Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease

Authors:

James C. Lee[1]*, Daniele Biasci[1]*, Rebecca Roberts[2], Richard B. Gearry[2], John C. Mansfield[3], Tariq Ahmad[4], Natalie J. Prescott[5], Jack Satsangi[6], David C. Wilson[7], Luke Jostins[8], Carl A. Anderson[9], the UK IBD Genetics Consortium[10], James A. Traherne[11], Paul A. Lyons[1], Miles Parkes[1], Kenneth G.C. Smith[1].

Affiliations:

[1] Department of Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge, UK.

[2] University of Otago, Department of Medicine, Christchurch, New Zealand

[3] Institute of Genetic Medicine, Newcastle University, UK.

[4] University of Exeter Medical School, Exeter, UK.

[5] Department of Medical and Molecular Genetics, Faculty of Life Science and Medicine, King's College London, 8th Floor Guy's Tower, Guy's Hospital, London, UK.

[6] Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK.

[7] Paediatric Gastroenterology and Nutrition, Child Life and Health, College of Medicine and Veterinary Medicine, University of Edinburgh, Royal Hospital for Sick Children, Edinburgh, UK.

[8] Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK

[9] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

[10] A full list of consortium members is provided in the Supplementary Note

[11] Department of Pathology, University of Cambridge, Cambridge, UK.

* Co-first author

Correspondence should be addressed to James C. Lee (jcl65@cam.ac.uk) or Kenneth G.C. Smith (kgcs2@cam.ac.uk).

Abstract

For any immune-mediated disease, the main determinant of patient well-being is not the diagnosis itself, but the course the disease takes over time (prognosis)[1-3]. This varies substantially between patients for reasons that are poorly understood. Familial studies support a genetic contribution to prognosis[4-6], but little evidence has been found for a proposed association between prognosis and the burden of susceptibility variants[7-13]. To better characterise how genetic variation influences disease prognosis, we performed a within-cases genome-wide association study in two cohorts of patients with Crohn's disease. We identified four genome-wide significant loci, none of which showed any association with disease susceptibility. Conversely, the aggregated effect of all 170 disease susceptibility loci was not associated with prognosis. Together, these data suggest that the genetic contribution to prognosis in Crohn's disease is largely independent from the contribution to disease susceptibility, and point to a biology of prognosis that could provide new therapeutic opportunities.

Despite the success of genome-wide association studies (GWAS) in immune-mediated disease, a genetic contribution to aspects of disease other than susceptibility remains largely unstudied. One of the most important aspects from a clinical perspective is prognosis. For example in Crohn's disease (CD), a relapsing-remitting form of inflammatory bowel disease (IBD), some patients experience frequent relapses and require treatment with increasingly potent immunosuppressants and/or surgery, while others achieve prolonged remission without any additional therapy[1]. Such variability in prognosis occurs in most immune-mediated diseases[2,3] and can make the difference between an excellent long-term outcome or progressive disability and death[1-3].

What determines prognosis is poorly understood, although a genetic contribution has been proposed based on similar patterns of disease behaviour often being observed within families[4-6]. To date, investigation into the genetics of prognosis has largely focused upon susceptibility variants[7-9], based on the hypothesis that disease development and prognosis share a common genetic architecture[14]. Attempts to prove this hypothesis, however, have largely failed to demonstrate replicable associations[10-13]. For example, while *NOD2* variants were initially associated with increased need for surgery in CD[15], this has since been attributed to their association with ileal disease[10,11,16] (the distribution of CD that is more commonly treated with surgery as the procedure carries a lower morbidity). Indeed, a sub-phenotype analysis of IBD recently showed that susceptibility variants only explain a minority of phenotypic variance (and mainly to disease location) and that after conditioning on this there was little/no association with disease behaviour[16].

An alternative possibility is that the genetic contribution to prognosis might be distinct from that which confers disease susceptibility. Indeed, the idea of a separate biology of prognosis would fit with the observation that CD8 T-cell exhaustion correlates with outcome, but not

diagnosis, in multiple immune-mediated diseases[17], and the identification of a SNP in *FOXO3* that associates with prognosis, but not susceptibility, in several diseases including CD[18].

To better understand the genetic contribution to prognosis in CD, we performed a within-cases GWAS in two CD cohorts and meta-analysed the results. To do this, we first identified patients at opposite ends of the prognostic spectrum for comparison. This approach enriches for alleles influencing prognosis and makes associated variants easier to detect because the odds ratios between extreme groups is higher than in the total sample[19]. Poor prognosis CD was defined as frequently-flaring, treatment-refractory disease that required consecutive treatment with ≥2 immunomodulators, or ≥2 abdominal surgeries, or a combination of the two[18]. Good prognosis CD was defined as indolent disease that had not required treatment with immunomodulators or surgery despite ≥4 years follow up[18] (median 12 years). These definitions were applied to two cohorts of CD cases: the first from a previous GWAS[20] (669 poor prognosis cases, 389 good prognosis cases) and the second who were genotyped using UK Biobank Axiom arrays (1093 poor prognosis cases, 583 good prognosis cases). Altogether, 19.2% of cases had good prognosis CD and 32.2% had poor prognosis CD. Patients with an intermediate phenotype were excluded (n=2,794; 48.6%). At genome-wide significance ($P < 5 \times 10^{-8}$) the combined dataset had ≥80% power to detect variants (minor allele frequency [MAF] >20%) with odds ratios ≥1.33.

