Resolving mechanisms of immune-mediated disease in primary CD4 T cells

Bourges C, Groff AF, Burren OS, Gerhardinger C, Mattioli K, Hutchinson A, Hu T, Anand T, Epping MW, Wallace C, Smith KGC, Rinn JL, Lee JC

Review timeline:

Submission date:	1st Feb 2020
Editorial Decision:	17th Feb 2020
Revision received:	4th Mar 2020
Accepted:	9th Mar 2020

Editor: Céline Carret

Transaction Report:

(Note: An earlier version of this manuscript was assessed by another journal and was then transferred to EMBO Molecular Medicine. As the original review of the manuscript was performed outside of EMBO Molecular Medicine's transparent review process policy, no Peer Review Process information is available for this article. With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 17th Feb 2020

Thank you for the submission of your manuscript to EMBO Molecular Medicine. I am happy to report that we have now received the enclosed recommendations from our editorial advisers.

As you will see the advisers are supportive of publication and I am pleased to inform you that we will be able to accept your manuscript pending the following final amendments:

1) Please address the minor changes commented by advisors #1 and #2. We would like to encourage you to provide the data suggested and modify figures and text as recommended by this advisor.

Please provide a point-by-point letter INCLUDING my comments as well as the advisor' reports and your detailed responses to their comments (as Word file).

***** Reviewer's comments *****
Advisor #1:

I had a look to the Ms and authors' reply to reviewers' critics.
In my opinion, this Ms merits publication since by using a seemingly new approach they show that it is possible to provide biologically relevant information related to non coding variants shown to affect a superenhancer at least in one proof of principle case.

It is a quite difficult Ms to read, though now improved, perhaps not as strong in message as the authors claim (i.e., to potentially decipher the meaning of potentially all non coding variants), i.e. some rewriting is needed to make the paper more accessible to the general readers and maybe tuning down the strong claims of the abstract. But still, as said above

this approach adds to the armamentarium of potential strategies to tackle such (tough) questions.

Thus I am inclining to be positive.

Advisor #2:

I had the opportunity to go over the manuscript and the response to reviewers. I think the authors have clarified all the points raised by the reviewers. I think the manuscript will be of substantial interest to the broader complex genetics/functional genomics community. The experiments are well presented and the results advance our knowledge, especially in the context of functional consequences at the TNFAIP3 locus. I enjoyed reading this manuscript and thought the results were very relevant.

I have a few comments that I think would make the manuscript stronger if the authors would be willing to address them (I appreciate that they have already done comprehensive work addressing the previous comments).

1) Overall, I think the approach is important and I was disappointed that the authors didn't provide more comprehensive comparison of MPRA results over other approaches that prioritise functional variants, e.g. they briefly mention conservation scores in the context of the NFkB binding at TNFAIP3 locus, similar analysis could be done across all the tested variants to provide an overview of how many variants could have been prioritised prior the MPRA approach.

Somewhat on this note, I didn't have the access to supplementary materials, I think one piece of information that would be critical to share with the community alongside the publication is a supplementary table with all the variants tested, the primer sequences, raw result values from their experiments and computed expression values. I can imagine many researchers might be interested to check if the loci they are working have been tested in this experiments and with what results.

2) In response to the Reviewer 1 comments [about performing multiple comparisons using MPRAs in primary cells vs. cell lines]:

I agree that the multiple comparisons are beyond the scope of the manuscript. However, I also agree with the initial intuition of the authors to include the comparison between Jurkat and T cells. Even if well established, this is an important message and it is critical we think about ways to follow up GWAS signals in the relevant cellular models. Therefore, I suggest including the comparison panel from Figure 2 in the manuscript, perhaps as a part of a supplementary figure.

3) In figures, the pie charts are used to depict posterior probabilities of fine-mapped SNPs and they are not a good representation, the posterior probability values are not directly depicted and difficult to infer from the pie charts themselves, I suggest a different representation or providing the exact values.

1st Revision - authors' response 4th Mar 2020

***** Reviewer's comments *****
Advisor #1:

I had a look to the Ms and authors' reply to reviewers' critics.
In my opinion, this Ms merits publication since by using a seemingly new approach they show that it is possible to provide biologically relevant information related to non coding variants shown to affect a superenhancer at least in one proof of principle case.

It is a quite difficult Ms to read, though now improved, perhaps not as strong in message as the authors claim (i.e., to potentially decipher the meaning of potentially all non coding variants), i.e. some rewriting is needed to make the paper more accessible to the general readers and maybe tuning down the strong claims of the abstract. But still, as said above this approach adds to the armamentarium of potential strategies to tackle such (tough) questions.

Thus I am inclining to be positive.

We are grateful for these constructive comments. We did not intend to imply that adapted MPRA could decipher the meaning of all non-coding variants, and have now re-written the abstract in line with the advisor's suggestions. We have also rewritten some of the other sections that we think may not have been fully accessible to general readers. Edits are tracked in the revised version of the manuscript.

