nature genetics

Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease

Gosia Trynka[1,36], Karen A Hunt[2,36], Nicholas A Bockett[2], Jihane Romanos[1], Vanisha Mistry[2], Agata Szperl[1], Sjoerd F Bakker[3], Maria Teresa Bardella[4,5], Leena Bhaw-Rosun[6], Gemma Castillejo[7], Emilio G de la Concha[8], Rodrigo Coutinho de Almeida[1], Kerith-Rae M Dias[6], Cleo C van Diemen[1], Patrick C A Dubois[2], Richard H Duerr[9,10], Sarah Edkins[11], Lude Franke[1], Karin Fransen[1,12], Javier Gutierrez[1], Graham A R Heap[2], Barbara Hrdlickova[1], Sarah Hunt[11], Leticia Plaza Izurieta[13], Valentina Izzo[14], Leo A B Joosten[15,16], Cordelia Langford[11], Maria Cristina Mazzilli[17], Charles A Mein[6], Vandana Midah[18], Mitja Mitrovic[1,19], Barbara Mora[17], Marinita Morelli[14], Sarah Nutland[20], Concepción Núñez[8], Suna Onengut-Gumuscu[21], Kerra Pearce[22], Mathieu Platteel[1], Isabel Polanco[23], Simon Potter[11], Carmen Ribes-Koninckx[24], Isis Ricaño-Ponce[1], Stephen S Rich[21], Anna Rybak[25], José Luis Santiago[8], Sabyasachi Senapati[26], Ajit Sood[18], Hania Szajewska[27], Riccardo Troncone[28], Jezabel Varadé[8], Chris Wallace[20], Victorien M Wolters[29], Alexandra Zhernakova[30], Spanish Consortium on the Genetics of Coeliac Disease (CEGEC)[31], PreventCD Study Group[31], Wellcome Trust Case Control Consortium (WTCCC)[31], B K Thelma[26], Bozena Cukrowska[32], Elena Urcelay[8], Jose Ramon Bilbao[13], M Luisa Mearin[33], Donatella Barisani[34], Jeffrey C Barrett[11], Vincent Plagnol[35], Panos Deloukas[11], Cisca Wijmenga[1,37] & David A van Heel[2,37]

Using variants from the 1000 Genomes Project pilot European CEU dataset and data from additional resequencing studies, we densely genotyped 183 non-*HLA* risk loci previously associated with immune-mediated diseases in 12,041 individuals with celiac disease (cases) and 12,228 controls. We identified 13 new celiac disease risk loci reaching genome-wide significance, bringing the number of known loci (including the *HLA* locus) to 40. We found multiple independent association signals at over one-third of these loci, a finding that is attributable to a combination of common, low-frequency and rare genetic variants. Compared to previously available data such as those from HapMap3, our dense genotyping in a large sample collection provided a higher resolution of the pattern of linkage disequilibrium and suggested localization of many signals to finer scale regions. In particular, 29 of the 54 fine-mapped signals seemed to be localized to single genes and, in some instances, to gene regulatory elements. Altogether, we define the complex genetic architecture of the risk regions of and refine the risk signals for celiac disease, providing the next step toward uncovering the causal mechanisms of the disease.

Celiac disease is a common, complex and chronic immune-mediated disease with a seroprevalence of ~1% in individuals of European ancestry[1,2]. In celiac disease, a T cell–mediated small intestinal immune response is generated against gliadin fragments from wheat, rye and barley cereal proteins, leading to villous atrophy. Association of celiac disease with *HLA* variants was first shown in 1972, and predisposing *HLA-DQA1* and *HLA-DQB1* alleles are necessary but not sufficient to cause disease. Recent genome-wide association studies (GWAS) identified a further 26 non-*HLA* risk loci as being associated with celiac disease[3–6]. Many of these loci are also associated with other autoimmune or chronic immune-mediated diseases (although sometimes with different markers and directions of effect[7]), with particular overlapping of associated loci having been observed between celiac disease, type 1 diabetes[8] and rheumatoid arthritis[9].

Currently unresolved issues regarding the genetic predisposition to celiac disease, which are also relevant in other immune-mediated diseases, include explaining the remaining major fraction of heritability, including rare and additional common risk variants, and the identification of causal variants and causal genes (or at least more finely localizing the risk signal). The Immunochip Consortium[10] was developed to explore these questions by taking advantage of emerging comprehensive datasets containing common, low-frequency and rare variants and a commercial offer of much lower per-sample custom genotyping costs for a very large project comprising related diseases.

The Immunochip, a custom Illumina Infinium High-Density array, was designed to densely genotype immune-mediated disease loci identified by GWAS of common variants using data

A full list of author affiliations appears at the end of the paper.

Received 15 April; accepted 5 October; published online 6 November 2011; doi:10.1038/ng.998

Table 1 Sample collections

Population sample	Cases[a]	Controls
UK	7,728	8,274[b]
The Netherlands	1,123	1,147
Poland	505	533
Spain—CEGEC[c]	545	308
Spain—Madrid[c]	537	320
Italy—Rome, Milan and Naples	1,374	1,255
India—Punjab	229	391
Total	12,041	12,228

The collections from the UK, The Netherlands, Poland, Spain (Madrid) and Italy contained essentially the same sample set as our 2010 GWAS of celiac disease[5] but had substantial additional samples from the UK and The Netherlands and excluded amplified DNA samples from the Spanish collections. The Indian collection was not previously studied. Our 2010 GWAS contained several collections not studied here. [a]Cases are defined as individuals with Celiac disease. [b]This data includes 5,430 UK 1958 Birth Cohort participants and 2,844 UK Blood Services Common Controls. [c]We considered the two Spanish population samples separately because the samples were genotyped in different laboratories.

from the 1000 Genomes Project and any other available disease-specific resequencing data. The 1000 Genomes Project pilot CEU low-coverage whole-genome–sequencing dataset captures 95% of the variants of minor allele frequency (MAF) = 0.05, and although it is underpowered to comprehensively detect variants of rarer allele frequency, the dataset still identifies 60% of variants of MAF = 0.02 and 30% of variants of MAF = 0.01 (ref. 11). The Immunochip Consortium selected 186 distinct loci containing markers reaching genome-wide significance ($P < 5 \times 10^{-8}$) from 12 diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis). We submitted all sample variants from the 1000 Genomes Project low-coverage pilot CEU population[11] (September 2009 release) that were in 0.1-cM (HapMap3 CEU) recombination blocks around each GWAS region lead marker for array design. We did not apply any filtering on correlated variants (linkage disequilibrium (LD)). Further case and control regional resequencing data were submitted by several groups (Online Methods and **Supplementary Note**), as well as a small amount of investigator-specific undisclosed content, including GWAS results of intermediate significance.

Most GWAS were performed using common SNPs (typically with MAF > 5%) further selected for low inter-marker correlation and/or even genomic spacing. In contrast to GWAS, the Immunochip Consortium represents an opportunity to in depth and comprehensively dissect the architecture of both rare and common genetic variation at immunobiologically relevant genomic regions in human diseases. Because of the presence of the majority of the polymorphic genetic variants from the 1000 Genomes Project pilot CEU dataset (as well as additional resequencing at some loci) in our final Immunochip dataset, the true causal variants at many risk loci may have been directly genotyped and analyzed.

RESULTS

Overview of the study design

We submitted a total of 207,728 variants for Immunochip assay design, and 196,524 variants passed manufacturing quality control at Illumina. After extensive and stringent data quality control (Online Methods), we analyzed a near-complete dataset (overall, there were only 0.008% missing genotype calls) comprising 12,041 cases with celiac disease, 12,228 controls (from seven geographic regions; **Table 1**) and 139,553 polymorphic (defined here as at least two observed genotype groups) markers. We assayed 634 biallelic SNPs in duplicate; at these SNPs, we observed 189 of the 15,384,884 (0.0012%) genotype calls to be discordant. Considering the intended 207,728 variants submitted for design and an observed ~9.1% non-polymorphic rate in our data after quality control filtering, we estimated that we had high quality genotype data on ~74% of the complete set of the 1000 Genomes Project pilot CEU true polymorphic variants at the fine-mapped regions.