Standard quality control checks for SNPs and samples were performed prior to imputation against the UK10K reference panel[21] (Online Methods). Following imputation quality control (Online Methods, Supplementary Table 1) we tested 7.0 million and 7.5 million SNPs for association in cohorts 1 and 2 respectively, and meta-analysed the results. Cluster plots for associated SNPs were inspected to confirm the genotyping and, if imputed, were genotyped

to confirm the imputation (Online Methods, Supplementary Table 1). The genomic control inflation factor (λ_{GC}) was 1.023 in the combined analysis (Supplementary Fig. 1).

Four loci passed genome-wide significance. These association signals were located in *FOXO3*, *XACT*, a region upstream of *IGFBP1*, and the MHC region (Table 1, Fig. 1A, Supplementary Table 2). The associated *FOXO3* haplotype, which was previously identified in a candidate gene study using the same definitions[18], regulates *FOXO3* expression and controls a TGFβ1-dependent pathway that limits inflammatory responses in monocytes[18]. By genotyping or imputing 115 of 118 SNPs at this locus we mapped the peak association signal to an intronic region that contained enhancer marks (H3K4me1, H3K27ac, p300 binding) and transcription factor binding sites, consistent with the reported eQTL[18] (Fig. 1B, Supplementary Fig. 2).

The second association signal was located within *XACT* (Fig. 1C), a gene which encodes a long non-coding RNA that is only expressed from the active X chromosome[22], and which resides over 630Kb from the nearest protein-coding gene. This association was detectable in both males and females (Supplementary Table 2). Although *XACT* has been studied in pluripotent stem cells[22], its function in differentiated cells is unknown. Given the association with CD prognosis, however, it was striking that across multiple human tissues, *XACT* was most highly expressed within the intestine (Supplementary Fig. 3).

The third association signal was located immediately upstream of *IGFBP1* in a region that also contains *IGFBP3* (Fig. 1D). The associated SNP at this locus (rs75764599:G>A, g.45899250G>A, hg19) was conspicuous for its lack of linkage disequilibrium (LD) with local SNPs, and had to be directly genotyped in cohort 1. Indeed, although long-range LD enabled imputation in cohort 2 (INFO_score 0.82), we genotyped the SNP in this cohort as well to confirm the association (Supplementary Table 1). This relative lack of LD is

sometimes observed at low frequency SNPs, and implies that rs75764599:G>A is likely to be the causal variant. IGFBP-1 and IGFBP-3 belong to a family of proteins that bind and prolong the half-lives of insulin-like growth factors I and II; proteins involved in processes including immunity and longevity[23]. The *IGFBP1* locus has also been associated with anti-citrullinated peptide antibodies in rheumatoid arthritis (RA)[24]. In RA, these antibodies predict a poor prognosis[24], suggesting that this region might influence prognosis in CD and RA through a common pathway, as was shown for *FOXO3*[18].

The final association signal was located in the MHC and stretched from *HLA-B* to the *HLA-DR* genes (Fig. 1E). Despite the breadth of this signal, conditioning on the lead SNP revealed only one associated haplotype (Fig. 2A). To assess whether specific HLA alleles were responsible for this genetic association, we imputed classical 4-digit HLA alleles and detected associations at multiple class I and class II genes. Notably, all of the associated alleles belonged to a single multigene haplotype, "ancestral MHC 8.1" (AH8.1; Fig. 2B). Indeed, the strength of each allelic association correlated with how specific that allele was to AH8.1 (Supplementary Fig. 4). To clarify the contribution of AH8.1 to the genetic signal, we performed cross-conditioning to assess the residual associations when the SNP analysis was conditioned on the lead AH8.1 allele (HLA-B*08:01) and vice versa. In both directions, cross-conditioning abrogated any association, demonstrating that inheritance of AH8.1 was the major contributor to the genetic signal (Fig. 2C, Supplementary Table 3). Carriers of this haplotype are known to exhibit impaired responses to vaccination[25] and defects in T-cell activation[26,27], consistent with HLA-specific differences in antigen presentation. The increased frequency of this haplotype in good prognosis CD is therefore consistent with the notion that differential T-cell activation influences prognosis in immune-mediated disease[17]. Interestingly, no association was observed at HLA-DRB1*01:03, the HLA allele that has by far the strongest association with CD susceptibility of any HLA allele or SNP[28] (Fig. 2B,

Supplementary Table 3). Similarly, the prognosis-associated MHC SNPs did not overlap with known disease susceptibility SNPs (Supplementary Fig. 5).

