ESHG

COMMENT

Evidence that *UBASH3* is a causal gene for type 1 diabetes

John A Todd[1]

In this issue of EJHG, Ge and Concannon [1], combined with results from their previous study [2], report that *UBASH3A*, which encodes Ubiquitin-associated and SH3 domain-containing protein A and functions as a negative regulator of NF-kB signalling in T cells on stimulation of the antigen T cell receptor [2], is almost certainly a causal gene in the chromosome 21q22.3 region for type 1 diabetes (T1D) and other autoimmune diseases (https://www.immunobase.org). Few genome-wide association study (GWAS) regions have yet converged on a causal gene yet, although as you will see below there are complexities, and there may be additional genes involved in the region.

The presence of the T1D-predisposing non-coding alleles of single nucleotide polymorphisms (SNPs) within the UBASH3 gene, is associated with its increased transcription thereby leading to reduced transcription of the interleukin-2 (IL-2) gene by effector T cells, leading to increased susceptibility to T1D [2]. These results are consistent with a model that T1D results from autoimmune destruction of the pancreatic insulin-producing beta cells in which deficiencies in IL-2 signalling and the functions of the autoimmune disease-protective T regulatory cells, which are wholly dependent IL-2 for their maintenance and function, are recognised as major aetiological pathways of this common multifactorial disorder. The latter consensus is a major justification for the ongoing trials of IL-2 therapy in T1D to improve the regulation of the immune and save some beta cells from autoimmune destruction and reduce or even halt the requirement for multiple daily insulin injections. These results provide further support for this model and the translation of ultra-low dose IL-2 into the T1D clinic [2, 3].

In the original T1D ImmunoChip SNP association study [4] four credible SNPs were identified in the *UBASH3A* region: rs11203202 C>G (where G is the minor, disease predisposing or risk allele), rs112203203 G>A, rs9981624 G>C and rs8004410 T>C. However, no evidence was presented for more than one association in the region and rs11203202 C>G was reported as the index SNP. The existence of multiple causal variants in a disease susceptibility chromosome region is common and will be become increasingly so as sample sizes increase. In the present study Ge and Concannon [1] investigated a fifth *UBASH3A* intron 9 non-coding SNP, rs1893592 A>C, because its minor C allele had been associated previously with protection from three other autoimmune diseases, or the major A allele is associated with susceptibility (https://www.immunobase.org/). They use the term "opposite", but it is just a switch from minor to major allele for susceptibility. They present some evidence from conditional and haplotype analyses that for the *UBASH3A* region there may be more than one non-coding causal variant, where rs1893592 may be still associated with T1D once the associations of rs112203203 and rs80054410 are taken into account (rs80054410 and rs11203203 association is probably the same signal since they are in almost completely linkage disequilibrium, $r^2 = 0.98$). They did not include rs11203202 or rs9981624 in their conditional analysis. Much larger case-control sample sets will be required to sort this out combining haplotype analysis and Bayesian modelling.

Alleles of these SNPs, rs11203203, rs11203202 and the newly analysed one, rs1893592 (which shows little linkage disequilibrium with the other two, $r^2 = 0.12$ and 0.02, respectively) have to be considered on the haplotypes they occur on when analysing disease association, or indeed, gene expression: a disease-protective allele of one credible SNP may occur on a haplotype that has the disease-susceptible allele of another credible SNP resulting in a possible neutral association of this haplotype with disease, or an eQTL or other trait. The authors carried out *UBASH3A* haplotype analysis in families and a simplified version (Table 1) shows the type of pattern of haplotype

✉ John A Todd
john.todd@well.ox.ac.uk

1 Wellcome Centre for Human Genetics, Nuffield Department of Medicine, Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK

associations observed if a region contains more than one causal variant. The informative haplotypes, H9 and H1, indicate that having susceptible alleles at both rs11203202 and rs11203203 overrides the protective effect of the minor C allele rs1893592 for H9, whereas for H1 being susceptible with the major C allele of rs1893592 neutralises the protective effects of rs11203202 and rs11203203. The statistical support presented for these differentiating these haplotype associations is not strong but coupled with their allele-specific mRNA expression data, which was obtained using specific *UBASH3A* PCR primers and quantitative PCR mRNA measurement, and consistent with previous mRNA expression data [5] (but see below), it is evident that the minor C allele of rs1893592 (or an allele of an untyped variant in linkage disequilibrium with it) is associated with decreased T1D disease susceptibility and lower *UBASH3A* expression, presumably leading to more *IL2* RNA and protein and increased protection from T1D. Note that their evidence for higher plasma IL-2 protein concentrations in the rs1893592 A/C heterozygotes is not as convincing as their mRNA data but their previous study on UBASH3A protein production indicates that alteration of *UBASH3A* mRNA expression can lead to protein production changes in the same direction [2].