We observed that 36 of the 183 non-*HLA* immune-mediated disease loci selected for dense 1000-Genomes–based genotyping using the Immunochip reached genome-wide significance ($P < 5 \times 10^{-8}$) for celiac disease in either the current study or in our previous GWAS[5] (the summary association statistics for all markers are available in T1DBase (see URLs)). All variants reaching genome-wide significance were common (MAF > 5%). We also observed marked enrichment for celiac disease association signals of intermediate significance (for example, rs6691768, at the *NFIA* locus, $P = 5.3 \times 10^{-8}$) at a proportion of the remaining 147 densely genotyped non-celiac autoimmune disease regions (**Supplementary Fig. 1**). Variants from three densely genotyped regions selected on Immunochip for a non–immune-mediated trait (bipolar disorder) showed no excess of association signals (**Supplementary Fig. 1**).

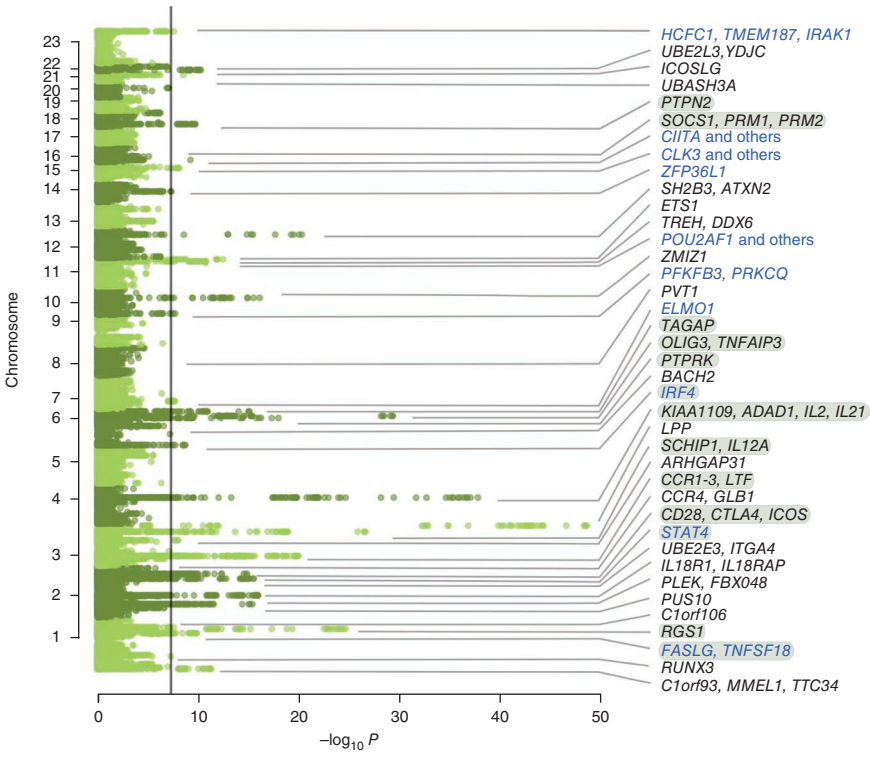

Figure 1 Manhattan plot of association statistics for previously known and newly discovered celiac disease risk loci. Newly discovered loci are indicated in blue; loci with multiple signals are shown in a gray highlighted box. The significance threshold used was $P = 5 \times 10^{-8}$.

Table 2 Risk variant signals at genome-wide significant celiac disease loci

Top variant[a]	Chr.	HapMap3 CEU LD block[b] position (*n* markers; size[c])	MAF[d]	*P*[e]	OR	The position of highly correlated variants[f] (*n* markers; size[c])	Localization relative to protein-coding genes[g]
rs4445406	1	2,396,747–2,775,531 (358; 379)	0.344	5.4×10^{-12}	0.87	2,510,162–2,710,035 (27; 200)	*C1orf93*, *MMEL1*, *TTC34*
rs72657048	1	25,111,876–25,180,863 (125; 69)	0.498	3.8×10^{-6}	0.92	25,162,321–25,177,139 (18; 15)	0–10 kb 5′ and the first exon of *RUNX3*
rs12068671	1	170,917,308–171,207,073 (355; 290)	0.185	1.4×10^{-10}	0.86	170,940,206–170,948,695 (11; 8)	35–43 kb 5′ of *FASLG*
Signal 2 rs12142280	1	"	0.180	$8.3 \times 10^{-9,e}$	0.87	171,129,607–171,131,275 (2; 2)	Intergenic region between *FASLG* and *TNFSF18*
rs1359062	1	190,728,935–190,814,664 (181; 86)	0.180	2.5×10^{-25}	0.77	190,786,488–190,811,722 (17; 25)	0–24 kb 5′ of and the first exon of *RGS1*
Signal 2 rs72734930	1	"	**0.022**	$3.7 \times 10^{-4,e}$	1.23	190,779,182 (1)	32 kb 5′ of *RGS1*
rs10800746	1	199,119,734–199,308,949 (331; 189)	0.305	2.6×10^{-8}	0.89	199,148,015 (1)	Ninth intron of *C1orf106*
rs13003464	2	60,768,233–61,745,913 (1,047; 978)	0.388	4.3×10^{-16}	1.17	61,040,333–61,058,360 (3; 18)	Exons 5–11 of *PUS10*
rs10167650	2	68,389,757–68,535,760 (357; 146)	0.266	1.3×10^{-4}	0.92	68,493,221–68,499,064 (4; 6)	Intergenic region between *PLEK* and *FBX048*
rs990171	2	102,221,730–102,573,468 (894; 352)	0.225	1.2×10^{-16}	1.20	102,338,297–102,459,513 (45; 121)	*IL18R1* and *IL18RAP*
rs1018326	2	181,502,502–181,972,196 (898; 470)	0.418	3.1×10^{-16}	1.16	181,708,291–181,803,246 (24; 95)	Intergenic region between *UBE2E3* and *ITGA4*
rs6715106	2	191,581,798–191,715,979 (203; 134)	0.058	8.4×10^{-9}	0.79	191,621,279–191,643,278 (4; 22)	Exons 6–14 of *STAT4*
Signal 2 rs6752770	2	"	0.296	$1.3 \times 10^{-6,e}$	1.10	191,681,808 (1)	Intron 3 of *STAT4*
Signal 3 rs12998748	2	"	0.119	$2.6 \times 10^{-4,e}$	0.90	191,656,882 (1)	Intron 3 of *STAT4*
rs1980422	2	204,154,625–204,524,627 (642; 370)	0.233	1.4×10^{-15}	1.19	204,318,641–204,320,303 (2; 2)	Intergenic region between *CD28* and *CTLA4*
Signal 2 rs34037980	2	"	0.217	$1.6 \times 10^{-5,e}$	0.91	204,470,572–204,478,299 (2; 8)	Intergenic region between *CTLA4* and *ICOS*
Signal 3 rs10207814	2	"	**0.039**	$1.3 \times 10^{-4,e}$	1.20	204,158,521–204,168,206 (5; 10)	111–121 kb 5′ of *CD28*
rs4678523	3	32,895,606–33,063,377 (260, 168 kb)	0.313	2.4×10^{-7}	1.11	33,012,725–33,012,756 (2; 31)	Intergenic region between *CCR4* and *GLB1*
rs2097282	3	45904804–46625997 (1,343; 721)	0.314	1.1×10^{-20}	1.20	46,321,275–46,377,631 (27; 56)	Intergenic region between *CCR3* and *CCR2*
Signal 2 rs7616215	3	"	0.361	$8.6 \times 10^{-9,e}$	1.12	46,162,711–46,180,690 (2; 18)	38–55 kb 3′ of *CCR1*
Signal 3 rs60215663	3	"	0.070	$4.8 \times 10^{-5,e}$	1.16	46,458,634–46,480,319 (7; 22)	Exons 2–13 of *LTF* (NM_002343.3)
rs61579022	3	120,587,671–120,783,345 (372; 196)	0.390	9.9×10^{-9}	1.11	120,601,187–120,605,968 (4; 5)	Intron 10 of *ARHGAP31*
[imm_3_ 161120372]	3	161,065,075–161,237,201 (423; 168)	0.111	2.6×10^{-27}	1.36	161,112,778–161,147,744 (4; 35)	Intergenic region between *SCHIP1* and *IL12A*
Signal 2 rs1353248	3	"	0.288	$9.8 \times 10^{-9,e}$	0.88	161,106,253 (1)	Intergenic region between *SCHIP1* and *IL12A*
Signal 3 rs2561288	3	"	0.455	$8.1 \times 10^{-8,e}$	1.12	161,136,316–161,168,494 (6; 32)	Intergenic region between *SCHIP1* and *IL12A*
rs2030519	3	189,552,054–189,622,323 (142; 70)	0.486	3.0×10^{-49}	0.76	189,587,750–189,602,595 (8; 15)	Intron 2 of *LPP*
rs13132308	4	123,192,512–123,784,752 (1,294; 592)	0.166	1.9×10^{-38}	0.71	123,269,042–123,770,564 (11; 502)	Multiple genes (*KIAA1109*, *ADAD1*, *IL2* and *IL21*)
Signal 2 rs62323881	4	"	0.073	$8.6 \times 10^{-5,e}$	1.15	123,257,527–123,722,990 (87; 465)	Multiple genes (*KIAA1109*, *ADAD1*, *IL2* and *IL21*)
rs1050976	6	315,547–402,748 (199; 87)	0.488	1.8×10^{-9}	0.89	353,079–355,417 (3; 2)	3′ UTR of *IRF4* (NM_002460.3)
Signal 2 rs12203592	6	"	0.183	$2.6 \times 10^{-4,e}$	0.91	341,321 (1)	Intron 4 of *IRF4* (NM_002460.3)
rs7753008	6	90,863,556–91,096,529 (341; 233)	0.380	2.7×10^{-7}	1.10	90,866,360–90,875,874 (5; 10)	Intron 2 of *BACH2* (NM_001170794.1)
rs55743914	6	127,993,875–128,382,483 (572; 389)	0.239	1.1×10^{-18}	1.21	128,332,892–128,335,255 (2; 2)	The last exon of *PTPRK* in the 3′ UTR (NM_002844.3)