We next investigated whether these associations were dependent on the criteria used to define prognosis. We found that the results could not be attributed to differences in disease location (Supplementary Table 4) or follow-up (Supplementary Table 5) and that 3 of the 4 associations remained genome-wide significant if another definition of poor prognosis was used; abdominal surgery within 2 years of diagnosis[29] (Supplementary Table 6). These associations are therefore likely to be involved in the biology that determines prognosis in CD, as was shown for *FOXO3*[18]. To further explore this underlying biology, we examined whether genes tagged by an extended list of prognosis-associated SNPs (meta $P < 10^{-5}$) were enriched within specific biological pathways (Online Methods). Strong enrichment was observed for pathways that regulate innate and adaptive immune responses, and responses to micro-organisms (Fig. 3A). Similar results were obtained if SNPs within the extended MHC were removed (Supplementary Table 7). To identify the cell-types involved, we examined for tissue-specific enrichment using an expression atlas of primary human cells[30]. Strong enrichment was observed in macrophages, followed by weaker signals in monocytes and dendritic cells (Fig. 3B, Supplementary Table 8). These mononuclear phagocytes play important roles in innate immunity and can initiate adaptive immune responses. Monocytes are also the cell-type in which the *FOXO3* variant was biologically relevant[18]. Finally, we performed a protein-protein interaction analysis and found that many genes at associated loci interact either directly or indirectly, suggesting that there may be underlying cellular pathways that are important in disease prognosis (Supplementary Fig. 6).

Interestingly, the 4 prognosis-associated haplotypes have not been associated with CD susceptibility in any GWAS, including a recent meta-analysis[31] (Supplementary Table 9).

Accordingly, if they were to influence disease susceptibility, their effect size would be negligible. Conversely, after stratifying on disease location – to control for associations of some loci (e.g. *NOD2*) with distributions of CD that could affect assessments of clinical outcome[10,11,16] – none of the 170 susceptibility variants[31,32] were associated with prognosis, even if more relaxed statistical thresholds were used (Online Methods, Supplementary Table 10). There was also no correlation between the power to detect effects at these SNPs and the observed *P* values, consistent with the hypothesis that susceptibility variants are not individually associated with prognosis (Supplementary Fig. 7). Similarly, genome-wide LD score regression confirmed that there was no significant overlap between the genetic bases of susceptibility and prognosis in CD if disease location was taken into account (Genetic Correlation = -0.51, Standard Error = 0.33, $P = 0.121$).

To determine whether susceptibility variants might collectively influence prognosis, we calculated genetic risk scores for each patient, but again no significant differences were observed between the good and poor prognosis subgroups (Fig. 4). This result did not change if an extended list of CD-associated SNPs was used[33], based on variants with meta $P < 10^{-4}$ (Supplementary Fig. 8). Together, these data suggest that susceptibility alleles do not meaningfully contribute to prognosis in CD, although a weak effect cannot be excluded due to the power of this analysis. If such an effect were to occur, however, it would be trivial compared to the effects of non-susceptibility loci that are associated with prognosis.

Collectively, these data establish that the main genetic contribution to prognosis comes from loci that are distinct from those that drive disease susceptibility, and demonstrate that disease initiation is not only temporally distinct from active symptomatic disease, but also appears to be governed by separate genetics. This provides a starting point for better understanding the biology that determines prognosis in CD, and could lead to new and potentially improved

opportunities for therapeutic intervention. This also has implications for "personalised medicine", although any genetic classifier would need to work in unselected CD cases (in whom the odds ratios at prognosis-associated variants will be smaller than were observed here). More generally, this work illustrates the value of re-analysing existing GWAS data using carefully selected subphenotypes. Indeed, providing sufficiently detailed clinical data are available, this approach should be broadly applicable, and could yield important new insights across multiple diseases.

Acknowledgements

We thank Lucy Hildyard, Emma Gray and other members of the Wellcome Trust Sanger Institute DNA team for their help with sample co-ordination, and Abigail Groff and Catherine Weiner for critical reading of the manuscript. This work was supported by the NIHR Cambridge Biomedical Research Centre (in particular John Todd and the NIHR BRC Genomics Theme), Crohn's and Colitis UK (Medical Research Award M/14/2), the Evelyn Trust (17/07), and the Medical Research Council (Programme Grant MR/L019027/1). J.C.L. is supported by a Wellcome Trust Intermediate Clinical Fellowship (105920/Z/14/Z) and D.B. by a Marie Curie PhD Fellowship (TranSVIR FP7-PEOPLE-ITN-2008 #238756). C.A.A. is supported by the Wellcome Trust (098051). K.G.C.S. is an NIHR Senior Investigator. This study makes use of data generated by the UK10K Consortium, derived from samples from ALSPAC and DTR cohorts. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust (WT091310).

