nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Fine-mapping, trans-ancestral, and genomic analyses identify causal variants, cells, genes, and drug targets for type 1 diabetes
Corresponding author name(s): Professor John Todd

Reviewer Comments & Decisions:

Decision Letter, initial version:

28th July 2020

Dear John,

Your Article "Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes" has now been seen by three referees. You will see from their comments below that, while they find your work of interest, they have raised some relevant concerns. We are interested in the possibility of publishing your study in Nature Genetics, but we would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. In this case, we think each referee has provided constructive feedback and highlighted different aspects of the analyses and their interpretation that would benefit from further clarification. Accordingly, we ask you that address all technical comments and concerns through appropriate revisions and additional analyses where warranted. We also ask that you provide more extensive analysis and reporting of the overlap between the current results and previous findings for other diseases and traits as requested by Reviewer #3. We hope you will find this prioritized set of referee points to be useful when revising your study. Please do not hesitate to contact me if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within 8-12 weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your

work.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665

Referee expertise:

Referee #1: Genetics, autoimmune diseases, statistical methods

Referee #2: Genetics, autoimmune diseases, regulatory variation

Referee #3: Genetics, immune-mediated diseases

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:

Robertson, Inshaw et al present an updated immunochip mapping study in T1D. They double the sample size relative to the previous study, identifying 36 new risk loci at genome-wide significance, and another 70 suggestive loci. The study is notable in including non-European ancestry samples, which are often difficult to ascertain in autoimmune disease. Fine-mapping reveals that about a third of the loci contain more than one independent effect, and credible interval variants are enriched in chromatin accessible in immune cell subpopulations, especially in stimulus-specific ATACseq peaks. The authors further show that a handful of T1D risk effects colocalize with chromatin QTLs and eQTLs, again consistent with the relative paucity of such effects mapped in autoimmune diseases. The authors take a closer look at the BACH2 locus, where they generate a mechanistic hypothesis that allele-specific ETS1 binding in T cells that drives the regulation of BACH2, which is required for naive state maintenance amongst other functions.

Overall, this is a solid paper, with post-GWAS analyses done thoughtfully and well to place results in context. A number of questions arose as I read the manuscript, which the authors may wish to consider.

Technical issues:

- Admixture in non-European ancestry samples. These are sourced to a large extent from the US, and are thus likely admixed. I am a little concerned that this may affect both the overall mapping efforts and the fine-mapping. For example, on p10 the authors highlight a variant that is rare in non-African populations but present in an appreciable MAF in samples classified as AFR. Different rates of

African/non-African admixture at this locus would confound this signal and/or inflate the effect size (the reported OR is 2.9 - high for common-ish variants). Can the authors demonstrate that admixture is equivalent between cases and controls here, or control for those differences?

The authors also fine-map associations with evidence of a single causal variant across ethnic groups using PAINTOR. That algorithm is quite sensitive to reference LD structure, and admixture will contribute noise to this. It is unclear from the methods how the authors mitigated this source of error.

- GUESSFM results in UBASH3A. The decrease in AIC of the three-variant model is relatively modest, and could be explained by the addition of more factors to the model. A known limitation in model comparison is that as one adds predictors to a model, one tends to see better fit, even if those predictors are random variables. AIC includes a term for complexity, but modeling noise may spuriously improve goodness of fit beyond that penalty. Some form of cross-validation of the model would go a long way to allaying that concern, especially in a cherry-picked example.

Discussion issues:

- Discussion of effect sizes (p10). The authors do not mention winner's curse, where robustly detected effects are likely to be over-estimated due to sampling noise (and hence detected robustly). See also admixture issues above.

- IL6ST association. The authors also list an ANKD55 eQTL as colocalizing in Table 2, with much stronger evidence than IL6ST (z = 58 vs z = 10, respectively). This is also seen In PMID:28218759, where both IL6ST and ANKD55 are prioritized as potential eQTLs in naive CD4+ T cells for Crohn disease, MS, and RA (all immunochip data), with ANKD55 having substantially stronger colocalization evidence. I understand the biological plausibility of the IL6ST story, but the data may support a more ambiguous reading.

- I am a little surprised not to see discussion on the relative paucity of T1D association overlaps with caQTL/eQTL (p16, credible interval SNPs overlap a caQTL in 11/52 loci, and only 5 of these have evidence of a shared effect). This seems to be an emerging issue in the field: bulk analyses demonstrate strong enrichment of heritability in accessible chromatin in immune subpopulations (especially CD4+ T cells), but colocalization identifies relatively few specific examples of shared associations between disease and molecular traits.

- minor: in the discussion the authors refer to "irritable bowel disease (IBD)" - this should be inflammatory bowel disease

- minor: the word causal is sometimes misspelled as casual (e.g. line 170 in Online Methods, bottom of p11 in main text). A search and replace is advised.

Reviewer #2:
Remarks to the Author:

Robertson and colleagues present the largest T1D genetic study to date, including over 61,000 participants across multiple ancestry groups. They identified a total of 152 T1D-associated regions,

including 36 that had not been previously significantly associated. Integration of these variants with chromatin accessibility data revealed that many variants fall within accessible chromatin in CD4+ T cells. Additionally, overlap of variants with chromatin accessibility and expression quantitative trait loci (caQTL and eQTL) yielded five variants associated with altered chromatin landscape, four of which were linked to known target genes. Further characterization of one of these variants at BACH2 revealed that the variant site interacts with the BACH2 promoter in naïve CD4+ T cells, is associated with decreased chromatin accessibility and BACH2 expression in naive CD4+ T cells, and disrupts ETS1 binding, indicating a potential mechanism for a candidate causal variant.

Overall, this paper is a well-executed study with an impressive sample size that identifies 36 novel regions associated with T1D, greatly expanding data on T1D risk and highlighting novel regions for mechanistic studies. Reassuringly, the authors identified most previously characterized T1D-associated regions in addition to novel regions.

The authors should address the following specific concerns:

1. On page 5, the authors state that "Approximately 50 chromosome regions are known to contain variants that alter T1D risk," but they mention at least 116 regions of the 152 identified were not novel associations. Please clarify the difference between "chromosome regions" and "regions." This ambiguous terminology is confusing in several places in the manuscript.

2. On page 15, the authors assess enrichment of credible variants in accessible chromatin of multiple cell types. According to the "ATAC-seq Enrichment Analysis" in the methods (beginning line 264), the authors conducted this analysis comparing T1D credible variants to variants from regions with similar LD structure and gene density. An additional important comparison is between T1D credible variants and locus-matched non-T1D-associated variants. The authors should include this analysis.

3. On page 16, the authors state that "Within-peak credible variants with consistent caQTL effects and allele-specific accessibility provide high priority candidate variants for functional follow-up, as the observed allele-specific accessibility is unlikely to be caused by LD with variants outside the peak of interest." Although variants within an accessibility peak may be causal candidates, it is possible that variants outside of the peak alter accessibility as well, and they should not be excluded as possible credible variants. On this note, are reported caQTLs also associated with local chromatin accessibility changes beyond the immediate accessibility peak? Furthermore, the authors should acknowledge that there could be important context-specific accessibility peaks that are not identified in available datasets, so variants not in the current set of peaks could still be causal.

4. In Table 2, rs71624119 colocalizes to an eQTL for both ANKRD55 and IL6ST. Is there available chromatin evidence that this region loops to both gene promoters?

