ARTICLES

nature genetics

Multiple common variants for celiac disease influencing immune gene expression

Patrick C A Dubois[1,39*], Gosia Trynka[2,39], Lude Franke[1,2], Karen A Hunt[1], Jihane Romanos[2], Alessandra Curtotti[3], Alexandra Zhernakova[4], Graham A R Heap[1], Róza Ádány[5], Arpo Aromaa[6], Maria Teresa Bardella[7,8], Leonard H van den Berg[9], Nicholas A Bockett[1], Emilio G de la Concha[10], Bárbara Dema[10], Rudolf S N Fehrmann[2], Miguel Fernández-Arquero[10], Szilvia Fiatal[5,11], Elvira Grandone[12], Peter M Green[13], Harry J M Groen[14], Rhian Gwilliam[15], Roderick H J Houwen[16], Sarah E Hunt[15], Katri Kaukinen[17], Dermot Kelleher[18], Ilma Korponay-Szabo[19,20], Kalle Kurppa[17], Padraic MacMathuna[21], Markku Mäki[17], Maria Cristina Mazzilli[22], Owen T McCann[15], M Luisa Mearin[23], Charles A Mein[3], Muddassar M Mirza[13], Vanisha Mistry[1], Barbara Mora[22], Katherine I Morley[15], Chris J Mulder[24], Joseph A Murray[25], Concepción Núñez[10], Elvira Oosterom[2], Roel A Ophoff[26–28], Isabel Polanco[29], Leena Peltonen[15,30], Mathieu Platteel[2], Anna Rybak[31], Veikko Salomaa[6], Joachim J Schweizer[23], Maria Pia Sperandeo[32], Greetje J Tack[24], Graham Turner[18], Jan H Veldink[9], Wieke H M Verbeek[24], Rinse K Weersma[33], Victorien M Wolters[16], Elena Urcelay[10], Bozena Cukrowska[34], Luigi Greco[32], Susan L Neuhausen[35], Ross McManus[18], Donatella Barisani[36], Panos Deloukas[15], Jeffrey C Barrett[15], Paivi Saavalainen[37,38], Cisca Wijmenga[2] & David A van Heel[1]

We performed a second-generation genome-wide association study of 4,533 individuals with celiac disease (cases) and 10,750 control subjects. We genotyped 113 selected SNPs with $P_{GWAS} < 10^{-4}$ and 18 SNPs from 14 known loci in a further 4,918 cases and 5,684 controls. Variants from 13 new regions reached genome-wide significance ($P_{combined} < 5 \times 10^{-8}$); most contain genes with immune functions (*BACH2, CCR4, CD80, CIITA-SOCS1-CLEC16A, ICOSLG* and *ZMIZ1*), with *ETS1, RUNX3, THEMIS* and *TNFRSF14* having key roles in thymic T-cell selection. There was evidence to suggest associations for a further 13 regions. In an expression quantitative trait meta-analysis of 1,469 whole blood samples, 20 of 38 (52.6%) tested loci had celiac risk variants correlated ($P < 0.0028$, FDR 5%) with *cis* gene expression.

Celiac disease is a common heritable chronic inflammatory condition of the small intestine induced by dietary wheat, rye and barley, as well as other unidentified environmental factors, in susceptible individuals. Specific *HLA-DQA1* and *HLA-DQB1* risk alleles are necessary, but not sufficient, for disease development[1,2]. The well-defined role of HLA-DQ heterodimers encoded by these alleles is to present cereal peptides to $CD4^+$ T cells, activating an inflammatory immune response in the intestine. A single genome-wide association study (GWAS) has been performed in celiac disease, which identified the *IL2-IL21* risk locus[1]. Subsequent studies probing the GWAS information in greater depth have identified a further 12 risk regions. Most of these regions contain a candidate gene that functions in the immune system, although only in the case of *HLA-DQA1* and *HLA-DQB1* have the causal variants been established[3–5]. Many of the known celiac disease–associated loci overlap with those of other immune-related diseases[6]. To identify additional risk variants, particularly those with smaller effect sizes, we performed a second-generation GWAS using more than six times as many samples as the previous GWAS and a denser genome-wide SNP set. We followed up promising findings in a large collection of independent samples.

RESULTS

Overview of study design

The GWAS included five European celiac disease case and control sample collections, including the celiac disease dataset reported previously[1]. We performed stringent data quality control (see Online Methods), including calling genotypes using a custom algorithm on both large sample sets and, where possible, cases and controls together (see Online Methods). We tested 292,387 non-*HLA* SNPs from the Illumina Hap300 marker set for association in 4,533 individuals with celiac disease and 10,750 control subjects of European descent (**Table 1**). A further 231,362 additional non-*HLA* markers from the Illumina Hap550 marker set were tested for association in a subset of 3,796 individuals with celiac disease and 8,154 controls. All markers

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

Received 4 December 2009; accepted 4 February 2010; published online 28 February 2010; corrected online 12 March 2010; doi:10.1038/ng.543

Table 1 Sample collections and genotyping platforms

		Celiac disease cases			Controls		
Collection	Country	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]
Stage 1: Genome-wide association							
1[d,e]	UK	778	737	Illumina Hap300v1-1	2,596[i]	2,596	Illumina Hap550-2v3
2[d,f]	UK	1,922	1,849	Illumina 670-QuadCustom_v1	5,069[i]	4,936	Illumina 1.2M-DuoCustom_v1
3[d]	Finland	674	647	Illumina 670-QuadCustom_v1	1,839[i]	1,829	Illumina 610-Quad
4[g]	The Netherlands	876	803	Illumina 670-QuadCustom_v1	960	846	Illumina 670-QuadCustom_v1
5[d]	Italy	541	497	Illumina 670-QuadCustom_v1	580	543	Illumina 670-QuadCustom_v1
Analysis of Hap300 markers[c]			4,533			10,750	
Analysis of additional Hap550 markers[c]			3,796			8,154	
Stage 2: Follow-up							
6	USA	987	973	Illumina GoldenGate	615	555	Illumina GoldenGate
7	Hungary	979	965	Illumina GoldenGate	1,126	1,067	Illumina GoldenGate
8[h]	Ireland	653	597	Illumina GoldenGate	1,499	1,456	Illumina GoldenGate
9	Poland	599	564	Illumina GoldenGate	745	716	Illumina GoldenGate
10	Spain	558	550	Illumina GoldenGate	465	433	Illumina GoldenGate
11[d]	Italy	1,056	1,010	Illumina GoldenGate	864	804	Illumina GoldenGate
12[d]	Finland	270	259	Illumina GoldenGate	653[j]	653	Illumina 610-Quad[j]
Subtotal			4,918			5,684	
Analysis of Hap300 markers and follow-up (91 SNPs)[c]			9,451			16,434	
Analysis of additional Hap550 markers and follow-up (40 SNPs)[c]			8,714			13,838	

[a]Sample numbers attempted for genotyping, before any quality control (QC) steps were applied. [b]Sample numbers after all quality control (QC) steps (used in the association analysis). [c]All platforms contain a common set of Hap300 markers; the Hap550, 610-Quad, 670-Quad and 1.2M contain a common set of Hap550 markers. [d]As an additional quality control step, we performed case-case and control-control comparisons for collection 1 versus 2, and collection 3 versus 12, for the 40 SNPs in **Table 2** and observed no markers with $P < 0.01$. We did observe (as expected) differences for collection 5 versus 11, from northern and southern Italy, respectively. [e]All 737 post-QC cases reported in a previous GWAS[1]. [f]690 of the post-QC cases and 1,150 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [g]498 of the post-QC cases and 767 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [h]352 of the post-QC cases and 921 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [i]Some of these data were generated elsewhere, and some prior quality control steps (information not available) had been applied. [j]Finnish stage 2 controls were individuals within the Finrisk collection for whom Illumina 610-Quad genotype data became available after the completion of stage 1.

were from autosomes or the X chromosome. Genotype call rates were >99.9% in both datasets. The overdispersion factor of association test statistics, $\lambda_{GC} = 1.12$, was similar to that observed in other GWASs of this sample size[7,8]. Findings were not substantially altered by imputation of missing genotypes for 737 cases with celiac disease genotyped on the Hap300 BeadChip and corresponding controls (**Table 1**, collection 1). Here we present results for directly genotyped SNPs, as around half the additional Hap550 markers cannot be accurately imputed from Hap300 data[9] (including the new *ETS1* locus reported in this study). Results for the top 1,000 markers are available in **Supplementary Data 1**; however, because of concerns regarding the detection of individuals' identities[10], results for all markers are available only on request to the corresponding author.