Author contributions

The experiment was conceived by J.C.L., M.P. and K.G.C.S.. J.C.L, D.B. and P.A.L. designed the analysis. D.B. performed the analysis with input from J.C.L., L.J., C.A.A., J.A.T and P.A.L.. Patient samples and phenotype data were provided by J.C.L., R.R., R.B.G., J.C.M., T.A., N.J.P., J.S., D.C.W., M.P. and other members of the UK IBD Genetics Consortium. J.C.L. and K.G.C.S. wrote the manuscript with input from D.B., P.A.L., and M.P.. All authors reviewed and approved the manuscript prior to submission.

Competing Financial Interests

The authors declare no competing financial interests

References

1 Jess, T. *et al.* Changes in clinical characteristics, course, and prognosis of inflammatory bowel disease during the last 5 decades: a population-based study from Copenhagen, Denmark. *Inflamm Bowel Dis* **13**, 481-489 (2007).

2 Pincus, T. Long-term outcomes in rheumatoid arthritis. *Br J Rheumatol* **34 Suppl 2**, 59-73 (1995).

3 Weinshenker, B. G. *et al.* The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability. *Brain* **112 (Pt 1)**, 133-146 (1989).

4 Satsangi, J., Grootscholten, C., Holt, H. & Jewell, D. P. Clinical patterns of familial inflammatory bowel disease. *Gut* **38**, 738-741 (1996).

5 Chataway, J. *et al.* Multiple sclerosis in sibling pairs: an analysis of 250 families. *J Neurol Neurosurg Psychiatry* **71**, 757-761 (2001).

6 Jawaheer, D., Lum, R. F., Amos, C. I., Gregersen, P. K. & Criswell, L. A. Clustering of disease features within 512 multicase rheumatoid arthritis families. *Arthritis Rheum* **50**, 736-741 (2004).

7 Weersma, R. K. *et al.* Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. *Gut* **58**, 388-395 (2009).

8 Hilven, K., Patsopoulos, N. A., Dubois, B. & Goris, A. Burden of risk variants correlates with phenotype of multiple sclerosis. *Mult Scler* (2015).

9 Chibnik, L. B. *et al.* Genetic risk score predicting risk of rheumatoid arthritis phenotypes and age of symptom onset. *PLoS One* **6**, e24380 (2011).

10 Ananthakrishnan, A. N. *et al.* Differential effect of genetic burden on disease phenotypes in Crohn's disease and ulcerative colitis: analysis of a North American cohort. *Am J Gastroenterol* **109**, 395-400 (2014).

11 Jung, C. *et al.* Genotype/phenotype analyses for 53 Crohn's disease associated genetic polymorphisms. *PLoS One* **7**, e52223 (2012).

12 Jensen, C. J. *et al.* Multiple sclerosis susceptibility-associated SNPs do not influence disease severity measures in a cohort of Australian MS patients. *PLoS One* **5**, e10003 (2010).

13 Scott, I. C. *et al.* Do Genetic Susceptibility Variants Associate with Disease Severity in Early Active Rheumatoid Arthritis? *J Rheumatol* **42**, 1131-1140 (2015).

14 Plomin, R., Haworth, C. M. & Davis, O. S. Common disorders are quantitative traits. *Nat Rev Genet* **10**, 872-878 (2009).

15 Helio, T. *et al.* CARD15/NOD2 gene variants are associated with familially occurring and complicated forms of Crohn's disease. *Gut* **52**, 558-562 (2003).

16 Cleynen, I. *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387**, 156-167 (2016).

17 McKinney, E. F., Lee, J. C., Jayne, D. R., Lyons, P. A. & Smith, K. G. T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *Nature* **523**, 612-616 (2015).

18 Lee, J. C. *et al.* Human SNP Links Differential Outcomes in Inflammatory and Infectious Disease to a FOXO3-Regulated Pathway. *Cell* **155**, 57-69 (2013).

19 Van Gestel, S., Houwing-Duistermaat, J. J., Adolfsson, R., van Duijn, C. M. & Van Broeckhoven, C. Power of selective genotyping in genetic association analyses of quantitative traits. *Behav Genet* **30**, 141-146 (2000).

20 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

21 Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).

22 Vallot, C. *et al.* Erosion of X Chromosome Inactivation in Human Pluripotent Cells Initiates with XACT Coating and Depends on a Specific Heterochromatin Landscape. *Cell Stem Cell* **16**, 533-546 (2015).

23 Franceschi, C. *et al.* Genes involved in immune response/inflammation, IGF1/insulin pathway and response to oxidative stress play a major role in the genetics of human longevity: the lesson of centenarians. *Mech Ageing Dev* **126**, 351-361 (2005).

24 Padyukov, L. *et al.* A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Ann Rheum Dis* **70**, 259-265 (2011).

25 Egea, E. *et al.* The cellular basis for lack of antibody response to hepatitis B vaccine in humans. *J Exp Med* **173**, 531-538 (1991).

