European Journal of Human Genetics (2007) 15, 980–987

www.nature.com/ejhg

ARTICLE

A comprehensive screen for SNP associations on chromosome region 5q31–33 in Swedish/Norwegian celiac disease families

Silja Svanstrøm Amundsen*[,1], Svetlana Adamovic[2], Åsa Hellqvist[2], Staffan Nilsson[3,4], Audur H Gudjónsdóttir[5], Henry Ascher[5,6], Johan Ek[7], Kristina Larsson[8], Jan Wahlström[2], Benedicte A Lie[9], Ludvig M Sollid[1,9] and Åsa Torinsson Naluai[2,4]

[1]*Institute of Immunology, University of Oslo, Oslo, Norway;* [2]*Department of Medical and Clinical Genetics, Institute of Biomedicine, Sahlgrenska Academy at Göteborg University, Göteborg, Sweden;* [3]*Chalmers University of Technology, Göteborg, Sweden;* [4]*Swegene Genomics and Bioinformatics Core Facilities, Sahlgrenska Academy at Göteborg University, Göteborg, Sweden;* [5]*Department of Pediatrics, The Queen Silvia Children's Hospital, The Sahlgrenska Academy, Göteborg University, Göteborg, Sweden;* [6]*The Nordic School of Public Health, Göteborg, Sweden;* [7]*Department of Pediatrics, Buskerud Hospital Trust, Drammen, Norway;* [8]*Department of Medical Sciences, Uppsala University, Uppsala, Sweden;* [9]*Institute of Immunology, Rikshospitalet-Radiumhospitalet Medical Center, Oslo, Norway*

Celiac disease (CD) is a gluten-induced enteropathy, which results from the interplay between environmental and genetic factors. There is a strong human leukocyte antigen (HLA) association with the disease, and HLA-DQ alleles represent a major genetic risk factor. In addition to HLA-DQ, non-HLA genes appear to be crucial for CD development. Chromosomal region 5q31–33 has demonstrated linkage with CD in several genome-wide studies, including in our Swedish/Norwegian cohort. In a European meta-analysis 5q31–33 was the only region that reached a genome-wide level of significance except for the HLA region. To identify the genetic variant(s) responsible for this linkage signal, we performed a comprehensive single nucleotide polymorphism (SNP) association screen in 97 Swedish/Norwegian multiplex families who demonstrate linkage to the region. We selected tag SNPs from a 16 Mb region representing the 95% confidence interval of the linkage peak. A total of 1404 SNPs were used for the association analysis. We identified several regions with SNPs demonstrating moderate single- or multipoint associations. However, the isolated association signals appeared insufficient to account for the linkage signal seen in our cohort. Collective effects of multiple risk genes within the region, incomplete genetic coverage or effects related to copy number variation are possible explanations for our findings.
European Journal of Human Genetics (2007) **15,** 980–987; doi:10.1038/sj.ejhg.5201870; published online 6 June 2007

Keywords: celiac disease; 5q31–33; genetic association; autoimmunity; HLA

***Correspondence:** S Svanstrøm Amundsen, Institute of Immunology, University of Oslo, Rikshospitalet-Radiumhospitalet Medical Centre, Sognsvannsveien 20, Oslo, Norway.
Tel: +47 23073500; **Fax:** +47 23073510;
E-mail: s.s.amundsen@medisin.uio.no
Received 1 March 2007; revised 24 April 2007; accepted 9 May 2007; published online 6 June 2007

Introduction

Celiac disease (CD) is a prevalent inflammatory disorder of the small intestine with a multifactorial etiology. Patients suffering from the disease are intolerant to wheat gluten and related cereal proteins of rye and barley. There is a clear correlation between the disease and the human leukocyte antigen (HLA)-DQ status; 90–95% of all patients carry the

gene pair encoding the DQ2 heterodimer (DQA1*05/DQB1*02), whereas most of the remaining patients are positive for DQ8 (DQA1*03/DQB1*0302).[1] The DQ2 and DQ8 molecules confer susceptibility by binding and presenting gluten-derived peptides to $CD4^+$ T cells in the small intestine. The chronic inflammation is accompanied by villous atrophy and crypt hyperplasia. Although the association with HLA is strong in CD, there are strong indications that non-HLA genes also contribute to disease susceptibility. One indication is the high frequency of the DQ2 heterodimer among healthy individuals (20–30%), but the chief argument is the large difference in concordance rates between monozygotic twins and HLA-identical siblings.[2]

