Presymptomatic Diagnosis of Celiac Disease in Predisposed Children: The Role of Gene Expression Profile

*†Martina Galatola, *†Donatella Cielo, *Camilla Panico, *Pio Stellato, ‡Basilio Malamisura, *Lorenzo Carbone, †§Carmen Gianfrani, *†Riccardo Troncone, *†Luigi Greco, and *†Renata Auricchio

ABSTRACT

Objective: The prevalence of celiac disease (CD) has increased significantly in recent years, and risk prediction and early diagnosis have become imperative especially in at-risk families. In a previous study, we identified individuals with CD based on the expression profile of a set of candidate genes in peripheral blood monocytes. Here we evaluated the expression of a panel of CD candidate genes in peripheral blood mononuclear cells from at-risk infants long time before any symptom or production of antibodies.

Methods: We analyzed the gene expression of a set of 9 candidate genes, associated with CD, in 22 human leukocyte antigen predisposed children from at-risk families for CD, studied from birth to 6 years of age. Nine of them developed CD (patients) and 13 did not (controls). We analyzed gene expression at 3 different time points (age matched in the 2 groups): 4–19 months before diagnosis, at the time of CD diagnosis, and after at least 1 year of a gluten-free diet. At similar age points, controls were also evaluated.

Results: Three genes (*KIAA*, *TAGAP* [T-cell Activation GTPase Activating Protein], and *SH2B3* [SH2B Adaptor Protein 3]) were overexpressed in patients, compared with controls, at least 9 months before CD diagnosis. At a stepwise discriminant analysis, 4 genes (*RGS1* [Regulator of G-protein signaling 1], *TAGAP*, *TNFSF14* [Tumor Necrosis Factor (Ligand) Superfamily member 14], and *SH2B3*) differentiate patients from controls before serum antibodies production and clinical symptoms. Multivariate equation correctly classified CD from non-CD children in 95.5% of patients.

Conclusions: The expression of a small set of candidate genes in peripheral blood mononuclear cells can predict CD at least 9 months before the appearance of any clinical and serological signs of the disease.

Key Words: celiac disease, celiac disease first-degree relatives, gene expression, presymptomatic diagnosis, risk factors

(*JPGN* 2017;65: 314–320)

What Is Known

- A panel of genes associated to celiac disease have a confirmed role in the pathogenesis of gluten-induced immune response.
- The expression of 12 genes clearly identify celiac from non-celiac disease subjects in mucosa and in peripheral blood monocytes.

What Is New

- At least 9 months before any presence of antibodies in the serum or clinical signs, the expression of a small set of celiac disease candidate genes in peripheral blood mononuclear cells predict the development of celiac disease in at-risk infants.

The estimated prevalence of celiac disease (CD) in first-degree relatives is as high as 10%, which is 10 times higher than in the general population (estimated around 1%) (1–4). Concordance in monozygotic twins is higher than 80%; there is a strong genetic component, which is unusual for a multifactorial disease (5). Recurrence in families is an important source of new patients and an early diagnosis in these at-risk families become an urgent challenge (3,4).

Recently the European Society for Paediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) recommended to limit the use of small intestinal biopsy for the diagnosis of CD, opening the perspective of non-invasive diagnosis in child or adolescent with typical CD symptoms, tissue-transglutaminase IgA higher than 10 times the reference value and presence of human leukocyte antigen (HLA)-DQ2/-DQ8 genes (6).

Unfortunately, the molecular features of CD genetics are still not fully elucidated. Although 95% of patients carry HLA-DQ2/-DQ8, this trait explains just 35% to 40% of the genetic variance (7,8). Repeated genome-wide association studies identified up to 57 single-nucleotide polymorphorphisms (SNPs) associated to CD, but they accounted for just a small fraction (approximately 6.5%) of the heredity of CD; together with HLA they explained about half the heredity (9–18).

Given the complexity of CD, the genetic data available are not yet sufficient for disease prediction. In an attempt to obtain functional data, we previously explored the expression of a set of CD-associated genes in peripheral blood monocytes (PBMs) easily available, rather than in the intestinal mucosa tissue (19). Indeed, Ontiveros et al (20) showed that plasma levels of inflammatory cytokines in whole blood collected after acute (3 days) oral gluten

Received August 9, 2016; accepted January 10, 2017.

From the *Department of Translational Medical Science, University of Naples Federico II, Naples, Italy, the †European Laboratory for Food-Induced Disease (ELFID), University of Naples Federico II, Naples, Italy, the ‡Pediatric Unit, University Hospital of Salerno, Cava de' Tirreni, and the §Institute of Protein Biochemistry-CNR, Naples, Italy.

Address correspondence and reprint requests to Martina Galatola, M.Sc, PhD, Department of Translational Medical Science, University of Naples Federico II, Via S. Pansini 5, 80131 Naples, Italy (e-mail: martinagalatola@hotmail.it).

This work was funded by the PREVENT-CD project: EU-FP6-2005-FOOD4B-contract no. 036383. The authors thank European Laboratory for Food-Induced Disease (ELFID). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (www.jpgn.org).

