Refinement in Localization and Identification of Gene Regions Associated with Crohn Disease

Heather Elding,[1] Winston Lau,[1] Dallas M. Swallow,[1] and Nikolas Maniatis[1,*]

The risk of Crohn disease (CD) has a large genetic component. A recent meta-analysis of 6 genome-wide association studies reported 71 chromosomal intervals but does not account for all of the known genetic contribution. Here, we refine localization of the previously reported intervals and also identify additional CD susceptibility genes using a mapping approach that localizes causal variants based on genetic maps in linkage disequilibrium units (LDU maps). Using 2 of the 6 cohorts, 66 of the 71 previously reported loci are confirmed and more precise location estimates for these intervals are given. We identify 78 additional gene regions that pass genome-wide significance, providing strong evidence for 144 genes. Additionally, 56 nominally significant signals, but with more stringent and precise colocalization, are identified. In total, we provide evidence for 200 gene regions confirming that CD is truly multifactorial and complex in nature. Many identified genes have functions that are compatible with involvement in immune/inflammatory processes and seem to have a large effect in individuals with extra ileal as well as ileal inflammation. The precise locations and the evidence that some genes reflect phenotypic subgroups will help identify functional variants and will lead to greater insight of CD etiology.

Crohn Disease (CD) is, like ulcerative colitis (UC), a major subtype of Inflammatory Bowel Disease (IBD [MIM 266600]). Although CD differs from UC in several respects (e.g., clinical manifestations, cytokine profile), they both involve chronic intestinal inflammation but with variable and somewhat overlapping manifestations. CD principally involves the ileum, but there is variation in disease phenotype. There are inflammatory, stricturing, or penetrating types that may affect different sites in the gastrointestinal tract (GI) and also differences in the age of onset. This variability results from the interaction of environmental factors, including gut microbiota and the immune and inflammatory mechanisms of the genetically susceptible host. CD is a polygenic disorder with some high-penetrance genes and an estimated individual sibling recurrence risk ratio (λs) ranging from 15 to 30.[1] Because of this high λs, CD is one of the most widely studied common multifactorial diseases. The genetic contribution was first explored by linkage analysis using families, which led to the important identification of the role of *NOD2* on chromosome 16q (MIM 605956). Further progress was subsequently made through genome-wide association studies (GWASs).[2–11] A recent meta-analysis explored 6 GWASs and identified a total of 71 signals of association with CD using more than 6,000 cases and 15,000 controls.[5] The authors estimated that these loci explain less than one quarter[5] of the reported heritability in liability. A GWAS utilizes hundreds of thousands of SNP markers across the genome. There is a lack of power to detect genome-wide association partly because of (1) the highly stringent significance thresholds required as a result of multiple testing, and (2) the unrealistic assumption that by testing each SNP one at time, one of the markers in the genotyping platform used will either be the causal or in almost complete linkage disequilibrium (LD) with the causal variant. An additional problem is that in order to achieve sufficient power, studies are often combined with the aim of increasing sample size. As a result, data sets often include cases with variable disease onset, variable phenotype definition, and sample collection in different geographic regions. Taken together, this not only means that rare variants of large effect are likely to be missed but that common variants with small effect or that apply only to a particular subgroup could also be missed.

Our recent study of the chromosome 16q linkage region using two of these GWASs illustrated how a multimarker approach based on high-resolution LD maps can provide additional study power. We were also able to explain heterogeneity in a genomic region first identified by linkage analysis.[12] By using UK data provided by the Wellcome Trust Case Control Consortium (WTCCC), we identified, on this chromosome arm, with high statistical confidence, several new gene regions that are involved in inflammation and immune dysregulation. These distinct signals were replicated with high precision of location using independent North American data provided by the National Institute of Diabetes and Digestive and Kidney Diseases IBD Genetics Consortium (NIDDK). The study highlighted the importance of genetic heterogeneity (i.e., the involvement of different risk genes in different individuals) within the extensively studied *NOD2* gene region and demonstrated the independent involvement of the nearby *CYLD* (MIM 605018).[12] We also illustrated the importance of accurate and detailed phenotype definition in revealing gene association. By examining the group of NIDDK cases that had involvement of nonileal intestinal regions, as well as the ileum, we were able to replicate association with *CDH1* (encoding E-cadherin [MIM 192090]), which is located coincident with a

[1]Research Department of Genetics Evolution & Environment, University College London, Gower Street, London WC1E 6BT, UK
*Correspondence: n.maniatis@ucl.ac.uk
http://dx.doi.org/10.1016/j.ajhg.2012.11.004.

previously described linkage peak in cases without *NOD2* mutations.

