Gut

Genome-wide association analysis of diverticular disease points towards neuromuscular, connective tissue and epithelial pathomechanisms

Journal:	*Gut*
Manuscript ID	gutjnl-2018-317619.R1
Article Type:	Original Article
Date Submitted by the Author:	n/a

Complete List of Authors:	Schafmayer, Clemens; Kiel University, Department of Visceral and Thoracic Surgery Harrison, James; University of Exeter, United Kingdom, University of Exeter Medical School Buch, Stephan; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden; Technische Universität Dresden (TU Dresden), Center for Regenerative Therapies Dresden (CRTD) Lange, Christina; Kiel University, Institute of Anatomy Reichert, Matthias; University of Homburg, Medical Department II, University Hospital Saarland Hofer, Philipp; Medizinische Universitat Wien Institut fur Krebsforschung, Department of Medicine I Cossais, François; Kiel University, Institute of Anatomy Kupcinskas, Juozas; Institute for Digestive Research, Lithuanian University of Health Sciences von Schönfels, Witigo; Kiel University, Department of Visceral and Thoracic Surgery Schniewind, Bodo; General Hospital Lüneburg Kruis, Wolfgang; Evangelic Hospital Köln-Kalk, Department of Internal Medicine, Gastroenterology and Pulmonology Tepel, Jürgen; Hospital Osnabrück Zobel, Myrko; Helios Hospital Weißeritztal Rosendahl, Jonas; University Clinic Halle, Clinic for Internal Medicine I, Jacobi, Thorsten; Diakonissenanstalt, Hospital Dresden Walther-Berends, Andreas ; Gastroenterology outpatient Center Fördepraxis Schröder, Michael; Center for Gastroenterology and Hepatology Vogel, Ilka; Community Hospital Kiel, Department of Surgery Sergeev, Petr; Hospital Riesa, Department of Internal Medicine II Boedeker, Hans; Hospital Freiberg, Department of Internal Medicine Hinrichsen, Holger; Gastroenterology Center Kiel, Volk, Andreas; Technische Universität Dresden (TU Dresden), Department of Visceral, Thoracic and Vascular Surgery Erk, Jens-Uwe; Diakonissenanstalt, Hospital Dresden Burmeister, Greta; Kiel University, Department of Visceral and Thoracic Surgery

Hendricks, Alexander; Kiel University, Department of Visceral and Thoracic Surgery
Hinz, Sebastian; Kiel University, Department of Visceral and Thoracic Surgery
Wolff, Sebastian; Krankenhaus Duren gGmbH
Böttner, Martina; Kiel University, Institute of Anatomy
Wood, Andrew; University of Exeter, United Kingdom, University of Exeter Medical School
Tyrrell, Jessica; University of Exeter, United Kingdom, University of Exeter Medical School
Beaumont, Robin; University of Exeter, United Kingdom, University of Exeter Medical School
Langheinrich, Melanie; University Hospital Erlangen, Department of Surgery
Kucharzik, Torsten; Stadtisches Klinikum Luneburg gGmbH, Department of Gastroenterology
Brezina, Stefanie; Institute of Cancer Research, Department of Medicine I, Medical University of Vienna
Huber-Schönauer, Ursula; Teaching Hospital of the Paracelsus Private Medical University of Salzburg, Department of Internal Medicine, Hospital Oberndorf
Pietsch, Leonora; Martin-Luther Universität Halle-Wittenberg, Medical Department 1, University Hospital Halle
Noack, Laura; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden
Brosch, Mario; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden; Technische Universität Dresden (TU Dresden), Center for Regenerative Therapies Dresden (CRTD)
Herrmann, Alexander; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden
Thangapandi, Veera; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden
Schimming, Hans; Helios Hospital Weißeritztal, Department of Gastroenterology
Zeissig, Sebastian; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden
Palm, Stefan; Outpatient Center for Gastroenterology
Focke, Gerd; Outpatient Center for Gastroenterology Dresden-Blasewitz
Andreasson, Anna; Stockholm University, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden; Stress Research Institute
Schmidt, Peter Thelin; Karolinska University Hospital, Karolinska Institutet
Weitz, Juergen; University Hospital Dresden, Department of General, Visceral and Thoracic Surgery
Krawczak, Michael; Kiel University, Institute of Medical Informatics and Statistics
Volzke, Henry; University of Greifswald, Institute for Community Medicine
Leeb, Gernot; Hospital Oberpullendorf, Department of Gastroenterology
Michl, Patrick; Martin-Luther-Universitat Halle-Wittenberg Medizinische Fakultat, Dept. of Internal Medicine I
Lieb, Wolfgang; Kiel University, Institute of Epidemiology and Biobank POPGEN
Grützmann, Robert; Universitätsklinikum Erlangen, Germany, Department of Surgery
Franke, Andre; Institute for Clinical Molecular Biology
Lammert, Frank; University of Homburg, Medical Department II, University Hospital Saarland
Becker, Thomas; Kiel University, Department of Visceral and Thoracic Surgery

	Kupcinskas, Limas; Lithuanian University of Health Sciences, Department of Gastroenterology and Institute for Digestive Research D'Amato, Mauro; Karolinska Institute, Department of Medicine Solna and Centre for Molecular Medicine Wedel, Thilo; Kiel University, Institute of Anatomy Datz, Christian; Hospital Oberndorf, Internal Medicine Gsur, Andrea; Institut of Cancer Research , Department of Medicine I, Medical University of Vienna Weedon, Mike; University of Exeter, Precision Medicine, Exeter Hampe, Jochen; Technische Universität Dresden (TU Dresden), Medical Department 1, University Hospital Dresden; Technische Universität Dresden (TU Dresden), Center for Regenerative Therapies Dresden (CRTD)
Keywords:	DIVERTICULAR DISEASE, INTESTINAL MOTILITY, GENETIC POLYMORPHISMS

TITLE PAGE

word count: 3994
Introduction/Methods/Results/Discussion: 3994

excluded from word count:
Abstract: 249 words; Significance of the study: 257 words; Keywords, Abbreviations, Acknowledgements: 312 words; Figure and Table headings

Genome-wide association analysis of diverticular disease points towards neuromuscular, connective tissue and epithelial pathomechanisms

Clemens Schafmayer[1*], James W. Harrison[2*], Stephan Buch[3,4*], Christina Lange[5], Matthias C. Reichert[6], Philipp Hofer[7], François Cossais[5], Juozas Kupcinskas[8], Witigo von Schönfels[1], Bodo Schniewind[9], Wolfgang Kruis[10], Jürgen Tepel[11], Myrko Zobel[12], Jonas Rosendahl[13], Thorsten Jacobi[14], Andreas Walther-Berends[15], Michael Schroeder[16], Ilka Vogel[17], Petr Sergeev[18], Hans Bödeker[19], Holger Hinrichsen[16], Andreas Volk[20], Jens-Uwe Erk[13], Greta Burmeister[1], Alexander Hendricks[1], Sebastian Hinz[1], Sebastian Wolff[10], Martina Böttner[5], Andrew R. Wood[2], Jessica Tyrrell[2], Robin N. Beaumont[2], Melanie Langheinrich[21], Torsten Kucharzik[9], Stefanie Brezina[7], Ursula Huber-Schönauer[22], Leonora Pietsch[13], Laura Sophie Noack[3], Mario Brosch[3], Alexander Herrmann[3], Raghavan Veera Thangapandi[3], H.-Wolfgang Schimming[12], Sebastian Zeissig[3], Stefan Palm[23], Gerd Focke[24], Anna Andreasson[25], Peter Thelin Schmidt[26], Jürgen Weitz[20], Michael Krawczak[27], Henry Völzke[28], Gernot Leeb[29], Patrick Michl[13], Wolfgang Lieb[30], Robert Grützmann[21], Andre Franke[31], Frank Lammert[6], Thomas Becker[1], Limas Kupcinskas[8], Mauro D'Amato[26], Thilo Wedel[5§], Christian Datz[22§], Andrea Gsur[7§], Michael N Weedon[2§], Jochen Hampe[3,4§#]

[*]C.S., J.W.H. and S.B. contributed equally to the study and the manuscript,

[§]T.W., C.D., A.G., M.W. and J.H. contributed equally to the study and the manuscript and assume equal responsibility for senior authorship

[1]Department of Visceral and Thoracic Surgery, Kiel University, Kiel, Germany
[2]University of Exeter Medical School, University of Exeter, United Kingdom
[3]Medical Department 1, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany
[4]Center for Regenerative Therapies Dresden (CRTD), Technische Universität Dresden (TU Dresden), Dresden Germany
[5]Institute of Anatomy, Kiel University, Kiel, Germany
[6]Medical Department II, University Hospital Saarland, Homburg, Germany
[7]Institute of Cancer Research, Department of Medicine I, Medical University Vienna, Vienna, Austria
[8]Department of Gastroenterology and Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas, Lithuania
[9]General Hospital Lüneburg, Germany
[10]Department of Internal Medicine, Gastroenterology and Pulmonology, Evangelic Hospital Köln-Kalk, Cologne, Germany
[11]Department of General and Thoracic Surgery, Hospital Osnabrück, Osnabrück, Germany
[12]Department of Gastroenterology, Helios Hospital Weißeritztal, Freital, Germany
[13]Medical Department 1, University Hospital Halle, Martin-Luther Universität Halle-Wittenberg, Halle, Germany
[14]Diakonissenanstalt, Hospital Dresden, Dresden, Germany
[15]Gastroenterology outpatient Center Fördepraxis, Kiel, Germany
[16]Center for Gastroenterology and Hepatology, Kiel, Germany
[17]Department of Surgery, Community Hospital Kiel, Kiel, Germany
[18]Department of Internal Medicine II, Hospital Riesa, Riesa, Germany
[19]Department of Internal Medicine, Hospital Freiberg, Germany
[20]Department of Visceral, Thoracic and Vascular Surgery, Technische Universität Dresden (TU Dresden), Dresden, Germany
[21]Department of Surgery, University Hospital Erlangen, Erlangen, Germany

[22]Department of Internal Medicine, Hospital Oberndorf, Teaching Hospital of the Paracelsus Private Medical University of Salzburg, Oberndorf, Austria
[23]Outpatient Center for Gastroenterology, Dippoldiswalde, Germany
[24]Outpatient Center for Gastroenterology Dresden-Blasewitz, Dresden, Germany
[25]The Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden; Stress Research Institute, Stockholm University, Stockholm, Sweden
[26]Department of Medicine Solna and Centre for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden
[27]Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany
[28]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
[29]Department of Gastroenterology, Hospital Oberpullendorf, Burgenland, Oberpullendorf, Austria
[30]Institute of Epidemiology & Popgen Biobank, Kiel University, Kiel, Germany
[31]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

[#]to whom correspondence should be addressed: Jochen Hampe, MD, Medical Department 1, Genomic Gastroenterology & Hepatology, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Fetscherstraße 74, 01307 Dresden / Germany, Phone +49-351-458 5643, Email: Jochen.Hampe@uniklinikum-dresden.de

Keywords: Diverticular disease, diverticulitis, GWAS, genetic association

Abbreviations:

GWAS - genome-wide association study

OR - odds ratio

CI - confidence interval

qPCR - quantitative polymerase chain reaction

RT-PCR - reverse transcription polymerase chain reaction

ICD - International Statistical Classification of Diseases and Related Health Problems

IBD - inflammatory bowel disease

IBS - irritable bowel syndrome

FUMA - FUnctional Mapping and Annotation of genetic associations

eQTL - expression quantitative trait loci

GTEx - Genotype-Tissue Expression

GARFIELD - GWAS analysis of regulatory of functional information enrichment with LD correction

ENCODE - ENCyclopedia Of DNA Elements

GENCODE - reference human genome annotation for The ENCODE Project

VEGAS2 - Versatile Gene-based Association Study - 2

ABSTRACT

Objective: Diverticular disease is a common complex disorder characterized by mucosal outpouchings of the colonic wall that manifests through complications such as diverticulitis, perforation and bleeding. We report the to date largest genome-wide association study (GWAS) to identify genetic risk factors for diverticular disease.

Design: Discovery GWAS analysis was performed on UK Biobank imputed genotypes using 31,964 cases and 419,135 controls of European descent. Associations were replicated in a European sample of 3,893 cases and 2,829 diverticula-free controls and evaluated for risk contribution to diverticulitis and uncomplicated diverticulosis. Transcripts at top 20 replicating loci were analyzed by real-time qPCR in preparations of the mucosal, submucosal and muscular layer of colon. The localization of expressed protein at selected loci was investigated by immunohistochemistry.

Results: We discovered 48 risk loci, of which 12 are novel, with genome-wide significance and consistent odds ratio in the replication sample. Nominal replication ($p<0.05$) was observed for 27 loci, and additional 8 in meta-analysis with a population-based cohort. The most significant novel risk variant rs9960286 is located near *CTAGE1* with a p-value of 2.3×10^{-10} and 0.002 ($OR_{allelic}$=1.14 [1.05-1.24]) in the replication analysis. Four loci showed stronger effects for diverticulitis, *PHGR1* (OR 1.32, CI 95% 1.12-1.56), *FAM155A-2* (OR 1.21, 95% CI 1.04-1.42), *CALCB* (OR 1.17, 95% CI 1.03-1.33) and *S100A10* (OR 1.17, 95% CI 1.03-1.33).

Conclusion: In silico analyses point to diverticulosis primarily as a disorder of intestinal neuromuscular function and of impaired connective fiber support, while an additional diverticulitis risk might be conferred by epithelial dysfunction.

SUMMARY BOX

What is already known on this subject?

