S1**) and was further used as a reference time point in all our assays (unless stated otherwise). The response is characterized by the upregulation of multiple inflammatory genes including *TNF*, *CXCL8*, *NFKB2* and *NFKBIA*. In order to validate our model system, we profiled transcriptomes of freshly isolated crypts from inflamed and non-inflamed colon epithelia derived from ulcerative colitis patients and exposed colon explants (i.e. intact biopsies) to bacterial lysates. Both crypts and explants exhibit similar expression patterns compared to those that we identified in the organoid model (**figure 1E**). We conclude that the responses in our *in vitro* system reflect the intestinal epithelial inflammatory responses *in vivo*.

Under physiological conditions the intestinal epithelium is exposed to bacterial antigens at the apical (luminal) surface only, and receptors that sense microbial antigens are not evenly distributed along the apical- basolateral axis (Abreu, 2010). To address the role of cell polarity in microbial responsiveness, we grew organoids as monolayers and exposed them from either the apical or basolateral side. In contrast to a robust response upon basolateral exposure, apical responses were attenuated under similar conditions (**figure 1F**, **figure S7.1C**). These findings implicate that in the *in vivo* situation, epithelial cells within the intestinal crypt may only respond upon disruption of the barrier that allows for passage of microbial molecules and subsequent basolateral exposure. This can occur when the barrier function of the intestinal epithelium is affected due to, for example, invasive microorganisms, genetic defects that are associated with increased permeability of the epithelial layer (Avitzur et al., 2014; Bigorgne et al., 2014) or disruption of the epithelia as was found in inflamed intestinal mucosa in Crohn's disease and ulcerative colitis. (Gassler et al., 2001; Vetrano et al., 2008).

We next performed an unsupervised clustering analysis of the IEC genes that responded to microbial exposure. The cluster that contained the highest upregulated genes, was

mainly involved in inflammatory pathways (**figure 1G**, **figure S7.1B** cluster1, **table S2 and S3**). Notably, many of these genes are regulated through NF κ B-signaling and by using a transcription factor binding motif enrichment analysis, we established that the NF κ B binding motif is in fact the most enriched motif within the promoters of upregulated genes ($p = 1.014 \times 10^{-18}$). Immunohistochemistry confirmed that, upon exposure, NF κ B is translocated to the nucleus and that inhibition of NF κ B-signaling blocks these epithelial responses (**figure 1H, I**). Altogether, we conclude that the inflammatory response upon microbial exposure of IECs is dependent on NF κ B-signaling.

To investigate whether epithelial cell responsiveness was limited to undifferentiated cells, we induced epithelial cell differentiation by removing nicotinamide, SB202190 and WNT3A from the culture medium (DM, differentiation medium, see methods). We determined that exposure of differentiated cells to bacterial lysates was associated with a vastly reduced response in comparison to undifferentiated organoids (**figure 1J**, **table S4**). By adding a Wnt agonist molecule (CHIR) to our cultures, we could further drive the organoid cells towards an undifferentiated, thus epithelial stem cell phenotype. (Yin et al., 2014). This condition resulted in an enhanced response, whereas induction of differentiation (i.e. decreasing the stemness) resulted in a decreased response to microbial stimulation (**figure S7.1D, E**). Next to the upregulation of inflammatory genes, microbial exposure resulted in the downregulation of the intestinal epithelial stem cell marker Leucine-rich repeat-containing G-protein coupled receptor 5 (LGR5) in undifferentiated colonic and ileal organoids (**figure S7.1D**).

We conclude that intestinal epithelial responsiveness is conserved along the longitudinal axis. In contrast, responses are exceedingly different when determined along the crypt-villus and apical-basolateral axis. Our results suggest that IECs are specifically responsive to bacterial antigens when exposed at the basolateral side and when the cells are in an undifferentiated state - as is the case at the bottom of the crypts.

Stem cells are a major contributor to the epithelial inflammatory response.

To further delineate the specific responses of the different epithelial cell types in their different states of differentiation along the crypt-villus axis, we performed single cell RNA sequencing on differentiated and undifferentiated colon organoid cultures that were exposed to bacterial lysates at the basolateral side. Through the analyzes of 911 cells we were able to identify seven cell clusters (**figure 2A**) that generally represent three major different states of the IECs. Both the stem cells and the transit amplifying cells (TA cells) can be divided into two clusters. The distinctive feature that determines this separation is the cycling activity of cells that is, amongst others, marked by the expression of KI67 (**figure 2A, B, C**). Differentiated cells appeared to be separated into three clusters. One cluster consisted of enterocytes in the early phase of differentiation, based on the lack of stemness markers and the relatively low expression of differentiation markers. The other two clusters contained differentiated enterocytes (**figure 2A, B, C**).

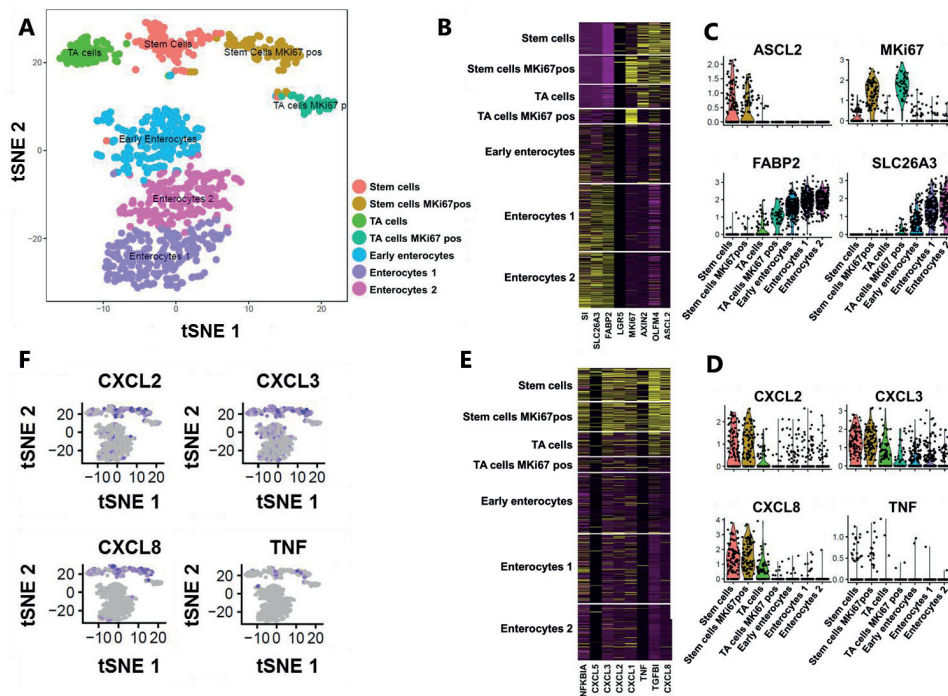


Figure 2. Single cell RNA-sequencing identifies epithelial stem cells as a major source of the inflammatory mediators upon microbial stimulation.

A. tSNE-plot of cell clusters representing different subtypes of IECs. Data were pooled from differentiated and undifferentiated colon organoids that were exposed to bacterial lysates. B. Heatmap depicting the expression of stemness and differentiation markers in each identified cell type. C. Expression of markers that are associated with different states of epithelial differentiation, KI67, SLC26A3, FABP2 and ASCL2 as determined per cell type. D. Expression of pro-inflammatory genes per cell type. E. Heatmap depicting the expression of inflammatory markers for each identified cell type F. Distribution of expression of CXCL2, 3, 8 and TNF by tSNE-plots.

Next, we projected the expression levels of multiple response genes on to the identified clusters. Cytokine expression was higher in the stem cell cluster compared to both the TA cell and the differentiated cell cluster. Multiple cytokines that are expressed upon exposure to bacterial antigens were only expressed within the stem cell clusters (**figure 2D, E**). Upregulation of the inflammatory pathway was displayed by all stem cells and was independent of Ki67 expression (**figure 2F**), thus excluding cycling activity as a crucial determinant for microbial responsiveness.

To further establish that intestinal stem cells are the main responder to microbial stimulation, we repeated single cell sequencing on exposed small intestinal organoids. Specifically, we analyzed 2027 cells and identified stem cells, TA cells and enterocytes as three separate clusters. These analyses confirmed that stem cells are the major inflammatory responders to microbial antigens in both the large and small intestine (**figure S7.2**).

Responsiveness of stem cells depends on signal transduction and post-transcriptional regulation

Our data show that the responsiveness of intestinal epithelial cells to bacterial antigens is lost upon differentiation of stem cells to enterocytes. Differences in the capacity of a cell to elicit an inflammatory response can be regulated at multiple levels. As the expression of specific TLRs in IECs remains similar in various states of differentiation (figure S7.3A), we first compared activation of the NFKB-cascade in stem cells versus differentiated cells. To this aim we used nuclear translocation of NFKB as a read-out for signal transduction upon the activation of microbial pattern recognition receptors (PRRs) and subsequent activation of the NFKB-pathway. Here we demonstrate that nuclear translocation of NFKB is present in cells of both differentiated and undifferentiated organoids that were exposed to bacterial antigens (figure 3A, B). Although the number

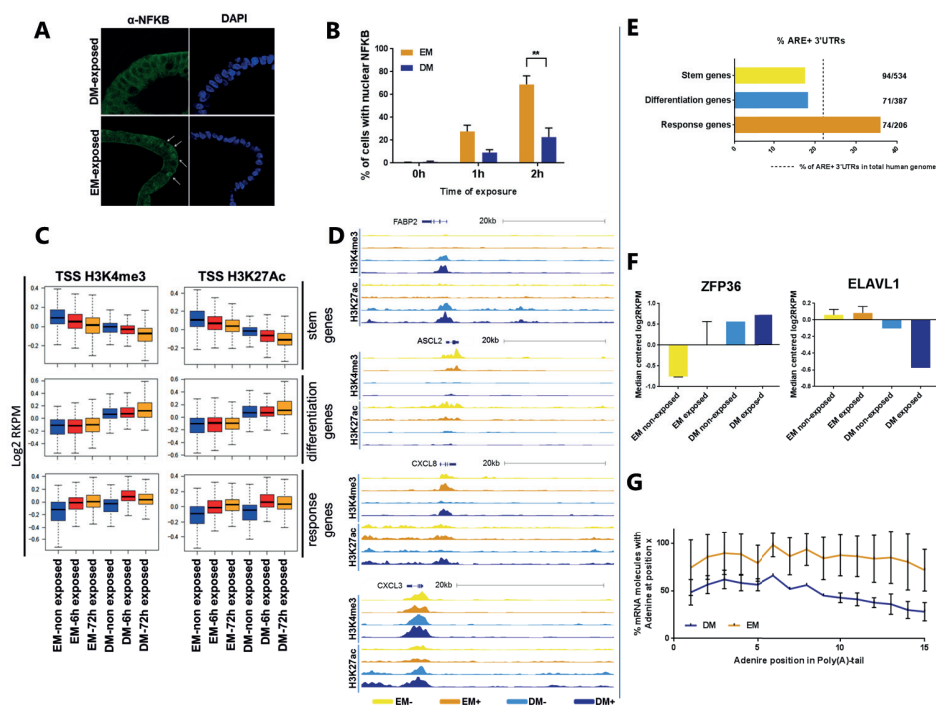


Figure 3. The mechanisms that determine responsiveness of IECs.

A. Immunohistochemistry of NFKB shows nuclear translocation for both differentiated (DM) and undifferentiated (EM) cells upon microbial exposure. B. Quantification of nuclear translocation of NFKB. The number of cells with nuclear NFKB is significantly lower in DM compared to EM (Mann-Whitney $p=0.0037$). Error bars indicate SEM. C. ChIP-seq signal of active histone modifications (H3K4me3, H3K27ac, $n=2$) in the 2kb window around the transcriptional start site of different gene groups – at the baseline and upon exposure to bacterial antigens for 6 and 72 hours. (values represent median centered and \log_2 transformed RPKM) D. ChIP tracks at four genomic regions at baseline (-) and upon 6 hours of exposure to bacterial lysates (+) E. Presence of ARE-elements in 3'UTR of stemness, differentiation and response genes. F. RNA expression of ZFP36 and ELAVL1 in non-exposed and 6 hour exposed organoids. Error bars indicate SEM. G. Poly(A)-assay on CXCL8 in exposed EM and DM colon organoids. Sanger sequencing signal for CXCL8 poly(A)-tails; normalized to the signal in the A-channel in the coding-body of CXCL8 mRNA. Error bars indicate SEM, $n=2$.

of cells that activate NF κ B is significantly higher in undifferentiated vs differentiated organoids (68,9% vs 22,7% respectively $p=0,0037$), the pathway is still activated in a substantial number of cells. Therefore, we postulated that the lack of an inflammatory response at the mRNA level as seen in differentiated epithelial cells, cannot be attributed to decreased sensing of antigens and activation of NF κ B-signaling alone.

Next, we addressed whether a decrease in local accessibility and activity of the chromatin of the promoters of the specific response genes could explain the diminished microbial responsiveness in differentiated epithelial cells. To address this, we performed chromatin immune-precipitation of histone modifications that can be found within the active gene promoters - H $_3$ K $_4$ me $_3$ and H $_3$ K $_27$ Ac (Heintzman et al., 2007). Notably, upon exposure to bacterial antigens, both differentiated and undifferentiated organoids showed an equal presence and induction of active chromatin marks within the promoters of response genes (**figure 3C, D**). These findings revealed that despite the fact that upregulation of the response-gene mRNA is limited to undifferentiated organoids (**figure 1J, figure 2**), differentiated cells still activate the NF κ B pathway and remain capable of promoter-activation of inflammatory genes.

The lack of increased mRNA levels of these genes in differentiated cells is therefore likely regulated at a posttranscriptional level. Specifically, a decrease in RNA stability and thus lifespan may explain the absence of inflammatory signals at the mRNA level in differentiated epithelial cells following mechanisms that have previously been established in immune cells. For example, inflammatory responses of macrophages are tightly regulated at a post-transcriptional level through AU-rich element (ARE) mediated mRNA decay (Carballo, 1998; Stoecklin and Anderson, 2006). In this process, RNA binding proteins bind to AREs in the 3'UTR (3'untranslated region) of mRNAs, after which exonucleases degrade the poly(A)-tail of the mRNA molecules. This in turn destabilizes the mRNA molecules and decreases their half-life.

We found that AREs are highly enriched in the response genes (i.e. genes that are upregulated upon bacterial antigen exposure in undifferentiated cells) compared to other gene types (**figure 3E**) (Bakheet et al., 2018). This finding indicates that genes involved in inflammatory responses are particularly prone to post-transcriptional regulation. Previously, it was shown that the Zinc Finger Protein 36 (ZFP36, also known as TTP) can bind to AREs and is a key player in poly(A)-mediated destabilization of mRNA (Lykke-Andersen and Wagner, 2005). We established that ZFP36 is highly expressed in differentiated cells (**figure 3F**). Similarly, we shown that ELAVL1, a molecule involved in mRNA stabilization (Dean et al., 2001), was down regulated in differentiated cells (**figure 3F**).

To further establish the role of posttranscriptional processes in the regulation of unresponsiveness by differentiated epithelial cells, we determined the poly(A)-tail length of the transcripts in our organoid model. Notably, degradation of the poly(A)

tail is an important step in the pathway of mRNA decay and therefore serves as a key determinant of mRNA half-life. We coupled the poly(A) tail length assay with the Sanger sequencing to determine the length of the poly(A)-tail in response genes (CXCL8) and in two housekeeping genes (GAPDH, ACTB). Whereas CXCL8 has a short tail (38 A-nucleotides called upon 6h exposure in DM), poly(A) tails of GAPDH and ACTB were longer (58 and 164 A-nucleotides called respectively) (figure S7.3B). We next compared the poly(A)-tail length in differentiated versus undifferentiated organoids upon exposure. We normalized the intensity of the sequencing signal of the poly(A)-

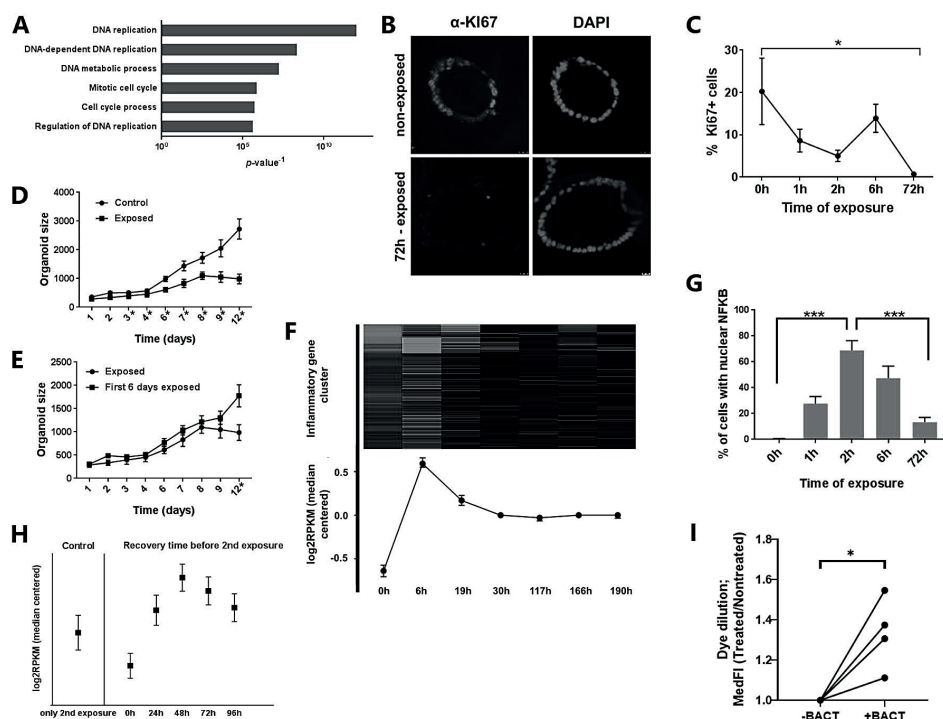


Figure 4 Stem cells develop a hypo-proliferative and hypo-responsive state upon prolonged exposure to bacterial antigens.

