



OPEN

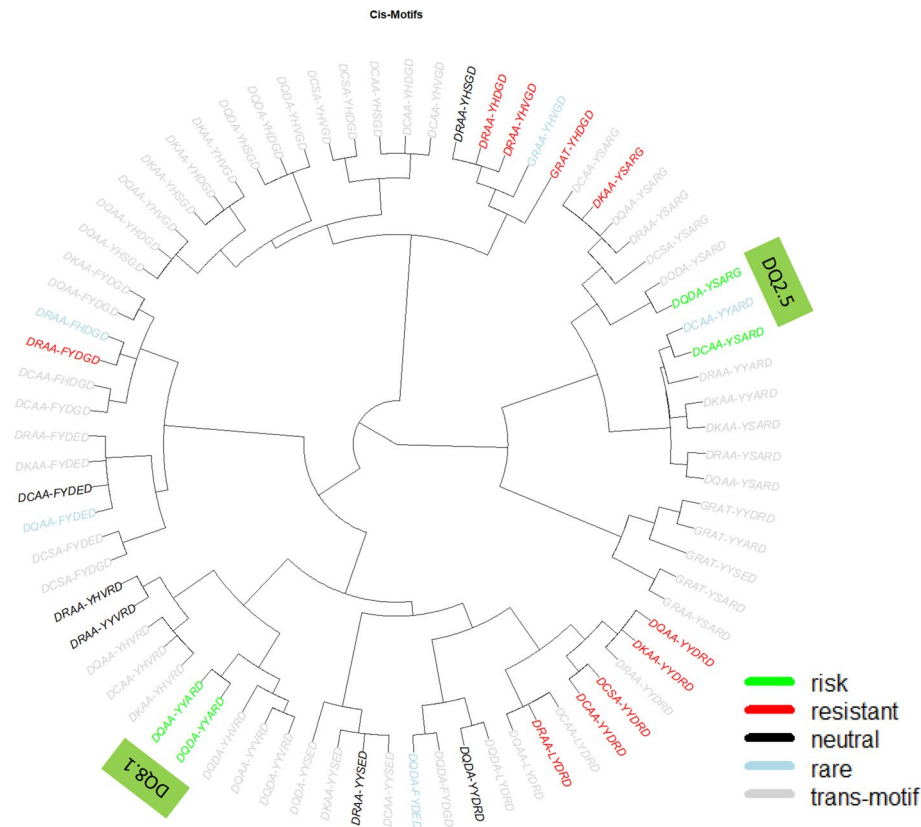
## Nine residues in HLA-DQ molecules determine with susceptibility and resistance to type 1 diabetes among young children in Sweden

Lue Ping Zhao<sup>1✉</sup>, George K. Papadopoulos<sup>2,10</sup>, Antonis K. Moustakas<sup>3</sup>, George P. Bondinas<sup>2</sup>, Annelie Carlsson<sup>4</sup>, Helena Elding Larsson<sup>5</sup>, Johnny Ludvigsson<sup>6</sup>, Claude Marcus<sup>7</sup>, Martina Persson<sup>8</sup>, Ulf Samuelsson<sup>6</sup>, Ruihan Wang<sup>9</sup>, Chul-Woo Pyo<sup>9</sup>, Daniel E. Geraghty<sup>9</sup> & Åke Lernmark<sup>5</sup>

HLA-DQ molecules account over 50% genetic risk of type 1 diabetes (T1D), but little is known about associated residues. Through next generation targeted sequencing technology and deep learning of DQ residue sequences, the aim was to uncover critical residues and their motifs associated with T1D. Our analysis uncovered ( $\alpha$ 1,  $\alpha$ 44,  $\alpha$ 157,  $\alpha$ 196) and ( $\beta$ 9,  $\beta$ 30,  $\beta$ 57,  $\beta$ 70,  $\beta$ 135) on the HLA-DQ molecule. Their motifs captured all known susceptibility and resistant T1D associations. Three motifs, "DCAA-YSARD" (OR = 2.10,  $p = 1.96 \times 10^{-20}$ ), "DQAA-YYARD" (OR = 3.34,  $2.69 \times 10^{-72}$ ) and "DQDA-YYARD" (OR = 3.71,  $1.53 \times 10^{-6}$ ) corresponding to DQ2.5 and DQ8.1 (the latter two motifs) associated with susceptibility. Ten motifs were significantly associated with resistance to T1D. Collectively, homozygous DQ risk motifs accounted for 43% of DQ-T1D risk, while homozygous DQ resistant motifs accounted for 25% protection to DQ-T1D risk. Of the identified nine residues five were within or near anchoring pockets of the antigenic peptide ( $\alpha$ 44,  $\beta$ 9,  $\beta$ 30,  $\beta$ 57 and  $\beta$ 70), one was the N-terminal of the alpha chain ( $\alpha$ 1), one in the CD4-binding region ( $\beta$ 135), one in the putative cognate TCR-induced  $\alpha\beta$  homodimerization process ( $\alpha$ 157), and one in the intra-membrane domain of the alpha chain ( $\alpha$ 196). Finding these critical residues should allow investigations of fundamental properties of host immunity that underlie tolerance to self and organ-specific autoimmunity.

Type 1 diabetes (T1D) is an autoimmune disease, due to the destruction of normal pancreatic islet beta-cells by the host immune system, and accounts for the majority of all diabetes among young children<sup>1,2</sup>. Further, earlier twin and family studies underline the major role of genetic factors together with environmental factors<sup>1-4</sup>. Using modern genotyping technologies, recent genome wide association studies (GWAS) have uncovered ~50 single nucleotide polymorphisms (SNPs) in multiple non-HLA immune genes explaining probably less than 25% T1D associations<sup>5-7</sup>. In contrast, class II genes in the human leukocyte antigen (HLA) system have been researched for several decades, and strongly contribute to the genetic association with T1D, accounting for ~50% of genetic association with T1D<sup>8,9</sup>. Class II genes include the multiallelic *HLA-DR\*AI-B1* haplotype coding for the DR $\alpha\beta$  heterodimer ( $\alpha$ -chain essentially monomorphic), and the *HLA-DQ\*AI-B1* haplotype coding the DQ $\alpha\beta$  heterodimer. Due to their genetic proximity, *HLA-DR* and *-DQ* genes are in linkage disequilibrium (LD), and thus empirically separating their individual associations with T1D is challenging<sup>9-11</sup>. Conventionally, DQ molecules exhibit

<sup>1</sup>Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave NE, Seattle, WA 98109, USA. <sup>2</sup>Laboratory of Biophysics, Biochemistry, Biomaterials and Bioprocessing, Department of Agriculture, Faculty of Agricultural Technology, University of Ioannina, Ioannina, Greece. <sup>3</sup>Department of Food Science and Technology, Faculty of Environmental Sciences, Ionian University, GR26100 Argostoli, Cephalonia, Greece. <sup>4</sup>Department of Pediatrics, Lund University, Lund, Sweden. <sup>5</sup>Department of Clinical Sciences, Lund University CRC, Skåne University Hospital, Malmö, Sweden. <sup>6</sup>Department of Clinical and Experimental Medicine, Crown Princess Victoria Children's Hospital, Region Östergötland and Div of Pediatrics, Linköping University, Linköping, Sweden. <sup>7</sup>Department of Clinical Science and Education Karolinska Institutet and Institution of Medicine, Clinical Epidemiology, Karolinska Institutet, Stockholm, Sweden. <sup>8</sup>Department of Medicine, Clinical Epidemiological Unit, Karolinska Institutet, Stockholm, Sweden. <sup>9</sup>Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. <sup>10</sup>George K. Papadopoulos is retired. ✉email: lzhao@fredhutch.org



**Figure 1.** Phylogenetic representation of all cis-DQ motifs that are risk to ( $p < 0.05$  and  $OR > 1$ , colored green), resistant to ( $p < 0.05$ ,  $OR < 1$ , colored red), neutral ( $p > 0.05$ , black), and rare (with fewer than 5 observed motif copies, colored blue), together with all trans-DQ motifs (colored gray).

stronger empirical associations with T1D. Our recent effort focused on identification of specific DQ risk residues among subjects with high-risk HLA-DQ haplotypes (DQ2.5 and DQ8.1)<sup>12</sup> as well as those resistant residues among subjects with low risk HLA-DQ haplotypes<sup>13</sup>, complementing earlier works on HLA-DR molecules<sup>14,15</sup>.

Integrating susceptible and resistant DQ residues, the current investigation was to explore their genotypic associations with T1D, especially, assessing heterodimeric associations with DQA1 and DQB1 alleles on the same chromosome (cis) and between homologous chromosomes (trans), with specific focus on the DQ2.5 and DQ8.1 haplotypes. Furthermore, we assessed if DR3 and DR4, two major risk alleles of HLA-DRB1, confounded HLA-DQ associations with T1D through haplotypic association analysis. To gain insight into autoimmunity, we centered our analysis on DQ motif associations with six biochemically defined islet autoantibodies measured at the time of T1D diagnosis<sup>16</sup>. To replicate discovered associations, we utilized a case-control study from the Type 1 Diabetes Genetic Consortium (T1DGC) to carry out T1D associations with discovered motifs<sup>17</sup>. Last, in a final analysis we estimated how much susceptibility and resistant DQ motifs contributed to either the excess or reduction of risk for T1D.

## Results

**Critical residues among subjects with risk or resistant DQ haplotypes.** *DQA1* and *DQB1*, which are in high LD and their alleles mostly form cis-haplotypes, are translated to form respectively the alpha and beta chain of the heterodimeric HLA-DQ molecule. DQ haplotypes were polymorphic and had 45 unique haplotypes in our case-control study, each of which consisted of unique residues (complete lists of amino acid sequences of mature molecules in Supplementary Figs. S1 and S2). Our phylogenetic analysis of sequence similarities showed that these DQ haplotypes were hierarchically organized into seven haplotype clusters; 1 to 7 (Fig. 1). Clusters 2 and 4 included susceptible motifs DQ2.5 and DQ8.1 (motifs colored green), and carriers of all DQ motifs in these two clusters were at high T1D risk. Among high-risk subjects, hierarchically organized haplotype (HOH) analysis identified three critical residues ( $\alpha 44$ ,  $\beta 57$ ,  $\beta 135$ ) that captured disease association with HLA-DQ among subjects with high-risk DQ haplotypes<sup>12</sup>. The remaining clusters included largely resistant DQ haplotypes, and the HOH analysis of subjects revealed seven critical residues ( $\alpha 1$ ,  $\alpha 156$ ,  $\alpha 196$ ,  $\beta 9$ ,  $\beta 30$ ,  $\beta 57$ ,  $\beta 70$ ) that accounted for autoimmunity in T1D<sup>13</sup>.

Residue	AA	High Risk Carriers					Low Risk Carriers				
		Control	Patient	OR	HS	p	Control	Patient	OR	HS	p
1) $\alpha$ 1	D	441 (34.67)	1404 (72.97)	2.10	20.28	1.84E-91	822 (64.62)	519 (26.98)	0.42	-19.97	9.79E-89
	G						9 (0.71)	1 (0.05)	0.07	-2.97	2.99E-03
2) $\alpha$ 44	C	153 (12.03)	485 (25.21)	2.10	9.26	1.96E-20	202 (15.88)	120 (6.24)	0.39	-8.90	5.42E-19
	K	114 (8.96)	46 (2.39)	0.27	-8.23	1.85E-16					
	Q	174 (13.68)	873 (45.37)	3.32	19.74	1.00E-86	70 (5.5)	63 (3.27)	0.60	-3.01	2.59E-03
	R						559 (43.95)	337 (17.52)	0.40	-16.20	5.20E-59
3) $\alpha$ 157	A	420 (33.02)	1300 (67.57)	2.05	18.15	1.27E-73	763 (59.98)	457 (23.75)	0.40	-19.98	7.83E-89
	D	21 (1.65)	104 (5.41)	3.27	5.33	9.57E-08	61 (4.80)	63 (3.27)	0.68	-2.12	3.43E-02
	S						7 (0.55)				
4) $\alpha$ 196	A	441 (34.67)	1404 (72.97)	2.10	20.28	1.84E-91	823 (64.70)	519 (26.98)	0.42	-20.02	3.63E-89
	T						8 (0.63)	1 (0.05)	0.08	-2.73	6.38E-03
5) $\beta$ 9	F						260 (20.44)	85 (4.42)	0.22	-13.87	9.44E-44
	L						9 (0.71)	1 (0.05)	0.07	-3.25	1.14E-03
6) $\beta$ 30	Y	441 (34.67)	1404 (72.97)	2.10	20.28	1.84E-91	562 (44.18)	434 (22.56)	0.51	-12.53	5.25E-36
	H						344 (27.04)	320 (16.63)	0.61	-7.34	2.09E-13
	S	231 (18.16)	532 (27.65)	1.52	6.37	1.85E-10					
7) $\beta$ 57	Y	210 (16.51)	872 (45.32)	2.75	17.83	4.45E-71	487 (38.29)	200 (10.40)	0.27	-17.91	1.02E-71
	A	392 (30.82)	1370 (71.21)	2.31	21.12	5.46E-99	5 (0.39)	7 (0.36)	0.93	-0.13	8.94E-01
	D	49 (3.85)	34 (1.77)	0.46	-3.54	3.93E-04	603 (47.41)	230 (11.95)	0.25	-20.83	2.23E-96
	S						19 (1.49)	17 (0.88)	0.59	-1.61	1.08E-01
8) $\beta$ 70	V						204 (16.04)	266 (13.83)	0.86	-1.76	7.80E-02
	E						61 (4.80)	79 (4.11)	0.86	-0.94	3.47E-01
	G						493 (38.76)	248 (12.89)	0.33	-16.58	8.99E-62
9) $\beta$ 135	R	441 (34.67)	1404 (72.97)	2.10	20.28	1.84E-91	277 (21.78)	193 (10.03)	0.46	-8.86	7.76E-19
	D	363 (28.54)	1357 (70.53)	2.47	21.65	6.06E-104	831 (65.33)	520 (27.03)	0.41	-20.28	1.84E-91
	G	78 (6.13)	47 (2.44)	0.40	-5.20	2.00E-07					

**Table 1.** Results from association analysis of nine selected residues ( $\alpha$ 1,  $\alpha$ 44,  $\alpha$ 157,  $\alpha$ 196,  $\beta$ 9,  $\beta$ 30,  $\beta$ 57,  $\beta$ 70,  $\beta$ 135) with T1D in the case-control study with 636 control and 962 patients with estimated allelic frequencies (%), odds ratio, hap-score and p-value among carriers of DQ2.5 and DQ8.1 and non-carriers.