(continued)

Table 2 Risk variant signals at genome-wide significant celiac disease loci (continued)

Top variant[a]	Chr.	HapMap3 CEU LD block[b] position (*n* markers; size[c])	MAF[d]	*P*[e]	OR	The position of highly correlated variants[f] (*n* markers; size[c])	Localization relative to protein-coding genes[g]
Signal 2 rs72975916	6	"	0.150	$1.2 \times 10^{-5,e}$	0.89	128,307,943–128,339,304 (15; 31)	*PTPRK* exons 28–30 in the 3′ UTR to 24 kb 3′
rs17264332	6	137,924,568–138,316,778 (864; 392)	0.211	5.0×10^{-30}	1.29	138,000,928–138,048,197 (6; 47)	Intergenic region between *OLIG3* and *TNFAIP3*
Signal 2 [imm_6_ 138043754]	6	"	0.190	$2.1 \times 10^{-7,e}$	0.88	138,015,797–138,043,754 (4; 28)	Intergenic between *OLIG3* and *TNFAIP3*
rs182429	6	159,242,314–159,461,818 (514; 220)	0.427	8.5×10^{-16}	1.16	159,385,965–159,390,046 (4; 4)	4 kb 5′ and 5′ UTR of *TAGAP* (NM_152133.1)
Signal 2 rs1107943	6	"	0.071	$2.8 \times 10^{-6,e}$	1.18	159,418,255 (1)	32 kb 5′ of *TAGAP* (NM_152133.1)
[1kg_7_ 37384979]	7	37,330,503–37,406,978 (213; 76)	0.101	2.1×10^{-8}	1.18	37,366,994–37,404,402 (31; 37)	Intron 1 of *ELMO1*
rs10808568	8	129,211,716–129,368,419 (400;157)	0.256	2.2×10^{-5}	0.91	129,333,242–129,345,888 (4; 13)	151–163 kb 3′ of *PVT1*
rs2387397	10	6,428,077–6,585,110 (411; 157)	0.229	1.9×10^{-8}	0.88	6,430,198 (1)	Intergenic region between *PFKFB3* and *PRKCQ*
rs1250552	10	80,690,408–80,774,414 (223; 84)	0.470	8.0×10^{-17}	0.86	80,728,033 (1)	Intron 14 of *ZMIZ1*
rs7104791	11	110,682,429–110,815,769 (3; 133)	0.209	1.9×10^{-11}	1.16	Not high-density genotyped	[region: *POU2AF1*, *C11orf93*]
rs10892258	11	117,847,131–118,270,810 (466; 424)	0.237	1.7×10^{-11}	0.86	118,080,536–118,085,075 (5; 5)	Intergenic region between *TREH* and *DDX6*
rs61907765	11	127,754,640–127,985,723 (480; 231)	0.213	3.4×10^{-13}	1.18	127,886,184–127,901,948 (6; 16)	5 kb 5′ and the first exon of *ETS1* (NM_001162422.1)
rs3184504	12	110,183,529–111,514,870 (938; 1,331)	0.488	5.4×10^{-21}	1.19	110,368,991–110,492,139 (4; 123)	5′ UTR and exons 1–3 of *SH2B3*; exons 2–25 and the 3′ UTR of *ATXN2*
rs11851414	14	68,238,574–68,387,815 (338; 149)	0.221	4.7×10^{-8}	1.13	68,329,159–68,341,722 (3; 13)	1 kb 5′ of and the first exon of *ZFP36L1*
rs1378938	15	72,397,784–73,270,664 (23; 873)	0.278	7.8×10^{-9}	1.13	Not high-density genotyped	[region including *CLK3*, *CSK* and multiple genes]
rs6498114	16	10,834,038–10,903,351 (8; 69)	0.246	5.8×10^{-10}	1.14	Not high-density genotyped	[region: *CIITA*]
rs243323	16	11,220,552–11,385,420 (446; 165)	0.300	2.5×10^{-5}	0.92	11,254,549–11,268,703 (12; 14)	11 kb 5′ of, 1 kb 3′ of and all of *SOCS1*
Signal 2 [imm_16_ 11281298]	16	"	**0.004**	$1.3 \times 10^{-4,e}$	1.70	11,281,298 (1)	Intergenic region between *PRM1* and *PRM2*
Signal 3 rs9673543	16	"	0.169	$2.0 \times 10^{-4,e}$	1.10	11,292,457 (1)	10 kb 5′ of *PRM1*
rs11875687	18	12,728,413–12,914,117 (411; 186)	0.150	1.9×10^{-10}	1.17	12,811,903–12,870,206 (16; 58)	Exons 2–5 of *PTPN2* (NM_ 080422.1)
Signal 2 rs62097857	18	"	**0.040**	$5.2 \times 10^{-5,e}$	1.20	12,847,758 (1)	Intron 2 of *PTPN2* (NM_080422.1)
rs1893592	21	42,683,153–42,760,214 (226; 77)	0.282	3.0×10^{-9}	0.88	42,728,136 (1)	Intron 9 of *UBASH3A* (NM_018961)
rs58911644	21	44,414,408–44,528,088 (239; 114)	0.193	6.2×10^{-7}	0.89	44,446,245–44,453,549 (8; 7)	18–25 kb 3′ of *ICOSLG*
rs4821124	22	20,042,414–20,352,005 (131; 310)	0.186	5.7×10^{-11}	1.16	20,250,903–20,313,260 (36; 62)	*UBE2L3*, *YDJC*
rs13397	X	152,825,373–153,043,675 (88; 218)	0.133	2.7×10^{-8}	1.18	152,872,114–152,937,386 (4; 65)	*HCFC1*, *TMEM187*, *IRAK1*

Non-*HLA* loci meeting genome-wide significance ($P < 5 \times 10^{-8}$) in the current Immunochip data set and in the previous GWAS and replication data set[5] are shown. Loci reported for the first time for celiac disease at genome-wide significance are shown in bold in the 'top variant' column.

