Supplementary Methods

Fine-mapping methods overview

The aim of Bayesian fine-mapping in complex disease is to define the number of independent signals at a locus, generate credible sets of variants that could potentially be causal for each signal, and assign posterior probabilities that each of those variants is causal. To ensure the reliability of our results, we used three different fine-mapping methods, and took a consensus of the results across methods. Our procedure for combining the results of the three methods is described in the Online Methods.

Our three methods use different models of genetic risk (multinomial logistic models for methods 1 and 2, liability threshold model for method 3), different ways of handling multiple phenotypes (using model selection for method 1, using correlated effect sizes for method 2 and analyzing each phenotype independently for method 3), different priors on effect sizes (a flat improper prior for method 1, a correlated normal prior for method 2 and a double exponential prior for method 3) and different ways of fitting the models and exploring the parameter space (a steepest descent approximation for method 1, a Laplace integral approximation for method 2 and an MCMC sampler for method 3). We describe each method below in a structured format that starts with a brief summary of the method, followed by its relationship to published methods, a detailed description, its criteria for including >1 signal in a locus, and software availability.

Summary of notation

We have phenotype data $\boldsymbol{Y}$, which is an $N_{ind} \times N_{pheno}$ matrix, where elements $Y_{jk} \in \{0,1\}$ are the values of phenotype *k* for individual j. We have covariate data $\boldsymbol{X} = (\boldsymbol{X}_{snp}\, \boldsymbol{X}_{pc})$ where $\boldsymbol{X}_{\boldsymbol{snp}}$ is an $N_{ind} \times N_{snp}$ matrix of genotype dosages and $\boldsymbol{X}_{pc}$ is an $N_{ind} \times N_{pc}$ matrix of confounders (measured in our case by principal components). We will write $\boldsymbol{X}_{j,snp}$ and $\boldsymbol{X}_{j,pc}$ to refer to the vectors of genotype dosages and confounders for individual j. In our case, $N_{pheno} = 2$, for Crohn's disease (CD) and ulcerative colitis (UC). We will consider each fine-mapping locus independently, so here N_{snp} is the number of variants in the fine-mapping locus.

We have parameters, $\theta \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the set of all possible parameter vectors. We also have models, $m \in \boldsymbol{M}$, where each model corresponds to some set of constraints, $\theta \in \Theta_m$. In all the cases below these models correspond to constraining some parameters to be equal to zero and others to be non-zero. We can define subsets of the model, $\boldsymbol{M}_i \subseteq \ \boldsymbol{M}$ as the set of all models where variant *i* is causal for at least one phenotype under consideration, and $\boldsymbol{M}_{\boldsymbol{S}} = \ \bigcup_{i \in \boldsymbol{S}} \boldsymbol{M}_i$ as the subset of models in which at least one of the variants in the set $\boldsymbol{S}$ is causal. We will write $\boldsymbol{\Theta}_{\boldsymbol{M}_i} = \ \bigcup_{m \in \boldsymbol{M}_i} \boldsymbol{\Theta}_m$ and $\boldsymbol{\Theta}_{\boldsymbol{M}_{\boldsymbol{S}}} = \ \bigcup_{i \in \boldsymbol{S}} \boldsymbol{\Theta}_{\boldsymbol{M}_i}$.

The aim of our fine-mapping approach is to identify a minimum set of variants $\boldsymbol{S}$ that collectively has at least a 95% posterior probability of being causal, i.e.

$$\Pr(\boldsymbol{M}_{\boldsymbol{S}} | \boldsymbol{X}, \boldsymbol{Y}) = \sum_{m \in \boldsymbol{M}_{\boldsymbol{S}}} \Pr(m \ | \boldsymbol{X}, \boldsymbol{Y}) \geq 0.95. \qquad (1)$$

In cases where we can assume that there is only one causal variant, the above equation can be written as a function of the marginal likelihoods and priors for each causal variant

$$\frac{\sum_{m \in \boldsymbol{M_S}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m) \Pr(m)}{\sum_{m \in \boldsymbol{M}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m) \Pr(m)} \geq 0.95. \quad (2)$$

All the methods below assume an equal prior for each model (i.e. no variants are assumed to be more likely than any others to be causal *a-priori*), so we can write this as

$$\frac{\sum_{m \in \boldsymbol{M_S}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m)}{\sum_{m \in \boldsymbol{M}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m)} \geq 0.95. \quad (3)$$

We define the marginal likelihood in terms of the data likelihood and the prior using:

$$\Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m) = \int_{\theta \in \boldsymbol{\Theta_m}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid \theta) \Pr(\theta \mid m) d\theta. \quad (4)$$

Method 1: Flat prior with steepest descent approximation

Summary: The first fine-mapping method assumes that $\Pr(\boldsymbol{X}, \boldsymbol{Y} \mid \theta)$ takes the form of a multinomial likelihood, and places a flat (improper) prior on all parameters. This method approximates the marginal likelihood as a penalized maximum likelihood, and is thus comparable to the Bayesian information criterion. This method can handle arbitrary phenotypes, and constructs a full risk model across multiple phenotypes and variants using a greedy search.

Relationship to published methods: This method is based on the same approximations used to carry out gene-based association testing in ref 18, and can be viewed as an extension of the previously published method of Onengut-Gumuscu *et al*[9]. We have extended this approach in three ways. First, it is generalized to use multinomial rather than binomial logistic regression, meaning it can handle multi-category phenotypes (in this case, CD and UC). Second, we use a Bayesian approach to detect additional signals, as opposed to the frequentist stepwise logistic regression used by Onengut-Gumuscu *et al*[9]. Finally, our method carries out a repositioning step, to ensure that a suboptimal model has not been selected due to collinearity (a potential problem that can arise during stepwise regression, see chapter 15 of Draper and Smith[61]).

Description: The parameter set is given by $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, where $\boldsymbol{\alpha}$ is a vector of intercepts for each phenotype (length N_{pheno}), and $\boldsymbol{\beta}$ is an $(N_{pc} + N_{snp}) \times N_{pheno}$ matrix of effect sizes, with elements β_{ik} being the effect size (log odds ratio relative to controls) of predictor *i* on each phenotype *k*. A particular model *m* is defined by which variants are causal for which phenotypes, so *m* is an $N_{snp} \times N_{pheno}$ matrix where $m_{ik} = 1$ if variant *i* is causal for phenotype *k* and $m_{ik} = 0$ otherwise. In the terminology given above, this means that $\theta \in \Theta_m$ if $\beta_{ik} \neq 0 \; \forall \; i, k : m_{ik} = 1$ (i.e. all causal variants have non-zero effect size) and $\beta_{ik} = 0 \; \forall \; i, k : m_{ik} = 0$ (i.e. all non-causal variants have zero effect size). A model is in the set of causal models for variant *i* ($m \in \boldsymbol{M_i}$) if the variant is causal for at least one phenotype (i.e. $\exists k : \; m_{ik} = 1$). Equivalently, we can state that $\theta \in \Theta_{\boldsymbol{M_i}}$ if $\boldsymbol{\beta_i} \neq \boldsymbol{0}$. Note that principal components are always included as fixed covariates (i.e. $\boldsymbol{\beta_i} \neq \boldsymbol{0} \; \forall \boldsymbol{i} \in \boldsymbol{pc}$).