**Four alpha and five beta residues were critically associated with T1D risk.** Integrating all risk and resistant residues in the DQ molecule led to four alpha residues ( $\alpha$ 1,  $\alpha$ 44,  $\alpha$ 157,  $\alpha$ 196) and five beta residues ( $\beta$ 9,  $\beta$ 30,  $\beta$ 57,  $\beta$ 70,  $\beta$ 135) that were critically associated with either susceptibility or resistance to T1D. To assess roles of these residues in T1D associations, we investigated residue-specific frequencies of corresponding amino acids between controls and patients, who had high and low DQ risk haplotypes (Table 1). Residue  $\alpha$ 44 took four possible amino acids (C, K, Q, R);  $\alpha$ 44C and Q were strongly associated with high T1D risks (OR=2.10 and 3.32,  $p=1.96 \times 10^{-20}$  and  $1.00 \times 10^{-86}$ , respectively),  $\alpha$ 44K was strongly associated with resistance to T1D risk (OR=0.27,  $p=1.85$ )<sup>13</sup> and  $\alpha$ 44R was absent among high-risk DQ subjects. Among low risk DQ subjects, the resistant  $\alpha$ 44C and Q had significant associations with resistance to T1D risk (OR=0.39 and 0.6,  $p=5.42 \times 10^{-19}$  and  $2.59 \times 10^{-3}$ , respectively), and the residue  $\alpha$ 44R was highly associated with the resistance to T1D (OR=0.40 and  $p$ -value =  $5.20 \times 10^{-59}$ ). Interestingly, the amino acid  $\alpha$ 44K was absent in low risk subjects. When contrasting the associations of  $\alpha$ 44C and Q between high and low risk subjects, their differences of risk and resistant associations were highly significant (exact Fisher's  $p$ -value =  $2.16 \times 10^{-31}$  and  $1.92 \times 10^{-19}$ , respectively). On the other hand, the beta chain residues  $\beta$ 57A (OR=2.31;  $p=5.46 \times 10^{-99}$ ) and D (OR=0.46;  $3.93 \times 10^{-4}$ ) had reverse associations with T1D risk among high-risk subjects. Among the low risk subjects, the first residue,  $\beta$ 57A, had no significant association to T1D risk (OR=0.93,  $p=0.89$ ), while the second residue,  $\beta$ 57D, was associated with resistance to T1D (OR=0.25,  $p=2.23 \times 10^{-96}$ ). Two amino acids,  $\beta$ 57S/V, were absent among high-risk subjects and had no associations ( $p=0.11$  and  $0.78$ , respectively). Last,  $\beta$ 135D and G had opposite associations with T1D high-risk subjects (OR=2.47 and 0.40,  $p=6.06 \times 10^{-104}$  and  $2.00 \times 10^{-7}$ , respectively). Among the low risk subjects,  $\beta$ 135D had significant resistance to T1D (OR=0.41,  $p=1.84 \times 10^{-91}$ ), while  $\beta$ 135G was absent. The disease associations with residue  $\beta$ 135D were completely opposite to each other between high and low risk subjects (Fisher's exact test;  $p=2.88 \times 10^{-117}$ ).

From the perspective of resistant residues,  $\alpha$ 1D and G among low risk subjects were significantly associated with resistance to T1D (OR=0.42 and 0.07,  $p=9.79 \times 10^{-89}$  and  $2.99 \times 10^{-3}$ , respectively). Among high-risk subjects, residue  $\alpha$ 1D had significant association with susceptibility to T1D (OR=2.10,  $p=1.84 \times 10^{-91}$ ), while  $\alpha$ 1G was absent. The association with  $\alpha$ 1D was significantly different between low and high-risk subjects (exact Fisher's

ID	cis Motif	Control	Patient	OR	Z	p	Sera	Allele Group	aa1	aa4	aa157	aa196	β9	β30	β57	β70	β135							
1	DCAA-FYDED	55 (4.32)	77 (4.00)	0.93	-0.45	6.53E-01	DQ4	*04:02~*04:02~*04:01~*04:02~*04:04~*04:02	.	C	A	.	F	.	.	E	.							
2	DCAA-YSARD	153 (12.03)	485 (25.21)	2.10	9.26	1.96E-20	DQ2.5	*05:01~*02:01~*05:05~*02:01	.	C	A	.	.	S	A	.	.							
3	DCAA-YYDRD	139 (10.93)	40 (2.08)	0.19	-10.41	2.32E-25	DQ7	allele group <sup>1</sup>	.	C	A	.	.	.	.	.	.							
4	DCSA-YYDRD	7 (0.55)						*05:03~*03:01	.	C	S	.	.	.	.	.	.							
5	DCAA-YSARG	78 (6.13)	41 (2.13)	0.35	-5.76	8.47E-09	DQ2.2	*02:01~*02:02	.	K	A	.	.	S	A	.	G							
6	DCAA-YYDRD	36 (2.83)	5 (0.26)	0.09	-6.36	1.99E-10	DQ9.2	*02:01~*03:03	.	K	A	.	.	.	.	.	.							
7	DQAA-YYARD	152 (11.95)	769 (39.97)	3.34	17.98	2.69E-72	DQ8.1	*03:01~*03:02~*03:01~*03:04~*03:01~*03:05	.	A	.	.	.	A	.	.								
8	DQAA-YYDRD	9 (0.71)						*03:01~*03:01~*03:01~*03:03	.	A	.	.	.	.	.	.								
9	DQDA-YSARG		6 (0.31)					*03:02~*02:02	.	.	.	.	.	S	A	.	G							
10	DQDA-YYARD	13 (1.02)	73 (3.79)	3.71	4.81	1.53E-06	DQ8.1	*03:02~*03:02~*03:02~*03:04	.	.	.	.	.	.	.	A	.							
11	DQDA-YYDRD	67 (5.27)	88 (4.57)	0.87	-0.87	3.82E-01		*03:02~*03:03~*03:03~*03:01 <sup>3</sup>	.	.	.	.	.	.	.	D	Q	D	A	Y	Y	D	R	D
12	DRAA-FYDGD	201 (15.8)	8 (0.42)	0.03	-16.39	2.25E-60	DQ6.2	*01:02~*06:02	.	R	A	.	.	F	.	.	G							
13	DRAA-LYDRD	9 (0.71)	1 (0.05)	0.07	-3.25	1.14E-03		*01:03~*06:01	.	R	A	.	.	L	.	.	.							
14	DRAA-YHDGD	117 (9.2)	44 (2.29)	0.25	-8.86	7.72E-19	DQ6.3	allele group <sup>2</sup>	.	R	A	.	.	H	.	.	G							
15	DRAA-YHSGD	16 (1.26)	15 (0.78)	0.62	-1.36	1.75E-01		*01:02~*05:02	.	R	A	.	.	H	S	G	.							
16	DRAA-YHVG D	149 (11.71)	180 (9.36)	0.80	-2.15	3.12E-02		*01:01~*05:01~*01:02~*05:01	.	R	A	.	.	H	V	G	.							
17	DRAA-YHVRD	52 (4.09)	80 (4.16)	1.02	0.10	9.22E-01	DQ6.4	*01:02~*06:04	.	R	A	.	.	H	V	.	.							
18	DRAA-YYSED	3 (0.24)	2 (0.10)	0.44	-0.92	3.55E-01		*01:02~*05:04	.	R	A	.	.	S	E	.	.							
19	DRAA-YYVRD	2 (0.16)	6 (0.31)	1.98	0.86	3.91E-01		*01:02~*06:09	.	R	A	.	.	V	.	.	.							
20	GRAT-YHDGD	8 (0.63)	1 (0.05)	0.08	-2.73	6.38E-03		*01:04~*05:03	G	R	A	T	Y	H	.	G	.							

1) allele group: \*05:05~\*03:01~\*05:05~\*03:19~\*05:05~\*03:09~\*05:10~\*03:01~\*05:09~\*03:01~\*06:01~\*03:01~\*04:01~\*03:03

2) allele group: \*01:01~\*05:03~\*01:03~\*06:03~\*01:02~\*06:03~\*01:07~\*05:03~\*01:10~\*06:03. The combination DQA1\*01:07-B1\*05:03 cannot appear, bound with an antigenic peptide, on the cell membrane of an APC because of the inability of α76Cys (instead of the near invariant Arg) to form a salt bridge with β57Asp<sup>70</sup>. However, the combination DQA1\*01:07-B1\*06:04, not found here in cis may be allowed in trans

3) motif ID=11: DQDA-YYDRD is selected as a reference motif

**Table 2.** Allelic association analysis of type 1 diabetes with DQA1-DQB1 motifs of (aa1, α44, α157, α196, β9, β30, β57, β70, β135): estimated frequency (%) among control and patient, estimated odds ratios, Z-score, and p-value, corresponding serotype, and equivalent allelic groups, in addition to profiles of amino acids across all motifs. Five rare motifs, with fewer five observed motif copies, are excluded. In total we have 1266 and 1921 motif copies observed in 636 controls and 962 patients, respectively, after excluding 6 and 3 motif copies from respective groups.

$p = 7.87 \times 10^{-102}$ ). The next residue α157 took three possible amino acids (A, D, S); the first two, α157A/D, had significant resistance associations (OR = 0.40 and 0.68,  $p = 7.83 \times 10^{-89}$  and 0.034, respectively), while α157S was absent among patients but had seven allelic copies among controls with low risk DQ haplotypes. Among high-risk subjects, α157A/D were significantly associated with susceptibility to T1D (OR = 2.05 and 3.27,  $p = 1.27 \times 10^{-73}$  and  $9.57 \times 10^{-8}$ , respectively), while the third residue, α157S, was absent. Again, T1D risk associations with α157A/D were significantly different between low- and high-risk subjects (exact Fisher's  $p = 9.18 \times 10^{-97}$  and  $4.93 \times 10^{-8}$ , respectively). The next residue, β9, had three amino acids associated with resistance (F, L, Y; OR = 0.22, 0.07 and 0.51,  $p = 9.44 \times 10^{-44}$ ,  $1.14 \times 10^{-3}$  and  $5.25 \times 10^{-36}$ , respectively) among low-risk subjects. Among high-risk subjects, β9F and L were absent, while β9Y was associated with T1D risk (OR = 2.10,  $p = 1.84 \times 10^{-91}$ ), and the association was significantly different between high and low risk (exact Fisher's  $p = 3.74 \times 10^{-66}$ ). Among low-risk subjects, residue β30 had two resistant amino acids β30H/Y (OR = 0.61 and 0.27,  $p = 2.09 \times 10^{-13}$  and  $1.02 \times 10^{-71}$ , respectively), and β30S was not observed in such carriers. Among high-risk subjects, β30S/Y had significant associations with susceptibility to T1D (OR = 1.52 and 2.75,  $p = 1.85 \times 10^{-10}$  and  $4.45 \times 10^{-71}$ , respectively), but β30H was not observed in these subjects. For amino acid β30Y, its association with T1D was significantly different between low- and high-risk subjects (Fisher's exact test;  $p = 7.62 \times 10^{-106}$ ). The ninth residue, β70, had two amino acids significantly associated with resistance, G and R (OR = 0.33 and 0.46,  $p = 7.76 \times 10^{-19}$  and  $1.84 \times 10^{-91}$ , respectively), and had one amino acid, β70E, with a neutral association (OR = 0.86,  $p = 0.34$ ) among low-risk subjects. No high-risk subjects were found with β70E or G, while β70R was significantly associated with T1D (OR = 2.10,  $p = 1.84 \times 10^{-91}$ ), and its association was significantly different between high- and low-risk subjects (exact Fisher's  $p = 2.19 \times 10^{-45}$ ).

In summary, all these nine residues had highly significant associations with T1D among subjects with either low- or high-risk DQ haplotypes. Strikingly, the same amino acids could exhibit paradoxical associations between low- and high-risk subjects. Further, some of common amino acids in the high-risk group are absent from the low-risk group, and vice-versa. Note that some motifs listed in Table 2 had fewer than 10 observed copies, e.g., ID 13, 18, 19 and 20, and their associated p-values were not robust enough and should be taken as suggestive values.

**Motifs of nine residues account for DQ susceptibility and resistance to type 1 diabetes risks.** Four alpha residues and five beta residues originated from transcription in cis and haplotypes of these nine amino acids are referred to as cis DQ motifs. In total, there were 25 unique DQ motifs: four susceptibility

motifs (highlighted in green), ten resistance motifs (highlighted in red), and six neutral motifs with insignificant associations ( $p$ -values  $> 0.05$ ) (Table 2). Five rare motifs with fewer than five observations were shown only in the Supplementary Table S1. The motif “DCAA-YSARD”, corresponding to DQA1\*05:01-B1\*02:01 and A1\*05:05-B1\*02:01, was significantly associated with susceptibility to T1D (OR = 2.10,  $p = 1.96 \times 10^{-20}$ ), and is equivalent to the traditionally known DQ2.5 serotype. The motifs “DQAA-YYARD” and “DQDA-YYARD”, corresponding to traditionally known DQ8.1 serotypes, had significant association with T1D susceptibility (OR = 3.34 and 3.71,  $p = 2.69 \times 10^{-72}$  and  $1.53 \times 10^{-6}$ , respectively). Last, the susceptible motif “DQDA-YSARG” was absent among controls and had six copies with allelic frequency of 0.31% among patients.

Among twenty HLA-DQ motifs, ten were associated with resistance to T1D. The motif “DRAA-FYDGD”, uniquely corresponding to the well-known protective DQ6.2 haplotype, conveyed significant resistance to T1D risk (OR = 0.03,  $p = 2.25 \times 10^{-60}$ ). The next resistance motif “DCAA-YYDRD”, motif 3, corresponding to a group of HLA-DQ haplotypes listed in the footnote, was significantly resistant to T1D (OR = 0.19,  $p = 2.32 \times 10^{-25}$ ). The third most protective motif, “DRAA-YHDGD”, was associated with resistance to T1D (OR = 0.25,  $p = 7.72 \times 10^{-19}$ ). The other five protective motifs were “DKAA-YSARG”, “DKAA-YYDRD”, “DRAA-LYDRD”, “DRAA-YHVGD” and “GRAT-YHDGD”, with  $p$ -values less than 0.05. Two motifs, “DCSA-YYDRD” and “DQAA-YYDRD”, were observed (with more than five copies) only among controls (Table 2).

Given the substantially different genetic distributions of residues between low- and high-risk subjects (Table 1), we assessed motif associations stratified by low or high DQ risks. As expected, distributions of DQ motifs were rather distinct between subjects with low and high DQ risks (Table S1). The most of commonly observed motifs were specific to subjects with high or low risk DQ haplotypes, and their associations with T1D were largely consistent with those in the combined low and high-risk subjects. For example, the motif “DCAA-YSARD” was observed only among high-risk subjects, and “DCAA-YYDRD” was observed only among low risk subjects. Hence, in all combined data, their respective association with susceptibility or resistance for T1D were largely unchanged. One exception, to this general observation, was the motif “DQDA-YYARD”, which was significantly associated with susceptibility to T1D risk among high-risk subjects (OR = 4.56,  $p = 4.30 \times 10^{-7}$ ) but was relatively uncommon among low risk subjects and showed no indication of associations with T1D risks (OR = 0.88,  $p = 0.87$ ).