[a]dbSNP130 ID. [b]Regions were first defined by LD blocks extending 0.1 cM to the left and right of the risk SNP, as defined by the HapMap3 CEU recombination map. For loci with multiple different previously reported risk SNPs for different diseases and overlapping blocks, the extended region is shown. All chromosomal positions are based on NCBI build 36 (hg18) coordinates. [c]Size in kb. [d]MAFs are shown for the European controls. See **Supplementary Table 4** for more detailed allele frequencies in the cases and controls according to collection. Low-frequency and rare variants are shown in bold. [e]According to a logistic regression association test. The tests for second (and third) independent signals are conditioned on the first (and second) reported variant(s). The per-locus significance thresholds for the second (and third) independent signals are shown in **Supplementary Table 3**. [f]Highly correlated variants are defined as $r^2 > 0.9$, according to hg18. [g]RefSeq track UCSC/hg18. Only the most significantly associated risk variant from each region and independent signal is shown. Variant names are shown as they are listed in dbSNP130 where available, and otherwise, the Illumina Immunochip manifest name is shown in brackets (**Supplementary Table 5** shows both names for the variants). Chr., chromosome.

We identified 13 new celiac risk loci ($P < 5 \times 10^{-8}$; **Fig. 1**, **Table 2** and **Supplementary Fig. 2**), 10 of which were immune-mediated disease loci selected for dense 1000-Genomes–based genotyping on the Immunochip. Several of these new loci were reported at lesser significance levels in our previous studies[5,9], and almost all of these loci have been reported in at least one other immune-mediated disease. These new loci, along with the *HLA* loci, bring the total number of reported (in the current and a previous study[5], which had an

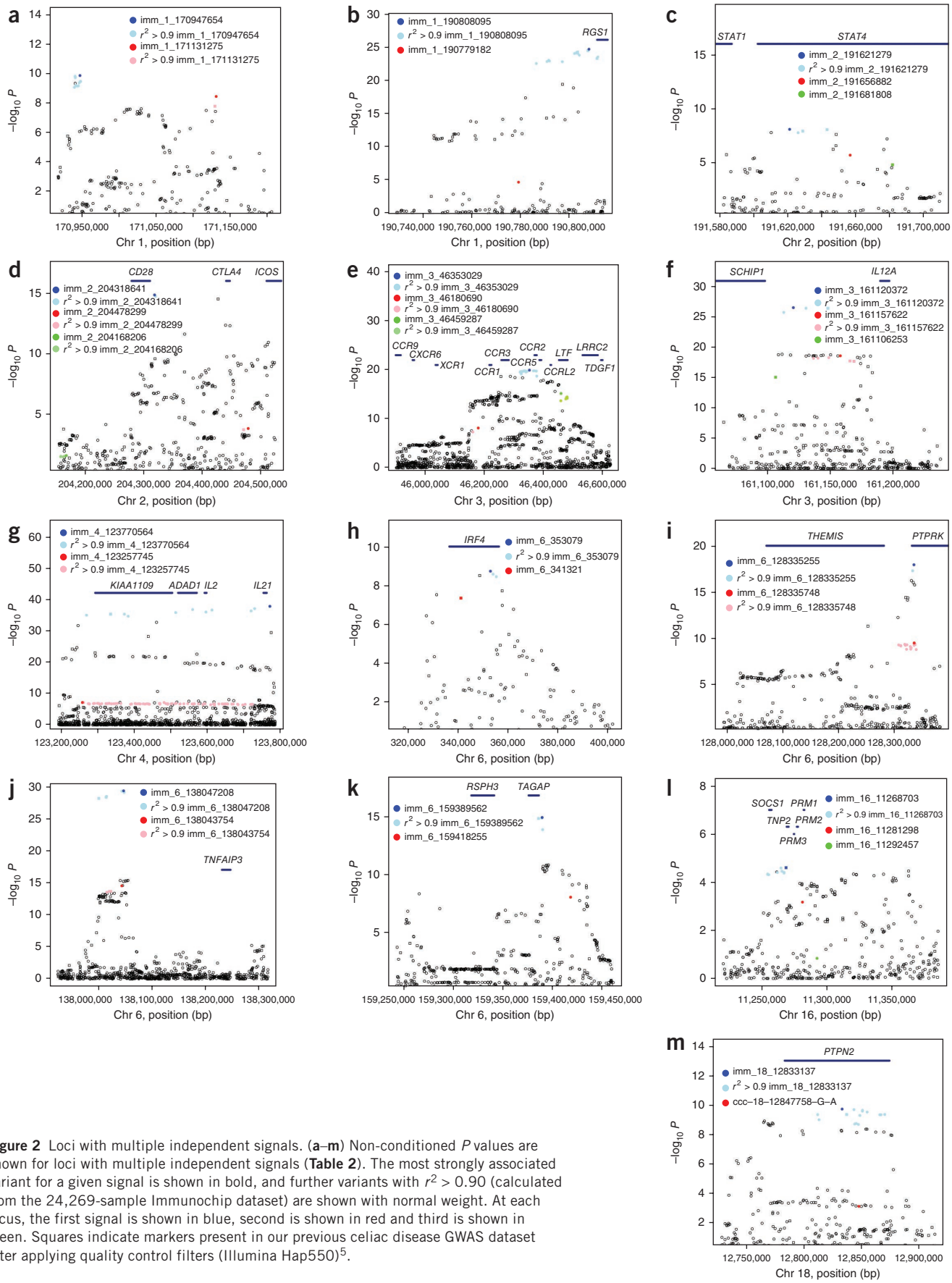

Figure 2 Loci with multiple independent signals. (**a**–**m**) Non-conditioned *P* values are shown for loci with multiple independent signals (**Table 2**). The most strongly associated variant for a given signal is shown in bold, and further variants with $r^2 > 0.90$ (calculated from the 24,269-sample Immunochip dataset) are shown with normal weight. At each locus, the first signal is shown in blue, second is shown in red and third is shown in green. Squares indicate markers present in our previous celiac disease GWAS dataset after applying quality control filters (Illumina Hap550)[5].

overlapping but slightly different sample set) genome-wide significant celiac disease loci to 40. Most of these loci contain candidate genes of immunological function, which is consistent with our previous findings at celiac disease loci[3–5].

The median of the effect sizes (odds ratios (ORs) and inverting protective effects) for the most significant marker per locus was 1.155 (range 1.124–1.360) for the top signals from 26 non-*HLA* loci measured using Illumina Hap300 and Hap550 LD-pruned tag SNPs in our 2010 celiac disease GWAS[5] and was 1.166 (range 1.087–1.408) for the corresponding most significant marker (for the same signal) per locus in the current high-density fine-mapping Immunochip data set (Wilcoxon test $P = 0.75$; **Supplementary Table 1**). Although we observed no difference in the effect sizes between the GWAS lead SNPs and the subsequent fine-mapped signals, we note that the resequencing of the cases in the current Immunochip dataset is limited (see the Discussion section).

In all, we report 57 independent celiac disease association signals (**Table 2**) from 39 separate loci, of which 18 (32%) were not efficiently ($r^2 > 0.9$; **Supplementary Table 2**) tagged by our previous GWAS[5] (Illumina Hap550 dataset after quality control filtering) markers.

Multiple independent common and rare variant signals

In contrast to most GWAS chips, the Immunochip contains a substantial proportion of polymorphic variants of low MAF. Of 139,553 variants in our 11,837 controls of European ancestry, 24,661 variants are low frequency (defined[11] as MAF = 5–0.5%) and a further 22,941 variants are rare (MAF < 0.5%). We investigated the possibility of the existence of multiple independently associated variants (of any allele frequency) at each locus using stepwise logistic regression conditioning on the most significant variant at the locus (Online Methods and **Supplementary Table 3**). This analysis is sensitive to genotype miscalling and missing data[12], hence our use of extremely rigorous quality control measures for the dataset and manual inspection of genotype clusters for all reported markers.