5. On page 20, the authors conclude that "the rs72928038 minor A allele disrupts ETS1 binding, which leads to decreased enhancer activity and BACH2 expression in naïve CD4+ T cells." Although the authors show that the variant is associated with decreased accessibility, decreased BACH2 expression, and decreased ETS1 binding, they have not performed genetic perturbation experiments to confirm that disrupted binding is the cause of decreased enhancer activity and BACH2 expression. Additionally, the authors seem to assume that rs72928038 is causal at this locus, without convincingly ruling out any expression effects caused by the presence of rs6908626. Instead of the current wording, please specify that disrupted ETS1 binding is a candidate mechanism or hypothesis for chromatin accessibility

and BACH2 expression that may confer T1D risk.

6. On page 22, the authors identify potential gene targets for therapeutic intervention. Please comment on the directionality of gene expression changes associated with variants affecting these gene targets. Are lower levels of expression and/or loss-of-function mutations associated with disease risk or disease protection? This more detailed information is important in considering which gene products may be prioritized as candidate targets to prevent or treat disease.

7. On page 23, the authors state that "rs72928038 is associated with changes in expression of 39 distal genes in whole blood." Are any complementary data available to assess if these genes are regulated by BACH2 (directly or indirectly)?

Minor concerns:

1. In Supplementary Figure 1, definitions for abbreviations are missing.

2. Supplementary Figure 2, 3, and 5 are not cited in the text.

3. In Figure 1a, the y-axis label, color legend, and numbers for variants in each group should be rotated to improve legibility.

4. In Figure 1c, exclude the data interpretation from the figure legend.

5. In Figure 1d, the meaning of black and white fill for each haplotype is not explained. Additionally, the wording in the figure legend "the effect estimate from the haplotype which contains the minor allele at rs11203203 (grey) only relative to the haplotype with major alleles at each variant is close to 0 (comparison of columns 1 and 3), whereas the non-zero effects of each of rs729984852 (blue, columns 2 and 6), rs13048049 (red, column 4) and rs7276555 (green, comparing columns 3 and 5) can be observed" is confusing and should be clarified.

6. On page 16, the authors conclude that "results indicate that T1D credible variants may contribute to islet autoimmunity, in part, by altering responses to T cell receptor signaling." This is concluded from data of T cells stimulated with anti-CD3/CD28 and IL-2 and therefore, a more accurate statement would be that variants may alter responses to T cell receptor signaling, co-stimulation and/or cytokine signaling.

7. In Figure 2a, 2b, and 3a, the gene map is not clear. A scale bar and decreasing the size of the arrows indicating transcriptional direction would make it more legible and clarify the intron/exon structure.

8. In Figure 3b, all colors used in the figure are not explained.

9. In Figure 3d, there is no measurement of peak accessibility on the y-axis, so it is not possible to assess the effect size for each genotype.

10. In Figure 3f, the scale bar is missing. Additionally, are all rows shown on the same vertical scale?

Reviewer #3:
Remarks to the Author:

The study of Robertson et al reports the large Immunochip study of type 1 diabetes (T1D). The study involves >60,000 participants (of them, ~16,000 T1D cases and ~7000 samples of non-European origin). They identified 152 FDR-significant loci (78 genome-wide significant), of which 36 were not reported earlier in relation to T1D. The authors further look at the overlap of associated loci with expression-QTLs and chromatin-accessibility QTLs. Based on the colocalization analysis with published functional genomic datasets and protein-protein interactions, they prioritized the potential drug targets.

This is a long-awaited Immunochip study. The other similar studies in most autoimmune diseases were published in the period of 2011-2013. For example, Jostins et al, Nat. Genetics, 2012 (Crohn's disease and ulcerative colitis), Trynka et al, Nat. Genetics, 2011 (Celiac disease), Cortes et al, Nat. Genetics, 2013 (Ankylosing spondylitis), Beecham et al, Nat. Genetics, 2013 (Multiple Sclerosis), Tsoi et al, Nat. Genetics, 2013 (Psoriasis), Eyre et al, Nat. Genetics, 2012 (Rheumatoid arthritis), and others. The T1D results remained unpublished so far, which was unfortunate, since these results were not included in the follow up analyses/reviews of the shared genetics of autoimmunity, which were published in past years. Good to see that this study is finally completed.

The statistical analysis of the associated loci was not straightforward. First, the authors report the FDR significant loci rather than genome-wide loci. It is not wrong per se, though I think in the abstract should clearly state the number of genome-wide significant loci. Secondly, the authors perform the selection of credible SNPs using GUESSFM method, which report multiple credible variants per locus that are likely to contribute to disease. The downstream analysis is done by using the set of 2431 T1D "credible variants".

Since all major Immunochip papers are published, this study would benefit from the more extensive analysis of the overlap between these results and other traits. The authors state in the discussion: "13 of the 36 novel regions associated with T1D risk are associated with other immune-related traits" This statement is clearly wrong; many other loci from table 1 are associated to various autoimmune/ immune-mediated diseases, though the LD between the T1D and other disease lead SNP is less than what was used by the authors in their table 1. I suggest that the author would add information on the top SNPs associated with other common diseases and LD with the top T1D SNP for each reported locus. This will greatly improve the value of this paper for the immunogenetic society.

The authors focused on the functional analysis of BACH2 locus in relation to enhancer accessibility and gene expression. It would be informative to present the current results of BACH2 analysis in the perspective of the previous studies of BACH2. The role of BACH2 in T1D was reported by the same group in 2008 (https://pubmed.ncbi.nlm.nih.gov/18978792/), since then many papers have been published in the field. It is not clear for me, how novel the results on BACH2 reported in this paper are. I should say, I am not an expert in these studies, but for me the role of BACH2 in relation to T1D is as classical as NOD2 in Crohn's disease. It would be helpful to be more clear in the discussion – what was known prior to this study on the role of BACH2

in T1D, and what were the novel findings added by this study.

Why is HLA excluded? Information on HLA fine-mapping and conditional analysis will be very valuable to present.

Links to supplementary tables are mixed up, for example, ST7 appears after ST8 and 9, I did not see the link to ST10 before the link to ST11.

ST11 is not clear to me: for example, how to read allele 1 and allele2 information for the lines with more than one SNP in the index column (i.e "rs549803996;rs1034010577")?

The results of the ATACseq enrichment analysis does not allow to make any conclusion: it is expected that enrichment would be observed for immune cells, since immunochip represents a selection of immune-related loci. Immunochip does not provide an opportunity for the genome-wide analysis, therefore enrichment results are biased.

I tried to follow the haplotype analysis using the github link (https://github.com/ccrobertson/t1d-immunochip-2020), but it is unclear which information is presented and where. The pictures are not explained and formatting do not allow them to be seen on a screen, for example: https://github.com/ccrobertson/t1d-immunochip-2020/blob/master/supplemental_data/guessfm_fine_mapping/SH2B3/chr12.111569952.C.T.png

https://github.com/ccrobertson/t1d-immunochip-2020/blob/master/supplemental_data/guessfm_fine_mapping/BACH2/chr6.90267049.G.A.png

Sometimes the writing flow is not logical, for example, in the middle of the first part of results, when the authors report the methodology of combining case-control and trio results, they suddenly relate to comparing the previous results in IL6R with the current findings. It would be more logical to create a separate paragraph comparing the current results with previous findings in those loci that have been established prior to this study.

Minor comments:

Why did the authors put separately the Finnish cohort? Supplementary figure 1 plots suggest they could be combined to the European cluster.

Supplementary table 4: would be informative to add the ancestry for each cohort.

"Of 52 ImmunoChip regions associated with T1D" – the authors mean 152?