For follow-up, we first inspected genotype clouds for the 417 non-*HLA* SNPs that met $P_{GWAS} < 10^{-4}$, being aware that top GWAS signals might be enriched for genotyping artifact, and excluded 22 SNPs from further analysis using a low threshold for possible bias. We selected SNPs from 113 loci for replication. Markers that passed design and genotyping quality control included (i) 18 SNPs from all 14 previously identified celiac disease risk loci (including a tag SNP for the major celiac disease–associated *HLA-DQ2.5cis* haplotype[1]); (ii) 13 SNPs from all 7 newly discovered regions with $P_{GWAS} < 5 \times 10^{-7}$; (iii) 86 SNPs from 59 of 68 newly discovered regions with $5 \times 10^{-7} < P_{GWAS} < 5 \times 10^{-5}$ in stage 1; and (iv) 14 SNPs from 14 of 30 newly discovered regions with $5 \times 10^{-5} < P_{GWAS} < 10^{-4}$ in stage 1 (for this last category, we mostly chose regions with immune system genes). Two SNPs were selected per region for regions with stronger association, regions with possible multiple independent associations and/or regions containing genes of obvious biological interest. We successfully genotyped 131 SNPs in 7 independent follow-up cohorts comprising 4,918 individuals with celiac disease and 5,684 control subjects of European descent (**Table 1**). Genotype call rates were >99.9% in each collection. Primary association analyses of the combined GWAS and follow-up data were performed with a two-sided $2 \times 2 \times 12$ Cochran-Mantel-Haenszel test. Finally, we examined associated risk loci for *cis* expression-genotype correlations; a summary of subjects used for expression quantitative trait locus (eQTL) analyses is reported in **Supplementary Table 1**.

Celiac disease risk variants

The *HLA* locus and all 13 other previously reported celiac disease risk loci showed evidence for association at a genome-wide significance threshold ($P_{combined} < 5 \times 10^{-8}$; **Table 2** and **Supplementary Fig. 1**). We note that some loci were previously reported using less stringent criteria (for example, the $P < 5 \times 10^{-7}$ recommended by the 2007 WTCCC study[11]); however, in the current, much larger sample set, all known loci meet recently proposed $P < 5 \times 10^{-8}$ thresholds[12,13].

We identified 13 new risk regions with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$) of association, including regions containing the *BACH2*, *CCR4*, *CD80*, *CIITA-SOCS1-CLEC16A*, *ETS1*, *ICOSLG*, *RUNX3*, *THEMIS*, *TNFRSF14* and *ZMIZ1* genes, which have obvious immunological functions (**Table 2** and **Supplementary Fig. 1**). A further 13 regions met 'suggestive' criteria for association ($10^{-6} > P_{combined} > 5 \times 10^{-8}$ and/or $P_{GWAS} < 10^{-4}$ and $P_{followup} < 0.01$; **Table 2** and **Supplementary Fig. 1**). These regions also contain multiple genes with immunological functions, including *CD247*, *FASLG-TNFSF18-TNFSF4*, *IRF4*, *TLR7-TLR8*, *TNFRSF9* and *YDJC*. Six of the 39 non-HLA regions show evidence for the presence of multiple independently associated variants in a conditional logistic regression analysis (**Supplementary Table 2**).

We tested the 40 SNPs with the strongest association (**Table 2**) from each of the known genome-wide significant, new genome-wide significant and new suggestive loci for evidence of heterogeneity across the 12 collections studied. Only the *HLA* region was significant (Breslow-Day test $P < 0.05$ per 40 tests, rs2187668 $P = 4.8 \times 10^{-8}$), which is consistent with the well-described North-South gradient in HLA allele frequency in European populations, and more specifically for *HLA-DQ* in celiac disease[14].

We observed no evidence for interaction between each of the 26 genome-wide significant non-*HLA* loci, which is consistent with what has been reported for other complex diseases so far. However, we did observe weak evidence for lower effect sizes at non-*HLA* loci in high risk *HLA-DQ2.5cis* homozygotes, similar to what has been observed in type 1 diabetes[7].

To obtain more insight into the functional relatedness of the celiac disease risk loci, we applied GRAIL, a statistical tool that uses text mining of PubMed abstracts to annotate candidate genes from loci associated with common disease risk[15,16]. To assess the sensitivity of this tool (using known loci as a positive control), we first

Table 2 Genomic regions with the strongest association signals for celiac disease