The purpose of this study is to analyze the whole genome using a gene-mapping approach based on LDU maps in order to identify additional genes for CD and provide the basis for new etiological insights into the disease. In order to accomplish this aim, we used 2 GWA data sets. From the UK WTCCC GWAS data[2] based on the Affymetrix 500K Array, we analyzed a total of 1,698 cases of CD and 2,948 controls. The CD cases consisted of individuals with any subtype of CD. The GWAS replication data consisted of 813 North American CD cases and 947 ethnically matched controls, and all cases exhibit inflammation of the ileum. These GWAS data are available from the NIDDK IBDGC with the genotype data derived from the Illumina HumanHap300 Array. This array consists of a much lower SNP resolution and a set of SNPs that only partially overlap with the WTCCC 500K Affymetrix Array. Although a smaller data set, the NIDDK data contains a richer phenotypic description of cases and includes the various additional locations of intestinal inflammation for each individual. Further details about the GWAS data sets are available in other reports.[2,11]

The mapping approach in this study uses genetic locations derived from a high-resolution LD map with distances in LD units (LDU).[13] High-density LDU maps were constructed using data from the HapMap Phase II Project. They can be visualized by plotting marker location in LDU against the marker distances in kb. Plotting these LDU maps reveals the nonlinear relationship between physical distance and the underlying LD together with the "block-step" structure of the region in question (see Figure 1). "Blocks" of LD represent areas of low haplotype diversity whereas "steps" define LD breakdown mainly caused by recombination because crossover profiles agree with LD patterns.[14]

The construction of the LDU maps is based on the Malécot Model, which describes the decline of pairwise SNP associations as a function of physical distance in kilobases (kb).[13] For the association mapping analyses, all the SNPs in the WTCCC and NIDDK data sets were assigned an LDU location from the HapMap LDU maps. We then divided each chromosome arm into nonoverlapping analytical windows, which were constrained to 10 LDU and a minimum of 60 SNPs. The windows were identical for both data sets. For each window along the whole genome, a composite-likelihood test[15] was used, in which all the SNPs within a window are simultaneously tested. Therefore, the method avoids the need for imputation. Because each window is a test, then the total number of tests performed is substantially lower than single SNP analysis. This has important implications in multiple testing. Each test is also based on the same Malécot model used to construct the LDU maps, although in this case, we model the decline of affection-status-by-SNP association as a function of genetic distance in LDU. The model uses the locations from the LDU map in order to estimate the most likely location ($\hat{S}$) of the causal agent. The method returns a p value for association of the window together with an estimated $\hat{S}$ for each window. In windows where $\hat{S}$ mapped to the end of a window-boundary, that window was extended in that direction to capture information on the far side to improve the accuracy of location. The construction of the LDU maps using the HapMap Phase II data has been described in detail elsewhere[16] and the association mapping methodology presented here using the LDU maps is given by Maniatis et al.[15] This method has been shown to increase power and localization over kb maps.[12,15]

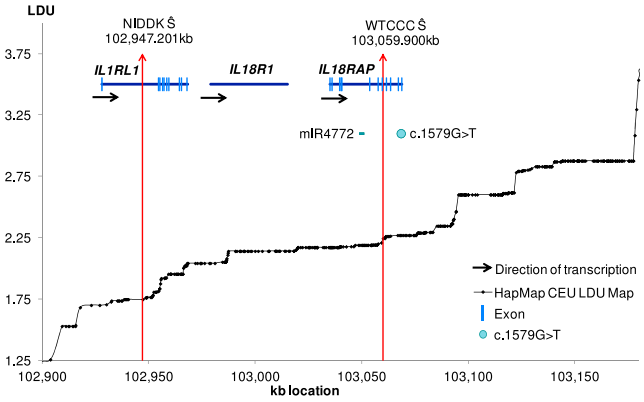

Figure 1. Localization within the *IL1RL1/IL18RAP* Region
The LD map of the region is shown by plotting LDU (y axis) against kb (x axis). The red vertical arrows are the estimated locations Ŝ for the WTCCC and NIDDK data sets. The WTCCC Ŝ is within an intron and between a miRNA (mIR4772) and coding rare variant associated with CD within *IL18RAP* (NM_003853.2) c.1579G>T (p.Val527Leu).