26 Modica, M. A., Cammarata, G. & Caruso, C. HLA-B8,DR3 phenotype and lymphocyte responses to phytohaemagglutinin. *J Immunogenet* **17**, 101-107 (1990).

27 Candore, G. *et al.* T-cell activation in HLA-B8, DR3-positive individuals. Early antigen expression defect in vitro. *Hum Immunol* **42**, 289-294 (1995).

28 Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet* **47**, 172-179 (2015).

29 Sands, B. E. *et al.* Risk of early surgery for Crohn's disease: implications for early treatment strategies. *Am J Gastroenterol* **98**, 2712-2718 (2003).

30 Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics* **14**, 632 (2013).

31 Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* (2015).

32 Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124 (2012).

33 Wei, Z. *et al.* Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet* **92**, 1008-1012 (2013).

Figure Legends

Figure 1. Within-cases GWAS identifies four loci that are associated with prognosis in CD.

(a) Plot of genome-wide association results. $-\log_{10}$(*P* values) from the Wald statistic (logistic regression model) are plotted against chromosomal position for the combined association analysis (n = 1,762 poor prognosis CD, 972 good prognosis CD). Each point represents a SNP. Dotted red line indicates genome-wide significance threshold. The four new loci identified in this study are indicated. **(b – e)** Characteristics of the associated genomic regions. Upper panel, chromosomal position; middle panel, $-\log_{10}$(*P* values) for individual SNPs at each locus (left y axis), rate of recombination indicated by red line (right y axis); lower panel, gene position within the locus (in Panel **e** only selected genes from class I and II regions are shown for clarity). SNPs are coloured according to LD with the most associated variant.

Figure 2. Association signal at the MHC is linked to the ancestral 8.1 haplotype

(a) The residual MHC association after conditioning on the lead SNP (rs9279411). **(b)** Allelic associations at class I and II HLA genes. For each allele, odds ratios (OR), 95% confidence intervals, and *P* values are shown. The significance threshold was corrected for the number of independent tests ($P < 2.5 \times 10^{-4}$). Alleles that are components of the ancestral MHC AH8.1 haplotype are shaded blue (full AH8.1 annotation: HLA A*01:01, C*07:01, B*08:01, DRB1*03:01, DRB3*01:01, DQA1*05:01, DQB1*02:01). HLA DRB1*0103 (the strongest IBD susceptibility allele, ref 28) is included for comparison (shaded grey). **(c)** The residual MHC association after cross-conditioning on the lead HLA allele (HLA-B*08:01). In panels **a** and **c** the $-\log_{10}(P)$ of SNPs are plotted against chromosomal position. Only selected genes from class I and II regions are shown for clarity.

Figure 3. Pathway analysis implicates regulation of immune responses and mononuclear phagocytes in CD prognosis

(a) Horizontal bar plot of –log10 (*P* values) for the top 40 most enriched pathways in an analysis of 29 LD-pruned prognosis-associated SNPs (meta $P < 1 \times 10^{-5}$) across 1,751 pathways annotated by Gene Ontology. (**b**) Bar plot of –log10 (*P* values) for 155 specific cell-types and conditions. Significant enrichment was observed in “monocyte-derived macrophages stimulated by M-CSF” and “monocyte-derived macrophages stimulated by M-

CSF and interferon-gamma". Other non-immune cell-types included neural, skin, lung, liver, stem cells, smooth muscle, stromal cells, bone marrow progenitors, endothelial and epithelial cells. Dotted lines represent Bonferroni-corrected significance thresholds. Analyses performed using SNPsea.

Figure 4. Distribution of CD susceptibility alleles does not differ between the prognostic subgroups

(a) "Box and Whiskers" plot of weighted Genetic Risk Scores between good prognosis and poor prognosis CD subgroups. Box represents median and interquartile range. Whiskers represent maximum and minimum values. Weights were based on beta coefficients calculated from odds ratios reported in Liu et al. (2015) for 170 CD susceptibility SNPs. n = 2,413 **(b)** Distribution of unweighted susceptibility allele counts between good prognosis and poor prognosis CD subgroups. Purple histogram bars represent the poor prognosis CD subgroup, yellow histogram bars represent the good prognosis CD subgroup. Statistical significance was assessed using unpaired two-tailed Student's t-test and was stratified for disease location.

Table 1. Association results for Crohn's disease prognosis loci

Chr	Position (Mb)	SNP	Risk allele	RAF (ind.)	OR	95% CI	*P* cohort 1	*P* cohort 2	*P* combined	Candidate gene or region
X	112.9	rs5929166	A	0.03	0.33	0.23-0.48	1.12 x 10-5	5.01 x 10-5	4.56 x 10-9	*XACT*
6	31.7	rs9279411	- [a]	0.15	0.60	0.50-0.71	1.35 x 10-5	7.22 x 10-5	5.46 x 10-9	MHC
6	109.0	rs147856773	GTG[b]	0.12	0.57	0.47-0.70	7.76 x 10-4	4.29 x 10-6	1.31 x 10-8	*FOXO3*
7	45.9	rs75764599	A	0.01	3.02	2.04-4.49	1.73 x 10-3[c]	6.25 x 10-6	4.32 x 10-8	*IGFBP1/ IGFBP3*

Crohn's disease prognosis SNPs that met genome-wide significance ($P < 5 \times 10^{-8}$) in the combined analysis and nominal significance ($P < 0.05$) in both individual cohorts.