**Subtle variations of residues alter association profiles with T1D.** The directly constructed profile of disease-associated motifs permitted us to examine how subtle variations of specific residues may alter their T1D associations (right panel of Table 2). The motif “DCAA-YSARD”, corresponding to DQ2.5, differed from the adjacent motif “DCAA-YYDRD” with only two residues at  $\beta 30$  and  $\beta 57$ , but their T1D associations were completely opposite and reverted from susceptibility to resistance (OR = 2.10 and 0.19,  $p = 1.96 \times 10^{-20}$  and  $2.32 \times 10^{-25}$ , respectively), the difference between the two was significant (Fisher’s exact test;  $p = 4.24 \times 10^{-39}$ ). The next motif pair “DQAA-YYARD”, corresponding to DQ8.1, and “DKAA-YYDRD”, differed only at  $\alpha 44$  and  $\beta 57$ , and their T1D associations were opposite (OR = 3.34 and 0.09,  $p = 2.69 \times 10^{-72}$  and  $1.99 \times 10^{-10}$ , respectively), the difference was significant (Fisher’s exact test;  $p = 5.04 \times 10^{-22}$ ). The next DQ8.1 motif “DQDA-YYARD” differed from the adjacent motif “DQDA-YYDRD” only at residue  $\beta 57$ , but the former was associated with susceptibility to T1D (OR = 3.71,  $p = 1.53 \times 10^{-6}$ ) and the other was neutral (OR = 0.87,  $p = 0.38$ ). The difference of the associations between these two motifs was significant (Fisher’s exact test;  $p = 7.48 \times 10^{-6}$ ). It was also of interest to contrast the uncommon motif 8, “DQAA-YYDRD”, with motif 7, “DQAA-YYARD”, and they were nearly identical except for  $\beta 57$ . In fact, all five alleles in both motifs had an identical alpha chain, DQA1\*03:01. Their associations differed significantly from each other (Fisher’s exact test;  $p = 1.16 \times 10^{-7}$ ).

**Genotypes of DQ motifs have profound associations with T1D.** Each subject carries a pair of DQ motifs, i.e., a motif genotype, and two motifs were jointly associated with T1D. The combination of 20 motifs (Tables 2 and S2) resulted in several rare genotypes with fewer than 5 observations. Upon merging of all rare genotypes into a single category, there was a total of 60 distinct genotypes (Table S2), and 25 were deemed to have neutral associations with corresponding  $p$ -values greater than 0.05. The genotype association results for 35 genotypes were sorted by  $Z$ -scores (Table 3). The first eight genotypes were significantly associated with susceptibility to T1D, with OR ranging from 2.46 to 14.54 ( $p$ -values ranged from 0.012 to  $4.14 \times 10^{-21}$ ). Next thirteen genotypes were significantly associated with resistance to T1D risk with OR ranging from 0.07 to 0.30 ( $p$ -values ranged from 0.038 to  $3.69 \times 10^{-9}$ ). The last fourteen genotypes, which had frequencies of five or more copies were observed among controls, and none were among patients, and were thus deemed to be highly resistant, perhaps protective, to T1D risks.

**DQ2.5 and DQ8.1 genotypic effects.** Out of 20 common DQ motifs, three motifs “DCAA-YSARD”, “DQAA-YYARD” and “DQDA-YYARD”, and the relatively uncommon “DQDA-YSARG”, were significantly associated with susceptibility to T1D. Given their susceptibilities, it was important to investigate specific genotype associations of these three DQ motifs, with themselves (homozygotes), with each other (doubly heterozygote), with neutral motifs (showing motif-specific effect), and with resistant motifs (showing additive motif effect); neutral motifs included all motifs with associated  $p$ -values greater than 0.05 and resistant motifs with  $p$ -values less than 0.05 and odds ratios less than 1. We conducted a univariate association analysis to compare each genotype compared to all other genotypes combined through a logistic regression model, estimating coefficient, odds ratio, standard error,  $Z$  score and  $p$ -value (Table 4).

Pairing the motif “DCAA-YSARD” with neutral motifs, itself, two DQ8.1 motifs and resistant motifs resulted in five possible genotypes (Table 4). The heterozygote with neutral motifs showed a significant association with susceptibility to T1D (OR = 2.54,  $p = 0.035$ ), which was largely due to the motif “DCAA-YSARD”. As

DQ Motif Genotype	Control	Patient	OR	Z	p
1 DCAA-YSARD/DQAA-YYARD	21 (3.3)	269 (27.96)	8.47	9.43	4.14E-21
2 DQAA-YYARD/DQAA-YYARD	5 (0.79)	97 (10.08)	12.83	5.56	2.64E-08
3 DQAA-YYARD/DRAA-YHVG	13 (2.04)	83 (8.63)	4.22	4.83	1.38E-06
4 DCAA-FYDED/DQAA-YYARD	9 (1.42)	48 (4.99)	3.53	3.47	5.22E-04
5 DCAA-YSARD/DCAA-YSARD	10 (1.57)	46 (4.78)	3.04	3.19	1.43E-03
6 DQAA-YYARD/DRAA-YHVR	11 (1.73)	41 (4.26)	2.46	2.66	7.91E-03
7 DCAA-YSARD/DQDA-YYARD	1 (0.16)	22 (2.29)	14.54	2.62	8.83E-03
8 DQAA-YYARD/DQDA-YYARD	1 (0.16)	20 (2.08)	13.22	2.52	1.17E-02
9 DCAA-YYDRD/DCAA-YYDRD	7 (1.1)	2 (0.21)	0.19	-2.08	3.77E-02
10 DCAA-FYDED/DRAA-YHDGD	8 (1.26)	2 (0.21)	0.17	-2.28	2.28E-02
11 DKAA-YYDRD/DQAA-YYARD	8 (1.26)	1 (0.1)	0.08	-2.35	1.87E-02
12 DQDA-YYDRD/DRAA-YHDGD	9 (1.42)	1 (0.1)	0.07	-2.48	1.32E-02
13 DCAA-YSARD/DCAA-YYDRD	16 (2.52)	6 (0.62)	0.25	-2.91	3.58E-03
14 DRAA-YHVG/DRAA-YHVG	13 (2.04)	3 (0.31)	0.15	-2.94	3.33E-03
15 DRAA-YHDGD/DRAA-YHVG	13 (2.04)	2 (0.21)	0.10	-3.01	2.62E-03
16 DCAA-YSARD/DRAA-YHDGD	16 (2.52)	4 (0.42)	0.17	-3.22	1.28E-03
17 DCAA-YYDRD/DRAA-YHVG	16 (2.52)	2 (0.21)	0.08	-3.32	8.86E-04
18 DCAA-YYDRD/DRAA-YHDGD	17 (2.67)	3 (0.31)	0.12	-3.43	6.02E-04
19 DRAA-FYDGD/DRAA-YHVG	25 (3.93)	1 (0.1)	0.03	-3.56	3.68E-04
20 DQAA-YYARD/DRAA-FYDGD	29 (4.56)	6 (0.62)	0.14	-4.44	9.18E-06
21 Rare Genotypes (<5 copies)	77 (12.11)	35 (3.64)	0.30	-5.90	3.69E-09
22 DCAA-FYDED/DRAA-FYDGD	8 (1.26)				
23 DCAA-YSARD/DRAA-FYDGD	21 (3.3)				
24 DCAA-YYDRD/DKAA-YSARG	5 (0.79)				
25 DCAA-YYDRD/DKAA-YYDRD	7 (1.1)				
26 DCAA-YYDRD/DRAA-FYDGD	22 (3.46)				
27 DCAA-YYDRD/DRAA-YHVR	9 (1.42)				
28 DKAA-YSARG/DQAA-YYDRD	6 (0.94)				
29 DKAA-YSARG/DRAA-FYDGD	14 (2.2)				
30 DKAA-YSARG/DRAA-YHDGD	11 (1.73)				
31 DKAA-YYDRD/DRAA-FYDGD	6 (0.94)				
32 DQDA-YYDRD/DRAA-FYDGD	8 (1.26)				
33 DRAA-FYDGD/DRAA-FYDGD	17 (2.67)				
34 DRAA-FYDGD/DRAA-YHDGD	16 (2.52)				
35 DRAA-FYDGD/DRAA-YHVR	10 (1.57)				

**Table 3.** Genotypic association analysis of type 1 diabetes with DQA1-DQB1 motifs of ( $\alpha$ 1,  $\alpha$ 44,  $\alpha$ 157,  $\alpha$ 196,  $\beta$ 9,  $\beta$ 30,  $\beta$ 57,  $\beta$ 70,  $\beta$ 135): estimated frequency (%) among control and patient, estimated odds ratios, Z-score, and p-value, corresponding serotype, and equivalent allelic groups. Listed are genotypes with p-values less than 0.05. Genotypic association results for all genotypes are listed in Table S2. The logistic regression analysis used the offset to compute odds ratio with the "virtual genotype" under the null as the reference.

expected, its homozygote had a significant association with susceptibility to T1D with OR = 5.21 ( $p = 8.97 \times 10^{-4}$ ), with an effect that was nearly doubled that of the heterozygote with neutral motifs (OR = 2.54). The heterozygote "DCAA-YSARD" (DQ2.5)/ "DQAA-YYARD" (DQ8.1) was associated with markedly increased susceptibility to T1D (OR = 14.52,  $p = 1.99 \times 10^{-10}$ ). Similarly, the heterozygote "DCAA-YSARD" (DQ2.5)/ "DQDA-YYARD" (DQ8.1) was associated with an even greater susceptibility (OR = 24.93,  $p = 2.95 \times 10^{-3}$ ), although it had a lower genotype frequency. Last, the heterozygote with resistant motifs had insignificant association with T1D (OR = 0.69,  $p = 0.35$ ), probably due to additive effect from both susceptibility and resistance motifs.

Pairing the motif "DQAA-YYARD" with neutral motifs, itself, "DQDA-YYARD" and resistant motifs, showed similar effects, albeit with some differences. The heterozygote with the neutral motifs captured the pure

First Motif	Second Motif	Control	patient	Coef	OR	SE	Z	p
DCAA-YSARD (DQ2.5)	Neutral	21 (3.3)	47 (4.89)	0.93	2.54	0.44	2.11	3.48E-02
	DCAA-YSARD	10 (1.57)	46 (4.78)	1.65	5.21	0.50	3.32	8.97E-04
	DQAA-YYARD	21 (3.3)	269 (27.96)	2.68	14.52	0.42	6.36	1.99E-10
	DQDA-YYARD	1 (0.16)	22 (2.29)	3.22	24.93	1.08	2.97	2.95E-03
DQAA-YYARD (DQ8.1)	Resistant	87 (13.68)	53 (5.51)	-0.37	0.69	0.39	-0.94	3.48E-01
	Neutral	34 (5.35)	122 (12.68)	1.40	4.07	0.40	3.47	5.14E-04
	DQAA-YYARD	5 (0.79)	97 (10.08)	3.09	21.99	0.58	5.33	9.64E-08
	DQDA-YYARD	1 (0.16)	20 (2.08)	3.12	22.67	1.08	2.88	3.99E-03
DQDA-YYARD (DQ8.1)	Resistant	86 (13.52)	162 (16.84)	0.76	2.13	0.38	2.00	4.51E-02
	Neutral	1 (0.16)	13 (1.35)	2.69	14.73	1.10	2.45	1.42E-02
	DQDA-YYARD	0 (0)	0 (0)					
Neutral	Resistant	10 (1.57)	18 (1.87)	0.71	2.04	0.53	1.34	1.79E-01
	Neutral	17 (2.67)	15 (1.56)	-0.13	0.88	0.35	-0.35	7.24E-01

**Table 4.** Results from the univariate association analysis for three high risk DQ motifs (DQ2.5: DCAA-YSARD, DQ8.1: DQAA-YYARD and DQ8.1: DQDA-YYARD), after accounting for all neutral and resistant motifs. Numbers and frequency percentages of genotype groups are listed for controls and patients, and so are estimated coef (log odds ratio), standard error, Z score, and p-value. The reference genotype is the specific genotype versus all others.

contribution of the motif “DQAA-YYARD” to T1D, with a significant association with susceptibility to T1D (OR = 4.07,  $p = 5.14 \times 10^{-4}$ ). The homozygote had much elevated association with T1D susceptibility (OR = 21.99,  $p = 9.64 \times 10^{-9}$ ). The heterozygote with another DQ8.1 “DQDA-YYARD” was also highly associated with increased susceptibility (OR = 22.67,  $p = 3.99 \times 10^{-3}$ ). Last, the heterozygote with resistant motifs had much weaker association (OR = 2.13,  $p = 0.014$ ), because resistant motifs compromised some of disease associations but not to the same extent (ORs from 0.03 to 0.35, Table 2).

Examining in the same manner another DQ8.1 motif, motif 10, i.e. “DQDA-YYARD”, its heterozygote with neutral motifs had significant association with T1D susceptibility (OR = 14.73,  $p = 0.0142$ ), which captured its sole contribution. There were no homozygotes in this relatively large study population, since the expected genotype frequency is in the order of  $1.44 \times 10^{-3}$ . The heterozygote with resistant motifs showed insignificant association with T1D risk (OR = 2.04,  $p = 0.18$ ), probably due to relatively low genotypic frequency.

By estimated odds ratios, the heterozygote “DCAA-YSARD/DQAA-YYARD” had an OR of 14.52, which was slightly greater than the product of individual contributions 10.34 ( $= 2.54 \times 4.07$ ) from “DCAA-YSARD/neutral” and “DQAA-YYARD/neutral”, respectively. Similarly, the heterozygote “DCAA-YSARD/DQDA-YYARD” had an OR of 24.93, which was slightly less than the product of individual contributions 37.41 ( $= 2.54 \times 14.73$ ) from “DCAA-YSARD/neutral” and “DQDA-YYARD/neutral”.

#### Expression in trans of DQA1 and DQB1 motifs enriches cellular diversity of DQ heterodimers but has limited associations with T1D.