Our recent work[12] showed that stratifying the analysis according to the occurrence of extraileal intestinal involvement, as opposed to ileal inflammation alone (a subdivision that appears to accompany genetic stratification[17]), further increases the power of detecting association. We stratified the NIDDK data based on detailed information regarding the location of individuals' intestinal inflammation. For all the analytical windows that showed nominal statistical evidence of association using the WTCCC data, we analyzed the same windows using the entire NIDDK data and also using a subset of the NIDDK data containing CD cases that reported extraileal in addition to ileal inflammation. Unlike the controls, the vast majority of cases in this subdivision were non-Jewish. We therefore analyzed only non-Jewish cases with extraileal inflammation and non-Jewish controls. This subdivision reduced the sample size from 813 to 277 CD cases and from 947 to 515 controls.

In an initial analysis, we looked for signals within the previously published intervals from the recent meta-analysis[5] and present p values for the relevant data sets. Because the majority of these intervals were more than 200 kb in length (defined around the most significant SNP using LD information[5]), it was difficult to meta-analyze

Table 1. Examples Taken from the Previously Reported Intervals, Showing the Whole-Genome Association Statistics and Closest Gene to the Estimated Location of the Causal Agent $\hat{S}$

Chr	Reported LD Interval (Mb)	Interval Span (Mb)	WTCCC $\hat{S}$	WTCCC p Value	NIDDK $\hat{S}$	NIDDK p Value	Gene (WTCCC/NIDDK If different)
1p31.3	67.36 – 67.77	0.41	67,684.8	2.1×10^{-11}	67,707.3	3.0×10^{-10}	*IL23R*[a]
2q12.1	102.80 – 103.30	0.50	103,059.8	1.8×10^{-08}	102,951.7	1.7×10^{-05}	*IL18RAP*[a]/*IL1RL1*[a]
2q37.1	234.15 – 234.57	0.42	234,144.8	5.6×10^{-25}	234,171.9	2.2×10^{-09}	*ATG16L1*[a]
5q31.1	129.38 – 132.02	2.64	131,748.2	2.7×10^{-22}	131,631.4	8.9×10^{-07}	*C5orf56*[a]/*SLC22A4*[a]
16q12.1	50.46 – 50.85	0.39	50,803.2	2.6×10^{-15}	50,803.2	1.1×10^{-09}	*NOD2*[a]
			50,846.3	1.6×10^{-13}	50,846.5	4.1×10^{-04}	*CYLD*

[a]At least one of the estimated locations $\hat{S}$ is located within the identified gene. All locations in kb are in NCBI 37. Full details on the entire list of 71 LD intervals are given in Table S1.

the results obtained using this mapping procedure. We therefore present the p values obtained from WTCCC and NIDDK separately. For the additional genes we identified, we present a meta-analysis p value for each analytical window that shows (1) evidence of nominal statistical significance for each data sample and (2) that the $\hat{S}$ estimates for both samples are within 150 kb of one another. These two criteria were used in order to reduce the possibility of heterogeneity and thus facilitate the meta-analysis of shared locations. In order to account for multiple testing, we used a Bonferroni p value threshold of 1×10^{-5}, which corresponds to the 4,993 tests, to the final meta p value. This number of tests included the total number of analytical windows across the genome (4,399 windows with identical boundaries for both data sets) plus the additional analyses performed for the NIDDK data using the extraileal subgroup.

In order to capture replications of slightly lower significance but with more precisely agreeing $\hat{S}$ values, we used a second threshold in which the $\hat{S}$ for both data sets had to lie within 80 kb of each other to be considered a replication and with a meta p value within the range of 1×10^{-3} and 1×10^{-5}.

For nominally significant windows of each data set, the closest gene of the estimated location $\hat{S}$ was retrieved and then listed in the tables. Using BioMart, the Gene Ontology (GO) annotations attributed to each gene were retrieved and are listed. Using the Cytoscape plugin BiNGO,[18] we also checked for overrepresentation of GO terms from the GO Slim GOA database using a hypergeometric test.

Our first goal was to determine whether the LDU mapping method could detect the 71 intervals previously identified from the most recent large meta-analysis of six independent data sets. Table 1 shows some examples and Table S1 presents the statistical evidence and location estimates, for significant signals within each of these 71 intervals.[5] Our analyses identified a total 66 out of the 71 intervals (loci) in one or both of these data sets using an uncorrected significance threshold of $p < 0.05$. Only five intervals showed no evidence of association for either WTCCC or NIDDK using the Malécot model. Notably, the first published results by the WTCCC[2] reported just nine loci, but Table S1 (available online) shows 28 signals for the WTCCC alone that passed our genome-wide significance threshold (1×10^{-5}). The majority of the 66 intervals (88%) were identified with the larger WTCCC data with the remaining eight intervals (12%) showing nominal evidence of association for the smaller NIDDK data set alone. Interestingly, these eight include regions (e.g., 10q22, 16p11) that have previously been implicated for pediatric IBD cases[7] (Table S1). This is consistent with the fact that the NIDDK data set probably includes a larger number of individuals with early onset CD (~37%) than WTCCC, where this information is not recorded. It is worth noting that one of the six contributory data sets to the most recent meta-analysis included entirely pediatric IBD cases.[5]