- Diverticular disease is among the most common diseases of the gastrointestinal tract.
- Despite its clinical importance and clear indications of familial clustering, only three loci (*ARHGAP15*, *FAM155A*, *COLQ*) of genome-wide significance have been reported so far. Recently, a replication analysis of a UK biobank GWAS by Maguire *et al.* identified 37 additional susceptibility loci with genome-wide significance and a replication of 8 of these loci in a Michigan population cohort.

What are the new findings?

- Here, we report the to date largest and most detailed genome-wide association study (GWAS) with a sample size of 451,099 individuals to identify genetic risk factors for diverticular disease.
- We report 48 loci with genome wide significance, of which 12 are novel. We were able to replicate 27 of these loci in specifically recruited replication samples from a gastrointestinal specialty service with colonoscopy data available in all controls. In addition, we replicated further 8 risk loci in a combined meta-analysis with data from a Michigan population cohort.
- The current study increases the number of replicated susceptibility loci for diverticular disease to 35, of which 25 loci had previously not been replicated.
- Results point to diverticular disease primarily as a disorder of intestinal neuromuscular function, impaired mesenteric vascular smooth muscle function and of impaired connective fiber support. Whilst diverticulitis risk might be conferred by epithelial dysfunction.

How might it impact on clinical practice in the foreseeable future?

- The results from this GWAS provide deep new insights into the colonic biology and disease pathophysiology of diverticular disease.

Diverticular disease is a common complex disorder characterized by mucosal outpouchings of the colonic wall at sites of relative weakness in the muscle layers close to penetrating blood vessels [1,2]. The incidence of diverticular disease has increased to 50% for individuals older than 60 years and a significant rise of incidence and hospitalization rates has been seen in younger age groups [3]. Although the majority of patients harboring diverticula remain asymptomatic throughout life, 10–25 % [4–8] experience complications such as acute diverticulitis, abscess, fistula formation, bleeding or perforation. These complications cause an annual mortality of ~1 per 100,000 [9] due to the need for inpatient treatment and sigmoid resection after repeated episodes of diverticulitis. Owing to its high prevalence and associated complications diverticular disease is the 5th most costly gastrointestinal disease in Western countries [10].

The pathogenesis of diverticular disease is thought to be a multifactorial process that involves lifestyle factors (smoking, physical inactivity, high body mass index [BMI]), structural and functional changes of the colonic wall, aging and a genetic predisposition [11]. In contrast to its high clinical and economic impact, diverticular disease is under-researched in terms of its pathophysiology [1]. Epidemiological [12] and twin studies [13] have estimated the heritability of diverticular disease at 40-53% percent. A previous genome-wide association study (GWAS) from Iceland identified associations of variants in *ARHGAP15* and *COLQ* with uncomplicated diverticular disease and variants in *FAM155A* with diverticulitis [14]. Additionally 37 susceptibility loci with genome-wide significance were identified in a recent study from Maguire *et al.* [15], with replication of 8 loci.

We report a total of 48 risk loci with genome-wide significance and consistent odds ratio in a replication sample of 3,893 cases and 2,829 diverticula-free controls as verified by colonoscopy. We were able to replicate 27 of these loci in specifically recruited replication samples from a gastrointestinal specialty service with colonoscopy data available in all controls. The large number of loci we identified and our functional follow-up provide novel insight into the pathophysiology of diverticular disease as a disorder of intestinal neuromuscular function, vascular smooth muscle function and impaired connective fibre support.

PATIENTS AND METHODS

Study participants

An individual was classified as a diverticular disease case if they matched hospital-bases ICD9 or ICD10 coding (562, K57) in the UK Biobank dataset (n = 31,964). Control individuals where classified on the basis of absence of a diverticular disease diagnosis (n = 419,135). Depth of ICD coding was insufficient to differentiate disease subtype diverticulosis (i.e. diverticular disease without inflammation) from diverticulitis in the UK Biobank dataset. Replication samples were obtained from Germany, Austria, Lithuania and Sweden from gastrointestinal specialty services. Details of recruitment and phenotype ascertainment for diverticulosis and diverticulitis for each cohort are described in the Supplementary Materials and Methods section. An overview of the study population is provided in Table 1.

Table 1: Study populations

Overview of the study populations used in the discovery and replication cohorts. All quantitative measures (age, BMI) are provided as medians and interquartile ranges. Patients from the Germany/North cohort were recruited through the popgen biobank as described previously [16].

Variable		Discovery	European replication cohort n= 6722 (Diverticular disease (DD) 3893, diverticula free controls 2829)					
		UKBB - 500k GWAS (n=451,052)	Germany (North) (n=2842)	Germany (West) (n=320)	Austria (Vienna) (n=1151)	Austria (Oberndorf) (n=1378)	Lithuania (n=773)	Sweden (n=258)
		DD / Controls 31,917 / 419,135	DD / Controls 2171/671	DD / Controls 227/93	DD / Controls 578/573	DD / Controls 387/991	DD / Controls 479/295	DD / Controls 51/207
male (%)	Case	46.5	46.9	54.6	60.4	57.4	43.0	39.2
	Control	45.7	54.7	48.4	58.1	48.5	39.5	37.2
Age (years)	Case	72 (68-76)	68 (59-75)	67 (58-75)	68 (61-74)	65 (57-72)	68 (61-74)	62 (56-67)
	Control	68 (60-73)	62 (57-68)	59 (44-64)	64 (56-72)	59 (53-67)	57 (46-67)	53 (44-61)
BMI	Case	27.9 (25-31)	26.9 (24-30)	27.1 (24-30)	28.1 (25-30)	28.0 (23-33)	28.6 (26-32)	24.0 (22-27)
	Control	26.6 (24-29)	25.7 (23-28)	27.0 (24-30)	27.4 (24-30)	27.3 (21-33)	26.6 (23-30)	25.0 (22-28)

GWAS analysis

Discovery GWAS analysis was performed on UK Biobank on Version 3 imputed genotypes using BOLT-LMM v2.34, which applies a linear mixed model to adjust for the effects of population structure and individual relatedness [17]. This enabled the inclusion of all related individuals in our white European subset allowing a sample size of 451,099 individuals as detailed in Supplementary Material and Methods.

Loci Discovery and Functional Annotation (FUMA)

Genomic risk loci, lead variants and candidate SNPs were derived from **FU**nctional **M**apping and **A**nnotation of genetic associations (FUMAv1.3.1) [18] based on GWAS summary statistics. Candidate SNP and gene positions are provided in Supplementary Table 1 and 2. Functional consequences were assessed using ANNOVAR, a tissue-specific *cis*-eQTL dataset (GTExV7, https://gtexportal.org) and 15-core chromatin states (ENCODE, 2012) as detailed in the Supplementary Material and Methods section.

Annotation of candidate genes

In order to identify candidate gene(s) at the respective genomic risk locus we followed i) a manually curated selection process based on local LD structure and supporting evidence from regulatory elements (eQTL and chromatin interaction), outlined in Supplementary Table 3 and ii) we performed hypothesis-free functional and gene annotations based on the genomic positions of risk loci using FUMA [18], as the manually curated selection process of candidate genes might not capture the full biology of the risk architecture, as detailed in Supplementary Material and Methods.

Replication genotyping and meta-analysis

Top GWAS associated loci (n = 51; $P<5\times10^{-8}$) were validated in a combined European sample of 3,893 cases and 2,829 diverticula-free controls based on colonoscopy (Table 1) using the most significant discovery variant or appropriate proxies when direct genotyping of a lead variant was not technical feasible. Logistic regression analyses were performed with PLINK [19], cohort-specific β effect estimates were combined with META [20]. For replication a nominal significance level of

$P<0.05$ and consistency in odds ratio direction between the discovery and replication stage was applied. Additional replication was achieved by including replication data presented by Maguire et al. [15] (Supplementary Table 4) from European samples (N=29,367) from the Michigan genome initiative (MGI) into a combined meta-analysis of all European replication cohorts (N=36,089 samples). Details on the genotyping, quality control and meta-analysis are provided in the Supplementary Materials and Methods section.

Table 2: GWAS and replication results: newly discovered and novel replicated risk loci

Results of GWAS analysis in diverticular disease. The column "Loci overlap" indicates an overlap of the respective risk loci to a risk locus recently identified in a GWAS by Maguire et al. [15]. The corresponding risk locus number is given as (*Mag*.#1-82), with bold print indicating prior attainment of genome-wide significance. Current GWAS risk loci are numbered descending by the P-value in the discovery analysis. Ranked discovery GWAS and replication tables are provided as Supplementary Table 5 and 6. Results are structured showing a) newly discovered diverticular disease risk loci with genome-wide significance replicated in European samples with diverticula-free controls (Table 1) and b) replicated in a meta-analysis of European samples of the current study and Michigan samples with population controls (MGI) and c) newly discovered diverticular disease risk loci currently lacking replication and d,e) showing novel replicated, previously discovered (Maguire et al.) diverticular disease risk loci replicated c) in European samples and d) replicated in a meta-analysis of European samples of the current study and Michigan samples with population controls (MGI). Bold gene symbols and bold *P* repl. indicate replication using an FDR of 0.1 after Benjamin-Hochberg correction. The lead candidate gene annotation corresponds to the curated candidate gene(s), which selection is described in Supplementary Material and Methods and in Supplementary Table 3. Candidate genes that harbor variants in LD (r^2 >0.8) to the respective lead variant at genomic risk locus are indicated with "LD", additional candidate genes with variants with $P < 1.0 \times 10^{-5}$ and r2 > 0.6 to independent significant lead variants are marked with an asterix(*). At intergenic annotated risk loci, if the lead variant or proxy variants are not mapping to a specific gene, closest neighboring candidate genes (<1 cM distance) are marked with (**). Candidate genes with eQTL variants affecting gene expression in sigmoid colon at FDR<0.05 or at nominal p_eQTL <0.05[1] are shown (data from GTExV7). Additional candidate genes mapped by 3D chromatin interactions are listed in Supplementary Table 2 for each risk locus. Rs-IDs of replication SNP which are proxies for the discovery variant are marked with an asterisk (*) and pairwise LD (r^2) to the discovery variant is provided. SNPs genotyped by TaqMan rather than iPLEX are indicated by a pound (#) sign. Variants at the *FAM155A-1* and *FAM155A-2* (Table 3) were in low LD (r^2=0.0043) and thus considered as individual loci. Odds ratios are based on the reference allele (RA). Reference allele frequencies (RAF%) are provide for cases/controls in the discovery GWAS. I^2 measure of the percentage of between-cohort heterogeneity. The direction of obtained odds ratio (OR dir.) between discovery and replication analysis is the consitent for all reported loci. The positions of lead variants where annotated according to Genome Reference Consortium Human Build 37 patch release 13. Gene annotation are based on RefSeq curated gene predictions from NCBI; pseudogenes were excluded from annotation.

Discovery GWAS in UK Biobank diverticular disease (DD) 31917, controls (CON) 419135											**Replication** **in European samples** diverticular disease (DD) 3893, diverticula free controls (CON) 2829						**with MGI** DD:4611 CON:31478	
Locus	Chr	Position	Lead variant	Lead variant location	Closest gene	Candidate genes at risk locus: **Lead candidate gene(s)**	Loci overlap	*P* GWAS	RA	RAF Ca|Co	rsID	RA	r^2	*P* repl.	OR (CI 95%)	OR dir.	I^2	*P* repl.
a) newly discovered, replicated diverticular disease risk loci																			
27	15	76826003	rs2056544	intronic	***SCAPER***	LD, eQTL: ***SCAPER***, *LOC101929439 (RP11-593F23.1)*, LD: *EFTA, C15orf27, ISL2*, RP11-593F23.1*, RCN2**	*Mag.#47*	9.8×10^{-11}	G	41.2|42.3	rs12443137*	G	1	**0.039**	0.93 (0.86-1.00)	same	43	**0.006**
29	18	20028737	rs9960286	intergenic	***CTAGE1***	***CTAGE1****	*Mag.#44*	2.3×10^{-10}	G	25.4|24.4	rs2009593*	G	0.97	**0.002**	1.14 (1.05-1.24)	same	0	**0.001**
37	22	40695172	rs6001870	intronic	***TNRC6B***	LD: ***TNRC6B***	*Mag.#82*	4.3×10^{-9}	C	36.1|35.1	rs5995842*#	G	0.98	**0.025**	1.09 (1.01-1.18)	same	2	**0.039**
40	4	15386383	rs4132788	intronic	***C1QTNF7***	LD, eQTL: ***C1QTNF7*** [1], LD: *RP11-665G4.1*	*Mag.#53*	1.1×10^{-8}	T	27.0|26.0	rs4515160*#	G	1	**0.019**	1.10 (1.02-1.20)	same	35	**0.046**
49	15	68238462	rs387505	intergenic	***PIAS1***	LD, eQTL: ***PIAS1*** [1]**, LD: *AC009292.2**, SKOR1***	*Mag.#54*	2.9×10^{-8}	T	44.9|43.8	rs387505#	T	1	**0.009**	1.10 (1.02-1.18)	same	0	**0.017**
b) newly discovered, replicated diverticular disease risk loci – replicated in a meta-analysis with data from MGI																			
41	5	122329729	rs34126945	intronic	***SNX24***	LD: ***SNX24***, *PPIC, SNX2, AC008669.1**	*Mag.#41*	1.2×10^{-8}	G	32.3|33.5	rs34126945	G	1	0.079	0.93 (0.86-1.01)	same	47	**0.004**
c) newly discovered diverticular disease risk loci – currently not replicated																			
35	17	76856966	rs1973232	intronic	*TIMP2*	LD: ***TIMP2***	*Mag.#48*	2.9×10^{-9}	G	19.3|18.5	rs9909232	A	0.92	0.764	1.01 (0.92-1.11)	same	0	0.401
38	20	37493576	rs208814	intronic	*PPP1R16B*	LD: ***PPP1R16B***, eQTL: ***FAM83D***	*Mag.#75*	9.0×10^{-9}	A	36.6|35.5	rs208814#	A	1	0.637	1.02 (0.94-1.10)	same	0	0.461
42	6	32609965	rs7990	exon	*HLA-DQA1*	LD: ***HLA-DQA1***, *BTNL2*, HLA-DRB9*, BTNL2** eQTL: *HLA-DQA2, HLA-DRB1, HLA-DOB, BAG6;*	N/A	1.4×10^{-8}	A	10.5|9.9	rs2395163*#	C	0.25	0.739	1.02 (0.92-1.12)	same	59	N/A
43	10	124168942	rs139760870	intronic	*PLEKHA1*	LD: ***PLEKHA1***, *HTRA1*, BTBD16**	*Mag.#52*	1.4×10^{-8}	A	5.4|5.9	rs117811194*	A	1	0.318	0.91 (0.76-1.09)	same	0	0.181
47	2	33361425	rs6714546	intronic	*LTBP1*	LD, eQTL: ***LTBP1*** [1]	*Mag.#52*	2.4×10^{-8}	A	30.1|29.1	rs6714546#	A	1	0.552	1.03 (0.94-1.11)	same	0	N/A
48	13	33727605	rs1473813	intronic	*STARD13*	LD: ***STARD13***	*Mag.#64*	2.9×10^{-8}	A	38.0|39.2	rs1473813#	A	1	0.507	0.98 (0.90-1.05)	same	34	0.276
d) novel replicated, previously discovered (Maguire *et al.*) diverticular disease risk loci																			