We observed two or more independent signals at 13 of the 36 high-density genotyped non-*HLA* loci (**Fig. 2**). Four of these loci each had three independent signals (*STAT4*, the chromosome 3 *CCR* region, *IL12A* and *SOCS1-PRM1-PRM2*; **Table 2**). We observed low frequency and/or rare variant signals at four separate loci (*RGS1*, *CD28-CTLA4-ICOS*, *SOCS1-PRM1-PRM2* and *PTPN2*). Notably, we saw the strongest effect (OR = 1.70) at the rare variant imm_16_11281298 (at the *SOCS1-PRM1-PRM2* locus) with genotype counts (AA/AG/GG) of 1/136/11,904 (MAF 0.57%) in all cases with celiac disease and 0/91/12,136 (MAF 0.37%) in all controls (the detailed genotype count and allele frequency data for the top signals by collection are shown in **Supplementary Table 4**).

We next performed haplotype analysis on all loci with multiple independent signals to investigate whether the multiple signals were a result of multiple causal effects or a single effect best tagged by several variants. For all but one locus (*PTPN2*), the haplotype association test results (data not shown) were of similar significance to those from the single SNP association tests, suggesting that for each signal, we genotyped either the causal variant or markers very strongly correlated with it. These findings contrast with those from a recent resequencing study[13], probably because of the much greater variant density of our study. However, at the *PTPN2* locus, the imm_18_12833137(T) + ccc-18-12847758-G-A(G) haplotype was considerably more strongly associated with disease ($P = 4.8 \times 10^{-14}$, OR = 0.84) than either SNP alone (imm_18_12833137, $P = 1.9 \times 10^{-10}$ and ccc-18-12847758-G-A, $P = 0.0008$).

At the *SOCS1* locus, the third independent signal, imm_16_11292457, showed association after conditioning on the two other signals ($P = 2.0 \times 10^{-4}$) but not in the single-SNP non-conditioned association analysis ($P = 0.15$). Further inspection identified the protective imm_16_11292457(A) allele to be correlated (in LD) with the risk (A) allele of the first signal, imm_16_11268703; thus, although there are indeed three independent signals, the effect of the third signal is only seen after conditioning on the first. A similar statistical effect (Simpson's paradox) was recently shown at a Parkinson's disease locus[14].

Fine mapping to localize causal signals

GWAS signals are typically reported within relatively large LD blocks. We tested whether our much denser genotyping strategy would allow finer-scale localization and the pinpointing of association signals. We found that markers strongly correlated ($r^2 > 0.9$) with the most significant independent variant clustered together and defined regions that are a median of 12.5 times smaller than the relevant HapMap3 CEU 0.1-cM LD blocks (**Table 2**, **Fig. 2** and **Supplementary Fig. 2**). Localization was highly successful for some regions (for example, *PTPRK* and *TAGAP*) but was not possible at others (for example, *IL2-IL21*). At many loci, the localized regions comprised only a handful of markers in close physical proximity to each other.

Considering the 36 loci genotyped at high density, we localized 29 of the total 54 independent non-*HLA* signals to a single gene (**Table 2** and **Supplementary Fig. 2**). We identified all markers strongly correlated ($r^2 > 0.9$) with the independent non-*HLA* variants reported in our analyses (**Table 2**), and using functional annotation (**Supplementary Table 2**), identified only a handful of markers in exonic regions, of which three are protein-altering variants (the non-synonymous SNPs imm_1_2516606 (*MMEL1*), imm_12_110368991 (*SH2B3*) and 1kg_X_152937386 (*IRAK1*)). In contrast, a number of signals appeared to be more finely localized around the transcription start site of specific genes (which we defined as the first exon and 10 kb 5′ of the first exon), including signals at *RUNX3*, *RGS1*, *ETS1*, *TAGAP* and *ZFP36L1*, and around the 3′ untranslated region (UTR) (and 10 kb 3′), including signals at *IRF4*, *PTPRK* and *ICOSLG*.

We saw overlap between multiple independent signal regions at some loci (**Fig. 2**), suggesting that causal variants might be functioning through a shared mechanism, for example, within a 2-kb region of the *PTPRK* 3′ UTR, within an 11-kb region 5′ of *IL12A* or within a 28-kb region of *TNFAIP3*. In contrast, we observed multiple independent signals that spread across the three immune genes of the *CD28-CTLA4-ICOS* region.

DISCUSSION

We show that fine mapping of GWAS regions using dense resequencing data, for example, from the 1000 Genomes Project (as we used here), is feasible and generates substantial additional information at many loci. We identify a complex architecture of multiple common and rare genetic risk variants for around one-third of the now 40 confirmed celiac disease loci. The design of our study allowed us to find many more complex regions than the ~10% with multiple signals seen in our previous study[5] and a recent large GWAS for human height[15]. It seems probable that if larger sample sizes than those used in the current study were to be tested, additional loci might be shown to have a similarly rich architecture with multiple risk variants. Multiple independent risk signals for celiac disease have also long been known to exist in the *HLA* region[16]. Our success in identifying multiple risk signals in celiac disease might be partly a result of the extensive selective pressures for haplotypic diversity that have taken place at immune gene loci[17]. Previous studies reported independently associated common and rare variants at individual loci for a

handful of phenotypes, for example, fetal hemoglobin[13], sick sinus syndrome[18], Crohn's disease[19] and hypertriglyceridemia[20]. To the best of our knowledge, this is the first study to have comprehensively surveyed the genetic architecture of all known risk loci for a trait.

In part, our identification of rare variants at risk regions relies on the prior discovery of a genome-wide significant common variant association signal at each locus. This then permits a per-locus correction rather than a genome-wide multiple-testing correction when searching for additional independent association signals. Only particularly strong rare variant signals would, on their own, generate significance levels reaching the genome-wide threshold typically used in GWAS ($P < 5 \times 10^{-8}$). Alternative methods, such as collapsing rare variant signals across a gene or functional categories of genes, have therefore been suggested as approaches to this problem[21]. Although a rare variant may occur on a recent haplotypic background and thus show LD at a substantially longer range than common variants, we deliberately restricted our search to around the common-variant LD blocks because to do otherwise would have incurred a considerably greater penalty from multiple testing. Therefore, although our study provides considerable support for exome and whole-genome sequencing efforts aimed at identifying rare risk variants (and those not necessarily restricted to GWAS loci) in common complex diseases, it further highlights the statistical challenges of establishing associations for rare variants.

We used a dense genotyping strategy and a stepwise conditional association analysis but did not identify any rare highly penetrant variants that might explain the genome-wide significant common SNP signals at any of the 39 loci. Our study does have limitations in this regard, particularly: (i) the restriction of the analysis to 0.1-cM LD blocks; (ii) the limited control resequencing sample size of the 1000 Genomes Project pilot CEU dataset; (iii) the limited case resequencing sample size; and (iv) case resequencing being limited to three loci for celiac disease and to selected loci for other immune diseases. We observed a weak trend toward a lower MAF ($P = 0.042$, Wilcoxon test; **Supplementary Table 1**) for the best fine-mapping SNP from the Immunochip analysis compared to the lead SNP from our 2010 tag SNP GWAS (determined by measuring the MAF in a subset of samples genotyped in both datasets). One signal showed a substantially higher MAF (>25% change) using fine mapping and four signals showed a substantially lower MAF using fine mapping (**Supplementary Table 1**), however, all fine-mapping variants corresponding to the lead GWAS SNPs remained common (MAF > 0.10). We suggest that these changes in the MAFs of the lead GWAS SNPs using fine mapping simply reflect a more precise measurement of common frequency risk haplotypes. Although we cannot exclude the possibility that a single high-penetrance lower-frequency variant explains most of the association signal at a locus, especially without more comprehensive resequencing of the cases, we found no evidence to support this possibility in the current fine-mapping analysis. Similarly, although our stepwise selection procedure cannot robustly refute the 'synthetic association' hypothesis—in particular, that a combination of multiple rare variants jointly explains the association signal[22]—we have so far not observed any evidence supporting this possibility.