Several genome-wide linkage screens have pointed out non-HLA candidate regions; some regions have shown linkage in several studies whereas others have not. The chromosomal region 5q31–33 was initially pointed out in Italian cohorts,[3,4] and thereafter in three subsequent linkage studies.[5–7] In addition, a pooled analysis using raw data from four independent genome scans, including the Swedish/Norwegian data, further confirmed linkage to this region.[8] Apart from the HLA region, 5q31–33 was the only region in this study that reached genome-wide level of significance as defined by Lander and Kruglyak,[9] with maximum linkage at marker D5S640; $Zlr = 4.39$, P-value $= 6 \times 10^{-6}$. Hence, 5q31–33 can be considered to be a confirmed linked region. A substantial increase of the linkage signal with a maximum *Zlr* score of 4.6 at marker rs1972644 (P-value $= 2 \times 10^{-6}$) was evident when linkage analysis in our Scandinavian cohort was refined with more densely spaced markers (both microsatellites and single nucleotide polymorphisms (SNPs)) (Adamovic *et al*, unpublished data). Marker rs1972644 is located approximately 1.8 Mb centromeric of D5S640. In the same study, several of the included markers demonstrated nominally significant genetic association with CD. However, no susceptibility gene(s) could be identified convincingly.

To identify the CD susceptibility gene(s) located at 5q31–33, we have performed an extensive SNP association screen of the 16 Mb region which defines the 95% confidence interval (CI) (position 131.895.324–148.053.211 bp) of the linkage peak found in our Swedish/Norwegian CD cohort. From our available multiplex cohort, we included only families who had previously shown genetic linkage to the region, that is 97 families. A total of 1404 SNPs were successfully genotyped, and the genotypes of these SNPs were used for single- and multipoint association analysis.

Materials and methods

Subjects

From the previously described 106 unrelated Swedish/Norwegian CD multiplex families (families with two or more affected children) used in the initial genome-wide linkage screen,[6] we included 97 families in the current SNP screen. A more detailed description of these 106 multiplex families is available elsewhere.[10] The 97 families were selected based on the identical-by-descent (IBD) status for each individual sib-pair. The IBD status for the entire 5q31–33 region was defined using genotype information available from the previous genotyped markers used in the genome-wide linkage screen[6] and the association study (Adamovic *et al*, unpublished data). Allegro v2 software[11] was run to extract phased haplotypes that were used to determine the IBD status for each family. All families where the sib-pairs did not share at least 80% of the 5q31–33 region IBD for at least one of the chromosomes were defined as 0 IBD families and were excluded from this study. All members of the 97 selected families were genotyped in our screen, except the affected siblings of the probands from the 21 families that showed two IBD over the whole region (based on the assumption that genotyping both sibs would not provide additional information since they share identical genotypes). In total, 372 individuals were genotyped. All patients fulfilled the European Society for Paediatric Gastroenterology and Nutrition diagnostic criteria.[12]

Definition of the chromosomal region and the SNP selection

The region of interest was limited to the region representing the 95% CI of the linkage peak at marker rs1972644. The 95% CI for the linkage peak was calculated by bootstrapping from the family scores (calculated using the Allegro v2 software[11]) and was found to cover a 16 Mb interval between chromosomal position 131.895.324 and 148.053.211 bp (data not shown). In the present study, we selected 1536 tag SNPs distributed within this 95% CI, of which 1372 SNPs were successfully genotyped. Tag SNP selection was performed by first obtaining SNP genotype information for Centre d'Etude du Polymorphisme Humain individuals (Utah residents with ancestry from northern and western Europe) from the HapMap phase I database release #16b (http://www.hapmap.org/). Of the 36551 SNPs reported by Illumina to locate within the interval of interest, the HapMap phase I provided genotype information for 5516 SNPs. An assay score value for each of these SNPs was provided by the Illumina SNP service. This score value indicates the likelihood for the individual SNP of being successfully genotyped in the Illumina assay plexing system. Therefore, following the Illumina service recommendation, only SNPs with an assay score >0.6 were considered for the final tag SNP selection (4733 SNPs of the 5516 SNPs with genotype information fulfilled the assay score criteria). The genotype information for these 4733 SNPs was used for further tag SNP selection applying the tagger function implemented in the Haploview version 3.32 software[13] (using $r^2 > 0.9$ and minor allele frequency

(MAF) >0.05) and by applying the pairwise tag SNP selection method.[14] In total, 1536 tag SNPs were identified using these criteria.

SNP genotyping and quality control

Illumina bead array SNP genotyping (Illumina, San Diego, CA, USA) was performed at the Wallenberg Consortium North SNP platform (University of Uppsala, Uppsala, Sweden). Recommended control samples were included in each run. Pedcheck[15] was run to reveal Mendelian misinheritance, and the SNPs were examine for deviation from Hardy–Weinberg equilibrium. In the statistical analysis we included 30 previously genotyped SNPs (Adamovic *et al*, unpublished data) leaving a total of 1404 SNPs. The SNPs were distributed throughout the region with an average spacing of approximately 11 kb; five gaps of $\sim$110 kb were seen, 66% of the SNPs were located less than 10 kb apart, 31% within the range of 10–50 kb and 2% within the range of 50–100 kb apart.