Chr.	Position (bp)	SNP	LD block[a,b] (Mb)	Minor allele	Minor allele freq[c]	P_{GWAS} 4,533 cases, 10,750 controls	$P_{follow-up}$ 4,918 cases, 5,684 controls	$P_{combined}$ 9,451 cases, 16,434 controls	Odds ratio[c] (95% CI)	Multiple independent association signals[d]	Refs.	RefSeq Genes in LD block	Genes of interest and GRAIL annotation[e]
Previously reported risk variants													
1	190803436	rs2816316	190.73–190.81	C	0.160	1.45×10^{-12}	1.56×10^{-6}	2.20×10^{-17}	0.80 (0.76–0.84)		22	1	*RGS1*
2	61040333	rs13003464	60.78–61.74	G	0.401	4.92×10^{-8}	1.57×10^{-6}	3.71×10^{-13}	1.15 (1.11–1.20)	Yes	32	8	*REL*, *AHSA2*
2	102437000	rs917997	102.22–102.57	A	0.236	5.97×10^{-15}	7.83×10^{-4}	1.11×10^{-15}	1.19 (1.14–1.25)		22	5	*IL18RAP*, *IL18R1*, *IL1RL1*, *IL1RL2*
2	181704290	rs13010713	181.50–181.97	G	0.448	2.02×10^{-8}	3.21×10^{-4}	4.74×10^{-11}	1.13 (1.09–1.18)		33	1	*ITGA4*, *UBE2E3*
2	204510823	rs4675374	204.40–204.52	A	0.223	8.80×10^{-8}	4.94×10^{-3}	5.79×10^{-9}	1.14 (1.09–1.19)		17	2	*CTLA4*, *ICOS*, *CD28*
3	46210205	rs13098911	45.90–46.57	A	0.097	2.53×10^{-11}	1.96×10^{-7}	3.26×10^{-17}	1.30 (1.23–1.39)	Yes	22	11	*CCR1*, *CCR2*, *CCRL2*, *CCR3*, *CCR5*, *CCR9*
3	161147744	rs17810546	161.07–161.23	G	0.125	4.56×10^{-18}	9.57×10^{-12}	3.98×10^{-28}	1.36 (1.29–1.44)	Yes	22	1	*IL12A*
3	189595248	rs1464510	189.55–189.62	A	0.485	9.49×10^{-24}	3.63×10^{-18}	2.98×10^{-40}	1.29 (1.25–1.34)		22	1	*LPP*
4	123334952	rs13151961	123.19–123.78	G	0.142	6.31×10^{-18}	4.45×10^{-11}	2.18×10^{-27}	0.74 (0.70–0.78)		1	4	*IL2*, *IL21*
6	32713862	rs2187668	Gene identified	A	0.258	$<10^{-50}$	$<10^{-50}$	$<10^{-50}$	6.23 (5.95–6.52)	(Yes)	1,3	6	*HLA-DQA1*, *HLA-DQB1*
6	138014761	rs2327832	137.92–138.17	G	0.216	1.41×10^{-14}	1.97×10^{-6}	4.46×10^{-19}	1.23 (1.17–1.28)		32	0	*TNFAIP3*
6	159385965	rs1738074	159.24–159.45	A	0.434	3.14×10^{-8}	1.56×10^{-8}	2.94×10^{-15}	1.16 (1.12–1.21)		22	2	*TAGAP*
12	110492139	rs653178	110.19–111.51	G	0.495	6.03×10^{-14}	1.47×10^{-8}	7.15×10^{-21}	1.20 (1.15–1.24)		22	13	*SH2B3*
18	12799340	rs1893217	12.73–12.91	G	0.165	5.52×10^{-7}	1.04×10^{-4}	2.52×10^{-10}	1.17 (1.12–1.23)		17	1	*PTPN2*
New loci with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$)													
1	2516606	rs3748816	2.40–2.78	G	0.339	4.93×10^{-7}	1.17×10^{-3}	3.28×10^{-9}	0.89 (0.85–0.92)			4	*TNFRSF14*, *MMEL1*
1	25176163	rs10903122	25.11–25.18	A	0.480	3.21×10^{-5}	8.44×10^{-7}	1.73×10^{-10}	0.89 (0.85–0.92)			1	*RUNX3*
1	199158760	rs296547	199.12–199.31	A	0.357	6.46×10^{-5}	1.34×10^{-5}	4.11×10^{-9}	0.89 (0.86–0.92)			2	?
2	68452459	rs17035378[f]	68.39–68.54	G	0.278	1.34×10^{-5}	1.41×10^{-4}	7.79×10^{-9}	0.88 (0.84–0.92)			2	*PLEK*
3	32990473	rs13314993[f]	32.90–33.06	C	0.464	6.87×10^{-6}	1.09×10^{-4}	3.27×10^{-9}	1.13 (1.08–1.17)			2	*CCR4*
3	120601486	rs11712165[f]	120.59–120.78	C	0.394	5.40×10^{-7}	1.72×10^{-3}	8.03×10^{-9}	1.13 (1.08–1.17)			5	*CD80*, *KTELC1*
6	90983333	rs10806425	90.86–91.10	A	0.397	9.46×10^{-6}	9.25×10^{-6}	3.89×10^{-10}	1.13 (1.09–1.17)			1	*BACH2*, *MAP3K7*
6	128320491	rs802734	127.99–128.38	G	0.311	1.36×10^{-6}	1.70×10^{-9}	2.62×10^{-14}	1.17 (1.12–1.22)	Yes		2	*PTPRK*, *THEMIS*
8	129333771	rs9792269	129.21–129.37	G	0.238	8.14×10^{-6}	1.00×10^{-4}	3.28×10^{-9}	0.88 (0.84–0.91)			0	?
10	80728033	rs1250552	80.69–80.76	G	0.466	5.80×10^{-8}	1.81×10^{-3}	9.09×10^{-10}	0.89 (0.86–0.92)			1	*ZMIZ1*
11	127886184	rs11221332[f]	127.84–127.99	A	0.237	4.74×10^{-11}	9.98×10^{-7}	5.28×10^{-16}	1.21 (1.16–1.27)	Yes		1	*ETS1*
16	11311394	rs12928822	11.22–11.39	A	0.161	1.07×10^{-5}	7.59×10^{-4}	3.12×10^{-8}	0.86 (0.82–0.91)			4	*CIITA*, *SOCS1*, *CLEC16A*
21	44471849	rs4819388	44.42–44.47	A	0.280	3.42×10^{-5}	1.66×10^{-5}	2.46×10^{-9}	0.88 (0.84–0.92)			2	*ICOSLG*
New loci with suggestive evidence (either $10^{-6} > P_{combined} > 5 \times 10^{-8}$ or $P_{GWAS} < 10^{-4}$ and $P_{follow-up} < 0.01$)													
1	7969259	rs12727642	7.84–8.13	A	0.185	3.06×10^{-5}	8.21×10^{-4}	9.11×10^{-8}	1.14 (1.09–1.20)			4	*PARK7*, *TNFRSF9*
1	61564451	rs6691768	61.52–61.62	G	0.378	2.63×10^{-5}	1.16×10^{-3}	1.19×10^{-7}	0.90 (0.87–0.94)			1	*NFIA*
1	165678008	rs864537	165.43–165.71	G	0.391	1.01×10^{-7}	9.25×10^{-2}	3.80×10^{-7}	0.91 (0.87–0.94)			1	*CD247*
1	170977623	rs859637	170.87–171.20	A	0.486	8.15×10^{-5}	5.68×10^{-3}	1.75×10^{-6}	1.10 (1.06–1.14)			1	*FASLG*, *TNFSF18*, *TNFSF4*
3	69335589	rs6806528[f]	69.27–69.37	A	0.097	4.84×10^{-5}	7.66×10^{-4}	1.46×10^{-7}	1.19 (1.12–1.27)			1	*FRMD4B*
3	170974795	rs10936599	170.84–171.09	A	0.252	2.99×10^{-7}	6.63×10^{-2}	4.57×10^{-7}	1.12 (1.07–1.16)			3	?
6	328546	rs1033180[g]	0.32–0.40	A	0.080	9.14×10^{-6}	1.48×10^{-3}	5.58×10^{-8}	1.21 (1.13–1.29)	Yes		1	*IRF4*[g]
7	37341035	rs6974491	37.32–37.41	A	0.170	1.37×10^{-5}	2.63×10^{-3}	1.56×10^{-7}	1.14 (1.09–1.20)			1	*ELMO1*
13	49733716	rs2762051	49.63–49.96	A	0.184	3.35×10^{-5}	5.06×10^{-3}	6.64×10^{-7}	1.13 (1.08–1.18)			0	?
14	68347957	rs4899260	68.24–68.39	A	0.263	4.55×10^{-5}	2.21×10^{-3}	3.92×10^{-7}	1.12 (1.07–1.16)			2	*ZFP36L1*
17	42220599	rs2074404	41.40–42.25	C	0.250	5.03×10^{-5}	5.96×10^{-3}	1.23×10^{-6}	0.90 (0.86–0.94)			10	?
22	20312892	rs2298428	20.14–20.35	A	0.201	2.49×10^{-7}	4.13×10^{-2}	1.84×10^{-7}	1.13 (1.08–1.19)			6	*UBE2L3*, *YDJC*
X	12881445	rs5979785	12.82–12.93	G	0.263	6.32×10^{-6}	2.18×10^{-3}	6.36×10^{-8}	0.88 (0.84–0.92)			1	*TLR7*, *TLR8*

[a]The most significantly associated SNP from each region is shown. [b]LD regions were defined by extending 0.1 cM to the left and right of the focal SNP as defined by the HapMap3 recombination map. All chromosomal positions are based on NCBI build-36 coordinates. [c]Minor allele in all samples in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls. [d]Evidence from logistic regression at a genome-wide significant or suggestive level of significance after conditioning on other associated SNPs (see **Supplementary Table 2**). *HLA* region not tested, but previously known. [e]Selected named genes within or adjacent to the same LD block as the associated SNPs; causality is not proven. In particular, other genes and other causal mechanisms may exist. Gene names underlined are identified from GRAIL[15,16] analysis (see Online Methods) with $P_{text} < 0.01$. [f]These markers were present on the Hap550 but not Hap300 SNP sets, and are not genotyped for 737 cases and 2,596 controls in the stage 1 GWAS, and combined dataset analyses. Only minor changes in *P* values were observed when these genotypes were imputed and included in analysis. [g]The *IRF4* region (specifically rs9738805, $r^2 = 0.08$ with rs1033180 in HapMap CEU) was previously identified as showing strong geographical differentiation[11]. Association with celiac disease was still observed after correction for population stratification using either a structured association approach[34] (corrected $P_{GWAS} = 5.16 \times 10^{-6}$, $478 \times 2 \times 2$ CMH test) or principal components correction (uncorrected $P_{GWAS} = 7.05 \times 10^{-6}$, corrected $P_{GWAS} = 2.28 \times 10^{-5}$, Cochran-Armitage trend tests combined using weighted *Z* scores; see Online Methods). However, definitive exclusion of population stratification would require family-based association studies.

performed a 'leave-one-out' analysis of the 27 genome-wide significant celiac disease loci (including *HLA-DQ*). GRAIL scores of $P_{text} < 0.01$ were obtained for 12 of the 27 loci (44% sensitivity; **Table 2**). Factors that limit the sensitivity of GRAIL include biological pathways being both known (a 2006 dataset is used to avoid GWAS-era studies) and published in the literature. We then applied GRAIL analysis, using the 27 known regions as a seed, to all 49 regions (49 SNPs) with $10^{-3} > P_{combined} > 5 \times 10^{-8}$ and obtained GRAIL $P_{text} < 0.01$ for 9 regions (18.4%). As a control, only 5.5% (279 of 5,033) of randomly selected Hap550 SNPs reached this threshold. In addition to the five