The authors report no conflicts of interest.

DOI: 10.1097/MPG.0000000000001519

challenge could differentiate patients with CD from controls who have adopted gluten-free diet (GFD) for at least 6 months.

These results encouraged us to explore the adoption of peripheral blood mononuclear cells (PBMCs) as a tool for diagnosis of CD instead of intestinal biopsy. A multicenter randomized study has been performed trying to reduce the risk of CD in children with a first-degree relative affected by CD. This study aimed to identify risk factors associated with CD (breast-feeding, time and quantity of gluten introduction, other environmental and genetic factors) by following them from birth to 5 years of age (21). This study allowed collecting serial blood samples from a subset of infants at risk for CD at fixed time points from birth up to 6 years of age, developing an adequate structure for a longitudinal study.

The aim of the present study is to longitudinally evaluate the expression of a panel of CD-candidate genes directly in PBMCs, in genetically predisposed children with a first-degree relative affected by CD, before the appearance of any clinical or serological markers of CD.

PATIENTS AND METHODS

Population Enrolled in the Study

Participants were infants from the Italian cohort of newborns from at-risk families (ie, families with at least a case of CD), bearing HLA-DQ2 or -DQ8, monitored from birth to 6 years of age, enrolled during the PREVENT-CD study (21,22). Subjects were monitored clinically and serologically (anti-tissue-transglutaminase antibody [anti-tTG]) every 3 to 6 months from birth to 6 years old. In children who developed anti-tTG antibodies, CD was confirmed by a small intestinal biopsy (21,22). We analyzed 9 children who developed CD at a median age of 30 months, and 13 age- and sex-matched controls (children from the same cohort who did not develop CD up to 6 years of age). Among infants who developed CD, 5 were asymptomatic, 4 had symptoms, 3 showed growth retardation, of which one was associated to diarrhea, the 4th had recurrent abdominal pain.

In the PREVENT-CD cohort studied in our center, 24 children became celiac before 6 years of age. Unfortunately, the number of samples of sufficient size and quality required for this study was lower because of progressive deterioration of samples for the long-time span from recruitment to the disclosure of the double-blind design of the study (>6 years). Therefore, only for the 9 patients here analyzed, we were able to use the biological material in all time points established in this work. No selection bias could be envisaged in this subsample.

Management of Samples

Genotyping was performed on DNA obtained from cord blood. Gene expression was assessed on blood sample of 9 subjects (CD patients) in 3 different time points: before antibody production (Time 1), which ranged from 4 to 13 months of age, median 12 months; at the age of anti-tTG positivity and confirmed diagnosis (Time 2), which ranged from 18 to 50 months, median 30 months of age; and at least 1 year after starting GFD (Time 3), which ranged from 36 to 48 months, median 36 months of age. Blood samples were collected at similar time points from 13 controls at Time 1, which ranged from 5 to 19 months of age, median 7 months; at Time 2, which ranged from 18 to 50 months, median 18 months of age; and at Time 3, which ranged from 23 to 62 months, median 36 months of age. None of the infants neither in the group who became celiac nor in the group that did not developed CD produced any measurable level of anti-tTG at the time T1 when Gene Expression was first examined.

In all CD patients, and in 6 of the 13 controls, a duodenal biopsy was performed according to PREVENT-CD protocol based on clinical or serological suspicion. Of these 6 controls, 1 had diarrhea, whereas the other 5 were asymptomatic but showed at least 1 positive value of CD-associated antibody (anti-gliadin antibodies [AGA], anti-tissue transglutaminase [anti-tTG], endomysial antibodies [EMA]) (22). In all controls, the diagnosis of CD up to 6 years of age was excluded. Features of each subject enrolled in this study are listed in Table 1.

Ethics Statement

The study protocol was approved by the ethics committee of the University of Naples Federico II. Biopsies were taken during routine hospital admission requested for diagnostic purposes: parents or guardian gave their informed consent about the endoscopic procedures and the biopsy. Patients did not undergo specimen sampling over and above those requested by the routine diagnostic procedures according to the ESPGHAN guidelines (6).

Isolation of Peripheral Blood Mononuclear Cells

Blood was obtained while children were in stable clinical conditions without any acute illness in the last 3 weeks. Blood samples (4–5 mL) were collected and PBMCs were isolated by Ficoll gradient according to the manufacturer's protocol. PBMCs were frozen and stored incryoprotective media containing 10% dimethylsulfoxide and fetal bovine serum.