Statistics

Based on the linkage signal seen in our population, we performed statistical calculations to predict the size of the association signals expected to be obtained in our material under the assumption of one gene effect and one founder. Using the estimated IBD sharing probabilities (z_0, z_1, z_2) generated by Genehunter v2.1[16] at the linkage peak, and assuming a single disease allele frequency (P), the penetrances (f_0, f_1, f_2) were derived by numerically solving the nonlinear equations that relate these quantities. The expected genotype and transmission configurations were calculated by application of Bayes' theorem when the disease model was fully specified through P, f_0, f_1 and f_2. Assuming the existence of one close SNP at linkage disequilibrium (LD)-distance $|D'|=1$ from the disease locus (ie no recombination observed between them) we varied the disease allele frequencies (0.02–0.3) and the associated marker allele frequency, by varying r^2 values (0.4–1).

Single-point association analysis was performed using the family-based association test, FBAT v5.5 assuming an additive risk model.[17] Haplotype comprising all SNPs across the entire region was constructed utilizing the Allegro v2 software.[11] Parental haplotypes-transmitted IBD to the affected individuals were defined as case haplotypes whereas the complementary never-transmitted parental haplotypes were defined as control haplotypes. Thereafter, these case and control haplotypes were explored by the HapMiner software v1.1.[18,19] HapMiner is a direct haplotype-mining program that utilizes a density-based clustering algorithm to assess association. In our analysis we fixed weight 1 ('the counting measure' which represents the total number of matching alleles within a given window size) and allowed for clustering of haplotypes with respect to weight 2 only ('the length measure' which represents the length of the longest continuous interval of matching alleles around a locus). In addition, we generated haplotypes of different SNP lengths: 7, 15, 19 and 27. Depending on whether the haplotype is under- or overrepresented in cases *versus* controls, either a negative or positive Z-score value is assigned (for simplicity, we refer to the absolute Z-score value ($|Z$-score$|$) in the text).

Results

Genotype quality control

Only two Mendelian inconsistencies were revealed in the data set, which indicates extremely accurate genotyping. The genotypes in the involved families were removed. SNPs that appeared nonpolymorphic ($n=108$) or that did not reach the genotype success score threshold at 90% ($n=56$) were removed from the data set before analysis. After exclusion of such SNPs, 1372 of the 1536 tag SNPs remained.

Prediction of association signal strength based on linkage results

The IBD sharing probabilities (z_0, z_1, z_2) were estimated to (0.12, 0.47, 0.41). Using these probabilities we estimated the strength of association signal we would expect to achieve if the linkage peak was caused by one gene effect and assuming one founder. In the examined models compatible with the IBD sharing probabilities the P-values for the expected outcomes in single SNP analysis varied between 10^{-17} and 10^{-6} (data not shown). The positive single- and multipoint association signals we obtained in this study were less significant (uncorrected P-value (P_{nc}-value) in the range of 0.05–0.001). Therefore, the observed strength of association signals seen in the SNP screen were not within the range of what we could expect to observe if the linkage signal were due to one gene effect, one founder and $r^2>0.4$ between the test marker and the disease locus.

Single-point association analysis

The single-point association analysis revealed 59 SNPs, which showed association with an P_{nc}-value below 0.05 (Supplementary Table 1). In total, eight SNPs displayed a significant P_{nc}-value<0.01, of these, the two most associated SNPs were located within the Jak and microtubule-interacting protein 2 (JAKMIP2) gene (rs12653715, observed/expected number of alleles (S/E(S)) = 5/16.5; $P_{nc}=0.0019$ and rs12655012, S/E(S) = 5/15.5; $P_{nc}=0.0032$; Supplementary Table 1). Except for two SNPs, all the associated SNPs ($P_{nc}<0.05$) were located within two broad regions; region 1, 133.913.277–136.848.899 ($\sim$3 Mb) and region 2, 141.997.428–147.856.177 ($\sim$6 Mb) (Figure 1).