7	19	38738130	rs4802297	intergenic	***PPP1R14A***	LD, eQTL: ***PPP1R14A,*** *C19orf33*, LD: *SPINT2*	***Mag.#6***	2.0×10^{-16}	G	49.8|48.2	rs12976534*#	G	1	**0.029**	1.08 (1.01-1.17)	same	15	**0.004**
9	7	102474903	rs72221075	intronic	***FBXL13***	LD: ***FBXL13,*** *FAM185A, LRRC17*	***Mag.#8***	7.7×10^{-15}	(-)	35.2|33.6	rs10257317*#	C	0.93	**3.9×10^{-4}**	1.14 (1.06-1.23)	same	0	**9.23×10^{-4}**
10	13	107897823	rs9520339	intronic	***FAM155A_1***	LD: ***FAM155A***	***Mag.#10***	1.1×10^{-14}	T	22.7|24.0	rs9520344*#	A	1	**0.018**	0.90 (0.83-0.98)	same	0	**0.010**
14	8	120456193	rs60869342	intergenic	***NOV***	LD, eQTL: ***NOV,*** *LD: ENPP2**	***Mag.#21***	4.4×10^{-13}	T	23.0|24.2	rs1381335*#	T	0.61	**3.0×10^{-4}**	0.85 (0.78-0.93)	same	21	**0.001**
15	21	47399453	rs111316530	intergenic	***COL6A1***	LD: ***COL6A1,*** *COL6A2*, AL592528.1, PCBP3*, AL133493.2*, FTCD**	***Mag.#14***	4.8×10^{-13}	(-)	15.5|14.5	rs7281388*	A	0.96	**0.009**	1.16 (1.04-1.30)	same	0	**0.031**
17	16	86233413	rs2280028	intergenic	***LINC01082***	LD: ***LINC01082***	***Mag.#15***	4.1×10^{-12}	A	13.2|14.2	rs2280028	A	1	**0.019**	0.88 (0.79-0.98)	same	0	**0.004**
19	6	98364895	rs9482094	intronic	***LOC101927314***	LD: ***LOC101927314*** *(RP11-436D23.1)*	***Mag.#19***	1.6×10^{-11}	A	35.9|37.2	rs4839715*#	A	1	**0.022**	0.91 (0.85-0.99)	same	0	**0.010**
20	3	151074941	rs3732760	intronic	***P2RY12***	LD: ***P2RY12,*** *P2RY14, MED12L, GPR87, P2RY13**	***Mag.#16***	1.7×10^{-11}	C	38.7|37.3	rs3732760	C	1	**0.026**	1.09 (1.01-1.18)	same	0	**0.049**
22	11	15065235	rs575909118	intergenic	***CALCB***	LD: ***CALCB,*** *CALCA*	***Mag.#25***	2.8×10^{-11}	T	28.6|27.5	rs12293178*	A	0.95	**0.020**	1.10 (1.02-1.19)	same	27	**0.022**
23	1	151970629	rs61814883	intergenic	***S100A10***	LD, eQTL: ***S100A10,*** *THEM4*	***Mag.#22***	3.2×10^{-11}	A	28.6|29.9	rs61814883	A	1	**0.003**	0.89 (0.82-0.96)	same	0	**0.030**
25	15	40649609	rs71472433	intergenic	***DISP2***	LD, eQTL: ***PHGR1***[1]***, DISP2***	***Mag.#17***	6.1×10^{-11}	C	17.6|16.6	rs71472433	C	1	**0.014**	1.14 (1.03-1.27)	same	0	**0.038**
31	10	18440444	rs1888693	intronic	***CACNB2***	LD: ***CACNB2***	***Mag.#29***	2.7×10^{-10}	A	33.2|34.3	rs1888693	A	1	**0.003**	0.89 (0.82-0.96)	same	0	**0.006**
45	1	221066373	rs2784255	intergenic	***HLX***	LD: ***HLX,*** *HLX-AS1*, eQTL: LINC01352 (RP11-295M18.2)*	***Mag.#34***	2.0×10^{-8}	C	47.3|48.5	rs2784255	C	1	**0.023**	0.92 (0.85-0.99)	same	0	**0.041**
46	3	5843836	rs7624168	intergenic	***EDEM1***	neighboring genes: ***EDEM1****, *GRM7, GRM7-AS3*	***Mag.#39***	2.3×10^{-8}	A	21.5|22.6	rs4684509*	G	1	**0.036**	0.91 (0.83-0.99)	same	8	0.389
e) novel replicated, previously discovered (Maguire *et al.*) diverticular disease risk loci – replicated in a meta-analysis with data from MGI																		
5	2	56093204	rs1802575	3´UTR	***EFEMP1***	LD: ***EFEMP1,*** eQTL: *RPS27A*	***Mag.#5***	3.7×10^{-19}	C	14.5|13.3	rs1802575	C	1	0.111	1.10 (0.98-1.25)	same	47	**0.030**
12	1	219294570	rs61823192	intronic	***LYPLAL1-AS1***	LD: ***LYPLAL1-AS1*** *(RP11-135J2.4), LYPLAL1*	***Mag.#9***	4.6×10^{-14}	T	2.5|3.0	rs61823192#	T	1	0.082	0.80 (0.62-1.03)	same	63	**0.003**
13	10	101391169	rs7098322	intergenic	***SLC25A28***	LD: ***SLC25A28,*** *COX15, ENTPD7, CUTC*	***Mag.#11***	6.0×10^{-14}	C	13.4|12.5	rs7091203*	A	1	0.076	1.11 (0.99-1.25)	same	33	**0.043**
21	5	64295363	rs10471645	intronic	***CWC27***	LD: ***CWC27***	***Mag.#28***	2.1×10^{-11}	T	17.5|16.5	rs2968205*	A	1	0.579	1.03 (0.93-1.14)	same	0	**0.046**
30	17	42312778	rs8074740	intergenic	***SLC4A1***	LD: ***UBTF,*** *ASB16, C17orf53, TMUB2, ATXN7L3, SLC4A1, AC003102.1, HDAC5;* eQTL: *ASB16-AS1*	***Mag.#23***	2.4×10^{-10}	A	33.4|32.1	rs4793086*	C	0.98	0.077	1.07 (0.99-1.16)	same	38	**0.019**

Table 3: GWAS and replication results: confirmed, previously replicated risk loci and currently not replicated risk loci

Results of GWAS analysis in diverticular disease. Overlap of risk loci to a previous GWAS by Maguire et al. [15] and Sigurdsson et al. [14] with the corresponding risk locus is indicated. Current GWAS risk loci are numbered descending by the P-value in the discovery analysis. Results are structured by showing a) previously (Maguire et al.; Sigurdsson et al.) replicated diverticular disease risk loci replicated (confirmed) in European samples with diverticula-free controls (by colonoscopy, Table 1) of the current study or b) confirmed in a meta-analysis of European samples of the current study and Michigan samples with population controls (MGI) and c) previously discovered (Maguire et al.) diverticular disease risk loci lacking replication. The direction of obtained odds ratio (OR dir.) between discovery and replication analysis is the same for all loci. Table headings are identical to those in Table 2.

Discovery GWAS in UK Biobank diverticular disease (DD) 31917, controls (CON) 419135											**Replication** **in European samples** diverticular disease (DD) 3893, diverticula free controls 2829						**with MGI** DD:4611 CON:31478		
Locus	Chr	Position	Lead variant	Lead variant location	Closest gene	Candidate genes at risk locus: **Lead candidate gene(s)**	Loci overlap	*P* GWAS	RA	RAF Ca|Co	rsID	RA	r^2	*P* repl.	OR (CI 95%)	OR dir.	I^2	*P* repl.
a) confirmed, previously (Maguire et al.; Sigurdsson et al.) replicated diverticular disease risk loci																				
1	2	144314247	rs6734367	intronic	***ARHGAP15***	LD: ***ARHGAP15***	***Mag.#1; Sigur.#1***	4.4×10^{-55}	T	20.1|17.8	rs6734367#	T	1	**1.5×10^{-7}**	1.29 (1.17-1.42)	same	0	**1.11×10^{-7}**
3	10	25819228	rs7077800	intronic	***GPR158***	LD, eQTL: ***GPR158***	***Mag.#4***	1.7×10^{-22}	T	45.7|47.6	rs7086249*	C	1	**0.003**	0.90 (0.83-0.96)	same	0	**7.24×10^{-6}**
4	3	15502681	rs7609897	intronic	***COLQ***	LD: ***COLQ***, *METTL6, HACL1, EAF1**	***Mag.#3; Sigur.# 2***	5.6×10^{-22}	T	19.9|21.4	rs7609897#	T	1	**0.009**	0.89 (0.81-0.97)	same	24	**2.81×10^{-4}**
8	9	136149229	rs505922	intronic	***ABO***	LD, eQTL: ***ABO***	***Mag.#13***	4.5×10^{-15}	C	30.2|31.8	rs687621*	G	0.96	**0.011**	0.91 (0.84-0.98)	same	0	**3.63×10^{-5}**
11	13	108215404	rs9555371	intronic	***FAM155A_2****	LD: ***FAM155A***	***Mag.#10 Sigur.#3***	1.2×10^{-13}	G	18.7|19.8	rs9555371#	G	1	**3.5×10^{-7}**	0.79 (0.72-0.87)	same	55	**1.30×10^{-7}**
16	7	73427600	rs112609918	intergenic	***ELN***	LD: ***ELN***	***Mag.#27***	2.9×10^{-12}	T	5.2|4.6	rs112609918*	T	1	**0.004**	1.39 (1.11-1.74)	same	0	**0.011**

32	4	95821419	rs3775010	intronic	***BMPR1B***	LD: ***BMPR1B***	***Mag.#32***	6.8×10^{-10}	C	37.0|35.8	rs972409*#	T	1	**0.021**	1.10 (1.01-1.19)	same	0	**7.95×10^{-5}**
36	11	70005374	rs875107	intronic	***ANO1***	LD: ***ANO1***, *RP11-805J14.3*	***Mag.#26***	3.7×10^{-9}	C	49.6|48.2	rs2276068*#	C	1	**0.001**	1.14 (1.05-1.23)	same	0	**3.13×10^{-4}**
b) confirmed, previously replicated (Maguire et al.) diverticular disease risk loci – replicated in a meta-analysis with data from MGI																		
2	1	234352899	rs4333882	intronic	***SLC35F3***	LD: ***SLC35F3***	***Mag.#2***	2.5×10^{-24}	G	20.7|19.2	rs4333882	G	1	0.106	1.08 (0.98-1.17)	same	28	**0.035**
28	7	96078564	rs3113037	intergenic	***SEM1***	LD: ***SEM1***, *SHFM1**	***Mag.#24***	1.0×10^{-11}	T	24.4|23.2	rs3113037	T	1	0.070	1.08 (0.99-1.18)	same	50	**0.002**
c) previously discovered (Maguire et al.) diverticular disease risk loci with genome-wide significance – currently not replicated																		
6	11	27748493	rs17309930	intergenic	*AC103796.1*	LD: ***BDNF***, *BDNF-AS, LIN7C*	***Mag.#7***	6.8×10^{-17}	A	19.5|20.8	rs962369*#	C	0.50	0.335	0.96 (0.88-1.04)	same	0	0.292
18	16	84857378	rs2131755	intronic	*CRISPLD2*	LD: ***CRISPLD2***	***Mag.#18***	1.5×10^{-11}	G	42.1|40.7	rs2131755	G	1	0.095	1.07 (0.99-1.15)	same	0	0.085
24	5	37772780	rs10472291	intergenic	*WDR70*	eQTL: ***GDNF***[**kl]; LD: *WDR70*	***Mag.#12***	3.8×10^{-11}	A	34.6|33.2	rs10472291#	A	1	0.143	1.06 (0.98-1.14)	same	33	0.263
33	8	122259074	rs4871180	intergenic	*HAS2*	neighboring genes: ***HAS2*****, *SNTB1***	***Mag.#30***	1.1×10^{-9}	T	25.6|24.4	rs4871180	T	1	0.364	1.04 (0.96-1.13)	same	4	0.308
34	9	78739440	rs147496465	intronic	*PCSK5*	LD: ***PCSK5***	***Mag.#35***	1.2×10^{-9}	(-)	45.6|46.9	rs7035893*	C	0.98	0.818	0.99 (0.92-1.07)	same	31	0.808
39	2	18937283	rs62125298	intergenic	*NT5C1B*	neighboring genes: ***NT5C1B*****, *RDH14***	***Mag.#37***	1.0×10^{-8}	T	16.1|17.0	rs4832619*#	G	1	0.714	0.98 (0.90-1.08)	same	15	0.307
51	8	116588546	rs2049865	intronic	*TRPS1*	LD: ***TRPS1***	***Mag.#31***	4.9×10^{-8}	C	41.0|42.1	rs6469600*#	C	1	0.550	0.98 (0.91-1.05)	same	20	0.356

mRNA expression analysis and immunohistochemistry

Colonic tissue samples were obtained during surgical resection. Characteristics of patients used for RT-PCR are provided in Supplementary Table 7. RT-primer sequences are provided in Supplementary Table 8. Layer- and disease specific expression analysis results are shown in Supplementary Table 9 and 10. Fluorescence immunohistochemistry was performed as previously described [21]. Details on sample processing are provided in the Supplementary Materials and Methods section.