Table 3 Celiac risk variants correlated with *cis* gene expression

SNP[a]	Chr.	SNP position[b]	Probe center position[b]	Illumina ArrayAddressID	Expression dataset[c]	Gene name	eQTL P[d]
Loci with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$)							
rs3748816	1	2516606	2412221	650452	HT-12	*PLCH2*	1.66×10^{-5}
rs3748816	1	2516606	2482955	6520725	Ref-8v2 + HT-12	*TNFRSF14*	1.30×10^{-3}
rs3748816	1	2516606	2510429	6250338	Ref-8v2	*C1orf93*	1.16×10^{-4}
rs3748816	1	2516606	2533115	2070246	Ref-8v2 + HT-12	*MMEL1*	1.03×10^{-20}
rs296547	1	199158760	198880146	1300279	Ref-8v2 + HT-12	*DDX59*	2.45×10^{-5}
rs842647	2	60972975	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	3.30×10^{-10}
rs13003464[e]	2	61040333	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	6.39×10^{-11}
rs3816281[f]	2	68461451	68461957	4810020	Ref-8v2 + HT-12	*PLEK*	7.97×10^{-26}
rs917997	2	102437000	102418571	6520180	Ref-8v2 + HT-12	*IL18RAP*	7.35×10^{-87}
rs13010713	2	181704290	181593865	1780433	HT-12	*UBE2E3*	4.93×10^{-5}
rs13098911	3	46210205	45964449	6550333	Ref-8v2 + HT-12	*CXCR6*	9.66×10^{-6}
rs13098911	3	46210205	46255176[g]	2190671	HT-12	*CCR3*	5.50×10^{-10}
rs13098911	3	46210205	46255176[g]	7570670	Ref-8v2	*CCR3*	5.69×10^{-4}
rs6441961[d]	3	46327388	46255176[h]	2190671	HT-12	*CCR3*	2.87×10^{-19}
rs6441961[d]	3	46327388	46255176[h]	7570670	Ref-8v2	*CCR3*	1.02×10^{-4}
rs11922594[f]	3	120608512	120683364[i]	6550288	Ref-8v2 + HT-12	*KTELC1*	5.09×10^{-17}
rs11922594[f]	3	120608512	120683364[i]	3850161	Ref-8v2 + HT-12	*KTELC1*	7.34×10^{-6}
rs10806425	6	90983333	90878075	3520349	HT-12	*BACH2*	1.92×10^{-3}
rs1738074	6	159385965	159380068	5890739	Ref-8v2 + HT-12	*TAGAP*	1.99×10^{-3}
rs1738074	6	159385965	159381094[j]	5360364	HT-12	*TAGAP*	3.23×10^{-4}
rs1738074	6	159385965	159381094[j]	4860242	HT-12	*TAGAP*	2.18×10^{-3}
rs1250552	10	80728033	80622540	2450131	Ref-8v2 + HT-12	*ZMIZ1*	1.80×10^{-3}
rs653178	12	110492139	110399552	6560301	Ref-8v2 + HT-12	*SH2B3*	9.24×10^{-12}
rs653178	12	110492139	110710447	840253	Ref-8v2 + HT-12	*ALDH2*	1.44×10^{-4}
rs653178	12	110492139	110894406[k]	2070736	HT-12	*TMEM116*	3.68×10^{-4}
rs653178	12	110492139	110894406[k]	3190129	Ref-8v2	*TMEM116*	1.51×10^{-3}
rs12928822	16	11311394	11335627	4540072	Ref-8v2 + HT-12	*C16orf75*	1.02×10^{-8}
rs4819388	21	44471849	44049567	7200373	Ref-8v2	*RRP1*	2.62×10^{-3}
Loci with suggestive evidence (either $10^{-6} > P_{combined} > 5 \times 10^{-8}$ or $P_{GWAS} < 10^{-4}$ and $P_{follow\text{-}up} < 0.01$)							
rs12727642	1	7969259	7956138	610193	Ref-8v2 + HT-12	*PARK7*	9.76×10^{-15}
rs864537	1	165678008	165710482[l]	6290400	Ref-8v2 + HT-12	*CD247*	1.77×10^{-9}
rs864537	1	165678008	165710482[l]	3890689	HT-12	*CD247*	2.93×10^{-7}
rs6974491	7	37341035	37157761	2750154	Ref-8v2 + HT-12	*ELMO1*	5.40×10^{-6}
rs2074404	17	42220599	41824345	3520672	Ref-8v2 + HT-12	*LRRC37A*	1.17×10^{-4}
rs2074404	17	42220599	42106695[m]	5260138	Ref-8v2 + HT-12	*NSF*	1.20×10^{-5}
rs2074404	17	42220599	42106695[m]	1410484	HT-12	*NSF*	4.28×10^{-4}
rs2074404	17	42220599	42223012	4070615	HT-12	*WNT3*	2.77×10^{-3}
rs2074404	17	42220599	42485154	4880037	HT-12	*LOC388397*	1.78×10^{-9}
rs2298428	22	20312892	20308188	1230242	Ref-8v2 + HT-12	*UBE2L3*	1.96×10^{-90}
rs5979785	X	12881445	12842944[n]	6480360	Ref-8v2 + HT-12	*TLR8*	3.88×10^{-13}
rs5979785	X	12881445	12842944[n]	3390612	Ref-8v2 + HT-12	*TLR8*	1.07×10^{-7}

See **Supplementary Figures 2** and **3** for detailed results and **Supplementary Table 3** for more details of Illumina expression probes.

[a]We tested the SNP with the strongest association from 34 of 39 non-HLA loci ($P_{combined} < 10^{-6}$, **Table 2**), Hap300 proxy SNPs for 4 further loci, and a second independently associated SNP from 6 loci, for correlation with gene expression in PAXgene blood RNA in up to 1,349 individuals. One locus (containing *ETS1*) where an adequate proxy SNP was not available was not included for the eQTL analysis. SNP-gene expression correlations were tested for probes within a 1-Mb window. Results are presented for SNPs showing significant correlations with *cis* gene expression after controlling false-discovery rate at 5% (corresponding to $P < 0.0028$). [b]All chromosomal positions are based on NCBI build-36 coordinates. Probe center position was determined by re-mapping probe sequences to the human transcriptome and calculated from the midpoint of the transcript start and transcript end positions in genomic coordinates. [c]'HT-12' comprise 1,240 individuals with blood gene expression assayed using Illumina Human HT-12v3 arrays; 'Ref-8v2' comprise 229 individuals with blood gene expression assayed using Illumina Human-Ref-8v2 arrays (see Online Methods). [d]Spearman rank correlation of genotype and residual variance in transcript expression. Meta-analysis eQTL P value shown if both datasets had identical probes. [e]Second, independently associated SNP from this locus. [f]Proxy SNP, $r^2 = 0.61$ in HapMap CEU with most associated SNP rs11712165. [g–n]Different Illumina probe sequences with the same probe center position.

'suggestive' loci shown in **Table 2**, GRAIL annotated four further interesting gene regions with lower significance in the combined association results: rs944141-*PDCD1LG2* ($P_{combined} = 4.4 \times 10^{-6}$), rs976881-*TNFRSF8* ($P_{combined} = 2.1 \times 10^{-4}$), rs4682103-*CD200-BTLA* ($P_{combined} = 6.8 \times 10^{-6}$) and rs4919611-*NFKB2* ($P_{combined} = 6.1 \times 10^{-5}$). There appeared to be further enrichment for genes of immunological interest that are not GRAIL-annotated in the $10^{-3} > P_{combined} > 5 \times 10^{-8}$ significance window, including rs3828599-*TNIP1* ($P_{combined} = 1.55 \times 10^{-4}$), rs8027604-*PTPN9* ($P_{combined} = 1.4 \times 10^{-6}$) and rs944141-*CD274* ($P_{combined} = 4.4 \times 10^{-6}$). Some of these findings, for which neither genome-wide significant nor suggestive association is achieved, are likely to comprise part of a longer tail of disease-predisposing common variants with weaker effect sizes. Definitive assessment of these biologically plausible regions would require genotyping and association studies using much larger sample collections than the present study.

We previously showed that there is considerable overlap between risk loci for celiac disease and type 1 diabetes[17], as well as between risk loci for celiac disease and rheumatoid arthritis[18], and more generally, there is now substantial evidence for shared risk loci between the common chronic immune-mediated diseases[6]. To update these observations, we searched 'A Catalog of Published Genome Wide Association Studies' (accessed 18 November 2009)[19] and the HuGE database[20]. We found some evidence (requiring a published association report of $P < 1 \times 10^{-5}$) of shared loci with at least one other inflammatory or immune-mediated disease for 18 of the current 27 genome-wide significant celiac disease risk regions. We defined shared regions as the broad linkage disequilibrium block; however, different SNPs are often reported in different diseases, and at only 3 of the 18 shared regions are associations across all diseases with the same SNP or a proxy SNP in $r^2 > 0.8$ in HapMap CEU. Currently, nine regions seem to be specific to celiac disease and might reflect distinctive disease biology, including the regions containing rs296547 and rs9792269 and the regions around *CCR4*, *CD80*, *ITGA4*, *LPP*, *PLEK*, *RUNX3* and *THEMIS*. In fact, locus sharing between diseases is probably greater because of both stochastic variation in results from sample size limitations and regions that have a genuinely stronger effect size in one disease and weaker effect size in another.

Genetic variation in *ETS1* has recently been reported to be associated with systemic lupus erythematosus (SLE) in the Chinese population, although it is not associated with SLE in European populations[21]. The most strongly associated celiac disease (European population) SNP, rs11221332, and the most strongly associated SLE (Chinese population) SNP, rs6590330, map 70 kb apart. Inspection of the HapMap phase II data shows broadly similar linkage disequilibrium patterns between Chinese (CHB) and European (CEU) populations in this region, with the two associated SNPs in separate nonadjacent linkage disequilibrium blocks. Thus, distinct common variants within the same gene can predispose to different autoimmune diseases across different ethnic groups.