The multinomial likelihood: The probability of an individual *j* having phenotype *k*, given their vector of predictors $\boldsymbol{X}_{j\cdot}$ and the parameter values $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, is given by

$$\Pr\left(Y_{jk} = 1, Y_{jk'} = 0 \ \forall k' \neq k \,\middle|\, \boldsymbol{X}_{j\cdot}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{exp(\alpha_k + \boldsymbol{X}_{j\cdot}\boldsymbol{\beta}_{\cdot k})}{\boldsymbol{1} + \sum_{\boldsymbol{k'=1}}^{\boldsymbol{N_{pheno}}} exp(\alpha_{k'} + \ \boldsymbol{X}_{j\cdot}\boldsymbol{\beta}_{\cdot k'})} \quad (5)$$

and the probability of being a control (i.e. all phenotypes being zero) is

$$\Pr\left(Y_{j\cdot} = 0 \,\middle|\, \boldsymbol{X}_{j\cdot}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{1}{\boldsymbol{1} + \sum_{\boldsymbol{k'=1}}^{\boldsymbol{N_{pheno}}} exp(\alpha_{k'} + \ \boldsymbol{X}_{j\cdot}\boldsymbol{\beta}_{\cdot k'})} \quad (6)$$

The overall likelihood is then given by

$$\Pr(\boldsymbol{Y} | \boldsymbol{X}, \theta) = \prod_{\boldsymbol{j=1}}^{\boldsymbol{N_{ind}}} \Pr\left(Y_{j\cdot} \,\middle|\, \boldsymbol{X}_{j\cdot}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) \ (7)$$

Prior and approximation to marginal likelihood: Equation (4) (and thus also (7)) can be estimated using the steepest descent approach if the likelihood function only has one global maximum and decays rapidly to zero away from the maximum[18]. All likelihood functions in the exponential family (including the linear, binomial, or multinomial models) meet these criteria. Assuming a flat (improper) prior on the parameters θ, the integral in (4) can be calculated as

$$\ln \Pr(\boldsymbol{X}, \boldsymbol{Y} \, | m) \approx \ln \Pr(\boldsymbol{Y} | \, \boldsymbol{X}, \theta^*) - \frac{|m| \ln(N_{ind})}{2} \quad (8)$$

in which θ^* is maximum likelihood estimate of the parameter values subject to $\theta^* \in \Theta_m$ and $|m|$ is the number of non-zero elements of *m* (i.e. the number of model parameters that are non-zero). This formulation is equivalent to using a Bayesian information criterion in model selection.

Model selection: We begin by selecting the optimal model m^*, which maximizes the marginal likelihood in (8). We do this using a greedy search, followed by a repositioning search. The greedy search starts by setting the initial model as $m^0 = \mathbf{0}$ (i.e. the null model, where only intercepts and principal components are non-zero). For each iteration (*t*), we calculate the marginal likelihood in (8) for every model m that adds one new causal variant (i) for one phenotype (k) to the model at the previous iteration; i.e. each model m where $m_{ik} = 1$ (the new causal variant/phenotype pair) and $m_{i'k'} = m_{i'k'}^{t-1}\ \forall(i', k') \neq (i, k)$ (the remainder of the model carried over from the last iteration). Note that we can add an extra variant that has not been associated with any phenotype (i.e. an entirely new causal variant), or add an additional phenotype to a variant that is already associated with one or more phenotypes. If none of these models improve the marginal likelihood then the algorithm terminates and returns the model m^{t-1}, otherwise it sets m^t to the model with the largest marginal likelihood and starts another iteration.

If there is only one causal variant in the model when the greedy search terminates (i.e. we only have one signal), then we set $m^* = m^t$, where t is the final iteration of the greedy search. If there is more than one causal variant in the final model (i.e. we have more than one signal), we undergo a repositioning stage to ensure that we have the optimal set of causal variants to explain the signals. At each stage, we select each causal variant in turn and see whether the marginal likelihood can be improved by swapping it with another variant. If so, we swap it with the variant that gives the largest improvement to the marginal likelihood. Once no variants can be swapped without decreasing the marginal likelihood, we finish the repositioning and set m^* to the final model selected.

Fine-mapping: If there is only one signal at a locus, fine-mapping can be carried out easily by plugging equation (8) into equation (3). When there are multiple signals, we carry out fine-mapping for each signal i conditional on the set of lead variants for the remaining signals (which we denote $cond$). This means that we restrict our set of possible models to only those models where the causal variants for the alternative signals are fixed, i.e. to the set

$$M_{cond} = \bigcap_{j \in cond} M_j \qquad (9)$$

and credible sets can then be calculated using the equation

$$\frac{\sum_{m \in M_S \cap M_{cond}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m)}{\sum_{m \in M_{cond}} \Pr(\boldsymbol{X}, \boldsymbol{Y} \mid m)} \geq 0.95. \quad (10)$$

This has the advantage of being fast and easy to calculate, and makes each credible set distinct and easily interpretable. However, it has the disadvantage of not accounting for uncertainty in the localization of alternative signals.

Criteria for including additional signals: Under the Bayesian information criterion, additional signals are included if they improve the log maximum likelihood by at least $\ln(N_{ind})/2$. In this case, $N_{ind} = 67{,}852$, which corresponds to a likelihood ratio chi-squared test statistic of 11.125 (i.e. $P = 8.52 \times 10^{-4}$).

Software availability: This method is implemented in R scripts, and the source code is available from https://github.com/hailianghuang/Fine-mapping.

Method 2: Empirical covariance prior with Laplace approximation

Summary: This fine-mapping method also uses a multinomial likelihood for $\Pr(\boldsymbol{X}, \boldsymbol{Y} \mid \theta)$, but places an empirical correlated normal prior on the effect sizes, with the hyperparameters of this prior learned across all known disease loci. This leverages the strong correlation in effect size between loci to improve power without needing to make assumptions about which loci are or are not shared (as all loci are assumed to be shared, just with potentially different effect sizes). This method directly calculates maximum *a-priori* model estimates (as opposed to approximating them with maximum likelihood estimates, as in the method above), and uses a Laplace approximation to convert these to marginal likelihoods.

Relationship to published methods: This method uses the empirical prior Bayesian association framework described in ref 49, and the fine-mapping approach is briefly described and assessed in the appendix to that paper. The extension from Bayesian multinomial logistic regression to multinomial fine-mapping is essentially identical to the extension from Bayesian binomial logistic regression to binomial fine-mapping detailed in Maller *et al*[6].

Description:

The parameter set is given by $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, as in the previous method. Unlike that method, each variant is either causal for all phenotypes, or for no phenotypes (so it is assumed that any causal variant has an effect on all phenotypes, though this effect is allowed to be very small for some phenotypes). Thus, a particular model *m* merely defines which variants are causal, so *m* is an N_{snp} vector where $m_i = 1$ if variant *i* is causal for any phenotype. In the terminology given above, this means that $\theta \in \Theta_m$ if $\boldsymbol{\beta_i} \neq \mathbf{0} \ \forall \ i : m_i = 1$ and $\boldsymbol{\beta_i} = \mathbf{0} \ \forall \ i : m_i = 0$. A model is in the set of causal models for variant *i* ($m \in \boldsymbol{M_i}$) if the variant is causal (i.e. $m_i = 1$). This is equivalent to the definition of causal model sets above, as $\theta \in \Theta_{\boldsymbol{M_i}}$ if $\boldsymbol{\beta_i} \neq \mathbf{0}$. As before, confounders are included as fixed covariates.

The multinomial likelihood: The multinomial likelihood is specified as in equation (7) above. This method uses the list of signals generated by the previous method. As in the previous method, when there is more than one independent signal at a locus we consider each signal separately by including all other signals in the region as fixed covariates. This means that, given a causal model m, all effect sizes are assumed to be zero other than those for the causal variant, the alternative signals that we are conditioning on, and the confounders, i.e. $\boldsymbol{\beta}_a = \mathbf{0}\ \forall a \notin \{i, pc, cond\}$. We can thus write the likelihood given a particular causal variant using a reduced subset of the data

$$\Pr(\boldsymbol{Y} \mid \boldsymbol{X}, \theta, \theta \in \boldsymbol{M}_i) = \Pr(\boldsymbol{Y} \mid \boldsymbol{X}_i, \boldsymbol{X}_{pc}, \boldsymbol{X}_{cond}, \boldsymbol{\alpha}, \boldsymbol{\beta}_i, \boldsymbol{\beta}_{pc}, \boldsymbol{\beta}_{cond}), \quad (11)$$

which reduces the complexity of the integrals below.