RNA Extraction and Gene Expression Studies

Total RNA was extracted from PBMCs using TRIZOL Reagent (Life Technologies, Foster City, CA). The quantity of RNA was measured using a Nanodrop spectrophotometer and RNA quality was analyzed using agarose gel electrophoresis in Tris/Borate/ethylenediaminetetraacetic acid buffer. The RNA (1 μg) was reverse transcribed into cDNA using the High Capacity cDNA Reverse Transcription kit (Life Technologies). The experiments were performed with a 7900HT Fast Real-Time polymerase chain reaction (PCR) system using the TaqMan Gene Expression Assay (Life Technologies), and approximately 40 ng of cDNA according to the manufacturer's instructions. The relative expression of each gene was obtained using the ΔΔCt method, and normalized to an endogenous housekeeping gene Glucoronidase (*GUSb*). We used GUSb as reference gene after it was identified as the most stable reference gene of 5 candidates (*β-actin*, *B2M*, *GAPDH*, *GUSb*, and *HPRT1*). In particular, the relative quantification (RQ) was calculated normalizing each sample against the mean value of T2 samples of unaffected subjects. The SDS software (ABI, version 1.4 or 2.4, Life Technologies, Foster City, CA) was used to analyze raw data, and statistical analysis was performed on GraphPad Prism 5.01H. Gene expression experiments were conducted according to minimum information for publication of quantitative real-time PCR experiments (MIQE) guidelines (*http://www.gene-quantification.de/miqe-bustin-et-al-clin-chem-2009.pdf*). The candidate genes evaluated, and their respective assays used for the analysis are listed in Supplemental Digital Content, Table 1, *http://links.lww.com/MPG/A885*.

Genotyping

Patients were genotyped for a set of candidate genes as described elsewhere and for HLA (23,24). Patients and controls were grouped into 5 HLA-haplotype classes, as reported previously (24). Nine non-HLA SNPs, located in the candidate genes and listed in the Supplemental Digital Content, Table 2 (*http://links.lww.com/MPG/A885*), were analyzed by genotyping assay.

TABLE 1. Individual characteristics of patients enrolled in the study

UIC	Sex	Status	T1 Before-CD	Age (mo) T2 At CD-diagnosis	T3 At-GFD	Histology*	At T1† (U/mL)	Anti-tTG T2† (U/mL)	T3† (U/mL)	HLA-DQ	Proband in family
P1	Male	CD	9	18	36	M3a/b/c	0.50	>100.00	5.20	DQ2.5/X	Mother
P2	Male	CD	13	30	47	M3a/b/c	0.10	67.10	0.90	DQ2.5/X	Father
P3	Female	CD	9	20	36	M3a/b/c	0.10	>100.00	0.10	DQ2.5/DQ7	Mother
P4	Female	CD	12	32	48	M3a/b/c	0.10	36.00	0.20	DQ2.2/X	Father + Sib
P5	Female	CD	12	50	NA	M1	0.10	>100.00	0.10	DQ2.5/DQ2.2	Mother
P6	Female	CD	12	25	36	M3a/b/c	0.10	8.00	0.10	DQ2.5/X	Sibs
P7	Female	CD	12	37	NA	M3a/b/c	0.10	>100.00	0.80	DQ2.5/DQ2.2	Mother
P8	Female	CD	9	20	37	M3a/b/c	0.10	>100.00	0.80	DQ2.5/DQ2.2	Father + Sib
P9	Female	CD	4	35	NA	M3a/b/c	0.10	>100.00	0.10	DQ2.2/DQ7	Sibs
P10	Male	No-CD	6	18	NA	M0	0.10	0.10	0.10	DQ2.5/DQ7	Sibs
P11	Male	No-CD	9	24	48	M0	0.10	0.10	0.10	DQ2.2/X	Mother
P12	Female	No-CD	9	25	62	M0	0.10	0.10	0.10	DQ2.5/DQ2.2	Mother+ Sib
P13	Female	No-CD	4	25	48	M1	0.10	0.20	0.50	DQ8/X	Mother
P14	Female	No-CD	6	19	27	M1	0.10	4.50	1.00	DQ2.2/DQ7	Father
P15	Female	No-CD	19	35	42	M1	0.10	4.80	0.60	DQ2.2/DQ7	Sibs
P16	Male	No-CD	9	18	35	NA	0.10	0.10	0.10	DQ2.2/X	Mother+ Sib
P17	Female	No-CD	5	12	23	NA	0.10	0.20	0.10	DQ2.5/DQ7	Mother
P18	Female	No-CD	8	12	23	NA	0.10	0.10	0.10	DQ2.2/DQ7	Sibs
P19	Female	No-CD	9	20	36	NA	0.10	0.10	0.10	DQ8/X	Sibs
P20	Male	No-CD	6	12	23	NA	0.10	0.10	0.10	DQ2.2/X	Mother
P21	Female	No-CD	9	17	61	NA	0.10	0.10	0.10	DQ2.5/DQ8	Sibs
P22	Male	No-CD	6	18	36	NA	0.10	2.40	0.30	DQ2.2/DQ2.2	Sibs

anti-tTG = anti-tissue transglutaminase; CD = celiac disease; GFD = gluten-free diet; T1 = before antibody production; T2 = age at appearance of antibodies and mucosal damage; T3 = at least 1 year after diagnosis on GFD; NA = not available; UIC = unique identification code.