Author Rebuttal to Initial comments

Response to Referees

Reviewer #1:
Remarks to the Author:

Robertson, Inshaw et al present an updated immunochip mapping study in T1D. They double the sample size relative to the previous study, identifying 36 new risk loci at genome-wide significance, and another 70 suggestive loci. The study is notable in including non-European ancestry samples, which are often difficult to ascertain in autoimmune disease. Fine-mapping reveals that about a third of the loci contain more than one independent effect, and credible interval variants are enriched in chromatin accessible in immune cell subpopulations, especially in stimulus-specific ATACseq peaks. The authors further show that a handful of T1D risk effects colocalize with chromatin QTLs and eQTLs, again consistent with the relative paucity of such effects mapped in autoimmune diseases. The authors take a closer look at the BACH2 locus, where they generate a mechanistic hypothesis that allele-specific ETS1 binding in T cells that drives the regulation of BACH2, which is required for naïve state maintenance amongst other functions.

Overall, this is a solid paper, with post-GWAS analyses done thoughtfully and well to place results in context. A number of questions arose as I read the manuscript, which the authors may wish to consider.

Technical issues:

- Admixture in non-European ancestry samples. These are sourced to a large extent from the US, and are thus likely admixed. I am a little concerned that this may affect both the overall mapping efforts and the fine-mapping. For example, on p10 the authors highlight a variant that is rare in non-African populations but present in an appreciable MAF in samples classified as AFR. Different rates of African/non-African admixture at this locus would confound this signal and/or inflate the effect size (the reported OR is 2.9 - high for common-ish variants). Can the authors demonstrate that admixture is equivalent between cases and controls here, or control for those differences?

Thank you for this thoughtful comment. Yes, our samples classified as “African Admixed (AFR)” ancestry represent African-Americans and are admixed, with varying proportions of European ancestry across participants. As described, we control for population stratification by first partitioning the total sample into major ancestry groups (Supplementary Figure 2), followed by adjustment for “within-ancestry” genetic principal components (PCs) in the association analyses. In addition, we visualized within-ancestry genetic PCs by case-control status to ensure appropriate matching by ancestry. Within these groups, the cases and controls overlap and are similarly distributed across PC1 and PC2 space (as shown in Supplementary Figure 3, AFR panel).

As you note, for the rare variant, rs190514104 (G>A) near *MTF1*, one possibility for the observed result could be confounding by heterogeneity in admixture across AFR subjects. As the prevalence of T1D is higher in European ancestry populations, if admixture were confounding

this association, one might anticipate that the African-enriched/specific allele would appear to be protective against T1D, which is not the case for rs190514104. However, given that our AFR sample sizes are still limited, we agree that this association requires replication in an independent set of samples (none available to our knowledge). We have now revised the text to emphasize the necessity of replication (p. 11, lines 209-213):

> "Considering the limited sample size, potential heterogeneity of the AFR cohort, and possible over-estimation of effect sizes due to "the winner's curse", this association requires replication in an independent cohort. Nonetheless, this example underscores the potential for emphasis on under-studied populations to enhance biological insight, even with limited sample sizes."

The authors also fine-map associations with evidence of a single causal variant across ethnic groups using PAINTOR. That algorithm is quite sensitive to reference LD structure, and admixture will contribute noise to this. It is unclear from the methods how the authors mitigated this source of error.

Thank you for this comment. We agree that external LD reference panels are often inadequate for admixed populations. In order to address this concern, we generated LD estimates from our own data sets separately for each ancestry group. We have updated the Online Methods to make this clear (Online Methods p8, line 175-177):

> "Input LD reference panels, as used by PAINTOR, were generated with the software LDstore version 1.1 (http://www.christianbenner.com/#)[17] separately for each ancestry group using imputed genotype data from unrelated cases and controls from this study."

- GUESSFM results in UBASH3A. The decrease in AIC of the three-variant model is relatively modest, and could be explained by the addition of more factors to the model. A known limitation in model comparison is that as one adds predictors to a model, one tends to see better fit, even if those predictors are random variables. AIC includes a term for complexity, but modeling noise may spuriously improve goodness of fit beyond that penalty. Some form of cross-validation of the model would go a long way to allaying that concern, especially in a cherry-picked example.

Thank you for noting that model comparison using AIC is an efficient approach to prioritize models, while recognizing that sampling noise may limit the strength of conclusions. We have approached this concern by subsampling without replacement 80% of our analysis cohort five times and repeating the GUESSFM procedure for the *USBASH3A* region in each of the five subsampled data sets. We provide the results in our Supplemental Data repository (https://github.com/ccrobertson/t1d-immunochip-2020/tree/master/supplemental_data/guessfm_UBASH3A_subsampling). In brief, neither rs11203203 nor any other single-SNP model was selected by GUESSFM in any of the five subsampled analyses, while at least one SNP included in the three-SNP model presented in Figure 1c (rs7276555 + rs13048049 + rs9984852) was always chosen by GUESSFM.

In addition, for all five subsampled data sets, the 3-SNP model (rs7276555 + rs13048049 + rs9984852) had lower AIC compared to the "top model" prioritised by GUESSFM in the same subsampled dataset. This suggests GUESSFM struggled to prioritise the three-SNP model in the

subsampled data (despite it being the best fit), likely due to the penalty for model complexity (GUESSFM performs model selection using the Bayesian Information Criterion).

Finally, in each of the five subsampled data sets, we compared AIC for the six models shown in Figure 1c. Again, the three-SNP model outperformed the single-SNP models (including only rs11203203) in all subsampled data sets.

These subsampling analyses support the inference that multiple variants in this locus may contribute to disease risk. Furthermore, the lead candidate for the true underlying model in this region remains the three-SNP model presented in Figure 1c (rs7276555 + rs13048049 + rs9984852).

We have revised the manuscript to provide these additional supporting analyses and to emphasize the primary message, which is the complexity of association in the region (p12, line 233-248).

> "For example, although stepwise regression analysis in the *UBASH3A* locus supported a single causal variant (**Supplementary Table 7**), GUESSFM fine-mapping and haplotype analyses indicated that the lead variant in this region (rs11203203 (G>A)) is unlikely to be causal. GUESSFM fine-mapping supported a three-variant model (rs9984852 (T>C), rs13048049 (G>A) and rs7276555 (T>C)) (**Figure 1b**), which had a better fit than the single variant model (AIC 45073 vs. 45138, **Figure 1c**). Haplotype analysis (Online Methods) demonstrated that when rs11203203 (A) is present without the GUESSFM-prioritised variants, there is no effect of rs11203203 on disease risk (**Figure 1d**). Resampling experiments consistently supported two or more causal variants in the region, with at least one of the three GUESSFM-prioritised variants more likely to be causal than rs11203203 (**Supplementary Table 10**). Given the complexity of association in the *UBASH3A* region, and likely at many loci, statistical methods designed to use univariable summary statistics alone are not sufficient to explore the genetic architecture of T1D."

Discussion issues:

- Discussion of effect sizes (p10). The authors do not mention winner's curse, where robustly detected effects are likely to be over-estimated due to sampling noise (and hence detected robustly). See also admixture issues above.

We agree that winner's curse is likely leading to over-estimation of effect sizes at novel loci and note that additional studies are needed for replication and more accurate estimation of effect sizes in these regions. We have added text to acknowledge the effect of winner's curse on our effect size estimates and emphasize the importance of future studies to refine our estimation of variant contribution to disease burden, especially in non-European groups.