A. Pathway analysis of down-regulated genes. Down-regulated genes are enriched for cell cycle pathways. B. Immunohistochemistry with α -KI67 of colon organoids that were exposed for 72h compared to non-exposed organoids. C. Quantification of IHC. The number of KI67+ cells decreases upon prolonged exposure (Kolmogorov-Smirnov $p=0.015$). D. Proliferation assay. Colon organoids were grown in the presence or absence of bacterial lysates for 12 days. Organoid size was significantly lower in exposed organoids. At time points that indicated with * the size was significantly different when comparing the two conditions. (Kolmogorov-Smirnov $p<0.05$) E. Proliferation assay. Colon organoids were grown in the presence of bacterial lysates. After 6 days of exposure, half of the organoids were further cultured without bacterial lysates. At time points that are indicated with * the size was significantly different between the two conditions. (Kolmogorov-Smirnov $p<0.05$) F. RNA-sequencing data of prolonged exposure. Heatmap and expression levels of the response gene cluster. G. Nuclear translocation in undifferentiated organoids upon short and prolonged exposure to bacterial lysates (Mann-Whitney 0h vs 2h $p=0.002$, 2h vs 72h $p=0.0003$). All error bars in this figure show the standard error of the mean. H. RNA-sequencing data of response gene cluster. Organoids were exposed for 72 hours, lysates were removed for 0 to 4 days, subsequently organoids were exposed for 6 hours to test their responsiveness. I. Cell proliferation assay using CellTrace after 6 days of growth (*paired t-test $p = 0.03365$).

tail to the adenine signal in the middle of the mRNA body to correct for differences in expression level. These experiments revealed that CXCL8 mRNA molecules as expressed in undifferentiated colon organoids have, on average, longer poly(A)-tails compared to those that were found in differentiated IECs. Furthermore, approximately half of the captured CXCL8 molecules in differentiated cells completely lacked a poly(A)- signal, whereas in undifferentiated cells >75% were found to have a clear poly(A) signal (**figure 3G**).

These findings establish that hypo-responsiveness to microbial stimulation of differentiated IECs results from both a decrease in NF κ B-mediated signal transduction as well as an increased poly(A)-mediated decay of response gene transcripts.

Stem cells mediate epithelial homeostasis after an encounter with bacterial antigens

How the inflammation affects proliferation of intestinal epithelial cells is not clearly understood. A gene set enrichment analysis (GSEA) by using publicly available transcriptomics datasets derived from chronically inflamed and control intestinal biopsies shows conflicting results between different studies (**figure S7.4C**) – and implicate both hypo and hyperproliferation. Therefore, we further delineated the intricate interplay between IEC microbial sensing, proliferation and differentiation, by a set of experiments focusing on the kinetics of epithelial responsiveness.

To this aim, we analyzed transcriptional profiles of intestinal organoids that were stimulated for a prolonged time. The changes in RNA expression that occur upon exposure to bacterial antigens are not limited to the upregulation of inflammatory genes. As such, we identified a clear decrease in activity of cell cycle pathways within the cluster of downregulated genes (**figure 4A, table S5**). These data suggest that proliferating cells – e.g. stem cells, are likely involved in and affected by the response - as they represent the proliferative compartment of the intestinal epithelium (van Es et al., 2012; Fevr et al., 2007; Korinek et al., 1998). Indeed, we found that the expression of the proliferation marker KI67 diminished over time as determined in undifferentiated organoids upon prolonged exposure to bacterial antigens (**figure 4B, C**).

Next, we performed a proliferation assay under similar conditions, and determined organoid size as a readout. In this case, prolonged exposure of organoids to the bacterial lysate resulted in decreased growth rates (**figure 4D, figure S7.4A, B**). Already upon three days of exposure we found the organoid size to be significantly smaller, pointing at cessation of proliferation under these conditions. We next tested whether this hypo-proliferative state was reversible. As shown in **figure 4E**, when we removed the bacterial lysates, growth rates increased and superseded that of organoids that were continuously exposed. These experiments establish that the hypo-proliferative state upon microbial exposure is reversible.

Upon prolonged exposure of the epithelium to bacterial antigens, the inflammatory gene clusters that were initially upregulated are quickly downregulated (**figure 4F**). This downregulation was coupled with decreased nuclear translocation of NFkB (**figure 4G**). Next, we examined if the responsiveness could be restored to baseline. To this aim we exposed the organoids for 3 days with antigens and subsequently removed the lysate for 1 to 4 days and followed by re-stimulation. Elimination of the stimulus restored the capacity of the epithelium to respond upon re-exposure. This reversibility of the hypo-responsive state occurred already after 1 day of removal and the extent of the response was comparable to the responsiveness at baseline (**figure 4H**)

Finally, to examine if the decreased growth rate was due to a decreased number of cell divisions, we performed a cell proliferation assay using a fluorescent dye (CellTrace) that is stably incorporated into the cells and dilutes upon cell division. The cells exposed to bacterial lysate demonstrated a higher concentration of dye after 6 days of culture (**figure 4I**), suggesting that impaired organoid growth can be, in part, explained by a decreased number of cell divisions in time. We conclude that the intestinal epithelium develops a hypo-responsive and a hypo-proliferative state upon prolonged exposure to bacterial antigens that is reversible upon elimination of the stimulus.

Discussion

In this study, we identified stem cells as the principal epithelial responders to bacterial antigens. Stem cells show, compared to other IECs, a higher rate of signal transduction to the nucleus. Furthermore, our data suggest that mRNA molecules of inflammatory genes are regulated through post-transcriptional processes that include poly(A)-mediated decay in differentiated IECs. Stem cells develop a hypo-responsive and hypo-proliferative state upon prolonged exposure to bacterial antigens, that is reversible through elimination of the stimulus.

We show that stem cell-responses to microbial stimulation are similar along the longitudinal GI-axes and that ISCs only respond to basolateral exposure. This underlines the mucosal paradigm of the intestinal epithelium that does not constantly elicit inflammatory responses to luminal antigens, and at the same time becomes activated when microbial antigens penetrate and reach the cells from its basolateral surface.

The role for the intestinal stem cells as the major responding cell to microbial stimulation is in line with recent reports that established other immune functions for these cells at the mucosal surfaces of the intestines. Specifically, T-cell activation was shown to be dependent on antigen presentation by ISCs via MHC-II expression (Biton et al., 2018). Subsequently, T-cells influence differentiation and self renewal of ISCs by production of pro- and anti-inflammatory cytokines respectively. Our study shows that next to antigen presentation, ISCs are a crucial cell type that mediates local immune responses upon exposure to bacterial antigens.

Human organoid cultures do not contain the full complexity of cell types that is found in the native epithelium (Fujii et al., 2018). By combining organoids that were grown under different culture conditions we were able to profile a heterogeneous population of IECs. However, no Paneth cell (PC) cluster (small intestine) or deep secretory cell cluster (colon) were present in our cultures (as previously reported, Fujii et al., 2018). Previously, we showed that PC degranulation does not directly occur upon stimulation with microbial antigens or bacteria (Farin et al., 2014), thus further supporting that ISCs are the main responders in our experiments.

We show that the hypo-responsive state of differentiated cells is characterized by an active chromatin landscape and that the response is inhibited both up and downstream of transcriptional activity. The transcription profiles (**figure S7.3**) suggest that relevant TLRs are expressed in both differentiated and undifferentiated cells. We show here that the hypo-responsive state that occurs in differentiated IECs is to a large extent mediated at a post-transcriptional level. The role of post-transcriptional regulation during inflammatory responses was previously established for immune cells (Carpenter et al., 2014). Support for the (clinical) relevance of these mechanism is provided by the discovery of the genetic association of ZFP36 family members (ZFP36L1, ZFP36L2) to Inflammatory Bowel Disease (IBD) (Jostins et al., 2012; Meddens et al., 2016). Specifically, ZFP36 was shown to be a key player in mediating protective, anti-inflammatory processes in a murine colitis model (Joe et al., 2014).

Finally, the expression of the response genes and ZFP36 are under the regulation of NF κ B and ZFP36 was shown to inhibit nuclear translocation of NF κ B (Chen et al., 2013; Schichl et al., 2009). Our results suggest that ZFP36 and its family members might play a pivotal role in the regulation of IEC responsiveness. We anticipate that different checkpoints within these post-transcriptional regulation pathways may represent some novel targets for the treatment of IBD, and other diseases that are associated with unopposed epithelial inflammatory responses.

This study emphasizes that the intestinal epithelial response to bacterial antigens is highly regulated at various levels. Whereas responsiveness may be similar when determined along the longitudinal intestinal axis, we show strong differences in responses that are associated with the state of differentiation and cellular side of stimulation. Finally, we identify the intestinal stem cells as the key epithelial responder upon bacterial stimulation.

Materials and Methods

Medical and Ethical guidelines

Biopsies were obtained by ileo-colonoscopies and gastroscopies that were performed as part of standard diagnostic procedures. Human Material Approval for this study was obtained by the Ethics Committee (Medisch Ethische Toetsings Commissie, METC) of the University Medical Center Utrecht (www.umcutrecht.nl/METC).

Organoid culturing

Colon and ileum biopsies were obtained by colonoscopy, duodenum biopsies were obtained by flexible gastroduodenoscopy. The biopsies were macroscopically and pathologically normal. Crypt isolation and culture of human intestinal cells from biopsies were performed as was described previously. (Dekkers et al., 2013; Sato et al., 2011) Composition of culture media for colon, ileum and duodenum were the same. The organoids were maintained long-term in expansion medium (EM), which is composed in Advanced medium/F12 (Gibco) containing RSPO1, noggin, EGF, A83-01, nicotinamide, SB202190, and WNT3A. For induction of differentiation, cultures were maintained for 5-7 days in differentiation medium (DM), which is EM without nicotinamide, SB202190, and WNT3A. We used conditioned media for RSPO1 (stably transfected RSPO1 HEK293T cells were kindly provided by Dr. C. J. Kuo, Department of Medicine, Stanford, CA), noggin, and WNT3A. The medium was changed every 2-3 days and organoids were passaged 1:4 approximately every 10 days (detailed media composition: **table S6**).

Matrigel-embedded organoids (3D) were cultured in 70% matrigel (BD Biosciences) diluted using growth factor lacking medium (GF-) consisted of Advanced DMEM supplemented with penicillin/streptomycin (GIBCO), 1M HEPES (GIBCO) and Glutamax 100x (GIBCO). Primary intestinal organoids were cultured at 37°C, in 5% CO₂. Intestinal epithelial monolayers (2D) were prepared as described previously. (Moon et al., 2014) Briefly, transwells (Corning Costar, Tewksbury, MA, USA) were coated with matrigel (1:40 in PBS+ with Ca/Mg, Sigma-Aldrich) for 1 hour at RT after which 2.5×10^5 single cells were seeded on a transwell insert in the corresponding 24 well plate. 100µL and 600µL medium was used in apical and basolateral compartment respectively. Monolayers were grown until confluent (determined by microscopy and trans-epithelial electrical resistance (TEER) measurement). Primary intestinal monolayers were cultured at 37°C, in 5% CO₂.

Exposure to bacterial antigens

Bacterial lysate was prepared from *E. coli* HST-08 Stellar competent cells. Bacteria were heat-inactivated for 20 min at 75°C. Subsequently, samples were sonicated (Covaris ultrasonicator, Duty cycle: 20%; Intensity: 10; Cycles burst: 500; time: 30 sec) in 13x65mm glass vial (Covaris), centrifuged at 10,000g, 30 min at 4°C and filter-sterilized. The amount of bacterial lysate used for organoid exposure was titrated (1µL to 20µL lysate per 500µL medium to determine the concentration that elicits half of the maximum CXCL-8 response on mRNA in 3D cultures. Same bacterial lysate concentrations were used in 3D, and 2D apical and basolateral exposure. In experiments in which cultures were exposed for multiple days, new bacterial lysate was added each time the medium was refreshed (approx. every 2 days).

RNA isolation and qPCR

RNA was isolated with TRIzol® LS (Ambion, cat. no. 10296-028), according to the manufacturers protocol. cDNA was synthesized by performing reverse-transcription (Invitrogen, Carlsbad, CA or iScript, Biorad, Hercules, CA). Messenger RNA (mRNA) abundances were determined by real-time PCR using validated primer pairs (Supplementary data) and SYBR Green (Bio-Rad, Hercules, CA). ACTB mRNA abundance was used for normalization. qPCR primers used: Lgr5 forward GAATCCCCTGCCAGTCTC, Lgr5 reverse ATTGAAGGCTTCGCAAATTCT, B-actin forward TGGCACCCAGCACAAATGAA, B-actin reverse CTAAGTCATAGTCCGCCTAGAAGCA, TNFα forward CGCTCCCCAAGAAGAC, TNFα reverse GGTTCGAGAAGATGATCTGA, NFκB1A forward GCAAATCCTGACCTGGTGT, NFκB1A reverse GCTCGTCCTCTGTGAACTCC, CXCL8 forward GGCACAAACTTTCAGAGACAG, CXCL8 reverse ACACAGAGCTGCAGAAATCAG.

Multiplex immunoassay

Medium from organoids was collected at the moment of harvesting (after 6h exposure) and was stored at -80°C. CXCL8 concentrations were measured using the Luminex technology as previously described (de Jager et al., 2005).

RNA seq

RNA was isolated with TRIzol® LS (Ambion, cat. no. 10296-028), according to the manufacturers protocol. Libraries were generated using NEXTflex™ Rapid RNA-seq Kit (Bio Scientific) and sequenced by the Nextseq500 platform (Illumina) to produce 75 bp single-end reads through the Utrecht DNA sequencing facility. Reads were aligned to the human reference genome GRCh37 using STAR. Differentially expressed genes in the transcriptome data were identified using the DESeq2 package with standard settings.

Explants and Crypts

Colon biopsies were washed and cultured in EM (described above) for 24h after which explants were stimulated with bacterial lysate for 6h. Next, RNA was isolated and sequenced as described above. Inflamed biopsies were taken from the sigmoid of ulcerative colitis patients, uninfamed biopsies were taken from colon ascending and were macroscopically normal. Crypts were isolated from the biopsies and RNA was sequenced.

Pathway analyses and TF motif enrichment

Pathway analyses and TF motif enrichments were performed on the genes from cluster 1 (table S1) (figure 1G) or all genes in exposed vs non exposed comparison with a negative fold change and $p < 0.05$ (figure 4A) were performed using the Toppfun platform. (Chen et al., 2009) Biological processes and binding motifs with highest enrichments (lowest p-values) are depicted.

Immunohistochemistry

Organoids were collected by carefully disrupting the matrigel and sequentially elimination of the matrigel through centrifugation (5 min, 2000rpm). Samples were fixed in 4% formaldehyde and embedded in 200µL 2% agarose in dH₂O before embedding the samples in paraffin. 5µm thick slides were deparaffinized and heat-mediated antigen retrieval was performed for 20 min in citrate antigen retrieval buffer pH6 (Sigma-Aldrich, C9999). Slides were blocked for 30 min in 5% BSA at RT and incubated ON at 4°C with primary antibodies (mouse α -NF κ B p65 L8F6 1:50 (CST 6956S, Cell signaling Technology) and rabbit α -Ki67 1:25 (AB16667, Abcam) in 5% BSA-PBS). Slides were incubated with secondary antibodies Alexa 488 donkey-anti-mouse (1:400 (A21202, Thermo Fisher Scientific)) and Alexa 647 donkey-anti-rabbit (1:400 (A31573, Thermo Fisher Scientific)) for 1 hour at rt. Images were captured with a 63x objective on a Leica TCS SP8 X confocal microscope.

NFKB-inhibition

Duodenum organoids were grown from single cells on EM medium. 8 Days after seeding, NFKB-inhibitors were added to the organoids: 5µM IMD 0354 (Abcam, ab144823) or 10µM TPCA 1 (Abcam, ab145522). 12 Hours after addition of inhibitors, organoids were exposed to bacterial lysate for 6h, after which RNA was isolated.

Single cell seq

Colon and ileum derived organoids were cultured for 10-14 days from single cells in 3 conditions: 1) on EM during the whole experiment, 2) EM was changed to DM for the last 24 hours before harvesting, 3) EM was changed for the last 4 days before harvesting. For each condition half of the organoids were exposed for 6 hours with bacterial lysates before harvesting. Next, cells were trypsinized and FACS-sorted using PI (Thermo Fisher Scientific - P3566) to eliminate dead cells into 384-well pre-indexed plates per condition. The plates were processed by Single Cell Discoveries as described previously using SORT-seq technology. (Muraro et al., 2016) The data were analyzed using the Seurat V2.3.4. after excluding mitochondrial and ribosomal gene and a set of unreliably mapped genes (UGDH-AS1, PGM2P2, LOC100131257, MALAT1, KCNQ1OT1, PGM5P2, MAB21L3, EEF1A1) with these parameters: CreateSeuratObject(min.cells = 3, min.genes = 1500), NormalizeData(normalization.method = "LogNormalize", scale.factor = 10000), FindVariableGenes(mean.function = ExpMean, dispersion.function = LogVMR, x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5), FindClusters(reduction.type = "pca", dims.use = 1:12, resolution = 0.6).

ChIP seq

Chromatin immunoprecipitation was performed using the MAGnify ChIP kit (Invitrogen, Carlsbad, CA) according to the manufacturers recommendations. μL α -acetylated histone 3 lysine 27 (H3K27ac) (ab4729; Abcam) or μL α -trimethylated lysine 4 (H3K4me3) (#39159; Active Motif) was used per immunoprecipitation. Captured DNA was purified using ChIP DNA Clean & Concentrator kit (Zymo Research). Libraries were prepared using the NEXTflex™ Rapid DNA Sequencing Kit (Bioo Scientific). Samples were PCR amplified, checked for the proper size range and for the absence of adaptor dimers on a 2% agarose gel, and barcoded libraries were sequenced 75 bp single-end on Illumina NextSeq500 sequencer. Sequencing reads were mapped against the reference genome (hg19 assembly, NCBI37) using the BWA package (mem -t 7 -c 100 -M -R)42. Multiple reads mapping to the same location and strand were collapsed to single read and used for peak-calling. Peaks/regions were called using Cisgenome 2.043 (-e 150 -maxgap 200 -minlen 200).

ARE-enrichment

ARE-enrichment was calculated using the ARED-plus database.(Bakheet et al., 2018) The percentage of response genes (genes upregulated upon 6h exposure to bacterial lysate), stem genes (genes upregulated in EM, table S7) or differentiation genes (genes upregulated in DM, table S7) that have at least one AU-rich element encoded in the 3'UTR was calculated per gene set.