Gene set and pathway analysis

We used two gene set and pathway analysis approaches (MSigDB [22] and VEGAS2pathway [23]) to determine if the polygenic signal measured in the diverticular disease associated genes clustered in specific biological pathways. Lead candidate genes (Table 2, 3) were tested for overrepresentation with gene sets curated in MSigDB6.1. Results are provided in Supplementary Table 11 and 12. VEGAS2pathway results are provided in Supplementary Table 13 and 14.

Enrichment analyses in cell lines and primary tissues.

We used GARFIELD to identify significant enrichment patterns in our GWAS findings with regulatory or functional annotations in cell lines and primary tissue derived from *ENCODE* and Roadmap epigenomics data (Supplementary Table 15). GWAS SNPs were pruned (LD r2 > 0.1) and then annotated based on functional information overlap. Further details are provided in the Supplementary Methods section.

RESULTS

Genome-wide association study and validation of the loci

We observed genome-wide significant association ($P<5\times10^{-8}$) with diverticular disease for 2,568 variants mapping to 51 independent genomic loci (Supplementary Table 1), of which 12 had not been previously discovered (Table 2). The resulting Manhattan plot is shown in Figure 1A. The genomic inflation factor (λ_{GC}) was 1.199 and after LD score regression, the intercept was 1.02 – an acceptable level for this size of study (QQ Plot in Supplementary Figure 1) [24]. The 51 loci were validated in a combined European sample of 3,893 cases and 2,829 diverticula-free controls based on colonoscopy (Table 1). The direction of genotypic effect between discovery and replication samples was consistent for 48 out of 51 loci (93.8%; P for binominal test = 1×10^{-9}) (Supplementary Table 5) and odds ratios were strongly correlated between both analyses (r = 0.87; P= 1.59×10^{-13}, Supplementary Figure 2). Nominal replication significance (P<0.05) and a consistent direction of effect between the two cohorts were observed for 27 loci within European colonoscopy cohorts (Supplementary Table 6). Additional replication was observed for further 8 loci in a combined meta-analysis of European colonoscopy cohorts with a European population cohort from Michigan (Table 2, 3). 36 out of 48 identified risk loci have been previously reported [15] with genome-wide significant association (Table 2, 3 and Supplementary Table 4). All previously replicated risk loci for diverticular disease (*ARHGAP15*, *FAM155A*, *COLQ)* and (*GPR158, ABO, ANO1/FADD, ELN, BMPR1B, SLC35F3, SEM1/SHFM1*) were identified both in the current GWAS and replication analyses with similar odds ratios to those reported by Sigurdsson et al. [14] and Maguire et al.[15] (Table 3). The most significant novel risk variant rs9960286 is located near *CTAGE1* (Cutaneous T Cell Lymphoma-Associated Antigen 1) with a p-value of 2.3×10^{-10} and 0.002 ($OR_{allelic}$=1.14 [1.05-1.24]) in the replication analysis. The most significant novel replicated risk variant rs60869342 is located in *NOV* (Nephroblastoma Overexpressed) with a p-value of 4.4×10^{-13} and 0.0003 ($OR_{allelic}$=0.85 [0.78-0.93]) in the replication analysis. Rs1381335 (r^2=0.81 to rs60869342) in *NOV* was reported previously by Maguire et al. [15] as risk locus # 21, however, without formal replication.

Post hoc analysis of diverticulitis risk

The 27 replicating loci within European colonoscopy cohorts were evaluated for their relative genetic impact on diverticulitis (N=1167) and uncomplicated diverticulosis (N=1756) in a subset of the replication samples with the respective subphenotype information (Supplementary Table 16). The majority of loci showed similar odds ratios for diverticulosis and diverticulitis indicating relevance of the underlying variants for both phenotypes (Figure 1B, Supplementary Table 17). Loci that show a similar odds ratio in this analysis provide not evidence for a specific diverticulitis risk. Based on a 95% confidence interval, four loci showed stronger effects for diverticulitis, namely variants at *PHGR1* (OR 1.32, 95% CI 1.12-1.56), *FAM155A-2* (OR 1.21, 95% CI 1.04-1.42), *CALCB* (OR 1.17, 95% CI 1.03-1.33) and the *S100A10* (OR 1.17, 95% CI 1.03-1.33) locus.

Real-time PCR and immunohistochemistry analysis of curated candidate genes

We next selected candidate genes for further experimental analysis as detailed for each locus in Supplementary Table 3. Except for locus #25 (Supplementary Table 1, (*PHGR1* and *DISP2*), a single curated candidate gene "lead candidate gene" was assigned to each locus, based on local LD structure and supporting evidence from regulatory elements (eQTL and chromatin interaction). To provide a first indication of the relevant microanatomical colonic compartment relevant for disease, transcripts encoded at the top 20 replicating loci (Supplementary Table 6) were analyzed by quantitative real-time PCR in RNA preparations of the mucosal, submucosal and muscular layer from seven control patients (Supplementary Tables 7A, 8). The majority of transcripts (13 out 18 at $p<0.05$) showed layer-specific expression patterns indicating the relevance of this higher histotopographical resolution as compared to total colonic expression (Supplementary Table 9, Supplementary Figure 3). A potential disease-specific regulation of transcripts within each the mucosal, submucosal and muscular layer was analyzed in 20 controls, 13 diverticulosis and 21 diverticulitis patients (Supplementary Table 7b). A trend for upregulation of *S100A10* (nominal p=0.003) in the submucosal layer in diverticulitis patients was noted, while overall a primary and strong disease-specific differential expression finding was not observed (Supplementary Table 10 and Supplementary Figure 4). To obtain further spatial resolution, the localization of expressed protein at selected novel loci with expression in all layers (COL6A1),

predominant expression in the mucosa (PHGR1), submucosa (GPR158, EFEMP1) and submucosa and muscle layer (ELN, CRISPDL2) was investigated by immunohistochemistry (Figure 2B-E). As epitomized for instance for GPR158, which localizes predominantly to enteric ganglia and mucosa or elastin (ELN), which localizes to the lamina propria, vessel walls and muscle, significant additional information is gained by this higher anatomical resolution.

Overlap with inflammatory bowel disease (IBD), irritable bowel syndrome (IBS) and monogenic syndromes

There was no overlap of the 2,568 genome-wide significant variants ($P<5\times10^{-8}$) for diverticular disease with the 634 reported risk variants ($p<9\times10^{-6}$) according to the GWAS catalogue [25] for inflammatory bowel disease (IBD), Crohn´s disease (CD) and ulcerative colitis (UC). Also, there was no overlap of the lead candidate genes at the 48 risk loci with in the GWAS catalogue reported risk genes for IBD, CD and UC [26], except for *HLA-DQA1*. However, the IBD lead variant rs6927022 at the *HLA-DQA1* locus was not in LD to the diverticular disease associated lead SNPs according to FUMA, thus pointing to a non-overlapping genetic risk structure. The percentage of individuals diagnosed with IBS among GWAS cases was 7.6% as compared to 3.1% among controls not diagnosed with diverticular disease. None of the 51 genome-wide significant lead variants for diverticular disease was significantly associated with the IBS phenotype in the UK Biobank (15,401 diagnoses of IBS vs. 406,175 controls without a diagnosis of diverticular disease and without a diagnosis of IBS, data not shown). In contrast, mutations in 12 of the lead candidate genes for diverticular disease are reported in OMIM [27] as autosomal dominant or recessive causative factors for 18 monogenic syndromes (Supplementary Table 18). Many of these genes fall into the broad categories of neuromuscular syndromes, connective tissue stability disorders and morphogenesis traits and are considered in depth in the Discussion. A hypothesis-free analysis of the overlap of the genomic risk locations for diverticular disease within 500 kb distance to the lead variant is provided in Supplementary Table 19.

Functional implications of curated candidate gene signature

Consistent with the overlap with monogenic syndromes, a gene set enrichment analysis (GSEA / MSigDB) [22] using the 48 lead candidate genes revealed significant enrichments for neuromuscular mechanisms, connective tissue strength and morphogenesis (Supplementary Table 11, Supplementary Figure 5) and significant overlap with extracellular matrix-associated proteins of the murine colon (Supplementary Table 12).

Functional implications based on in silico analysis of the global diverticulosis risk signature

We performed additional hypothesis-free functional and gene annotations based on the genomic positions of risk loci using FUMA [18] as the curated candidate genes might not capture the full biology of the risk architecture. Positional gene mapping aligned SNPs to 176 genes, eQTL gene mapping matched *cis*-eQTL SNPs to 269 genes whose expression levels they influence (snp-gene pairs with FDR<0.05), with 21 genes specifically affected in sigmoid colon (Supplementary Table 20). Chromatin interaction mapping annotated SNPs to 977 genes based on 3D DNA–DNA interactions. This resulted in 1080 unique mapped genes (Supplementary Table 2 and Supplementary Figure 6). The majority of these mapped genes were protein coding genes (61%), while 39% were RNA and pseudogenes. A graphical representation of all mapped genes is given as circular plots for each chromosome carrying a risk locus in Supplementary Figure 7.

Using a broad definition of candidate variants, namely a p-value cut-off of 1.0×10^{-5} and $r^2\geq0.6$ to an independent significant SNP at the diverticular disease risk locus, most variants were located either intronic or intergenic (Supplementary Table 1). 18 variants, of which 9 were genome-wide significant, constituted exonic nonsynonymous variants (Supplementary Figure 8, Supplementary Table 17. Based on the Combined Annotation-Dependent Depletion (CADD) score, the most likely variants with functional consequences were rs1042917 (COL6A2) and rs17855988 (ELN) with CADD scores of 25.8 and 23.2, respectively (Supplementary Table 21). Detailed fine-mapping plots of each risk locus are provided in Supplementary Figure 9 showing local LD structure to the lead variant and annotation of variants by potential pathogenic and functional consequence assessed by

CADD score and Regulome score and presences of cis-eQTL variants in sigmoid colon tissue. At genomic risk locus #15 (in Table 2), our annotated candidate gene was COL6A1 with the lead SNP located intronic to the gene, instead of COL6A2 as implicated by the functional effect of the candidate SNP rs1042917. The proteins synthesized by both genes are subunits of collagen VI, thereby pointing to a consistent functional mechanism. The identification of the mechanistically causal variants at each risk locus will, however, require further experimentation in model organisms and human tissue.

Interestingly, 94.6% (4738 of 5007 SNPs) of candidate SNPs were located at sites of open chromatin (Supplementary Figure 10). Because the majority of lead variants were located in non-coding regions and thus not directly amendable to functional interpretation, we utilized GARFIELD to analyse enrichment statistics for the diverticular disease GWAS risk dataset with cell-specific coding, non-coding and functional elements from the GENCODE, ENCODE and Roadmap projects [18]. A graphical summary of the enrichment of DNAse I hypersensitive sites is provided in Supplementary Figure 11. As reported in detail in Supplementary Table 15, regulatory elements from fibroblasts, fetal muscle and brain were particularly enriched in the genetic risk structure of diverticular disease. To further mine the genomic locations for functional implications, we performed a VEGAS2Pathway analysis [23], which pointed to processes involved in cell and organ differentiation and extracellular matrix among the top five identified pathways (Supplementary Table 13, 14).