*For the diagnosis of CD Marsh classification has been applied, all biopsied controls have a normal duodenal mucosa with no atrophy (Marsh lesion stage M0-M1).

†Reference values of anti-tTG: Negative: <4.0 U/mL; Doubt: >4.0 and <10.0 U/mL; Positive: >10.0 U/mL.

The SNPs were genotyped using TaqMan technology (Life Technologies) using TaqMan SNP Genotyping Assays on a 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster City, CA); the final volume was 15 μL, containing master mix, TaqMan assays and about 60 ng of genomic DNA template.

Statistical Analysis

Non-parametric rank sum test was adopted to compare gene expression because of the small sample size. Frequencies were compared using the χ^2-test, with probability for null hypothesis equal to 0.05. A discriminant analysis was performed to estimate the contribution of the genotype and or the expression of each gene to distinguish patients from controls. Wilks' lambda provides an estimate of the cumulative discriminating capacity between 2 groups produced by the multivariate combination of variables, ranging from 1 = complete overlap to 0 complete distance.

By multiplying the standardized value of each variable included in the stepwise discriminant equation by its respective regression coefficient we obtain a discriminant score D, which express the spatial location of each individual to the right or to the left: it is then possible to assign the individual to the group of patients or the group of controls, only on the basis of the genetic data, blind to the final conclusive diagnosis. Hence, the percentage of correct classification may be obtained. Statistical analyses were performed with the SPSS 17.0 (SPSS Inc, Chicago, IL).

RESULTS

Human Leukocyte Antigen Genotyping

Due to the a priori selection of genetically predisposed children, the distribution of HLA genotypes did not differ significantly between patients and controls. Three patients but no controls are, however, homozygote for DQ2. Five of the 13 children in the latter group have an HLA low-risk class (DQ8 or heterozygote for DQ2).

Genotyping of Candidate Genes

Although there were few chances to observe significant differences between the 2 groups, because of the small sample size, 3 of 9 candidate genes explored showed a different distribution of the risk allele. For ReticuloEndotheliosis Viral Oncogene homolog (*cREL*) and SH2B Adaptor Protein 3 [*SH2B3*] genes the risk allele ''A'' is more frequent in patients, than in controls, whereas for Tumor Necrosis Factor Receptor Superfamily member 14 (TNFRSF14) the ''A'' allele is less frequent (Supplemental Digital Content, Table 2 and Fig. 1, *http://links.lww.com/MPG/A885*).

Gene Expression Analysis of Candidate Genes

Time 1 (Before Symptoms or Anti-tTG Production)

KIAA, T-cell Activation GTPase Activating Protein (*TAGAP*), and SH2B3 expression was higher in patients before CD diagnosis, with differences of RQ of 0.6-, 0.5-, and 0.7-fold higher in patients versus controls, respectively. The expression of Tumor Necrosis Factor (Ligand) Superfamily member 14 (*TNFSF14*) and Regulator of G-protein signaling 1 (*RGS1*) was higher in controls (Fig. 1A), with differences of RQ level of 1.5, and 1.2-fold higher in controls versus patients. The differences were statistically significant for TAGAPSH2B3 and TNFSF14 (Mann-Whitney U test, $P < 0.05$), but not significant for KIAA and RGS1 (Supplemental Digital Content, Table 3 and Fig. 1, *http://links.lww.com/MPG/A885*).