Exploring the function of celiac disease risk variants

Celiac disease risk variants in the *HLA* genes alter protein structure and function[4]. However, we identified only four nonsynonymous SNPs with evidence for association with celiac disease ($P_{GWAS} < 10^{-4}$) from the other 26 genome-wide significant associated regions (rs3748816-*MMEL1*, rs3816281-*PLEK*, rs196432-*RUNX3*, rs3184504-*SH2B3*). Although comprehensive regional resequencing is required to test the possibility that coding variants contribute to the observed association signals, more subtle effects of genetic variation on gene expression are the more likely functional mechanism for complex disease genes. With this in mind, we performed a meta-analysis of new and published genome-wide eQTL datasets comprising 1,469 human whole blood (PAXgene) samples reflecting primary leukocyte gene expression. We applied a new method, transcriptional components, to remove a substantial proportion of inter-individual nongenetic expression variation and performed eQTL meta-analysis on the residual expression variation (Online Methods).

We assessed 38 of the 39 genome-wide significant and suggestive celiac disease–associated non-*HLA* loci (**Table 2**) for *cis* expression-genotype correlations. We tested the SNP with the strongest association from each region. However, for five regions the most associated SNP was not genotyped in the eQTL samples (Hap300 data); instead, for four of these, we tested a proxy SNP ($r^2 > 0.5$ in HapMap CEU). In addition, for six loci showing evidence of multiple independent associations in conditional regression analyses, we tested a second SNP that showed independent association with celiac disease for eQTL analysis. In total, we assessed 44 independent non-*HLA* SNP associations in peripheral whole blood samples genotyped on the Illumina Hap300 BeadChip and either Illumina Ref8 or HT12 expression arrays, correlating each SNP with data from gene probes mapping within a 1-Mb window.

We identified significant (Spearman $P < 0.0028$, corresponding to 5% false-discovery rate) eQTLs at 20 of 38 (52.6%) non-*HLA* celiac loci tested (**Table 3** and **Supplementary Figs. 2** and **3**). Some loci had evidence of eQTLs with multiple probes, genes or SNPs (**Table 3**). We assessed whether the number of SNPs with *cis*-eQTL effects out of the 44 SNPs that we tested was significantly higher than expected. On average, eQTL SNPs had a substantially higher minor allele frequency (MAF) than non-eQTL SNPs in the 294,767 SNPs tested. To correct for this, we selected 44 random SNPs that had an equal MAF distribution and determined for how many of these MAF-matched SNPs eQTLs were observed. There were a significantly higher number of eQTL SNPs ($P = 9.3 \times 10^{-5}$, 10^6 permutations) among the celiac disease–associated SNPs than expected by chance (22 observed eQTL SNPs versus 7.8 expected eQTL SNPs). Therefore, the celiac disease–associated regions are greatly enriched for eQTLs. These data indicate that some risk variants might influence celiac disease susceptibility through a mechanism of altered gene expression. Candidate genes with a significant eQTL where the peak eQTL signal and peak case-control association signal are similar (**Supplementary Fig. 3**) include *MMEL1*, *NSF*, *PARK7*, *PLEK*, *TAGAP*, *RRP1*, *UBE2L3* and *ZMIZ1*.

We also assessed the coexpression of genes that mapped within 500 kb of SNPs that showed the strongest case-control association from the 40 genome-wide significant and suggestive celiac disease loci in an analysis of the 33,109 human Affymetrix Gene Expression Omnibus dataset. This analysis loses power to detect tissue-specific correlations from the use of numerous tissue types, but it greatly gains power from the large sample size. We detected several distinct coexpression clusters (Pearson correlation coefficient between genes >0.5), including four clusters of immune-related genes that contain at least one gene from 37 of the 40 genome-wide significant and suggestive loci (**Fig. 1**). These data further demonstrate that genes from celiac disease risk loci map to multiple distinct immunological pathways involved in disease pathogenesis.

DISCUSSION

We previously reported that most celiac genetic risk variants mapped near genes that are functional in the immune system[22], and this remains true for the 13 new genome-wide significant and 13 new suggestive risk variants from the current study. We can now refine

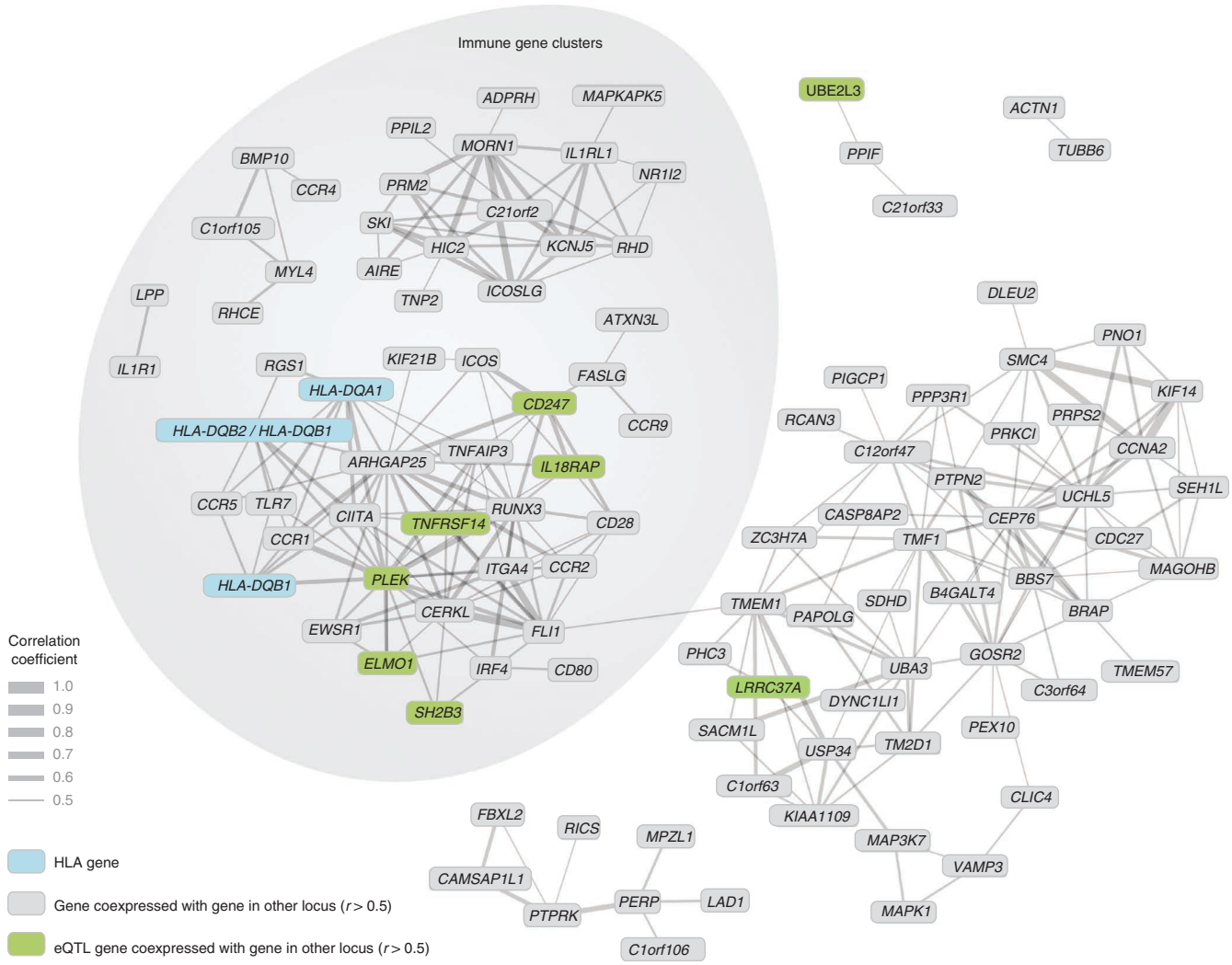

Figure 1 Coexpression analysis of genes mapping to 40 genome-wide significant and suggestive celiac disease regions in 33,109 heterogenous human samples from the Gene Expression Omnibus. Genes mapping within a 1-Mb window of associated SNPs (**Table 2**) were tested for interaction with genes from other loci. Interactions with Pearson correlation > 0.5 are shown ($P < 10^{-100}$). Only the genes known to contain causal mutations (*HLA-DQA1*, *HLA-DQB1*) were analyzed from the HLA region; HLA-DQB2/HLA-DQB1 is a single expression probeset mapping to both genes. No probe for *THEMIS* was present on the earlier version of the U133 array; however, in a subset analysis of U133 Plus2.0 data, *THEMIS* is coexpressed in the major immune gene cluster.

these observations and highlight specific immunological pathways that are relevant to the pathogenesis of celiac disease.