Heterodimeric formation of the DQ molecule with doubly heterozygous DQA1 and DQB1 genotypes could produce four different DQ molecules: two cis-formations with alleles on the same chromosome and two trans-formations across homologous chromosomes, which may not be observed experimentally, as some are structurally prohibited<sup>18</sup> (see also Fig. 1). For homozygotes or semi-homozygotes, DQA1-B1 motifs on different chromosomes, formed trans-DQ motifs (whenever structurally allowed) that were the same as those cis-DQ motifs. In total, in trans-formation created 73 DQ motifs, of which 48 were new -DQ motifs, several of which were structurally forbidden<sup>18–20</sup> (see Supplementary Table S3). In a systematic assessment of all trans-motifs, we evaluated potential conditional association of each trans-motif with T1D, after adjusting for all cis-motifs. After excluding trans-motifs with fewer than five copies, we reported conditional associations with 45 trans-motifs. Most of trans-motifs did not exhibit any meaningful adjusted associations, with a few exceptions, e.g., “DQDA-YYARD” ( $p = 0.01$ ) (see Supplementary Table S3). The many permitted trans-motifs are indicated in the column “Prmt”. In the table, estimated coefficient quantifies adjusted log odds ratio for cis-motifs, together with estimated standard error, Z-score and p-value. Should there be an interest in marginally estimated odds ratio, it was possible to take the ratio of patient and control trans-motif frequencies. For example, the trans-motif 24 “DQAA-YSARD” had the marginal odds ratio of 8.37, but the adjusted odds ratio was 1.26, which was not significantly different from null ( $p$ -value = 0.387), because associated cis-motifs explained most of the risk associations. In summary, trans-motifs did not appear to exhibit additional associations, after accounting for all 25 cis-motifs. However, this empirical result from the conditional adjust-

ID	DR	DQ	Control	Patient	OR	Z	p
1	*01	DRAA-YHVGD	136 (10.69)	173 (8.99)	0.84	-1.59	1.11E-01
2	*03	DCAA-YSARD	153 (12.03)	485 (25.21)	2.10	9.26	1.96E-20
3	*04:01	DQAA-YYARD	71 (5.58)	593 (30.82)	5.52	17.69	4.94E-70
4	*04:01	DQDA-YYARD	7 (0.55)	45 (2.34)	4.25	3.94	7.98E-05
5	*04:01	DQDA-YYDRD	41 (3.22)	55 (2.86)	0.89	-0.57	5.67E-01
6	*04:02	DQAA-YYARD	7 (0.55)	27 (1.40)	2.55	2.25	2.47E-02
7	*04:03	DQAA-YYARD	11 (0.86)	2 (0.10)	0.12	-3.31	9.18E-04
8	*04:04	DQAA-YYARD	61 (4.8)	146 (7.59)	1.58	3.10	1.91E-03
9	*04:05	DQDA-YYARD	3 (0.24)	23 (1.20)	5.07	2.97	2.99E-03
10	*04:07	DQDA-YYDRD	9 (0.71)	1 (0.05)	0.07	-3.28	1.05E-03
11	*07	DKAA-YSARG	78 (6.13)	41 (2.13)	0.35	-5.76	8.47E-09
12	*07	DKAA-YYDRD	36 (2.83)	5 (0.26)	0.09	-6.36	1.99E-10
13	*08	DCAA-FYDED	55 (4.32)	76 (3.95)	0.91	-0.52	6.00E-01
14	*09	DQDA-YYDRD	10 (0.79)	28 (1.46)	1.85	1.72	8.57E-02
15	*11	DCAA-YYDRD	89 (7.00)	20 (1.04)	0.15	-8.99	2.57E-19
16	*12	DCAA-YYDRD	29 (2.28)	14 (0.73)	0.32	-3.67	2.45E-04
17	*13	DCAA-YYDRD	11 (0.87)	5 (0.26)	0.30	-2.38	1.73E-02
18	*13	DRAA-YHDGD	91 (7.15)	39 (2.03)	0.28	-7.28	3.38E-13
19	*13	DRAA-YHVRD	52 (4.09)	80 (4.16)	1.02	0.10	9.22E-01
20	*14	DRAA-YHDGD	20 (1.60)	4 (0.21)	0.13	-4.46	8.09E-06
21	*15	DRAA-FYDGD	197 (15.49)	8 (0.42)	0.03	-16.24	2.49E-59
22	*16	DRAA-YHSGD	14 (1.10)	13 (0.68)	0.61	-1.29	1.97E-01

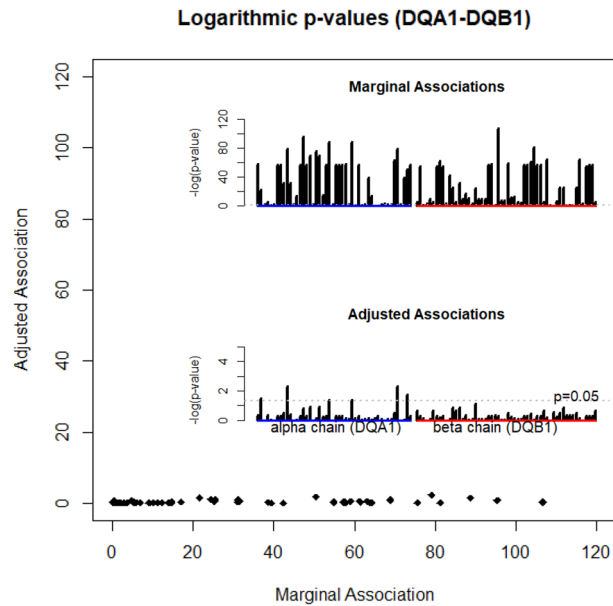
**Table 5.** Haplotype association analysis of type 1 diabetes with HLA-DRB1 and DQA1/B1: estimated haplotype frequency (%) among control and patient, estimated odds ratio, Z-score, and p-value, for those haplotypes with five or more copies.

ment did not exclude their potential biological associations between trans DQ motifs and T1D as the cellular expression of DQ heterodimers are enriched.

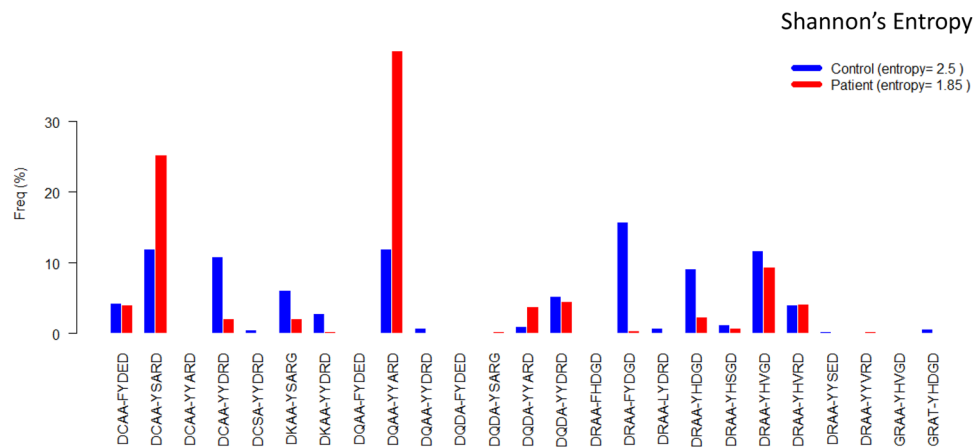
**DQ associations may be influenced by its linkage-disequilibrium with DRB1.** *DQA1* and *DQB1* were genetically nearby *DRB1*, and in high LD with each other, i.e., *DR3* and *DQ2.5* were linked, and so were *DR4* and *DQ8.1*. With DQ motifs of nine residues, we performed haplotype association analysis of DQ motifs and *DRB1*, with six *DR4* subtypes specified by four-digit resolution. Clearly, *DR3* was uniquely linked with the motif “DCAA-YSARD” (*DQ2.5*), and their joint haplotype had significant association with T1D susceptibility (OR = 2.10,  $p = 1.96 \times 10^{-20}$ ) (Table 5). Similarly, *DRB1\*04:02*, *\*04:03*, *\*04:04*, *\*04:05* and *\*04:07* were linked DQ motifs with variable associations. Hence, these empirical haplotype associations prohibited separating *DRB1* and DQ associations. However, the same *DRB1\*04:01* was on the same haplotypes with DQ motifs “DQAA-YYARD”, “DQDA-YYARD”, or “DQDA-YYDRD”. These three haplotypes had variable associations (OR = 5.52, 4.25 and 0.89,  $p = 4.94 \times 10^{-70}$ ,  $7.98 \times 10^{-5}$  and 0.57, respectively), implying that DQ motifs dictated T1D association in the background of *DRB1\*04:01*. *DRB1\*04:03* reduced T1D susceptibility when in LD with *DQ8.1* (OR = 0.12, while for the *DQ8.1* alleles collectively OR = 3.34) even though the number of patients and controls with this haplotype was very low. In the case of the *DRB1\*15* molecule, the OR is identical to the very tightly linked *DQA1\*01:02-B1\*06:02* haplotype. In fact, the number of patients and controls is nearly identical concerning two nearly congruent groups.

**Motifs of nine residues account for all DQ associations with T1D.** Besides selected residues on alpha and beta chains, there were 95 polymorphic amino acids across DQ haplotypes in this study population. Due to LD, many residues were expected to have significant marginal associations. We evaluated these marginal associations via a logistic regression of T1D with polymorphic residues. By the log-likelihood ratio test, p-values were computed for each residue, and were shown in the top panel of the bar plot in the Fig. 2. Evidently, many p-values greatly exceeded the threshold of 0.001, and some were approaching  $10^{-100}$  or even less, showing strong marginal associations with T1D. By adjusting for the motifs of selected nine residues, we repeated the same logistic regression analysis, estimating p-values from log-likelihood ratio test (the lower bar plot of the Fig. 2). Indeed, most p-values were insignificant, i.e.,  $p > 0.05$ , except that several p-values were slightly less than 0.05.





**Figure 2.** Logarithmic p-values computed for association of type 1 diabetes with the individual residue, with (y-axis) and without (x-axis) adjusting for DQA1-DQB1 motifs.

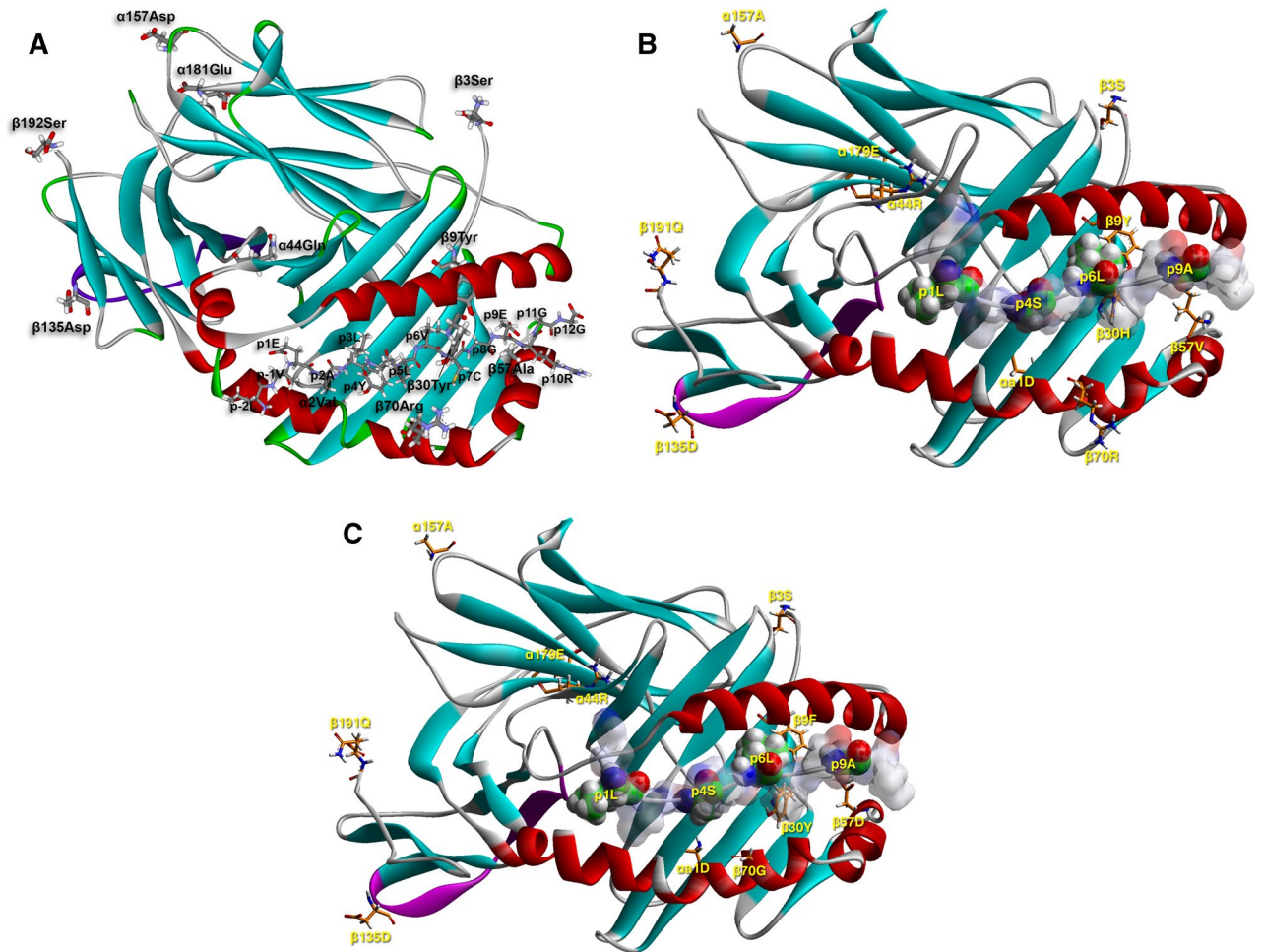


**Figure 3.** Frequencies of all DQ motifs among 636 controls and 962 patients involved in BDD study.

We plotted adjusted p-values against marginal p-values as an x–y plot, showing that adjusting motifs accounted for most of the marginal associations (Fig. 2).

**DQ motifs are less diverse among patients than among controls.** DQ motifs of selected nine residues retained all susceptibility and resistance to T1D associations that were identified by DQ haplotypes but had reduced polymorphisms due to merging of some DQ haplotypes, especially those rare DQ haplotypes with relatively common ones. The frequencies of 25 DQ motifs among controls and patients (Fig. 3) indicates that the two motifs “DCAA-YSARD” and “DQAA-YYARD”, corresponding to DQ2.5 and DQ8.1, had exceptionally higher frequencies among patients than controls. Further, controls tended to have modest frequencies across many motifs. For this reason, DQ motifs among controls, with a Shannon’s entropy value of 2.50, tended to be more diverse than those among patients with an entropy value of 1.85, where the entropy was used to quantify diversity of genetic polymorphisms<sup>21</sup>.