Haplotype association analysis

We performed haplotype association analysis by HapMiner utilizing four different haplotype lengths (7, 15, 19 and

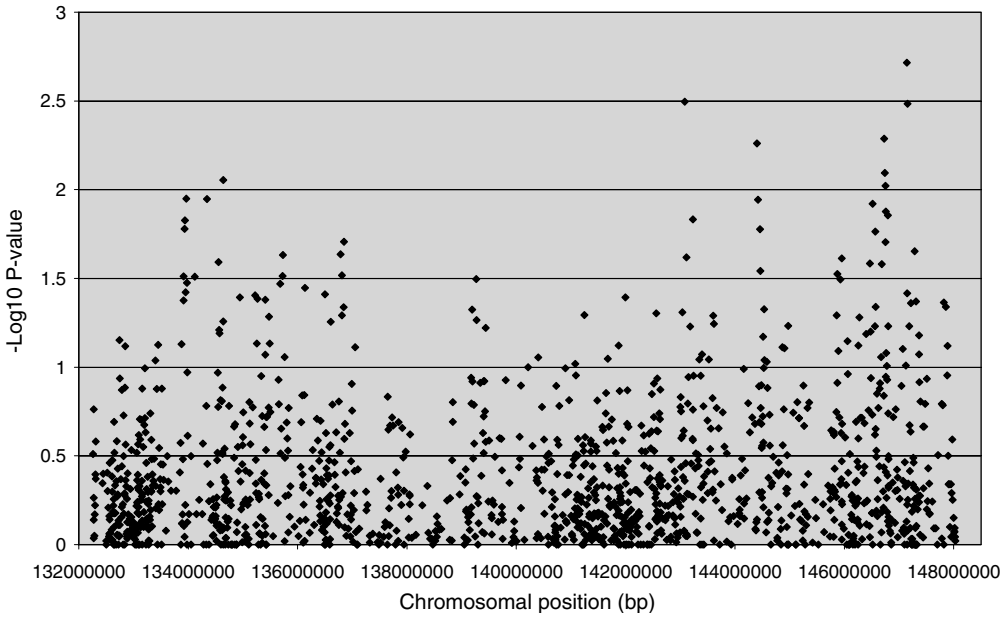

Figure 1 Single-point association for all tested single-nucleotide polymorphisms (SNPs). Each dot represents the –log 10 of the *P*-value generated with FBAT statistics for one single SNP at its chromosomal location. The chromosomal positions correspond to the human July 2003 (hg16) assembly of the University of California Santa Cruz database (http://genome.ucsc.edu/).

27). In all four analyses, numerous haplotypes displayed moderate statistical significant association with |*Z*-score| between 2.0 and 3.0 (corresponding to a P_{nc}-value within the range of 0.05–0.003) (Supplementary Table 2). In fact, only 18 haplotypes located within seven regions demonstrated stronger association than what was seen for the single SNP displaying a |*Z*-score|>3.5 (corresponding to a $P_{nc}<0.0005$) with any haplotype length (Figure 2). A detailed description of these 18 haplotypes is given in Table 1. Overall, the analyses of haplotypes of various lengths reflected the same associated haplotypes within each of the seven candidate regions. All the 59 single-point-associated SNPs were present on associated haplotypes, of which 15 SNPs colocalized to the seven candidate regions. Consequently, most of the associated single SNPs did not displayed stronger association as part of a haplotype. For instance the two strongest associated SNPs located within the JAKMIP2 gene did not display stronger association as a haplotype. Only one of the seven regions did not display association at the single SNP level (region 2 in Table 1). The strongest haplotype associations were seen within three chromosomal regions. The associated haplotypes with highest |*Z*-score| (between 3.92 and 4.04 corresponding to a statistical significance level of P_{nc}-value of 8.8×10^{-5}–5.2×10^{-5}) include region 1 covering the hypothetical protein FLJ23312 (AK26965), region 3 harboring the σ-GTPase activating protein 26 (ARHGAP26) gene and region 4 with the minor histocompatibility antigen HB-1 (Table 1).

Discussion

In this study we have performed an extensive SNP association screen testing 1404 SNPs within the 95% CI of the linkage peak at chromosome 5q31–33 previously obtained in our Swedish/Norwegian CD family cohort. We were unable to identify an association signal of any of the markers that alone could explain the linkage signal observed in the same patient cohort.

Under the assumption of a single gene effect and one founder mutation it is possible to predict the strength of the association signal which would explain the observed magnitude of the linkage peak in our material. This prediction indicates a single-point association signal with a *P*-value less than 10^{-6}. It should however be noted that this prediction is hampered with some limitations. The calculation is based on one causal gene variant located at the maximum of the linkage peak, and by assuming complete LD between the marker and disease variant ($|D'|=1$), but varying r^2 (0.4–1). Therefore, if the disease gene is located more distant from the peak, or if the risk variant is less correlated with the tested markers, the predicted signal would be weaker. Moreover, our prediction does not take into account the uncertainty which exists for the IBD sharing probabilities as this would require much more extensive calculations. Linkage of chromosomal region 5q and CD has been confirmed in several studies.[3–8] This speaks against a false-positive linkage signal in our cohort. It is possible, however, that the linkage signal in our cohort is overestimated, which would lead to a falsely inflate magnitude of the predicted