Unlike the recent meta-analysis, where a particular gene or genes within these intervals was identified through a series of in silico analyses, we present the genes closest to the location estimates we have obtained using this method (Table 1; Table S1). In several instances, these agree with the previously identified genes (e.g., *STAT3* [MIM 102582], *IL23R* [MIM 607562]; Table S1), but in other instances there are differences. For several regions, we obtain two different estimates of localization ($\hat{S}$) for WTCCC and NIDDK, both within the same interval. Heterogeneity in some regions is expected and location estimates can help identify this. For example, for the interleukin-rich interval (2q12.1), the WTCCC data set yielded an estimate within *IL18RAP* (MIM 604509) but for NIDDK the $\hat{S}$ is within *IL1RL1* (MIM 601203; Table 1). Both genes have been previously suggested as candidates within the 2q12.1 interval,[5] but an independent fine mapping study has shown strong association to the *IL18RAP* rs917997 SNP for both CD and UC,[19] which is approximately 10 kb away from WTCCC $\hat{S}$ (Table 1; Figure 1). A recent study[20] using deep sequencing identified a rare coding variant of possible functional significance within *IL18RAP* (NM_003853.2:c.1579G>T [p.Val527Leu]), which is only 8 kb away from the WTCCC location, though it is probable that other functional variants are also involved. As far as

Table 2. Whole-Genome Association Statistics and the Closest Gene to the Estimated Location of the Causal Agent for the 78 Gene-Regions that Passed Our Genome-Wide Significance Threshold 1×10^{-5} with $\hat{S}$ Locations of 150 kb or Less between the Two Data Sets

Chr	Gene region (WTCCC/NIDDK if different)
1	*NFIA*[b], ***PGM1***[a]***/EFCAB7***[a,b], *RORC*[a]*/THEM4*, ***KCNK1***[a,b], ***LGALS8***[a], ***ACTN2***[a]***/HEATR1***
2	*LOC400940*, ***FAM49A***, *DNAJC27-AS1*[a]*/ADCY3*[a], *DCTN1*[a]*/RTKN*[a,b], ***intergenic (2q14.1)***, *STK39*[a], *ZNF804A*[a], *ERBB4*[a]
3	*STAC/ARPP21*, ***ITGA9***[a], ***EPHA6***[b], *COL8A1*
4	*TEC*[a]*/SLAIN2*, *LNX1/RPL21P44*[b], ***PDGFRA***[b], ***COL25A1***[a,b], *SNX25*[a]
5	***MCC***[a,b], *MEGF10/PRRC1*[b]
6	***GPLD1***[a], *GABBR1*[a], *EYS*[a], *FNDC1*[a], *T*[b],
7	***MACC1/TMEM196***, *NPY*[b], ***INHBA-AS1/GLI3***, *SEMA3C*[a], *TSPAN12/KCND2*[a], ***intergenic***[b] **(7q31.33)**, *PODXL*[b]
8	*MTUS1*[a], *LOC100507632*, *STMN2*[b], ***intergenic (8q21.13)***, *DCAF13/FZD6*[b], ***intergenic***[b] ***(8q23.3)***
9	*DOCK8*[a,b]
10	***GPR158***[a,b], ***FAM13C***[a], *ARID5B/C10orf107*, ***PLAU***[b], *KCNMA1*[a], *XPNPEP1*, ***NRAP***[a]
11	*OR52N4/OR52N2*[b], *BDNF*, ***OR9G4***, ***KIRREL3***[a]
12	*CCDC91*[b], ***TMEM117***[a,b], ***C12orf74***[b], ***ANO4***[a]
13	***MIR4703***
14	*PNN/GEMIN2*[a,b], ***NRXN3***[a,b]
15	*OTUD7A*[a], *PGBD4/CHRM5*[a]
16	*RBFOX1*[b], ***RBFOX1***[a], *BRD7*, ***HSD11B2***[b], ***IRF8***
17	*GRB2*[a]*/MIR3678*
18	*GATA6*, ***RIT2***[a,b], *IER3IP1*[b]
20	*HAO1*[a,b], *LOC339568*, ***CD40/SLC12A5***[a,b]
22	*CSF2RB*[a]*/LOC1005006241*
X	*FMR1NB*

bold denotes that the signal is replicated using the data stratified according to non-Jewish ileal and extraileal inflammation.
[a]At least one of the estimated locations $\hat{S}$ is located within the identified gene.
[b]The signal is significant in the pooled NIDDK data as well as the stratified non-Jewish ileal and extraileal data set. All locations in kb are in NCBI 37. Full details (p values and estimates of $\hat{S}$) are given in Table S2.