Prior distribution on parameters: To calculate the marginal likelihood, we next need to define the prior distributions for all the parameters. We assume that intercepts and confounding effect sizes are all independently normally distributed with a mean of zero and unit variance (in practice the results are invariant to these values, providing the variance is not vanishingly small). For associated variants, we assume they are drawn from a correlated multivariate normal prior

$$\boldsymbol{\beta}_i \sim MVN(0, \boldsymbol{\Sigma}), \quad (12)$$

where $\boldsymbol{\Sigma}$ is a hyperparameter representing the covariance matrix for the effect sizes across different phenotypes at the same variant, such that $\Sigma_{ab} = cov[\beta_{.a,}\, \beta_{.b}]$.

Laplace approximation to the marginal likelihood: The marginal likelihood for each SNP is then calculated using the Laplace approximation:

$$\Pr(\boldsymbol{Y},\boldsymbol{X}\,|m) \propto \Pr(\boldsymbol{Y}\mid \boldsymbol{m},\boldsymbol{X})$$

$$= \int_{\theta\in\boldsymbol{\Theta}_m} \Pr(\boldsymbol{Y}\mid\boldsymbol{X},\theta)\,\Pr(\theta|\boldsymbol{\Sigma})d\theta$$

$$\approx (2\,\pi)^{(2+N_{pca}+N_{cond})}|H(\theta^*)|^{\frac{1}{2}}\Pr(\boldsymbol{Y}\mid\boldsymbol{X},\theta^*)\,\Pr(\theta^*|\boldsymbol{\Sigma}) \qquad (13)$$

where θ^* is the maximum posterior estimate of θ, calculated using Newton's method, and $H(\theta)$ is the Hessian of the log joint likelihood, i.e.

$$H_{ab}(\theta) = \frac{\partial^2 \ln[\Pr(\boldsymbol{Y}\,|\boldsymbol{X},\theta)]}{\partial\theta_a\theta_b} + \frac{\partial^2 \ln[\Pr(\theta|\boldsymbol{\Sigma})]}{\partial\theta_a\theta_b} \qquad (14)$$

Credible sets can then be generated using equation (4).

Criteria for including additional signals: This method takes the number of independent signals from method 1. However, the actual signals generated can differ between methods 1 and 2: while method 2 does not propose new signals, it can move a signal to a new lead SNP if that lead SNP gives a higher value of equation (13).

Software availability: This method is implemented in the c++ program Trinculo[49], and source code and binaries are available from http://sourceforge.net/projects/trinculo/.

Method 3: Bayesian LASSO

Summary: This method uses a liability threshold model to describe the relationship between genotype and phenotype. Unlike the other two methods it does not attempt to calculate marginal likelihoods directly. Instead, it uses an MCMC algorithm to sample values of $\theta \in M$, and directly measures the proportion of samples that fall into different causal models. It considers multiple causal variants at once, and uses a Bayesian LASSO to enforce sparsity on the number of signals. However, unlike the previous two methods, this method does not consider different phenotypes simultaneously, and thus can only fine-map one phenotype at a time.

Relationship to published methods: The Bayesian LASSO element of the method is based on previous work on LASSO priors to allow effective shrinkage of sparse genetic effect sizes[62]. The use of MCMC to handle multiple signals is similar in approach to other published MCMC-based fine-mapping methodologies, such as the piMASS method[63]. The synthesis of these ideas into a fine-mapping method is described in a pre-print[64].

Description: The parameter set is given by $\theta = (\alpha, \boldsymbol{\beta}, \boldsymbol{L})$, where α is an intercept for the liability and $\boldsymbol{\beta}$ is a column vector of effect sizes (increase in liability per unit increase in predictor) for both the confounders ($\boldsymbol{\beta_{pc}}$) and the genotype dosages ($\boldsymbol{\beta_{snp}}$), and $\boldsymbol{L}$ is a vector of liabilities for each individual. There is also a series of hyperparameters, $\boldsymbol{\lambda}$ and $\boldsymbol{\tau^2}$, which define the priors on effect sizes. Note that, in order to account for uncertainty in the hyperparameters, the MCMC samples from both the parameters and the hyperparameters (with hyperpriors set on the latter). This method only analyses one phenotype for each locus, i.e. *k* is set to a particular value for each locus. After defining the clusters (discussed in *Defining clusters of SNPs*), we merge overlapping cluster pairs (sharing one or more variants) if one of them is associated with CD and the other is associated with UC. These merged clusters were assigned the phenotype of IBD.

We define a causal model *m* as in method 2, whereby $m_i = 1$ if variant *i* is causal for the phenotype *k* under consideration and $m_i = 0$ otherwise. A model is in the set of causal models for variant *i* ($m \in \boldsymbol{M_i}$) if the variant is causal (i.e. $m_i = 1$). This is comparable to the first two methods, as $\theta \in \Theta_{\boldsymbol{M_i}}$ if $\beta_i \neq 0$. Note that, unlike the previous two methods, the causal model *m* is not an internal state of the method, as the method does not constrain effect sizes to be zero or non-zero, instead it uses a Bayesian LASSO to shrink effect sizes to zero.

Liability threshold model: The proposed model is based on the standard assumption of an underlying (unobserved) normally distributed liability L_j for each individual *j*, with threshold *T*, such that individuals for which $L_j > T$ are affected and individuals for which $L_j \leq \mathrm{T}$ are healthy. This method hypothesizes that the liability is influenced by a set of (known) confounders and by a set of (unknown) potentially causal variants $\boldsymbol{C}$, such that the liability of individual *j* as

$$L_j = \alpha + \boldsymbol{X_{j,C}\beta_C} + \boldsymbol{X_{j,pc}\beta_{pc}} + \varepsilon_j \qquad (15)$$

where α is the population mean, and ε_i is an error term, assumed to have a mean of 0 and a variance σ_E^2. Following Sorensen and Gianola (2002)[65], the values of *T* and σ_E^2 are fixed at 0 and 1, respectively. We describe the set $\boldsymbol{C}$ as "potentially causal", as it is set to always contain exactly 20 causal variants (i.e. enough to contain all detectable independent signals, as in practice the number observed is always many fewer than 20), but these values are allowed to shrink to zero in the LASSO stage, meaning that many of these signals will have $\beta_i \cong 0$.

Priors and hyperpriors on the parameters: We define uniform (improper) priors on the population mean, α, and the effect sizes for confounders $\boldsymbol{\beta}_{pc}$ (i.e. $\Pr(\alpha) \propto 1; \Pr(\boldsymbol{\beta}_{pc}) \propto 1$). Following Fang *et al.*[62], the prior distribution of the effect of variant $i \in \mathbf{C}$, β_i, is assumed to follow a zero-centered double-exponential distribution:

$$\Pr(\beta_i) = \frac{\lambda_i}{2} e^{-\lambda_i |\beta_i|} \quad (16)$$

The double-exponential prior is used in Bayesian LASSO, and ensures that small effects shrink to zero. In order to allow efficient sampling from the double exponential, we introduce an exponentially distributed random nuisance variable τ_i^2, which is equal to the prior variance on β_i, such that $\Pr(\beta_i|\tau_i^2) = \phi(\beta_i|0, \tau_i^2)$ and $\Pr(\tau_i^2|\lambda_i) = \frac{\lambda_i}{2}e^{-\lambda_i\tau_i^2/2}$. Finally, we place a gamma distributed hyperprior on λ_i^2, such that $\Pr\left(\frac{\lambda_i^2}{2}\right) = gamma(\delta, \delta)$, where δ is a very small number (we used 10^{-5} in this study).