However, on inspection of results in other databases, there is a peculiarity. In (https://molgenis58.target.rug.nl/bloodeqtlbrowser/)[6] the rs1893592 minor C allele is associated with increased, not decreased, *UBASH3A* expression in whole blood ($P = 6.4 \times 10^{-92}$). In the GTEx RNA-seq database [7], the C allele is also associated with increased expression of the gene, in whole blood ($P = 10^{-11}$) and in several tissues, and also in the database https://molgenis58.target.rug.nl/biosqtlbrowser/ [8, 9]. In another study in purified CD4$^+$ and CD8$^+$ T cells the C allele of rs1893592 was associated with increased *UBASH3A* expression [10]. *UBASH3A* is very specifically expressed in CD4$^+$ and CD8$^+$ T cells with very low levels of mRNA in other cells and tissues (http://biogps.org/#goto = welcome) such that one explanation is that if microarray/ RNA-seq or qPCR eQTL or allele-specific gene expression analyses are carried out with very low amounts of target mRNA artefactual results can result, owing to reverse transcriptase or PCR polymerase biases or random priming. They can be highly statistically significant, as many results owing to technical issues and batch effects are. Nevertheless, on average about 12% of blood cells are T cells and therefore it is difficult to understand how a true eQTL signal could be lost or reversed. Another explanation is that there is allele swapping, which is still unfortunately a problem and must always be carefully verified (by Sanger sequencing, for example, as Concannon and colleagues have done; although note that in Fig. 6A of Ge et al. [2] for SNP rs80054410 they switched from the convention T>C to A>G). The former explanation for genes that are very specifically expressed in certain cell types and even then at low or modest levels could be a problem for many genes and published results.

This question applies to the other *UBASH3A* SNP: the minor A allele of rs11203203 shows the opposite *UBASH3A* gene expression to that reported by Ge et al. [2]: decreased in whole blood in both studies, GTEx and https://molgenis58.target.rug.nl/bloodeqtlbrowser/. And yet in lymphoblastoid cell lines, which express very little *UBASH3A* mRNA, the minor A allele rs11203203, respectively, is associated with increased expression, as reported by Ge et al. [2]. In the more recent large-scale whole blood eQTL analyses https://molgenis58.target.rug.nl/biosqtlbrowser/ [8, 9] the mRNA level associated with the alleles of rs11203203 is again in the opposite direction of Ge et al. [2]. Ge et al. did not test *UBASH3A* eQTL expression for one of the other top T1D-associated SNPs rs11203202 but in the databases the minor T1D risk allele G is associated with decreased *UBASH3A* expression, not increased.

Table 1 Haplotype mapping of the UBASH3A region suggests that multiple causal variants are present

Haplotype	Overall disease haplotype association	rs11203202 disease association	rs11203203 disease association	rs1893592 disease association
H2	S	S	S	S
H3	P	P	P	P
H9	S	S	S	P
H1	N	P	P	S

S susceptible, *P* protective, *N* neutral

Note that the rs80054410 SNP does not provide additional information for this haplotype comparison

Is it possible that rs1893592 is causal? It is not in much LD with any other SNP, $r^2 = 0.5$ being the highest value. In RegulomeDB there is no data for rs1893592 i.e. it does not coincide with known transcription factor (TF) binding site or epigenetic mark that could regulate gene expression. SNPs rs11203202 and rs11203203 have scores of 3a (less likely to affect TF binding) and 5 (minimal binding evidence), respectively, whereas rs80054410, which lies with 176 bases from rs11203203 in *UBASH3A* intron 4, has stronger evidence of function with a score of 2b (likely to affect binding), suggesting it could be causal, affecting TF binding and *UBASH3A* transcription. SNP rs9981624 is 356 bases from rs11203202 and they are in complete LD ($r^2 = 1.0$) with a RegulomeDB score of 3a. Of course these SNPs could be tagging a structural variant or indel not being typed directly by the existing arrays.