DISCUSSION

In this study we report the largest and most detailed genome-wide analysis to date for diverticular disease. We discovered 48 risk loci with genome-wide significance and consistent odds ratio in a replication sample. 27 of these loci replicate at a nominal significance level of $p<0.05$. Among these loci, 12 are novel risk loci for diverticular disease and 5 of the novel loci were also replicated in a European clinical cohort with detailed phenotyping and colonoscopy data for all controls. The three previously known risk loci [14] *ARHGAP15*, *COLQ* and *FAM155A* are among the validated loci and support the robustness of the phenotype and analysis on both the previous study and our analysis. A recent study by Maguire *et al.* [15], who analyzed a smaller UK-Biobank data set (n = 409,728 individuals) compared to the current study (n = 451,099) identified 40 loci with genome-wide significance using GWAS results publicly available from the Roslin Gene Atlas. There was an overlap for 36 out of 48 identified loci with genome-wide significance between the studies. Maguire *et al.* were able to replicate 8 loci in an independent European population cohort from Michigan. We replicated further 8 risk loci in a meta-analysis approach integrating data from this Michigan cohort. The current study thus increases the number of replicated susceptibility loci for diverticular disease to 35, of which 25 loci had previously not been replicated. A limitation of the discovery study is that controls were 4 years younger than the cases. The modest lower age of controls increases the chance to include yet undiagnosed cases in the control sample thereby potentially reducing the statistical power of the GWAS analysis. We based the functional interpretation of the GWAS results both on curated candidate genes and on more inclusive automated analysis tools such as GARFIELD, VEGAS2 and FUMA. Both analysis strategies point to diverticular disease as foremost a disorder of intestinal neuromuscular function and impaired connective fiber support. Many of the risk genes implicated in polygenic diverticular disease also have been implicated in monogenic neuromuscular and connective tissue disorders, as will be detailed below, which was consistent with the pathway analyses. These findings provide a specific molecular basis for the previously suggested mechanisms of structural weakness of the intestinal wall and dysregulated intestinal motility. Additional risk loci point towards a relevance of intestinal epithelial and vascular function, while a prominent immune signature was not apparent in the data.

Neuromuscular mechanisms

A number of candidate genes point towards a dysfunction of the enteric nervous system and the neuromuscular junction in the large bowel. Mutations in *COLQ* cause myasthenic congenital syndrome and the gene product anchors asymmetric acetylcholine (ACh) esterase in the basal lamina of the motoric endplate [28]. *COL6A1* encodes the alpha 1 subunit of collagen VI (ColVI) [29]. ColVI is required for the structural and functional integrity of the neuromuscular junction [30]. Mutations in glial cell line-derived neurotrophic factor (*GDNF*) have been suggested to act in concert with *RET* mutations to produce aganglionic megacolon (Hirschsprung´s disease), which is characterized by congenital absence of intrinsic ganglion cells in the myenteric and submucosal plexuses of the gastrointestinal tract [31]. Impaired GDNF function has been shown at gene and protein level not only to occur in diverticular disease but also during early stages of diverticula formation [32]. Plausible links to neuronal physiology are also evident for *GPR158*, a G-protein coupled orphan receptor [33] and brain derived neurotropic factor (*BDNF*).

Three identified genes point to calcium sensitization and calcium-dependent signaling in gastrointestinal smooth muscle [34]: Inhibiting myosin light chain phosphatase activity with protein kinase C-potentiated phosphatase inhibitor protein-17 kDa (CPI-17, *PPP1R14A*) is considered one of the primary mechanisms underlying myofilament Ca2+ sensitization [35]. Further, for ANO1 (Anoctamin 1), a Calcium activated chloride channel, a role of in mediating cholinergic neurotransmission in the murine gastric fundus has been shown [36]. *CACNB2* (Cav1.2) encodes for the beta-2 subunit of a calcium-dependent calcium channel. The expression of Cav1.2 channels in colonic smooth muscle cells is key to colonic motility, decreased in colonic inflammation and a potential treatment target for motility disorders [37]. Taken together these data give further evidence for disturbed enteric neuromuscular functions as a relevant mechanism of diverticular disease [2,38].

Neuromuscular development

HLX is a homeobox transcription factor gene conserved across species [27]. Mutations in *HLX* have been observed in two fetuses with congenital diaphragmatic hernia and *HLX* homozygous null mice have a short bowel and reduced muscle cells in the diaphragm [39,40]. HLX homozygous null animals exhibiting abnormal developmental of the enteric nervous system [39].

Connective tissue function and morphogenesis

A second common functional theme of the identified risk loci is connective fiber function based in pathway, molecular function and syndrome associations. For instance, elastin (*ELN*) encodes a protein that is one of the two components of elastic fibers which confer elasticity to organs and tissues. Mutations in *ELN* cause autosomal dominant cutis laxa [41]. Mutations in bone morphogenetic protein receptor type 1B (BMPR1B) underlie autosomal recessive Hunter-Thompson [42] type of acromesomelic dysplasia. EGF Containing Fibulin Extracellular Matrix Protein 1 (*EFEMP1*) has been associated with polygenic susceptibility to inguinal hernia [43] and varicose veins [44]. *EFEMP1* encodes fibulin-3, an extracellular matrix protein. Efemp1(-/-) mice developed multiple large hernias including inguinal hernias. Histological analysis of Efemp1(-/-) mice revealed a marked reduction of elastic fibers in fascia [45]. The fibulin family of protein has been associated with further connective tissue disorders. Mutations in fibulin-5 have been identified in patients with cutis laxa and mutations in fibrillin 1 cause Marfan syndrome. Interestingly, the N-terminal region of fibrillin-1 mediates a bipartite interaction with LTBP1 [46]. Variants in Cysteine Rich Secretory Protein LCCL Domain Containing 2 (*CRISPLD2*) have been associated with non-syndromic orofacial cleft [47,48]. A further example without association to genetic syndromes includes tissue inhibitor of metalloproteinases 2 (TIMP2), a peptidase involved in degradation of the extracellular matrix. The S100A10 protein regulates the remodelling of the extracellular matrix through plasmin-dependent activation of MMP-9 and plasminogen-dependent macrophage tissue invasion [49,50].

Mesenteric vascular function

Diverticula occur predominantly at sites of preformed weakness in the intestinal wall, namely at sites of vascular entry through the muscle layer. In the interaction between muscular layer and the vessel, vascular biology and contractility may play an additional role. Calcitonin related polypeptide beta (*CALCB*), which plays a role in mesenteric vascular smooth muscle function [51] and protein phosphatase 1 regulatory subunit 16B (*PPP1R16B*), which regulates endothelial cell function [52] may provide a potential mechanistic basis for altered vascular biology at these entry points.

Epithelial function and risk of diverticulitis

Interestingly, only one of the identified candidate genes – namely *PHGR1* – has a clear and exclusive link to epithelial function. Proline-, histidine-, and glycine-rich protein 1 mRNA and protein are found to be expressed specifically in epithelial cells of intestinal mucosa as shown previously [53] and in our immunohistochemistry analyses in Figure 2 with the highest expression in the most mature and differentiated cells. *PHGR1* showed the strongest effect size (odds ratio 1.3 in comparison to uncomplicated diverticulosis) among the few loci associated with a higher risk of diverticulitis suggesting that for this complication of diverticular disease, indeed epithelial cell function may play a key role.

In summary, the novel genetic risk signature indicates that diverticular disease is a disorder of impaired intestinal neuromuscular function, impaired mesenteric vascular smooth muscle function and of impaired connective fiber support. We observe an intriguing convergence of previous monogenic findings with the polygenic risk signature of diverticular disease through the overlap with syndromic neuromuscular, connective tissue and morphogenesis disorders. Through the phenotype and the established cell biology of the Mendelian syndromes, inference of the functional implication of the novel risk loci for instance at the motoric end plate is possible. The manifestation of the inflammatory complication – diverticulitis – in turn may be triggered by epithelial dysfunction in the context of altered colon anatomy. These findings provide a deeper understanding of colonic biology and disease

pathophysiology and open a new path for a functional dissection and therapeutic tackling of this common disease.

ACKNOWLEDGEMENTS

The authors wish to thank all study participants, researchers, clinicians and administrative staff who contributed to this study.

Contributors

JWH, SB, CL, FC: performed the experiments, analysed the data and wrote the manuscript; JWH, AH, SB, AR, WR, NB: performed the bioinformatic analyses; CL, FC, MB: performed real-time PCR, histological, immunohistochemical analyses; CS, FL, LK, MZ, WvS, MCR, JR, TB coordinated, managed collection of samples, performed phenotyping; WL coordinated and supervised collection of samples; SN, UH-S, FR, PH, BS, WK, JT, MZ, JR, AW-B ,TJ, JK, MS, IV, PS, HB, HH, AV, J-UE, GB, AH, S H, SW, ML, TK, SB, UH-S, LP, LSN, H-WS, SZ, SP GF, AA, PTS, GL, J W, FL, TB, LK, PM, RG, VM: obtained the samples, performed phenotyping, interpretation of data; MK, MD'A, SZ, AF, MB, HV, WK, FL, RVT, JT gave conceptual advice, participated in the discussions, interpretation of the results, editing of the manuscript; CS, JH, MW, CD, AG, TW: conceived the experimental and analytical design, analysed data, wrote and reviewed the manuscript. All authors critically revised and contributed to the final manuscript.

Funding

The work presented in this manuscript was supported by the German Research Council (DFG, Ha3091/9-1, WE2366/5-1) and the Austrian Science Fund (FWF, I1542-B13). Further support was received from SPAR Austria and from institutional funds from the Christian-Albrechts-University Kiel. The recruitment of the West German cohort was supported by a grant from the Faculty of Medicine, Saarland University (HOMFOR grant T201000747) to Matthias C. Reichert. This study was supported by a grant of the Research Council of Lithuania No. SEN-06/2015/PRM15-135. Anna Andreasson, Peter Thelin Schmidt were supported by the Stockholm County Council (ALF project). Mauro D'Amato was supported by the Swedish Research Council (VR grant 2017-02403). Data access to the UK biobank data was granted under project numbers 22691 and 9055.

Conflicts of interest

The authors declare no conflicts.

REFERENCES

1 Tursi A. Diverticulosis today: unfashionable and still under-researched. *Therap Adv Gastroenterol* 2016;**9**:213–28. doi:10.1177/1756283X15621228

2 Pfützer RH, Kruis W. Management of diverticular disease. *Nat Rev Gastroenterol Hepatol* 2015;**12**:629–38. doi:10.1038/nrgastro.2015.115

3 Etzioni DA, Mack TM, Beart RW, *et al.* Diverticulitis in the United States: 1998-2005: changing patterns of disease and treatment. *Ann Surg* 2009;**249**:210–7. doi:10.1097/SLA.0b013e3181952888

4 Weizman A V, Nguyen GC. Diverticular disease: epidemiology and management. *Can J Gastroenterol* 2011;**25**:385–9

5 Warner E, Crighton EJ, Moineddin R, *et al.* Fourteen-year study of hospital admissions for diverticular disease in Ontario. *Can J Gastroenterol* 2007;**21**:97–9

6 Stollman N, Raskin JB. Diverticular disease of the colon. *Lancet (London, England)* 2004;**363**:631–9. doi:10.1016/S0140-6736(04)15597-9

7 Jacobs DO. Clinical practice. Diverticulitis. *N Engl J Med* 2007;**357**:2057–66. doi:10.1056/NEJMcp073228

8 Sheth AA, Longo W, Floch MH. Diverticular disease and diverticulitis. *Am J Gastroenterol* 2008;**103**:1550–6. doi:10.1111/j.1572-0241.2008.01879.x

9 Peery AF, Dellon ES, Lund J, *et al.* Burden of gastrointestinal disease in the United States: 2012 update. *Gastroenterology* 2012;**143**:1179-87.e1-3. doi:10.1053/j.gastro.2012.08.002

10 Sandler RS, Everhart JE, Donowitz M, *et al.* The burden of selected digestive diseases in the United States. *Gastroenterology* 2002;**122**:1500–11. doi:S0016508502749652 [pii]

11 von Rahden BHA, Germer C-T. Pathogenesis of colonic diverticular disease. *Langenbeck's Arch Surg* 2012;**397**:1025–33. doi:10.1007/s00423-012-0961-5

12 Strate LL, Erichsen R, Baron JA, *et al.* Heritability and familial aggregation of diverticular

disease: a population-based study of twins and siblings. *Gastroenterology* 2013;**144**:736–742.e1; quiz e14. doi:10.1053/j.gastro.2012.12.030

13 Granlund J, Svensson T, Olén O, *et al.* The genetic influence on diverticular disease--a twin study. *Aliment Pharmacol Ther* 2012;**35**:1103–7. doi:10.1111/j.1365-2036.2012.05069.x

14 Sigurdsson S, Alexandersson KF, Sulem P, *et al.* Sequence variants in ARHGAP15, COLQ and FAM155A associate with diverticular disease and diverticulitis. *Nat Commun* 2017;**8**:15789. doi:10.1038/ncomms15789

15 Maguire LH, Handelman SK, Du X, *et al.* Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease. *Nat Genet* 2018;**50**:1359–65. doi:10.1038/s41588-018-0203-z

16 Buch S, Stickel F, Trépo E, *et al.* A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. *Nat Genet* 2015;**47**:1443–8. doi:10.1038/ng.3417

17 Loh P-R, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;**47**:284–90. doi:10.1038/ng.3190

18 Watanabe K, Taskesen E, van Bochoven A, *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826. doi:10.1038/s41467-017-01261-5

19 Chang CC, Chow CC, Tellier LC, *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7. doi:10.1186/s13742-015-0047-8

20 Liu JZ, Tozzi F, Waterworth DM, *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010;**42**:436–40. doi:10.1038/ng.572

21 Cossais F, Leuschner S, Barrenschee M, *et al.* Persistent Increased Enteric Glial Expression of S100β is Associated With Low-grade Inflammation in Patients With Diverticular Disease. *J Clin Gastroenterol* Published Online First: March 2018. doi:10.1097/MCG.0000000000001011

22 Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-

based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545–50. doi:10.1073/pnas.0506580102

23 Mishra A, MacGregor S. A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Res Hum Genet* 2017;**20**:1–9. doi:10.1017/thg.2016.100

24 Bulik-Sullivan BK, Loh P-R, Finucane HK, *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;**47**:291–5. doi:10.1038/ng.3211

25 MacArthur J, Bowler E, Cerezo M, *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;**45**:D896–901. doi:10.1093/nar/gkw1133

26 Ellinghaus D, Jostins L, Spain SL, *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* 2016;**48**:510–8. doi:10.1038/ng.3528

27 Hamosh A, Scott AF, Amberger J, *et al.* Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* 2000;**15**:57–61. doi:10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G

28 Ohno K, Brengman J, Tsujino A, *et al.* Human endplate acetylcholinesterase deficiency caused by mutations in the collagen-like tail subunit (ColQ) of the asymmetric enzyme. *Proc Natl Acad Sci U S A* 1998;**95**:9654–9.