We identified 13 new loci for celiac disease at genome-wide significance, most of which have been reported previously at lesser significance levels or in another immune-mediated disease. The Illumina Hap550 chip (used in our 2010 GWAS) would have detected 10 of the 13 new loci and, in total, 39 of the 57 independent non-*HLA* signals that we report here. A current genotyping platform, the Illumina Omni2.5 chip, would have detected 12 of the 13 new loci and, in total, 50 of the 57 independent non-*HLA* signals that we report here. However, neither of these chips would have provided the finer-scale localization of the Immunochip. The 13 new loci contain many candidate genes with an immunological function ($P = 0.0002$ for enrichment of the Gene Ontology term 'immune system process'[23]), which is in line with expectations based on our previous studies. We also found evidence suggesting that substantial additional signals exist at other immune-mediated disease loci that are below the genome-wide significance threshold applied to the current dataset. It is a point of debate whether such strict ($P < 5 \times 10^{-8}$) criteria should apply; for example, an analyst might apply a higher Bayesian prior at a locus already reported in another immune-mediated disease. Alternatively, an Immunochip-wide *P* value with a Bonferroni correction for independent SNPs, as was used recently in the Cardiochip custom genotyping project[24], of $P < 1.9 \times 10^{-6}$ (Online Methods) would yield 16 new celiac disease loci in addition to the 13 we identified here. These 16 loci also mostly contain immune system genes. An analysis of these signals of intermediate significance would gain substantial additional power in a meta-analysis across the several hundred thousand samples from multiple immune-mediated disease collections currently being run on the Immunochip.

We found that our previous GWAS using tag SNPs gave very similar estimates of effect size as our current fine-mapping experiment (**Supplementary Table 1**), which is in contrast to a simulation study that suggested that GWAS markers often underestimate risk[14]. However, we found substantial evidence for multiple additional signals at known loci and report many new loci. In individuals of European ancestry, the 39 non-*HLA* loci explain 13.7% of the genetic variance of celiac disease (*HLA* variants account for a further ~40%). We also show a long list of effects of weaker significance, which will explain substantial additional heritability.

Only one of the variants reported here was discovered in a disease-specific resequencing study: ccc-18-12847758-G-A (rs62097857), a marker identified by the WTCCC's resequencing of cases with Crohn's disease and controls (**Supplementary Note**) and that is also present in the Watson genome. We submitted for Immunochip analysis ~4,000 variants from high-throughput resequencing of pools of 80 cases with celiac disease for extended genomic regions at three loci (*RGS1*, *IL12A* and *IL2-IL21*; **Supplementary Note**). These loci did not contribute any signals in addition to those obtained from the 1000 Genomes Project pilot CEU variants, although they did increase the number of variants correlated with each signal (that is, the set of markers that probably contains the causal variant(s)) and more precisely define the boundaries of the signal localization. We note that larger-scale resequencing of cases (for example, using many hundreds of samples) would identify a spectrum containing more rare variants than the current study, and this method has previously been used with success at selected genes and phenotypes.

The possibility of performing fine-scale mapping of GWAS regions using, for example, 1000 Genomes Project data, has been discussed as a natural follow-up strategy for such studies[25,26] and has been used recently to identify risk variants in *APOL1* in African-Americans with renal disease[27]. To our knowledge, our current report is the first to test such a strategy on a large scale in a complex disease. At multiple regions, we were able to refine the signal to a handful of variants over a few kb or tens of kb, although some regions (for example, *IL2-IL21*) were resistant to this approach, presumably because of the presence of particularly strong LD. Most GWAS report signals mapping to an LD block based on HapMap recombination rates (with a sample size of 60 families from the CEU dataset). In our data, where we have both much denser genotyping than GWAS chips (with a mean of 13.6× the genotyping density at celiac loci compared to the Illumina Hap550 chip)

and nearly 25,000 genotyped samples for the LD calculations, we are able to observe much finer-scale recombination and more precisely estimate the boundaries of no or minimal recombination intervals. Our findings are similar in terms of genotyping density and the resulting fine-mapped region size and lack of haplotype-specific effects to an earlier study of the *IL2RA* locus in type 1 diabetes[26]. At the majority of regions, we saw a tight block of highly correlated variants rather than a gradual decay of correlation (for example, see the plots for *IL12A* and *PTPRK* in **Fig. 2**). At many loci, we defined a handful of likely candidates as the causal variant(s) to be taken forward to functional studies, although we may have missed candidate variants at some regions as a result of the sample size of the 1000 Genomes Project pilot CEU dataset (60 individuals), the status of the individuals in this dataset as controls and our estimate that ~25% of these variants were excluded from our final dataset. These variants could be assessed by imputation methods[28], but our approach, particularly in regard to the more sensitive conditional regression analysis, has been to prefer the more accurate direct genotyping of all assayable variants. As much larger reference datasets based on whole-genome resequencing become available (for example, from the 1000 Genomes Project), these datasets could be imputed into our Immunochip dataset, including variants with substantially lower frequency[29]. We also investigated whether our use of multiple ethnic subgroups within Europe (for example, Southern European Spanish compared to Northern European UK populations) or the relatively small Indian collection we used contributed to fine mapping and found that, in most instances, the same degree of localization was possible with just the UK collection alone (data not shown).

Our data suggest that most common risk variants function by influencing regulatory regions, which is consistent with variants previously reported in other immune-mediated diseases and in complex traits in general[11]. The exception is the *SH2B3* non-synonymous SNP imm_12_110368991 (rs3184504) reported in our 2008 celiac GWAS[4], which, even with fine mapping of 938 polymorphic variants from the *SH2B3* region, remains the strongest signal at this locus, suggesting it may be the causal variant. The same variant has been associated with other immune diseases and a functional immune phenotype[5]. Notably, we observed a common ~980-bp intergenic deletion between *IL2* and *IL21* (DGV40686, accurately genotyped by Infinium assay with a control MAF = 7.3%) that correlated with the second independent signal at this region, although we have no evidence to suggest causality at this location.

Our fine-scale localization approach identified probable causal genes at multiple loci and at eight genes signals localized around the 5′ or 3′ regulatory regions. For example, at the *THEMIS-PTPRK* locus, two independently associated sets of variants cluster in the 3′ UTR of *PTPRK* (one, imm_6_128332892 (rs3190930), is located in a predicted binding site for the microRNA hsa–miR-1910). *PTPRK*, a TGF-β target gene, is involved in $CD4^+$ T cell development, and a deletion mutation in *PTPRK* causes T helper cell deficiency in the LEC rat strain[30]. The signal at *TAGAP* is within a 4-kb region immediately 5′ of the transcription start site and presumably contains promoter elements. At *ETS1*, the signal comprises six variants overlapping the promoter and first exon of the T cell expressed isoform NM_001162422.1, and one of these variants (imm_11_127897147 (rs61907765)) has predicted regulatory potential and overlaps multiple transcription factor binding sites (UCSC GenomeBrowser ChipSeq and ESPERR tracks (see URLs); **Supplementary Table 2**). We observed similarly notable variants in regulatory regions of *RUNX3* (imm_1_25165788 (rs11249212)) and *RGS1* (imm_1_190807644 (rs1313292) and imm_1_190811418 (rs2984920)) (**Supplementary Table 2**). A similar approach to identify the functional potential of risk variants was recently successfully used to define a causal variant in *TNFAIP3* for systemic lupus erythematosus[31]. Although we localized signals at many loci, and although recent research suggests the causal gene is often located near the most strongly associated variant[15], only more detailed functional studies (for example, transcription factor binding assays[31] and transcriptional activity assays of constructs with individual single nucleotide alterations at risk SNPs[32]) will show precisely which gene variants might be causal.