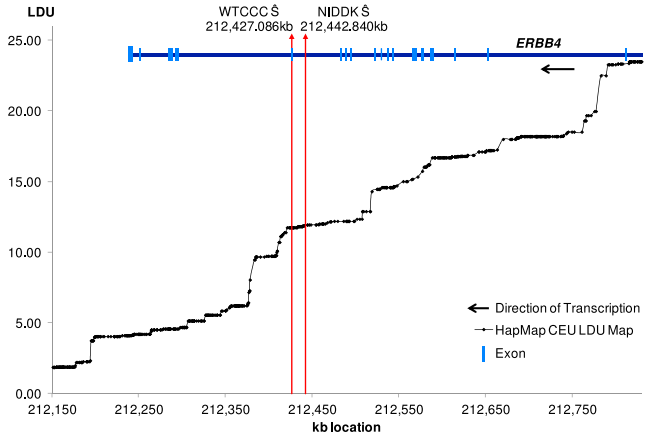

Figure 2. Localization within the *ERBB4* Region
The LD map of the region is shown by plotting LDU (y axis) against kb (x axis). The red vertical arrows are the estimated locations $\hat{S}$ for the WTCCC and NIDDK data sets.

the second gene is concerned, the importance of *IL1RL1* to inflammatory processes has also been documented[21] for a variety of human pathologies including celiac disease.[22] The LDU map clearly shows that there are several recombination hot spots between *IL18RAP* and *IL1RL1* (Figure 1) making it likely that these two signals are independent and that there is genetic heterogeneity within these previously identified intervals. This result is similar to that previously noted in the case of *NOD2* and *CYLD*[12] (Table 1).

As well as our 66 localization estimates, we identified 134 additional signals, of which 78 passed our genome-wide significance threshold (1×10^{-5}; Table 2; Table S2). Of these 78 signals, 64 were replicated using the complete NIDDK data set, of which half (32) were also replicated in the smaller subset of the data (ileal and extraileal). However, notably, in half of these (16) the association showed higher significance in the smaller subset despite the substantial decrease in the number of cases and controls (see entries in bold marked with b in Table 2). Furthermore, the remaining 14 (22%) of the WTCCC signals were only replicated using this subset of the NIDDK data.

For the majority (74%) of the 78 signals, the location estimates from both data sets point to the same gene, with approximately half of the signals being intragenic and half intergenic locations (gene with "a" indicates that at least one of the signals is within the gene; Table 2). In all signals, including the ~50% that reside outside of genes, the closest gene is considered as a candidate. It should however be noted that in some instances, functional variation in *cis*-elements that regulate genes far away from the location determined here may be involved.[23]

Several of the additional genes we identified have been reported to be involved in inflammatory and/or immune dysregulation conditions (e.g., *DOCK8* [MIM 611432], *ITGA9* [MIM 603963]). In other instances, the genes have been previously implicated in colonic inflammation or epithelial cell morphology based on functional studies. By way of illustration, Figure 2 presents an example of the estimated locations ($\hat{S}$) for the genomic region that harbors *ERBB4* ([MIM 600543], 2q34 window). For both data sets, this window was significantly associated with CD with a meta p value of 3×10^{-6} (Table 2). Figure 2 also shows the LDU map starting from the 3′ region and spanning halfway along the gene, which stretches across more than 1 megabase (Mb). The map shows numerous short LD blocks across the region. However, despite the large LD breakdown in the area, both estimates of $\hat{S}$ are close to each other and within the same intron. Interestingly, *ERBB4* (which encodes Receptor protein-tyrosine

Table 3. List of Genes for the 56 Signals with p Values 10^{-3}–10^{-5} and with $\hat{S}$ Locations of 80 kb or Less between the Two Data Sets