**Motifs of nine residues capture DQ associations with islet autoantibodies.** Our study measured six autoantibodies among newly diagnosed patients, i.e., IAA, GADA, IA-2A, ZnT8RA, ZnT8WA and ZnT8QA, which were scored as positive or negative. Focusing on 11 DQ motifs observed more than 10 times among patients, we evaluated the association of each DQ motif with the positivity for each type of autoantibody (Table 6). T1D subjects with the motif “DCAA-YSARD”, corresponding to DQ2.5, tended to have an increased



**Figure 4.** A–C. Molecular depiction of a T1D-susceptible HLA-DQ molecule (DQ8—InsB11-23/B24Gly; 1jk8.pdb, A.), a neutral one (DQB1\*06:04—InsB5-15, B.), and a T1D-resistant molecule (DQB1\*06:02—InsB5-15, C.) in complex with a so-restricted antigenic peptide from insulin (parts B. and C. obtained upon molecular simulation reported in<sup>23</sup> based on the crystal structure of the HLA-DQB1\*06:02—hypocretin 1–13 complex, 1uvq.pdb<sup>69</sup>). Binding has been demonstrated to the respective epitopes, and in the first case several CD4<sup>+</sup> T cell clones, specific to this peptide and restricted to HLA-DQ8 have been isolated from patients. The binding register has not been questioned, yet there is evidence that cognate TCRs from T cell clones isolated from patients with T1D recognize an epitope with a shifted register by one residue towards the peptide C-terminus<sup>38</sup>. The HLA-DQ-insulin peptide complexes are depicted in their secondary structure formation ( $\alpha$ -helix in red,  $\beta$ -sheet in turquoise,  $\beta$ -turn, random coil or any other form, such as poly-proline II helix of the antigenic peptide backbone, in grey), and the  $\beta$ 134–148 CD4-binding stretch in purple. The HLA-DQ AA residues in question are depicted in stick form (atom color convention: carbon, grey; oxygen, red; nitrogen, blue; sulfur, yellow; hydrogen, white). The antigenic peptide is shown in stick form in (A) (thinner sticks, with the same color convention), and in space-filling form in (B.) and (C.) with the p1Leu, p4Ala, p6Leu and p9Ala anchor residues opaque, and the remaining residues non-transparent surfaces colored by atom charge (red, negative; blue positive, partial charges colored with hues in-between), in order to appreciate the positioning and orientation of the selected  $\beta$ -chain residues; same atomic color conventions as in (A) (except for carbon that is in green).

frequency of GADA (OR = 1.56;  $p = 6.35 \times 10^{-8}$ ) but a depressed frequency of IA-2A (OR = 0.59,  $p = 6.55 \times 10^{-11}$ ). Among subjects with the motif “DQAA-YYARD” corresponding to DQ8.1, patients had a reversed association pattern, i.e., depressed GADA (OR = 0.87;  $p = 4.50 \times 10^{-3}$ ) but elevated IA-2A (OR = 1.52,  $p = 2.18 \times 10^{-9}$ ) frequency. Importantly, subjects with the motif “DQDA-YYARD”, corresponding to a slightly different DQ8.1 allele, had an elevated IA-2A positivity (OR = 5.67,  $p = 5.42 \times 10^{-4}$ ), while its association with GADA was neutral (OR = 0.91,  $p = 0.67$ ). Lastly, the motif “DRAA-YHVRD” corresponding to DQ6.4 (itself a neutral allele regarding T1D susceptibility) seemed to associate with increased frequency of ZnT8A (RA, WA and QA; OR = 1.92, 1.72 and 1.57,  $p = 6.25 \times 10^{-3}$ , 0.017 and 0.041, respectively), while having negative associations with GADA and IA-2A frequencies. In order to delineate possible pathways for T1D manifestation and resistance via islet-autoantigen-specific epitope presentation restricted by susceptible and resistant motifs, respectively, we have chosen three examples of: a susceptible (DQ8—InsB11-23), a neutral (DQB1\*06:04—InsB5-15) and a resistant (DQB1\*06:02—InsB5-15) heterodimer (Fig. 4A–C, and Supplementary Fig. S3A–C)<sup>22,23</sup>.

ID	DQA1-DQB1 Motif	Sero-type	IAA			GADA			IA2A			ZnT8RA			ZnT8WA			ZnT8QA		
			OR	HS	p	OR	HS	p	OR	HS	p	OR	HS	p	OR	HS	p	OR	HS	p
1	DCAA-FYDED	DQ4	1.36	1.36	1.73E-01	0.59	-2.47	1.34E-02	0.80	-0.88	3.76E-01	0.76	-1.29	1.97E-01	0.81	-0.94	3.49E-01	0.86	-0.64	5.23E-01
2	DCAA-YSARD	DQ2.5	0.93	-0.86	3.92E-01	1.56	5.41	6.35E-08	0.59	-6.53	6.55E-11	0.88	-1.69	9.17E-02	0.89	-1.60	1.10E-01	0.82	-2.43	1.53E-02
3	DCAA-YYDRD		1.02	0.06	9.49E-01	1.48	1.11	2.68E-01	1.38	0.70	4.81E-01	1.29	0.76	4.45E-01	1.13	0.38	7.00E-01	1.21	0.57	5.71E-01
4	DCAA-YSARG	DQ2.2	0.99	-0.04	9.65E-01	1.96	1.93	5.34E-02	1.42	0.81	4.18E-01	0.99	-0.03	9.74E-01	0.88	-0.40	6.86E-01	1.29	0.82	4.15E-01
5	DQAA-YYARD	DQ8.1	1.02	0.44	6.63E-01	0.87	-2.84	4.50E-03	1.52	5.98	2.18E-09	0.94	-1.32	1.85E-01	1.10	1.94	5.21E-02	1.04	0.68	4.95E-01
6	DQDA-YYARD	DQ8.1	1.04	0.16	8.70E-01	0.91	-0.42	6.72E-01	5.67	3.46	5.42E-04	1.40	1.44	1.50E-01	1.06	0.28	7.82E-01	0.99	-0.05	9.57E-01
7	DQDA-YYDRD		1.85	2.97	3.02E-03	1.36	1.35	1.75E-01	1.40	1.14	2.54E-01	1.75	2.48	1.30E-02	1.11	0.51	6.11E-01	1.46	1.79	7.32E-02
8	DRAA-YHDGD		1.10	0.30	7.63E-01	0.83	-0.61	5.39E-01	0.73	-0.93	3.50E-01	0.85	-0.56	5.78E-01	0.64	-1.49	1.36E-01	0.94	-0.19	8.47E-01
9	DRAA-YHSGD		1.42	0.67	5.04E-01	1.27	0.44	6.63E-01	0.36	-2.01	4.40E-02	0.28	-2.34	1.95E-02	0.23	-2.49	1.28E-02	0.31	-1.64	1.00E-01
10	DRAA-YHVG D		0.69	-2.37	1.79E-02	0.74	-2.20	2.81E-02	0.91	-0.58	5.64E-01	1.22	1.40	1.61E-01	0.91	-0.71	4.79E-01	0.96	-0.29	7.71E-01
11	DRAA-YHVRD	DQ6.4	0.81	-0.88	3.79E-01	0.54	-2.80	5.17E-03	0.60	-2.11	3.47E-02	1.92	2.73	6.25E-03	1.72	2.39	1.68E-02	1.57	2.05	4.07E-02

**Table 6.** Association analysis of DQ motifs with elevated autoantibody levels among all patients (with five or more subjects): estimated odds ratio, haplotype score, and p-value, across all motifs with six autoantibodies (IAA, GADA, IA2A, ZnT8RA, ZnT8WA and ZnT8QA).

Motif 1	Motif 2	Control	Patient	Coef	SE	Z	P-value	OR	AF
Neutral	Neutral	17 (2.67)	15 (1.56)	0.00				1.00	
	Resistant	108 (16.98)	54 (5.61)	-0.57	0.39	-1.45	1.47E-01	0.57	-0.07
	Risk	59 (9.28)	187 (19.44)	1.28	0.38	3.33	8.80E-04	3.59	0.19
Resistan	Resistant	231 (36.32)	16 (1.66)	-2.54	0.44	-5.80	6.53E-09	0.08	-0.25
	Risk	183 (28.77)	234 (24.32)	0.37	0.37	1.01	3.13E-01	1.45	0.11
Risk	Risk	38 (5.97)	456 (47.40)	2.61	0.39	6.65	2.91E-11	13.60	0.43

**Table 7.** Estimated attributable fractions for all susceptible and resistant DQ motif genotypes: genotype frequency (%) among control and patient, estimated coef, standard error, Z-score, p-value, estimated odds ratio and attributable fraction. Notations and Conventions: As in Supplementary Fig. S1. Additional symbols are: For the putative homodimerization patch of  $\beta$ 49-55 residues are marked with @ at the bottom of the sequence and highlighted in light green. Homodimerization of HLA-DQ has never been observed in any of the respective crystal structures. Yet, F(ab')<sub>2</sub> fragment of anti-HLA-DQ antibodies differentially ligate HLA-DQ molecules on the surface of human monocytes and lead to differential (compared to HLA-DR and -DP) downstream signal transduction.

**Estimated attributable fractions of risk and protection from respectively susceptibility and resistance motifs.** To quantify an overall T1D risk or protection from DQ motifs, we computed attributable risk and attributable resistance, respectively (methodological details of which are given in “Methods”). Briefly, we pooled DQ motifs into three groups: resistant, neutral or risk. Each individual had pairs of resistant, neutral or risk motifs. Using a logistic regression model on genotypes, we estimated regression coefficients corresponding to five genotypes, treating homozygous neutral motifs as the reference (Table 7). Motif frequencies were estimated from the controls (Table 7). In comparison with the reference genotype, in subjects with the homozygous resistant motif, the attributable resistance fraction was 25%, while its heterozygote with neutral motif had the resistance fraction of 7%. On the other hand, for subjects with a homozygous risk motif, the attributable risk fraction was 43%, and its heterozygote with the neutral motif had 19% attributable risk fraction. The heterozygote of resistant and risk motifs had 11% attributable risk fraction.

**Structural properties of nine DQ residues.** Of the nine residues, least is known about the putative structural role of  $\alpha$ 1Asp/Gly, the second residue of the mature HLA-DQ alpha chain. It is a residue without counterpart in HLA-DR alpha chains<sup>21</sup>. Of the remaining residues  $\alpha$ 44 is near pocket 1, while  $\beta$ 9,  $\beta$ 30,  $\beta$ 57, and  $\beta$ 70 are part of pockets 9, 6, 9 and 4 respectively, thus being among the main determinants for the binding of autoantigenic epitopes to the respective HLA-DQ molecules<sup>20,22–24</sup>. Further,  $\beta$ 70 may also be considered as a most probable TCR contact residue and  $\beta$ 57 as a potential such residue<sup>23</sup>. Residue  $\beta$ 135 is part of the  $\beta$ 134-148 peptide stretch shown to bind to co-receptor CD4 both in HLA-DR- as well as in mouse I-A-expressing APCs<sup>24,25</sup>. Along the same lines,  $\alpha$ 157 is a participant in the formation of the cognate TCR-induced pMHCII homodimer of heterodimers<sup>26,27</sup>, while  $\alpha$ 196 is within the membrane of the APC, contributes to signal transmission to such a cell, once engagement with cognate TCR is made<sup>28</sup>.

## Discussion

HLA-DQ genes are known to significantly contribute to susceptibility to islet autoimmunity and further to the progression to clinical onset of T1D, independently or jointly with HLA-DRB1. The major finding through exploration of the protein sequences of DQA1 and DQB1<sup>12,13</sup> was that we identified, respectively, three and seven critical amino acid residues associated with T1D onset<sup>14,15</sup>. Integrating both sets of critical residues resulted in

four alpha residues ( $\alpha 1$ ,  $\alpha 44$ ,  $\alpha 157$ ,  $\alpha 196$ ) and five beta residues ( $\beta 9$ ,  $\beta 30$ ,  $\beta 57$ ,  $\beta 70$ ,  $\beta 135$ ) forming motifs that was associated with either susceptibility or resistance to T1D. The importance of both DQ $\alpha$  and DQ $\beta$  chains in determining T1D susceptibility and resistance is demonstrated not just in the respective motif consisting of 4  $\alpha$ -chain and 5  $\beta$ -chain residues but also in all DQ $\alpha\beta$  combinations revealed by next generation targeted sequencing<sup>11,12</sup>. The deep sequencing approach reveals the importance of HLA-DQA-B combinations in T1D both susceptibility and resistance to considerations of -DQB or -DQA alone. This is evident, for example, in motif 5 (DQA1\*02:01-B1\*02:02), a T1D mildly resistant allele compared to the rare motif 9 (DQA1\*03:02-B1\*02:02) with only six copies among T1D patients, despite their sharing of an identical  $\beta$ -chain.

When interpreting association results of nine identified residues with respect to their binding properties, we should be mindful of two residue sets: one set includes all monomorphic and conserved residues, and another set includes those monomorphic residues in their respective clusters of DQ haplotypes. The former set of residues essentially defines the overall DQ molecule structure, and the latter set defines substructures of DQ heterodimers, e.g., DQ2.5 or DQ8.1 heterodimers. Hence, they remain structurally and functionally important, even though they may have limited associations with T1D risk.

Specifically, the newly identified DQB1\*02:02, previously scored as DQB1\*02:01 with low resolution methods, was found in a T1D-resistant combination (motif 5, Table 2, DQA1\*02:01-B1\*02:02) in 78 controls and 41 patients. The same  $\beta$ -chain by contrast, forms a rare haplotype in combination with a different  $\alpha$ -chain, i.e. DQA1\*03:02-B1\*02:02, found only in six patients but in no controls. Similar striking findings were observed with DQB1\*03:01 and DQB1\*03:03. When either of them pairs with DQA1\*03:01 they form rare haplotypes found only in 9 controls (motif 8, Table 2). Yet when the first of them pairs with DQA1\*03:03 and the second with DQA1\*03:02 they result in neutral haplotypes, found in 67 controls and 88 patients (motif 11, Table 2).

We also demonstrate that the DQ motifs associated with T1D appeared to be mediated through the frequency of islet autoantibodies. In particular, subjects with the DQ\*DCAA-YSARD motif, corresponding to DQ2.5, more often had GADA ( $p = 6.35 \times 10^{-8}$ ) than IA-2A ( $p = 6.55 \times 10^{-11}$ ). Meanwhile, subjects with DQ\*DQAA-YYARD and DQ\*DQDA-YYARD, corresponding to DQ8.1, more often had IA-2A ( $p = 2.18 \times 10^{-9}$  and  $5.42 \times 10^{-4}$ , respectively), but less often GADA ( $p = 4.50 \times 10^{-3}$ ). Other motifs had more variable associations with the different autoantibodies. To assess motif associations with the number of autoantibodies at diagnosis, we compared patients with 2, 3 or 4 autoantibodies with patients with 0 or 1 autoantibody, which revealed a single significant association (see Supplementary Table S4). In this neutral/mildly resistance motif, DQ\*DQDA-YYDRD, the association with T1D was significant only when all four autoantibodies were present. These data suggest that the presence of all four autoantibodies overcame the resistance mechanism(s) of the DQ7 and DQ9 haplotypes (both  $\beta 57\text{Asp}^+$ ) represented in this motif<sup>29</sup>. Hence, it cannot be excluded that other non-HLA genetic factors important to the pathogenesis contributed to autoantigen spreading and appearance of multiple autoantibodies.