Poly(A)-assay

Colon organoids were cultured from single cells for 10-14 days, either on EM or organoids were differentiated for the 5-7 days on DM. All organoids were exposed to bacterial lysates for 6h before RNA isolation (described above). GI-tailing and reverse transcription were done using the Poly(A) tail length assay kit (ThermoFisher Scientific, # 764551KT) and according to the manufacturers protocol. Next, amplification of individual genes was done with a gene specific forward primer (CXCL8 CTTGTTCATTGCCAGCTGTGT, GAPDH CAACGAATTTGGCTACAGCA, B-actin ATCCTAAAAGCCACCCCACT) and a GI-tail reverse primer (Poly(A) tail length assay kit) for 35 cycles in Platinum™ PCR SuperMix High Fidelity (Invitrogen, # 12532016). Samples were purified using the QIAquick PCR Purification Kit (Qiagen, # 28104). Forward primers that were used for amplification were also used for Sanger Sequencing.

Relative number of molecules with a adenine base at any position in the poly(A) tail was calculated against the Sanger-signal in the A-channel in the mRNA-body. The average signal of the called A-peaks on position 100 to 200 of the molecule was used to account for expression level differences (AVsig). Adenine signal was divided by AVsig at each position along the poly(A)-tail.

Growth curves

500 Single cells were seeded in 5 μL matrigel droplets in a 96-well plate (Costar). Images were taken daily, using the Evos microscope (Thermo Scientific) and were analyzed using ImageJ and Excel.

Organoid proliferation measurement

Ileum organoids were seeded single cells stained or unstained 5-7 days prior to FACS analysis. Stained cells were stained before seeding using 20 μM CellTrace Violet (Invitrogen), by incubation for 10 minutes at 37°C. After 2 days of seeding half of the organoids started bacterial lysate stimulation. On the day of FACS analysis, the organoids were harvested as single cells, and wash in PBS with Ca²⁺ and Mg²⁺ supplemented with 2mM EDTA and 0.5% BSA. A life/dead staining using PI was included. Unstained organoids were harvested and half of these were stained as on day 1 before FACS. FACS was performed using a BD FACSCanto II (BD-Biosciences). BD FACSDiva software was used for gating, cells were gated for PI staining and Cell Trace Violet staining. Data was analyzed using FlowJo_V10 software.

GSEA

GSEA(Subramanian et al., 2005) was performed using gene expression datasets from intestinal biopsies obtained from ulcerative colitis patients (datasets available at GSE11223, GSE75214,

GSE9452, GSE38713, GSE6731) and exposed and non-exposed colon organoids. In each dataset, expression in inflamed colon biopsies was compared to expression in healthy colon biopsies. Significance of the enrichment was calculated based on 1000 cycles of permutations.

References

- Abreu, M.T. (2010). Toll-like receptor signalling in the intestinal epithelium: how bacterial recognition shapes intestinal function. *Nat. Rev. Immunol.* 10, 131–144.
- Anderle, P., Sengstag, T., Mutch, D.M., Rumbo, M., Praz, V., Mansourian, R., Delorenzi, M., Williamson, G., Roberts, M.-A., Stenberg, P., et al. (2005). Changes in the transcriptional profile of transporters in the intestine along the anterior-posterior and crypt-villus axes. *BMC Genomics* 6, 69.
- Avitzur, Y., Guo, C., Mastropaolo, L.A., Bahrami, E., Chen, H., Zhao, Z., Elkadri, A., Dhillon, S., Murchie, R., Fattouh, R., et al. (2014). Mutations in tetratricopeptide repeat domain 7A result in a severe form of very early onset inflammatory bowel disease. *Gastroenterology* 146, 1028–1039.
- Bakheet, T., Hitti, E., and Khabar, K.S.A. (2018). ARED-Plus: an updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs. *46*, 2017–2019.
- Bigorogne, A.E., Farin, H.F., Lemoine, R., Mahlaoui, N., Lambert, N., Gil, M., Schulz, A., Philippet, P., Schlessner, P., Abrahamsen, T.G., et al. (2014). TTC7A mutations disrupt intestinal epithelial apical-basal polarity. *J. Clin. Invest.* 124, 328–337.
- Biton, M., Haber, A.L., Rogel, N., Yilmaz, O.H., Regev, A., Xavier, R.J., Biton, M., Haber, A.L., Rogel, N., Burgin, G., et al. (2018). Article T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. *Cell* 1–14.
- Carballo, E. (1998). Feedback Inhibition of Macrophage Tumor Necrosis Factor- Production by Tristetraprolin. *Science* (80-.). 281, 1001–1005.
- Carpenter, S., Ricci, E.P., Mercier, B.C., Moore, M.J., and Fitzgerald, K.A. (2014). Post-transcriptional regulation of gene expression in innate immunity. *Nat. Rev. Immunol.* 14, 361–376.
- Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, 305–311.
- Chen, Y.L., Jiang, Y.W., Su, Y.L., Lee, S.C., Chang, M.S., and Chang, C.J. (2013). Transcriptional regulation of tristetraprolin by NF- κ B signaling in LPS-stimulated macrophages. *Mol. Biol. Rep.* 40, 2867–2877.
- Damen, G.M., Hol, J., de Ruiter, L., Bouquet, J., Sinaasappel, M., van der Woude, J., Laman, J.D., Hop, W.C.J., Büller, H. a. Escher, J.C., et al. (2006). Chemokine production by buccal epithelium as a distinctive feature of pediatric Crohn disease. *J. Pediatr. Gastroenterol. Nutr.* 42, 142–149.
- Dean, J.L.E., Wait, R., Mahtani, K.R., Sully, G., Clark, A.R., and Saklatvala, J. (2001). The 3' Untranslated region of tumor necrosis factor alpha mRNA is a target of the mRNA-stabilizing factor HuR. *Mol. Cell. Biol.* 21, 721–730.
- Dekkers, J.F., Wiegerinck, C.L., de Jonge, H.R., Bronsveld, I., Janssens, H.M., de Winter-de Groot, K.M., Brandsma, A.M., de Jong, N.W.M., Bijvelds, M.J.C., Scholte, B.J., et al. (2013). A functional CFTR assay using primary cystic fibrosis intestinal organoids. *Nat. Med.* 19, 939–945.
- van Es, J.H., Haegerbarth, A., Kujala, P., Itzkovitz, S., Koo, B.-K., Boj, S.F., Korving, J., van den Born, M., van Oudenaarden, A., Robin, S., et al. (2012). A critical role for the Wnt Effector Tcf4 in adult intestinal homeostatic self-renewal. *Mol. Cell. Biol.* 32, 1918–1927.
- Farin, H.F., Karthaus, W.R., Kujala, P., Rakhshandehroo, M., Schwank, G., Vries, R.G.J., Kalkhoven, E., Nieuwenhuis, E.E.S., and Clevers, H. (2014). Paneth cell extrusion and release of antimicrobial products is directly controlled by immune cell-derived IFN- γ . *J. Exp. Med.* 211, 1393–1405.
- Fevr, T., Robine, S., Louvard, D., and Huelsken, J. (2007). Wnt / -Catenin Is Essential for Intestinal Homeostasis and Maintenance of Intestinal Stem Cells. *†*. 27, 7551–7559.
- van der Flier, L.G., and Clevers, H. (2009). Stem Cells, Self-Renewal, and Differentiation in the Intestinal Epithelium. *Annu. Rev. Physiol.* 71, 241–260.
- Fujii, M., Matano, M., Toshimitsu, K., Takano, A., Mikami, Y., Nishikori, S., Sugimoto, S., and Sato, T. (2018). Human Intestinal Organoids Maintain Self-Renewal Capacity and Cellular Diversity in Niche-Inspired Culture Condition. *Cell Stem Cell* 23, 787–793.e6.
- Gassler, N., Rohr, C., Schneider, A., Kartenbeck, J., Bach, A., Obermüller, N., Otto, H.F., and Autschbach, F. (2001). Inflammatory bowel disease is associated with changes of enterocytic junctions. *Am. J. Physiol. - Gastrointest. Liver Physiol.* 281.
- Haber, A.L., Biton, M., Rogel, N., Herbst, R.H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T.M., Howitt, M.R., Katz, Y., et al. (2017). A single-cell survey of the small intestinal epithelium. *Nature* 551, 333–339.
- Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K. a. et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and

- enhancers in the human genome. *Nat. Genet.* 39, 311–318.
- de Jager, W., Prakken, B.J., Bijlsma, J.W.J., Kuis, W., and Rijkers, G.T. (2005). Improved multiplex immunoassay performance in human plasma and synovial fluid following removal of interfering heterophilic antibodies. *J. Immunol. Methods* 300, 124–135.
- Jain, U., Lai, C.-W., Xiong, S., Goodwin, V.M., Lu, Q., Muegge, B.D., Christophi, G.P., VanDussen, K.L., Cummings, B.P., Young, E., et al. (2018). Temporal Regulation of the Bacterial Metabolite Deoxycholate during Colonic Repair Is Critical for Crypt Regeneration. *Cell Host Microbe* 24, 353–363.e5.
- Joe, Y., Uddin, M.J., Zheng, M., Kim, H.J., Chen, Y., Yoon, N.A., Cho, G.J., Park, J.W., and Chung, H.T. (2014). Tristetraprolin mediates anti-inflammatory effect of carbon monoxide against DSS-induced colitis. *PLoS One* 9, 3–8.
- Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–124.
- Kaiko, G.E., Ryu, S.H., Koues, O.I., Collins, P.L., Solnica-Krezel, L., Pearce, E.J., Pearce, E.L., Oltz, E.M., and Stappenbeck, T.S. (2016). The Colonic Crypt Protects Stem Cells from Microbiota-Derived Metabolites. *Cell* 167, 1137.
- Korinek, V., Barker, N., Moerer, P., Donselaar, E. Van, Huls, G., Peters, P.J., and Clevers, H. (1998). Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Development* 125, 379–383.
- Lykke-Andersen, J., and Wagner, E. (2005). Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes Dev.* 19, 351–361.
- Meddens, C.A., Harakalova, M., Dungen, N., Foroughi Asl, H., Hijma, H.J.H., Cuppen, E.P.J.G.E., Börkegen, J., Asselbergs, F.W.F., Nieuwenhuis, E.E.E.S.S., Mokry, M., et al. (2016). Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease. *Genome Biol.*
- Menckeburg, C.L., Hol, J., Simons-Oosterhuis, Y., Raatgeep, H.C., de Ruitter, L.F., Lindenbergh-Kortleve, D.J., Korteland-van Male, a. M., El Aidy, S., van Lierop, P.P.E., Kleerebezem, M., et al. (2014). Human buccal epithelium acquires microbial hyporesponsiveness at birth, a role for secretory leukocyte protease inhibitor. *Gut* 64, 884–893.
- Middendorp, S., Schneeberger, K., Wiegerinck, C.L., Mokry, M., Akkerman, R.D.L., Van Wijngaarden, S., Clevers, H., and Nieuwenhuis, E.E.S. (2014). Adult stem cells in the small intestine are intrinsically programmed with their location-specific function. *Stem Cells.*
- Moon, C., Vandussen, K.L., Miyoshi, H., and Stappenbeck, T.S. (2014). Development of a primary mouse intestinal epithelial cell monolayer culture system to evaluate factors that modulate IgA transcytosis. *Mucosal Immunol.* 7, 818–828.
- Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., et al. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* 3, 385–394.e3.
- Nusse, Y.M., Savage, A.K., Marangoni, P., Rosendahl-Huber, A.K.M., Landman, T.A., De Sauvage, F.J., Locksley, R.M., and Klein, O.D. (2018). Parasitic helminths induce fetal-like reversion in the intestinal stem cell niche. *Nature* 559, 109–113.
- Price, A.E., Shamardani, K., Lugo, K.A., Deguine, J., Roberts, A.W., Lee, B.L., and Barton, G.M. (2018). A Map of Toll-like Receptor Expression in the Intestinal Epithelium Reveals Distinct Spatial, Cell Type-Specific, and Temporal Patterns. *Immunity* 49, 560–575.e6.
- Rakoff-Nahoum, S., Pglino, J., Eslami-Varzaneh, F., Edberg, S., and Medzhitov, R. (2004). Recognition of comensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* 118, 229–241.
- Rescigno, M., and Nieuwenhuis, E.E. (2007). The role of altered microbial signaling via mutant NODs in intestinal inflammation. *Curr Opin Gastroenterol* 23, 21–26.
- Sato, T., Stange, D.E., Ferrante, M., Vries, R.G.J., Van Es, J.H., Van den Brink, S., Van Houdt, W.J., Pronk, A., Van Gorp, J., Siersema, P.D., et al. (2011). Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* 141, 1762–1772.
- Schichl, Y.M., Resch, U., Hofer-Warbinek, R., and de Martin, R. (2009). Tristetraprolin impairs NF- κ B/p65 nuclear translocation. *J. Biol. Chem.* 284, 29571–29581.
- Schmitt, M., Schewe, M., Sacchetti, A., Feijtel, D., van de Geer, W.S., Teeuwssen, M., Sleddens, H.F., Joosten, R., van Royen, M.E., van de Werken, H.J.G., et al. (2018). Paneth Cells Respond to Inflammation and Contribute to Tissue Regeneration by Acquiring Stem-like Features through SCF/c-Kit Signaling. *Cell Rep.* 24, 2312–2328.e7.
- Stoecklin, G., and Anderson, P. (2006). Posttranscriptional Mechanisms Regulating the Inflammatory Response. *Adv. Immunol.* 89, 1–37.
- Stumpo, D.J., Lai, W.S., and Blakeshear, P.J. (2010). Inflammation: Cytokines and RNA-based regulation. *Wiley Interdiscip. Rev. RNA* 1, 60–80.

- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550.
- Vetrano, S., Rescigno, M., Cera, M.R., Correale, C., Rumio, C., Doni, A., Fantini, M., Sturm, A., Borroni, E., Repici, A., et al. (2008). Unique role of junctional adhesion molecule-a in maintaining mucosal homeostasis in inflammatory bowel disease. *Gastroenterology* 135, 173–184.
- Yin, X., Farin, H.F., van Es, J.H., Clevers, H., Langer, R., and Karp, J.M. (2014). Niche-independent high-purity cultures of Lgr5 + intestinal stem cells and their progeny. *Nat. Methods* 11, 106–112.



Discussion

8

This thesis presents an overview of the studies that we have done on multiple aspects of complex genetic diseases, ranging from chromatin conformation to intestinal stem cell biology. During this process we have faced challenges, set boundary conditions, encountered technical limitations and learned from the continuous scientific developments in our field. In this final chapter, I will discuss the implications and limitations of our findings regarding the choices we have made from a scientific and clinical perspective.

Complicating complex genetics

The multitude of genetic associations that have been found in the past, revealed that the genetic background of complex genetic diseases is heterogeneous and consists of numerous SNPs that each account for a minor increase in risk. The identification of variants that affect the amino acid sequence of a gene has been limited, which leaves us with many associated genetic loci and tens to hundreds of candidate genes per disease.¹ Translation of these findings into the pathological mechanisms and to the patients' benefit is the major challenge in this field. In this thesis, instead of addressing this challenge, we added another layer of complexity that resulted in the association of hundreds of extra putative candidate genes for IBD, CKD and CAD. Although this may seem counter intuitive and for some even counter productive, it has brought us closer to the underlying pathological mechanisms and to novel treatment potential.

First, in chapter 3, we show that the SNPs that have been associated to IBD are enriched in active enhancer elements. This supports the hypothesis that many genetic variants that contribute to the pathogenesis of IBD alter the function of regulatory elements, rather than protein coding sequences. The involvement of DRE in the pathogenesis of complex genetic diseases is further established in chapters 4 and 6, by the finding that associated variants can alter the activity of regulatory elements. When aiming at developing novel treatments or unravelling pathogeneses, this paradigm that involves more than half of the associated variants, cannot be ignored although it inevitably leads to a more complicated model of the genetic background of these diseases.² Albert Einstein supposedly once said: *'Everything should be made as simple as possible, not simpler'*.

Second, a complete overview of the genes that are involved in complex genetic diseases, provides a good basis for the identification of targetable pathways and processes that underlie the pathogeneses. The 4C-seq analyses on CKD and CAD lead to the identification of novel pathways. The pathway analysis in CAD shows that oxidative stress is a shared process between many candidate genes and the mevalonate pathway was identified in CKD. The main pathways that are shared among the IBD candidate genes were already previously identified based on classical candidate gene approaches. However, chromatin conformation capture pinpoints the cell types in which these pathways are likely dysregulated, thereby unravelling an important part of the pathogenesis. Next to pathways, we studied the transcription factors that regulate the candidate genes, in order to identify shared 'key' regulators. Targeting of these regulators could be a strategy that

leads to targeting many dysregulated genes at once. As for the pathways, 4C-seq enabled the identification of cell type specific key regulators, which is valuable when developing drugs and delineating pathogenic mechanisms. Identification of key regulators and pathways shows that integration of the growing knowledge on the genetics of complex genetic diseases can lead to transcending findings and strategies.

Third, the newly identified candidate genes provide novel insight into the genetics of diseases that have both monogenic and complex genetic backgrounds. Many complex genetic diseases share phenotypic characteristics with monogenic diseases that usually have an earlier onset and a more severe disease course.^{3,4} In general, variants that are associated to complex genetic diseases are common and often do not affect the protein coding sequences. In contrast, variants that cause monogenic diseases are usually rare and deteriorate protein function.⁵ The results in chapters 4,5 and 6 show that many newly identified candidate genes are also known as genes that cause monogenic diseases. This suggests that the same genes may be affected in these diseases, although through a different genetic mechanism. For example, mutations in *IL10RA* are known to abrogate *IL10* signaling in monogenic early onset IBD. Our results now suggest that *IL10RA* expression is dysregulated in patients with complex genetic forms of IBD. The 4C-seq analyses performed in this thesis support a model in which monogenic and complex genetic diseases together form a spectrum of diseases. This paradigm can be used for the interpretation of diagnostic whole exome sequencing results, as well as for the prioritization of candidate genes and pathways in complex genetic diseases.

Alternatives for 4C-seq

We have used circular chromatin conformation capture-sequencing (4C-seq) to systematically analyze chromatin interactions at single regulatory regions. The incorporation of such analyses in the follow-up of a GWAS calls for a database that covers all chromatin interactions in any cell type or state. 4C-seq is a technically challenging method that is limited by the need to individually assay each region of interest. We have partially overcome this by applying a multiplexing approach, which enabled us to assay dozens of regions at once.⁶ However, this is not applicable to the scale of the whole human genome and cannot be used to assay all nuclear interactions. To enable the assessment of the effects of GWAS-variants on regulatory elements alternative strategies need to be developed. In chapter 4, we evaluated whether the presence of insulator elements could be used to predict chromatin interactions *in silico*. However, we found that the majority of interactions bypassed multiple insulators and therefore this approach cannot be applied.

