




RESEARCH NOTE

Meta-analysis of crowdsourced data compendia suggests pan-disease transcriptional signatures of autoimmunity [version 1; referees: 2 approved]

William W. Lau^{1*}, Rachel Sparks^{2*}, OMiCC Jamboree Working Group,
John S. Tsang ²

¹Office of Intramural Research, Center for Information Technology, National Institutes of Health, Bethesda, Maryland, USA

²Systems Genomics and Bioinformatics Unit, Laboratory of Systems Biology, National Institutes of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

* Equal contributors

v1 First published: 20 Dec 2016, 5:2884 (doi: [10.12688/f1000research.10465.1](https://doi.org/10.12688/f1000research.10465.1))
Latest published: 20 Dec 2016, 5:2884 (doi: [10.12688/f1000research.10465.1](https://doi.org/10.12688/f1000research.10465.1))

Abstract

Background: The proliferation of publicly accessible large-scale biological data together with increasing availability of bioinformatics tools have the potential to transform biomedical research. Here we report a crowdsourcing Jamboree that explored whether a team of volunteer biologists without formal bioinformatics training could use OMiCC, a crowdsourcing web platform that facilitates the reuse and (meta-) analysis of public gene expression data, to compile and annotate gene expression data, and design comparisons between disease and control sample groups.



Methods: The Jamboree focused on several common human autoimmune diseases, including systemic lupus erythematosus (SLE), multiple sclerosis (MS), type I diabetes (DM1), and rheumatoid arthritis (RA), and the corresponding mouse models. Meta-analyses were performed in OMiCC using comparisons constructed by the participants to identify 1) gene expression signatures for each disease (disease versus healthy controls at the gene expression and biological pathway levels), 2) conserved signatures across all diseases within each species (pan-disease signatures), and 3) conserved signatures between species for each disease and across all diseases (cross-species signatures).



Results: A large number of differentially expressed genes were identified for each disease based on meta-analysis, with observed overlap among diseases both within and across species. Gene set/pathway enrichment of upregulated genes suggested conserved signatures (e.g., interferon) across all human and mouse conditions.

Conclusions: Our Jamboree exercise provides evidence that when enabled by appropriate tools, a "crowd" of biologists can work together to accelerate the pace by which the increasingly large amounts of public data can be reused and

Open Peer Review

Referee Status:  

	Invited Referees	
	1	2
version 1 published 20 Dec 2016	 report	 report

- 1 **Hans Lehrach** , Max Planck Institute for Molecular Genetics (MPIMG) Germany,
Hans-Jörg Warnatz, Max Planck Institute for Molecular Genetics (MPIMG) Germany
Thomas Risch, Max Planck Institute for Molecular Genetics (MPIMG) Germany
- 2 **Markus Riester** , Novartis Institutes for BioMedical Research USA

Discuss this article

Comments (0)

meta-analyzed for generating and testing hypotheses. Our encouraging experience suggests that a similar crowdsourcing approach can be used to explore other biological questions.

Corresponding author: John S. Tsang (john.tsang@nih.gov)

How to cite this article: Lau WW, Sparks R, OMiCC Jamboree Working Group and Tsang JS. **Meta-analysis of crowdsourced data compendia suggests pan-disease transcriptional signatures of autoimmunity [version 1; referees: 2 approved]** *F1000Research* 2016, 5:2884 (doi: [10.12688/f1000research.10465.1](https://doi.org/10.12688/f1000research.10465.1))

Copyright: © 2016 Lau WW *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Grant information: This research was funded by the Intramural Programs of the National Institute of Allergy and Infectious Diseases (NIAID) and the Center for Information Technology (CIT) at the National Institutes of Health.

Competing interests: No competing interests were disclosed.

First published: 20 Dec 2016, 5:2884 (doi: [10.12688/f1000research.10465.1](https://doi.org/10.12688/f1000research.10465.1))

Introduction

The volume of large-scale biological data in the public domain is increasing at an unprecedented rate; as a result, data reuse is becoming an increasingly viable means to generate and test hypotheses¹ (Figure 1). The reusability of public data, however, depends on the quality and availability of the associated meta-data and annotations. Given a research goal, for example, to generate gene expression signatures for a biological phenotype, one has to first identify and annotate relevant public data, followed by the construction of comparison group pairs (or CGP - see Figure 1 - e.g., a group of samples corresponding to the phenotype of interest versus a group of control samples) and subsequent bioinformatics analyses. Bench scientists are uniquely empowered with biological knowledge to identify and annotate relevant public data and form proper comparisons. Recently, there have also been a variety of crowdsourcing efforts, including hackathons, datathons and open challenges, in which diverse groups of individuals work together to accelerate the pace of pursuing common goals^{2,3}. Thus, we were interested in assessing what could be accomplished by harnessing the collective biological knowledge of a group of biologists to explore, identify, and annotate public datasets

when empowered with a user-friendly web platform and a shared scientific goal; would this approach accelerate the pace by which useful biological comparison groups could be constructed and utilized? What would be the specific strengths and hurdles, from both a social and scientific perspective? Towards addressing these questions, we conducted a crowdsourcing “Jamboree” exercise within the NIH immunological community to test the hypothesis that the use of OMiCC⁴ (<https://omicc.niaid.nih.gov>), an open, programming-free web platform that enables a crowdsourcing approach to public gene expression data reuse, can facilitate the rapid assembly of a large data compendium followed by bioinformatics analyses to generate biological hypotheses. Select aspects of this exercise, particularly on how it provides evidence that a tool such as OMiCC can enable biologists without bioinformatics training to directly explore public data, have been highlighted elsewhere⁵ and for which this work serves as a companion (also see supplemental website to ref. 5 - <https://omicc.niaid.nih.gov/2016-nih-jamboree-analysis/report.html>); here we focus on the post-Jamboree data quality control, analysis, and observations, as well as discussing the utility and caveats of this approach.