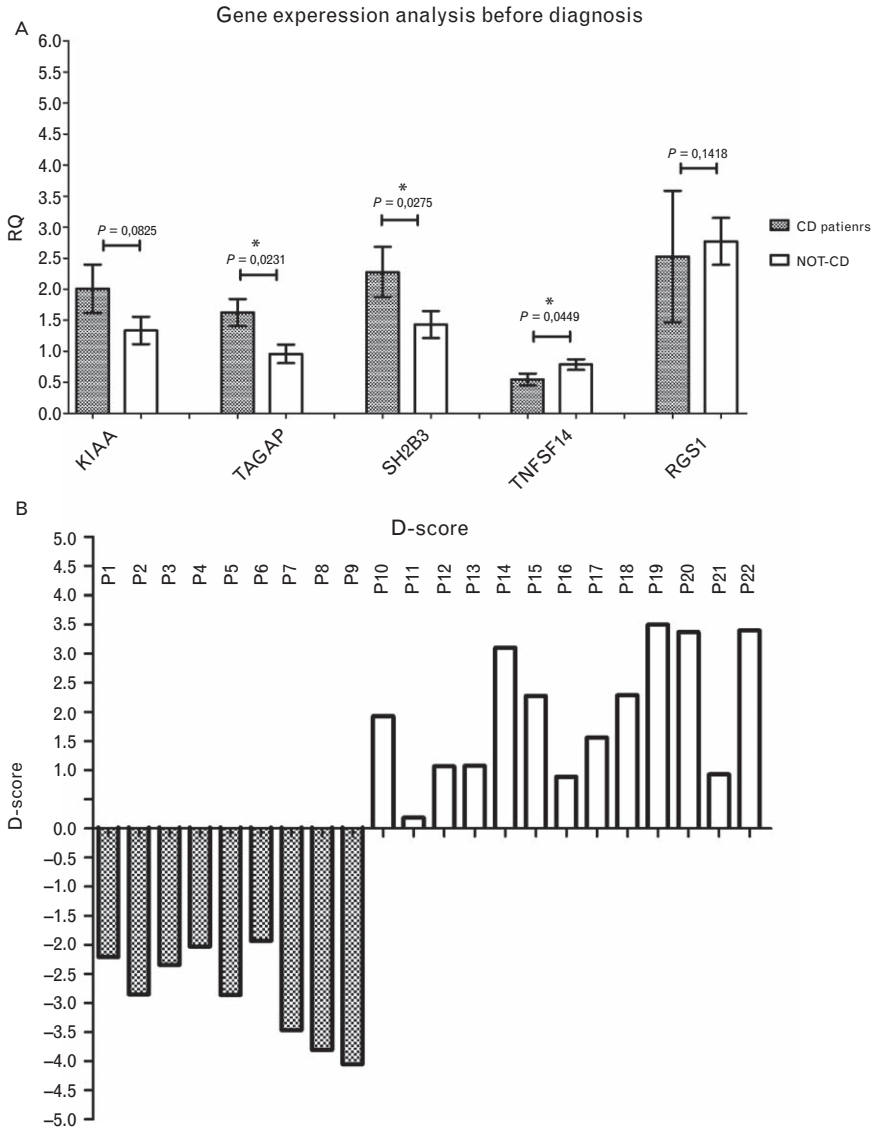

FIGURE 1. (A) mRNA expression of candidate genes in PBMC at least 9 months before the diagnosis (Time 1). KIAA, TAGAP, and SH2B3 expression was higher in 9 patients than in 13 controls, with the differences of RQ level of 0.6, 0.5, and 0.7-fold change, respectively. *TNFSF14* and RGS1 expression was lower in patients than in controls, with differences of RQ level of 1.5, and 1.2. *Mann-Whitney *U* test $P < 0.05$. (B) Distribution of the discriminant score in patients and in controls. The D-score clearly separated the 2 groups of subjects evaluated. Only 9 patients had a negative D-score. PBMC = peripheral blood mononuclear cells; RGS1 = Regulator of G-protein signaling 1; TAGAP = T-cell Activation GTPase Activating Protein; TNFSF14 = Tumor Necrosis Factor (Ligand) Superfamily member 14.

Time 2 (Diagnosis)

At moment of CD diagnosis, only 2 genes, *SH2B3* and *TNFRSF14*, showed significant higher expression in patients versus controls, with a differences of RQ level of 0.5, and 0.6-fold (Mann-Whitney test $P = 0.02$ and 0.04, respectively) (Supplemental Digital Content, Fig. 1, *http://links.lww.com/MPG/A885*).

Time 3 (1 Year After Diagnosis)

The expression of the genes examined did not differ significantly between the 2 groups at time point 3 (data not shown).

Multivariate Analysis of Candidate Genes Genotyping

Considering the observed differences in genotype distribution in patients versus controls (Supplemental Digital Content, Table 2, *http://links.lww.com/MPG/A885*), we explored the combined profile of the 9 candidates' genes in the patients compared with controls. In a stepwise multivariate discriminant analysis, 5 genotypes significantly discriminated between CD and controls, with a correct classification of 91% (Table 1 panels A and B).

TABLE 2. Discriminant analysis of candidate genes genotype

Step	Genotype	Wilks' lambda	Variance ratio F Statistic	*P*
(A) Stepwise discriminat analysis of genotypes				
1	SH2B3	0.604	13.091	0.002
2	RGS1	0.48	10.292	0.001
3	TAGAP	0.429	7.973	0.001
4	cREL	0.35	7.908	0.001
5	LPP	0.307	7.233	0.001

		Predicted group		
	Status	CD	Not CD	Total
(B) Classification by discriminat equation of genotypes				
Original group	CD	8 (89%)	1 (11.1%)	9
	Not CD	1 (7.7%)	12 (92.3%)	13
Overall correct classification = 91%				

Step	Genotype	Wilks' lambda	Variance ratio F Statistic	*P*
(C) Stepwise discriminat analysis of gene expression at T1				
1	TAGAP	0.742	6.938	0.016
2	TNFSF14	0.542	8.019	0.003
3	SH2B3	0.446	7.453	0.002
4	RGS1	0.366	7.377	0.001

		Predicted group		
	Status	CD	Not CD	Total
(D) Classification by discriminat equation of gene expression at T1				
Original group	CD	9 (100%)	0 (0%)	9
	Not CD	1 (7.7%)	12 (92.3%)	13
Overall correct classification = 95%				

(A) Five genes (*SH2B3*, *RGS1*, *TAGAP*, *cREL*, and *LPP*) were selected for discriminating capacity, with a *P* value <0.001. (B) Results of the prediction analysis: 92.3% of controls and 89% patients were correctly classified. Panels 1 (C and D) Discriminant analysis of candidate genes gene expression before diagnosis (Time 1). (C) Four genes significantly contributed to lower Wilks' lambda ($P < 0.001$); *TAGAP*, *TNFSF14*, *SH2B3*, and *RGS1* were selected for discriminating capacity. (D) Results of the prediction analysis: 92.3% of controls and 100% of patients were correctly classified.