We conclude that dense fine mapping of regions identified through GWAS can uncover a complex genetic architecture of independent common and rare variants and can often successfully localize risk variant signals to a small set of SNPs to be taken forward to functional assays. Denser fine-mapping studies using larger resequencing sample sizes from both cases and controls over broader regions might provide further resolution of GWAS signals.

URLs. Database of Genomic Variants, http://projects.tcag.ca/variation/?source=hg18; T1Dbase, http://www.t1dbase.org; UCSC Genome Browser, http://genome.ucsc.edu/; ESPERR, http://www.bx.psu.edu/files/projects/esperr/; SIFT, http://sift.jcvi.org/; BioGPS, biogps.gnf.org; PreventCD consortium, www.preventceliacdisease.com; Wellcome Trust Case Control Consortium, http://www.wtccc.org.uk/; European Genome-Phenome Archive, http://www.ebi.ac.uk/ega/; R, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank Coeliac UK for assistance with direct recruitment of individuals with celiac disease and the clinicians from the UK (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis and K. Moriarty) who recruited individuals with celiac disease to provide blood samples as described in our previous studies. We thank the Dutch clinicians for recruiting individuals with celiac disease to provide blood samples as described in our previous studies (C.J. Mulder, G.J. Tack, W.H.M. Verbeek, R.H.J. Houwen and J.J. Schweizer). We thank the genotyping facility of the University Medical Center Groningen (UMCG) (P. van der Vlies) for help in generating some of the Immunochip data and S. Jankipersadsing and A. Maatman at the UMCG for preparation of the samples. We thank R. Scott for preparing samples for genotyping and the staff at the University of Pittsburgh Genomics and Proteomics Core Laboratories for performing the genotyping. We thank C. Wallace for assistance with Immunochip SNP selection and J. Stone for coordinating the Immunochip design and production at Illumina. We thank the members of each disease consortium who initiated and sustained the cross-disease Immunochip project. We thank all individuals with celiac disease and all controls for participating in this study.

Funding was provided by the Wellcome Trust (084743 to D.A.v.H.), by grants from the Celiac Disease Consortium and an Innovative Cluster approved by the Netherlands Genomics Initiative. Partial funding was provided by the Dutch Government (BSIK03009 to C. Wijmenga) and the Netherlands Organisation for Scientific Research (NWO, grant 918.66.620 to C. Wijmenga). Funding was also provided by the US National Institutes of Health grant 1R01CA141743 (to R.H.D.) and Fondo de Investigación Sanitaria grants FIS08/1676 and FIS07/0353 (to E.U.). This research utilized resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, the National Human Genome Research Institute, the National Institute of Child Health and Human Development and the Juvenile Diabetes Research Foundation International and is supported by the US National Institutes of Health grant U01-DK062418. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS-CC collection), which is funded by the Wellcome Trust grant 076113/C/04/Z and by US National Institute for Health Research program grant to the National Health Service Blood and Transplant

(RP-PG-0310-1002). The collection was established as part of the WTCCC[33]. We acknowledge the use of DNA from the British 1958 Birth Cohort collection, which is funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. S.S. is supported by a Senior Research Fellowship from the Council for Scientific and Industrial Research (CSIR), New Dehli, India.

AUTHOR CONTRIBUTIONS

D.A.v.H. and C. Wijmenga led the study. D.A.v.H., K.A.H., G.T. and C. Wijmenga wrote the paper. K.A.H., G.T., V.Mistry, N.A.B., J.R., M.P., M.Mitrovic, R.H.D. and K.F. performed DNA sample preparation and genotyping assays. D.A.v.H., V.P., K.A.H. and G.T. performed the statistical analysis. All other authors contributed primarily to the sample collection and phenotyping. P.D. led the formation of the Immunochip Consortium, and SNP selection was performed by J.C.B. and C. Wallace. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bingley, P.J. *et al.* Undiagnosed coeliac disease at age seven: population based prospective birth cohort study. *Br. Med. J.* **328**, 322–323 (2004).
2. West, J. *et al.* Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. *Gut* **52**, 960–965 (2003).
3. van Heel, D.A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*. *Nat. Genet.* **39**, 827–829 (2007).
4. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008).
5. Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
6. Trynka, G. *et al.* Coeliac disease-associated risk variants in *TNFAIP3* and *REL* implicate altered NF-κB signalling. *Gut* **58**, 1078–1083 (2009).
7. Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10**, 43–55 (2009).
8. Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).
9. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
10. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
11. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
12. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
13. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
14. Spencer, C., Hechter, E., Vukcevic, D. & Donnelly, P. Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genet.* **7**, e1001337 (2011).
15. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
16. van Heel, D.A., Hunt, K., Greco, L. & Wijmenga, C. Genetics in coeliac disease. *Best Pract. Res. Clin. Gastroenterol.* **19**, 323–339 (2005).
17. Zhernakova, A. *et al.* Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* **86**, 970–977 (2010).
18. Holm, H. *et al.* A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
19. Lesage, S. *et al. CARD15/NOD2* mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* **70**, 845–857 (2002).
20. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).
21. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44**, 293–308 (2010).
22. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
23. Zheng, Q. & Wang, X.J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* **36**, W358–W363 (2008).
24. Lanktree, M.B. *et al.* Meta-analysis of dense gene-centric association studies reveals common and uncommon variants associated with height. *Am. J. Hum. Genet.* **88**, 6–18 (2011).
25. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728–731 (2008).
26. Lowe, C.E. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the *IL2RA* region in type 1 diabetes. *Nat. Genet.* **39**, 1074–1082 (2007).
27. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
28. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
29. Jostins, L., Morley, K.I. & Barrett, J.C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* **19**, 662–666 (2011).
30. Asano, A., Tsubomatsu, K., Jung, C.G., Sasaki, N. & Agui, T. A deletion mutation of the protein tyrosine phosphatase kappa (*Ptprk*) gene is responsible for T-helper immunodeficiency (thid) in the LEC rat. *Mamm. Genome* **18**, 779–786 (2007).
31. Adrianto, I. *et al.* Association of a functional variant downstream of *TNFAIP3* with systemic lupus erythematosus. *Nat. Genet.* **43**, 253–258 (2011).
32. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
33. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[1]Genetics Department, University Medical Center and University of Groningen, Groningen, The Netherlands. [2]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [3]Department of Gastroenterology, Vrije Universiteit (VU) Medical Center, Amsterdam, The Netherlands. [4]Fondazione Istituto Di Ricovero e Cura a Carattere Scientifico (IRCCS) Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy. [5]Department of Medical Sciences, University of Milan, Milan, Italy. [6]Genome Centre, Barts and the London School of Medicine and Dentistry, John Vane Science Centre, Charterhouse Square, London, UK. [7]Universitat Rovira I Virgili, Department of Paediatric Gastroenterology, Hospital Univesitari de Sant Joan de Reus, Reus, Spain. [8]Immunology Department, Hospital Clínico S. Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Madrid, Spain. [9]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [10]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA. [11]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [12]Department of Gastroenterology, University Medical Center and Groningen University, Groningen, The Netherlands. [13]Immunogenetics Research Laboratory, Hospital de Cruces, Barakaldo, Bizkaia, Spain. [14]European Laboratory for Food Induced Disease, University of Naples Federico II, Naples, Italy. [15]Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [16]Nijmegen Institute for Infection, Inflammation and Immunity (N4i), Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [17]Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy. [18]Dayanand Medical College and Hospital, Ludhiana, Punjab, India. [19]University of Maribor, Faculty of Medicine, Center for Human Molecular Genetics and Pharmacogenomics, Maribor, Slovenia. [20]Juvenile Diabetes Research Foundation, Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [21]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA. [22]University College London Genomics, Institute of Child Health, University College London, London, UK. [23]Pediatrics Gastroenterology Department, Hospital La Paz, Madrid, Spain. [24]Pediatric Gastroenterology Department, La Fe University Hospital, Valencia, Spain. [25]Department of Gastroenterology, Hepatology and Immunology, Children's Memorial Health Institute, Warsaw, Poland. [26]Department of Genetics, University of Delhi, South Campus, New Delhi, India. [27]Department of Pediatrics, The Medical University of Warsaw, Warsaw, Poland. [28]Department of Pediatrics, University of Naples Federico II, Naples, Italy. [29]Department of Paediatric Gastroenterology, University Medical Centre Utrecht, Utrecht, The Netherlands. [30]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands. [31]A full list of members is provided in the **Supplementary Note**. [32]Department of Pathology, Children's Memorial Health Institute, Warsaw, Poland. [33]Department of Paediatrics, Leiden University Medical Centre, Leiden, The Netherlands. [34]Department of Experimental Medicine, Faculty of Medicine, University of Milano-Bicocca, Monza, Italy. [35]University College London Genetics Institute, University College London, London, UK. [36]These authors contributed equally to this work. [37]These authors jointly directed this work. Correspondence should be addressed to D.A.v.H. (d.vanheel@qmul.ac.uk) or C.W. (c.wijmenga@medgen.umcg.nl).