> p11, line 209-213 (same as above response for Reviewer 1):
>
> "Considering the limited sample size, potential heterogeneity of the AFR cohort, and possible over-estimation of effect sizes due to "the winner's curse", this association requires replication in an independent cohort. Nonetheless, this example underscores the potential for

emphasis on under-studied populations to enhance biological insight, even with limited sample sizes."

p31, line 528-534:

"Additionally, the effect sizes of novel loci are likely over-estimated due to winner's curse, particularly those identified in non-European groups where sample sizes remain modest, such as rs190514104 (G>A) near *MTF1*. While this analysis is the largest and most comprehensive study prioritising novel therapeutic interventions in T1D according to genetic evidence, extension of future genetic studies to genome-wide analyses and continuing efforts to expand cohorts from diverse populations will further define the genetic landscape of T1D."

- IL6ST association. The authors also list an ANKD55 eQTL as colocalizing in Table 2, with much stronger evidence than IL6ST (z = 58 vs z = 10, respectively). This is also seen in PMID:28218759, where both IL6ST and ANKD55 are prioritized as potential eQTLs in naive CD4+ T cells for Crohn disease, MS, and RA (all immunochip data), with ANKD55 having substantially stronger colocalization evidence. I understand the biological plausibility of the IL6ST story, but the data may support a more ambiguous reading.

IL6ST is an appealing target since mechanisms of IL6 signaling in human biology have been thoroughly studied; as a result, drugs already exist to target this pathway. Additionally, our Priority Index analyses, which incorporates all credible variants and protein-protein interaction networks, prioritized *IL6ST* but not *ANKRD55*; however, our results do not exclude alternative targets or mechanisms. We have modified the text to acknowledge this ambiguity (p30, line 504-511):

"Both the *IL6ST* and *IL6R* regions were identified here as T1D-associated at genome-wide significance for the first time (Table 1), and both *IL6ST* and *IL6R* were prioritised by our Priority Index analysis. *IL6ST* is implicated by QTL colocalisation and the lead T1D variant near *IL6R* (rs2229238) is an eQTL for *IL6R* expression in whole blood (formal colocalisation was not assessed as the *IL6R* region is not densely covered on the ImmunoChip). We cannot say, based on current evidence, that *IL6ST* and *IL6R* are T1D causal genes. The associations in each region may be unrelated and due to different causal genes; for example, the association near *IL6ST* also colocalizes with an eQTL for *ANKRD55*."

We have also added additional discussion of the fine-mapping and regulatory effects of T1D-associated variants in the *IL6ST/ANKRD55* region (p28, line 463-472):

"In the 5q11.2 region, fine-mapping and caQTL colocalization point to the within-peak variant, rs7731626, as a potential causal variant for T1D, complementary to a recent regulatory QTL fine-mapping study that highlighted the same variant as likely functional in T cells[48]. Additionally, the T1D association signal colocalises with eQTLs for both *ANKRD55* and *IL6ST*, mirroring multiple sclerosis, Crohn's disease, and rheumatoid arthritis[46]. The region overlapping rs7731626 loops to the *IL6ST* promoter in $CD4^+$ T cells according to promoter capture Hi-C data[42]. This variant is also intronic within *ANKRD55*. Although we did

not find evidence that rs7731626 loops to the canonical transcription start site for *ANKRD55*, publicly available nascent RNA-sequencing data suggest it overlaps the 5' end of the transcriptionally active region of *ANKRD55* in human T cells[49], suggesting a potential regulatory role. "

- I am a little surprised not to see discussion on the relative paucity of T1D association overlaps with caQTL/eQTL (p16, credible interval SNPs overlap a caQTL in 11/52 loci, and only 5 of these have evidence of a shared effect). This seems to be an emerging issue in the field: bulk analyses demonstrate strong enrichment of heritability in accessible chromatin in immune subpopulations (especially CD4+ T cells), but colocalization identifies relatively few specific examples of shared associations between disease and molecular traits.

Thank you for the insightful comment. In addition to the limitations of bulk analyses, that ignore cell type-specific, time-sensitive, or stimulus-dependent effects, recent work has shown that colocalisation methods may have high false negative rates due, in part, to low statistical power (Hukku, A. *et. al*. *bioRxiv* (2020) doi:10.1101/2020.07.01.182097). We were stringent in defining colocalisation, as we intended to prioritise putative causal variants, genes, and mechanisms with strong and reproducible evidence of shared regulatory effects, as opposed to a longer list of candidates with likely high false positive rates. These sites could represent targets for experimental follow-up. We have added text to the discussion to acknowledge this issue (p27, line 449-455):

"Despite enrichment of credible variants in $CD4^+$ T cell open chromatin, only five of 52 fine-mapped T1D associations could be explained by a colocalised caQTL. This is consistent with previous work exploring functional effects of variants associated with immune traits [46]. Possible explanations for this apparent discrepancy include limited power in QTL discovery due to small sample sizes or imprecise cell types[46,47]. Nevertheless, while this approach may lack sensitivity, the five regions showing colocalisation between caQTL and T1D associations prioritize variants with regulatory effects that could be realistic targets for experimental follow-up."

- minor: in the discussion the authors refer to "irritable bowel disease (IBD)" - this should be inflammatory bowel disease

Thank you. This has been corrected.

- minor: the word causal is sometimes misspelled as casual (e.g. line 170 in Online Methods, bottom of p11 in main text). A search and replace is advised.

Thank you. This has been corrected.

Reviewer #2:
Remarks to the Author:

Robertson and colleagues present the largest T1D genetic study to date, including over 61,000 participants across multiple ancestry groups. They identified a total of 152 T1D-associated regions, including 36 that had not been previously significantly associated. Integration of these variants with chromatin accessibility data revealed that many variants fall within accessible chromatin in CD4+ T cells. Additionally, overlap of variants with chromatin accessibility and expression quantitative trait loci (caQTL and eQTL) yielded five variants associated with altered chromatin landscape, four of which were linked to known target genes. Further characterization of one of these variants at BACH2 revealed that the variant site interacts with the BACH2 promoter in naïve CD4+ T cells, is associated with decreased chromatin accessibility and BACH2 expression in naive CD4+ T cells, and disrupts ETS1 binding, indicating a potential mechanism for a candidate causal variant.

Overall, this paper is a well-executed study with an impressive sample size that identifies 36 novel regions associated with T1D, greatly expanding data on T1D risk and highlighting novel regions for mechanistic studies. Reassuringly, the authors identified most previously characterized T1D-associated regions in addition to novel regions.

The authors should address the following specific concerns:

1. On page 5, the authors state that "Approximately 50 chromosome regions are known to contain variants that alter T1D risk," but they mention at least 116 regions of the 152 identified were not novel associations. Please clarify the difference between "chromosome regions" and "regions." This ambiguous terminology is confusing in several places in the manuscript.

Thank you for pointing out this ambiguity. We defined a "novel" T1D-associated region as one meeting genome-wide significance ($p < 5x10^{-8}$) where the lead variant was not within 250kb of a previously identified genome-wide significant variant in well-powered T1D studies (PMIDs: 17554260, 17554300, 17632545, 18198356, 18840781, 18978792, 19430480, 21980299, 22293688, 25751624, 31152121). We identified 152 regions associated with T1D at FDR < 0.01 (Supplementary Table 8 and Supplementary Table 9); however, only 78 of these regions reached genome-wide significance ($p < 5x10^{-8}$) and are, therefore, potentially novel T1D regions. Of these 78 regions, 36 had not been associated previously with T1D at $p < 5x10^{-8}$; in the remaining 42 regions, the most associated variant was within 250kb of the lead variant in a previous T1D study.

We have added text to clarify our definition of a novel region (p8, line 170-173):

> "When comparing these 78 regions to previous T1D studies[1-3,8-15], 36 novel regions associated with T1D at genome-wide significance for the first time (**Table 1**), while, in the remaining 42 regions, the lead variant was within 250kb of the lead variant in a previous T1D study."