Another *in silico* alternative that is often used for the annotation of regulatory effects to GWAS associations are expression quantitative trait loci (eQTLs). Although eQTLs do not address whether a correlation between expression levels and a SNP is directly causal, this method has a high specificity, is widely used and has resulted in the identification of

many variants with regulatory phenotypes. As a SNP affects expression levels only under specific conditions, eQTLs are highly context dependent.⁷ Therefore, the chance to detect an eQTL is low and eQTLs have a high false negative discovery rate. Whereas an eQTL might only be detected under rare conditions (and need genotyping and transcriptional profiling of large cohorts), chromatin conformation is more stable and the interaction between a gene and a regulatory variant can be detected even under conditions where the variant does not cause a phenotype. Therefore, chromatin conformation capture (3C) based techniques can be beneficial, because they enable to study the role of DRE in many cell types without the need for genotyping of large cohorts and profiling of transcriptomes under rare conditions.

Although the high throughput variants of chromatin conformation capture (3C) do not have the same resolution as 4C-seq⁸, they enable the generation of widely applicable datasets. Publically available chromatin conformation datasets are systematically generated with high throughput techniques like HiC and 5C⁹ and are starting to be used for interpretation of GWAS results.¹⁰

DNA regulatory elements; identification and definitions

In this thesis, we aimed to identify candidate genes based on the effect of associated variants on DNA regulatory elements (DRE). Although there are many different types of DRE, we chose to focus on enhancer elements, because they are among the most studied regulatory elements, they provide insight into cell type specific mechanisms and their state of activity can be identified.¹¹ This has resulted in the identification of many novel candidate genes, pathways, key regulators and cell type specific pathological mechanisms. However, by omitting other types of DRE, we have probably missed candidate genes and mechanisms. We have not studied the effect of variants on promoters, insulators and silencers, nor of alternative splice sites and untranslated regions (UTRs).¹² As promoters, splice sites and UTRs are found close to or within the gene body, the genes that are affected by variants in these elements will have been detected by classical candidate gene approaches. This is due to the proximity between the variant and the candidate genes on the linear genome. The genes that are affected by variants in insulators and silencers will not be pick-up in classical approaches.

The definition of active enhancers that was used in this thesis is based on one characteristic of enhancers, namely the presence of activating histone modifications H3K27Ac and H3K4me1 at flanking nucleosomes.¹³ However, an active enhancer-sequence itself is known to be devoid of nucleosomes and therefore accessible to DNase¹⁴; bind transcription factors and the flanking nucleosomes are enriched for non canonical histone subunits (H2A.Z and H3.3)¹⁵. None of these characteristics are completely specific and sensitive for the identification of active enhancers.¹¹ Transcription factor binding can co-occur with the binding of repressive transcription factors, that also bind to nucleosome free DNA.¹⁶ Non canonical histone subunits are found at transcriptionally active chromatin,

that is not limited to active enhancers.¹⁷ The use of activating histone modifications could result in the false positive identification of poised enhancers (that have bivalent active/repressive modifications)¹⁸ and a false negative identification of latent enhancers (that are not labelled by activating marks)¹⁹. However, for the identification of candidate genes, the identified enhancers have to interact with promoters, which is another key characteristic of active enhancers, that reduces false positive results.

For further improvement of enhancer identification, the combination of characteristics can be sourced. However, also the dynamics of these factors and the synergy between enhancers plays an important role and could be valuable addition. A major limiting factor of a straightforward identification of DRE is that the function that these characteristics play in activating transcription remains elusive. It is therefore paramount that the biology of DRE functioning is further elucidated. This will not only improve their identification, but also provide novel therapeutic targets and strategies.

Intestinal stem cells in the pathogenesis of IBD

In chapter 7, we identify intestinal stem cells as a major source of the intestinal response to microbial antigens. We show that responsiveness, proliferation and differentiation are intertwined processes that are tightly regulated to maintain epithelial homeostasis. We and others have previously shown that the intestinal epithelium plays an important role in the pathogenesis of IBD.²⁰⁻²² As the processes that we describe in chapter 7 are involved in responsiveness and epithelial homeostasis, they might be the pathways that are affected in IBD.

We show that in healthy intestinal organoids, stem cells are the major responsive cell type in contrast to differentiated IECs. Our data suggest that ARE-mediate decay is activated in differentiated cells, which results in rapid degradation of inflammatory mRNA molecules. The native intestinal epithelium is known to be intrinsically hyper-responsive in IBD patients even in absence of lesions.^{20,23} Through GWASs and 4C-seq approaches, we and others identified genes involved in post-transcriptional regulation as IBD-candidate genes.^{24,25} Furthermore, it has been shown that affecting post-transcriptional regulation in mice results in intestinal inflammation.^{26,27} Therefore, inefficient degradation of inflammatory mRNA molecules in differentiated IECs, might contribute to the inflammatory phenotype in IBD. However, Parikh et al. recently performed single cell RNA-seq of intestinal biopsies of inflamed epithelia in IBD patients and showed that specifically undifferentiated IECs upregulate CXCL1, 2 and 3.²³ These results suggest that also in IBD, the stem cells might be a major source of cytokines. This implies that the hypo-responsive state of stem cells that occurred upon prolonged exposure of healthy organoids, might be affected in IBD. It would therefore be valuable to study whether ARE-mediated decay is contributing to the hypo-responsive state of stem cells as well as to the general hypo-responsiveness that was observed in differentiated IECs.

Together with the upregulation of inflammatory genes, we found that proliferation is downregulated upon exposure of ISCs to microbial antigens. This might be a mechanism through which stem cells can dedicate their activity to immune cell recruitment and prevent the generation of hyper-responsive daughter cells. We have performed GSEAs on many IBD datasets to test whether proliferation of IECs is affected in IBD. These analyses showed hyper-proliferation in some datasets and hypo-proliferation in others.²⁸⁻³² Aberrant proliferation will affect intestinal homeostasis irrespective of the direction of the defect (i.e. up or down). When the induction of the hypo-proliferative state is affected, hyper-proliferation might occur and could result in hyperplasia as is often seen in IBD.³³ As we show that the proliferating IECs are responsive, this might result in increased numbers of cytokine producing cells in the intestines of IBD patients. On the other hand, prolonged hypo-proliferation could be due to a continuous responsive state. This might contribute to impaired regeneration of the intestinal epithelium. WNT-signaling is known to be affected in ISCs of IBD patients through the downregulation of HB-EGF²³, which might result in hypo-proliferation. Furthermore, PGE₂, that is known to inhibit crypt regeneration, is locally increased in inflamed intestinal mucosa.³⁴

Altogether, these results suggest that aberrant intestinal kinetics of responsiveness might contribute to the pathogenesis of IBD. To further address which mechanisms are affected in IBD, organoids derived from IBD patients could be used. It has previously been shown that IBD-organoids can be efficiently grown and differentiated.^{35,36} Exposing IBD organoids from a substantial group of donors to bacterial lysates can reveal how post-transcriptional regulation, proliferation, differentiation and responsiveness are affected in IBD.

The future pipe lines

The novel findings on the genetic background of IBD, CAD and CKD need to be further established. The key pathways and upstream regulators should be validated in the relevant cells types of affected individuals. In vitro characterization of the phenotypes that are related to these main 'switches' could provide a platform to test the efficacy of compounds that target these pathways and regulators. This way, the elucidation of the pathological mechanisms can move side-by-side with the development of diagnostic and therapeutic strategies. We have identified multiple transcription factors that turned out to be key regulators of a large number of candidate genes and were identified as a candidate genes themselves. This suggests that the expression of these TFs is dysregulated and could thereby amplify the effect of the genetic variants that affect the regulated genes. These genes could form a lead of departure in the prioritization of functional testing of pathways and regulators.

The experimental set up that was used to unravel epithelial responsiveness and homeostasis provides a design that can be used to study the involvement of the identified processes in the context of IBD. As organoids have the same genetic background as

their donor, organoids derived from IBD-patients could be exposed to bacterial antigens to further delineate the role of the intestinal epithelium in the pathogenesis. Studying organoids under conditions that activate disease-related mechanisms can be used to test novel therapeutics.³⁷ The growing evidence for the major involvement of the intestinal epithelium in IBD make the intestinal organoids (possibly in co-culture with immune cells) an exquisite system to create an in vitro pipeline for drug testing under a plethora of disease-relevant conditions.

References

- Hindorf, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci.* 106, 9362 LP-9367 (2009).
- Maurano, M. T. et al. Systematic Localization of Common Disease-Associate Variation in Regulatory DNA. *Science* (80-.). 337, 1190–1195 (2012).
- Verbsky, J. W. Monogenic causes of inflammatory disease in rheumatology. *Curr. Opin. Rheumatol.* 24, 506–514 (2012).
- Kaser, A., Zeissig, S. & Blumberg, R. S. Inflammatory bowel disease. *Annu. Rev. Immunol.* 28, 573–621 (2010).
- Thomas, P. D. & Kejariwal, A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. U. S. A.* 101, 15398 LP-15403 (2004).
- Meddens, C. A., van der List, A. C. J., Nieuwenhuis, E. E. S. & Mokry, M. Non-coding DNA in IBD: from sequence variation in DNA regulatory elements to novel therapeutic potential. *Gut* [gutjnl-2018-317516](https://doi.org/10.1136/gutjnl-2018-317516) (2019). doi:10.1136/gutjnl-2018-317516
- Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* (80-.). 343, 1246949 (2014).
- Wit, E. De & Laat, W. De. A decade of 3C technologies-insights into nuclear organization. *Genes Dev.* 11–24 (2012). doi:10.1101/gad.179804.111.GENES
- Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801 (2018).
- Boyd, M. et al. Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nat. Commun.* 9, (2018).
- Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–86 (2014).
- Riethoven, J.-J. M. in (ed. Ladunga, I.) 33–42 (Humana Press, 2010). doi:10.1007/978-1-60761-854-6_3
- Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318 (2007).
- Boyle, A. P. et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* 132, 311–322 (2008).
- He, H. H. et al. Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.* 42, 343–347 (2010).
- Preger-Ben Noon, E., Davis, F. P. & Stern, D. L. Evolved Repression Overcomes Enhancer Robustness. *Dev. Cell* 39, 572–584 (2016).
- Huang, C. et al. H3.3-H4 tetramer splitting events feature cell-type specific enhancers. *PLoS Genet.* 9, e1003558–e1003558 (2013).
- Bernstein, B. E. et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* 125, 315–326 (2006).
- Ostuni, R. et al. Latent Enhancers Activated by Stimulation in Differentiated Cells. *Cell* 152, 157–171 (2013).
- Damen, G. M. et al. Chemokine production by buccal epithelium as a distinctive feature of pediatric Crohn disease. *J. Pediatr. Gastroenterol. Nutr.* 42, 142–149 (2006).
- Mokry, M. et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. *Gastroenterology* 146, 1040–1047 (2014).
- Kaser, A. et al. XBP1 Links ER Stress to Intestinal Inflammation and Confers Genetic Risk for Human Inflammatory Bowel Disease. *Cell* 134, 743–756 (2008).
- Parikh, K. et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* 567, 49–55 (2019).
- Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–24 (2012).

25. Meddens, C. A. et al. Systematic analysis of chromatin interactions at disease associated loci links novel candidate genes to inflammatory bowel disease. *Genome Biol.* (2016). doi:10.1186/s13059-016-1100-3
26. Kontoyiannis, D. et al. Genetic Dissection of the Cellular Pathways and Signaling Mechanisms in Modeled Tumor Necrosis Factor–induced Crohn’s-like Inflammatory Bowel Disease. *J. Exp. Med.* 196, 1563–1574 (2002).
27. Joe, Y. et al. Tristetraprolin mediates anti-inflammatory effect of carbon monoxide against DSS-induced colitis. *PLoS One* 9, 3–8 (2014).
28. Noble, C. L. et al. Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. *Gut* 57, 1398–405 (2008).
29. Planell, N. et al. Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut* 62, 967–976 (2013).
30. Wu, F. et al. Genome-wide gene expression differences in Crohn’s disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflamm. Bowel Dis.* 13, 807–821 (2007).
31. Vancamelbeke, M. et al. Genetic and Transcriptomic Bases of Intestinal Epithelial Barrier Dysfunction in Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* 23, 1718–1729 (2017).
32. Olsen, J. et al. Diagnosis of ulcerative colitis before onset of inflammation by multivariate modeling of genome-wide gene expression data. *Inflamm. Bowel Dis.* 15, 1032–1038 (2009).
33. Fruin, A. B., El-Zammer, O., Stucchi, A. F., O’Brien, M. & Becker, J. M. Colonic Metaplasia in the Ileal Pouch Is Associated With Inflammation and Is Not the Result of Long-Term Adaptation. *J. Gastrointest. Surg.* 7, 246–254 (2003).
34. Jain, U. et al. Temporal Regulation of the Bacterial Metabolite Deoxycholate during Colonic Repair Is Critical for Crypt Regeneration. *Cell Host Microbe* 24, 353–363.e5 (2018).
35. Dotti, I. et al. Alterations in the epithelial stem cell compartment could contribute to permanent changes in the mucosa of patients with ulcerative colitis. *Gut* [gutjnl-2016-312609](https://doi.org/10.1136/gutjnl-2016-312609) (2016). doi:10.1136/gutjnl-2016-312609
36. Suzuki, K. et al. Single cell analysis of Crohn’s disease patient-derived small intestinal organoids reveals disease activity-dependent modification of stem cell properties. *J. Gastroenterol.* 53, 1–13 (2018).
37. Dekkers, J. F. et al. A functional CFTR assay using primary cystic fibrosis intestinal organoids. *Nat. Med.* 19, 939–45 (2013).

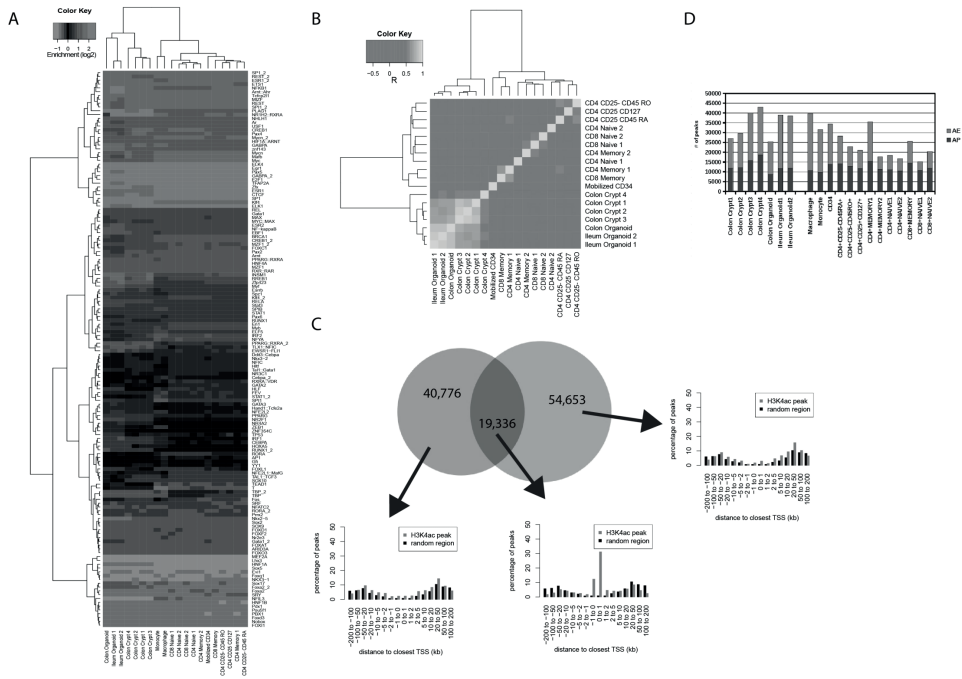


Supplementary material

9

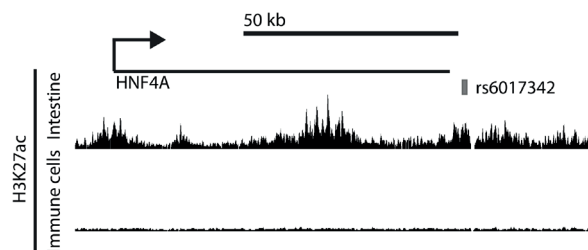
Chapter 3

Supplementary tables and full color images are accessible via: <https://www.sciencedirect.com/science/article/pii/S0016508513017393#appsec1>



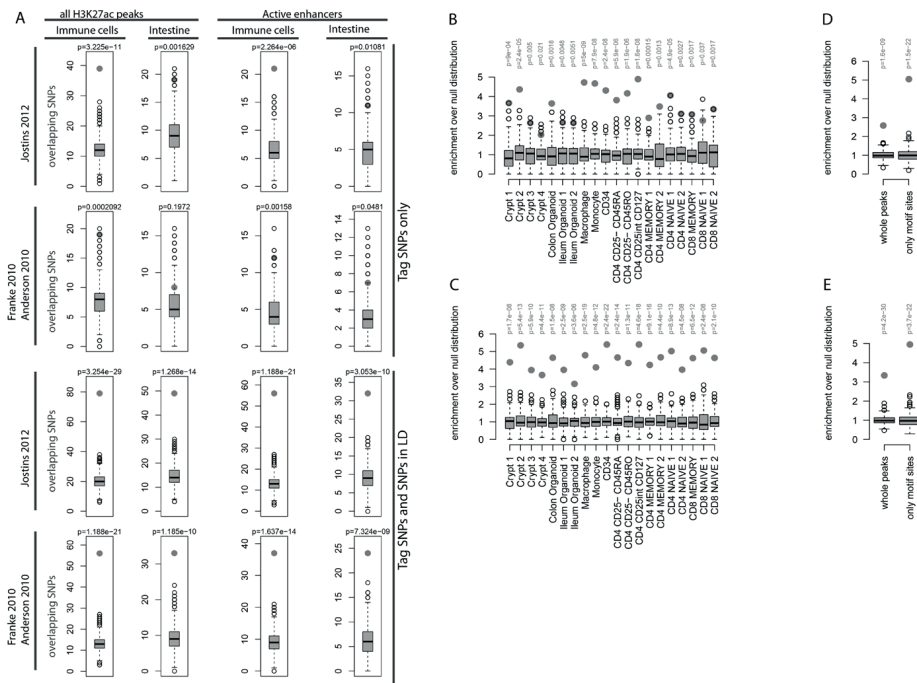
Supplementary Figure 3.1

A) Hierarchical clustering of DREs derived from intestinal epithelium (specified in **Suppl. Table 1**) and immune cells. Clustering is based on known transcription factor binding motif enrichment over random regions. B) Hierarchical clustering of H3K27ac profiles derived from intestinal epithelium and immune cells. Clustering is based on pairwise Pearson's correlation values of the first 10 principal components calculated from normalized H3K27ac signal values on DRE. C) Overlap of DRE identified in immune cells and intestinal epithelium and distribution of separate DRE sets with the respect to closest transcriptional start site. D) Number of H3K27ac peaks identified in intestinal epithelium and immune cells and their location with respect to transcriptional start sites (TSS). Active enhancers (AE) are located distal (> 2.5 kilobase) and active promoters (AP) < 2.5 kilobase from the closest TSS.