Chr	Gene region (WTCCC/NIDDK if different)
1	*LOC400752*, *DAB1*[a], ***GADD45A***, ***intergenic (1p31.1)***, *intergenic*[b] (1q31.3), *REN*[a], *DNAH14*[a], ***H3F3A***[b]
2	*ETAA1*[b], *IL1R2*[a], ***LRP2***[a], *NYAP2*
3	*CACNA2D3*, *ROBO2*[a], *BTLA*[a], *AGTR1*[a]
4	*HSP90AB2*[b], ***intergenic (4q13.2)***, *PRDM5*[a]
5	***CDH12***, ***intergenic (5q21.1)***, *FAM71B/ITK*[a,b]
6	*SOX4*, *BEND3*[a], *MARCKS*, *STL*[b], *IFNGR1*, *intergenic* (6q25.3), *SNX9*[a,b]
7	***MACC1***, ***NPSR1***[a], ***IQUB***[b], *KEL*[a]
8	*SLA*[a]
9	*C9orf85/C9orf57*, ***GNA14***[a,b], ***intergenic (9q33.1)***
10	*FAS*[b]
11	*MOB2*[a], ***MICALCL***[a]
12	***CAND1***, *KITLG*[b], *LOC100128554*[b]
13	***GPR12***, *NBEA*[a], ***intergenic (13q31.3)***, ***intergenic (13q31.3)***
14	*intergenic*[b] *(14q31.1)*
15	***ATP8B4***[a]
16	*CDH8*[a], *CNTNAP4*, *WWOX*[a,b]
20	*MACROD2*[a]
21	*DSCAM*[a]
22	*ISX*[b]
X	***intergenic (Xq23)***

bold denotes that the signal is replicated using the data stratified according to non-Jewish ileal and extraileal inflammation.
[a]At least one of the estimated locations $\hat{S}$ is located within the identified gene.
[b]The signal is significant in the Pooled NIDDK data as well as the stratified non-Jewish ileal and extraileal data set. All locations in kb are in NCBI 37. Full details (p values and estimates of $\hat{S}$) are given in Table S3.

kinase erbB-4) is expressed at high levels in the inflamed colonic mucosa of CD cases.[24] Using adult mouse colon, it was also shown that ERBB4 is an important regulator in the epithelial response to inflammation and injury.[24] *ERBB4* expression has been linked to a number of cellular processes such as cell survival, proliferation, and tumorigenesis in different tissues.[25,26] A recent study suggested that this elevated *ERBB4* expression could lead to colitis-associated development of colorectal tumors.[27] *ERBB4* has also been implicated, together with E-cadherin (encoded by *CDH1* on 16q22.1, previously identified as a risk gene for CD[12] and ulcerative colitis[28]), in the suppression of anoikis[29] (programmed cell death as a result of cell detachment from the extracellular matrix), suggesting that there may be interaction between these two genes. Despite the fact that *ERBB4* is a very large gene (>1 Mb), the estimated locations of the causal agent for both data sets (WTCCC and NIDDK) are very close and within the same intron, making it an excellent target for follow-up fine mapping studies and resequencing.

Table 3 shows the list of 56 of the 134 additional signals that showed association with CD with p values that ranged between 1×10^{-3} and 1×10^{-5}. These p values were below the Bonferroni threshold, but the windows were selected on more stringent criteria for colocalization (i.e., estimates of localization in the two data sets were not more than 80 kb apart from each other). Detailed information (p values, location estimates) for the all the genes listed in Table 3 are given in the Table S3. As in Table 2, most of the signals in Table 3 give estimates of $\hat{S}$ that point to the same gene. Several of these genes are also very interesting. For example, the window harboring *BTLA* (MIM 607925) gave location estimates that are just 2 kb apart. A study using a mouse model of colitis has shown functional evidence of *BTLA* involvement in colitis.[30]

In summary, using a powerful multimarker association mapping approach based on LDU maps, we confirm 66 of the 71 previously reported loci from meta-analysis and 134 additional gene regions that are associated with Crohn disease, providing evidence for 200 gene regions that include CD susceptibility loci. We provide more precise location estimates for the 66 previously published intervals, as well as locations for the 134 additional signals. This is a major step forward in identifying the relevant genes and functional variants and thus elucidating the genetics of CD etiology. The very large numbers of genes listed in Tables 2 and 3 confirm that CD is truly polygenic and complex in nature. Many genes show functions that are compatible with involvement in immune and/or inflammatory processes as well as integrity of the intestinal epithelium and differentiation. The groups of genes were enriched for GO terms such as signal transducer activity, receptor activity, and binding. These functions are important in inflammatory processes (see Supplemental Information; Tables S4 and S5).