We have also now included a separate section describing the associated regions based on FDR<0.01 entitled "**Additional regions identified using alternative inheritance models and metric of statistical significance**" (p10, line 194-219). We hope this will help the reader to better follow the flow of analyses.

2. On page 15, the authors assess enrichment of credible variants in accessible chromatin of multiple cell types. According to the "ATAC-seq Enrichment Analysis" in the methods (beginning line 264), the authors conducted this analysis comparing T1D credible variants to variants from regions with similar LD structure and gene density. An additional important comparison is between T1D credible variants and locus-matched non-T1D-associated variants. The authors should include this analysis.

Thank you for this suggestion. To address this concern, we ran additional enrichment analysis using GoShifter software (Trynka, G. *et. al. Am. J. Hum. Genet.* (2015)). As you describe, our first approach (now referred to in the manuscript as the 'SNP-matching' approach) uses variants from across the genome matched based on LD structure and gene density to generate a null distribution of SNPs overlapping accessible chromatin. In contrast, GoShifter generates a null distribution within each locus. Specifically, GoShifter "locally shifts sites of tested features within each locus to generate a null distribution of annotations overlapping associated variants by chance." We now include the results from this analysis in Supplementary Figure 10b, and a comparison of the two approaches in Supplementary Figure 10c.

We have updated the main text to reflect this change (p18, line 307-309):

> "T1D credible variants were enriched in open chromatin in multiple primary immune cell types based on two complementary enrichment analysis approaches (Online Methods)."

We have also updated the Online Methods (p12, line 266-273):

> "To examine enrichment of T1D credible variants (group marginal posterior probability > 0.8 from GUESSFM) in open chromatin, for each cell type, we applied two complementary approaches: SNP-matching using a custom script and GoShifter[33] (http://software.broadinstitute.org/mpg/goshifter/). In the SNP-matching approach, we randomly sampled variants from across the genome matched based on LD structure and gene density to generate a null distribution of SNPs overlapping accessible chromatin (detailed algorithm is described below). In contrast, GoShifter generates a null distribution within each locus. For details on the GoShifter algorithm, please refer to Trynka et al, 2015[33]."

3. On page 16, the authors state that "Within-peak credible variants with consistent caQTL effects and allele-specific accessibility provide high priority candidate variants for functional follow-up, as the observed allele-specific accessibility is unlikely to be caused by LD with variants outside the peak of interest." Although variants within an accessibility peak may be causal candidates, it is possible that variants outside of the peak alter accessibility as well, and they should not be excluded as possible credible variants.

Thank you for this suggestion. While we view allele-specific accessibility at a variant as suggestive evidence of the variant being biologically relevant (with the most straightforward mechanism being the variant influences accessibility in the region within which it resides), we

acknowledge there are other possible mechanisms underlying allele-specific accessibility, including functional variants outside the peak. We have removed the quoted statement to avoid making this claim in absence of supporting data.

On this note, are reported caQTLs also associated with local chromatin accessibility changes beyond the immediate accessibility peak?

We have seen evidence occasionally of a SNP associated with chromatin accessibility in multiple locations in a region (perhaps supporting the concept of "regulatory hubs" of co-regulated active chromatin); however, this was not the case for any of the five regions highlighted as having colocalizing caQTLs in Table 2.

Furthermore, the authors should acknowledge that there could be important context-specific accessibility peaks that are not identified in available datasets, so variants not in the current set of peaks could still be causal."

Thank you. We agree that these analyses do not conclusively identify (or exclude) variants as causal. We take advantage of prior knowledge that $CD4^+$ T cells are implicated in the T1D disease process (as supported by mouse models, favorable patient response to CD3 inhibition, MHC class II alleles as strongest predictors of T1D risk, and enrichment of T1D-associated variants in active chromatin from $CD4^+$ T cells) to prioritize putative causal variants and generate hypotheses for their effects on gene regulation in this cell type. We have added the following text to acknowledge the possibility of alternative causal variants acting through other cell types and contexts (p20, line 339-342):

"While we cannot exclude the possibility that additional variants are acting through alternative cell types and contexts, regions with support for a common variant underlying association with T1D, chromatin accessibility in $CD4^+$ T cells, and gene expression in whole blood provide candidate causal T1D variants and genes for mechanistic investigation."

4. In Table 2, rs71624119 colocalizes to an eQTL for both ANKRD55 and IL6ST. Is there available chromatin evidence that this region loops to both gene promoters?

Thank you for this question. We were also interested in whether chromatin looping supports the observed QTLs in this region, particularly since colocalizing eQTLs for *ANKRD55* and *IL6ST* were found previously for multiple sclerosis, Crohn's disease, and rheumatoid arthritis (Chun, S. *et al*. *Nat*. *Genet*. 2017;49(4):600-605, PMID: 28218759), and a recent fine-mapping study (Kundu, K. *et al*. *bioRxiv doi:10.1101/2020.01.15.907436*) highlighted the same variant, rs7731626, as likely functional and potentially causal based on independent QTL data sets. We did not address this in the initial version of the manuscript due to space constraints. However, we now include this discussion (p28, line 463-472):

"In the 5q11.2 region, fine-mapping and caQTL colocalization point to the within-peak variant, rs7731626, as a potential causal variant for T1D, complementary to a recent regulatory QTL fine-mapping study that highlighted the same variant as likely functional in T

cells[48]. Additionally, the T1D association signal colocalises with eQTLs for both *ANKRD55* and *IL6ST*, mirroring multiple sclerosis, Crohn's disease, and rheumatoid arthritis[46]. The region overlapping rs7731626 loops to the *IL6ST* promoter in CD4+ T cells according to promoter capture Hi-C data[42]. This variant is also intronic within *ANKRD55*. Although we did not find evidence that rs7731626 loops to the canonical transcription start site for *ANKRD55*, publicly available nascent RNA-sequencing data suggest it overlaps the 5' end of the transcriptionally active region of *ANKRD55* in human T cells[49], suggesting a potential regulatory role. "

5. On page 20, the authors conclude that "the rs72928038 minor A allele disrupts ETS1 binding, which leads to decreased enhancer activity and BACH2 expression in naïve CD4+ T cells." Although the authors show that the variant is associated with decreased accessibility, decreased BACH2 expression, and decreased ETS1 binding, they have not performed genetic perturbation experiments to confirm that disrupted binding is the cause of decreased enhancer activity and BACH2 expression. Additionally, the authors seem to assume that rs72928038 is causal at this locus, without convincingly ruling out any expression effects caused by the presence of rs6908626. Instead of the current wording, please specify that disrupted ETS1 binding is a candidate mechanism or hypothesis for chromatin accessibility and BACH2 expression that may confer T1D risk.

Thank you for requesting greater clarity in language. We do intend to present this as a *candidate* mechanism and have edited the text to make this clear (p23, line 391-395):

"These data prioritise rs72928038 as a functionally relevant variant in T cells and provide preliminary support for a candidate regulatory mechanism underlying the 6q15 region association with T1D. Specifically, we hypothesize that the rs72928038 minor A allele disrupts ETS1 binding, which leads to decreased enhancer activity and *BACH2* expression in naïve CD4+ T cells."

6. On page 22, the authors identify potential gene targets for therapeutic intervention. Please comment on the directionality of gene expression changes associated with variants affecting these gene targets. Are lower levels of expression and/or loss-of-function mutations associated with disease risk or disease protection? This more detailed information is important in considering which gene products may be prioritized as candidate targets to prevent or treat disease.