29 Bönnemann CG. The collagen VI-related myopathies: muscle meets its matrix. *Nat Rev Neurol* 2011;**7**:379–90. doi:10.1038/nrneurol.2011.81

30 Cescon M, Gregorio I, Eiber N, *et al.* Collagen VI is required for the structural and functional integrity of the neuromuscular junction. *Acta Neuropathol* Published Online First: May 2018. doi:10.1007/s00401-018-1860-9

31 Angrist M, Bolk S, Halushka M, *et al.* Germline mutations in glial cell line-derived

neurotrophic factor (GDNF) and RET in a Hirschsprung disease patient. *Nat Genet* 1996;**14**:341–4. doi:10.1038/ng1196-341

32 Barrenschee M, Wedel T, Lange C, *et al.* No neuronal loss, but alterations of the GDNF system in asymptomatic diverticulosis. *PLoS One* 2017;**12**:1–13. doi:10.1371/journal.pone.0171416

33 Khrimian L, Obri A, Ramos-Brossier M, *et al.* Gpr158 mediates osteocalcin's regulation of cognition. *J Exp Med* 2017;**214**:2859–73. doi:10.1084/jem.20171320

34 Perrino BA. Calcium sensitization mechanisms in gastrointestinal smooth muscles. *J Neurogastroenterol Motil* 2016;**22**:213–25. doi:10.5056/jnm15186

35 Kitazawa T, Eto M, Woodsome TP, *et al.* Phosphorylation of the myosin phosphatase targeting subunit and CPI-17 during Ca2+ sensitization in rabbit smooth muscle. *J Physiol* 2003;**546**:879–89.

36 McErlain TL, Bhraonain EPN, Kelly RTS. The role of Ano1 in mediating cholinergic neurotransmission in the murine gastric fundus. *J Physiol* 2018;:epub. doi:10.1113/JP276383

37 Shi XZ, Pazdrak K, Saada N, *et al.* Negative transcriptional regulation of human colonic smooth muscle Cav1.2 channels by p50 and p65 subunits of nuclear factor-κB. *Gastroenterology* 2005;**129**:1518–32. doi:10.1053/j.gastro.2005.07.058

38 Böttner M, Wedel T. Abnormalities of neuromuscular anatomy in diverticular disease. *Dig Dis* 2012;**30**:19–23. doi:10.1159/000335699

39 Farrell SA, Sodhi S, Marshall CR, *et al.* HLX is a candidate gene for a pattern of anomalies associated with congenital diaphragmatic hernia, short bowel, and asplenia. *Am J Med Genet* 2017;**173**:3070–4. doi:10.1002/ajmg.a.38354

40 Zhu Q, High FA, Zhang C, *et al.* Systematic analysis of copy number variation associated with congenital diaphragmatic hernia. *Proc Natl Acad Sci* 2018;**115**:5247–52. doi:10.1073/pnas.1714885115

41 Zhang MC, He L, Giro M, *et al.* Cutis laxa arising from frameshift mutations in exon 30 of the elastin gene (ELN). *J Biol Chem* 1999;**274**:981–

6.http://www.ncbi.nlm.nih.gov/pubmed/9873040

42 Ullah A, Umair M, Muhammad D, *et al.* A novel homozygous variant in BMPR1B underlies acromesomelic dysplasia Hunter-Thompson type. *Ann Hum Genet* 2018;**82**:129–34. doi:10.1111/ahg.12233

43 Jorgenson E, Makki N, Shen L, *et al.* A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. *Nat Commun* 2015;**6**:10130. doi:10.1038/ncomms10130

44 Ellinghaus E, Ellinghaus D, Krusche P, *et al.* Genome-wide association analysis for chronic venous disease identifies EFEMP1 and KCNH8 as susceptibility loci. *Sci Rep* 2017;**7**:45652. doi:10.1038/srep45652

45 McLaughlin PJ, Bakall B, Choi J, *et al.* Lack of fibulin-3 causes early aging and herniation, but not macular degeneration in mice. *Hum Mol Genet* 2007;**16**:3059–70. doi:10.1093/hmg/ddm264

46 Robertson IB, Dias HF, Osuch IH, *et al.* The N-Terminal Region of Fibrillin-1 Mediates a Bipartite Interaction with LTBP1. *Structure* 2017;**25**:1208–1221.e5. doi:10.1016/j.str.2017.06.003

47 Chiquet BT, Yuan Q, Swindell EC, *et al.* Knockdown of Crispld2 in zebrafish identifies a novel network for nonsyndromic cleft lip with or without cleft palate candidate genes. *Eur J Hum Genet* Published Online First: June 2018. doi:10.1038/s41431-018-0192-5

48 Chiquet BT, Lidral AC, Stal S, *et al.* CRISPLD2: a novel NSCLP candidate gene. *Hum Mol Genet* 2007;**16**:2241–8. doi:10.1093/hmg/ddm176

49 Gong Y, Hart E, Shchurin A, *et al.* Inflammatory macrophage migration requires MMP-9 activation by plasminogen in mice. *J Clin Invest* 2008;**118**:3012–24. doi:10.1172/JCI32750

50 Madureira PA, O'Connell PA, Surette AP, *et al.* The biochemistry and regulation of S100A10: a multifunctional plasminogen receptor involved in oncogenesis. *J Biomed Biotechnol* 2012;**2012**:353687. doi:10.1155/2012/353687

51 Chauhan M, Yallampalli U, Banadakappa M, *et al.* Involvement of Receptor Activity-Modifying Protein 3 (RAMP3) in the Vascular Actions of Adrenomedullin in Rat Mesenteric Artery Smooth Muscle Cells. *Biol Reprod* 2015;**93**:116. doi:10.1095/biolreprod.115.134585

52 Boratkó A, Csortos C. TIMAP, the versatile protein phosphatase 1 regulator in endothelial cells. *IUBMB Life* 2017;**69**:918–28. doi:10.1002/iub.1695

53 Oltedal S, Skaland I, Maple-Grødem J, *et al.* Expression profiling and intracellular localization studies of the novel Proline-, Histidine-, and Glycine-rich protein 1 suggest an essential role in gastro-intestinal epithelium and a potential clinical application in colorectal cancer diagnostics. *BMC Gastroenterol* 2018;**18**:1–15. doi:10.1186/s12876-018-0752-8

FIGURES

Figure 1: GWAS results

Principal findings of genetic analyses: ***Panel A:*** Manhattan plot of genome-wide association results for diverticular disease. P values ($-\log_{10}$) are shown for SNPs that passed quality control. The genome-wide significance threshold (5×10^{-8}) is shown as a black line. Gene names for loci with consistent effect and a known gene annotation are included in the panel. Gene names for newly discovered loci (as detailed in Table 2) are printed in bold. ***Panel B:*** Forrest plot with 95% confidence intervals of the relative impact of the 27 replicating variants on diverticulitis versus diverticulosis risk. ORs greater than one indicate a higher impact on diverticulitis risk. The respective reference allele is provided in Supplementary Table 17. ***Panel C:*** Locus plot for diverticular disease risk locus *GPR158*. The −log10 (P values, mixed model association test) are plotted against SNP genomic position based on NCBI Build 37, with the names and location of nearest genes shown at the bottom. The variant with the lowest P value (lead variant) in the discovery analysis in the region is marked by a purple diamond. SNPs are coloured to reflect correlation with the most significant SNP, with red denoting the highest LD (r^2 >0.8) with the lead SNP. The association signal is confined to a single association peak located intronic in *GPR158*. Estimated recombination rates from the 1000 Genomes Project (hg19/genomes March 2012 release, EUR population) are plotted in blue to reflect the local LD structure. Gene annotations were obtained from the UCSC Genome Browser. The plot was generated using LocusZoom. ***Panel D:*** Locus plot for diverticular disease risk locus *FAM155A*: The variant with the lowest P value in the *FAM155A-1* region is marked by a purple diamond. For the *FAM155A* gene, two independent association signals (termed *FAM155A-1* and *FAM155A-2*) with low pairwise LD (r^2=0.0043) were considered as individual loci. SNPs are coloured to reflect correlation with the most significant SNP at *FAM155A-1*, with red denoting the highest LD (r2 >0.8) and dark blue the lowest LD (r2 < 0.2) with the lead SNP.

Figure 2: Expression of risk genes

Layer-specific expression pattern of novel risk genes for diverticular disease. ***Panel A:*** Normalized mRNA expression in the mucosal (left, green), submucosal (middle, red) and muscular (blue, right)

layers in control colon (n=7). ***Panels B-E:*** Fluorescence immunohistochemical analysis of expression in control colon in the mucosa (B), submucosal (C), muscular layer (D) and in myenteric ganglia (E). The respective target gene antibody is labelled in red, with DAPI (blue) for nuclear staining and alpha smooth muscle actin (smooth muscle marker, C, D) and Protein Gene Product 9.5 (neuronal marker, E) in green. It is evident, that risk genes show different expression patterns within the colonic wall and are localized to specific structures such as blood vessels, lamina propria, epithelium, smooth muscle or nerve cells. Scale bars are added in white (50 µm).