It is important to realize that antigen-specificity in HLA-DQ- and -DR-directed autoimmune response (as well as resistance to it) is paramount for an organ-specific autoimmune disease, such as T1D. This is reflected by the observation that 4 of the 9 selected HLA-DQ residues are in anchoring pockets involved directly in antigen binding ( $\beta 9$ ,  $\beta 30$ ,  $\beta 57$ ,  $\beta 70$ ), another one ( $\alpha 1$ ) is near pocket 1, and two of the first four residues ( $\beta 57$ ,  $\beta 70$ ) in probable cognate TCR engagement and activation<sup>23,30,31</sup>. The remaining polymorphic residues (except  $\alpha 1$  about which there is hardly any information), are involved in other accessory functions of the HLA-DQ heterodimer:  $\alpha 135$  in CD4 co-receptor binding, as part of the  $\beta 134$ -148 stretch<sup>32,33</sup>,  $\alpha 157$  in the formation of the putative homodimer of heterodimers that presumably facilitates CD4 binding<sup>34,35</sup> and consequent cognate TCR activation and HLA-DQ signal transduction in the specific T cell and APC respectively<sup>23</sup>. Last,  $\alpha 196$  is within the single membrane spanning domain of the respective chain and is involved in consequent signal transduction after the crucial step of pMHCII recognition by TCR, as just described<sup>23,36</sup>. The structural depictions of a susceptible, a resistant and a neutral HLA-DQ heterodimer happen to have bound peptide fragments from the same antigenic stretch of the insulin B-chain peptide 5–23<sup>21,22,37,38</sup>. Within the insulin peptide B5–23, the InsB11–23 epitope binds to DQ8 and is recognized by cognate CD4<sup>+</sup> Th1 cells<sup>37,38</sup> and the InsB5–15 epitope that binds in an identical register to the T1D-resistant heterodimer containing DQB1\*06:02, as well as the neutral DQB1\*06:04<sup>22</sup>. While not all epitopes have been compared in one-to-one comparisons between a resistant and a neutral/susceptible allele, it has already been shown that DQB1\*06:02 can out-compete DQ8 for the epitope InsB5–23 by forming the DQB1\*06:02—InsB6–14 complex, with the peptide shown only as the core nonamer<sup>22</sup>. The same is probably the case for the neutral allele DQB1\*06:04, which binds the same preproinsulin epitopes as DQB1\*06:02, albeit at slightly higher IC50 values, i.e. with lower affinity<sup>22</sup>. It will naturally be of interest to determine how newly-identified risk and resistant alleles in the Swedish population interact with T1D autoantigenic epitopes<sup>12,13</sup>.

Two recent studies have added significant knowledge to the area of HLA-DR-DQ-restricted proinsulin and GAD65 epitopes<sup>26,27</sup>. The first study deals with proinsulin-specific epitopes and their cognate CD4<sup>+</sup> T cell clones found in healthy children homozygous for either HLA-DR3-DQ2, DR4-DQ8 or being DR3-DQ2/DR4-DQ8 heterozygotes. It is based on the observation that IAA represents the first appearing autoantibody in about 50% of children with these high-risk HLA-DR-DQ haplotypes who later on developed additional islet autoantibodies followed later by clinical onset of T1D<sup>39,40</sup>. Remarkably, the epitopes were identified using PBMCs from these healthy children and relevant antigen-presenting cells supplemented with various proinsulin antigenic peptides. IL-2 combined with IL-7 was used to stimulate both CD4<sup>+</sup> T effectors as well as CD4<sup>+</sup> T regulatory cells (Tregs). The second study deals with the effects of GAD65-alum treated newly-diagnosed T1D patients by stimulating frozen and thawed PBMCs in culture with whole GAD65 and antigen-specific bifunctional Th1/Th2 cell lines generated subsequently<sup>27</sup>. It is noteworthy that the T cell lines generated from heterozygous HLA-DR3-DQ2/DR4-DQ8 individuals were greatly restricted in the number of recognized epitopes, distinctly fewer than the sum of those recognized by T cell lines restricted to either HLA-DR3-DQ2 or HLA-DR4-DQ8<sup>27</sup>. The specific HLA class II restricting element(s) of the respective CD4<sup>+</sup> T cell lines and clones were not determined. Interestingly, a recent study showed that GAD65- and IGRP (islet glucose-reactive protein)-specific CD4<sup>+</sup> T effector

and T regulatory cells (Tregs) were present at higher frequencies in the peripheral blood of HLA-DRB1\*15:01-DQA1\*01:02-B1\*06:02 positive individuals (i.e. negatively associated with T1D) compared to subjects with neutral or T1D-susceptible alleles<sup>28</sup>. Furthermore, cloned Tregs carrying these specificities and restricted to HLA-DRB1\*15:01 suppressed so-restricted and specific Teffectors. It will be of interest to determine whether HLA-DQ alleles linked with resistance to T1D can also give rise to similar Tregs. A similar effect has already been noted in the case of Goodpasture's syndrome with higher frequency of Tregs specific for the Collagen IV autoantigen  $\alpha 3_{135-145}$  in HLA-DRB1\*01:01+ (disease-resistant) individuals, compared to HLA-DRB1\*15:01+ (disease-susceptible) individuals, regarding the same autoantigen specificity, albeit with binding register shifted by one residue<sup>41</sup>. This would extend the "epitope-stealing" effect, long hypothesized for T1D and already documented for HLA-DQA1\*01:02-B1\*06:02 over HLA-DQ8-specific proinsulin epitopes<sup>25,37,42</sup>. If DQA1\*01:02-B1\*06:02-restricted Tregs can be shown to have similar suppressive effects as those restricted to DRB1\*15:01 above, then the contributions of the former molecule to T1D resistance would be multiple. The amino acid motifs identified in the present study should prove useful to identify these mechanisms.

All our considerations thus far, have assumed equal levels of membrane HLA-DQ protein expression on APCs. It was shown over two decades ago, that the density of APC, the level of expression of HLA-DQ on the membrane of such APC, as well as the level of soluble peptide antigen present, would influence the proliferation of restricted and specific CD4+ T effector cells<sup>43</sup>. While the upstream regions of HLA-DQA and -DQB genes are replete with regulatory nucleotide sequences, very few studies exist on the regulation of HLA-DQ membrane protein expression by the various different alleles<sup>44-46</sup>. Notwithstanding the fact that in a disease situation (i.e. the islet  $\beta$  cells under autoimmune attack, as well as the proximal pancreatic lymph nodes) the cytokine and chemokine milieu is an unknown and may differentially influence the expression of HLA-DR-DQ heterodimers even in an allele-specific manner. The situation has become more complex by the fact that human, as well as mouse, islet  $\beta$  cells, are closely approximated by islet macrophages that receive antigenic peptide fragments from these  $\beta$  cells and possess MHCII molecules through which such antigens are presented<sup>47,48</sup>. It is also of interest that Tregs, shown to be suppressive over T effector cells specific for T1D autoantigens, seem to be consistently absent from islets in autopsy material from newly diagnosed T1D patients<sup>49,50</sup>.

We have limited our discussion to possible autoantigenic epitopes and their interactions with HLA-DQ molecules and alleles that confer susceptibility or resistance to, or are neutral to T1D. Unfortunately nothing can be said about the interaction of diabetogenic pHLA-DQ heterodimers with cognate TCRs, as there is no single complex with solved 3-dimensional coordinates. Such complex would provide information about the possible engagement of the amino acid residue motifs identified in the present study. Using TCR-HLA-DR-DQ complexes with CNS antigenic epitopes in multiple sclerosis as a guide (the only available such complexes from organ-specific autoimmune diseases), the TCR orientation with respect to pMHCII is not the canonical diagonal, but expected to be off-diagonal<sup>24</sup>. The fact that this interaction, that leads to autoimmunity, is different from the canonical one is also reflected in the irregular form of immunological synapse formed by autoimmune CD4+ T cell clones from patients with multiple sclerosis and T1D, compared to such clones specific for influenza virus peptides: there was no accumulation of pMHCII complexes in the central Supramolecular Activation Cluster of the synapse; nor for that matter, any transport of TCR-pMHCII complexes into this central cluster of the synapse<sup>51</sup>. In another example, the key HLA class II molecule accounting for over 90% of patients with celiac disease is DQ5.2 (A1\*05:01-B1\*02:01), while the very closely related DQ2.2 (A1\*02:01-B1\*02:02) is neutral. The  $\alpha 22Y/F$  dimorphism (DQ5.2 compared to DQ2.2, respectively), differentiate between the binding at p3 of gluten peptides in DQ2.2 compared to by DQ2.5. The latter combinations select for a very focused set of cognate TCRs that are also pathogenic for self resulting in small intestinal tissue destruction<sup>52</sup>. The small set of gluten peptide-DQ2.2 complexes surprisingly resulted in a wide set of cognate TCRs, none of which were pathogenic<sup>52</sup>.

A potential weakness to the present studies is that all children are Swedish, only 8% in the present cohort were born to parents born in another country along with the grandparents<sup>29</sup> and that the environmental exposure would be relatively homogeneous. It cannot be excluded that the pattern of amino acid residues observed as well as the presence of autoantibodies at the time of clinical onset may reflect a Swedish environment. The exposures of these children diagnosed with T1D are not only long term but also multiplex as the incidence rate for IAA-first in DR4-DQ8 children showed a peak at 1-3 years of age in the TEDDY study as related to prolonged shedding of enterovirus<sup>40,53,54</sup> and other studies have shown that viral exposure of islet autoantibody positive children may accelerate progression to clinical onset<sup>55</sup>. It cannot be excluded that the two potential endotypes in T1D etiology and multiple environmental exposures leading to the eventual clinical onset of T1D have been important to the frequency of specific amino acid motifs.

Three additional limitations of the study are worth noting. The current study uses the retrospective case-control approach, drawing patients from clinics around Sweden and geographically matched controls. Inevitably, hospital-referred patterns may bias patient selection, and some controls, if any, may have not developed T1D at the time of the study. However, both issues are unlikely to alter the results of assessing the genetic susceptibility to T1D, because HLA genes are time-invariant, T1D is relatively uncommon among adults and healthcare system provides universal care in Sweden. Another limitation is associated with measuring autoantibodies among patients at the time of diagnosis. As noted above, autoantibodies are important biomarkers predicting T1D onset<sup>54,56</sup> but have variable temporal patterns. In the absence of observing such temporalities, the current investigation of genetic associations with autoantibodies is inevitably limited and needs to be cautiously interpreted, even though observed associations still remain meaningful. The third limitation concerns the possibility of genetic heterogeneity due to immigrants that may confound the association results. In the BDD study, 8% of patients are children born to immigrants<sup>16,29</sup>. Even though immigrant status information is not controlled, it is expected that normal controls may have a comparable percentage of immigrants. Given the relatively small percentage, the mixture of immigrants with native Sweden is not expected to substantially alter association results observed here.

Recently, a novel reductionist conditional approach was used to search for critical residues<sup>57</sup>. The key idea is to repeat residue-specific association tests, while adjusting (conditioning on) selected residues<sup>57</sup>. Their conditional analysis led to identification of  $\beta 57$  as a critical residue for T1D. However, this approach fails to differentiate between empirically observed associations that result from genetic causal associations or from LD-induced associations. Hence their results have invited much controversy. In contrast to the reductionist approach, our approach, being a holistic, starts with motifs of all polymorphic residues, retaining their sequence structure, and eliminates those residues that do not associate with T1D within their respective clusters. Consequently, our primary results concern T1D associations with motifs of selected residues, and these motifs explain nearly all DQ haplotype associations with T1D. By the same reason, this holistic approach is probably preferred to other residue-specific association methods.

In T1D, recent data in the TEDDY study<sup>53</sup> suggest that prolonged shedding of enterovirus B in young DR4-DQ8.1 positive children precede the appearance of insulin autoantibodies (IAA) as a biomarker for autoimmunity against the beta cells. In older children, prolonged shedding of adenovirus F in DR3-DQ2.5 positive children preceded the appearance of GAD65 autoantibodies (GADA) as a first biomarker for beta cell autoimmunity. The intersection between prolonged virus exposure and appearance of a first autoantibody marker is yet to be dissected. However, it cannot be excluded that the motifs described in the present paper contribute to antigen presentation that involves a competition between a viral peptide or either (pro)insulin or GAD65 to trigger an autoimmune response rather than a viral immune response. It is therefore of interest that the susceptibility motifs were more diverse among controls than among patients (entropy = 2.5 versus 1.85, respectively). Out of 25 unique motifs, three motifs “DCAA-YSARD”, “DQAA-YYARD”, and “DQDA-YYARD” exhibited significant susceptibility to T1D (OR = 2.10, 3.34 and 3.71,  $p = 1.96 \times 10^{-20}$ ,  $2.69 \times 10^{-72}$  and  $1.53 \times 10^{-6}$ , respectively) with one uncommon motif “DQDA-YSARG” was observed seven times only among patients. On the other hand, eight motifs “DCAA-YYDRD”, “DKAA-YSARG”, “DKAA-YYDRD”, “DRAA-FYDGD”, “DRAA-LYDRD”, “DRAA-YHDGD”, “DRAA-YHVGD” and “GRAT-YHDGD” were negatively associated with autoimmunity against  $\beta$  cells resulting in T1D (OR = 0.19, 0.35, 0.09, 0.03, 0.07, 0.25, 0.80 and 0.08,  $p = 2.32 \times 10^{-25}$ ,  $8.47 \times 10^{-9}$ ,  $1.99 \times 10^{-10}$ ,  $2.25 \times 10^{-60}$ ,  $1.14 \times 10^{-3}$ ,  $7.72 \times 10^{-19}$ , 0.0312 and  $6.38 \times 10^{-3}$ , respectively), with two motifs “DCSA-YYDRD” and “DQAA-YYDRD” respectively observed 7 and 9 times among only controls. It is obvious that the lower HLA-DQ motif entropy among T1D patients reflects the fact that susceptibility to T1D requires certain structural features on the part of the susceptibility-conferring HLA-DQ molecules. These certainly would have to do with the relevant autoantigenic epitopes, as well as the molecules that the T1D-causing pMHCII complexes interact with such cognate TCRs and CD4 co-stimulatory molecules, and perhaps other molecules in the respective signal transducing pathways.