Thank you for this excellent suggestion. We have extended our discussion of the target prioritization results (p26, line 424-432):

"T1D susceptibility alleles may alter expression of gene targets in either direction, and gene regulatory effects may be seen across multiple major immune cell populations or be restricted to a single cell type (**Supplementary Figure 13**). For example, alleles leading to increased risk of T1D appear to cause increased expression of *MAPK3* and *DGKQ* but decreased expression of *TYK2* across multiple major immune populations. Meanwhile, for *RGS1*, the

T1D risk allele leads to decreased expression across most immune cell types but increased expression specifically in $CD8^+$ T cells. The directionality and cell type-specificity of gene regulatory effects associated with T1D risk alleles may inform therapeutic target considerations."

7. On page 23, the authors state that "rs72928038 is associated with changes in expression of 39 distal genes in whole blood." Are any complementary data available to assess if these genes are regulated by BACH2 (directly or indirectly)?

Thank you for this suggestion. Unfortunately, there are no data available to directly address this question since it is difficult to definitively determine which genes are regulated by a given transcription factor. However, bioinformatic analyses based on chromatin immunoprecipitation sequencing (ChIP-seq) of BACH2 binding in humans and mice suggests that this could be the case. We queried the 39 trans-genes associated with rs72928038 using Binding Analysis for Regulation of Transcription (BART) software (https://zanglab.github.io/bart/index.htm), which uses a catalog of transcription factor ChIP-seq data (obtained through the Cistrome Data Browser http://cistrome.org/) to identify regulators of user-supplied gene sets. The 39 trans-genes associated with rs72928038 are enriched for nearby BACH2 binding sites (Fisher's exact $p = 9x10^{-30}$); however, there are other transcription factors with significant enrichment of binding sites near this set of genes, including FOXP3 ($p = 4x10^{-60}$) and ETS1 ($p = 8x10^{-116}$).

Minor concerns:

1. In Supplementary Figure 1, definitions for abbreviations are missing.

Thank you for catching this. We have made the changes requested to the figure legend (now Supplementary Figure 2).

2. Supplementary Figure 2, 3, and 5 are not cited in the text.

Thank you for catching this. We now reference all Supplementary Figures in the main text.

3. In Figure 1a, the y-axis label, color legend, and numbers for variants in each group should be rotated to improve legibility.

Thank you for this suggestion. We have made the changes requested.

4. In Figure 1c, exclude the data interpretation from the figure legend.

Thank you for catching this. We have made the changes requested.

5. In Figure 1d, the meaning of black and white fill for each haplotype is not explained. Additionally, the wording in the figure legend "the effect estimate from the haplotype which contains the minor allele at rs11203203 (grey) only relative to the haplotype with major alleles at each variant is close to 0 (comparison of columns 1 and 3), whereas the non-zero effects of each of rs729984852 (blue, columns 2 and 6), rs13048049 (red, column 4) and rs7276555 (green, comparing columns 3 and 5) can be observed" is confusing and should be clarified.

Thank you for catching this. We have updated the legend to more thoroughly explain the meaning of the figure (p14, line 258-262):

"d) Analysis of haplotypes associated with T1D in the *UBASH3A* region; the most common haplotype (H1: T-G-G-T for rs7276555- rs13048049-rs11203203-rs9984852) is presented on the far left; alternative haplotypes (H2-H6) are shown with white squares highlighting the differentiating alleles (C, A, A, or C, respectively); the frequency and effect estimates (log odds ratio and 95% confidence interval) for association with T1D relative to the baseline haplotype (H1) are shown above the grid; for example, the log odds ratio for T1D risk for haplotype H3 (T-G-A-T) relative to the baseline haplotype (H1) is close to zero and not statistically significant."

6. On page 16, the authors conclude that "results indicate that T1D credible variants may contribute to islet autoimmunity, in part, by altering responses to T cell receptor signaling." This is concluded from data of T cells stimulated with anti-CD3/CD28 and IL-2 and therefore, a more accurate statement would be that variants may alter responses to T cell receptor signaling, co-stimulation and/or cytokine signaling.

Thank you for catching this. We have made the wording change requested (p19, line 321-322).

7. In Figure 2a, 2b, and 3a, the gene map is not clear. A scale bar and decreasing the size of the arrows indicating transcriptional direction would make it more legible and clarify the intron/exon structure.

Thank you for catching this. We have made the changes requested.

8. In Figure 3b, all colors used in the figure are not explained.

Thank you for catching this. We have made the changes requested.

9. In Figure 3d, there is no measurement of peak accessibility on the y-axis, so it is not possible to assess the effect size for each genotype.

Thank you for catching this. Peak accessibility is quantified based on the number of transposase cut sites in a fixed region (peak boundaries) and then normalized using the "TMM method" and transformed using mean-variance modeling (Online Methods). We now make a note of this in the

figure legend and refer readers to the Online Methods for details.

10. In Figure 3f, the scale bar is missing. Additionally, are all rows shown on the same vertical scale?

Thank you for catching this. We have added a scale bar to this plot. Also, all tracks are on the same vertical scale. We now indicate this in the legend.

Reviewer #3:
Remarks to the Author:

The study of Robertson et al reports the large Immunochip study of type 1 diabetes (T1D). The study involves >60,000 participants (of them, ~16,000 T1D cases and ~7000 samples of non-European origin). They identified 152 FDR-significant loci (78 genome-wide significant), of which 36 were not reported earlier in relation to T1D. The authors further look at the overlap of associated loci with expression-QTLs and chromatin-accessibility QTLs. Based on the colocalization analysis with published functional genomic datasets and protein-protein interactions, they prioritized the potential drug targets.

This is a long-awaited Immunochip study. The other similar studies in most autoimmune diseases were published in the period of 2011-2013. For example, Jostins et al, Nat. Genetics, 2012 (Crohn's disease and ulcerative colitis), Trynka et al, Nat. Genetics, 2011 (Celiac disease), Cortes et al, Nat. Genetics, 2013 (Ankylosing spondylitis), Beecham et al, Nat. Genetics, 2013 (Multiple Sclerosis), Tsoi et al, Nat. Genetics, 2013 (Psoriasis), Eyre et al, Nat. Genetics, 2012 (Rheumatoid arthritis), and others. The T1D results remained unpublished so far, which was unfortunate, since these results were not included in the follow up analyses/reviews of the shared genetics of autoimmunity, which were published in past years. Good to see that this study is finally completed.

The statistical analysis of the associated loci was not straightforward. First, the authors report the FDR significant loci rather than genome-wide loci. It is not wrong per se, though I think in the abstract should clearly state the number of genome-wide significant loci.

> Thank you for this suggestion. We have updated the abstract to include the number of genome-wide significant loci based on a significance threshold of $p < 5x10^{-8}$ (p4, line 71-72).

Secondly, the authors perform the selection of credible SNPs using GUESSFM method, which report multiple credible variants per locus that are likely to contribute to disease. The downstream analysis is done by using the set of 2431 T1D "credible variants".

Since all major Immunochip papers are published, this study would benefit from the more extensive analysis of the overlap between these results and other traits. The authors state in the discussion: "13 of the 36 novel regions associated with T1D risk are associated with other immune-related traits" This statement is clearly wrong; many other loci from table 1 are associated to various autoimmune/ immune-mediated diseases, though the LD between the T1D and other disease lead SNP is less than what was used by the authors in their table 1. I suggest that the author would add information on the top SNPs associated with other common diseases and LD with the top T1D SNP for each reported locus. This will greatly improve the value of this paper for the immunogenetic society.