CD = celiac disease; LPP = Lipoma Preferred Partner; RGS1 = Regulator of G-protein signaling 1; SH2B3 = SH2B Adaptor Protein 3; TAGAP = T-cell Activation GTPase Activating Protein; TNFSF14 = Tumor Necrosis Factor (Ligand) Superfamily member 14.

Multivariate Analysis of Candidate Genes Expression

Since gene expression is inter-correlated within the same cell type; the expression of each candidate gene may not provide an accurate picture of eventual differences between groups. Some genes may be over-expressed due to the effect of another gene in the same pathway. In a stepwise discriminant analysis (Table 2 panels A and B), the combined expression of 4 genes (*TAGAP*, *TNFS14*, *SH2B3*, and *RGS1*) significantly discriminated between patients and controls before diagnosis (Time 1); in fact, it correctly predicted 9 of 9 patients and 12 of 13 controls (overall 95% correct prediction). Figure 1B shows the discriminant score obtained by combining the expression of the 4 selected genes in each individual case. All of the children with CD had a negative D-score and all those without CD had a positive D-score, no child was wrongly classified as ''CD.''

TABLE 3. Discriminant analysis combining genotype and gene expression and of candidate genes before diagnosis (Time 1)

Step	Genotype	Wilks' lambda	Variance ratio F Statistic	*P*
(A) Stepwise discriminat analysis of combined genotype and gene expression at T1 of candidate genes				
1	SH2B3 genotype	0.604	13.091	0.002
2	RGS1 genotype	0.480	10.292	0.001
3	cREL expression at T1	0.340	11.626	0.000
4	TNFRSF14 genotype	0.260	12.066	0.000
5	SH2B3 expression at T1	0.217	11.523	0.000
6	TNFSF14 genotype	0.190	10.656	0.000

Predicted group

	Status	CD	Not CD	Total
(B) Classification by discriminat equation of combined genotype and gene expression at T1 of candidate genes				
Original group	CD	9 (100%)	0 (0%)	9
	Not CD	0 (0%)	13 (100%)	13

Predicted group by Jacknife iterative exclusion

	Status	CD	Not CD	Total
(C) Classification by discriminat equation of combined genotype and gene expression at T1 of candidate genes				
Original group	CD	7 (77.8%)	2 (22.2%)	9
	Not CD	2 (15.4%)	11 (84.6%)	13

(A) The genotype or the expression of 5 genes significantly contributed to lowering Wilks' lambda in a stepwise process (SH2B3 genotype, RGS1 genotype, cREL expression at T1, TNFRSF14 genotype, and TNFSF14 genotype) and were selected for discriminating capacity. (B) Computing the relative membership probability by the discriminant score, 100% of patients were correctly classified. (C) Cross validation analysis 84.6% of controls and 77.8% patients were correctly classified.

CD = celiac disease; RGS1 = Regulator of G-protein signaling 1; SH2B3 = SH2B Adaptor Protein 3; TAGAP = T-cell Activation GTPase Activating Protein; TNFSF14 = Tumor Necrosis Factor (Ligand) Superfamily member 14

Multivariate Analysis Combining Genotype With Gene Expression

Although a discriminant function based on a relatively small sample performs well in prediction, it may not be sufficiently reliable for diagnostic purposes. To enhance the robustness of the model, we pooled the genotyping and gene expression data at Time 1, and applied the stepwise multivariate approach reported above. Table 3 shows the performance of genotypes and expression of the candidate genes.

The genotypes of 2 genes (*SH2B3* and *RGS1*) were the most powerful in discriminating between patients and controls. The expression of *cREL* increased the discriminating capacity. The next most powerful discriminator was the TNFRSF14 genotype, followed by SH2B3 expression, and finally the *TNFSF14* genotype

(Table 3A). This combination resulted in a low Wilks' lambda (0.604–0.190); consequently, this discriminant equation allowed to correctly classify 100% of subjects (Table 3B).

To verify the robustness of the model, we adopted an auto-exclusion strategy (each individual was predicted by an equation excluding his own data, iteratively) to obtain an unbiased estimate of the predicting capacity of the discriminant function. In this case, only 2 of 9 CD patients were not predicted as CD, and only 2 of 13 controls were wrongly predicted as CD; 81.8% of CD patients and controls were correctly classified (Table 3C).