## Methods

**Study design and populations.** We used a case-control study design to evaluate genetic associations with T1D. A total of 962 patients (cases) were from the nation-wide Swedish Better Diabetes Diagnosis (BDD) study<sup>16,58,59</sup>, which involves participation in 2005–2010 of all 42 pediatric clinics in Sweden. The American Diabetes Association and World Health Organization criteria were used for the diagnosis of diabetes and to classify the disease<sup>60</sup>. Included patients had one or several autoantibodies against either insulin (IAA), GAD65 (GADA), IA-2 (IA-2A), and three variants (amino acid 325 being either R, W or Q, ZnT8-RA, ZnT8-WA or ZnT8-QA, respectively) at the time of clinical diagnosis<sup>16,58</sup>. Patients diagnosed with diabetes at 9 months–18 years of age were sequentially enrolled in the BDD study<sup>16,29,61</sup>. A total of 636 nation-wide and geographically representative controls were analyzed at the same time<sup>62</sup>. The demographic characteristics of the BDD patients are detailed elsewhere<sup>16,29,59</sup>. Approximately 8% of BDD participants are immigrants<sup>16,29</sup>. The informed consent was taken from participants/guardians for young participants (<18 years old). The Karolinska Institute Ethics Board approved the BDD study (2004/1:9).

All experiments (DNA extraction, genotyping, measuring islet autoantibodies) and analytic methods are carried out in clinical laboratories (CLIA certified or equivalent in Sweden), following relevant guidelines and regulations.

**DNA extraction.** The plasmid Max isolation kit (Qiagen, Bothell, Washington, USA) was used to isolate DNA according to the manufacturer’s instructions from frozen whole blood samples of cases and controls.

**HLA next generation targeted sequencing (NGTS) analysis.** The NGTS HLA typing approach utilized PCR based amplification of HLA and sequencing using Illumina MiSeq technology as described in detail<sup>63,64</sup>. Briefly, the laboratory steps consisted of consecutive amplicon-based PCR with bar coding incorporated in the PCRs for individual sample tracking followed by application to the MiSeq. Robust assays for each of the target loci for all class II genes were purchased from Scisco Genetics Inc., Seattle WA (<https://sciscogenetics.com>). This system employs amplicons individually extending 4–500 bp segments covering each of exons 1–4 of HLA-DQA1 and -DQB1 and MiSeq read depths of over 100-fold coverage for each amplicon were obtained. The analytical tools to define haplotypes and genotypes were made available as part of the genotyping kit assay from Scisco Genetics Inc. (Seattle, WA). To date these tools have been tested—with 100% accuracy—on >2000 control samples genotyped with the present NGS approach<sup>63,64</sup>.

**Amino acid sequences of DQA1 and DQB1 alleles.** As NGTS sequenced DNA nucleotides of selected exons and used sequences to infer DQA1 and DQB1 alleles, we determined HLA-DQA1 and -DQB1 genotypes at the high resolution of 6 digits or higher, based on the HLA nomenclature of IMGT (<https://raw.githubusercontent.com/ANHIG/IMGTHLA/Latest/alignments/DQA1>). Further, we determined physical positions of individual aminoacids in the alpha chain from codon  $\alpha 1$  to  $\alpha 232$ , and in the beta chain from codon  $\beta 1$  to  $\beta 237$ .

We have adopted the numbering system as first suggested<sup>65</sup>, and later improved<sup>20</sup>, that essentially allows for structural equivalence among residues of various MHC II alleles, regardless of gene locus or species, and based on the structure of the first published MHC II allele HLA-DR1<sup>34,66</sup>.

**Islet autoantibodies.** GADA, IA-2A, IAA, and three variants of ZnT8A (ZnT8-RA, ZnT8-WA or ZnT8-QA, respectively) were determined in quantitative radio-binding assays using in house standards to determine levels as previously described in detail<sup>16,58</sup>.

**Cis- and trans-heterodimer through interlocus recombination.** High LD enables us to haplotype HLA-DQA1 and DQB1 genotypes with exceptionally high posterior probabilities (minimum value > 0.97). Let  $(h/\bar{h}) = (\hat{a}_A\hat{a}_B/\hat{a}_A\hat{a}_B)$  denote a pair of -DQA1 and -DQB1 haplotypes, in which  $\hat{a}_A\hat{a}_B$  and  $\hat{a}_A\hat{a}_B$  are two cis-haplotypes (or cis-heterodimers). In contrast, two trans-haplotypes are  $\hat{a}_A\hat{a}_B$  and  $\hat{a}_A\hat{a}_B$ . In case that -DQA1 and -DQB1 are doubly homozygous, all four haplotypes are identical. For a semi-heterozygous, two trans-haplotypes are identical to two cis-haplotypes. Thirdly, a doubly heterozygous DQA1 and DQB1 would lead to four different haplotypes. Under the assumption that the "interlocus recombination" is a completely random process, alleles of DQA1 and DQB1 are always recombined to form cis- and trans-heterodimers within cells, even though some trans-heterodimers may be structurally prohibited or unexpressed<sup>18</sup>. Lastly, it is expected that some trans-heterodimers, from doubly heterozygotes, are novel and unobserved in the general population, while many others could be observed in the general population. There are still some that were probably expressed at the mRNA level but were not permitted to form proteins<sup>18</sup>.

**Conditional analysis strategy.** In the empirical association analysis of T1D disease outcome ( $Y$ ) with one primary covariable  $X$  in light of another correlated variable  $Z$  that may potentially confound the association of interest, the conventional approach is to adjust the confounding variable in the association analysis via the following logistic regression model for modeling the disease probability:

$$\Pr(Y = 1|X, Z) = 1/\{1 + \exp[-\text{adjust}(X, \beta) - \text{test}(Z, \gamma)]\},$$

in which the first component "adjust( $X, \beta$ )" is a function of the primary covariates to be adjusted for and is indexed by the parameter  $\beta$ , and the second component "test( $Z, \gamma$ )" is a function of the test covariates indexed by the parameter  $\gamma$ . Here the first step is to fit a logistic regression model with the "adjust( $X, \beta$ )". Then fixing this component, the second step is to systematically test the disease association with the second component test( $Z, \gamma$ ). Despite the intuitive appeal of the conditional analysis for adjusting the primary variable, it also has several intrinsic limitations, since this adjustment is empirical rather than not biological (see "Discussion"). Nevertheless, it is helpful to perform such a conditional analysis for the following scenarios:

- (1) Assessing disease associations with trans-motifs, after adjusting for the cis-motifs present.
- (2) Assessing disease associations with individual residues, after adjusting for cis-motifs

**Shannon's entropy.** HLA-DQ motifs are polymorphic, and have a degree of diversity. Because of their categorical nature, we use the information entropy, a basic information unit in the information theory, which is also known as Shannon's entropy introduced by Claude Shannon<sup>21</sup>. Let  $f_1, f_2, \dots, f_q$  denote allelic frequencies for  $q$  alleles. The corresponding Shannon's entropy is defined as

$$\text{Entropy} = - \sum_{j=1}^q f_j \log(f_j).$$

**Attributable risk/protection fractions.** Attributable fraction is used to quantify an overall fraction of risk (protection) that can be accounted by a binary exposure factor, i.e., exposed versus non-exposed, and depends on the magnitude of the association and frequency of the exposure in the population<sup>67</sup> which can be computed as:

$$AF_{jk} = \frac{f_{jk}(OR_{jk} - 1)}{1 + f_{jk}|OR_{jk} - 1|},$$

which may be referred to as attributable risk fraction ( $AF > 0$ ) or attributable protective fraction ( $AF < 0$ ).

**Structural properties of HLA-DQ residues.** All depictions of determined structures of various HLA-DQ alleles have been carried out using the WebLabViewer v. 3.5 and the DSViewerPro v. 6.0, 3-D molecular rendering software of Accelrys (currently Dassault Systèmes, BIOVIA, San Diego, CA, USA, <https://www.3ds.com/products-services/biovia>), based on coordinates freely available in the Protein Data Bank. The DQ8 (A1\*03:01-B1\*03:02)—insulin B11-23 complex structure is obtained from the determined crystal structure<sup>22</sup>. The modeled structures of the complexes of DQB1\*06:02 (A1\*01:02-B1\*06:02)—insulin B4-16, and DQB1\*06:04 (A1\*01:02-B1\*06:04)—insulin B4-16 are obtained from the crystal structure of DQB602-hypocretin complex, as already described<sup>23</sup>.

**Statistical analysis and software.** Upon extracting aminoacids for each DQ haplotype, we compute distances between all unique sequences pairwise by the stringdist function of R package, with the Levenshtein distance, and then apply the hierarchical clustering algorithm using the hclust function of R package (version 4.0.2, <https://www.r-project.org>) with the agglomeration method of ward.D2, i.e., clustering two sequences closer together, if their sum of squared distances is relatively small. For the association analysis, we use the logistic regression function in “glm” of R package. To assess their associations with qualitatively determined elevations of islet autoantibodies (IAA, GADA, IA-2A, ZnT8RA, ZnT8WA, and ZnT8QA), we also use haplo.cc. To address the instability of “small sample sizes” in the diplotypic association analysis, we compare frequency counts of one diplotype versus those of all other diplotypes combined, and compute the corresponding odds ratios, 95% confidence interval, and the Fisher’s exact test p-values.

With respect to p-values and multiple comparison issues, we took an unconventional approach, presenting the p-values without the multiple comparison correction. While the conventional approach was to compute corrected p-values for number of comparisons by, say the Bonferroni correction or false discovery rate<sup>68</sup>, its intension was to control the overall false positive error rates of all comparisons. In the current context with the analysis of HLA-DQ haplotypes/diplotypes, the roles of both genes in T1D were implicated in multiple empirical and functional studies. Further, the empirical haplotype association analysis suggested that DQ associations are highly significant at the 5% level, even if multiple comparisons with 45 comparisons are considered. Hence, the haplotypic and diplotypic association analyses aimed to uncover which haplotypes/diplotypes may explain the overall association, i.e., our explorations were in the alternative hypothesis domain. Finally, p-values with no multiple comparison corrections have clear and simple interpretations, free from varying haplotypes/diplotypes in various explorations. Throughout the analyses, we used the threshold of 0.05 to highlight those p-values to be positively or negatively associated, while being mindful that some p-values, close to 0.05, could be falsely labeled. Naive applications of any multiple comparisons without varying correction factors could obscure the interpretations of results.

When computing haplotype-specific odds ratios, we intentionally avoided choosing a specific reference haplotype, because any choice of a particular reference haplotype could have undesired implications in the interpretations of the results. Instead, we considered two different strategies. One way was to compute the odds ratio of one haplotype versus all other haplotypes. Alternatively, we choose a “virtual haplotype” that has no association with T1D. As expected, the ratio of its corresponding haplotype frequency among patients over that among controls equals one. Hence, the expected odds ratio for any given haplotype is thus computed as the ratio of two observed haplotype frequencies among patients and controls<sup>14</sup>.

Received: 11 September 2020; Accepted: 4 March 2021

Published online: 23 April 2021

## References

- Knip, M. *et al.* Environmental triggers and determinants of type 1 diabetes. *Diabetes* **54**(Suppl 2), S125–136 (2005).
- Levin, L. *et al.* Analysis of HLA genes in families with autoimmune diabetes and thyroiditis. *Hum. Immunol.* **65**, 640–647. <https://doi.org/10.1016/j.humimm.2004.02.026> (2004).
- Redondo, M. J. *et al.* Heterogeneity of type 1 diabetes: Analysis of monozygotic twins in Great Britain and the United States. *Diabetologia* **44**, 354–362. <https://doi.org/10.1007/s001250051626> (2001).
- Rewers, M. & Ludvigsson, J. Environmental risk factors for type 1 diabetes. *Lancet* **387**, 2340–2348. [https://doi.org/10.1016/S0140-6736\(16\)30507-4](https://doi.org/10.1016/S0140-6736(16)30507-4) (2016).
- Bradfield, J. P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7**, e1002293. <https://doi.org/10.1371/journal.pgen.1002293> (2011).
- WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447**, 661–678. <https://doi.org/10.1038/nature05911> (2007).
- Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864. <https://doi.org/10.1038/ng2068> (2007).
- Howson, J. M., Walker, N. M., Clayton, D. & Todd, J. A. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes Obes. Metab.* **11**(Suppl 1), 31–45. <https://doi.org/10.1111/j.1463-1326.2008.01001.x> (2009).
- Pociot, F. & Lernmark, A. Genetic risk factors for type 1 diabetes. *Lancet* **387**, 2331–2339. [https://doi.org/10.1016/S0140-6736\(16\)30582-7](https://doi.org/10.1016/S0140-6736(16)30582-7) (2016).
- Ilonen, J. *et al.* Genetic susceptibility to type 1 diabetes in childhood: Estimation of HLA class II associated disease risk and class II effect in various phases of islet autoimmunity. *Pediatr. Diabetes* **17**(Suppl 22), 8–16. <https://doi.org/10.1111/pedi.12327> (2016).
- Mychaleckyj, J. C. *et al.* HLA genotyping in the international Type 1 Diabetes Genetics Consortium. *Clin. Trials* **7**, S75–S77. <https://doi.org/10.1177/1740774510373494> (2010).
- Zhao, L. P. *et al.* Motifs of three HLA-DQ amino acid residues (alpha44, beta57, beta135) Capture Full association with the risk of type 1 diabetes in DQ2 and DQ8 children. *Diabetes* **69**, 1573–1587. <https://doi.org/10.2337/db20-0075> (2020).
- Zhao, L. P. *et al.* Next generation HLA sequence analysis uncovers seven HLA-DQ amino acid residues and six motifs resistant to childhood type 1 diabetes. *Diabetes* **69**, 2523–2535 (2020).
- Zhao, L. P. *et al.* Eleven amino acids of HLA-DRB1 and fifteen amino acids of HLA-DRB3, 4, and 5 include potentially causal residues responsible for the risk of childhood type 1 diabetes. *Diabetes* **68**, 1692–1704. <https://doi.org/10.2337/db19-0273> (2019).
- Zhao, L. P. *et al.* Next-generation sequencing reveals that HLA-DRB3, -DRB4, and -DRB5 may be associated with islet autoantibodies and risk for childhood type 1 diabetes. *Diabetes* **65**, 710–718. <https://doi.org/10.2337/db15-1115> (2016).
- Delli, A. J. *et al.* Zinc transporter 8 autoantibodies and their association with SLC30A8 and HLA-DQ genes differ between immigrant and Swedish patients with newly diagnosed type 1 diabetes in the Better Diabetes Diagnosis study. *Diabetes* **61**, 2556–2564. <https://doi.org/10.2337/db11-1659> (2012).
- Rich, S. S. *et al.* Overview of the Type 1 Diabetes Genetics Consortium. *Genes Immun* **10**(Suppl 1), S1–4. <https://doi.org/10.1038/gene.2009.84> (2009).
- Kwok, W. W., Kovats, S., Thurtell, P. & Nepom, G. T. HLA-DQ allelic polymorphisms constrain patterns of class II heterodimer formation. *J Immunol* **150**, 2263–2272 (1993).