> We previously examined overlap and concordance of effects between T1D associations and associations with 15 immune-mediated diseases (Onengut-Gumuscu S, *et al*. *Nat Genet* 2015;47:381-386, PMID: 25751624). We found that diseases with well-powered association studies shared many regions with T1D. However, considering the direction of effects provided additional insight about shared biology. For example, juvenile idiopathic arthritis

(JIA) shared fewer regions with T1D than ulcerative colitis, but the direction of the effect for the SNPs were consistent between T1D and JIA, whereas the directions of effect between T1D and UC differed. Thus, our conclusion at that time was diseases with characteristic autoantibodies (T1D, JIA and RA) were more genetically similar to T1D compared to auto-inflammatory disorders (UC and CD). Subsequent work explored evidence for shared causal variants across autoimmune diseases, including T1D (Fortune, MD, *et al*. *Nat Genet* 2015; 47(7):839-46, PMID: 26053495), and demonstrated that, while many ImmunoChip regions are associated with multiple autoimmune diseases, the association patterns suggest that causal variants often differ between diseases. Together, these results emphasized the importance of evaluating both the evidence supporting shared causal variants and consistency in direction of effects when drawing conclusions about shared genetic etiology between immune diseases.

To provide and updated assessment of these findings, we queried the Open Targets Genetics (https://genetics.opentargets.org) database for immune-related diseases with associations in the same regions as each of our T1D lead variants (for 143 regions presented in Supplementary Table 8). We considered all association studies available through the UK Biobank or GWAS Catalog (as of November 2020) for the following 14 diseases: celiac disease, Crohn's disease, Grave's disease, Hashimoto thyroiditis, juvenile idiopathic arthritis, multiple sclerosis, primary biliary cholangitis, primary sclerosing cholangitis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes, ulcerative colitis, and vitiligo. For traits with multiple studies reported, the lead variant in a region was taken to be the variant with the lowest p-value across all studies. We provide the linkage disequilibrium (R-squared) between the T1D lead variants from the current study (as determined through ImmunoChip genotyping followed by TOPMed imputation) and lead variants for previous immune disease association studies (see Attached Figure). However, since most previous immune disease association studies did not include genotype imputation (and none were based on TOPMed imputation), we feel that the conclusions that can be drawn from these results are limited. This issue also limited our ability to determine whether the direction of effects for the updated set of T1D lead variants was consistent with other immune diseases, as association statistics for the majority of lead variants from the current study are not available for the remaining immune diseases. As we feel meaningful conclusions will require expanded analysis of the other immune diseases, we have chosen not to include these results in the manuscript.

Attached Figure Legend: Evidence supporting shared effects between T1D and 14 immune-related diseases. T1D lead variants, and their corresponding candidate genes, are indicated on the x-axis; 14 immune-related diseases under consideration are on the y-axis. A square indicates that the corresponding disease has a genome-wide significant association in the region, with a lead variant in moderate to high linkage disequilibrium (R-squared>0.5) with the lead T1D variant from this study. The R-squared between lead variants is provided within the square. Red squares indicate concordant direction of effect (the T1D lead variant risk allele also increases risk of the immune-related trait). Blue squares indicate discordant direction of effect (the T1D lead variant risk allele is protective against the immune-related trait). Grey squares indicate that summary statistics for association between the T1D lead variant and the immune-related disease are not available, likely due to lack of imputation in previous immune disease association studies. Data used to generate this plot were obtained using the Open Targets Genetics API.

The authors focused on the functional analysis of BACH2 locus in relation to enhancer accessibility and gene expression. It would be informative to present the current results of BACH2 analysis in the perspective of the previous studies of BACH2. The role of BACH2 in T1D was reported by the same group in 2008 (https://pubmed.ncbi.nlm.nih.gov/18978792/), since then many papers have been published in the field. It is not clear for me, how novel the results on BACH2 reported in this paper are. I should say, I am not an expert in these studies, but for me the role of BACH2 in relation to T1D is as classical as NOD2 in Crohn's disease. It would be helpful to be more clear in the discussion – what was known prior to this study on the role of BACH2

> Genetic variation in the 6q15 region has been associated previously with T1D risk, yet there has been limited statistical analyses to suggest that rs72928038 may be implicated in several autoimmune diseases, including T1D (*e.g.*, Marquez A, *et al.*, *Genome Med* 2018; 10:97. PMID: 30572963; Shooshtari P, *et al. Am J Hum Genet* 2017;101:75-86, PMID: 28686857). There was a non-significant association of rs72928038 with progression from single to multiple islet autoantibodies in TrialNet participants (Steck AK, *et al. J Clin Endocrinol Metab* 2017;102:2873-2880, PMID: 28520980). However, to our knowledge, this is the first demonstration that rs72928038 can be functionally prioritized as a potential causal variant with a proposed mechanism via regulation of *BACH2* expression. We are the first to show that rs72928038 is located within a T cell-specific accessible enhancer region, is a colocalizing caQTL, exhibits dramatic allele-specific accessibility, and has loss of ETS1 binding. Thus, although the *BACH2* locus has been implicated in autoimmunity and T1D risk, our report provides unique functional and analytic data supporting the role of rs72928038 in T1D.

Why is HLA excluded? Information on HLA fine-mapping and conditional analysis will be very valuable to present.

> We have published extensively using ImmunoChip data to investigate the HLA region and its association with T1D as well as multiple other autoimmune diseases (Hu X, *et al.*, *Nat Genet* 2015;47:898-905, PMID: 26168013; Lenz TL *et al.*, *Nat Genet* 2015;47:1085-1090, PMID: 26258845, Gutierrez-Arcelus M, *et al.*, *Nat Rev Genet* 2016;17:160-174, PMID: 26907721) as well as utilizing whole genome sequence data in the NHLBI TOPMed program (Gutierrez-Arcelus M, *et al.*, *Nat Genet* 2020;52:247-253). Additional work has focused on HLA associations with T1D in non-European populations (Howson *et al.*, *Diabet Med*. 2013;30(6):710-716, PMID: 23398374; Onengut-Gumuscu *et al.*, *Diabetes Care* 2019;42(3):406-415. PMID: 30659077). The focus of the current manuscript is on T1D region discovery and integrative approaches to fine-mapping to prioritize causal variants and genes, particularly in non-coding regions. Given our previous publications on ImmunoChip-based examination of HLA in T1D and advancement using sequence data, we felt that addressing the HLA region by SNPs (instead of sequencing), the lack of classical HLA typing (relying on different levels of imputation from SNPs in non-European ancestry populations), and lack of impact of HLA conditioning in prior work, we preferred not to include HLA in this manuscript.

Links to supplementary tables are mixed up, for example, ST7 appears after ST8 and 9, I did not see the link to ST10 before the link to ST11.

Thank you for catching this! We have corrected the order and added a reference to Supplementary Table 6 (formerly ST10).

ST11 is not clear to me: for example, how to read allele 1 and allele2 information for the lines with more than one SNP in the index column (i.e "rs549803996;rs1034010577")?

Thank you for pointing out this ambiguity. When two rsIDs are listed, it is because the same SNP was annotated with multiple databases, leading to multiple rsIDs from different dbSNP builds. We now report only dbSNP build 150 rsIDs.

The results of the ATACseq enrichment analysis does not allow to make any conclusion: it is expected that enrichment would be observed for immune cells, since immunochip represents a selection of immune-related loci. Immunochip does not provide an opportunity for the genome-wide analysis, therefore enrichment results are biased.

Thank you for this comment. To address this concern, we performed additional enrichment analyses using GoShifter software. Our first method (now referred to in the manuscript as the 'SNP-matching' approach) used variants from across the genome, with matching based on LD structure and gene density, in order to generate a null distribution of SNPs overlapping accessible chromatin. In contrast, GoShifter generates a null distribution within each locus. GoShifter "locally shifts sites of tested features within each locus to generate a null distribution of annotations overlapping associated variants by chance." We now include the results from these additional analyses in Supplementary Figure 10b, and provide a comparison of the two approaches in Supplementary Figure 10c.