One key pathway worth highlighting is T-cell development in the thymus. The rs802734 linkage disequilibrium block contains the recently identified gene *THEMIS* (thymus-expressed molecule involved in selection). *THEMIS* has a key regulatory role in both positive and negative T-cell selection during late thymocyte development[23]. Furthermore, the rs10903122 linkage disequilibrium block contains *RUNX3*, a master regulator of $CD8^+$ T lymphocyte development in the thymus[24,25]. *TNFRSF14* (LIGHTR, rs3748816 linkage disequilibrium block) has widespread functions in peripheral leukocytes and a crucial role in promoting thymocyte apoptosis[26]. The *ETS1* transcription factor (rs11221332 linkage disequilibrium block) is also active in peripheral leukocytes; however, it is also a key player in thymic $CD8^+$ lineage differentiation, acting in part by promoting *RUNX3* expression[27].

The importance of the thymus in the pathogenesis of autoimmune diseases has been previously emphasized by the established role of thymectomy in the treatment of myasthenia gravis. In type 1 diabetes, disease-associated genetic variation in the insulin gene *INS* causes altered thymic insulin expression and subsequent T-cell tolerance for insulin as a self-protein[28]. However, the importance of thymic T-cell regulation in the etiology of celiac disease has not been previously recognized. It is conceivable that the associated variants might alter biological processes before thymic MHC-ligand interactions. Alternatively, it is now clear that exogenous antigen presentation and selection occurs in the thymus through migratory dendritic cells; this has been demonstrated for skin and has been hypothesized for food antigens[29,30]. These findings suggest that it would be worthwhile to investigate immunological and pharmacological modifiers of T-cell tolerance more generally in autoimmune diseases.

A second pathway worth noting is the innate immune detection of viral RNA. Although the association signal at rs5979785 ($P_{combined} = 6.36 \times 10^{-8}$) in the *TLR7-TLR8* region is just outside our genome-wide significance threshold, we observe a strong effect of rs5979785 on *TLR8* expression in whole blood. Both TLRs recognize viral RNA. Taken together with the recent observation that rare loss-of-function mutations in the enteroviral response gene *IFIH1* are protective against

type 1 diabetes[31], these findings implicate viral infection (and the nature of the host response to infection) as a putative environmental trigger that could be common to these autoimmune diseases.

A third pathway involves T- and B-cell co-stimulation (or co-inhibition). This class of molecules controls the strength and nature of the response to T-cell or B-cell (immunoglobulin) receptor activation by antigens. We observe multiple regions with genes (*CTLA4-ICOS-CD28*, *TNFRSF14*, *CD80*, *ICOSLG*, *TNFRSF9*, *TNFSF4*) from this class of ligand-receptor pairs, indicating that fine control of the adaptive immune response might be altered in individuals at risk of celiac disease.

A final pathway involves cytokines, chemokines and their receptors. Our previous report discussed the function of the 2q11–12 interleukin receptor cluster (*IL18RAP* and so on), the 3p21 chemokine receptor cluster (*CCR5* and so on) and the loci containing *IL2-IL21* and *IL12A*[22]. We now report additional loci containing *TNFSF18* and *CCR4*.

We estimate that the current celiac disease variants, including the major celiac disease-associated *HLA* variant, *HLA-DQ2.5cis*, less common celiac disease–associated haplotypes in the *HLA* (*HLA-DQ8*; *HLA-DQ2.5trans*; *HLADQ2.2*), and the additional 26 definitively implicated loci explain about 20% of total celiac disease variance, which would represent 40% of genetic variance, assuming a heritability of 0.5. A long tail of common variants with low effect size, along with highly penetrant rare variants (both at the established loci and elsewhere in the genome), might contribute substantially to the remaining heritability.

We observed different haplotypes within the *ETS1* region associated with celiac disease in Europeans and SLE in the Chinese population. For some autoimmune diseases studied in European origin populations, although the same linkage disequilibrium block has been associated, the association is with a different haplotype. In some cases, the same variants are associated, but the direction of association is opposite (for example, rs917997-*IL18RAP* in celiac disease versus type 1 diabetes). We believe further exploration of these signals might reveal critical differences in the nature of the immune system perturbation between these diseases.

Previously, investigators have observed that only a small proportion of GWAS signals involve coding variants and have suggested that these variants might instead influence regulation of gene expression. Here we show that over half the variants associated with celiac disease are correlated with expression changes in nearby genes. This mechanism is likely to explain the function of some risk variants for other common, complex diseases. Further research is needed to definitively determine at each locus both the variants that can cause celiac disease and their functional mechanisms.

METHODS

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

Accession numbers. Expression data are available in GEO (http://www.ncbi.nlm.nih.gov/geo/) as GSE20142 and GSE20332.

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 UK clinicians (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 the celiac disease blood samples described in our previous studies[1,22]. We thank the genotyping facility of the UMCG (J. Smolonska and P. van der Vlies) for generating part of the GWAS and replication data and the gene expression data; R. Booij and M. Weenstra for preparation of Italian samples; H. Ahola, A. Heimonen, L. Koskinen, E. Einarsdottir and K. Löytynoja for their work on Finnish sample collection, preparation and data handling; and E. Szathmári, J.B.Kovács, M. Lörincz and A. Nagy for their work with the Hungarian families. The Health2000 organization, Finrisk consortium, K. Mustalahti, M. Perola, K. Kristiansson and J. Koskinen are thanked for providing the Finnish control genotypes. We thank D.G. Clayton and N. Walker for providing T1DGC data in the required format. We thank the Irish Transfusion Service and Trinity College Dublin Biobank for control samples and V. Trimble, E. Close, G. Lawlor, A. Ryan, M. Abuzakouk, C. O'Morain and G. Horgan for celiac disease sample collection and preparation We acknowledge DNA provided by Mayo Clinic Rochester and thank M. Bonamico and M. Barbato (Department of Paediatrics, Sapienza University of Rome, Italy) for recruiting individuals. We thank Polish clinicians for recruitment of individuals with celiac disease (Z. Domagala, A. Szaflarska-Poplawska, B. Oralewska, W. Cichy, B. Korczowski, K. Fryderek, E. Hapyn, K. Karczewska, A. Zalewska, I. Sakowska-Maliszewska, R. Mozrzymas, A. Zabka, M. Kolasa and B. Iwanczak). We thank M. Szperl for isolating DNA from blood samples provided by the Children's Memorial Health Institute (Warsaw, Poland). Dutch and UK genotyping for the second celiac disease GWAS was funded by the Wellcome Trust (084743 to D.A.v.H.). Italian genotyping for the second celiac disease GWAS was funded by the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009 to C.W.) and by the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.). E.G. is funded by the Italian Ministry of Healthy (grant RC2009). L.H.v.d.B. acknowledges funding from the Prinses Beatrix Fonds, the Adessium foundation and the Amyotrophic Lateral Sclerosis Association. L.F. received a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031) and a VENI grant from NWO (ZonMW grant 916.10.135). P.C.A.D. is an MRC Clinical Training Fellow (G0700545). G.T. received a Ter Meulen Fund grant from the Royal Netherlands Academy of Arts and Sciences (KNAW). The gene expression study was funded in part by COPACETIC (EU grant 201379). This study makes use of data generated by the Wellcome Trust Case-Control Consortium 2 (WTCCC2). A full list of the WTCCC2 investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the WTCCC2 project was provided by the Wellcome Trust under award 085475. This research utilizes 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 (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD) and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418. We acknowledge the use of BRC Core Facilities provided by the financial support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. We acknowledge funding from the NIH: DK050678 and DK081645 (to S.L.N.), NS058980 (to R.A.O.); and DK57892 and DK071003 (to J.A.M.). The collection of Finnish and Hungarian subjects with celiac disease was funded by the EU Commission (MEXT-CT-2005-025270), the Academy of Finland, Hungarian Scientific Research Fund (contract OTKA 61868), the University of Helsinki Funds, the Competitive Research Funding of the Tampere University Hospital, the Foundation of Pediatric Research, the Sigrid Juselius Foundation and the Hungarian Academy of Sciences (2006TKI247 to R.A.). Funding for the collection and genotyping of the Polish samples was provided by UMC Cooperation Project (6/06/2006/NDON). R.M. is funded by Science Foundation Ireland. C. Núñez has a FIS contract (CP08/0213). The Dublin Centre for Clinical Research contributed to collection of samples from affected individuals and is funded by the Irish Health Research Board and the Wellcome Trust. Finally, we thank all individuals with celiac disease and control individuals for participating in this study.