The odds ratio is presented with respect to the minor allele and the risk of poor prognosis CD.

Allele frequency data is presented for the good prognosis CD cohort.

[a] AG deletion

[b] Insertion

[c] rs75764599 could not be imputed in cohort 1 because of low linkage disequilibrium at this locus, and was therefore directly genotyped using a TaqMan SNP genotyping assay.

Chr, Chromosome; RAF, Risk Allele Frequency; ind., Indolent (good prognosis) CD; OR, odds ratio; 95% CI, 95% confidence interval for OR.

Online Methods

Study subjects

Following ethical approval by Cambridge MREC (reference: 03/5/12) and Mater Health Services, Christchurch Hospital, two cohorts of unrelated CD cases of Northern European descent were considered (cohort 1: n = 1,748, cohort 2: n = 3,999). These cases were enrolled by the UK IBD Genetics Consortium (n = 5,521) and the University of Otago, New Zealand (n = 226) and all provided written informed consent. All cases were treated using a treatment strategy in which immunosuppression is incrementally escalated, but only in response to persistently flaring disease. Subgroups of patients with a poor prognosis or a good prognosis were identified using phenotype data. Poor prognosis CD was defined as disease that had required ≥ 2 immunomodulators or ≥ 2 abdominal surgeries or a combination of these (e.g. 1 immunomodulator and 1 intestinal resection). Following quality control, 1762 cases met this definition (cohort 1: n = 669, cohort 2: n = 1,093). Good prognosis CD was defined as disease of ≥ 4 years duration (median 12 years) that had not required immunomodulators or intestinal resections. Following quality control, 972 cases met this definition (cohort 1: n = 389, cohort 2: n = 583). Patients who did not meet either criteria, or where additional immunomodulators or surgery had only been required because of drug intolerance or complications of earlier surgery respectively, were excluded.

Genotyping

Samples in cohort 1 were genotyped as part of an earlier GWAS[20]. For this cohort, normalised intensity data were downloaded from the European Genome-Phenome Archive (EGA). Samples in cohort 2 were genotyped using the UK Axiom Biobank array

(Affymetrix). 27 SNPs (including rs75764599; *IGFBP1* locus) demonstrated modest evidence of association in the cohort 2 ($P < 5 \times 10^{-5}$) but could not be imputed in cohort 1. These markers were directly genotyped in cohort 1 using TaqMan SNP genotyping assays. Imputed low frequency SNPs (MAF 1-5%) that showed evidence of association were also directly genotyped to confirm the imputation.

Data processing

1. Cohort 1

Marker Quality Control: Markers that failed quality control (QC) checks in the original GWAS study[20] were excluded. Genotypes were called from the normalised intensity data using CHIAMO. SNPs with overall missingness > 0.05 or differential missingness > 0.02 were excluded. Because there were no healthy controls in this study (in whom deviation from Hardy–Weinberg equilibrium [HWE] is usually assessed[34]) only SNPs with marked deviation from HWE were excluded ($P < 10^{-10}$).

Sample Quality Control: Samples that failed QC criteria in the original study[20] were excluded. The reported gender of each case was checked against the genotype-inferred gender using Plink[35]. Samples with gender mismatch were excluded. Samples with a genotype missingness rate > 0.05 or heterozygosity rate greater than two standard deviations from the mean were also excluded[34]. To identify related samples, 119,811 LD-independent SNPs were selected and pairwise identity-by-descent was estimated using Plink[35] (pi-hat threshold 0.1875). For any related samples, one case was randomly selected and kept and the other(s) were excluded.

Geographical outliers. A pruned version of the dataset, containing 56,919 LD-independent SNPs was merged with the 1000 Genomes Project dataset[36]. Principal Component Analysis (PCA) was used to identify and exclude geographical outliers.

In total, 62 samples in cohort 1 failed QC.

2. Cohort 2

In addition to standard pre-GWA QC measures, additional quality controls specific to the Axiom genotyping array were recommended by the manufacturers (Affymetrix). These were performed prior to standard QC.

Axiom-specific sample QC. Genotyping arrays were batched by processing date, and analysed separately as per the manufacturer's instructions. The Dish Quality Control (DQC) metric (the recommended quality metric for Axiom genotyping arrays) was computed for each sample. Samples with DQC < 0.82 were excluded, as per the manufacturer's recommendations. Samples with a preliminary call rate ≤ 97% at a set of 20,000 validated autosomal markers were also excluded. Plates with an average preliminary call rate of passing samples < 98.5% were excluded, as per the manufacturer's recommendations.