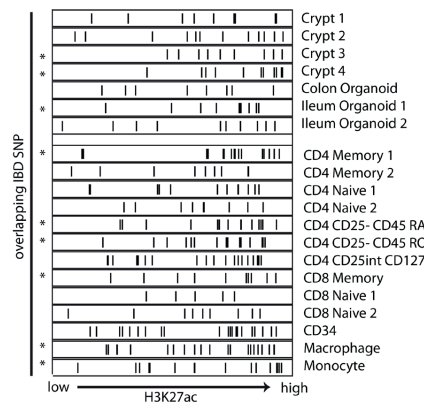
Supplementary Figure 3.2

Browser view of the H3K27ac signal in intestinal epithelium and immune cells depicting the intestine-specific active regulatory element overlapping with an example of an IBD-associated SNP¹, which is located close to the HNF4A gene known to play role in intestinal inflammation².



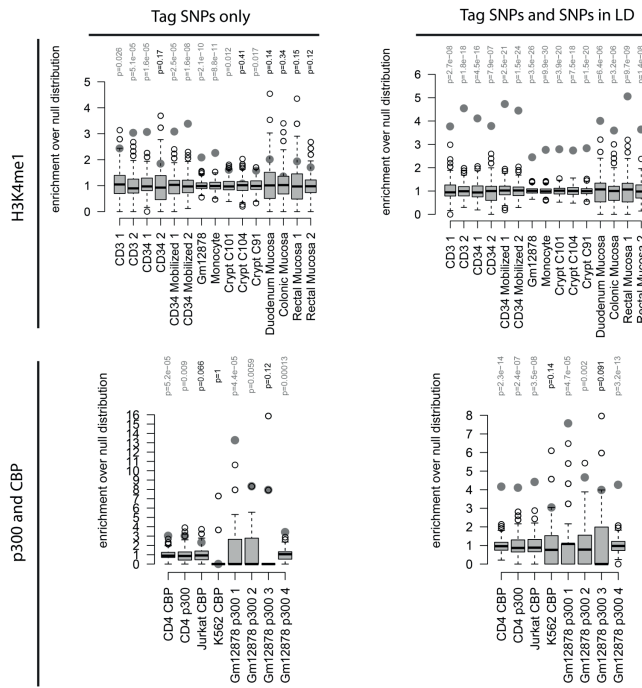
Supplementary Figure 3.3

A) Co-localization of IBD-associated SNPs from the different GWAS^{3, 4} with H3K27ac-marked regulatory elements in intestinal epithelium and immune cells. The number of IBD-associated SNPs (gray dot) overlapping with the regulatory regions by themselves or by SNPs in strong LD ($r_2 > 0.8$) compared to 10,000 random SNP sets (grey boxes). B-F) Co-localization of IBD-associated SNPs with DRE normalized to null distribution calculated from 10,000 random SNP sets. B) in separate samples. C) as B including SNPs in strong LD ($r_2 > 0.8$). E) with known transcription factor binding motifs located on DREs F) as E including SNPs in strong LD ($r_2 > 0.8$). p-Value for all the panels was calculated from binominal cumulative distribution function. Statistically significant co-localizations are depicted in gray.



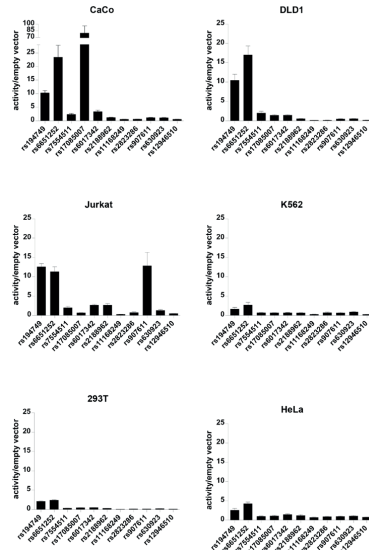
Supplementary Figure 3.4. Activity of overlapping DRE in separate samples.

DREs from separate samples are ranked according to the H3K27ac signal. DRE overlapping with IBD-associated SNPs are indicated by black lines. * $p < 0.05$, calculated using Wilcoxon signed rank test with continuity correction.



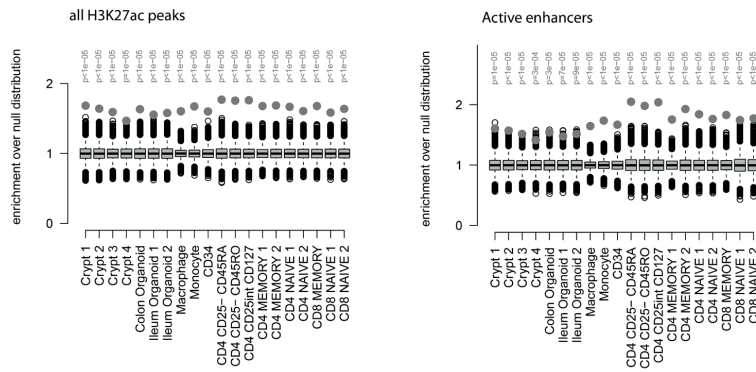
Supplementary Figure 3.5

Co-localization of IBD-associated SNPs with H3K4me1, p300 and CBP datasets normalized to null distribution calculated from 10,000 random SNP sets (grey boxes). p-Values for all the panels were calculated with binominal cumulative distribution function. Statistically significant co-localizations are depicted in gray.



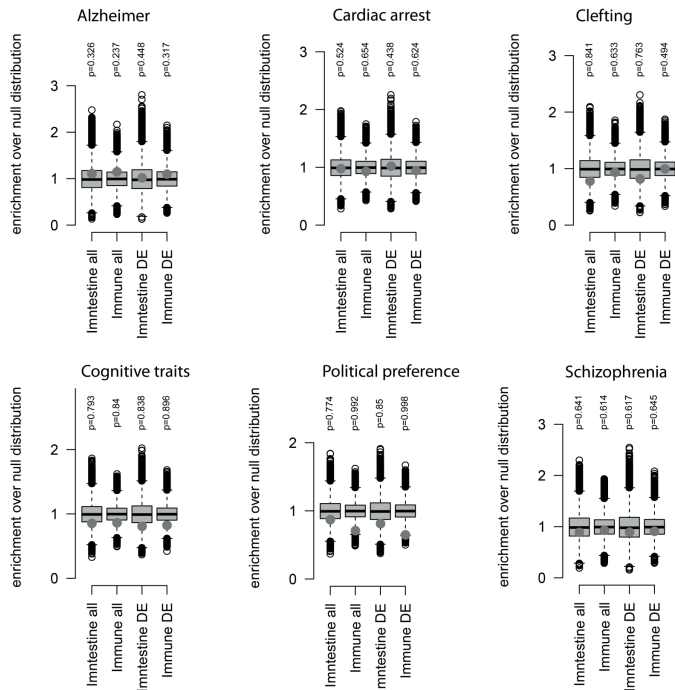
Supplementary Figure 3.6

Luciferase reporter assay performed in six different cell lines. Mean of at least four replicate experiments is shown. Error bars represent SEM values. Values are normalized over control pGL4.10 vector with minimal TATA box and with a CMV-driven Renilla as transfection control.



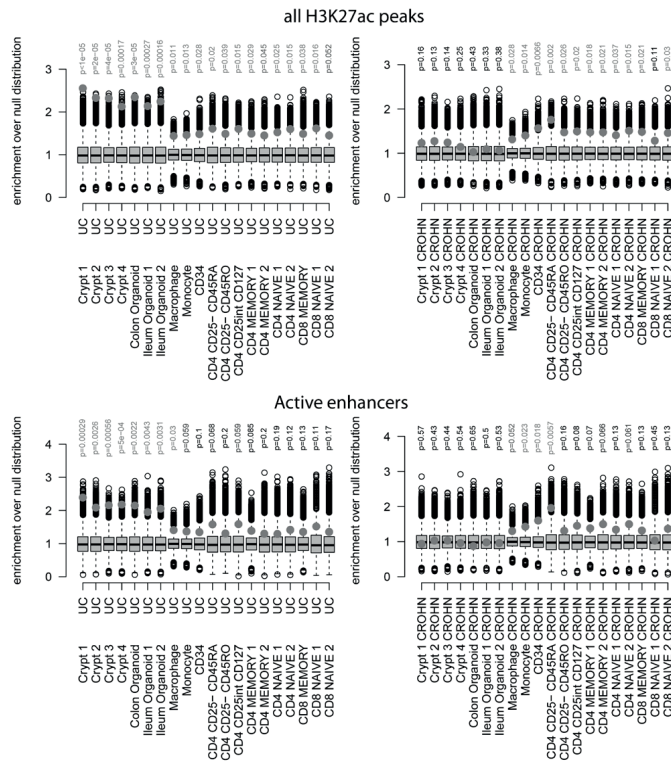
Supplementary Figure 3.7

Enrichment of regulatory elements identified in different samples from intestinal epithelium (specified in **Suppl. Table 1**) and immune cells in IBD-associated loci (gray dot) normalized to the average counts in 100,000 matched control sets (grey boxes). p-Values were calculated with a permutation test. Statistically significant co-localizations are depicted in gray.



Supplementary Figure 3.8

Enrichment or depletion of regulatory elements identified in intestinal epithelium and immune cells in loci associated with different traits (gray dot) normalized to the average counts in 100,000 matched control sets (grey boxes). p-Values were calculated with a permutation test



Supplementary Figure 3.9

Enrichment of regulatory elements identified in different samples from intestinal epithelium (specified in **Suppl. Table 1**) and immune cells in CD- and UC-associated loci (gray dot) normalized to the average counts in 100,000 matched control sets (grey boxes). P-values were calculated with a permutation test. Statistically significant co-localizations are depicted in gray.

Supplementary data references

1. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;491:119-24.
2. Ahn SH, Shah YM, Inoue J, et al. Hepatocyte nuclear factor 4 alpha in the intestinal epithelial cells protects against inflammatory bowel disease. *Inflamm Bowel Dis* 2008;14:908-920.
3. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47 (vol 43, pg 246, 2011). *Nat Genet* 2011;43:919-919.
4. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010;42:1118-+.

Supplementary methods

Crypt isolation

Intestinal crypts were isolated as described previously¹. In brief, fibrous and muscular layers of intestinal samples were removed and the samples were cut into small pieces. Biopsies and intestinal fragments were extensively washed with ice cold PBS. The samples were incubated in cold chelation buffer (5.6 mmol/L Na₂HPO₄, 8.0 mmol/L KH₂PO₄, 96.2 mmol/L NaCl, 1.6 mmol/L KCl, 43.4 mmol/L sucrose, 54.9 mmol/L d-sorbitol, 0.5 mmol/L dl-dithiothreitol and 2 mmol/L EDTA) for 30-60 minutes. After removal of the buffer, tissue fragments were vigorously resuspended in cold chelation buffer. The intestinal fragments were allowed to sediment and the supernatant was

inspected by inverted microscopy to check for presence of crypts. The resuspension/sedimentation procedure was typically repeated 1-4 times. The supernatants containing crypts were collected. Isolated crypts were pelleted, washed with cold chelation buffer, and centrifuged at 150–200g for 3 minutes to separate crypts from single cells. All steps were performed on ice.

Organoid cultures

Human organoids were maintained as describe previously¹. In brief, organoids were embedded into matrigel (BD Biosciences) and maintained in medium containing Rspo-1, Noggin, EGF, ALK4/5/7 inhibitor A83-01, Nicotinamide, SB202190 and Wnt3. We used conditioned media for Rspo-1, Noggin and Wnt3a (stably transfected Rspo-1 HEK293T cells were kindly provided by Dr. C. J. Kuo, Department of Medicine, Stanford, CA).

ChIP-seq

Chromatin IP was performed using MAGnify ChIP kit (Invitrogen) according manufacturers' recommendations; with these exceptions: crypts or organoids were cross-linked by 2% formaldehyde in PBS for 10 minutes, reaction was quenched with excess of glycine. Samples were washed with ice cold PBS and subsequently fragmented in 500 µl of lysis buffer. 20 µl of chromatin was used per single IP. One µl anti H3K27ac (ab4729, Abcam) was used per IP. Sequencing libraries from immunoprecipitated chromatin were prepared as described previously². In brief, the chromatin was additionally sheared, blunt ended and phosphorylated. The sequencing adaptors were ligated and the resulting library was PCR amplified using ligation-mediated PCR. The libraries were sequenced on SOLiD 5500 or WildFire sequencer in a multiplexed way to produce 40-50-bp long reads. Sequencing reads were mapped against the reference genome (hg19 assembly, NCBI³⁷) using BWA³ package (-c, -l 25, -k 2, -n 10). Multiple reads mapping to same location and strand have been collapsed to single read and only uniquely placed reads were used for peak-calling. Cisgenome v.2⁴ software package (-e 50, -maxgap 200, -minlen 200) was used for peak-calling from the ChIP-seq against the common input sample. All called peaks have FDR < 0.05. H3K27ac peaks from immune cells were called in the same way from publically available datasets^{5,6}.

Overlapping of SNPs with active DRE and DE

Coordinates H3K27ac peaks identified in individual samples were first stretched to at least 500 base pairs and then combined into three lists (intestinal, immune cells and all). Within these lists overlapping peaks were combined into a single peak – active DRE. Peaks with distance to closest annotated start site (HG19, annotation files downloaded from Cisgenome⁴ website) larger than 2,5kb we considered as active distal enhancers (DE). SNP falling within peak coordinates was considered as overlapping SNP.

Random matched SNP sets (500 unique random SNPs per each GWAS SNP) were generated from variants present on Human Omni1S genotyping chip (Illumina). Each random SNP has similar minor allele frequency (+/- 5%) as GWAS SNP based on frequencies from ftp://ftp.ensembl.org/pub/release-72/variation/gvf/homo_sapiens/1000GENOMES-phase_1_EUR.gvf.gz. Each random SNP has similarly located closest annotated transcription start site as GWAS SNP based on hg19, annotation files downloaded from Cisgenome⁴ website. Similar location of closest TSS is defined by belonging to the same “location bin” as GWAS SNP - 12 location bins are separated by {-200 k bp, -100 k bp, -25 k bp, -10 k bp, -5 k bp, 0 bp, 5 k bp, 10 k bp, 25 k bp, 100 k bp, 200 k bp}.

LD information (based on Utah residents with Northern and Western European ancestry from the CEPH collection “CEU dataset”) for GWAS SNPs and for random matched SNPs was accessed from the HapMap website⁷: http://hapmap.ncbi.nlm.nih.gov/downloads/ld_data/latest/ on 8th of August 2013. SNPs with $R^2 > 0.8$ were considered as SNPs in strong LD.

To depict the significance of overlap of regulatory elements with GWAS SNPs, we calculated the p-value using binominal cumulative distribution function $b(x; n, p)$ (as done previously⁸) using R^9 `pbinom()` function. With the probability of success calculated from 10,000 randomized datasets picked from random matched SNP sets.

Overlapping of DRE and DE with susceptibility loci

Susceptibility locus was defined as 500,000 base pair long genomic region with susceptibility SNP in the middle. DRE and DE falling within locus coordinates were considered as overlapping with a susceptibility locus. We compared the number of DRE overlapping the susceptibility locus and random matched loci. The 500,000 base pair size was selected based on cutoff used in recent IBD GWAS⁹⁰. This distance bin also contains substantial part of distal chromatin interaction of promoters with regulatory elements⁹¹.

Primers used for cloning for luciferase reporter assay:

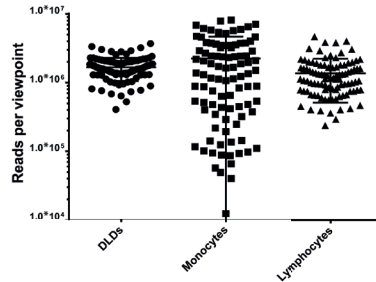
Primer sequences are accessible via: <https://www.sciencedirect.com/science/article/pii/S0016508513017393#appsec1>

Supplementary methods references

1. Sato T, Stange DE, Ferrante M, et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* 2011;141:1762-72.
2. Mokry M, Hatzis P, Schuijers J, et al. Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res* 2012;40:148-58.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754-60.
4. Jiang H, Wang F, Dyer NP, et al. CisGenome Browser: a flexible tool for genomic data visualization. *Bioinformatics* 2010;26:1781-2.
5. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010;28:1045-8.
6. Pham TH, Benner C, Lichtinger M, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* 2012;119:e161-71.
7. International HapMap C, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52-8.
8. Maurano MT, Humbert R, Rynes E, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 2012;337:1190-1195.
9. R_Development_Core_Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.
10. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012;491:119-24.
11. Sanyal A, Lajoie BR, Jain G, et al. The long-range interaction landscape of gene promoters. *Nature* 2012;489:109-U127.

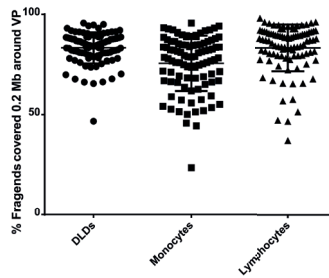
Chapter 4

Supplementary figures



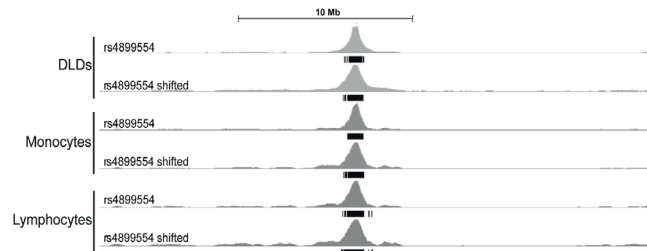
Supplementary figure 4.1. Number of reads per dataset.

Number of sequencing reads are shown for each of the 92 viewpoints in the three assayed cell types. Each dot represents one of the 276 viewpoints. The number of reads per dataset, although high in all cell types, is more variable in monocytes. This is due to the usage of a different pcr-multiplexing strategy in monocytes (see materials and methods for details). Datasets that consist of less than 1×10^5 reads, do not compromise on complexity.