AUTHOR CONTRIBUTIONS

D.A.v.H. and C.W. designed, co-ordinated and led the study. Experiments were performed in the labs of C.W., D.A.v.H., C.A.M., P.D. and P.M.G. Major contributions were: (i) DNA sample preparation: P.C.A.D., G.T., K.A.H., J.R., A.Z. and P.S.; (ii) genotyping: P.C.A.D., G.T., K.A.H., A.C., J.R. and R.G.; (iii) expression data generation: H.J.M.G., L.H.v.d.B., R.A.O., R.K.W. and L.F.; (iv) case-control association analyses: P.C.A.D., G.T., L.F., J.C.B. and D.A.v.H.; (v) expression analyses: L.F., G.A.R.H. and R.S.N.F.; (vi) manuscript preparation: P.C.A.D., G.T., L.F., R.S.N.F., G.A.R.H., J.C.B., C.W. and D.A.v.H. Other authors contributed variously to sample collection and all other aspects of the study. 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://npg.nature.com/reprintsandpermissions/.

1. 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).
2. van Heel, D.A. & West, J. Recent advances in coeliac disease. *Gut* **55**, 1037–1046 (2006).
3. Sollid, L.M. *et al.* Evidence for a primary association of celiac disease to a particular HLA-DQ α/β heterodimer. *J. Exp. Med.* **169**, 345–350 (1989).
4. Kim, C.Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L.M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. *Proc. Natl. Acad. Sci. USA* **101**, 4175–4179 (2004).
5. Henderson, K.N. *et al.* A structural and immunological basis for the role of human leukocyte antigen DQ8 in celiac disease. *Immunity* **27**, 23–34 (2007).
6. 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).
7. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
8. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
9. Anderson, C.A. *et al.* Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* **83**, 112–119 (2008).
10. Jacobs, K.B. *et al.* A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat. Genet.* **41**, 1253–1257 (2009).
11. 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).
12. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
13. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
14. Karell, K. *et al.* HLA types in celiac disease patients not carrying the DQA1*05–DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. *Hum. Immunol.* **64**, 469–477 (2003).
15. Raychaudhuri, S. *et al.* Genetic variants at *CD28, PRDM1* and *CD2/CD58* are associated with rheumatoid arthritis risk. *Nat. Genet.* **41**, 1313–1318 (2009).
16. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
17. 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).
18. Coenen, M.J. *et al.* Common and different genetic background for rheumatoid arthritis and coeliac disease. *Hum. Mol. Genet.* **18**, 4195–4203 (2009).
19. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
20. Yu, W., Clyne, M., Khoury, M.J. & Gwinn, M. Phenopedia and Genopedia: Disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* **26**, 145–146 (2010).
21. Han, J.W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1234–1237 (2009).
22. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008).
23. Allen, P.M. Themis imposes new law and order on positive selection. *Nat. Immunol.* **10**, 805–806 (2009).
24. Sato, T. *et al.* Dual functions of Runx proteins for reactivating CD8 and silencing CD4 at the commitment process into CD8 thymocytes. *Immunity* **22**, 317–328 (2005).
25. Woolf, E. *et al.* Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proc. Natl. Acad. Sci. USA* **100**, 7731–7736 (2003).
26. Wang, J. & Fu, Y.X. LIGHT (a cellular ligand for herpes virus entry mediator and lymphotoxin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC/self-peptide. *J. Immunol.* **170**, 3986–3993 (2003).
27. Zamisch, M. *et al.* The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8 T cell differentiation in the thymus. *J. Exp. Med.* **206**, 2685–2699 (2009).
28. Vafiadis, P. *et al.* Insulin expression in human thymus is modulated by *INS* VNTR alleles at the *IDDM2* locus. *Nat. Genet.* **15**, 289–292 (1997).
29. Bonasio, R. *et al.* Clonal deletion of thymocytes by circulating dendritic cells homing to the thymus. *Nat. Immunol.* **7**, 1092–1100 (2006).
30. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat. Rev. Immunol.* **9**, 833–844 (2009).
31. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
32. Trynka, G. *et al.* Coeliac disease-associated risk variants in *TNFAIP3* and *REL* implicate altered NF-κB signalling. *Gut* **58**, 1078–1083 (2009).
33. Garner, C.P. *et al.* Replication of celiac disease UK genome-wide association study results in a US population. *Hum. Mol. Genet.* **18**, 4219–4225 (2009).
34. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).

[1]Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [2]Genetics Department, University Medical Center and Groningen University, Groningen, The Netherlands. [3]The Genome Centre, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [4]Division of Biomedical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [5]Department of Preventive Medicine, University of Debrecen, Debrecen, Hungary. [6]National Institute for Health and Welfare, Helsinki, Finland. [7]Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy. [8]Department of Medical Sciences, University of Milan, Milan, Italy. [9]Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands. [10]Clinical Immunology Department, Hospital Clínico San Carlos, Madrid, Spain. [11]Public Health Research Group of Hungarian Academy of Sciences, Medical & Health Science Center, University of Debrecen, Debrecen, Hungary. [12]Unitá di Aterosclerosi e Trombosi, I.R.C.C.S Casa Sollievo della Sofferenza, S. Giovanni Rotondo, Foggia, Italy. [13]NIHR GSTFT/KCL Comprehensive Biomedical Research Centre, King's College London School of Medicine, Guy's Hospital, London, UK. [14]Department of Pulmonology, University Medical Center and Groningen University, Groningen, The Netherlands. [15]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [16]Department of Paediatric Gastroenterology, University Medical Centre Utrecht, Utrecht, The Netherlands. [17]Paediatric Research Centre, University of Tampere Medical School and Tampere University Hospital, Tampere, Finland. [18]Department of Clinical Medicine, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland. [19]Heim Pal Childrens Hospital, Budapest, Hungary. [20]Department of Pediatrics, Medical and Health Science Center, University of Debrecen, Hungary. [21]Gastrointestinal Unit, Mater Misericordiae University Hospital, Dublin, Ireland. [22]Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy. [23]Department of Paediatrics, Leiden University Medical Centre, Leiden, The Netherlands. [24]Department of Gastroenterology, VU Medical Center, Amsterdam, The Netherlands. [25]Division of Gastroenterology and Hepatology, Department of Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. [26]Department of Medical Genetics and [27]Rudolf Magnus Institute, University Medical Center Utrecht, Utrecht, The Netherlands. [28]Center for Neurobehavioral Genetics, University of California, Los Angeles, California, USA. [29]Pediatric Gastroenterology Department, Hospital La Paz, Madrid, Spain. [30]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland. [31]Department of Gastroenterology, Hepatology and Immunology, Children's Memorial Health Institute, Warsaw, Poland. [32]European Laboratory for Food Induced Disease, University of Naples Federico II, Naples, Italy. [33]Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands. [34]Department of Pathology, Children's Memorial Health Institute, Warsaw, Poland. [35]Department of Population Sciences, Beckman Research Institute of the City of Hope, Duarte, California, USA. [36]Department of Experimental Medicine, Faculty of Medicine University of Milano-Bicocca, Monza, Italy. [37]Department of Medical Genetics and [38]Research Program for Molecular Medicine, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland. [39]These authors contributed equally to this work. Correspondence should be addressed to D.A.v.H. (d.vanheel@qmul.ac.uk) or regarding expression analyses to L.F. (lude@cleverfranke.com).

ONLINE METHODS

Subjects. Written informed consent was obtained from all subjects, with Ethics Committee/Institutional Review Board approval. All individuals are of European ancestry. Affected celiac individuals were diagnosed according to standard clinical, serological and histopathological criteria, including small intestinal biopsy. DNA samples were from blood, lymphoblastoid cell lines or saliva. A more detailed description of subjects is provided in a **Supplementary Note**.

GWAS genotyping. For an overview, see **Table 1**. UK(1) case and control genotyping has been described[1,7]. Illumina 670-Quad and 1.2M-Duo (custom chips designed for the WTCCC2 and comprising Hap550/1M and common CNV content) and 610-Quad genotyping was performed in London, Hinxton and Groningen. Bead intensity data was normalized for each sample in BeadStudio, *R* and theta values exported and genotype calling performed using a custom algorithm[1,35]. A detailed description of genotype calling steps is provided in a **Supplementary Note**.

Quality control steps were performed in the following order. First, very low call rate samples and SNPs were excluded. SNPs were excluded from all sample collections if any collection showed call rates <95% or deviation from Hardy-Weinberg equilibrium ($P < 0.0001$) in controls. Samples were excluded for call rate <98%, incompatible recorded gender and genotype-inferred gender, ethnic outliers (identified by multi-dimensional scaling plots of samples merged with HapMap Phase II data), duplicates and first-degree relatives. We excluded 22 of 417 SNPs showing apparent association ($P_{GWAS} < 10^{-4}$) after visual inspection of *R* theta plots suggested possible bias.

The over-dispersion factor of association test statistics (genomic control inflation factor), λ_{GC}, was calculated using observed versus expected values for all SNPs in PLINK.