The story, however, for this chromosome region may not end at *UBASH3A*. There is some evidence that the the 5' region

of *UBASH3A* containing the SNPs rs11203202 and rs9981624 (365 bases apart and $r^2 = 0.39$ but both equally associated with T1D [4] (https://www.chicp.org/)) makes physical contact with the nearby SLC37A1 gene in untreated and unactivated CD4$^+$ T cells (https://www.chicp.org/). *SLC37A1* encodes a key enzyme in metabolism: it translocates glucose-6-phosphate from the cytoplasm into the lumen of the endoplasmic reticulum for hydrolysis into glucose. If this contact *UBASH3A-SLC37A1* is true and functional then you would predict that these SNPs would also show up as eQTLs for *SLC37A1*, and this may be the case. In both databases https://molgenis58.target.rug.nl/bloodeqltbrowser/ and GTEx (https://www.gtexportal.org/home/) the minor, T1D-risk minor A allele of rs11203203 and the minor C allele of rs9981624 show some evidence of association with decreased expression of *SLC37A1*. Another adjacent gene, *TMPRSS3* (encoding a serine protease), also shows evidence of eQTLs across databases and tissues and whole blood: in https://molgenis58.target.rug.nl/bloodeqltbrowser/ the A allele of rs11203203 (T1D predisposing) is associated with increased *TMPRSS3* mRNA ($P = 8.18 \times 10^{-14}$) and the G allele of rs11203202 (T1D predisposing) with increased mRNA ($P = 1.55 \times 10^{-13}$). In GTEx there is eQTL evidence for three tissues but not in whole blood, and the same direction of effect is present in https://molgenis58.target.rug.nl/biosqltbrowser/ [8, 9].

Both *SLC37A1* and *TMPRSS3* are expressed in tissues and blood cells much more widely and at higher mRNA levels than *UBASH3A*, suggesting that the potential problem of inadequate starting material giving rise to artefactual eQTL results might not apply for *SLC37A1* and *TMPRSS3*. We cannot conclude, however, that these *SLC37A1* or *TMPRSS3* eQTLs are real and even if further work shows that they are, this does necessarily mean that they are causal effects in T1D, which would require further experimentation. Nevertheless, we know from other common disease susceptibility regions that multiple genes as well as multiple causal variants are involved in disease aetiology, e.g., *CTLA4-CD28* [10], for example, when genes in a region are co-regulated by the same transcription factors.

Finally, analysis of the associations of these SNPs with other traits and diseases in the available GWAS analyses of the UK Biobank reveal interesting insights (https://biobankengine.stanford.edu/search). The expected decreased risk associated with the rs1893592 minor C allele for a number of autoimmune diseases is evident in the UK Biobank, but the most significant association is with increased lymphocyte percentage, albeit for a very small effect size in this huge sample size of over 300,000 participants (beta = 0.01; $P = 4.96 \times 10^{-7}$).

In conclusion and notwithstanding the questions around the eQTL findings, *UBASH3A* could well be a causal gene for T1D and other autoimmune diseases. One of the mechanisms-of-action may be reduced negative regulation of IL-2 production. Concannon and colleagues have reported previously that UBASH3A may inhibit the NF-κB signalling via the IκB kinase complex in a ubiquitin-dependent mechanism [2]. There may be novel opportunities to inhibit this pathway to promote T1D-protective IL-2 production. From a burden-of-proof perspective it would be good to repeat the allele-specific mRNA measurements, best done in heterozygous individuals with known haplotypes allowing for multiple effects, and also in larger sample sizes attempt to replicate the control of IL-2 production by *UBASH3A* SNPs.

Compliance with ethical standards

Conflict of interest The author declare that they have no conflict of interest.

References

1. Ge, Y and Concannon, P. Molecular-genetic characterization of common, noncoding UBASH3A variants associated with type 1diabetes. *Eur. J. Hum. Genet.* 2018;26:1060–64.
2. Ge Y, Paisie TK, Newman JRB, McIntyre LM, Concannon P. UBASH3A mediates risk for type 1 diabetes through inhibition of T-cell receptor-induced NF-kappaB signaling. Diabetes. 2017;66:2033–43.
3. Todd JA, Evangelou M, Cutler AJ, et al. Regulatory T cell responses in participants with type 1 diabetes after a single dose of interleukin-2: a non-randomised, open label, adaptive dose-finding trial. PLoS Med. 2016;13:e1002139.
4. Onengut-Gumuscu S, Chen WM, Burren O, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47:381–6.
5. Raj T, Rothamel K, Mostafavi S, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014;344:519–23.
6. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238–43.
7. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.
8. Bonder MJ, Luijk R, Zhernakova DV, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49:131–8.
9. Zhernakova DV, Deelen P, Vermaat M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49:139–45.
10. Kasela S, Kisand K, Tserel L, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+versus CD8+T cells. PLoS Genet. 2017;13:e1006643. journal.pgen.1006643, 1-21