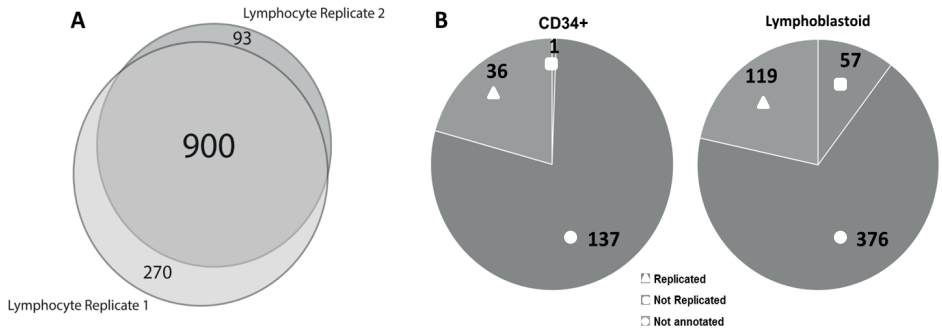
Supplementary figure 4.2. Percentage of fragends covered in a region of 0.2 Mb surrounding the viewpoint.

This quality measure provides an indication of the complexity of the sequenced libraries, libraries with a percentage of $>40\%$ are considered to be of high complexity.¹⁷ Each dot represents one of the 276 viewpoints.



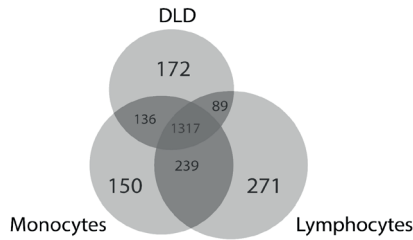
Supplementary figure 4.3. Different primer pairs in the same region give similar signals.

The 4C signal is shown for one viewpoint for which two primer pairs were designed (2620 bp apart). The presented primer pairs give similar signals within each cell type.



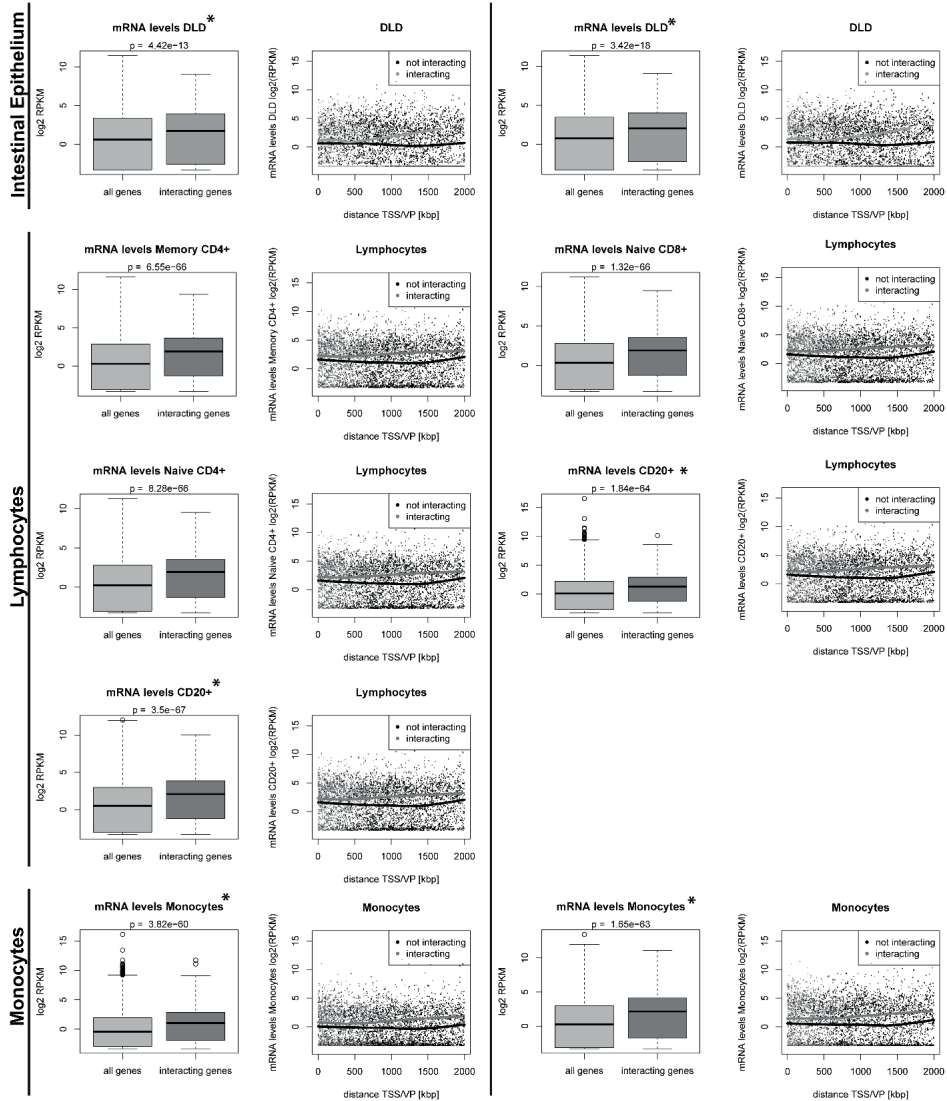
Supplementary figure 4.4. Validation dataset and overlap with Hi-C datasets.

A) Replication of the 4C-seq experiment on a 4C-template that was prepared from lymphocytes of a different donor. 91% of the candidate genes that were identified in the 2nd replicate were also identified in the dataset that is used throughout this study. This validated the high reproducibility of the 4C-technique; not only in technical, but also in biological duplicates. The replication dataset can be found in Supplementary table 2. B) To study the reproducibility of our results in available chromosome interaction datasets we intersected our data with two Hi-C datasets¹ that were created in CD34⁺ leukocytes and a lymphoblastoid cell line. First, we identified all genes in the Hi-C datasets that were found to interact within 5 kB of the genomic locus of the 4C viewpoints (i.e. loci of the SNPs that we identified as being active by H3K27Ac marks, Supplementary table 2), this resulted in 174 and 552 genes in the CD34⁺ leukocytes and lymphoblastoid cell line respectively. Next we checked which of these genes were annotated differently in the 4C-seq datasets (genes indicated with circle). Finally, we determined how many of these genes were also present in the 4C-datasets (genes indicated with triangle). This confirmed a high reproducibility by showing that 99% (CD34⁺) and 87% (lymphoblastoid) of the genes that were found by Hi-C were also found in the 4C-data presented here (these percentages do not include the genes that were differently annotated between the Hi-C and 4C data and could therefore not be studied in this comparison).



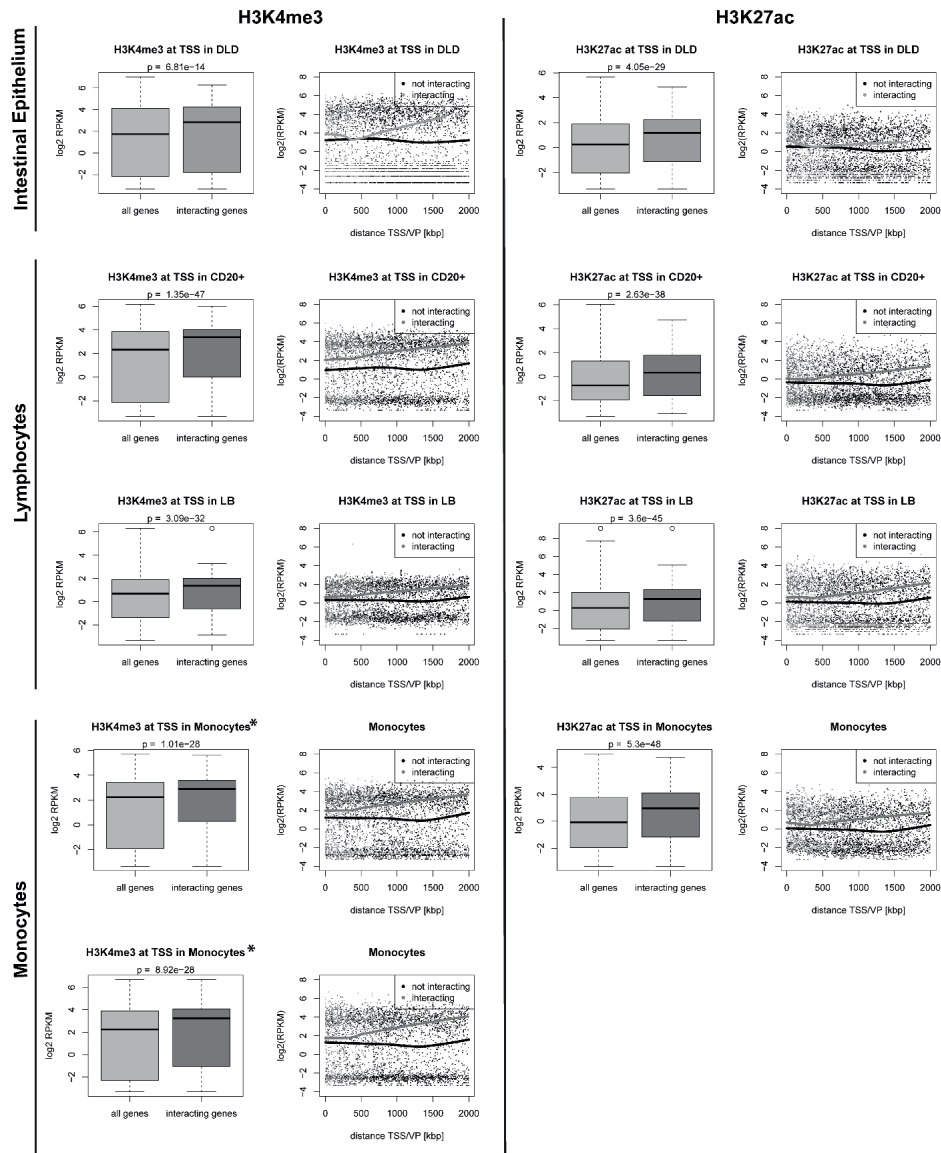
Supplementary figure 4.5. Overlap between interactions identified in different cell types.

This figure shows the number of genes that co-localize with a significant 4C signal in the three cell types. The surface of the circles corresponds with the genes unique for one cell type and the genes that overlap only two cell types. The number of genes shared by all three cell types is depicted in the center of the diagram. All cell types show a distinctive set of genes. As expected, monocytes and lymphocytes share more genes compared to DLD-1. In order to prevent possible biases based on cell type differences based on expression or enhancer activity, all genes that colocalize with 4C signal in any of the 92 viewpoints in any of the 3 cell types (regardless of enhancer activity of gene expression) have been used in this analysis.



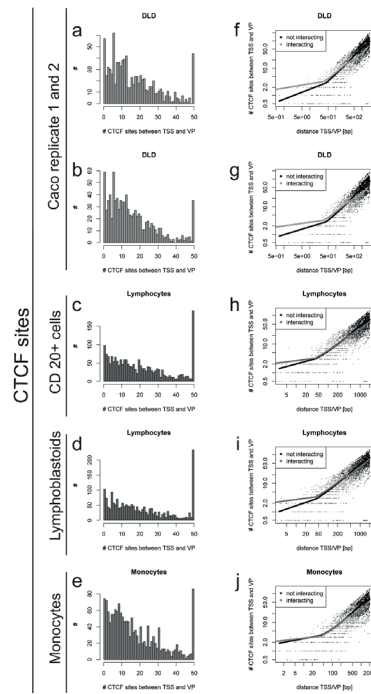
Supplementary figure 4.6. RNA expression.

Boxplots show that the mean RNA expression of regulated genes is higher than expression in all annotated genes. Each plot shows one RNA-seq data set. For all data sets, the genes interacting with the assayed regulatory elements are significantly higher expressed than all annotated genes. P-values are based on Mann-Whitney-U tests. The dot charts show expression levels of genes in the vicinity of the viewpoints. Genes that are interacting with the assayed enhancer are expressed at higher levels. *Biological duplicates for the same cell type.



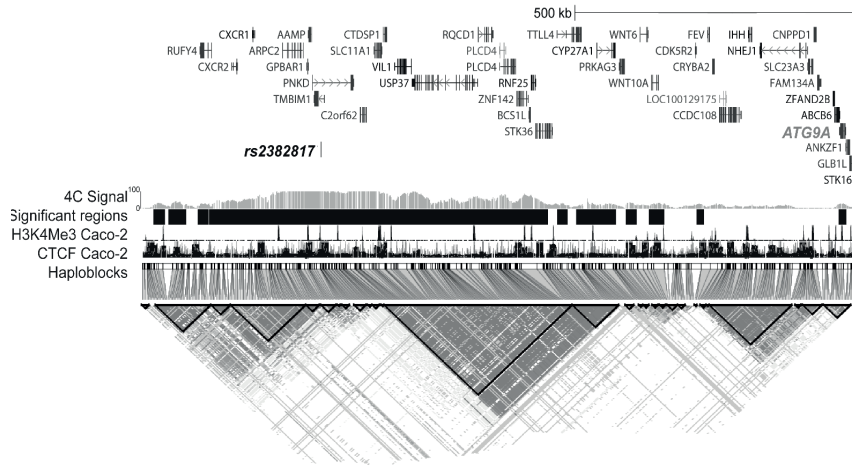
Supplementary figure 4.7. H3K4me3 and H3K27ac occupancy.

TSS occupancy by H3K4me3 and TSS occupancy by H3K27ac in different cell types. Both H3K4me3 and H3K27ac occupancy is significantly higher at TSSs of genes detected with 4C-seq, p-values are based on Mann-Whitney-U tests. Dot plots depict TSS occupancy by H3K4Me3 and H3K27Ac in genes in the vicinity of the viewpoint (interacting and not interacting with the assayed enhancer). *Biological duplicates for the same cell type.



Supplementary figure 4.8. Insulators between viewpoint and candidate genes

A-E) Number of CTCF bound-loci between viewpoint and interacting gene. F-J) The number of CTCF occupied sites between a gene and the viewpoint does not differ for genes interacting with the viewpoint, compared to genes that do not interact with the viewpoint (i.e. genes that do not co-localize with significant 4C signal).



Supplementary figure 9. ATG9A is identified as a novel IBD candidate gene.

4C-seq results of rs2382817 in DLDs identify ATG9A as novel candidate gene. ATG9A localizes to a different haploblock than the IBD-associated SNP (rs2382817). 4C signal of the rs2382817 locus is depicted on the y-axes as the percentage of fragends covered per pixel (see methods for details). Regions that are significantly enriched in 4C signal ($P < 10^{-8}$) are depicted by the black bars. ChIP-seq signals for H3K4Me3 and CTCF (in Caco-2 cell line, publicly available datasets from Encode consortium, see methods for additional information) and haploblocks (data retrieved from Haploview, see methods for additional information) are shown.

Supplementary methods

Circular Chromatin Conformation Capture-Template preparation

4C-chromatin was prepared as described previously.² In brief, 10×10^6 cells were used for chromatin preparation per cell type (monocytes, PBLs and DLD-1). Cells were crosslinked in 2% formaldehyde, lysed in lysis buffer and chromatin was isolated. Chromatin was digested with DpnII (NEB, #R0543L). After inactivation, the samples were diluted and ligated by T₄ DNA ligase. Thereafter the second digestion was done using CviQI (NEB, #R069S) and inactivated by phenol:chloroform extraction. Finally, the chromatin was diluted, ligated and purified. Digestion and ligation quality were analyzed for the proper fragment lengths on agarose gels.

Primer design

Primer sequences are listed in **Supplementary table 1**. Primers were designed as was described previously.² In brief, primers were designed in a window of 5 kbp up- and downstream from the associated SNP. Forward and reverse primers were designed at least 300 bp apart. Forward (reading) primers were designed on top of the first restriction enzyme site. The reverse (non-reading) primer was designed close to (max 100 bp away from) the second restriction enzyme site. In case no primer pair could be designed within the initial window, the window was extended 5 kbp up- and downstream ($n=22$). If still no primer could be designed, we selected a primer pair that was less than 300 bp, but at least 240 bp, apart ($n=2$).

Circular Chromatin Conformation Capture- Sequencing (4C-seq) library preparation

4C-sequencing library preparation was performed as described previously,² with minor adaptations in order to make the protocol compatible with the large number of viewpoints: the PCR of 4C template was performed with 800 ng to 1,6 µg of 4C template per reaction. 4 to 10 primer pairs were multiplexed in the initial PCR reaction (primer sequences are listed in **Supplementary table 1**). Primer pairs were pooled according to primer efficiency. In reactions in which ≥ 6 primer pairs were used, PCR products were purified after an initial PCR reaction of 6 cycles (reaction volume = 200 µL) and divided among 8-10 PCR reactions containing single primer pairs for another 26 cycles (reaction volume = 25 µL). In PCR reactions in which < 5 primer pairs were used, thermal cycling was limited to one reaction of 28 cycles.

Thereafter, PCR products derived from the same cells were pooled in equimolar amounts and a final 6 cycle PCR reaction containing 20 ng of pooled PCR product (reaction volume = 100 µL) was performed with primers that contained sequencing adaptor sequences (**Supplementary table 1**). All fragments > 700 bp were removed using size selection on a 1% agarose gel followed by gel extraction of the selected products (Qiagen, #28704). Quality measures for the 4C library preparation and sequencing can be found in **Supplementary figure 4.1-4.3**.

eQTL analyses (STAGE)

The STAGE study was used to investigate the association between the identified GWAS loci and gene expression. The STAGE dataset consists of seven vascular and metabolic tissue samples of well-characterized coronary artery disease (CAD) patients gathered during coronary artery bypass grafting (CABG) as described³. Patients were included if they were eligible for CABG and had no other severe systemic diseases (e.g., widespread cancer and active systemic inflammatory disease). Fasting white blood cells were obtained for DNA and RNA isolation. The Ethical committee of the Karolinska Hospital approved this study and patients gave written consent (Dnr 004-02).

Genome-wide Human SNP array 6.0 (Affymetrix) was used for genotyping. From total 909,622 SNPs, 530,222 autosomal SNP passed quality control filters (minor allele frequency MAF $< 5\%$, Hardy-Weinberg equilibrium (HWE) $P < 1e-6$, and call rate of 100%). A custom-made HuRSTA-2a520709 was used for gene expression profiling from 109 genotyped patients (WB, $n=102$). A total 19,610 gene expression profiles were obtained. The missing autosomal SNP in STAGE study were imputed using IMPUTE2⁴ after pre-phasing with SHAPEIT2⁴ using 1000 Genomes (phase

1, version 3)⁵. A total of 5,473,585 autosomal SNPs were selected after filtering out low quality imputed genotypes (INFO score < 0.3).