Overview of MCMC algorithm: The aim of the MCMC algorithm is to sample from the joint distribution of the parameters of interest (the vector $\boldsymbol{\beta_{snp}}$), as well as all of the unobserved nuisance parameters, given by:

$$\Pr(\boldsymbol{\beta_{snp}}, \boldsymbol{\beta_{pc}}, \boldsymbol{C}, \boldsymbol{\lambda}, \boldsymbol{\tau^2}, \boldsymbol{L}, \alpha \; |\boldsymbol{X}, \boldsymbol{Y}) \qquad (17)$$

The MCMC algorithm has two stages: a Metropolis-Hastings stage, to sample the potential causal variants $\boldsymbol{C}$, and a Gibbs sampler stage, to select values of the other parameters. The MH stage is carried out once, followed by a single Gibbs sampler update of $\alpha, \boldsymbol{\beta}_{pc}$ and $\boldsymbol{L}$, followed by 50 rounds of Gibbs sampler updates on $\boldsymbol{\beta_{snp}}, \boldsymbol{\lambda}$ and $\boldsymbol{\tau^2}$. The complete process is repeated 1 million times. We initiate the chain by assigning values randomly to all variables (within their legal boundaries). The first 500,000 iterations are used as burn-in and ignored when compiling the summary statistics, and manual inspection of parameter traces was used to check that the chain was well mixed and independent of the starting values. We will denote the parameters at iteration *t* as $\boldsymbol{\beta_{snp}^t}$, and the total number of iterations considered as N_t.

Defining clusters of SNPs: The MH stage of the MCMC serves two purposes: firstly, to search for independent signals in each locus, and secondly, to identify variants in high

LD for each signal from which to fine-map the causal variant. In order to ensure that the MH stage searches the entire space of variants, we first cluster all variants in the region into a series of high-LD clusters H, such that each variant belongs to exactly one cluster $h \in H$. We hierarchically cluster variants that are in high LD using (1-r^2) as distance measure and the "single linkage" approach implemented with the R "hclust" package. We define clusters such that variants in distinct clusters will never have $r^2 > Q$ (though note that variants within clusters may have $r^2 < Q$). We tested Q values of 0.9 and 0.5; larger values can artificially split credible sets across clusters, but smaller values can artificially lump different signals into the same cluster. We tested for the former by fitting a multivariate regression model with all detected clusters (with posterior > 0.5) and checking for pairwise collinearity between clusters. We used Q = 0.9 if the model fit did not show evidence of collinearity, otherwise we used Q = 0.5.

Variant sampling using the Metropolis-Hastings algorithm: During the MH stage, we enforce that each variant in $\boldsymbol{C}$ comes from a different cluster. At each round of the MCMC chain, we sequentially attempt to swap each of the variants in $\boldsymbol{C}$, by proposing a new variant to replace it. Half of the time the new proposed variant is selected uniformly from the same cluster as the original variant, and half of the time a new cluster is selected uniformly from all clusters that do not contain a causal variant and a new variant is uniformly selected from within this cluster. In other words, this means that the MCMC chain spends half of its time searching for the best possible variants within clusters, and half of its time for the best possible clusters. When a substitute variant is selected, the probability to "accept" it is

$$min\left(1, \frac{\Pr(\boldsymbol{L}|\boldsymbol{X}_{.i^{t+1}}, \dots)}{\Pr(\boldsymbol{L}|\boldsymbol{X}_{.i^{t}}, \dots)}\right) \qquad (18)$$

where

$$\Pr(\boldsymbol{L}|\ \ldots) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{1}{2}\sum_{j=1}^{N_{ind}}\left(L_j - \alpha - \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C - \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}\right)^2\right) \tag{19}$$

and i^{t+1} and i^{t} correspond to the “new” and “old” variants included in $\boldsymbol{C}$, respectively.

Parameter updates using the Gibbs sampler: We set β_i= 0 $\forall i \in snp \backslash \boldsymbol{C}$, and elements of the effect size vector $\boldsymbol{\beta_C}$ are sampled from normal distributions with mean

$$\overline{\beta_\iota} = \left(\sum\nolimits_{j=1}^{N_{ind}} X_{ji}^2 + \frac{1}{\tau_i^2}\right)^{-1} \sum\nolimits_{j=1}^{N_{ind}} X_{ji}\left(L_j - \alpha - \boldsymbol{X}_{j,C\,\setminus\,i}\boldsymbol{\beta}_{C\,\setminus\,i} - \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}\right), \forall i \in \boldsymbol{C} \tag{20}$$

and variance

$$\sigma_{\beta_i}^2 = \left(\sum\nolimits_{j=1}^{N_{ind}} X_{ji}^2 + \frac{1}{\tau_i^2}\right)^{-1}, \forall i \in \boldsymbol{C}\,, \tag{21}$$

where subscript $\boldsymbol{C} \setminus i$ indicates all potential causal variants, excluding the variant under consideration *i*.

Elements of $\boldsymbol{\tau^2}$are sampled from inverse Gaussian distributions

$$(\tau_i^2 | \beta_i, \lambda_i) \sim InvGauss\left(\sqrt{\frac{\lambda_i^2}{\beta_i^2}}, \lambda_i^2\right), i \in \boldsymbol{C}. \tag{22}$$

The hyper-parameters $\boldsymbol{\lambda}$ are sampled from gamma distributions

$$(\lambda_i^2 | \tau_i^2) \sim Gamma\left(1, \frac{\tau_i^2}{2}\right), i \in \boldsymbol{C}. \tag{23}$$

Elements of the confounding effect sizes $\boldsymbol{\beta_{pc}}$ are sampled from normal distributions with mean

$$\overline{\beta_\iota} = \left(\sum\nolimits_{j=1}^{N_{ind}} X_{ji}\right)^{-1} \sum\nolimits_{j=1}^{N_{ind}} \left(L_j - \alpha - \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C - \boldsymbol{X}_{j,pc\,\setminus\,i}\boldsymbol{\beta}_{pc\,\setminus\,i}\right), \forall i \in pc \tag{24}$$

and variance

$$\sigma_{\beta_i}^2 = \left(\sum\nolimits_{j=1}^{N_{ind}} X_{ji}\right)^{-1}, \forall i \in pc, \qquad (25)$$

where subscript $pc \setminus i$ indicates all confounders, excluding the confounder under consideration *i*.

The population mean, α, is sampled from a normal distribution with mean

$$\bar{\alpha} = \frac{1}{N_{ind}} \sum\nolimits_{j=1}^{N_{ind}} \left(L_j - \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C - \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}\right) \qquad (26)$$

and variance $1/N_{ind}$.

For affected individuals ($Y_{jk} = 1$), the liabilities, L_j, are sampled from the truncated normal distributions (such that $L_j > T$) with density

$$\Pr\left(L_j \middle| Y_{jk} = 1, \dots\right) = \frac{\phi\left(L_j \middle| \alpha + \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C + \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}, \boldsymbol{\sigma}_{\boldsymbol{E}}^2\right)}{1 - \Phi_T\left(\alpha + \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C + \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}, \boldsymbol{\sigma}_{\boldsymbol{E}}^2\right)} \qquad (27)$$

For unaffected individuals ($Y_{jk} = 0$), the liability, L_j, are sampled from the truncated normal distributions (such that $L_j \leq T$) with density

$$\Pr\left(L_j \middle| Y_{jk} = 0, \dots\right) = \frac{\phi\left(L_j \middle| \alpha + \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C + \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}, \boldsymbol{\sigma}_{\boldsymbol{E}}^2\right)}{\Phi_T\left(\alpha + \boldsymbol{X}_{j,C}\boldsymbol{\beta}_C + \boldsymbol{X}_{j,pc}\boldsymbol{\beta}_{pc}, \boldsymbol{\sigma}_{\boldsymbol{E}}^2\right)} \qquad (28)$$

In these, $\Phi_T(\cdot)$ corresponds to the cumulative density from $-\infty$ to *T*.