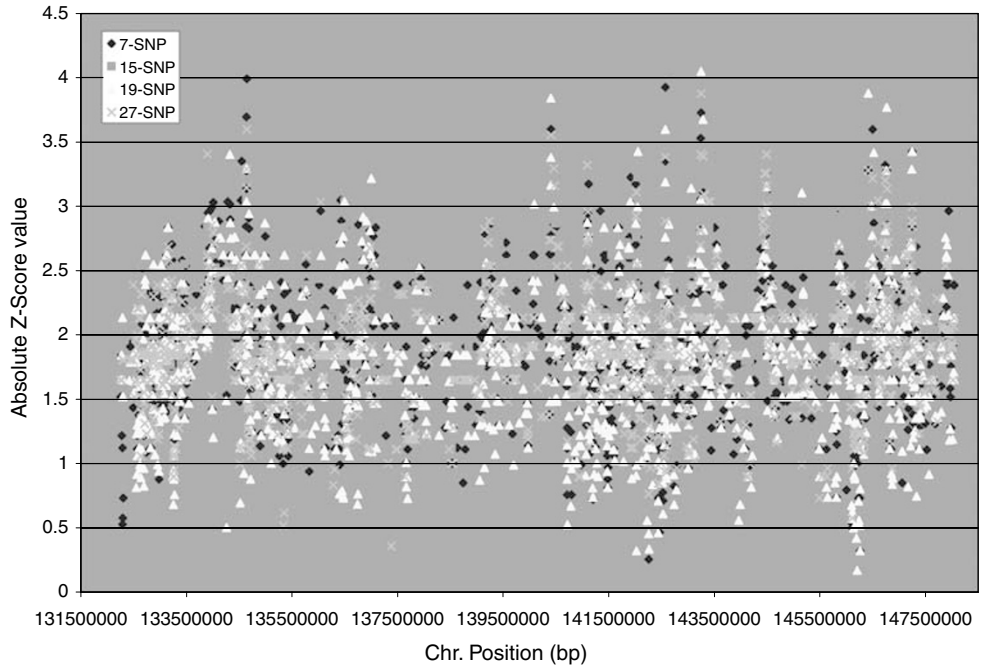

Figure 2 Associations of haplotypes of various lengths (7, 15, 19 and 27 single-nucleotide polymorphisms (SNPs)) generated by HapMiner. Each dot represents the absolute *Z*-score (|*Z*-score|) value of one haplotype. The chromosomal positions correspond to the human July 2003 (hg16) assembly of the University of California Santa Cruz database (http://genome.ucsc.edu/).

Table 1 The strongest associated haplotypes within the seven candidate regions obtained by HapMiner (|Z-score| > 3.5)

Region	*Haplotype*	*Case/ control*	Z-score
Region 1			
Haplotype 1	rs2059779-rs4976278-rs631607-rs7709668-rs2114961-rs1019737-rs1650464	22/3	3.69153
Alleles	**A-C-A-C-T-G-C**		
Haplotype 2	rs4976278-rs631607-rs7709668-rs2114961-rs1019737-rs1650464-rs1834902	22/2	3.98590
Alleles	**C-A-C-T-G-C-G**		
Haplotype 3	rs489441-rs647161-rs3749751-rs6869259-rs4246778-rs7723887-rs4976276-rs10515468-rs683756-rs2059779-rs4976278-rs631607-rs7709668-rs2114961-rs1019737-rs1650464-rs1834902-rs681139-rs652371-rs1427909-rs677394-rs4976285-rs2261050-rs2260940-rs10515469-rs3853685-rs10515470	19/2	3.59397
Alleles	A/G-A/C-T/C-A/G-A/G-C/T-A/T-**C**-A/G-A(C)-**C-A-C-T-G-C-G-T**-T/A-A/C-C/G-G/T-A/T-G/A-G/A-G/C-T/G		
Region 2			
Haplotype 1	rs151988-rs155806-rs3756325-rs31874-rs31872-rs31743-rs246584	31/8	3.59331
Alleles	**G-A-G-T-G-G-C**		
Haplotype 2	rs11954514-rs2073512-rs2531350-rs3822346-rs11748559-rs166567-rs246002-rs151988-rs155806-rs3756325-rs31874-rs31872-rs31743-rs246584-rs702390-rs31801-rs3733699-rs7722330-rs31844	33/8	3.83776
Alleles	**A**(C)-C/T-G/C-C/T-**T**(C)-**T**(A)-A/G-**G-A-G-T-G-G**(A)-**C-A**(G)-**A**(G)-**A**(T)-**T**(G)-**T**(C)		
Haplotype 3	rs3822356-rs1583005-rs778596-rs11954514-rs2073512-rs2531350-rs3822346-rs11748559-rs166567-rs246002-rs151988-rs155806-rs3756325-rs31874-rs31872-rs31743-rs246584-rs702390-rs31801-rs3733699-rs7722330-rs31844-rs246723-rs6884127-rs7700833-rs2740583-rs998794	27/6	3.55130
Alleles	A/G-A/G-A/G-A(C)-C/T-G/A-A/T-**T(C)-T(A)**-A(G)-**G-A-G-T-G-G-C-A-A(G)-A(T)-T(G)-T-G**-T/C-C-T/A-A(G)		
Haplotype 4	rs1583005-rs778596-rs11954514-rs2073512-rs2531350-rs3822346-rs11748559-rs166567-rs246002-rs151988-rs155806-rs3756325-rs31874-rs31872-rs31743-rs246584-rs702390- rs31801-rs3733699-rs7722330-rs31844-rs246723-rs6884127-rs7700833-rs2740583-rs998794- rs2907320	27/6	3.55130
Alleles	A/G-A/G-A(C)-C/T-G/A-A/T-**T(C)-T(A)**-A(G)-**G-A-G-T-G-G-C-A-A(G)-A(T)-T(G)-T-G**-T/C-C-T/A-A(G)-C(T)		