Identified loci from GWAS for IBD were matched with imputed and genotyped SNPs and were selected for eQTL discovery. We used Matrix-eQTL⁶ for investigating association between gene expression and SNPs. We compared the amount eQTLs between 'SNP-candidate gene'-pairs and 'SNP-control gene'-pairs. Control genes are genes within the same locus that are not interacting with the IBD associated locus. An empirical FDR was estimated for each eQTL-gene by shuffling patient IDs 1000 times on genotype data as described previously⁷.

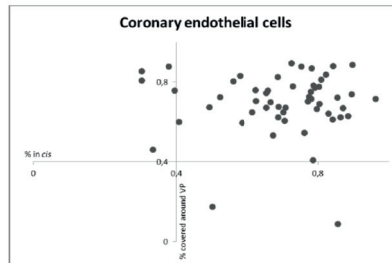
References

1. Mifsud, B. et al. Sup Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47, 598–606 (2015).
2. van de Werken, H. J. G. et al. 4C technology: protocols and data analysis. *Methods in enzymology* 513, (Elsevier Inc., 2012).
3. Hägg, S. et al. Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study. *PLoS Genet.* 5, (2009).
4. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–9 (2012).
5. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65 (2012).
6. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–8 (2012).
7. Foroughi Asl, H. et al. Expression quantitative trait Loci acting across multiple tissues are enriched in inherited risk for coronary artery disease. *Circ. Cardiovasc. Genet.* 8, 305–15 (2015).

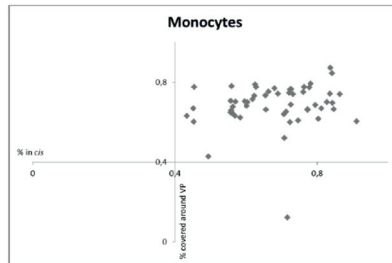
Chapter 5

All supplementary tables are accessible at:

<https://www.ahajournals.org/doi/suppl/10.1161/CIRCGENETICS.116.001664>



Supplementary figure 5.1

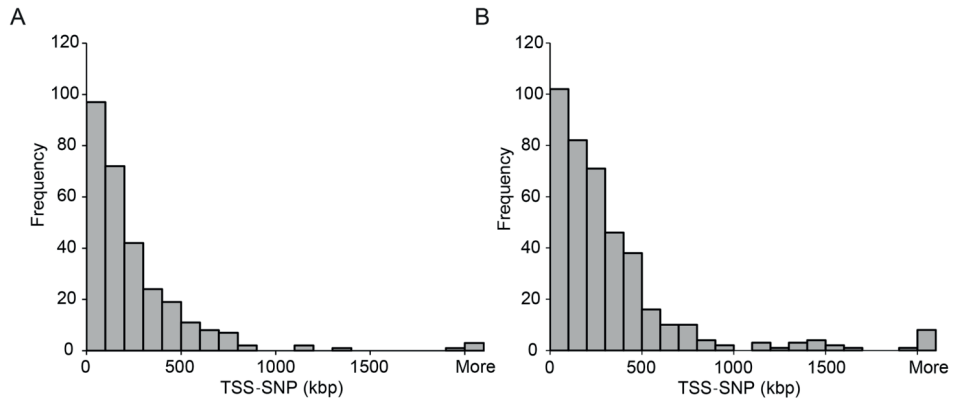


Supplementary figure 5.2

Chapter 6

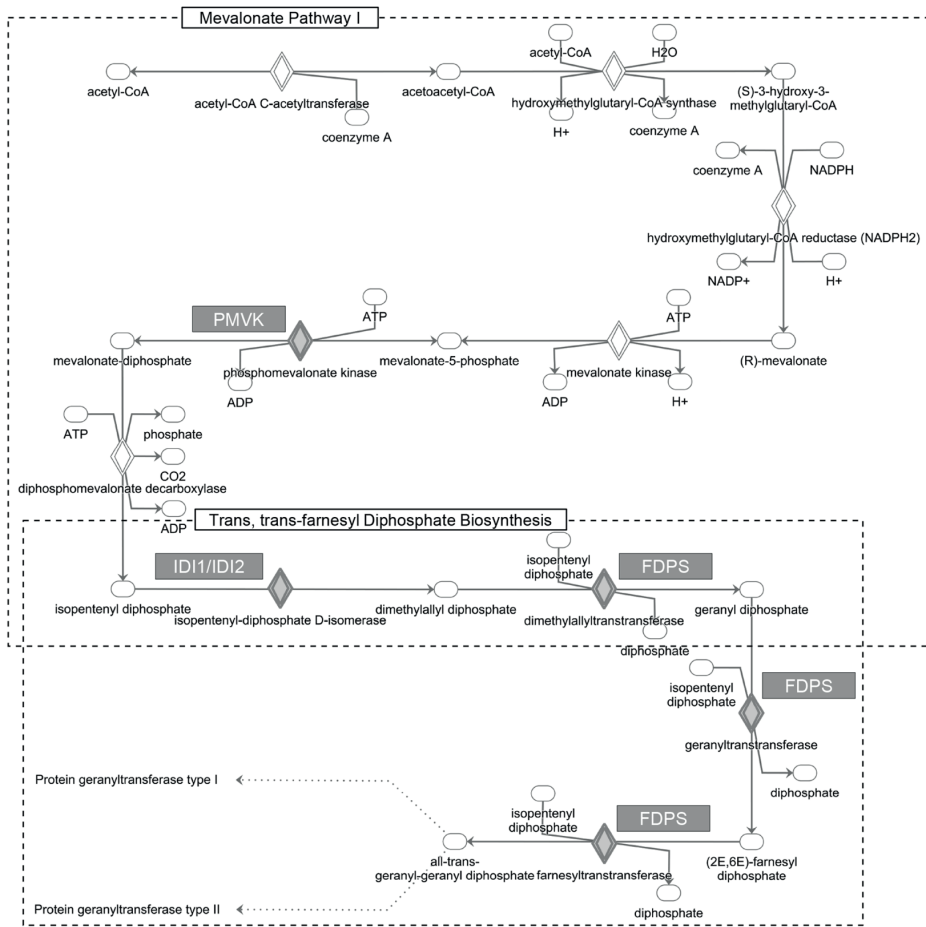
All supplementary tables are accessible at:

<https://jasn.asnjournals.org/content/jnephrol/suppl/2017/11/01/ASN.2016080875.DCSupplemental/ASN.2016080875SupplementaryData1.pdf>



Supplemental figure 6.1: The majority of genes identified with 4C-seq were positioned up to 500kbp from the CKD associated SNP locus.

Graph shows the distance from the SNP to the TSS of the target genes found with 4C in HRGECs (X-axis), expressed as the number of target genes (frequency on Y-axis) per 100kbp. The majority of target genes was positioned up to 500kbp from the SNP, but occasionally candidate genes were found at locations over 1.5mb from the SNP locus (A). The TSS of the majority of candidate genes found with 4C in HRPTECs was positioned up to 500kbp from the SNP position, but some candidate genes were found at locations over 1.5mb from the SNP locus (B).

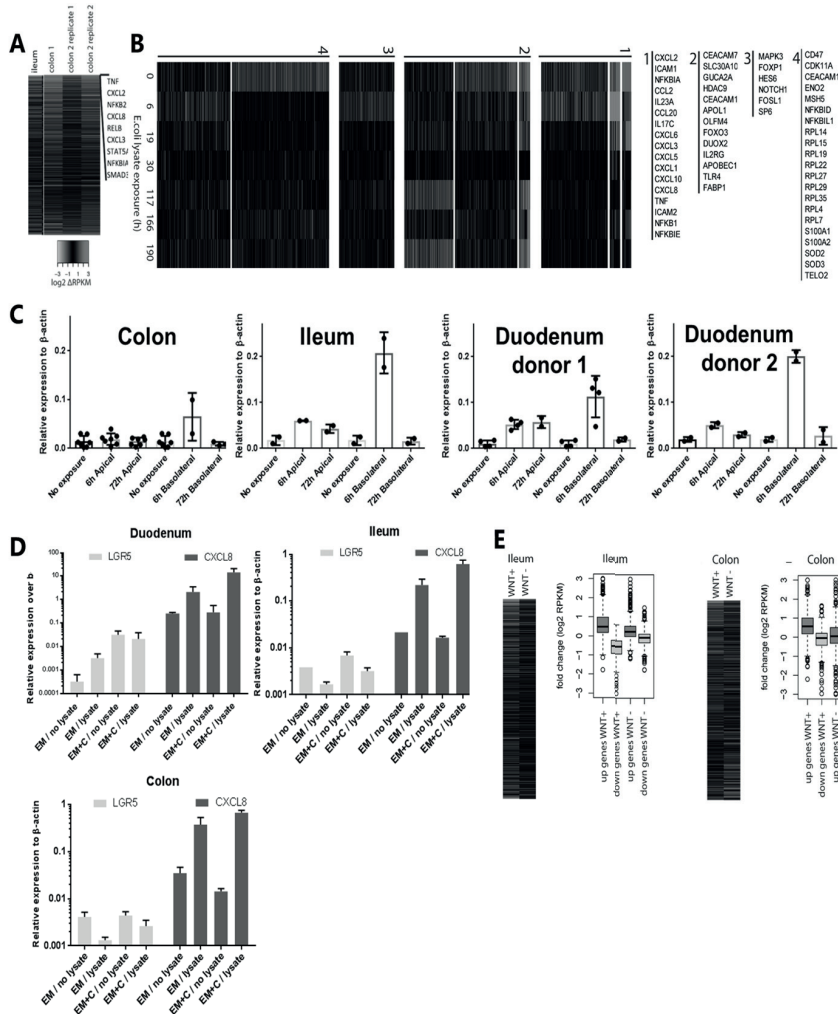


Supplemental figure 6.2

Target genes of DREs colocalizing with CKD-associated SNPs. Gene names depicted in the rectangles are enriched in the mevalonate pathway and the trans, trans-farnesyl diphosphate biosynthesis pathway (adapted from IPA).

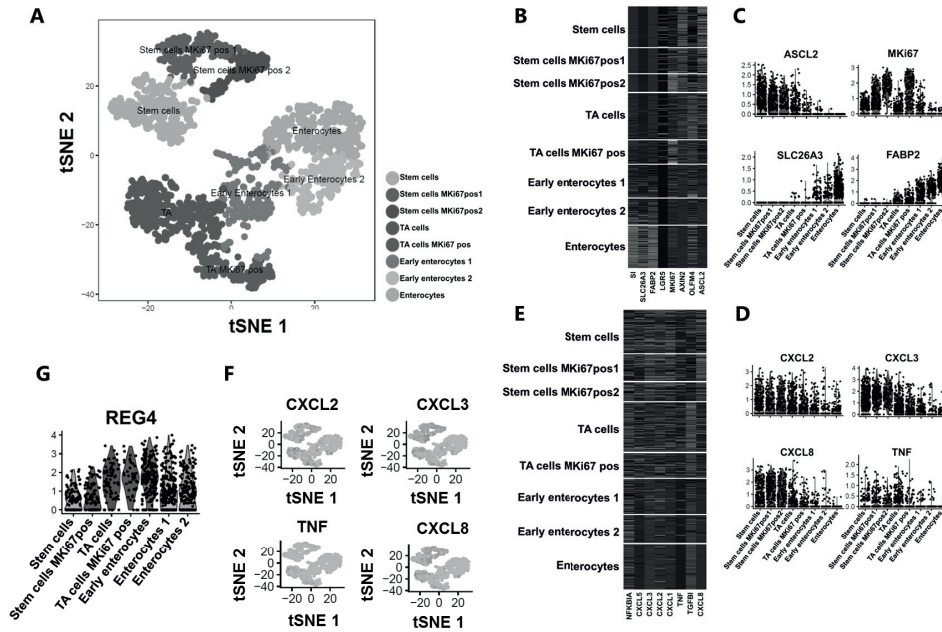
Chapter 7

Full color figures and supplementary tables can be requested via: c.a.meddens@umcutrecht.nl



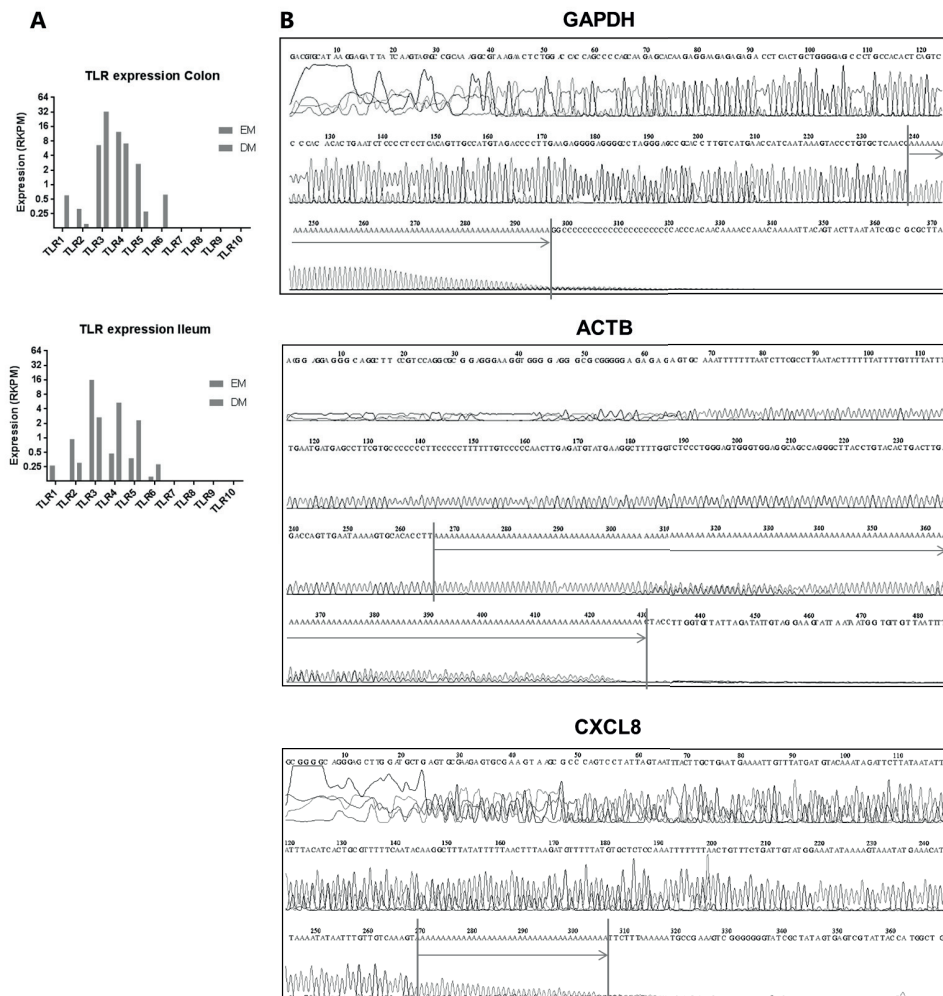
Supplementary figure 7.1.

A. Heatmap of RNA-sequencing data of 1 ileum and 2 colon derived organoid lines (+1 technical colon duplicate) that were exposed for 6h with bacterial lysates. These data show that the epithelial response to bacterial antigens is conserved between large and small intestine, between different healthy donors and between different experiments with the same organoid line. B. Heatmap of RNA-sequencing data of colon organoids that were exposed for 0 to 190 hours. For each cluster key genes are listed. C. Heatmap of RNA-sequencing data of ileum and colon organoids. Data show that the elimination of WNT from the culture media (and thereby a reduce in stemness) reduces the responsiveness of both ileum and colon organoids. D. qPCR of LGR5 and CXCL8. Addition of CHIR induced stemness (marked by LGR5-expression) in duodenum and ileum organoids and induced the responsiveness. CHIR did not induce stemness in colon organoids. E. qPCR of CXCL8 on organoids grown as monolayers. Basolateral 6h exposure of colon, ileum and duodenum organoids results in clear response through upregulation of CXCL8. Apical exposure does not significantly induce CXCL8 expression except for 1 donor of duodenum organoids.



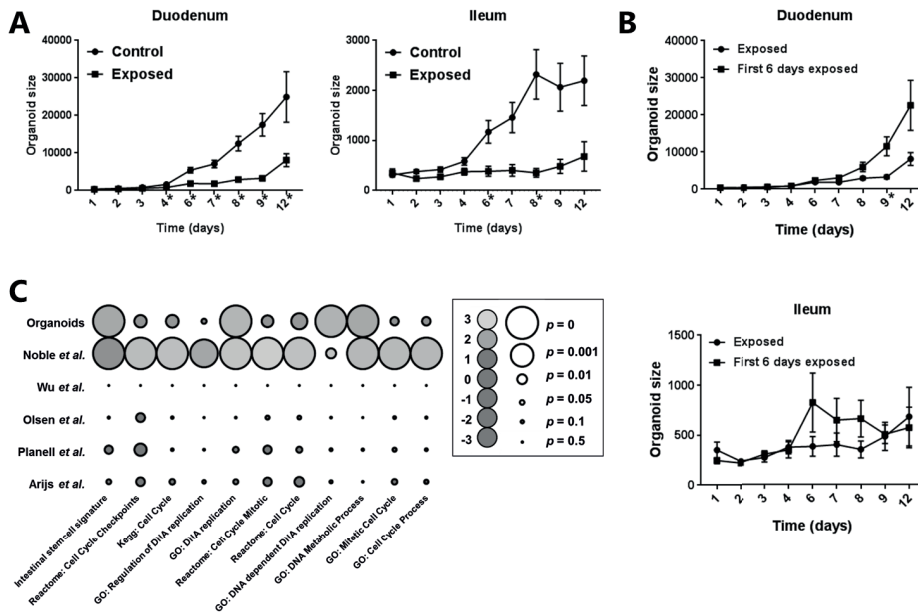
Supplementary figure 7.2. Stem cell data of exposed ileum organoids

A. tSNE-plot of 8 cell clusters that represent the 3 main cell types. Pooled data from differentiated and undifferentiated colon organoids exposed to bacterial lysates. B. Heatmap of expression of stemness and differentiation markers in each identified cell type. C. Expression of KI67, SLC26A3, FABP2 and ASCL2 per cell type. D. Expression of pro-inflammatory genes per cell type. These data confirm the finding in colon organoids that stem cells are a main source of the inflammatory response to bacterial antigens. E. Heatmap of inflammatory markers in each identified cell type. F. Distribution of expression of CXCL2, 3, 8 and TNF α on tSNE-plot. B. Expression of REG4 in exposed colon organoids. This does not correlate with expression of inflammatory markers.