Summarizing the results. The posterior probability that variant *i* is causal is given by the proportion of MCMC iterations where variant i is in the set of potentially causal variants for that iteration ($\boldsymbol{C^t}$), or

$$\Pr(\boldsymbol{M}_i | \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{N_t} \sum_t \mathrm{I}(i \in \boldsymbol{C}^t). \quad (29)$$

The posterior probability for each cluster *h* is then given by $\sum_{i \in h} \Pr(\boldsymbol{M}_i | \ \boldsymbol{X}, \boldsymbol{Y})$. We report a positive signal for any cluster that has a posterior probability greater than 50%, and calculate credible sets using equations (1) and (29). Finally, to eliminate spuriously associated clusters, we carried out a joint frequentist logistic regression to calculate *P* values for each variant, and only retained signals that achieved $p < 10^{-4}$ in this analysis.

Criteria for including additional signals: This method declares a new signal in any cluster where both the posterior of that cluster is greater than 50% (i.e. in at least 50% of MCMC samples, a causal variant from that cluster is included in $\boldsymbol{C}$) and has a logistic regression *P* value of less than 10^{-4}.

Software availability. The algorithm is implemented in the BayesFM program which can be downloaded from https://sourceforge.net/projects/bayesfm-mcmc-v1-0/

Supplementary Notes

Three regions that were not resolved in fine-mapping:

- Chr1:22.6-22.8Mb: This region contains two UC signals of similar significance. The first UC signal has *P* value of 2.7×10^{-13} with AF of 5%, and the second signal has 4.8×10^{-13} with AF=16%. R^2 between the two signals is 0.3. Jointly fitting the two signals has *P* value of 1×10^{-13}. Because the models are almost equally significant, our model selection procedure is unable to conclude which model is clearly the best.
- Chr4:112.9-123.6Mb: The best model in this region has one IBD and one CD signal, both identified only by “Method 3: Bayesian LASSO”. The next best model has an additional IBD signal identified using “Method 1: Flat prior with steepest descent approximation”. Because neither of these models includes signals identified by more than one method, we conservatively flag this region as “unresolved”.
- Chr16:85.9-86.0Mb: Fine-mapping has identified two models of similar significance for this region. Model 1 (P=6.0×10^{-26}) has a CD signal and an IBD signal, whereas model 2 (P=5.6×10^{-26}) has the same IBD signal, but a different CD signal. Because the models are almost equally significant, our model selection procedure is unable to conclude which model is clearly the best.

Duplicated insertion variant:

We discovered an indel alignment error in credible variants in signal 1 in region 104 (the *TNSF15*/*TNSF8* region). This credible set consists of four variants, three of which are insertions in nearly perfect LD ($r^2 > 0.999$). The local reference sequence is '`TAAAT`', and the alternate alleles for the three variants are '`T(AGA)AAAT`', '`TA(GAA)AAT`' and '`TA(GA)AAT`'. The first two of these are the same variant with the position shifted by 1bp, and the third appears to be a historical annotation error. If we combine the three posteriors into the current canonical variant in dbSNP (chr9:117571293 / rs35396782), we will have a two-variant credible set with chr9:117571293 having 92% posterior probability and rs4372078 having 3% posterior probability.

Supplementary Box

Strawberry notch homologue 2 (*SBNO2*, OMIM*615729), interleukin-10 (*IL10*, OMIM*124092) and interleukin-19 (*IL19*, OMIM*605687)

Fine-mapping of locus #165 (Chr. 19; containing *ABCA7*, *HMHA1*, *GPX4*, *POLR2E*, *SBNO2* and *STK11*) identified a primary signal with a credible set of three variants that are located in three separate introns of the *SBNO2* gene and that are all within 653 bp of each other. While these common variants (MAF ~0.22) do not overlap known regulatory sequences, their tight clustering within a single gene supports *SBNO2* as the best candidate causal gene. Furthermore, there is a secondary association signal in this region with a credible set of 10 variants, led by rs72977562, which is located in a gut enhancer element (H3K27ac) 6 kb before the TSS of *SBNO2*. While there is limited information on the function of this gene, it has been reported that it contributes to the downstream anti-inflammatory effects of IL-10[66] and may be involved in sepsis[67], bone homeostasis[68] and inflammatory response in the central nervous system[69]. *IL10* is one of three candidate genes in locus #23 (Chr. 1; *MAPKAPK2*, *IL10*, *IL19*), where the primary signal is defined by a credible set of two variants separated by 4kb, one in the third intron of *IL10* and the other ~1 kb downstream, and both overlap the H3K4me1 peaks. The secondary association signal in this locus is defined by a credible set of four variants within a 4kb interval between *IL10* and *IL19*. *IL19* is a member of the *IL10* cytokine family whose function in the inflammatory response in humans is poorly understood. In fact, *IL19* has been reported both to have an anti-inflammatory role[70] or a pro-inflammatory role[71] in IBD. Without functional data to establish causality it is not possible to definitively state whether the causal gene in this locus is *IL10* or *IL19*, however, this fine mapping clearly indicates the need to better understand the *SBNO2*/*IL10*/*IL19* axis in IBD, in particular in the monocyte/macrophage lineage where the latter two genes are primarily expressed[70].

Leucine-rich repeat kinase 2 (*LRRK2*, OMIM*609007)

Association to the chromosome 12 region containing the *LRRK2* and *MUC19* genes was identified in the first CD meta-analysis[72]. The primary association signal has a credible set of 20 variants extending over a ~65kb interval: 10 variants within the *LRRK2* gene (including one located within 3'UTR), 8 within the *MUC19* gene, and 2 intergenic. The secondary association signal, however, is fine-mapped to a single variant (rs7307562), which is located in the 39th intron of *LRRK2*, suggesting this as the more likely gene. *LRRK2* was first described as a gene for autosomal dominant Parkinson disease (*PARK8*; OMIM*607060). Functional studies demonstrated that *LRRK2* might play a role in inflammatory signaling pathways relevant to Crohn's disease[73]. Specifically, the promoter of *LRRK2* contains a conserved binding site for interferon (IFN) response factors, and it was shown that stimulation with IFN resulted in the up-regulation of the expression of *LRRK2*, which activated NFkB in an IKK-dependent manner. More recently it was shown that *LRRK2* deficiency conferred enhanced susceptibility to experimental colitis in mice[74]. In addition, it was reported that a coding variant (Met2397Thr) in this gene led to lower *LRRK2* protein levels and was therefore likely pathogenic[74], but that variant is more than ten orders of magnitude less significant than the rs7307562 intronic variant in our dataset, and can be explained by the residual signal derived from the modest correlation between these variants (r^2=0.38).

Kinase suppressor of Ras-1 (*KSR1*, OMIM*601132) and PR Domain containing protein 1 (*PRDM1*, OMIM*603423)

Fine-mapping of the locus #157 (containing 4 genes) identified a single association to CD. The credible set for this region consists of seven common variants (MAF ~0.30) across 25kb, all within the *KSR1* gene and more than 50 kb from any of the other genes in the region. With multiple predicted isoforms, these variants are clustered within the first introns of some isoforms and upstream of the transcription start site of others. Two of the variants within the credible set are predicted to impact transcription factor binding sites detected by ENCODE ChIP-seq experiments: rs8075695 in a PRDM1 site and rs2948542 in a TAL1 site. *KSR1* is a reasonable candidate gene for this locus as previous functional studies suggested that it had a role to play in protecting against intestinal inflammation. Specifically, Polk and colleagues observed that *KSR1* is activated in inflamed mucosa and using *KSR1*-deficient mice demonstrated that *KSR1* protects intestinal epithelium from cytokine-mediated apoptosis during inflammation[75]. Interestingly, there may be a link between *KSR1* and another IBD gene as *PRDM1* itself is implicated as the most likely causal gene in locus #77, which contains two well-defined associations (including a single-variant credible set).