Axiom-specific marker QC: Genotypes were called separately for each batch using the *apt-probeset-genotype* utility with default parameters. QC metrics for each marker were computed using the *Ps_Metrics* function in *SNPolisher* in R and used to classify markers based on the quality of signal. Markers classified in categories *PolyHighRes*, *NoMinorHom* and *MonoHighRes* were selected for further analysis. 500 probe-sets from each selected category were randomly selected and inspected to confirm the classification.

Standard sample and marker quality control. In addition to the Axiom-specific QC measures, standard pre-GWA QC steps were applied as described for cohort 1.

In total, 71 samples in cohort 2 failed QC.

Genotype imputation

Genotype imputation was performed using IMPUTE2[37] after estimating haplotypes with SHAPEIT2[38] and the UK10K reference panel[21]. Default SHAPEIT2 parameters were modified (*-states 500 -burn 10 -prune 10 -main 50*) to improve accuracy by increasing the number of conditioning states[39]. Samples from each cohort were pre-phased and imputed in a single batch, to avoid batch effects attributable to the imputation process. To make the computation feasible, the dataset was divided into ~3000 overlapping 1Mb regions. IMPUTE2 was used with the option *-buffer 2000* to leverage on a genomic window of 5Mb (1Mb analysed region plus 2 Mb buffer regions on either side). Results were controlled and reassembled in R. Phasing and imputation of the X Chromosome was performed using the *-chrX* flag and the gender of each sample was provided. Imputed variants with MAF < 1% and/or INFO score < 0.8 where excluded. To estimate imputation accuracy, imputed genotype calls at 99,124 SNPs were compared with direct genotyping data in 880 cases (Illumina Immunochip[32]). Mean concordance was 99.3% (Supplementary Table 1).

Power calculations

Power calculations were performed using the GPC function in the GeneticsDesign package in R. The prevalence of poor prognosis CD was calculated using the total number of poor prognosis cases (prior to QC) as a proportion of all of the cases considered (1,848/5,747).

Statistical analysis

Association tests and meta-analysis. A frequentist association test between genotypes and the binary phenotype was performed using SNPTEST under an additive genetic model. Ten principal components were included as covariates in the logistic regression model in order to control for population stratification, although genomic inflation was acceptable even before this correction was applied (unadjusted λ_{GC} <1.05). The λ_{GC} in the combined analysis following principal component correction was 1.023 (1.012 and 1.022 for cohorts 1 and 2 respectively; Supplementary Fig. 1). Genotype uncertainty, generated by the calling algorithm or by imputation, was factored into the association test using the *-method score* option in SNPTEST. For markers on the X chromosome, the association test was performed assuming a standard model of complete X inactivation, an equal effect size in men and women, and providing gender information for each sample. In this model, male genotypes were encoded as 0 / 1 and females as 0 / ½ / 1. . Meta-analysis was performed using a fixed-effects model and default parameters in META[40]. Cluster plots for associated SNPs were visually inspected to verify genotype calls and, if imputed, were directly genotyped to confirm the result (TaqMan SNP Genotyping Assay, Supplementary Table 1). Cross-conditioning analysis of the MHC region was performed in SNPTEST using the combined dataset and including the lead HLA allele (HLA-B*08:01) as a covariate. Cross-conditioning analysis of HLA allelic associations was performed using logistic regression in the combined dataset and including genotype at the lead SNP (rs9279411) as a covariate (Supplementary Table 2).

Zero-inflated Poisson regression analysis. To determine whether the results were influenced by differences in the length of follow up in the poor prognosis CD subgroup, a zero-inflated Poisson regression model was fitted to the count data of total treatment escalations per patient

(immunomodulators and surgery) using the *zeroinfl* function in pscl in R[41]. This was confirmed to be the most appropriate model for the data using Vuong's closeness test. Poisson regression was then performed and disease duration was included within the regression terms to assess if prognosis-associated SNPs were associated with treatment escalation rate independent of disease duration (Supplementary Table 5).

Disease-associated SNP analysis. 170 CD susceptibility loci were identified from a recent large meta-analysis[31] (Supplementary Table 10). Because several of these SNPs have been associated with specific anatomical distributions of CD, which can confound assessments of disease course if there are any differences in disease location between the prognostic subgroups[7,10,11,16], we stratified this analysis for disease location in addition to the top 10 principal components. Disease location data was available for 88.2% of samples (n = 2,413). Samples for whom disease location data were not available were excluded. To improve the power to detect smaller effects, a candidate gene analysis was performed including only the 170 susceptibility variants, and correcting for multiple testing using the Bonferroni method or false discovery rate (FDR $q < 0.25$). This analysis was predicted to have ≥ 80% power to detect variants (MAF > 20%) with odds ratios ≥ 1.25, and should have been adequately powered to identify effects that were previously reported in unstratified analyses[42]. For this analysis, approximate Bayes factors were also calculated using the method described in ref 43. The prior variance was based on there being an equivalent effect size as was observed in a recent susceptibility meta-analysis[31].