Supplementary figure 7.3

A. mRNA expression of TLR in differentiated (DM) and undifferentiated (EM) organoids. B. Sanger sequencing of poly(A)-tail length for GAPDH (58nt), β -ACTIN (164nt) and CXCL8 (38nt). Experiments were done in differentiated colon organoids that were exposed to bacterial lysates for 6h.



Supplementary figure 7.4

A. Proliferation assay. Small intestinal organoids were grown in the presence or absence of bacterial lysates for 12 days. Organoid size was significantly lower in exposed organoids. At timepoints that indicated with * the size was significantly different between the two conditions (Kolmogorov-Smirnov $p < 0.05$). In both organoid lines, growth is inhibited upon exposure to bacterial antigens. B. Proliferation assay. Small intestinal organoids were grown in the presence of bacterial lysates. After 6 days of exposure, half of the organoids were further cultured without bacterial lysates. At timepoints that indicated with * the size was significantly different between the two conditions. (Kolmogorov-Smirnov $p < 0.05$). In ileum organoids, although an initial growth acceleration is seen, this is not maintained throughout the whole experiment. In duodenum organoids, removal of the initial stimulus results in reversal of the hypo-proliferative state. C. Gene set enrichment analysis of inflamed vs healthy colon biopsies for gene sets involved in proliferation. Upper panel consist of expression data from exposed vs non exposed colon organoids. A positive normalized enrichment score indicates that the gene set is upregulated under inflamed conditions.



Samenvatting
Dankwoord
Curriculum vitae
List of publications

10

Summary

Delineating the pathogenesis of complex genetic diseases is complicated by the great variety of genetic loci, genes, cell types, environmental factors and tissues that are involved. The genetic background of complex genetic diseases has been intensively studied through genome wide association studies (GWASs). This led to the association of many genetic loci to a multitude of diseases. The identification of causal variants and affected genes has proven difficult, thereby leaving some pathogenic mechanisms unresolved and potential therapeutic targets unrevealed. In this thesis, we study the involvement of DNA regulatory elements (DRE) in the pathogenesis of complex genetic diseases and epithelial mechanisms that are potentially involved in inflammatory bowel disease (IBD).

In the first part of this thesis we studied loci associated to IBD, CVD (cardiovascular diseases) and CKD (chronic kidney disease). We show that disease associated loci are enriched for active DRE and we used chromatin conformation capture to study the 3D configuration of these loci. Through this approach we identified many novel candidate genes, pathways and key regulators of IBD, CVD and CKD.

Genetic studies on IBD have shown that the intestinal epithelium plays an important role in the pathogenesis of this complex genetic disease. Therefore, in the second part of this thesis, we use human intestinal organoids to study the interaction between bacterial antigens and the intestinal epithelium. We show that intestinal stem cells are a major source of the inflammatory response and that the responsiveness of the epithelium is regulated at the post-transcriptional level.

Samenvatting

De pathogenese van complex genetische ziekten is een samenspel tussen vele genetische loci, genen, celtypen, omgevingsfactoren en weefsels. De genetica van complex genetisch ziekten is uitgebreid bestudeerd door middel van 'genome wide associated studies' (GWASs). Dit heeft geleid tot de associatie van een grote hoeveelheid loci met verschillende ziekten. Het is echter moeilijk gebleken om causale varianten en betrokken genen te identificeren. Hierdoor is een aantal pathologische mechanismen nog niet blootgelegd en blijven sommige potentiële nieuwe therapeutische strategieën uit. In dit proefschrift wordt de rol van regulatoire elementen in de pathogenese van complex genetische ziekten onderzocht, te samen met epitheliale mechanismen die mogelijk betrokken zijn bij inflammatoire darmziekten.

In het eerste gedeelte van dit proefschrift hebben we loci die geassocieerd zijn met inflammatoire darmziekten, cardiovasculaire ziekten en chronische nierziekten in detail bestudeerd. We laten zien dat deze loci vaak co-lokaliseren met regulatoire elementen in het DNA. We hebben 'chromatin conformation capture' gebruikt om de 3D configuratie van deze loci in kaart te brengen. Door middel van deze techniek hebben we nieuwe kandidaat-genen, pathologische mechanismen en cruciale regulatoire factoren geïdentificeerd.

Studies naar de genetica van inflammatoire darmziekten tonen aan dat het darmepitheel een belangrijke rol speelt in de pathogenese van deze complex genetische ziekte. In het tweede gedeelte van dit proefschrift gebruiken we humane intestinale organoïden om de interactie tussen bacteriële antigenen en het darmepitheel te bestuderen. Dit heeft geleid tot de identificatie van stamcellen als bron van de intestinale inflammatoire respons. Tot slot tonen we aan dat de respons van de verschillende epitheelcellen wordt gereguleerd op het niveau van post-transcriptionele regulatie.

Dankwoord

In de wetenschap zoekt men, door te twijfelen, naar zekerheden. Tijdens de afgelopen jaren deed ik soms het tegenovergestelde: ik zette zowel mijn twijfels als mijn zekerheden opzij. Dankzij alle lieve en sterke mensen om mij heen heb ik inmiddels gelukkig weer genoeg zekerheden verzameld om uitgebreid te kunnen twijfelen. Deze mensen wil ik graag met een persoonlijk bericht bedanken en dat zal ik dan ook buiten dit dankwoord doen.

Lieve **Edward**, dankzij jou heb ik mijn onderzoek op mijn eigen(wijze) manier kunnen doen. De marathon die wij samen filosofeerden, kookten, bekookstooften, worstelden, componeerden, bediscussieerden en bewonderden hoop ik nooit meer mee te maken en had ik voor geen goud willen missen.

Dear **Michal**, our discussions and optimistic conversations on the implications of our results are among the most fun and dearest moments of my PhD. I have always felt that the two of us are a remarkable but great team. Thank you for your support, input and patience.

Beste **Leonie**, dank voor alles waarover ik me, dankzij jou, geen zorgen heb hoeven maken.

Lieve **Hankje**, ik kan bij jou altijd rekenen op een warm welkom, een nuchter kijk en goed glas wijn of lekker kopje thee. Ik hoop dat we er nog vele samen zullen drinken. Ik vind het heel fijn dat je vandaag in mijn commissie wilt zitten.

Lieve **Sabine F.**, dank voor je kritische blik en inbreng in mijn werk tot op het laatste moment! Ook wil ik je bedanken voor je interesse in wat mij bezighoudt, je lieve berichtjes en voor je steun de afgelopen jaren.

Dear **Dermot** and **Scott**, thank you for making the effort to be on my committee. I feel honored to have you in Utrecht for this special day and I hope to continue working together in the future.

Beste **Jeffrey**, ik heb de afgelopen jaren genoten van de gezellige momenten bij de koffieautomaat en je scherpe blik langs de zijlijn. Ik vind het een eer dat je als Professor Beekman deel uitmaakt van mijn commissie!

Beste **Prof. Geijssen** en **Prof. Cuppen**, bedankt dat jullie deel uit willen maken van mijn commissie. Ik kijk uit naar de discussie die we zullen hebben.

Lieve **Maaïke**, ik geloof dat ik geen tegenpool van mijzelf ken waar ik zo goed mee samen kan werken als met jou. Dank voor jouw eindeloze inzet en optimisme. Ook jouw proefschrift gaat er komen, daar twijfel ik niet aan!

Beste **Sabine F., Sabine M., Caroline, Jorik, Imre, Suze, Gautam, Marliek en Anke**, bedankt voor jullie input, kritische blik en hulp op duizend-en-één vlakken. Lieve Suze, dank voor alle koffietjes op de momenten waarop ze zo hard nodig waren.

Beste **Beekmannen**, dank voor de gezelligheid op het lab, de inhoudsloze en diepgaande gesprekken in de kweek. **Marne**, dank voor fijne koffiemomenten. **Annelot**, bedankt voor het stellen van vragen die niet iedereen durft te stellen.

Beste **Hemme, Jet, Miguel, Amy, Anne-Claire, Erik en Ninke**; zonder jullie hadden mijn promotie en mijn proefschrift er heel anders uitgezien. Ik heb heel veel van jullie geleerd en genoten van jullie inzet, vragen en ideeën. Ik vind het fantastisch dat alle studenten waarmee ik heb samengewerkt als auteur op één van de artikelen staan: well done!

Arjan, Jorg, Saskia, Maarten, Evelyn, Folkert, Hester, Caroline, Magdalena, Noortje, Nico, colleagues from the Clevers lab and collaborators from the VEO-IBD consortium, thank you for working together, sharing your expertise and for all your valuable input.

Lieve **Carien**, dank voor alle gesprekken die, of ze nou over 4C of de liefde gingen, altijd diepgang hadden.

Lieve Suermannen: **Peter-Paul, Lena, Morsal, Anne en David**. Van intervisie in Utrecht en Krav Maga in Doorn tot Halloween op Rockefeller, ik heb een fantastische tijd met jullie gehad en hoop dat we elkaar op geplande en ongeplande momenten tegen te blijven komen! Lieve **Lisan**, bedankt voor je betrokkenheid en je liefdevolle doortastendheid. Ik heb me gesterkt gevoeld, doordat ik wist dat je altijd voor me klaar stond.

Geny, Monique, Christine en het **Lunter-team**, in het bijzonder **Anne, Aart, Jet en Eliane**, dank voor alles wat ik samen met jullie beleefd en geleerd heb.

Lieve **buren**, ik had me geen fijnere burens kunnen wensen. Dank voor de fijne avonden aan lange tafels en voor de meest gezellige en kleurrijke steeg van de zeven.

Lieve **Aad**, dank voor je enthousiasme en het plezier dat we kijkend en fotograferend hebben gehad. Ik vind het heel bijzonder dat jij ook op deze dag wil fotograferen.

Lieve **Bieneke**, bedankt voor alle virtuele knuffels!

Lieve **Boys**, lieve **Gerwin, Gerjan, Abel, Jasper en Stephan**, bedankt voor de vriendschap die er voor mijn gevoel altijd is en altijd is geweest.

Lieve **Yolande**, bedankt voor je betrokkenheid en de vele fijne avonden in het theater.

Lieve **Carmen**, mooie lieve vrouw! Dank voor je vriendschap en alles wat we samen hebben gedeeld.

Lieve **Gerda**, ik weet niet hoe ik jou kan bedanken voor een jaar lang elke dag een gesproken gedicht. Het meest bijzondere cadeau dat ik ooit kreeg. Dank voor je zoekende kijk op het leven en onze bijzondere vriendschap.

Allerliefste **Knokploegers**, you mean the world to me. Lieve **Redmar**, wanneer ik met jou wandel kom ik altijd op plekken waar ik al eens geweest ben, maar zie ik dingen die ik nooit eerder zag. Bedankt voor je afleiding van wat ik al kende en nieuwsgierigheid naar het onbekende dat voor het oprapen ligt. Lieve **Mantre**, bedankt dat jouw huis voor mij altijd als thuis heeft gevoeld. Jij hebt mij geleerd hoe belangrijk onzinnige dingen zijn en wat een kracht het is als je daarvan kunt genieten. Ik hoop dat we ons hele leven samen nagels lakken, flitspilsjes drinken en onder één dekentje in de bioscoop zullen zitten. Lieve **Edward**, dankjewel voor het samen zoeken zonder vinden, het vinden zonder oordeel en het koken zonder recept. Bedankt dat het goed komt, altijd. Lieve **Carlijn**, dankjewel dat jij ook niet altijd het antwoord weet. Zeker op de vele momenten dat er geen antwoord is, geeft mij dat rust en vertrouwen. Ik geniet van de wereld die we delen, waarover we ons verwonderen en waar we samen de draak mee steken. Lieve **Laura**, bedankt voor alles wat je voor me hebt gedaan en betekent. Ik ben trots dat jij mijn zusje bent. Bedankt dat je er altijd was, ook wanneer het eigenlijk niet kon.

Lieve **Hanske & PP**, ik voel me gezegend met zulke fijne vrienden, zo dichtbij. Lieve Hanske, bedankt voor alle hardlooptrucs die zoveel meer voor me betekenen dan alleen samen sporten. Lieve PP, bedankt voor je wervelende energie en optimisme. Ik vind het fantastisch dat je op deze bijzondere dag naast me staat.

Lieve **familie**, bedankt dat jullie alles zijn wat familie onmisbaar maakt. Ik houd van jullie. Lieve **Aleid**, bedankt voor alle kaartjes, alle diners, concerten en het vele fijne samenzijn. Lieve **Thijs**, bedankt voor de manier waarop jij vragen stelt en mij bewust maakt van de mijn impliciete aannames. Ik ben blij dat jij, Marjo en Joshua weer dichtbij zijn! Lieve **Gilles**, bedankt voor je rustige aanwezigheid en vrolijke noot. Lieve lieve zussen, jullie maken onderdeel uit van wie ik ben en we delen alles wat dat met zich mee brengt. Lieve **Laura**, we lijken meer op elkaar dan ons lief is en kijken bewonderend naar de eigenschappen die we niet delen. Bedankt voor het vertrouwen dat je in mij hebt. Je bent een prachtige sterke vrouw, bij wie de wereld aan haar voeten ligt. Lieve **Charlotte**, bedankt dat je er bent om samen van het leven te genieten, er om te huilen en te lachen. Lieve **Anne**, ik bewonder jou om wie je bent en wie je durft te laten zien. Ik vind je sprankelend, mooi en echt. Bedankt dat je er voor me bent, zonder jezelf te verliezen. Lieve **Marjolein**, bij jou voel ik me altijd fijn. We kennen elkaars wereld en hebben aan een half woord genoeg. Bedankt voor het maken van de lay-out van dit proefschrift; een berg die voor mij niet te overzien was, maar dankzij jou gelukt is en zelfs leuk werd. Ik ben gelukkig dat je op deze bijzondere dag naast me zal staan.

Lieve **pap** en **mam**, jullie hebben me geleerd hoe het voelt om geliefd te zijn en daarmee de basis gelegd voor alle liefdevolle relaties die mijn leven maken tot wat het is. Bedankt voor jullie onuitputtelijke onvoorwaardelijkheid.

Lieve **David**, met jou wil ik zijn.

Curriculum vitae

Claartje Meddens was born on December 24th 1986 in the Netherlands. While studying for a major degree in biomedical sciences, she took multiple courses in a broad range of topics (i.a. philosophy, mechanical engineering, theater arts). After finishing a research master in molecular biology in 2010 she decided to combine a career in research with a clinical career by starting her medical training at the Selective Utrecht Medical Master (SUMMA)-program. After obtaining her medical degree Claartje was awarded an Alexandre Suerman stipend, which enabled her to work full-time on her PhD-research at the lab of Prof. Dr. Nieuwenhuis in the Wilhelmina Children's Hospital and Regenerative Medicine Center Utrecht. Her research focused on pathogenic mechanisms in complex genetic diseases, with special attention for inflammatory bowel disease (IBD). She studied how genomic variants in non coding DNA contribute to the development of IBD, cardiovascular disease and chronic kidney disease by studying the 3D conformation of DNA in cell types relevant for the disease. Furthermore, she has been working on epigenetic and transcriptional processes in the inflamed intestinal epithelium. Claartje is a member of the international Very Early Onset-IBD consortium.

After having finished her thesis, Claartje continues to work as a medical doctor in pediatrics in the St. Antonius hospital in Nieuwegein.

Next to being passionate about her work, Claartje is an enthusiast cook and hosts a 'living room restaurant', loves to listen to music, expresses herself through photography and goes for an early morning run with friends.

List of publications

Claartje A. Meddens, Amy C. J. van der List, Edward E. S. Nieuwenhuis, and Michal Mokry. **2019**. “Non-Coding DNA in IBD: From Sequence Variation in DNA Regulatory Elements to Novel Therapeutic Potential.” *Gut* 68(5):928–41.

Claartje A. Meddens*, Maarten M. Brandt*, Laura Louzao-Martinez, Noortje A. M. van den Dungen, Nico R. Lansu, Edward E. S. Nieuwenhuis, Dirk J. Duncker, Marianne C. Verhaar, Jaap A. Joles, Michal Mokry, and Caroline Cheng. **2018**. “Chromatin Conformation Links Distal Target Genes to CKD Loci.” *Journal of the American Society of Nephrology: JASN* 29(2):462–76.

Claartje A. Meddens*, Haitjema, Saskia*, Sander W. van der Laan, Daniel Kofink, Magdalena Harakalova, Vinicius Tragante, Hassan Foroughi Asl, Jessica van Setten, Maarten M. Brandt, Joshua C. Bis, Christopher O’Donnell, Caroline Cheng, Imo E. Hofer, Johannes Waltenberger, Erik Biessen, J. Wouter Jukema, Pieter A. F. M. Doevendans, Edward E. S. Nieuwenhuis, Jeanette Erdmann, Johan L. M. Björkegren, Gerard Pasterkamp, Folkert W. Asselbergs, Hester M. den Ruijter, and Michal Mokry. **2017**. “Additional Candidate Genes for Human Atherosclerotic Disease Identified Through Annotation Based on Chromatin Organization.” *Circulation: Cardiovascular Genetics* 10(2).

Claartje A. Meddens, Magdalena Harakalova, Noortje A. M. van den Dungen, Hassan Foroughi Asl, Hemme J. Hijma, Edwin P. J. G. Cuppen, Johan L. M. Björkegren, Folkert W. Asselbergs, Edward E. S. Nieuwenhuis, and Michal Mokry. **2016**. “Systematic Analysis of Chromatin Interactions at Disease Associated Loci Links Novel Candidate Genes to Inflammatory Bowel Disease.” *Genome Biology* 17(1):247.

Michal Mokry, Sabine Middendorp, Caroline L. Wiegerinck, Merlijn Witte, Hans Teunissen, Claartje A. Meddens, Edwin Cuppen, Hans Clevers, and Edward E. S. Nieuwenhuis. **2014**. “Many Inflammatory Bowel Disease Risk Loci Include Regions That Regulate Gene Expression in Immune Cells and the Intestinal Epithelium.” *Gastroenterology* 146(4):1040–47.

Gianpiero Spedale, Claartje A. Meddens, Maria J. E. Koster, Cheuk W. Ko, Sander R. van Hooff, Frank C. P. Holstege, H. Th Marc Timmers, and W. W. M. Pim Pijnappel. **2012**. “Tight Cooperation between Mot1p and NC2 β in Regulating Genome-Wide Transcription, Repression of Transcription Following Heat Shock Induction and Genetic Interaction with SAGA.” *Nucleic Acids Research* 40(3):996–1008.