ONLINE METHODS

Subjects. Written informed consent was obtained from all subjects with approval from the ethics committee or institutional review board of all participating institutions. All subjects, except those from the Indian population sample, were of European ancestry. DNA samples were taken from blood, lymphoblastoid cell lines or saliva.

Individuals affected with celiac disease were diagnosed according to standard clinical criteria, compatible serology and, in all cases, small intestinal biopsy; most cases were diagnosed using the revised European Society for Paediatric Gastroenterology, Hepatology and Nutrition criteria as a minimum requirement[34]. More specific requirements were as follows: cases from the UK[3–5] (hospital outpatients, $n = 1{,}145$) required a Marsh-classified stage 3 intestinal biopsy (HLA-DQ2.5cis tag SNP rs2187668 MAF = 0.4699); additional cases from the UK[4,5] (Celiac UK members, $n = 6{,}583$) had a self-reported diagnosis by intestinal biopsy (note the MAF of rs2187668 (0.4803) was similar as that in hospital outpatient cases from the UK, as compared to that in the combined UK controls (MAF = 0.1419)); cases from Italy (Milan)[5,35] and Poland[5] required a Marsh-classified stage 3 intestinal biopsy and positive endomysial or tissue transglutaminase antibodies; cases from Spain (CEGEC)[36] required at least a Marsh-classified stage 2 intestinal biopsy; cases from The Netherlands[5] required a Marsh-classified stage 3 intestinal biopsy or a Marsh-classified stage 2 intestinal biopsy with a compatible *HLA-DQ* type; cases from India (Punjab) required a Marsh-classified stage 3 intestinal biopsy and strongly positive tissue transglutaminase antibodies; and cases from Italy (Naples or Rome) required an abnormal intestinal biopsy and positive tissue transglutaminase antibodies[37].

The UK 1958 Birth Cohort and the UK Blood Services Common Controls were unselected population controls. Polish controls and Italian (Naples) controls excluded samples with positive celiac serology. Spanish (Madrid) controls were unselected blood donors and hospital employees. Spanish (CEGEC), Italian (Rome) and Indian (Punjab) controls were unselected blood donors. Italian controls (Milan) were unselected healthy individuals. Controls from The Netherlands were unselected blood donors and population controls.

SNP selection. All 1000 Genomes Project low-coverage whole-genome–sequencing pilot CEU variants within 0.1 cM of the lead SNP for each disease and region were selected. The September 2009 release comprising 60 CEU individuals was used (~5× mean read depth for whole-genome sequencing), and the markers selected were called in at least two of the Broad Institute, Sanger Institute and University of Michigan algorithms. Additional genomic region resequencing content was submitted for Immunochip analysis at specific loci from cases with celiac disease, Crohn's disease and type 1 diabetes and controls (**Supplementary Note**).

Genotyping. Samples were genotyped using the Immunochip according to Illumina's protocols (at labs in London, UK, Hinxton, UK, Groningen, The Netherlands, and Charlottesville, Virginia, USA). NCBI build 36 (hg18) mapping was used (Illumina manifest file Immuno_BeadChip_11419691_B.bpm).

Data quality control. Samples and variants with very low call rates were first excluded (after repeated testing of the samples). The Illumina GenomeStudio GenTrain2.0 algorithm was used to cluster an initial 2,000 UK samples. Subsequently, with additional sample data (case and control data were analyzed together), clusters were re-adjusted or excluded (manual or automated) for variants with low quality statistics (call rate <99.5%, a low GenCall score or many no calls with high intensity). This method produced better results than the GenoSNP or Illuminus clustering algorithms (data not shown). A cluster set based on 172,242 autosomal or X-chromosome variants (available on request) was then applied to all samples. Samples were excluded for call rate <99.5% across 172,242 markers. We then removed 15,657 non-polymorphic markers (that is, where only one of three expected genotype clouds was observed) that reflected a combination of ethnic-specific variants, allele-specific assay failure and substantial false-positive rates in early next-generation sequencing SNP calling algorithms.

Samples were excluded for incompatible recorded and genotype-inferred gender, duplicates and first- or second-degree relatives. Potential ethnic outliers were identified by multi-dimensional scaling plots of samples merged with HapMap3 data; the subset of SNPs common to HapMap3 and Immunochip accurately identified the different HapMap3 population samples. We considered the European and Indian collections separately.

Stepwise conditional logistic regression is sensitive to missing data and subtle genotyping error, so we therefore desired an ultra–high-quality dataset. Markers were excluded from all sample collections for deviation from Hardy-Weinberg equilibrium in controls ($P < 0.0001$) and/or differential missingness in no-call genotypes between cases and controls ($P < 0.001$) in any of the seven collections. Finally, we required a per-SNP call rate of >99.95% (a maximum of 12 no-call genotypes from 24,269 samples per autosomal marker), generating a data set of 139,553 markers (of which all but 372 indels are SNPs).

We visually inspected the intensity plot genotype clouds for all the markers listed in **Table 2** (as well as additional potential loci with $P < 1.9 \times 10^{-6}$) and confirmed all of these markers to be high quality. Genotype data has been deposited at the European Genome-Phenome Archive (see URLs), which is hosted by the European Bioinformatics Institute, under accession number EGAS00000000053.

Statistical analyses. Analyses were performed with PLINK v1.07 (ref. 38) using logistic regression tests with gender as a covariate and collection membership (**Table 1**) as a factorized covariate. Stepwise conditional logistic regression was performed in the order of markers with the smallest *P* value. Graphs were plotted in R (see URLs) and using a modified version of LocusZoom[39].

We permuted disease status for the dataset at each region to establish locus-wide statistical significance thresholds for defining independently associated SNPs. For each locus, defined by the LD boundaries (**Table 2**), we calculated the fifth percentile based on the nominal *P* value distribution for 1,000 permutations and controlling for multiple marker testing. This approach proved slightly more stringent than a per-locus Bonferroni correction for independent (using an estimate for independence as a pairwise $r^2 < 0.05$) variants (**Supplementary Table 3**). We estimated that our dataset contained 26,146 completely uncorrelated variants (using pairwise $r^2 < 0.05$ and a sliding 1,000-SNP window).

The fraction of additive variance was calculated using a liability threshold model[40] assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the UK dataset. Genetic variance was calculated assuming 50% heritability.

34. Anonymous. Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition. *Arch. Dis. Child.* **65**, 909–911 (1990).
35. Romanos, J.. *et al.* Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. *J. Med. Genet.* **46**, 60–63 (2009).
36. Plaza-Izurieta, L. *et al.* Revisiting genome wide association studies (GWAS) in coeliac disease: replication study in Spanish population and expression analysis of candidate genes. *J. Med. Genet.* **48**, 493–496 (2011).
37. Megiorni, F. *et al.* HLA-DQ and risk gradient for celiac disease. *Hum. Immunol.* **70**, 55–59 (2009).
38. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
39. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
40. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).