Gene mapping by linkage has a long history, but association mapping was in its infancy when high-throughput SNP genotyping was developed. There is currently some skepticism regarding the success of genome-wide association mapping, but here we demonstrate its great potential in dissecting the genetics of common disease—at least for Crohn, for which relatively highly penetrant genes are known to exist (as indicated by high λ_S estimates). We also demonstrate that the use of smaller data sets and detailed phenotypic information can detect a great proportion of the genes involved, as well as genetic heterogeneity. Here we get a glimpse of the way in which different genes are likely to be involved in different but overlapping pathways. Many of the replicated signals were more significant using a subset of the North American data (NIDDK), which contained individuals with ileal CD who had involvement of at least one extraileal intestinal location. Despite the substantial decrease in sample size, analysis of this subphenotype using our mapping method yielded much higher significance than the analysis of the full data. This result shows that accurate and detailed

phenotype information is extremely important in genetic studies especially when assessing a trait that shows clinical variation.

The majority of the 200 signals unambiguously identify single genes. Their confirmation in independent cohorts makes it likely that most have an effect (in many instances small, though sometimes large) on the risk of CD. The genomic locations we provide are estimates of the positions of the variants of functional significance assuming a monophylogenetic origin (i.e., no substantial allelic heterogeneity). We have previously shown the greater power of LDU compared to physical maps in detecting association and refining location using the same multimarker methodology.[31] The close distance between the location estimates in the two different cohorts (10 kb or less in 30% of the additional signals) suggests that these allelic variants will often be the same, but in some instances there are clear indications of independent allelic variants. In either case, the localization offers much greater precision than other methods for future in depth analysis using bioinformatics and resequencing.

The identified genes will facilitate more powerful experimental research designs including a range of functional studies and analytical pathway analyses for CD. Given the high-resolution estimates, the results presented here will allow more targeted, intelligently designed resequencing studies, where carefully characterized, homogeneous subgroups of cases can be more thoroughly investigated. Such research strategies should move us closer to more effective, personalized medical treatment for individuals with Crohn disease.

Supplemental Data

Supplemental Data includes five tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We are very grateful to the Wellcome Trust, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the International HapMap Project for making their invaluable data available to the scientific community. The data from the Wellcome Trust Case-Control Consortium (WTCCC)[2] was funded by the Wellcome Trust. A full list of the investigators who contributed to the generation of the data is available on the WTCCC website. The NIDDK IBD Genetics Consortium Crohn Disease GWAS was conducted by Judy H. Cho, Yale University; Steven Brant, Johns Hopkins University; Richard Duerr, University of Pittsburgh; Huiying Yang, Cedars-Sinai Medical Center; John Rioux, University of Montreal; and Mark Silverberg, University of Toronto, with support from the NIDDK. This manuscript was not prepared in collaboration with the laboratories of any of the investigators responsible for generating the data and does not necessarily reflect the views or opinions of these investigators or the NIDDK. The data sets used were obtained from the database of Genotypes and Phenotypes (dbGaP) at accession number phs000130. We thank Toby Andrew for his useful comments and suggestions and the Annals of Human Genetics for the financial support for H.E.'s PhD degree.

Received: July 16, 2012
Revised: August 3, 2012
Accepted: November 5, 2012
Published: December 13, 2012

Web Resources

The URLs for data presented herein are as follows:

Gene Ontology, http://www.geneontology.org/
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), www.niddk.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
UCSC Genome Browser, http://genome.ucsc.edu/
Wellcome Trust Case Control Consortium (WTCCC), http://www.wtccc.org.uk/

References

1. Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J., and Krawczak, M. (2005). Genetics of Crohn disease, an archetypal inflammatory barrier disease. Nat. Rev. Genet. *6*, 376–388.
2. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
3. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.
4. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461–1463.
5. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. *42*, 1118–1125.
6. Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K., Albrecht, M., Mayr, G., De La Vega, F.M., Briggs, J., et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat. Genet. *39*, 207–211.
7. Imielinski, M., Baldassano, R.N., Griffiths, A., Russell, R.K., Annese, V., Dubinsky, M., Kugathasan, S., Bradfield, J.P., Walters, T.D., Sleiman, P., et al.; Western Regional Alliance for Pediatric IBD; International IBD Genetics Consortium; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. (2009). Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat. Genet. *41*, 1335–1340.
8. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A., et al. (2007). Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. *3*, e58.