> Abstract
>
> "Deriving mechanisms of immune-mediated disease from GWAS data remains a formidable challenge, with attempts to identify causal variants being frequently hampered by strong linkage disequilibrium. To determine whether causal variants could be identified from their functional effects, we adapted a massively-parallel reporter assay for use in primary CD4 T-cells, the cell-type whose regulatory DNA is most enriched for immune-mediated disease SNPs. This enabled the effects of candidate SNPs to be examined in a relevant cellular context, and generated testable hypotheses into disease mechanisms. To illustrate the power of this approach, we investigated a locus that has been linked to 6 immune-mediated diseases but cannot be fine-mapped. By studying the lead expression-modulating SNP, we uncovered an NF-κB-driven regulatory circuit which constrains T-cell activation through the dynamic formation of a super-enhancer that upregulates TNFAIP3 (A20), a key NF-κB inhibitor. In activated T-cells, this feedback circuit is disrupted – and super-enhancer formation prevented – by the risk variant at the lead SNP, leading to unrestrained T-cell activation via a molecular mechanism that appears to broadly predispose to human autoimmunity."

Advisor #2:

I had the opportunity to go over the manuscript and the response to reviewers. I think the authors have clarified all the points raised by the reviewers. I think the manuscript will be of substantial interest to the broader complex genetics/functional genomics community. The experiments are well presented and the results advance our knowledge, especially in the context of functional consequences at the TNFAIP3 locus. I enjoyed reading this manuscript and thought the results were very relevant.

I have a few comments that I think would make the manuscript stronger if the authors would be willing to address them (I appreciate that they have already done comprehensive work addressing the previous comments).

We are grateful for these positive comments and for the time the advisor has taken in reading our manuscript and making constructive suggestions.

1) Overall, I think the approach is important and I was disappointed that the authors didn't provide more comprehensive comparison of MPRA results over other approaches that prioritise functional variants, e.g. they briefly mention conservation scores in the context of the NFkB binding at TNFAIP3 locus, similar analysis could be done across all the tested variants to provide an overview of how many variants could have been prioritised prior the MPRA approach.

We have now used 2 *in silico* methods to assess how many of these variants could have been prioritised by other means (DeepSEA, which uses evolutionary conservation and

predicted chromatin effects, and RegulomeDB, which uses a variety of public datasets). The results of these analyses are presented in a new Expanded View Dataset (Dataset EV1). Between these 2 methods, the lead MPRA SNP was classified as the most significant functional SNP at 3 of the 14 loci (2 by DeepSEA and 1 by RegulomeDB). DeepSEA also predicted that 3 other lead SNPs would be functionally significant, but ranked other candidate SNPs ahead of them (most of which had no expression-modulating effect in the MPRA). These *in silico* methods have been reported to be better at predicting negative effects compared with positive effects, and consistent with this, rs1736137 was included in the SNPs classified as non-functional. This is the variant shown in Fig 3A, which was previously identified as a causal variant by fine-mapping (and which has a highly significant expression-modulating effect in MPRA). Overall, these data indicate that while other methods can sometimes identify true expression-modulating variants, they are often misleading and cannot substitute for experimental methods. We have added the following text to the Results:

> "To determine whether these variants could have been prioritised by other means, we compared the MPRA results with *in silico* methods designed to identify functional variants – DeepSEA (Zhou & Troyanskaya, 2015) and RegulomeDB (Dong & Boyle, 2019) (Dataset EV1). Considering these approaches together, the lead MPRA SNP was predicted to be the most functionally significant variant at 3/14 loci (2 by DeepSEA, 1 by RegulomeDB). DeepSEA also predicted that 3 more lead SNPs would be functionally significant, but prioritised other candidate SNPs at these loci (most of which had no expression-modulating effect in CD4 T cells). At the remaining 8 loci, the lead MPRA SNP was not predicted to have an expression-modulating effect – consistent with these methods being better at predicting negative effects than positive effects (Dong & Boyle, 2019) and highlighting the value of studying disease-associated loci in relevant primary cells."

Somewhat on this note, I didn't have the access to supplementary materials, I think one piece of information that would be critical to share with the community alongside the publication is a supplementary table with all the variants tested, the primer sequences, raw result values from their experiments and computed expression values. I can imagine many researchers might be interested to check if the loci they are working have been tested in this experiments and with what results.

Much of this data was already included in the supplementary tables (now Datasets EV2 and EV3) which represent the meta-analysis results for every variant tested in resting and stimulated CD4 T cells respectively. This will enable researchers to examine the results at their loci of interest. The raw and processed sequencing values, and all of the sequences tested, are provided in our GEO submission (GSE135925). We can also provide these files separately (e.g. as additional Extended View Datasets) if this is felt to be necessary, but they are very large files, and it would probably be easier for interested researchers to access them from GEO. Let us know what you would prefer.

2) In response to the Reviewer 1 comments [about performing multiple comparisons using MPRAs in primary cells vs. cell lines]:

I agree that the multiple comparisons are beyond the scope of the manuscript. However, I also agree with the initial intuition of the authors to include the comparison between Jurkat and T cells. Even if well established, this is an important message and it is critical we think about ways to follow up GWAS signals in the relevant cellular models. Therefore, I suggest including the comparison panel from Figure 2 in the manuscript, perhaps as a part of a supplementary figure.