DISCUSSION

Children from families at risk for CD have a consistent risk of occurrence, from 1% to >20% depending on their HLA (5); they contribute notably to the growth of the celiac population. In fact, in Europe, with a population around 740,000,000, 5 to 7 million CD patients are likely to generate several tens of thousands of affected children each year. Over and above the HLA risk, we recently showed that, within each HLA risk class, the presence of an ''at-risk allele'' of only 3 candidate genes significantly increases the precision of the estimate of risk of occurrence, with a refined range of risk estimate ranging between 0.04 and 0.08 in DQ8-carrying individuals and between 0.17 and 0.23 in DQB1∗02 homozygote (23). The HLA genotype in addition to the genotype of the candidate genes, however, explains less than half the variance of CD heredity. The other half resides in the modulation of the action of genes (expression and regulation). The ''missing heredity'' of the disease could be explained by the action of epigenetic mechanisms such as DNA methylation, histone modifications, and microRNA regulation, by the interaction between genes and environmental risk factors, but this field of study is still the subject of just a few scientific articles (25,26).

The aim of this study was to predict which children from at-risk families will eventually develop CD from birth to 6 years of age, well in advance of the production of auto-antibodies (anti-tTG) or the appearance of clinical symptoms. We confirmed the importance of the HLA haplotype, and found that the genotype of 5 candidate genes (*SH2B3*, *RGS1*, *TAGAP*, *cREL*, and *LPP*) would seem increases (>90%) the risk of developing CD. But most children in at-risk families share their genotypes with affected relatives, and those selected for HLA-DQ2-DQ8 positivity should have the vast majority of genes equally distributed; hence, it is more informative to explore the expression of genes rather than their genotype.

From a simple blood sample, we were able to predict which children would develop CD within 18 to 24 months in this selected cohort. In fact, 5 genes (*KIAA*, *TAGAP*, *SH2B3*, *TNFSF14*, and *RGS1*) were differentially expressed long before any sign of the disease, and therefore identified with remarkable accuracy the children who did eventually developed CD. To increase the robustness of the prediction we combined the genotype of the child, obtained at birth, with gene expression at 4 to 16 months, and obtained a combination of genotypes and gene expression profiles that significantly increased (up to 100%) the ability to predict the outcome of the child.

Because the discriminant equation by which we classified patients at least 9 months before the CD diagnosis was developed from the data obtained in the same cohort, we adopted an auto-exclusion strategy (''jack-knifing'') in which we excluded iteratively from the equation the individuals to be classified. Notably, the differential gene expression (CD vs controls) observed before the appearance of specific antibodies, decrease at moment of diagnosis, and was no longer present 12 or more months after the diagnosis and the start of the GFD. The before/after study design provides the most efficient matched control of the gene expression profile observed before diagnosis. The reversibility of the expression of some genes in CD patients treated with a GFD is in line with the reversibility of deranged tight junction gene expression in CD children after 2 years of GFD (26,27). The reversibility of gene expression may indirectly confirm a diagnosis of CD, and consequently the evaluation of the expression can be considered a diagnostic marker (26). Studies of the potential role in CD diagnosis of these and other gene expressions, however, indicate that, unlike constitutive genetic profiles, a number of epigenetic regulation mechanisms of these candidate genes, may play a pivotal role in the etiology of CD (26–29).

This study has the limit of a small sample size, which however is counterbalanced by the longitudinal study design of a peculiar cohort of children at high risk of developing CD in the first years of their life. The longitudinal study design allows to observe important modifications of the gene expression in the same individual, which was analyzed 3 times: before diagnosis, at diagnosis, and after diagnosis, multiplying by 3 the power of the study. It is noteworthy that this model is applied in this context for the first time, since the follow-up of babies at risk for CD is not yet a common practice. We have also to acknowledge that to recruit infants from at-risk families, which have to undergo a surveillance lasting 6 years, with regular blood sampling on a scheduled basis, was made possible only by the PREVENT-CD longitudinal study. We can assure that no selection bias was applied to recruit these infants, other than the technical availability of the repeated blood samples.

Another limitation of this study is the small number of genes explored; this was a strategic choice because we preferred to explore genes with robust replication in several Genome Wide Association Studies (10–16), with a strong biological implication in the gluten-induced immune response (17), and explored in previous studies from our group in several models (19,23). Finally, it is important to underline that the selection of this candidate genes is not biased for the present cohort; it was identified in previous studies.

Given that epigenetic mechanisms are highly cell-type specific, analyses conducted with mixed cell tissues (eg, mucosal biopsies) carry the risk to be confounded by differences in cell type composition, particularly when comparing inflamed versus non-inflamed tissues. In our study design, we minimized this bias because we did not compare an inflamed tissue with a healthy one since we worked with PBMCs, when no major sign of antibody production or inflammation was present; and we compared the same cell type in the same individual before and after the occurrence of a major gluten-induced auto-immune event.

In conclusion, our preliminary work may pave the way for a molecular diagnosis of CD, before the onset of antibody production, in infants who carry a genetic risk profile. To consolidate these preliminary results, we are engaged in continuing longitudinal studies of at-risk infants, expanding not only the sample size but also the study of the ''regulome'' of these subjects. The diagnostic procedures for chronic diseases are slowly shifting from invasive, tissue-based techniques to less invasive molecular methodologies. The time is now ripe for such a common disease as CD to gradually shift to a molecular diagnosis for the vast majority of patients.