Table 1 (*Continued*)

Region	*Haplotype*	*Case/control*	Z-score
Region 3			
Haplotype 1	rs1438733-rs258799-rs258800-rs3776273-rs853173-rs258790-rs3776249	19/1	3.92090
Alleles	**A-T-G-C-A-T-C**		
Haplotype 2	rs3776307-rs7705006-rs246595-rs3776293-rs246607-rs246599-rs1438733-rs258799-rs258800-rs3776273-rs853173-rs258790-rs3776249-rs853165-rs2107622-rs258759-rs258770-rs3776230-rs867924	19/2	3.59397
Alleles	**T**(G)-A/G-**G**(T)-**G**(A)-**C**(T)-A/G-T/C-**A-T-G-C-A-T**(C)-**C-G**(A)-A/G-**C**(T)-**T**(C)-**T**(C)		
Region 4			
Haplotype 1	rs918377-rs161557-rs918378-rs880814-rs7718152-rs161553-rs1421776	26/48	−3.52671
Alleles	**A-C-G-T-G-C-T**		
Haplotype 2	rs161557-rs918378-rs880814-rs7718152-rs161553-rs1421776-rs1363134	30/54	−3.72212
Alleles	**C-G-T-G-C-T-C**		
Haplotype 3	rs325247-rs325238-rs4912658-rs4912950-rs3805462-rs918377-rs161557-rs918378-rs880814-rs7718152-rs161553-rs1421776-rs1363134-rs161551-rs10515527-rs315183-rs315171-rs315199-rs318373	35/62	−4.04446
Alleles	**T(C)-C(T)-T(C)-T(G)-A(G)-A(G)-C(T)-G-T-G-C-T-C(G)-C(T)-T(A)-C-G-G(A)-C(T)**		
Haplotype 4	rs325238-rs4912658-rs4912950-rs3805462-rs918377-rs161557-rs918378-rs880814-rs7718152-rs161553-rs1421776-rs1363134-rs161551-rs10515527-rs315183-rs315171-rs315199-rs318373-rs312561	36/60	−3.67234
Alleles	**C**(T)-**T**(C)-**T**(G)-**A**(G)-**A**(G)-**C**(T)-**G**(A)-**T-G-C-T-C-C**(T)-**T**(A)-**C-G-G**(A)-**C**(T)-**C**		
Haplotype 5	rs867924-rs7715450-rs2059129-rs188975-rs325247-rs325238-rs4912658-rs4912950-rs3805462-rs918377-rs161557-rs918378-rs880814-rs7718152-rs161553-rs1421776-rs1363134-rs161551-rs10515527-rs315183-rs315171-rs315199-rs318373-rs312561-rs315198-rs10515530-rs312563	28/53	−3.87224
Alleles	**C**(T)-C/T-T/C-T/C-**T**(C)-**C**(T)-**T**(C)-**T**(G)-**A**(G)-**A**(G)-**C-G-T-G-C-T-C-C**(T)-**T-C-G-G-C**(T)-**C**-T(A)-**G-C**(A)		
Region 5			
Haplotype 1	rs31902-rs389586-rs7379893-rs4913020-rs31913-rs248017-rs419844-rs11167889-rs10515546-rs953996-rs1363643-rs1541664-rs1422839-rs9324976-rs10515548	58/26	3.54899
Alleles	A/G-**T-T-A-A-C-C-C-C-T-G-G-C-A-A**(G)		
Region 6			
Haplotype 1	rs1383169-rs6580448-rs1480149-rs1480150-rs1480159-rs1480160-rs1480161	31/8	3.59331
Alleles	**C-T-G-A-C-G-C**		
Region 7			
Haplotype 1	rs1019916-rs1010966-rs1368285-rs4705159-rs7703154-rs918799-rs4705161-rs1424293-rs11954789-rs12659468-rs6884181-rs918798-rs2400295-STKEXON7_CT-rs918797-rs6866098-rs4705167-rs10515584-rs10515585	87/107	−3.87382
Alleles	A/G-A/G-G/T-G/A-A/C-**A**(T)-A/G-**A-T-A-C-C-G**(A)-C/T-**A**(G)-**T**(A)-**C**(T)-**C**(T)-A/G		
Haplotype 2	rs1010966-rs1368285-rs4705159-rs7703154-rs918799-rs4705161-rs1424293-rs11954789-rs12659468-rs6884181-rs918798-rs2400295-STKEXON7_CT-rs918797-rs6866098-rs4705167-rs105150584- rs10515585-rs723698	88/107	−3.76539
Alleles	A/G-G/T-G/A-A/C-**A**(T)-A/G-**A-T-A-C-C-G**(A)-C/T-**A**(G)-**T**(A)-**C**(T)-**C**(T)-A/G-T/C		