We very much appreciate this comment, and agree that this is an important message. We have now included the comparison panel in Appendix Figure S1 and added text to explain this:

> "The effects observed in resting and stimulated CD4 T cells were highly correlated (Appendix Fig S1E), but these effects did not correlate particularly well with results

obtained in Jurkat cells (an immortalised CD4 T cell line) – reinforcing the value of using an appropriate cellular model when studying human disease (Appendix Fig S1F)."

3) In figures, the pie charts are used to depict posterior probabilities of fine-mapped SNPs and they are not a good representation, the posterior probability values are not directly depicted and difficult to infer from the pie charts themselves, I suggest a different representation or providing the exact values.

We have now added the exact posterior probability values to these figure panels.

The authors performed the requested editorial changes.

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Corresponding Author Name: James Lee
Journal Submitted to: EMBO Molecular Medicine
Manuscript Number: EMM-2020-12112

Reporting Checklist For Life Sciences Articles (Rev. June 2017)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

A- Figures

1. Data

The data shown in figures should satisfy the following conditions:

- the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
- figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
- graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
- if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified
- Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

2. Captions

Each figure caption should contain the following information, for each panel where they are relevant:

- a specification of the experimental system investigated (eg cell line, species name).
- the assay(s) and method(s) used to carry out the reported observations and measurements
- an explicit mention of the biological and chemical entity(ies) that are being measured.
- an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.

- the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
- a statement of how many times the experiment shown was independently replicated in the laboratory.
- definitions of statistical methods and measures:
  - common tests, such as t-test (please specify whether paired vs. unpaired), simple $\chi 2$ tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - exact statistical test results, e.g., P values = x but not P values < x;
  - definition of 'center values' as median or average;
  - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

In the pink boxes below, please ensure that the answers to the following questions are reported in the manuscript itself. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable). We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?	MPRA sample size was chosen to ensure that the results were reproducible, and to control for inter-individual differences in CD4 T cell composition and the reduced dynamic range expected with a stronger promoter. No formal power calculation was performed as it was not possible to estimate the likely effect size using a newly-adapted technique. Comparative experiments (between 2 or more groups) were performed with 6 samples per group to provide 80% power to detect a Standradised Effect Size = 2, alpha = 0.05.
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	NA
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	No samples were excluded from any analysis
3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.	Internal controls were used wherever possible to minimise potential biases related to comparisons between unrelated samples (e.g. comparing between each allele in a heterozygous sample).
For animal studies, include a statement about randomization even if no randomization was used.	NA
4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.	NA
4.b. For animal studies, include a statement about blinding even if no blinding was done	NA
5. For every figure, are statistical tests justified as appropriate?	Yes, the statistical tests used are indicated in the figure legends. Exact P values are provided in Appendix Table S4.
Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.	The distribution of MPRA barcode counts have previously been shown to be best modelled by a beta-binomial distribution - hence we used a statistical method that uses this assumption (QuASAR-MPRA). For other data we assumed a Gaussian distribution and used a Shapiro-Wilk test to confirm this.
Is there an estimate of variation within each group of data?	Yes, the variation within groups of data are represented by SD or SEM, as indicated in the figure legend.

USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-report

http://grants.nih.gov/grants/olaw/olaw.htm
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm
http://ClinicalTrials.gov
http://www.consort-statement.org
http://www.consort-statement.org/checklists/view/32-consort/66-title

http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumo

http://datadryad.org

http://figshare.com

http://www.ncbi.nlm.nih.gov/gap

http://www.ebi.ac.uk/ega

http://biomodels.net/

http://biomodels.net/miriam/
http://jij.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

Is the variance similar between the groups that are being statistically compared?	Yes

C- Reagents

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right).	Catalog numbers, suppliers and dilutions are provided for all antibodies
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	The source of Jurkat cells is provided in the acknowledgements. These had been authenticated by the supplier and were tested for mycoplasma contamination.

* for all hyperlinks, please see the table at the top right of the document

D- Animal Models

8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA
10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.	NA

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	Cambridgeshire Regional Ethics Committee (REC:08/H0308/176 and REC:08/H0306/21).
12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.	This statement is included in the manuscript
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	NA
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	NA
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	NA
16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list.	NA
17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines.	NA

F- Data Accessibility

18: Provide a "Data Availability" section at the end of the Materials & Methods, listing the accession codes for data generated in this study and deposited in a public database (e.g. RNA-Seq data: Gene Expression Omnibus GSE39462, Proteomics data: PRIDE PXD000208 etc.) Please refer to our author guidelines for 'Data Deposition'. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Functional genomics data e. Proteomics and molecular interactions	A Data Availability section is provided at the end of the Materials & Methods listing the accession codes for data generated in this study and deposited in GEO.
19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).	Data is deposited in GEO - per Data Availability section
20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right).	See answers to 19 and 20.
21. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biomodels (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information.	NA

G- Dual use research of concern

22. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could.	No