Acknowledgments: The authors thank Jean Ann Gilder (Scientific Communication srl, Naples, Italy) for revising and editing the manuscript.

REFERENCES

1. Green PH, Cellier C. Celiac disease. *N Engl J Med* 2007;357:1731–43.
2. Di Sabatino A, Corazza GR. Coeliac disease. *Lancet* 2009;373:1480–93.

3. Mustalahti K, Catassi C, Reunanen A, et al. The prevalence of celiac disease in Europe: results of a centralized, international mass screening project. *Ann Med* 2010;42:587–95.
4. Tucci F, Astarita L, Abkari A, et al. Celiac disease in the Mediterranean area. *BMC Gastroenterol* 2014;14:24.
5. Greco L, Romino R, Coto I, et al. The first large population based twin study of coeliac disease. *Gut* 2002;50:624–8.
6. Husby S, Koletzko S, Korponay-Szabó IR, et al. Working Group of European Society of Paediatric Gastroenterology and Nutr European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease. *J Pediatr Gastroenterol Nutr* 2012;54:136–60.
7. Schuppan D, Junker Y, Barisani D. Celiac disease: from pathogenesis to novel therapies. *Gastroenterology* 2009;137:1912–33.
8. Sollid LM, Jabri B. Triggers and drivers of autoimmunity: lessons from coeliac disease. *Nat Rev Immunol* 2013;13:294–302.
9. Romanos J, van Diemen CC, Nolte IM, et al. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. *Gastroenterology* 2009;137:834–40.
10. van Heel DA, Franke L, Hunt KA, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* 2007;39:827–9.
11. Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 2008;40:395–402.
12. Castellanos-Rubio A, Martin-Pagola A, Santín I, et al. Combined functional and positional gene information for the identification of susceptibility variants in celiac disease. *Gastroenterology* 2008;134: 738–46.
13. Romanos J, Barisani D, Trynka, et al. Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. *J Med Genet* 2009;46:60–3.
14. Trynka G, Zhernakova A, Romanos J, et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kB signaling. *Gut* 2009;58:1078–83.
15. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010;42:295–302.
16. Trynka G1, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* 2011;43:1193–201.
17. Kumar V, Gutierrez-Achury J, Kanduri K, et al. Systematic annotation of celiac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. *Hum Mol Genet* 2015;24:397–409.
18. Dieli-Crimi R, Cénit MC, Núñez C. The genetics of celiac disease: a comprehensive review of clinical implications. *J Autoimmun* 2015;64:26–41.
19. Galatola M, Izzo V, Cielo D, et al. Gene expression profile of peripheral blood monocytes: a step towards the molecular diagnosis of celiac disease? *Plos One* 2013;8:e74747.
20. Ontiveros N, Tye-Din JA, Hardy MY, et al. Ex-vivo whole blood secretion of interferon (IFN)-γ and IFN-γ-inducible protein-10 measured by enzyme-linked immunosorbent assay are as sensitive as IFN-γ enzyme-linked immunospot for the detection of gluten-reactive T cells in human leucocyte antigen (HLA)-DQ2·5(+)-associated coeliac disease. *Clin Exp Immunol* 2014;175:305–15.
21. HogenEsch CE, Rosén A, Auricchio R, et al. The Prevent CD Study design: towards new strategies for the prevention of coeliac disease. *Eur J Gastroenterol Hepatol* 2010;22:1424–30.
22. Vriezinga SL, Auricchio R, Bravi E, et al. Randomized feeding intervention in infants at high risk for celiac disease. *N Engl J Med* 2014;371:1304–15.
23. Izzo V, Pinelli M, Tinto N, et al. Improving the estimation of celiac diseases ibling risk by non-HLA genes. *PLoS One* 2011;6:e26920.
24. Monsuur AJ, de Bakker PI, Zhernakova A, et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoSOne* 2008;3:e2270.
25. Bourgey M, Calcagno G, Tinto N, et al. HLA related genetic risk for coeliac disease. *Gut* 2007;56:1054–9.
26. Fernandez-Jimenez N, Castellanos-Rubio A, Plaza-Izurieta L, et al. Co-regulation and modulation of NFκB-related genes in celiac disease: uncovered aspects of gut mucosal inflammation. *Hum Mol Genet* 2014;23:1298–310.
27. Plaza-Izurieta L, Fernandez-Jimenez N, Irastorza I, et al. Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease. *Eur J Hum Genet* 2015;23:1100–5.
28. Jauregi-Miguel A, Fernandez-Jimenez N, Irastorza I, et al. Alteration of tight junction gene expression in celiac disease. *J Pediatr Gastroenterol Nutr* 2014;58:762–7.
29. Zilbauer M, Zellos A, Heuschkel R, et al. Epigenetics in paediatric gastroenterology, hepatology and nutrition—current trends and future perspectives. *J Pediatr Gastroenterol Nutr* 2016;62:521–9.