When two alternative alleles existed on the haplotype within a cluster, the most frequent allele is always depicted first. Alternative alleles are further separated by slash when both of the two alternative alleles occur quite frequently on the haplotypes, whereas the allele in parenthesis denotes an allele that only occurs on a few haplotypes within the cluster (the rare variant). The column case/control denotes the number of case *versus* control haplotypes present in the cluster. $|Z\text{-score}| > 3.5$ corresponds to a P_{nc}-value < 0.0005.

association signals. Despite these constrains in our predictions, there appears to be a striking discrepancy between the predicted association signal and the signals we observe. One obvious explanation for this discrepancy could be that this region contains several susceptibility genes that collectively contribute with moderate risk to CD susceptibility; effects that with our limited sample size would be difficult to detect.

Inability to detect 'the true association signal' could be another reason for not achieving a strong association signal in this study. The HapMap project provides a description of the LD composition of the human genome as to minimize the number of tag SNPs needed for association studies. It is important to note that the HapMap phase I provide genotype data mostly for common SNPs with MAF > 0.05 selected in accordance with the 'common disease/common

variant' hypothesis. Due to this underlying selection bias of high-frequent SNPs, tag SNPs selected from HapMap data have shown to have limited capacity to tag rare variants (variants with MAF < 5%).[20,21] Possible risk factors caused by low frequent variants would thus easily be overlooked. However, the most important parameters that influence how well tag SNPs cover genetic variation are SNP density and the pattern of LD. Tag SNPs selected in high-LD regions are robust to variable marker density whereas low-LD regions are not.[20,22] In the phase I project, genotype information of SNPs with a density of one SNP per 5 kb is available. This spacing has been estimated and later shown to enable sufficient coverage of approximately 75–80% of the genome if tag SNPs are selected using the pairwise algorithm and $r^2 = 0.8$.[20,23] Marker density and local LD pattern do therefore not make the transferability straight forward, and the tagging performance would be variable between different chromosomal regions. In our study, tag SNPs were selected from the, at that time, available phase I HapMap data (release #16b) by applying the pairwise algorithm, $r^2 > 0.9$ and MAF > 0.05. By only including relatively common SNPs, we ensured sufficient power in the study. It is also worth to note that HapMap provided genotypes for about 15% (ie 5516 SNPs) of the total amount of SNPs reported within the 95% CI of interest. The HapMap data do most likely therefore not provide coverage of all the genetic variation within our region of interest.

The apparent discrepancy between linkage and association signals as we experienced in this study could also be related to DNA copy number variation (CNV). CNVs represent a huge source of genetic diversity that in many cases will have functional implications.[24] CNVs, however, are often located in regions of complex genomic structure that are poorly covered by genotyped SNPs.[24,25] There is often a low LD between many CNVs and SNPs, which is reducing the likelihood of detecting a disease associated CNV in an association study. Furthermore, SNPs located within CNVs may give incorrect genotyping thereby perturbing proper analysis.[26] It is interesting that one of the strongest haplotype associations (the haplotypes in region 2 shown in Table 1) is located in the ~5 Mb gap where no single SNP associations were seen. This ~5 Mb region contains several annotated segmental duplications. Whether the CNVs located within this region have had any influence in our SNP association analysis is however unknown.