We have updated the main text to reflect this change (p18, line 307-309):

"T1D credible variants were enriched in open chromatin in multiple primary immune cell types based on two complementary enrichment analysis approaches (Online Methods)."

We have also updated the Online Methods (p12, line 266-273):

"To examine enrichment of T1D credible variants (group marginal posterior probability > 0.8 from GUESSFM) in open chromatin, for each cell type, we applied two complementary approaches: SNP-matching using a custom script and GoShifter[33] (http://software.broadinstitute.org/mpg/goshifter/). In the SNP-matching approach, we randomly sampled variants from across the genome matched based on LD structure and gene density to generate a null distribution of SNPs overlapping accessible chromatin (detailed algorithm is described below). In contrast, GoShifter generates a null distribution within each locus. For details on the GoShifter algorithm, please refer to Trynka et al, 2015[33]."

I tried to follow the haplotype analysis using the github link (https://github.com/ccrobertson/t1d-immunochip-2020), but it is unclear which information is presented and where. The pictures are not

explained and formatting do not allow them to be seen on a screen, for example: https://github.com/ccrobertson/t1d-immunochip-2020/blob/master/supplemental_data/guessfm_fine_mapping/SH2B3/chr12.111569952.C.T.png

https://github.com/ccrobertson/t1d-immunochip-2020/blob/master/supplemental_data/guessfm_fine_mapping/BACH2/chr6.90267049.G.A.png

> This repository is provided as a resource for the reader to explore code and additional results from the fine-mapping analyses, without excessive Supplemental Tables and Supplemental Figures. Unfortunately, the GitHub platform is not optimized for image viewing. For access to the image files, we recommend cloning the repository to your local computer, which will allow you to open any image (.png file) with your preferred desktop image viewer. We have now added instructions for doing this on the main page of the repository (https://github.com/ccrobertson/t1d-immunochip-2020).
>
> For a description of the results presented in each of the subdirectories, please see the "Details about subdirectory content" section of this page: https://github.com/ccrobertson/t1d-immunochip-2020/tree/master/supplemental_data.

Sometimes the writing flow is not logical, for example, in the middle of the first part of results, when the authors report the methodology of combining case-control and trio results, they suddenly relate to comparing the previous results in IL6R with the current findings. It would be more logical to create a separate paragraph comparing the current results with previous findings in those loci that have been established prior to this study.

> Thank you for this comment. We have relocated the presentation of this finding to improve the flow for the reader (now p8, line 170-178).

Minor comments:

Why did the authors put separately the Finnish cohort? Supplementary figure 1 plots suggest they could be combined to the European cluster.

> Ancestry stratification groups were determined using k-means clustering of individuals based on the first 10 principal components. In exploratory analyses, the Finnish individuals consistently clustered away from the remaining European individuals, likely due to the unique population history of Finland (Jakkula E *et al*., *Am J Hum Genet* 2008;83:787-794, PMID: 19061986). Although visualization with only the first two principal components showed overlap between Finns and other European groups, clustering based on additional principal components suggested sufficient substructure to warrant additional stratification into Finnish and non-Finnish Europeans.

Supplementary table 4: would be informative to add the ancestry for each cohort.

Please see Supplementary Table 1 for numbers of samples in each cohort by ancestry group.

"Of 52 ImmunoChip regions associated with T1D" – the authors mean 152?

Thank you for your question. The statement "Of 52 ImmunoChip regions associated with T1D" is correct and reflects the number of regions we had sufficient information content to fine-map. Our fine-mapping focused on the densely genotyped regions on the ImmunoChip, termed "ImmunoChip regions". In other areas of the ImmunoChip, sparse genotyping could lead to less accurate fine-mapping. Thus, of the 152 T1D-associated regions defined by FDR criterion, only 52 overlapped with the "ImmunoChip regions".

Decision Letter, first revision:

Our ref: NG-A55094R

23rd December 2020

Dear John,

Your revised manuscript "Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes" (NG-A55094R) has been seen by the original referees. As you will see from their comments below, they are generally satisfied with the responses to their previous comments, and therefore we will be happy in principle to publish your study in Nature Genetics pending final revisions to satisfy the referees' remaining requests and to comply with our editorial and formatting guidelines.

Note that we will send you a checklist detailing these editorial and formatting requirements in about a week. Please do not finalize your revisions or upload the final materials until you receive this additional information.

In recognition of the time and expertise our reviewers provide to Nature Genetics's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes". For those reviewers who give their assent, we will be publishing their names alongside the published article.

While we prepare these instructions, we encourage the Corresponding Author to begin to review and collect the following:

-- Confirmation from all authors that the manuscript correctly states their names, institutional affiliations, funding IDs, consortium membership and roles, author or collaborator status, and author contributions.

-- Declarations of any financial and non-financial competing interests from any author. For the sake of transparency and to help readers form their own judgment of potential bias, the Nature Research Journals require authors to declare any financial and non-financial competing interests in relation to the work described in the submitted manuscript. This declaration must be complete, including author initials, in the final manuscript text.

If you have any questions as you begin to prepare your submission please feel free to contact our Editorial offices at genetics@us.nature.com. We are happy to assist you.

Thank you again for your interest in Nature Genetics.

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665

Reviewer #1 (Remarks to the Author):

The authors have responded to my comments fairly and adequately. I still worry a little about confounding by admixture, but I acknowledge it is difficult to perform proper admixture mapping on the ImmunoChip. No further concerns.

Reviewer #2 (Remarks to the Author):

Overall, these revisions successfully addressed our comments, and this study should be of broad interest to the field.

A few remaining minor concerns include the following:

1. On line 300, the authors state that “ATAC-seq offers a high-resolution map of accessible chromatin with regulatory function.” A more accurate statement is that ATAC-seq offers a high-resolution map of accessible chromatin that has "potential regulatory function."

2. On line 372-376, the authors imply that variants altering chromatin accessibility are more likely to be functionally relevant than those that do not, despite other functional mechanisms through which variants may act (such as altering TF binding). The authors may consider softer wording for this statement on line 372, such as “These data help to prioritize…” Additionally, the authors may consider similar wording on line 391.

3. In Figure 3g, it is unclear if the LocusCompare plots are missing data, as the colors in the r^2 legend do not appear in the figure.

4. Due to the quality of the image, the results of the EMSA in Supplementary Figure 12 are not totally clear, and the authors may consider omitting this.

Reviewer #3 (Remarks to the Author):

I am satisfied with the answers and don't have additional major comments on the paper. Personally, I think additional figure is valuable to add to the supplementary, since it provides more extensive and complete information on top of results presented in the main text.

Final Decision Letter:

In reply please quote: NG-A55094R1 Todd

5th May 2021

Dear John,

I am delighted to say that your manuscript "Fine-mapping, trans-ancestral, and genomic analyses identify causal variants, cells, genes, and drug targets for type 1 diabetes" has been accepted for publication in an upcoming issue of Nature Genetics.

Prior to setting your manuscript, we may make minor changes to enhance the lucidity of the text and with reference to our house style. We therefore ask that you examine the proofs most carefully to ensure that we have not inadvertently altered the sense of your text in any way.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A55094R1) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we will be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more

about Transformative Journals

<B>Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January 2021.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily

reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to http://www.nature.com/protocolexchange/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A55094R1).

Sincerely,
Kyle

Kyle Vogan, PhD
Senior Editor
Nature Genetics
https://orcid.org/0000-0001-9565-9665