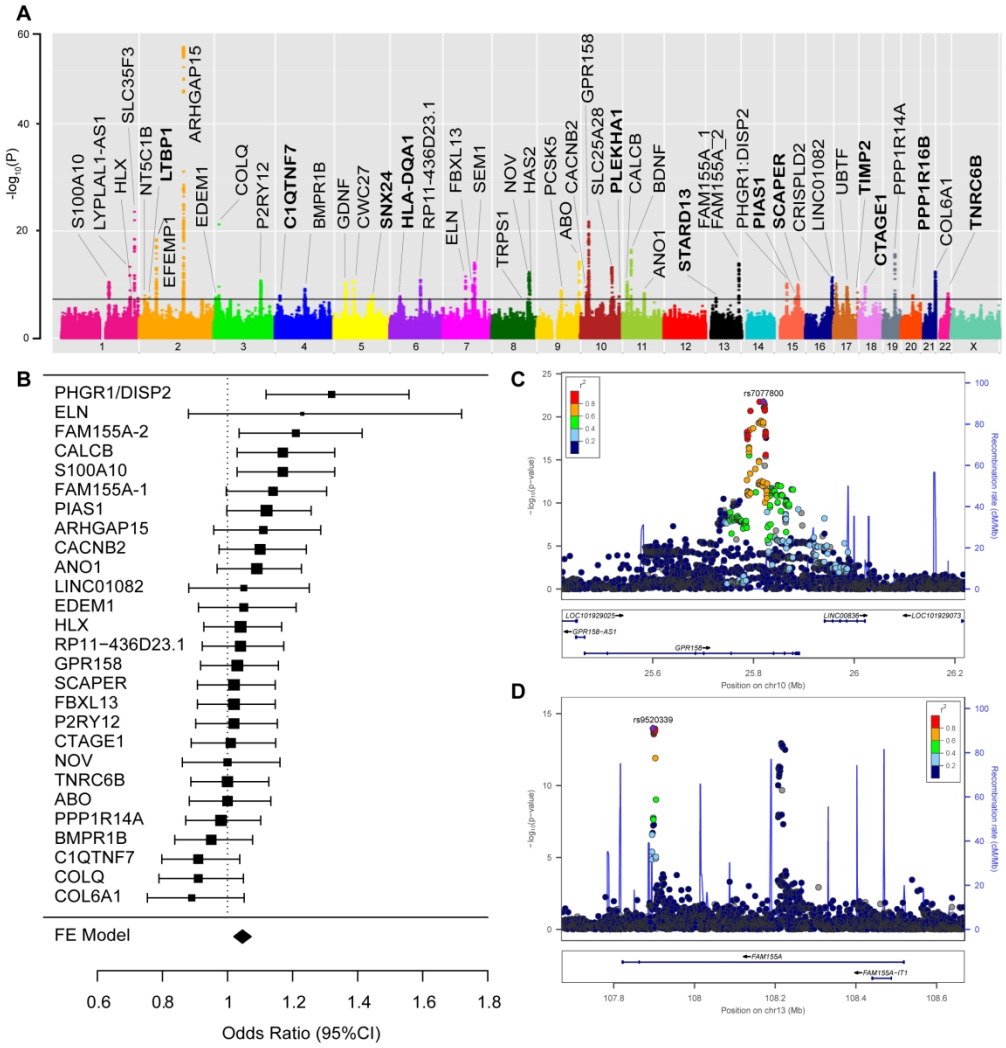

Figure 1: GWAS results

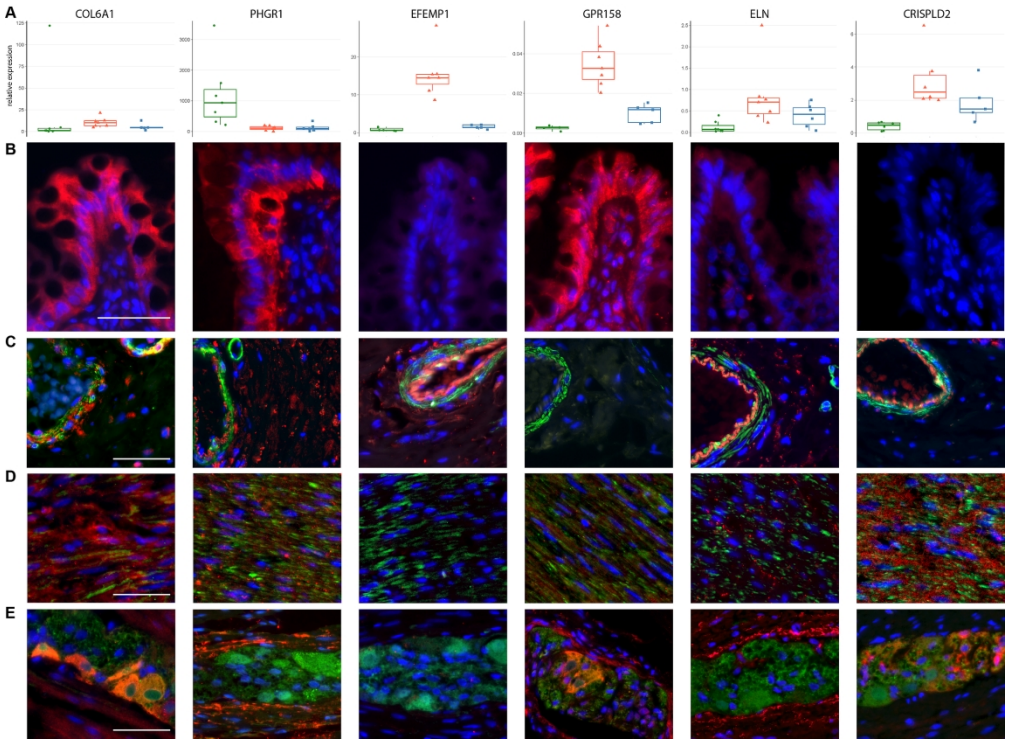

Figure 2: Expression of risk genes

230x169mm (300 x 300 DPI)

SUPPLEMENTARY MATERIALS AND METHODS

Phenotype definition in the UK Biobank

An individual was classified as a diverticular disease case if they matched hospital-bases ICD9 or ICD10 coding (562 and K57 respectively) for primary (n = 16,560), secondary (n = 13,375) and self-reported diverticular disease diagnosis (n = 1,982; information collected and placed within the coding tree during the verbal interview at the assessment clinic by a trained nurse) in the UK Biobank dataset. Control individuals where classified on the basis of absence of a diverticular disease diagnosis (n = 419,135). We did not have complete information on the proportion of diverticulitis patients in the UK Biobank as depth of ICD coding was insufficient to differentiate disease subtype diverticulosis (i.e. diverticular disease without inflammation) from diverticulitis. The majority of cases had diagnosis K57.3 coding for both disease subtypes diverticulosis without perforation or abscess (K57.30, K57.31) and diverticulitis without perforation or abscess (K57.32, K57.33) not allowing this differentiation at available coding depth K57.3.

Recruitment and phenotyping of the German replication samples

The Northern and Western German samples were phenotyped as follows: Controls and cases with uncomplicated diverticulosis were defined by manual review of colonoscopy (complete colonoscopy required) and patient records at participating hospitals and gastroenterology outpatient services. In addition, the patient questionnaire was reviewed for absence of diverticulitis (including a hospital or outpatient diagnosis of diverticulitis or episodes of lower left quadrant paint and fever in diverticulosis patients). Cases with diverticulitis were defined either as patients with colonoscopy proven diverticular disease, that required antibiotics due to diverticulitis or as patients diagnosed by abdominal ultrasound or CT scan. To obtain patient information on additional cases with diverticulitis, hospital or medical office information systems were screened for ICD- Code K57.x and OPS Code 5-455.7 for resection of the sigma and patient records where then further reviewed as described above. All patients were of self-reported Caucasian ancestry. Patients with carcinoma or inflammatory bowel

disease (IBD) were excluded from all groups. Patients from the Germany / North cohort were recruited through the popgen biobank as described previously [1,2]. Patients from the Germany / West cohort were recruited at the Department of Medicine II, Saarland University Medical Center, Homburg between 2012 and 2017. All German participants provided written informed consent. The study protocol was approved by the Research Ethics Committees of the Saarland University (approval 63/11), the Medical faculty of Christian-Albrechts-University Kiel, Germany (A 156/03) and the Medical Faculty of the Technische Universität Dresden (EK470122013).

Austrian replication samples

For the Austria / Vienna cohort, the ongoing molecular epidemiology colorectal cancer study of Austria (CORSA) was used. More than 16,000 Caucasian participants were recruited since May 2003 through the province-wide screening program "Burgenland Prevention Trial of Colorectal Disease with Immunological Testing" (B-PREDICT). All habitants of Burgenland aged between 40 and 80 years are invited to take part in the screening program. Participants with a positive faecal occult blood test receive further diagnostic workup including colonoscopy. Results of colonoscopies are collected in a central database and standardized documentation guidelines are followed. Demographic and anthropometric factors, dietary and smoking habits are assessed by questionnaire. All subjects gave written informed consent and the study was approved by the institutional review board "Ethikkommission Burgenland (EK33/2010)". Ascertainment of the diverticulosis / diverticulitis phenotype in the CORSA study population was based on database review of colonoscopy and clinical data for the years 2003 to 2009. Individuals with colon cancer or IBD were excluded. Genotypes for replication analysis were extracted from available QC'ed Axiom Genome-Wide CEU1 array (Affymetrix, Santa Clara, CA) data which was imputed to 1000 Genomes Project Phase 3 reference (using IMPUTE2). For the Austria / Oberndorf samples, a systematic, prospective recruitment of all patients undergoing screening colonoscopy at the Krankenhaus Oberndorf is being performed since 2007. Patients were recruited on site according to the same phenotypic criteria as for the German

cohorts. The study was approved by the local ethics committee (Ethikkommission des Landes Salzburg, approval no. 415-E/ 1262/2-2010) and informed consent was obtained from all participants.

Lithuanian replication cohort

Patients were recruited at the Department of Gastroenterology at the Lithuanian University of Health Sciences, Kaunas in Lithuania between 2012 and 2017 from patients referred for colonoscopy according to the same criteria as the Germany / West samples as described previously [3]. All patients have signed an informed consent form to participate in the study. The study protocol was approved by the Regional Kaunas Ethics Committee (BE-10-2). The study was performed according to the Declaration of Helsinki.

Swedish replication samples

The population-based colonoscopy study (PopCol) was performed at Ersta hospital in Stockholm, Sweden from 2002 to 2006, where 3356 randomly selected adults from the general population were sent an Abdominal Symptom Questionnaire and 2293 responders were contacted for further investigations. Of the 745 individuals (426 women and 319 men) who underwent an ileo-colonoscopy as part of the study, 130 individuals (17.4 %) had diverticulosis. No individual presented with diverticulitis. Illumina OmniExpressExome-8 v1 genotypes were extracted from available QC'ed and imputed data, which have already been used and described in previous publications [4,5]. Study approval was obtained from the local Ethics committee (No 394/01, Karolinska Institutet Huddinge Regional Ethics Board, Sweden) and written informed consent was obtained from all participants. The PopCol study is described in detail in Kjellström *et al.* Eur J Gastroenterol Hepatol. 2014 Mar; 26(3):268-75.

GWAS analysis

Discovery GWAS analysis was performed on UK Biobank on Version 3 imputed genotypes. Genome-wide association tests were performed using BOLT-LMM v2.34, which applies a linear mixed model to adjust for the effects of population structure and individual relatedness[6]. This enabled the inclusion of all related individuals in our white European subset allowing a sample size of up 451,099 individuals, as opposed to a maximal set of 379,767 unrelated individuals. We limited our analysis to 11,977,111 genetic variants centrally imputed using both the Haplotype Reference Consortium imputation reference panel and a combined UK10K and 1000 Genomes reference panel with a minimum minor allele frequency (MAF) >0.1% and imputation quality score (INFO) >0.4.

Data availability

The GWAS summary statistics is publicly available on our group website http://www.t2diabetesgenes.org/data/.

The data from UK Biobank reported in this paper are available via application directly to the UK Biobank (http://www.ukbiobank.ac.uk/).

Validation genotyping

The 51 loci were validated in a combined European sample of 3,893 cases and 2,829 diverticula-free controls based on colonoscopy (Table 1) using the most significant discovery variant. When direct genotyping of a lead variant was not technical feasible, appropriate proxies were selected instead, defined as the variant with the next-lowest P-value within 250 kb of the index SNP (Table 2,3 Supplementary Table 5). Genotyping of SNPs was performed using the Agena® iPLEX Gold chemistry MassARRAY platform and TaqMan® technology from Life Technologies on an automated platform as described previously [2]. The choice of genotyping technology per variant was based on technical considerations of assay design feasibility and is indicated in Table 2 and 3. Genomic DNA was amplified with the GenomiPhi (Amersham) whole-genome amplification kit and fragmented at 99 °C for 3 min. All data were logged and managed with a database-driven laboratory information management system (LIMS) [7]. Individual samples with >5% missing data were excluded from

further analyses. SNPs that had >5% missing data or deviated from Hardy-Weinberg equilibrium (exact $P < 10^{-6}$ in controls) were excluded and replaced. Logistic regression analyses under additive model of inheritance were performed with PLINK [8] adjusting for gender and age for Taqman and Plex genotyped cohorts (i.e. German data sets (Germany North and West), the Austrian data set (Austria Oberndorf) and the Lithuanian data set). For cohorts Austria (Vienna) and Sweden, 1000 genomes imputed GWAS data was available, here association tests were performed using SNPTEST (v2.5) [9] adjusting for gender and age. Study-specific β effect estimates from all European replication/differentiation cohorts were then combined by fixed-effect meta-analysis using an inverse variance–weighted method implemented in META (1.6.0) [10]. For replication a nominal significance level of $P<0.05$ and consistency in odds ratio direction between the discovery and replication stage was applied.

To increase statistical power to replicate risk loci with lower observed odds ratios the sample size of the replication cohort was increased by including replication data for diverticulitis presented by Maguire et al. from European samples (N=29,367) from the Michigan genome initiative (MGI) into a meta-analysis of all European replication cohorts (N=36,089 samples). Study-specific z-scores for each allele were combined across samples in a weighted sum, with weights proportional to the square-root of the sample size for each study implemented in METAL (http://csg.sph.umich.edu/abecasis/Metal/).

Sudy-specific effective sample sizes were calculated as 4 / (1/[# of cases] + 1/[# of controls]).

mRNA expression analysis

Colonic tissue samples were obtained during surgical resection for the controls and non-inflamed diverticulosis during partial colectomy for nonobstructive colorectal carcinoma with tissue obtained distal from the tumor at a distance >5cm from any additional pathology or the tumour. Anorectal evacuation and colonic motility disorders were excluded. Diverticulitis samples were obtained from patients operated after two or more attacks of diverticulitis during elective surgery. Full-thickness specimens were harvested from sites adjacent to colonic diverticula. All specimens were immediately

transferred from the operating room to the laboratory for tissue processing in PBS (phosphate-buffered saline, pH 7.2). The study of human tissue received approval from the Local Ethics Committee of the Faculty of Medicine, Kiel University, Germany (B299/07). Patient characteristics are provided in Supplementary Table 7.

For dissection of colonic layers, full-thickness rectangular tissue blocks (30 mm x 10 mm) were pinned out flat on a cork plate by fine needles as described previously [11]. Mucosa, submucosa, and muscularis propria were separated using microsurgical scissors and immediately frozen in isopentane after resection and stored at −70°C until further processing. RNA was extracted using the NucleoSpin® RNA Kit (Macherey-Nagel, Düren, Germany). Reverse transcription was carried out using the M-MLV Reverse Transcriptase RNase (H-) (Promega, Mannheim, Germany) according to the manufacturer's protocol. Duplicate real-time quantitative PCR reactions were performed with qPCR Master Mix Plus (Eurogentec, Seraing, Belgium) using an ABI Prism 7500 fast Real-Time PCR cycler (Life Technologies). The housekeeping gene *HPRT* was used for normalization. The primer/probe sets (Eurogentec) or TaqMan assays (Thermo Fisher Scientific, Waltham, MA, USA) used are listed in Supplementary Table 8. Quality filtering: samples showing deviations at duplicate Ct values of >1.5 units were excluded from analysis. Potential outliers were not excluded from non-parametric group comparison analysis.

Statistical analysis of mRNA expression data and creation of box plots with jittered points were performed using ggpubr: 'ggplot2' package in r (http://www.sthda.com/english/rpkgs/ggpubr/). For the comparison of two groups a non-parametric Mann-Whitney U test (i.e. unpaired two-samples Wilcoxon test) was used. For the comparison of three groups a non-parametric Kruskal-Wallis test was used. Differences were considered significant after correction for multiple testing at $P<0.0025$ (0.05/20 tests). All results are expressed as medians with interquartile ranges.