9. Mathew, C.G. (2008). New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat. Rev. Genet. *9*, 9–14.
10. McGovern, D.P., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., Ippoliti, A., Vasiliauskas, E., Berel, D., Derkowski, C., et al.; International IBD Genetics Consortium. (2010). Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum. Mol. Genet. *19*, 3468–3476.
11. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W., et al. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet. *39*, 596–604.
12. Elding, H., Lau, W., Swallow, D.M., and Maniatis, N. (2011). Dissecting the genetics of complex inheritance: linkage disequilibrium mapping provides insight into Crohn disease. Am. J. Hum. Genet. *89*, 798–805.
13. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X., and Morton, N.E. (2002). The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc. Natl. Acad. Sci. USA *99*, 2228–2233.
14. Webb, A.J., Berg, I.L., and Jeffreys, A. (2008). Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc. Natl. Acad. Sci. USA *105*, 10471–10476.
15. Maniatis, N., Collins, A., and Morton, N.E. (2007). Effects of single SNPs, haplotypes, and whole-genome LD maps on accuracy of association mapping. Genet. Epidemiol. *31*, 179–188.
16. Lau, W., Kuo, T.Y., Tapper, W., Cox, S., and Collins, A. (2007). Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics *23*, 517–519.
17. Ahmad, T., Armuzzi, A., Bunce, M., Mulcahy-Hawes, K., Marshall, S.E., Orchard, T.R., Crawshaw, J., Large, O., de Silva, A., Cook, J.T., et al. (2002). The molecular classification of the clinical manifestations of Crohn's disease. Gastroenterology *122*, 854–866.
18. Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448–3449.
19. Zhernakova, A., Festen, E.M., Franke, L., Trynka, G., van Diemen, C.C., Monsuur, A.J., Bevova, M., Nijmeijer, R.M., van 't Slot, R., Heijmans, R., et al. (2008). Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. Am. J. Hum. Genet. *82*, 1202–1210.
20. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al.; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.
21. Akhabir, L., and Sandford, A.J. (2011). Genome-wide association studies for discovery of genes involved in asthma. Respirology *16*, 396–406.
22. Amundsen, S.S., Rundberg, J., Adamovic, S., Gudjónsdóttir, A.H., Ascher, H., Ek, J., Nilsson, S., Lie, B.A., Naluai, A.T., and Sollid, L.M. (2010). Four novel coeliac disease regions replicated in an association study of a Swedish-Norwegian family cohort. Genes Immun. *11*, 79–86.
23. Jones, B., and Swallow, D. (2011). The impact of cis-acting polymorphisms on the human phenotype. The HUGO Journal *5*, 13–23.
24. Frey, M.R., Edelblum, K.L., Mullane, M.T., Liang, D., and Polk, D.B. (2009). The ErbB4 growth factor receptor is required for colon epithelial cell survival in the presence of TNF. Gastroenterology *136*, 217–226.
25. Erlich, S., Goldshmit, Y., Lupowitz, Z., and Pinkas-Kramarski, R. (2001). ErbB-4 activation inhibits apoptosis in PC12 cells. Neuroscience *107*, 353–362.
26. Starr, A., Greif, J., Vexler, A., Ashkenazy-Voghera, M., Gladesh, V., Rubin, C., Kerber, G., Marmor, S., Lev-Ari, S., Inbar, M., et al. (2006). ErbB4 increases the proliferation potential of human lung cancer cells and its blockage can be used as a target for anti-cancer therapy. International Journal of Cancer *119*, 269–274.
27. Frey, M.R., Hilliard, V.C., Mullane, M.T., and Polk, D.B. (2010). ErbB4 promotes cyclooxygenase-2 expression and cell survival in colon epithelial cells. Laboratory Investigation; a journal of technical methods and pathology *90*, 1415–1424.
28. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A., et al. (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat. Genet. *43*, 246–252.
29. Kang, H.G., Jenabi, J.M., Zhang, J., Keshelava, N., Shimada, H., May, W.A., Ng, T., Reynolds, C.P., Triche, T.J., and Sorensen, P.H. (2007). E-cadherin cell-cell adhesion in ewing tumor cells mediates suppression of anoikis through activation of the ErbB4 tyrosine kinase. Cancer Res. *67*, 3094–3105.
30. Steinberg, M.W., Turovskaya, O., Shaikh, R.B., Kim, G., McCole, D.F., Pfeffer, K., Murphy, K.M., Ware, C.F., and Kronenberg, M. (2008). A crucial role for HVEM and BTLA in preventing intestinal inflammation. J. Exp. Med. *205*, 1463–1476.
31. Andrew, T., Maniatis, N., Carbonaro, F., Liew, S.H.M., Lau, W., Spector, T.D., and Hammond, C.J. (2008). Identification and replication of three novel myopia common susceptibility gene loci on chromosome 3q26 using linkage and linkage disequilibrium mapping. PLoS Genet. *4*, e1000220.