19. Sollid, L. M., Qiao, S. W., Anderson, R. P., Gianfrani, C. & Koning, F. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics* **64**, 455–460. <https://doi.org/10.1007/s00251-012-0599-z> (2012).
20. Bondinas, G. P., Moustakas, A. K. & Papadopoulos, G. K. The spectrum of HLA-DQ and HLA-DR alleles, 2006: a listing correlating sequence and structure with function. *Immunogenetics* **59**, 539–553. <https://doi.org/10.1007/s00251-007-0224-8> (2007).
21. Zhao, J., Boerwinkle, E. & Xiong, M. An entropy-based statistic for genomewide association studies. *Am. J. Hum. Genet.* **77**, 27–40. <https://doi.org/10.1086/431243> (2005).
22. Lee, K. H., Wucherpfennig, K. W. & Wiley, D. C. Structure of a human insulin peptide-HLA-DQ8 complex and susceptibility to type 1 diabetes. *Nat. Immunol.* **2**, 501–507. <https://doi.org/10.1038/88694> (2001).
23. Ettinger, R. A., Papadopoulos, G. K., Moustakas, A. K., Nepom, G. T. & Kwok, W. W. Allelic variation in key peptide-binding pockets discriminates between closely related diabetes-protective and diabetes-susceptible HLA-DQB1\*06 alleles. *J. Immunol.* **176**, 1988–1998. <https://doi.org/10.4049/jimmunol.176.3.1988> (2006).
24. Rossjohn, J. *et al.* T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**, 169–200. <https://doi.org/10.1146/annurev-immunol-032414-112334> (2015).
25. van Lummel, M. *et al.* Epitope stealing as a mechanism of dominant protection by HLA-DQ6 in type 1 diabetes. *Diabetes* **68**, 787–795. <https://doi.org/10.2337/db18-0501> (2019).
26. Ihanntola, E. L. *et al.* Characterization of proinsulin T cell epitopes restricted by type 1 diabetes-associated HLA class II molecules. *J. Immunol.* **204**, 2349–2359. <https://doi.org/10.4049/jimmunol.1901079> (2020).
27. Arif, S. *et al.* GAD-alum immunotherapy in type 1 diabetes expands bifunctional Th1/Th2 autoreactive CD4 T cells. *Diabetologia* <https://doi.org/10.1007/s00125-020-05130-7> (2020).
28. Wen, X. *et al.* Increased islet antigen-specific regulatory and effector CD4(+) T cells in healthy individuals with the type 1 diabetes-protective haplotype. *Sci. Immunol.* <https://doi.org/10.1126/sciimmunol.aax8767> (2020).
29. Delli, A. J. *et al.* Type 1 diabetes patients born to immigrants to Sweden increase their native diabetes risk and differ from Swedish patients in HLA types and islet autoantibodies. *Pediatr. Diabetes* **11**, 513–520. <https://doi.org/10.1111/j.1399-5448.2010.00637.x> (2010).
30. Beringer, D. X. *et al.* T cell receptor reversed polarity recognition of a self-antigen major histocompatibility complex. *Nat. Immunol.* **16**, 1153–1161. <https://doi.org/10.1038/ni.3271> (2015).
31. Painter, C. A. & Stern, L. J. Conformational variation in structures of classical and non-classical MHCII proteins and functional implications. *Immunol. Rev.* **250**, 144–157. <https://doi.org/10.1111/imr.12003> (2012).
32. Cammarota, G. *et al.* Identification of a CD4 binding site on the beta 2 domain of HLA-DR molecules. *Nature* **356**, 799–801. <https://doi.org/10.1038/356799a0> (1992).
33. König, R., Shen, X. & Germain, R. N. Involvement of both major histocompatibility complex class II alpha and beta chains in CD4 function indicates a role for ordered oligomerization in T cell activation. *J. Exp. Med.* **182**, 779–787. <https://doi.org/10.1084/jem.182.3.779> (1995).
34. Stern, L. J. *et al.* Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* **368**, 215–221. <https://doi.org/10.1038/368215a0> (1994).
35. Cochran, J. R., Cameron, T. O. & Stern, L. J. The relationship of MHC-peptide binding and T cell activation probed using chemically defined MHC class II oligomers. *Immunity* **12**, 241–250. [https://doi.org/10.1016/s1074-7613\(00\)80177-6](https://doi.org/10.1016/s1074-7613(00)80177-6) (2000).
36. Dixon, A. M. & Roy, S. Role of membrane environment and membrane-spanning protein regions in assembly and function of the Class II Major Histocompatibility complex. *Hum. Immunol.* **80**, 5–14. <https://doi.org/10.1016/j.humimm.2018.07.004> (2019).
37. Eerligh, P. *et al.* Functional consequences of HLA-DQ8 homozygosity versus heterozygosity for islet autoimmunity in type 1 diabetes. *Genes Immunity* **12**, 415–427. <https://doi.org/10.1038/gene.2011.24> (2011).
38. Chow, I. T. *et al.* Discriminative T cell recognition of cross-reactive islet-antigens is associated with HLA-DQ8 transdimer-mediated autoimmune diabetes. *Sci. Adv.* **5**, eaaw9336. <https://doi.org/10.1126/sciadv.aaw9336> (2019).
39. Ilonen, J. *et al.* Patterns of beta-cell autoantibody appearance and genetic associations during the first years of life. *Diabetes* **62**, 3636–3640. <https://doi.org/10.2337/db13-0300> (2013).
40. Krischer, J. P. *et al.* The 6 year incidence of diabetes-associated autoantibodies in genetically at-risk children: the TEDDY study. *Diabetologia* **58**, 980–987. <https://doi.org/10.1007/s00125-015-3514-y> (2015).
41. Ooi, J. D. *et al.* Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. *Nature* **545**, 243–247. <https://doi.org/10.1038/nature22329> (2017).
42. Nepom, G. T. A unified hypothesis for the complex genetics of HLA associations with IDDM. *Diabetes* **39**, 1153–1157 (1990).
43. Reichstetter, S. *et al.* MHC-peptide ligand interactions establish a functional threshold for antigen-specific T cell recognition. *Hum. Immunol.* **60**, 608–618. [https://doi.org/10.1016/s0198-8859\(99\)00038-5](https://doi.org/10.1016/s0198-8859(99)00038-5) (1999).
44. Beaty, J. S., Sukiennicki, T. L. & Nepom, G. T. Allelic variation in transcription modulates MHC class II expression and function. *Microbes Infect./Inst. Pasteur* **1**, 919–927 (1999).
45. Beaty, J. S., West, K. A. & Nepom, G. T. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. *Mol Cell Biol* **15**, 4771–4782. <https://doi.org/10.1128/mcb.15.9.4771> (1995).
46. Andersson Svard, A. *et al.* Decreased HLA-DQ expression on peripheral blood cells in children with varying number of beta cell autoantibodies. *J. Transl. Autoimmun.* **3**, 100052. <https://doi.org/10.1016/j.jtauto.2020.100052> (2020).
47. Vomund, A. N. *et al.* Beta cells transfer vesicles containing insulin to phagocytes for presentation to T cells. *Proc. Natl. Acad. Sci. U S A* **112**, E5496–5502. <https://doi.org/10.1073/pnas.1515954112> (2015).
48. Wan, X. *et al.* The MHC-II peptidome of pancreatic islets identifies key features of autoimmune peptides. *Nat. Immunol.* **21**, 455–463. <https://doi.org/10.1038/s41590-020-0623-7> (2020).
49. Willcox, A., Richardson, S. J., Bone, A. J., Foulis, A. K. & Morgan, N. G. Analysis of islet inflammation in human type 1 diabetes. *Clin. Exp. Immunol.* **155**, 173–181. <https://doi.org/10.1111/j.1365-2249.2008.03860.x> (2009).
50. Sebastiani, G. *et al.* Regulatory T-cells from pancreatic lymphnodes of patients with type-1 diabetes express increased levels of microRNA miR-125a-5p that limits CCR2 expression. *Sci. Rep.* **7**, 6897. <https://doi.org/10.1038/s41598-017-07172-1> (2017).
51. Schubert, D. A. *et al.* Self-reactive human CD4 T cell clones form unusual immunological synapses. *J. Exp. Med.* **209**, 335–352. <https://doi.org/10.1084/jem.20111485> (2012).
52. Ting, Y. T. *et al.* A molecular basis for the T cell response in HLA-DQ2.2 mediated celiac disease. *Proc. Natl. Acad. Sci. USA* **20**, 5. <https://doi.org/10.1073/pnas.1914308117> (2020).
53. Vehik, K. *et al.* Prospective virome analyses in young children at increased genetic risk for type 1 diabetes. *Nat Med* **25**, 1865–1872. <https://doi.org/10.1038/s41591-019-0667-0> (2019).
54. Krischer, J. P. *et al.* Predicting islet cell autoimmunity and Type 1 diabetes: An 8-Year TEDDY study progress report. *Diabetes Care* **42**, 1051–1060. <https://doi.org/10.2337/dc18-2282> (2019).
55. Stene, L. C. *et al.* Enterovirus infection and progression from islet autoimmunity to type 1 diabetes: the Diabetes and Autoimmunity Study in the Young (DAISY). *Diabetes* **59**, 3174–3180. <https://doi.org/10.2337/db10-0866> (2010).
56. Ziegler, A. G. *et al.* Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children. *JAMA J. Am. Med. Assoc.* **309**, 2473–2479. <https://doi.org/10.1001/jama.2013.6285> (2013).
57. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat. Genet.* **47**, 898–905. <https://doi.org/10.1038/ng.3353> (2015).

58. Andersson, C. *et al.* Triple specificity of ZnT8 autoantibodies in relation to HLA and other islet autoantibodies in childhood and adolescent type 1 diabetes. *Pediatr. Diabetes* **14**, 97–105. <https://doi.org/10.1111/j.1399-5448.2012.00916.x> (2013).
59. Persson, M. *et al.* The Better Diabetes Diagnosis (BDD) study: A review of a nationwide prospective cohort study in Sweden. *Diabetes Res. Clin. Pract.* **140**, 236–244. <https://doi.org/10.1016/j.diabres.2018.03.057> (2018).
60. Classification and Diagnosis of Diabetes. *Diabetes Care* **38**, S8–S16. <https://doi.org/10.2337/dc15-S005> (2015).
61. Carlsson, A. *et al.* Low risk HLA-DQ and increased body mass index in newly diagnosed type 1 diabetes children in the Better Diabetes Diagnosis study in Sweden. *Int. J. Obes. (Lond)* **36**, 718–724. <https://doi.org/10.1038/ijo.2011.122> (2012).
62. Gyllenberg, A. *et al.* Age-dependent variation of genotypes in MHC II transactivator gene (CIITA) in controls and association to type 1 diabetes. *Genes Immun.* **13**, 632–640. <https://doi.org/10.1038/gene.2012.44> (2012).
63. Smith, A. G. *et al.* Next generation sequencing to determine HLA class II genotypes in a cohort of hematopoietic cell transplant patients and donors. *Hum. Immunol.* **75**, 1040–1046. <https://doi.org/10.1016/j.humimm.2014.08.206> (2014).
64. Nelson, W. C. *et al.* An integrated genotyping approach for HLA and other complex genetic systems. *Hum. Immunol.* <https://doi.org/10.1016/j.humimm.2015.05.001> (2015).
65. Fremont, D. H., Monnaie, D., Nelson, C. A., Hendrickson, W. A. & Unanue, E. R. Crystal structure of I-Ak in complex with a dominant epitope of lysozyme. *Immunity* **8**, 305–317. [https://doi.org/10.1016/s1074-7613\(00\)80536-1](https://doi.org/10.1016/s1074-7613(00)80536-1) (1998).
66. Brown, J. H. *et al.* Three-dimensional structure of the human class II histocompatibility antigen HLA-DRI. *Nature* **364**, 33–39. <https://doi.org/10.1038/364033a0> (1993).
67. Porta, M. *A Dictionary of Epidemiology* 5th edn. (Oxford University Press, Oxford, 2008).
68. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**, 279–284. [https://doi.org/10.1016/s0166-4328\(01\)00297-2](https://doi.org/10.1016/s0166-4328(01)00297-2) (2001).
69. Siebold, C. *et al.* Crystal structure of HLA-DQ0602 that protects against type 1 diabetes and confers strong susceptibility to narcolepsy. *Proc Natl Acad Sci USA* **101**, 1999–2004. <https://doi.org/10.1073/pnas.0308458100> (2004).
70. Matsuoka, T., Tabata, H. & Matsushita, S. Monocytes are differentially activated through HLA-DR, -DQ, and -DP molecules via mitogen-activated protein kinases. *J. Immunol.* **166**, 2202–2208. <https://doi.org/10.4049/jimmunol.166.4.2202> (2001).
71. H. Miyadera, L. B. Bungener, S. Kusano, S. Yokoyama, K. Tokunaga, B. G. Hepkema, (2015) Questionable expression of unstable DQ heterodimer containing HLA-DQA1\*01:07. *Tissue Antigens* **86** (6):413–418

## Acknowledgements

The authors would also like to thank Dr. Gun Forsander at Department of Pediatrics, the Queen Silvia Children's Hospital, Sahlgrenska University Hospital, Gothenburg, Sweden and Dr. Sten A. Ivarsson Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden, for their clinical expertise in conduct of the BDD study. Equally, the authors would like to thank Anita Ramelius, Linda Faxius, Ida Jönsson and Rasmus Bennet all at the Department of Clinical Sciences, Lund University/CRC, in Malmö, Sweden for expert technical assistance.

## Author contributions

L.P.Z., Å.L., G.K.P., D.G. and S.A. researched and analyzed the data and wrote the manuscript. I.K. researched data (control subjects) and reviewed the manuscript. M.K., B.X., R.W., C.-W.P. and W.C.N. contributed to the next-generation sequencing, researched data and reviewed the manuscript. A.K.M. and G.P.B. carried out molecular simulations and graphical representations of select HLA-DR molecules. A.C.; H.E.L.; J.L.; M.P.; U.S.; E.Ö. and C.M. designed the BDD study, researched data, contributed to discussion and reviewed the manuscript. All authors used appropriate clinical, epidemiological, genetic, and statistical methods to carry out all investigation in accordance with the best practice in biomedical research per relevant guidelines and regulations. Å.L. is the guarantor of this work, had full access to all the data, and takes full responsibility for the integrity of data and accuracy of data analysis.

## Funding

Open access funding provided by Lund University. The study was supported by a grant (1R01DK117276) from the National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), a grant from the European Foundation for the Study of Diabetes (EFSD) Clinical Research Grants Programme 2013 and in part the Swedish Child Diabetes Foundation (Barndiabetesfonden), Swedish Foundation for Strategic Research (Dnr IRC15-0067), Swedish Research Council, Strategic Research Area (Dnr 2009–1039), the Skåne County Council for Research and Development as well as the Swedish Association of Local Authorities and Regions (SKL).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-86229-8>.

**Correspondence** and requests for materials should be addressed to L.P.Z.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021