Mothers against decapentaplegic, drosophila (*SMAD3*, OMIM *603109)

A CD locus on Chr. 15 harbors three genes *SMAD3*, *AAGAB*, and *IQCH.* Previous analyses prioritized *SMAD3* within this locus based on the GRAIL text-mining algorithm, and now our fine-mapping focuses the primary association on rs17293632, which is in an intron of *SMAD3* and has the best posterior probability (40%) within the 5-variant credible set. This variant disrupts an AP-1 binding motif that is highly conserved across vertebrates[23]. ChIP-seq experiments in HeLa cells, which are fortuitously heterozygous at this site, showed major allelic imbalance with almost non-existent binding to the CD risk allele. Furthermore, a second independent association at this locus is resolved to a single protein-coding variant I170V in *SMAD3*, identifying with near certainty that this is the relevant gene. *SMAD3* is a key signaling molecule in the Transforming Growth beta (TGF-β) pathway, which is implicated in various immune abnormalities and is associated with other diseases including tumor development and fibrotic complications. In addition, mucosal inflammation in CD is characterized by reduced activity of TGF-β1 due to high levels of *SMAD7* mediated through *SMAD3*. An oral *SMAD7* antisense oligonucleotide (Mongersen) has recently shown great promise in a phase 2 clinical trial for Crohn's disease[76,77]. *SMAD7* (itself recently associated to IBD but not in a high density genotyping region on the Immunochip and therefore not fine-mapped here) is an inhibitor of *SMAD3*. Mongersen, by targeting *SMAD7*, restores *SMAD3* signaling and TGF-β activation, leading to suppression of inflammation. Consistent with this, genetic disruption of *SMAD3*, as identified here, constitutes an increased risk for CD.

Interferon-induced helicase c domain-containing protein 1 (*IFIH1,* OMIM*606951)

The UC genetic signal from the locus on Chr. 2 that harbors the candidate genes *DPP4, GCG, FAP, IFIH1, GCA* and *KCNH7* has here been reduced to a single rare missense variant (I923V) in *IFIH1*. The *IFIH1* gene encodes Melanoma Differentiation-Associated Protein 5 (MDA5) which is part of the RIG-I-Like receptor family and is a cytoplasmic viral RNA receptor involved in activating type I interferon signaling. We queried all our single variant credible sets for evidence of published associations to other diseases. By doing this we found that the single rare missense variant (I923V) in *IFIH1* associated to UC is also associated to type 1 diabetes (T1D) although with an opposite direction of effect[37]. This supports the notion that studies of shared genetic variation across immune-mediated diseases should, in many instances, ultimately point to the same variant. The identified variant has shown to be involved in the alteration of RNA secondary structure and reduces poly(I:C)-induced interferon-β production in subjects with T1D[78]. The isoleucine to valine change does not directly disrupt the RNA helicase domain of *IFIH1* that binds to dsRNA but electronic magnifier and ATP hydrolysis assays showed functional consequences due to impaired filament assembly and reduced kinetic stability[79]. This variant (I923V) was not genotyped in the previous IBD association study and a common variant (rs2111485), in LD ($r^2 > 0.9$) with a *IFIH1* coding variant (rs1990760, A946T), was reported[3]. This common variant is independent of I923V and is marginally significant (P= 7.8x10^{-6} versus the threshold at 1.35x10^{-6}). Another independent signal (rs72871627) with marginal significance (P= 3.5x10^{-6}) is a T1D fine-mapped association[9] and is additionally associated with psoriasis, ankylosing spondylitis and primary sclerosing cholangitis[80]. All three independent signals (I923V, rs2111485/A946T and rs72871627) are protective for psoriasis and T1D, and risk for IBD, ankylosing spondylitis and primary sclerosing cholangitis[9,80].

Interleukin receptor 2 alpha (*IL2RA*, *CD25* OMIM* 147730).

The *IL2RA* signal, first identified in 2012[3], has been fine-mapped to a single intronic variant, which has also been associated with T1D although with opposite direction of effect[38]. *IL2RA* encodes CD25, a constituent part of the *IL2* receptor that is expressed in both soluble and membrane-bound forms. The same single variant has been reported as the most associated variant in a 2013 Sardinian study of immune cell levels[81], demonstrating association to variation in levels of T-cells expressing *CD25*, with the allele increasing risk to Crohn's disease corresponding to higher levels of CD25 expressing T-cells. Animal models have suggested that CD25 prevents autoimmunity via a cell contact, cytokine independent mechanism and that the immunosuppressive effects of CD25/CD4 positive T cells can be overcome by microbial activation of Toll-like receptors and subsequent maturation of dendritic cells[82]. However *IL6* deficient mice were unable to overcome this Treg mediated suppression leading to decreased autoimmunity and increased infection suggesting that anti-IL6 therapy may be an appropriate therapy for immune-mediated diseases[83]. In addition to the T1D association genetic variants in *IL2RA* have also been associated with Graves disease and multiple sclerosis.

Regulator of telomere elongation helicase 1 (*RTEL1*, OMIM* 608833) and tumor-necrosis factor receptor superfamily, member 6b (*TNFRSF6B*, OMIM* 603361))

The original association of this locus in IBD susceptibility identified a region possibly containing more than 20 genes. Fine-mapping (locus #175) identified a single signal, rs6062496, that disrupts the binding of Early B Cell Factor 1 (EBF1), overlaps DNaseI hypersensitivity clusters, and is highly conserved across species. *RTEL1* is a DNA helicase involved in protecting telomeres during replication. Read-through transcription of *RTEL1* into tumor-necrosis factor receptor superfamily, member 6b (*TNFRSF6B*) results in a non-coding transcript. *TNFRSF6B* (also known as Decoy Receptor 3 [*DCR3*]) neutralizes three different TNF ligands all of which have been associated with IBD susceptibility: FasL (*TNFSF6*), LIGHT (*TNFSF14*), and TL1A (*TNFSF15*). Manipulation of both *TNFSF14* and *TNFSF15* have been proposed as potential therapeutic strategies for IBD[84,85]. In addition *DCR3* plays a role in monocyte development, and also monocyte adhesion to endothelial cells via up-regulation of *ICAM-1*, *VCAM-1*, and *IL8* expression[86]. *DCR3* is overexpressed in gastrointestinal malignancies[87] and has also been associated with multiple sclerosis, rheumatoid arthritis, and glomerulonephritis.

References

61. Draper, N. R. & Smith, H. *Applied Regression Analysis*. (1998). doi:10.1002/9781118625590
62. Fang, M. *et al.* Improved LASSO priors for shrinkage quantitative trait loci mapping. *Theor. Appl. Genet.* **124,** 1315–1324 (2012).
63. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5,** 1780–1815 (2011).
64. Fang, M. & Georges, M. BayesFM: a software program to fine-map multiple causative variants in GWAS identified risk loci. *bioRxiv* (2016). doi:10.1101/067801
65. Sorensen, D. & Gianola, D. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. (Springer Science & Business Media, 2002).
66. Kasmi, El, K. C. *et al.* Cutting edge: A transcriptional repressor and corepressor induced by the STAT3-regulated anti-inflammatory signaling pathway. *J. Immunol.* **179,** 7215–7219 (2007).
67. Bhatty, M., Fan, R., Muir, W. M., Pruett, S. B. & Nanduri, B. Transcriptomic analysis of peritoneal cells in a mouse model of sepsis: confirmatory and novel results in early and late sepsis. *BMC Genomics* **13,** 509 (2012).
68. Maruyama, K. *et al.* Strawberry notch homologue 2 regulates osteoclast fusion by enhancing the expression of DC-STAMP. *Journal of Experimental Medicine* **210,** 1947–1960 (2013).
69. Grill, M. *et al.* Strawberry notch homolog 2 is a novel inflammatory response factor predominantly but not exclusively expressed by astrocytes in the central nervous system. *Glia* **63,** 1738–1752 (2015).
70. Cantó, E. *et al.* Interleukin-19 impairment in active Crohn's disease patients. *PLoS ONE* **9,** e93910 (2014).
71. Fonseca-Camarillo, G., Furuzawa-Carballeda, J., Granados, J. & Yamamoto-Furusho, J. K. Expression of interleukin (IL)-19 and IL-24 in inflammatory bowel disease patients: a cross-sectional study. *Clin. Exp. Immunol.* **177,** 64–75 (2014).
72. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40,** 955–962 (2008).
73. Gardet, A. *et al.* LRRK2 is involved in the IFN-gamma response and host response to pathogens. *J. Immunol.* **185,** 5577–5585 (2010).
74. Liu, Z. *et al.* The kinase LRRK2 is a regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel disease. *Nat. Immunol.* **12,** 1063–1070 (2011).
75. Yan, F. *et al.* Kinase suppressor of Ras-1 protects intestinal epithelium from cytokine-mediated apoptosis during inflammation. *J. Clin. Invest.* **114,** 1272–1280 (2004).
76. Fiocchi, C. TGF-beta/Smad signaling defects in inflammatory bowel disease: mechanisms and possible novel therapies for chronic inflammation. *J. Clin. Invest.* **108,** 523–526 (2001).
77. Monteleone, G. *et al.* Mongersen, an oral SMAD7 antisense oligonucleotide, and Crohn's disease. *N. Engl. J. Med.* **372,** 1104–1113 (2015).