In common diseases where several genes are thought to contribute with modest risk to disease susceptibility, the use of too conservative correction procedures would lead to increased type II errors rather than facilitate identification of the real disease variant (especially in moderate-sized sample sets as is used in most studies).[27] As the single-point association signals we obtained were predominantly of moderate significance, correction for multiple testing would clearly not assist discrimination between true- and false-positive association signals. The number of associated SNPs did not exceed the number of false positives expected by chance for this number of tested SNPs (both at the 5 and 1% significance level), which in theory means that all associations could potentially be false positive findings. Single-point association studies are statistically weak and have clear limitations. Haplotype association analyses are more powerful as they can tag the founder chromosome more efficiently. This type of analysis is often done to provide additional support for the single marker associations. Our haplotype association analysis revealed seven regions with increased association signals. Except for one region, they all harbored markers with single-point association signals. Notably, however none of the haplotype association signals would have remained significant after correction for multiple testing. Hence, this analysis did not bring clarity into identifying the underlying causal variant(s). Whether any or more of these regions are involved in CD genetics should be scrutinized by replication attempts.

In conclusion, our comprehensive association analysis of a region with one of the strongest linkage signals seen in CD has failed to establish markers that demonstrate convincing association with the disease. Collective effects of multiple risk genes within the region, incomplete genetic coverage or effects related to CNV appear to be possible explanations for our findings.

Acknowledgements

This work was supported by grants from the Research Council of Norway, the Swedish Medical Research Council and the Swedish Research Council. We thank all the families participating in the study as well as Britt-Marie Käck and The Celiac Society in Sweden for help with collecting families and blood samples. We also thank Ann-Christine Syvänen and the Wallenberg Consortium North (WCN) SNP Technology Platform at Uppsala University, Sweden and the Swegene Genomics and Bioinformatics Core Facilities in Göteborg, Sweden.

References

1 Sollid LM: Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol* 2002; **2**: 647–655.
2 Greco L, Romino R, Coto I *et al*: The first large population based twin study of coeliac disease. *Gut* 2002; **50**: 624–628.
3 Greco L, Corazza G, Babron MC *et al*: Genome search in celiac disease. *Am J Hum Genet* 1998; **62**: 669–675.
4 Greco L, Babron MC, Corazza GR *et al*: Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families. *Ann Hum Genet* 2001; **65**: 35–41.
5 Liu J, Juo SH, Holopainen P *et al*: Genomewide linkage analysis of celiac disease in Finnish families. *Am J Hum Genet* 2002; **70**: 51–59.
6 Naluai AT, Nilsson S, Gudjonsdottir AH *et al*: Genome-wide linkage analysis of Scandinavian affected sib-pairs supports presence of susceptibility loci for celiac disease on chromosomes 5 and 11. *Eur J Hum Genet* 2001; **9**: 938–944.
7 Zhong F, McCombs CC, Olson JM *et al*: An autosomal screen for genes that predispose to celiac disease in the Western counties of Ireland. *Nat Genet* 1996; **14**: 329–333.

8 Babron MC, Nilsson S, Adamovic S *et al*: Meta and pooled analysis of European coeliac disease data. *Eur J Hum Genet* 2003; **11**: 828–834.
9 Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241–247.
10 Gudjonsdottir AH, Nilsson S, Ek J, Kristiansson B, Ascher H: The risk of celiac disease in 107 families with at least two affected siblings. *J Pediatr Gastroenterol Nutr* 2004; **38**: 338–342.
11 Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A: Allegro version 2. *Nat Genet* 2005; **37**: 1015–1016.
12 Walker-Smith J, Guandalini S, Schmitz J, Schmerling D, Visakorpi J: Report of working group of European Society of Paediatric Gastroenterology and Nutrition: revised criteria for diagnosis of coeliac disease. *Arch Dis Child* 1990; **65**: 909–911.
13 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
14 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
15 O'Connell JR, Weeks DE: PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998; **63**: 259–266.
16 Markianos K, Daly MJ, Kruglyak L: Efficient multipoint linkage analysis through reduction of inheritance space. *Am J Hum Genet* 2001; **68**: 963–977.
17 Horvath S, Xu X, Laird NM: The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet* 2001; **9**: 301–306.
18 Li J, Zhou Y, Elston RC: Haplotype-based quantitative trait mapping using a clustering algorithm. *BMC Bioinformatics* 2006; **7**: 258.
19 Li J, Jiang T: Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics* 2005; **21**: 4384–4393.
20 Montpetit A, Nelis M, Laflamme P *et al*: An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet* 2006; **2**: e27.
21 Zeggini E, Rayner W, Morris AP *et al*: An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat Genet* 2005; **37**: 1320–1322.
22 Ke X, Durrant C, Morris AP *et al*: Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum Mol Genet* 2004; **13**: 2557–2565.
23 International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
24 Redon R, Ishikawa S, Fitch KR *et al*: Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
25 Kehrer-Sawatzki H: What a difference copy number variation makes. *Bioessays* 2007; **29**: 311–313.
26 Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ: Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004; **36**: 861–866.
27 Shephard N, John S, Cardon L, McCarthy MI, Zeggini E: Will the real disease gene please stand up? *BMC Genet* 2005; **6** (Suppl 1): S66.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)