Follow-up genotyping. For an overview, see **Table 1**. Finnish controls (12) were genotyped on the 610-Quad BeadChip; other samples were genotyped using Illumina GoldenGate BeadXpress assays in London and Groningen. Genotyping calling was performed in BeadStudio for combined cases and controls in each separate collection, with the exception of the Finnish collection, and whole genome amplified samples (89 Irish cases and 106 Spanish controls). Quality control steps were performed as for the GWAS. In total, 131 of 144 SNPs passed quality control and visual inspection of genotype clouds.

SNP association analysis. Analyses were performed using PLINK v1.07 (ref. 36), mostly using the Cochran-Mantel-Haenzel test. Logistic regression analyses were used to define the independence of association signals within the same linkage disequilibrium block, with group membership included as a factorized covariate.

Genotype imputation was performed for samples genotyped on the Hap300 using BEAGLE and CEU, TSI, MEX and GIH reference samples from HapMap3. Association analysis was performed using logistic regression on posterior genotype probabilities, with group membership included as a factorized covariate.

Structured association tests were performed using PLINK as described using genetically matched cases and controls within collections identified by identity by state similarity across autosomal non-HLA SNPs[34] (settings–ppc 0.001–cc, clusters constrained by the five collections). Principal components analysis was performed using EIGENSTRAT and a set of 12,810 autosomal non-HLA SNPs chosen for low LD and ancestry information[37,38]; association tests were corrected for the top 10 principal components and combined using weighted *Z* scores.

The fraction of additive variance was calculated using a liability threshold model[39] assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the combined replication panel. Genetic variance was calculated assuming 50% heritability.

GRAIL analysis. We performed GRAIL analysis (http://www.broadinstitute.org/mpg/grail/grail.php) using HG18 and Dec2006 PubMed datasets, default settings for SNP rs number submission, and the 27 genome-wide significant celiac disease risk loci (most associated SNP) as seeds. As a query, we used either associated SNPs or 101 × 50 randomly chosen Hap550 SNP datasets (5,050 SNPs, of which 5,033 mapped to the GRAIL database).

Identification of transcriptional components. We noted that the power of eQTL studies in humans is limited by substantial observed inter-individual variation in expression measurements due to nongenetic factors, and therefore developed a method, 'transcriptional components', to remove a large component of this variation (manuscript in preparation). Expression data from 42,349 heterogeneous human samples hybridized to Affymetrix HG-U133A (GEO accession number: GPL96) or HG-U133 Plus 2.0 (GEO accession number: GPL570) Genechips were downloaded[40]. Samples missing data for >150 probes were excluded, and only probes available on both platforms were analyzed, resulting in expression data for 22,106 probes and 41,408 samples. We performed quantile normalization using the median rank distribution[41] and $\log_2$ transformed the data, ensuring an identical distribution of expression signals for every sample, discarding previous normalization and transformation steps.

Initial quality control (QC) was performed by applying principal component analysis (PCA) on the sample correlation matrix (pair-wise Pearson correlation coefficients between all samples). The first principal component (PC), explaining ~80–90% of the total variance[42,43], describes probe-specific variance. 6,375 samples with correlation $R < 0.75$ of the sample array with this PC were considered outliers of lesser quality and excluded from analysis. We excluded entire GEO datasets where >25% of the samples were outliers (probably expression ratios versus a reference, not absolute data). The final dataset comprised 33,109 samples (17,568 GPL96 and 15,541 GPL570 samples), and we repeated the normalization and transformation on the originally deposited expression values of these post-quality control samples.

We next applied PCA on the pairwise 22,106 × 22,106 probe Pearson correlation coefficient matrix assayed on the 33,109 sample dataset (our fast C++ tool, *MATool*, is available upon request), attempting to simplify the structure of the data. Here, PCA represents a transformation of a set of correlated probes into sets of uncorrelated linear additions of probe expression signals (eigenvectors) that we name transcriptional components (TCs). Each TC is a weighted sum of probe expression signals and eigenvector probe coefficients. These TC scores can be calculated for each observed expression array sample (reflecting the TC activity per sample).

Subjects for expression-genotype correlation. We obtained peripheral blood DNA and RNA (PAXgene) from Dutch and UK individuals who were disease cases or controls for GWAS studies (**Supplementary Table 1**). All samples had been genotyped for a common SNP set on Illumina platforms. Analysis was confined to 294,767 SNPs that had a MAF ≥ 5%, call-rate ≥ 95% and exact HWE $P > 0.001$. RNA from the samples was hybridized to either Illumina HumanRef-8 v2 arrays (229 samples, Ref-8v2) or Illumina HumanHT-12 arrays (1,240 samples, HT-12), and raw probe intensity extracted using BeadStudio. The Ref-8v2 samples were jointly quantile normalized and $\log_2$ transformed, as were the HT-12 samples. Subsequent analyses were also conducted separately for these datasets, up to the eventual eQTL mapping, which uses a meta-analysis framework, combining eQTL results from both arrays. HT-12 and Ref-8v2 arrays are different, but share many probes with identical probe sequences. Illumina sometimes use different probe identifiers for the same probe sequences; in meta-analysis and **Table 3**, the label HT-12 was used if both HT-12 and Ref-8v2 had the same sequence.

Re-mapping of probes. If probes mapped incorrectly or cross-hybridized to multiple genomic loci, it might be that an eQTL would be detected that would be deemed a *trans*-eQTL. To prevent this, we used a mapping approach versus a known reference that we developed for high-throughput short sequence RNAseq data[44]. We took the DNA sequence as synthesized for each cDNA probe and aligned it against a transcript masked gDNA genome combined with cDNA sequences. A more detailed description of probe re-mapping is provided in a **Supplementary Note**. Probes that did not map or that mapped to multiple different locations were removed.

Affymetrix transcriptional components applied to Illumina expression data. TC scores can be inferred in new (non-Affymetrix) datasets for every new individual sample. For the Illumina samples (used for the *cis*-eQTL mapping), only Illumina probes that could be mapped to any of our 22,106 Affymetrix

probes were used (www.switchtoi.com/probemapping.ilmn). The TC score of sample i for the j^{th} TC is defined as: $TCscore_{ij} = \sum_{t=1}^{t=n} a_{ti} \times v_{tj}$, where v_{tj} is defined as the t^{th} Affymetrix probe coefficient for the j^{th} TC; a_{ti} is the Illumina expression measurement for the t^{th} mapped probe for sample i. We inferred the Illumina TC scores for the top 1,000 TCs.

Removal of transcriptional component effects from Illumina expression data. Because our Illumina eQTL dataset (n = 1,469) is much less heterogeneous than the Affymetrix dataset (n = 33,109), we expect that some TCs will hardly vary. We therefore performed a PCA on the covariance matrix of the top 1,000 inferred TC scores for the Illumina dataset to effectively compress the TC data into a small set of 'aggregate TCs' (aTCs). As aTCs are orthogonal, we used linear regression to eliminate the effect of the top 50 aTCs. We correlated the TC-scores for each peripheral blood sample with probe expression levels. We then used the resulting residual gene expression data for subsequent *cis*-eQTL mapping.

***cis*-eQTL mapping.** We used the residual gene expression data in a meta-analysis framework, as described[45,46]. In brief, analyses were confined to those probe-SNP pairs for which the distance from probe transcript mid-point to SNP genomic location was less than 500 kb. To prevent spurious associations due to outliers, a nonparametric Spearman's rank correlation analysis was performed. When a particular probe-SNP pair was present in both the HT12 and H8v2 datasets, an overall, joint P value was calculated using a weighted Z-method (square root of the dataset sample number). To correct for multiple testing, we controlled the false-discovery rate (FDR). The distribution of observed P values was used to calculate the FDR, by permuting expression phenotypes relative to genotypes 1,000 times within the HT12 and H8v2 dataset. Finally, we removed any probes from analysis which contained a known SNP (1000Genomes CEU SNP data, April 2009 release).

35. Franke, L. *et al.* Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am. J. Hum. Genet.* **82**, 1316–1333 (2008).
36. 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).
37. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
38. Yu, K. *et al.* Population substructure and control selection in genome-wide association studies. *PLoS One* **3**, e2551 (2008).
39. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
40. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
41. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
42. Sherlock, G. Analysis of large-scale gene expression data. *Brief. Bioinform.* **2**, 350–362 (2001).
43. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
44. Heap, G.A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. *Hum. Mol. Genet.* **19**, 122–134 (2010).
45. Heap, G.A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2**, 1 (2009).
46. Franke, L. & Jansen, R.C. eQTL analysis in humans. *Methods Mol. Biol.* **573**, 311–328 (2009).

doi:10.1038/ng.543