78. Chistiakov, D. A., Voronova, N. V., Savost'Anov, K. V. & Turakulov, R. I. Loss-of-function mutations E6 27X and I923V of IFIH1 are associated with lower poly (I: C)–induced interferon-β production in peripheral blood mononuclear cells of type 1 diabetes patients. *Human immunology* **71,** 1128–1134 (2010).
79. Peisley, A. *et al.* Cooperative assembly and dynamic disassembly of MDA5 filaments for viral dsRNA recognition. *P Natl Acad Sci Usa* **108,** 21010–21015 (2011).
80. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* (2016). doi:10.1038/ng.3528
81. Orrù, V. *et al.* Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155,** 242–256 (2013).
82. Pasare, C. & Medzhitov, R. Toll pathway-dependent blockade of CD4+CD25+ T cell-mediated suppression by dendritic cells. **299,** 1033–1036 (2003).
83. Powrie, F. & Maloy, K. J. Regulating the regulators. (2003). doi:10.1126/science.1082031
84. Krause, P. *et al.* The tumor necrosis factor family member TNFSF14 (LIGHT) is required for resolution of intestinal inflammation in mice. *Gastroenterology* **146,** 1752–62.e4 (2014).
85. Shih, D. Q. *et al.* Inhibition of a novel fibrogenic factor Tl1a reverses established colonic fibrosis. *Mucosal Immunol* **7,** 1492–1503 (2014).
86. Yang, C.-R., Hsieh, S.-L., Ho, F.-M. & Lin, W.-W. Decoy receptor 3 increases monocyte adhesion to endothelial cells via NF-kappa B-dependent up-regulation of intercellular adhesion molecule-1, VCAM-1, and IL-8 expression. *J. Immunol.* **174,** 1647–1656 (2005).
87. Bai, C., Connolly, B. & Metzker, M. L. Overexpression of M68/DcR3 in human gastrointestinal tract tumors independent of gene amplification and its location in a four-gene cluster. in (2000).

Members of the International Inflammatory Bowel Disease Genetics Consortium

Clara Abraham[36], Jean-Paul Achkar[37,38], Tariq Ahmad[39], Leila Amininejad[40,41], Ashwin N Ananthakrishnan[30,42], Vibeke Andersen[9,10,43], Carl A Anderson[7], Jane M Andrews[44], Vito Annese[45,46], Guy Aumais[34,47], Leonard Baidoo[31], Robert N Baldassano[48], Peter A Bampton[49], Murray Barclay[50], Jeffrey C Barrett[7], Theodore M Bayless[51], Johannes Bethge[52], Alain Bitton[53], Gabrielle Boucher[8], Stephan Brand[54], Berenice Brandt[52], Steven R Brant[51], Carsten Büning[55], Angela Chew[20,56], Judy H Cho[35], Isabelle Cleynen[11], Ariella Cohain[57], Anthony Croft[58], Mark J Daly[1,2], Mauro D'Amato[12,13], Silvio Danese[59], Dirk De Jong[60], Martine De Vos[61], Goda Denapiene[62], Lee A Denson[63], Kathy L Devaney[30], Olivier Dewit[64], Renata D'Inca[65], Marla Dubinsky[66], Richard H Duerr[31,32], Cathryn Edwards[67], David Ellinghaus[15], Jonah Essers[68,69], Lynnette R Ferguson[70], Eleonora A Festen[28], Philip Fleshner[17], Tim Florin[71], Denis Franchimont[40,41], Andre Franke[15], Karin Fransen[72], Richard Gearry[50,73], Michel Georges[3,4], Christian Gieger[74], Jürgen Glas[54,75], Philippe Goyette[8], Todd Green[2,68], Anne M Griffiths[76], Stephen L Guthery[77], Hakon Hakonarson[48], Jonas Halfvarson[16], Katherine Hanigan[58], Talin Haritunians[17], Ailsa Hart[78], Chris Hawkey[79], Nicholas K Hayward[80], Matija Hedl[36], Paul Henderson[81,82], Xinli Hu[83], Hailiang Huang[1,2], Jean-Pierre Hugot[84], Ken Y Hui[35], Marcin Imielinski[48], Andrew Ippoliti[17], Laimas Jonaitis[85], Luke Jostins[5,6], Tom H Karlsen[86,87,88], Nicholas A Kennedy[21], Mohammed Azam Khan[89,90], Gediminas Kiudelis[85], Krupa Krishnaprasad[91], Subra Kugathasan[92], Limas Kupcinskas[93], Anna Latiano[45], Debby Laukens[61], Ian C Lawrance[19,20], James C Lee[25], Charlie W Lees[21], Marcis Leja[94], Johan Van Limbergen[76], Paolo Lionetti[95], Jimmy Z Liu[7], Edouard Louis[22], Gillian Mahy[96], John Mansfield[97], Dunecan Massey[25], Christopher G Mathew[26,33], Dermot PB McGovern[17], Raquel Milgrom[98], Mitja Mitrovic[72,99], Grant W Montgomery[80], Craig Mowat[100], William Newman[89,90], Aylwin Ng[30,101], Siew C Ng[102], Sok Meng Evelyn Ng[36], Susanna Nikolaus[52], Kaida Ning[36], Markus Nöthen[103], Ioannis Oikonomou[36], Orazio Palmieri[45], Miles Parkes[25], Anne Phillips[100], Cyriel Y Ponsioen[104], Urõs Potocnik[99,105], Natalie J Prescott[26], Deborah D Proctor[36], Graham Radford-Smith[58,106], Jean-Francois Rahier[107], Soumya Raychaudhuri[83], Miguel Regueiro[31], Florian Rieder[37], John D Rioux[8,34], Stephan Ripke[1,2], Rebecca Roberts[50], Richard K Russell[81], Jeremy D Sanderson[108], Miquel Sans[109], Jack Satsangi[21], Eric E Schadt[57], Stefan Schreiber[15,52], Dominik Schulte[52], L Philip Schumm[110], Regan Scott[31], Mark Seielstad[111,112], Yashoda Sharma[36], Mark S Silverberg[98], Lisa A Simms[58], Jurgita Skieceviciene [85], Sarah L Spain[26,27], A. Hillary Steinhart[98], Joanne M Stempak[98], Laura Stronati[113], Jurgita Sventoraityte[93], Stephan R Targan[17], Kirstin M Taylor[108], Anje ter Velde[104], Emilie Theatre[3,4], Leif Torkvist[114], Mark Tremelling[115], Andrea van der Meulen[116], Suzanne van Sommeren[28], Eric Vasiliauskas[17], Severine Vermeire[11,29], Hein W Verspaget[116], Thomas Walters[76,117], Kai Wang [48], Ming-Hsi Wang [37,51], Rinse K Weersma[28], Zhi Wei[118], David Whiteman[80], Cisca Wijmenga[72], David C Wilson[81,82], Juliane Winkelmann[119,120], Ramnik J Xavier[2,30], Bin Zhang[57], Clarence K Zhang[121], Hu Zhang[122,123], Wei Zhang[36], Hongyu Zhao[121], Zhen Z Zhao[80]