Immunohistochemistry

Fluorescence immunohistochemistry was performed as previously described [12]. Briefly, colonic specimens were fixed in 4% paraformaldehyde for 24 h. Paraffin-embedded tissue sections were pre-treated with citrate buffer and incubated overnight with following primary antibodies: mouse anti-Col6a1 (B-4, Santa Cruz Biotechnology, Santa Cruz, USA), rabbit anti-PHGR1 (gift of Oddmund Nordgård, Stavanger University Hospital, Norway), rabbit anti-GPR158 (ABIN 2890856, Antibodies-online.com, Aachen, Germany), mouse anti-EFEMP1 (mAB3-5, Santa Cruz Biotechnology), rabbit anti-CRISPLD2 (NBP1-85143, Novusbio, Littleton, USA), rabbit anti-Elastin (Ab21610, Abcam, Cambridge, U.K.), rabbit anti-a-SMA (Ab5694, Abcam), mouse anti-a-SMA (M0851, Dakocytomation, Glostrup, Denmark), rabbit anti-PGP9.5 (RA95101, UltraClone, Isle of Wight, U.K.) and mouse anti-PGP9.5 (BM699, Acris, Herford, Germany). Anti-rabbit AlexaFluor488, anti-rabbit Alexafluor555, anti-mouse AlexaFluor488 and anti-mouse Alexafluor555 (Life Technologies, Karlsruhe, Germany) were used as secondary antibodies. All antibodies were diluted in antibody diluent (Life Technologies). Nuclei were counterstained with DAPI (Roche, Mannheim, Germany). Image acquisition was performed on a fluorescence inverted microscope (Axiovert 200 M, Zeiss, Gottingen, Germany) coupled to an AxioCam MR3 camera (Zeiss) using Axiovision software (version 4.7, Zeiss).

Loci Discovery and Functional Annotation (FUMA)

Genomic risk loci and lead variants were derived from FUnctional Mapping and Annotation of genetic associations (FUMAv1.3.1, http://fuma.ctglab.nl) [13] based on GWAS summary statistics obtained from BOLT-LMM. Independent significant SNPs were identified using the SNP2GENE function and were defined as SNPs with a P-value of $<5\times10^{-8}$ and independence to other genome wide significant SNPs at $r^2 < 0.6$ based on reference panel 1000 Genomes phase 3. Unique genomic risk loci were identified as LD blocks of independent significant SNPs that are >250kb apart, closer blocks were merged into a single locus. For each genomic risk locus one or more lead SNPs were identified among the independent significant SNPs and were defined as those that were independent from each other at

r^2<0.1. The independent significant variant with the lowest p-value at each risk locus was classified as the top lead variant for that respective locus and was followed up in replication genotyping (N=51 top lead variants) (Supplementary Table 5). Based on these independent significant SNPs, candidate SNPs used in subsequent functional annotations were identified as all SNPs that had a P-value of $<1\times10^{-5}$, MAF > 0.01 and were in LD of $r^2\geq0.6$ with at least one of the independent significant SNPs (Supplementary Table 1).

In order to identify candidate gene(s) at the respective genomic risk locus we followed i) **a manually curated selection process** based on local LD structure and supporting evidence from regulatory elements (eQTL and chromatin interaction) as detailed below and in Supplementary Table 3 and ii) we performed hypothesis-free functional and gene annotations based on the genomic positions of risk loci using **FUMA** [21] as manually curated selection process of candidate genes might not capture the full biology of the risk architecture. In the manually curated selection process, except for locus #25 (*PHGR1* and *DISP2*), a single curated candidate gene was assigned to each locus. The FUMA based approach resulted in 1080 unique mapped candidate genes (Supplementary Figure 6 & 7 and Supplementary Table 2).

Manually curated selection process (Supplementary Table 3): For lead variants located intronic, exonic, in the 3´or 5´UTR to a single annotated gene, the respective gene was identified as the candidate gene (30 out of 51 loci). For lead variants located upstream or downstream to a single annotated gene, this gene was assigned as the candidate gene, if the respective gene contains a variant with at least an r^2>0.5 to the lead variant at the locus (12 out of 51 loci). For loci, where variants in more than one neighbouring or overlapping transcript showed significant LD (r^2>0.5) to the lead variant, the transcript with higher expression in the tissue of interest was selected (3 out of 51 loci). For loci, where variants in more than one neighbouring or overlapping transcript showed significant LD (r^2>0.5) to the lead variant and no clear differences in expression were evident, the curated candidate gene was selected if additional regulatory evidence was present, i.e. an eQTL in a tissue of interest according to GTEx_v7 or a chromatin interaction pointed to a particular gene (1 out of 51 loci). If the lead variant was not located in a gene region and did not show significant LD to a variant in a neighbouring transcript, the impact of the lead variant and the variants in LD to the lead variant on

regulatory elements (eQTLs or chromatin interaction) was evaluated. If such elements were identified, the closest respective gene was annotated as the curated candidate gene (Criterion 5: 4 out of 51 loci). For the remaining loci, the closest transcript to the lead variant was annotated as the curated candidate gene if the distance to the variant was less than 1MB (Criterion 6: 1 out of 51 loci).

Using **FUMA**, all candidate SNPs were by default mapped to Ensembl genes (build 85) using ANNOVAR. The maximum physical distance to map SNPs to genes was 10kb. Intergenic SNPs were mapped to the two closest up and down stream genes thus with possible assignment to multiple genes. Candidate SNP and gene positions are referring to the human reference assembly (GRCh37/hg19) and are provided in Supplementary Table 1. Functional consequences of candidate SNPs were assessed using ANNOVAR, a tissue-specific *cis*-eQTL dataset (GTExV7, https://gtexportal.org) and 15-core chromatin states (ENCODE Project Consortium, 2012) [16]. Candidate genes with eQTL variants affecting gene expression in sigmoid colon at FDR<0.05 or at nominal p_eQTL <0.05 are shown in Table 2 and Table 3 and Supplementary Table 16. Enhancer and promoter regions were obtained from Roadmap Epigenomics Projects for 111 epigenomes [16]. Those regions were predicted using DNase peaks and core 15-state chromatin state model.

Gene set and pathway analysis

We used two gene set and pathway analysis approaches (MSigDB [14] and VEGAS2pathway [15]) to determine if the polygenic signal measured in the diverticular disease associated genes clustered in specific biological pathways. First, positional candidate genes from genomic risk loci showing consistency in effect direction between both discovery and replication stage, as outlined in Supplementary Table 5, were tested for overrepresentation with gene sets from the C5 collection: GO Biological Processes and gene sets from C2 sub-collection CP: Canonical pathways, curated in the Molecular Signatures Database (MSigDB 6.1; http://software.broadinstitute.org/gsea/msigdb/). A hypergeometric over-representation p-value: (k, K, n, N) was calculated for each gene set from K (the number of genes in the set), k (the number of genes in the intersection of the query set), n (the number of genes in comparison) and N (the number of all known human gene symbols). To control the false

positive error rates a FDR ($p < 0.05$) threshold was applied for significance. Results are provided in Supplementary Table 9 and 11. Secondly, we used VEGAS2pathway (https://vegas2.qimrberghofer.edu.au/) an extension of the VEGAS2 approach (VErsatile Gene-based Association Study, VEGAS2v02) to test hypothesis-free for pathway and gene set enrichment using the GWAS summary statistics obtained from BOLT-LMM v2.34. VEGAS2Pathway is a two-step pathway analysis strategy. Firstly, we calculated the gene-based test statistics for all genes using VEGAS2, which accounts for the LD between the SNPs within a gene through simulation. Variants lying within 50 kb on either side of a gene's transcription site (hg19 annotated RefSeq genes from UCSC table browser) were assigned to the respective gene to compute its association p value. This selection criterion was used to balance between inclusions of possible cis-regulatory variants and maintaining specificity of a gene. Secondly, for each of a set of pre-specified gene-sets, the relevant gene-based results were carried forward to compute a pathway-based test for gene-sets from the Gene Ontology, curated gene-sets from MSigDB; containing canonical pathways and gene-sets from BIOCARTA, REACTOME, KEGG databases, PANTHER, and pathway commons databases. Gene-sets were filtered to include only those with size between 10 and 1,000 genes. Overall, there were 6,212 gene-sets, including 18,399 genes with 511,336 annotations. Results are provided in Supplementary Table 11 and 12.

Enrichment analyses in cell lines and primary tissues.

GARFIELD (**G**WAS **A**nalysis of **R**egulatory or **F**unctional **I**nformation **E**nrichment with **LD** correction). The GARFIELD (http://europepmc.org/preprints/ppr7035) approach is independent from FUMA annotated genes and VEGAS2pathway results. GARFIELD used the whole number of GWAS SNPs as input then performs greedy pruning of GWAS SNPs (LD $r2 > 0.1$) and then annotates them based on functional information overlap. Functional enrichment analysis of diverticular disease variants in DNaseI Hypersensitive sites from ENCODE and Roadmap Epigenomics data are provided in Supplementary Table 13. GARFIELD allows for parallel enrichment analyses at multiple p-value sub-thresholds, which improves power to define statistically significant enrichment patterns by

increasing the number of variants tested. GARFIELD uses a nonparametric approach to weight GWAS findings with regulatory or functional annotations to find features relevant to a phenotype of interest. GARFIELD accounts for LD, minor allele frequency, matched genotyping variants and local gene density with the application of permutations to derive statistical significance. GARFIELD quantifies enrichment using odds ratios (OR) at various GWAS p-value cutoffs and assesses their significance by employing generalized linear model testing, while accounting for minor allele frequency, distance to nearest transcription start site and number of LD proxies (r2 > 0.8). The fold enrichment at various GWAS p-value cutoffs is indicated by color coding as described in the figure legend of Supplementary Figure 11. Fold enrichment values are shown in black and blue, for the GWAS P-value thresholds <1×10-8 and <1×10-5, respectively. The innermost and outermost dots along the inside edge denote significant enrichment for the cell type at <1×10-5 and < 1×10-8, respectively.

URLs:

ANNOVAR: http://annovar.openbioinformatics.org/en/latest/
BOLT-LMM: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/
FUMA: http://fuma.ctglab.nl/
GARFIELD: http://europepmc.org/preprints/ppr7035
GSEA/MSigDB 6.1: http://software.broadinstitute.org/gsea/msigdb/
GWAS Catalog: https://www.ebi.ac.uk/gwas/
LOCUSZOOM: http://locuszoom.org/
META: https://mathgen.stats.ox.ac.uk/genetics_software/meta/meta.html
METAL: https://genome.sph.umich.edu/wiki/METAL
OMIM: https://omim.org/
PLINK: http://zzz.bwh.harvard.edu/plink/
R, ‘ggplot2’ package: http://www.sthda.com/english/rpkgs/ggpubr/
SNPTEST: https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html
UCSC: https://genome.ucsc.edu/
VEGAS2: https://vegas2.qimrberghofer.edu.au/

ACKNOWLEDGEMENTS

The authors wish to thank all study participants, researchers, clinicians and administrative staff who contributed to this study. We thank Julia Wilking, Inka Geurink, Katrin Neblung-Masuhr, Sanaz Sedghpour Sabet, Tanja Wesse, Anja Tanck, Frank Lichte and Karin Stengel for expert technical assistance and Oddmund Nordgård (Stavanger University Hospital, Norway) for kindly providing the anti-PHGR1 antibody.

REFERENCES

1 Buch S, Schafmayer C, Völzke H, *et al.* A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nat Genet* 2007;**39**:995–9. doi:10.1038/ng2101

2 Buch S, Stickel F, Trépo E, *et al.* A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. *Nat Genet* 2015;**47**:1443–8. doi:10.1038/ng.3417

3 Reichert MC, Kupcinskas J, Krawczyk M, *et al.* A Variant of COL3A1 (rs3134646) Is Associated With Risk of Developing Diverticulosis in White Men. *Dis Colon Rectum* 2018;**61**:604–11. doi:10.1097/DCR.0000000000001001

4 Bonfiglio F, Zheng T, Garcia-Etxebarria K, *et al.* Female-Specific Association Between Variants on Chromosome 9 and Self-Reported Diagnosis of Irritable Bowel Syndrome. *Gastroenterology* 2018;**155**:168–79. doi:10.1053/j.gastro.2018.03.064

5 Henström M, Diekmann L, Bonfiglio F, *et al.* Functional variants in the sucrase–isomaltase gene associate with increased risk of irritable bowel syndrome. *Gut* 2018;**67**:263–70. doi:10.1136/gutjnl-2016-312456

6 Loh P-R, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;**47**:284–90. doi:10.1038/ng.3190

7 Hampe J, Wollstein A, Lu T, *et al.* An integrated system for high throughput TaqMan based SNP genotyping. *Bioinformatics* 2001;**17**:654–5.

8 Chang CC, Chow CC, Tellier LC, *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7. doi:10.1186/s13742-015-0047-8

9 Johnson EO, Hancock DB, Levy JL, *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet* 2013;**132**:509–22. doi:10.1007/s00439-013-1266-7

10 Liu JZ, Tozzi F, Waterworth DM, *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010;**42**:436–40. doi:10.1038/ng.572

11 Böttner M, Barrenschee M, Hellwig I, *et al.* The enteric serotonergic system is altered in patients with diverticular disease. *Gut* 2012.

12 Cossais F, Leuschner S, Barrenschee M, *et al.* Persistent Increased Enteric Glial Expression of S100β is Associated With Low-grade Inflammation in Patients With Diverticular Disease. *J Clin Gastroenterol* Published Online First: March 2018. doi:10.1097/MCG.0000000000001011

13 Watanabe K, Taskesen E, van Bochoven A, *et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;**8**:1826. doi:10.1038/s41467-017-01261-5

14 Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545–50. doi:10.1073/pnas.0506580102

15 Mishra A, MacGregor S. A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Res Hum Genet* 2017;**20**:1–9. doi:10.1017/thg.2016.100

16 Schmitt AD, Hu M, Jung I, Xu Z et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* 2016 Nov 15;17(8)