HLA allelic association analysis. The MHC region on chromosome 6 (chr6:25092012-35092011) was extracted from the post-QC dataset and used to impute 199 classical HLA alleles using HLA*IMP:02[44]. Alleles imputed with confidence > 80% were retained for association testing. Imputed HLA alleles for both cohorts were combined into a single dataset

and univariate logistic regression was performed without covariates. Nominal *P* values were calculated from the Wald statistic[45]. The statistical significance threshold was adjusted using a Bonferroni correction for multiple testing. Reference allele frequency and haplotype data was obtained from the National Bone Marrow Donor Program https://bioinformatics.bethematchclinical.org/hla-resources/haplotype-frequencies (Six-Locus High Resolution HLA A~C~B~DRB3/4/5~DRB1~DQB1 Frequencies).

RNA Expression Analysis

Raw RNA sequencing data (fastq files) were downloaded from publically available repositories: GSE45326[46] (GEO) and E-MTAB-513 (ArrayExpress). Reads were aligned to the UCSC Refseq genes (hg19) using Star[47] and quantified, normalised and analysed using Cufflinks v2.2.1[48] (Seven Bridges Genomics platform, https://www.sbgenomics.com/).

Pathway Analysis

(i) *SNPsea:* Pathways and cell-types that were likely to be affected by prognosis-associated loci, were identified used SNPsea[49]. In separate analyses we assessed the enrichment of an LD-pruned list of 34 prognosis-associated SNPs ($r^2 > 0.6$, meta $P < 10^{-5}$) in 1751 Gene Ontology pathways and 155 primary human cell-types and conditions (comprising 745 individual samples)[30]. Empirical P values were calculated by permutation using null SNP sets (matched for the number of linked genes) from a list of LD-pruned SNPs (subset of SNPs in 1000 Genomes Project). Reference data including the null SNP sets, NCBI gene intervals, Gene Ontology pathways, and SNP linkage interval data were provided in SNPsea software.

(ii) *DAPPLE (Disease Association Protein-Protein Link Evaluator):* Physical interactions between proteins encoded by genes at prognosis-associated loci were assessed using DAPPLE[50], based on a list of SNPs with meta $P < 10^{-4}$.

Genetic risk scores. Weighted Genetic Risk Scores were calculated using the PredictABEL package[51] using SNPs and their corresponding beta coefficients as reported in Liu et al. 2015 (for the 170 CD GWAS hits; Fig. 4) or Wei et al. 2013 (who generated a 573 SNP classifier from 10,799 SNPs with $P < 1 \times 10^{-4}$ in a cohort of ~17,000 CD cases and ~22,000 controls; Supplementary Fig. 8). Unweighted susceptibility allele counts were calculated by summing the number of risk alleles without including a beta coefficient. Statistical significance was assessed using an unpaired two-tailed Student's t test. Data were stratified for disease location and the results were combined using META[40].

LD score regression

LD score regression[52] was performed using LD Hub[53], a centralised database of summary-level GWAS results that provides a web interface for LD score regression. For this analysis the MHC region was removed – as recommended by LD Hub (because its complex genetic architecture is not adequately captured by simple LD). Genetic correlation was then calculated between the prognosis meta-analysis results (stratified for disease location) and the summary statistics from the largest CD susceptibility meta-analysis to date[31].

Data availability

Genotyping data that support the findings of this study have been deposited in the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/home) with the accession codes: EGAD00000000005 [cohort 1] and EGAS00001002147 [cohort 2]). Summary

statistics can also be downloaded from

ftp://ftp.sanger.ac.uk/pub/project/humgen/summary_statistics/human/2016-10-12/CD_prognosis_GWA_results.csv.zip.

Methods-only References

34 Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-1573 (2010).

35 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

36 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).

37 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

38 O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* **10**, e1004234 (2014).

39 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).

40 Liu, J. Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-440 (2010).

41 Zeileis, A., Kleiber, C. & Jackman, S. Regression Models for Count Data in R. *Journal of Statistical Software* **27** (2008).

42 Cleynen, I. *et al.* Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project. *Gut* **62**, 1556-1565 (2013).

43 Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81**, 208-227 (2007).

44 Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput Biol* **9**, e1002877 (2013).

45 Wellek, S. & Ziegler, A. Cochran-Armitage test versus logistic regression in the analysis of genetic association studies. *Hum Hered* **73**, 14-17 (2012).

46 Nielsen, M. M. *et al.* Identification of expressed and conserved human noncoding RNAs. *RNA* **20**, 236-251 (2014).

47 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

48 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).

49 Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30**, 2496-2497 (2014).

50 Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).

51 Kundu, S., Aulchenko, Y. S., van Duijn, C. M. & Janssens, A. C. PredictABEL: an R package for the assessment of risk prediction models. *Eur J Epidemiol* **26**, 261-264 (2011).

52 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-1241 (2015).

53 Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* (2016).