[36]Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA. [37]Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA. [38]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA. [39]Peninsula College of Medicine and Dentistry, Exeter, UK. [40]Department of Gastroenterology, Erasmus Hospital, Brussels, Belgium. [41]Department of

Gastroenterology, Free University of Brussels, Brussels, Belgium. [42]Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [43]Institute of Regional Health Research , University of Southern Denmark, Odense, Denmark. [44]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, Adelaide, Australia. [45]Unit of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievo della Sofferenza (IRCCS-CSS) Hospital, San Giovanni Rotondo, Italy. [46]Strutture Organizzative Dipartimentali (SOD) Gastroenterologia 2, Azienda Ospedaliero Universitaria (AOU) Careggi, Florence, Italy. [47]Department of Gastroenterology, Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada. [48]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [49]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide, Australia. [50]Department of Medicine, University of Otago, Christchurch, New Zealand. [51]Meyerhoff Inflammatory Bowel Disease Center, Department of medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [52]Department for General Internal Medicine, Christian-Albrechts-University, Kiel, Germany. [53]Division of Gastroenterology, Royal Victoria Hospital, Montréal, Québec, Canada. [54]Department of Medicine II, Ludwig-Maximilians-University Hospital Munich-Grosshadern, Munich, Germany. [55]Department of Gastroenterology, Campus Charité Mitte, Universitatsmedizin Berlin, Berlin, Germany. [56]IBD unit , Fremantle Hospital, Fremantle, Australia. [57]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA. [58]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia. [59]IBD Center, Department of Gastroenterology, Istituto Clinico Humanitas, Milan, Italy. [60]Department of Gastroenterology and Hepatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [61]Department of Hepatology and Gastroenterology, Ghent University Hospital, Ghent, Belgium. [62]Center of hepatology, Gastroenterology and Dietetics, Vilnius University, Vilnius, Lithuania. [63]Pediatric Gastroenterology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [64]Department of Gastroenterology, Université Catholique de Louvain (UCL) Cliniques Universitaires Saint-Luc, Brussels, Belgium. [65]Division of Gastroenterology, University Hospital Padua, Padua, Italy. [66]Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, California, USA. [67]Department of Gastroenterology, Torbay Hospital, Torbay, Devon, UK. [68]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [69]Pediatrics, Harvard Medical School, Boston, Massachusetts, USA. [70]Faculty of Medical & Health Sciences, School of Medical Sciences, The University of Auckland, Auckland, New Zealand. [71]Department of Gastroenterology, Mater Health Services, Brisbane, Australia. [72]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands. [73]Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand. [74]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany. [75]Department of Preventive Dentistry and Periodontology, Ludwig-Maximilians-University Hospital Munich-Grosshadern, Munich, Germany. [76]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada. [77]Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah, USA. [78]Department of Medicine, St Mark's Hospital, Harrow, Middlesex, UK. [79]Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham, UK. [80]Molecular Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia. [81]Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, Edinburgh, UK. [82]Child Life and Health, University of Edinburgh, Edinburgh, Scotland, UK. [83]Division of Rheumatology Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [84]Université Paris Diderot, Sorbonne Paris-Cité, Paris, France. [85]Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania. [86]Research Institute of Internal Medicine, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway. [87]Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway. [88]K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [89]Genetic Medicine, Manchester Academic Health Science Centre, Manchester, UK. [90]The Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK. [91]QIMR Berghofer Medical Research Institute, Royal Brisbane Hospital, Brisbane, Australia. [92]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA. [93]Department of Gastroenterology, Kaunas University of Medicine, Kaunas, Lithuania. [94]Faculty of medicine, University of Latvia, Riga, Latvia. [95]Dipartimento di Neuroscienze, Psicologia, Area del Farmaco e Salute del Bambino

(NEUROFARBA), Universitˆ di Firenze Strutture Organizzative Dipartimentali (SOD) Gastroenterologia e Nutrizione Ospedale pediatrico Meyer, Firenze, Italy. [96]Department of Gastroenterology, The Townsville Hospital, Townsville, Australia. [97]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK. [98]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada. [99]Center for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Maribor, Slovenia. [100]Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK. [101]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [102]Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong. [103]Department of Genomics Life & Brain Center, University Hospital Bonn, Bonn, Germany. [104]Department of Gastroenterology, Academic Medical Center, Amsterdam, The Netherlands. [105]Faculty for Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia. [106]Department of Gastroenterology, Royal Brisbane and Womens Hospital, Brisbane, Australia. [107]Department of Gastroenterology, Université Catholique de Louvain (UCL) Centre Hospitalier Universitaire (CHU) Mont-Godinne, Mont-Godinne, Belgium. [108]Department of Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St-Thomas Hospital, London, UK. [109]Department of Digestive Diseases, Hospital Quiron Teknon, Barcelona, Spain. [110]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA. [111]Human Genetics, Genome Institute of Singapore, Singapore. [112]Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA. [113]Department of Biology of Radiations and Human Health, Agenzia nazionale per le nuove tecnologie l'energia e lo sviluppo economico sostenibile (ENEA), Rome, Italy. [114]Department of Clinical Science Intervention and Technology, Karolinska Institutet, Stockholm, Sweden. [115]Gastroenterology & General Medicine, Norfolk and Norwich University Hospital, Norwich, UK. [116]Department of Gastroenterology, Leiden University Medical Center, Leiden, The Netherlands. [117]Faculty of medicine, University of Toronto, Toronto, Ontario, Canada. [118]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA. [119]Institute of Human Genetics, Technische Universität München, Munich, Germany. [120]Department of Neurology, Technische Universität München, Munich, Germany. [121]Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, USA. [122]Department of Gastroenterology, West China Hospital, Chengdu, Sichuan, China. [123]State Key Laboratory of Biotherapy, Sichuan University West China University of Medical Sciences (WCUMS), Chengdu, Sichuan, China

Acknowledgements (original data): We thank all the subjects who contributed samples and the physicians and nursing staff who helped with recruitment globally. UK case collections were supported by the National Association for Colitis and Crohn's disease; Wellcome Trust grant 098051 (L.J., C.A.A., J.C.B.); Medical Research Council UK; the Catherine McEwan Foundation; an NHS Research Scotland career fellowship (R.K.R.); Peninsula College of Medicine and Dentistry, Exeter; the National Institute for Health Research, through the Comprehensive Local Research Network, and through Biomedical Research Centre awards to Guy's & Saint Thomas' National Health Service Trust, King's College London, Addenbrooke's Hospital, University of Cambridge School of Clinical Medicine and to the University of Manchester and Central Manchester Foundation Trust. The British 1958 Birth Cohort DNA collection was funded by Medical Research Council grant G0000934 and Wellcome Trust grant 068545/Z/02, and the UK National Blood Service controls by the Wellcome Trust. The Wellcome Trust Case Control Consortium projects were supported by Wellcome Trust grants 083948/Z/07/ Z, 085475/B/08/Z and 085475/Z/08/Z. North American collections and data processing were supported by funds to the National Institute of Diabetes, Digestive and Kidney diseases (NIDDK) IBD Genetics Consortium, which is funded by the following grants: DK062431 (S.R.B.), DK062422 (J.H.C.), DK062420 (R.H.D.), DK062432 (J.D.R.), DK062423 (M.S.S.), DK062413 (D.P.M.), DK076984 (M.J.D.), DK084554 (M.J.D. and D.P.M.) and DK062429 (J.H.C.). The funding source which enabled our recruitment of New Zealand subjects for the IIBDGC was the New Zealand Ministry of Business, Innovation and Employment.