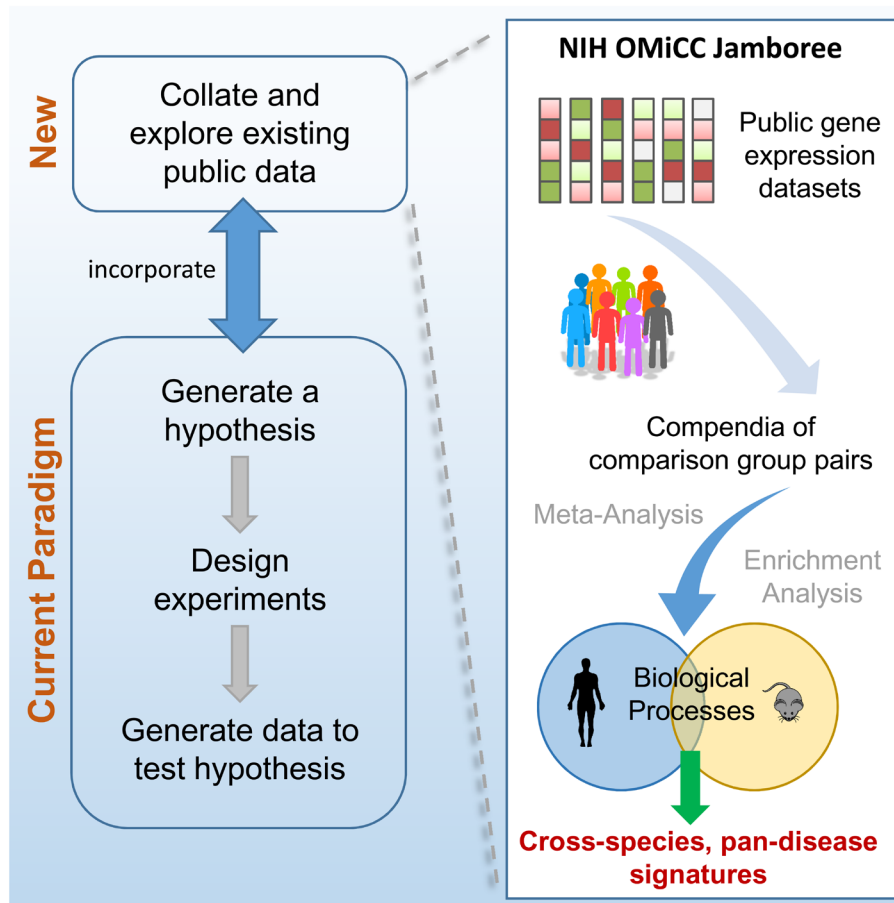


Figure 1. New research paradigm incorporating exploration and reuse of public data. Increasing availability of public data opens new opportunities for biologists to generate hypotheses. The NIH OMiCC Jamboree was a social experiment to assess whether a group of biologists without computational experience can identify and annotate public datasets and construct CGPs using the OMiCC tool. This paper describes the data analysis, including meta-analysis and gene set enrichment analysis, to derive gene expression signatures across human and mouse.

For this crowdsourcing experiment, we focused on assessing the gene expression patterns of and shared signatures among several common human autoimmune and inflammatory diseases and the corresponding mouse models. Mouse models of human diseases can be informative for studying disease mechanisms, but may not accurately reflect the underlying biology in humans^{6,7}. We were particularly interested in determining whether we could detect shared gene expression signatures among diseases (pan-disease signatures), including type I diabetes (DM1), multiple sclerosis (MS), rheumatoid arthritis (RA), sarcoidosis (sarcoid), Sjögren's syndrome (SS), and systemic lupus erythematosus (SLE), as well as among their mouse models. We chose these diseases because they have reasonably well-established mouse models and both human and mouse gene expression data are available publicly. A prior study has also evaluated pan-disease transcriptional signatures and found conserved signals across RA, SLE and SS⁸. Here we are including more diseases and are additionally interested in assessing whether human and mouse have shared pan-disease signatures. Given that data from mouse are often generated from non-blood tissues while those from humans usually come from blood, such cross-species comparisons could also point to potential links between blood and non-blood tissues. Cross-species comparisons of gene expression signatures have been performed previously in sepsis, for example, where both conserved and divergent signals have been detected^{6,9,10}. While our analyses are motivated by these questions, our primary goal here is not to validate previous findings or to generate new biological knowledge per se, but to use this exercise as a proof-of-concept to illustrate the potential utility of data reuse with crowdsourcing.

Methods

Crowdsourcing: team composition and responsibilities

The Jamboree was advertised on the NIH Immunology Listserv, which is primarily subscribed by local researchers to disseminate and share immunology-focused information. No inclusion or exclusion criteria were applied to the identification of the participants. The Jamboree involved a half-day group training session using the OMiCC platform followed by a day-long Jamboree, during which 29 volunteer biologists were separated into *ten* 2- or 3-member teams to search OMiCC for public gene expression datasets of DM1, MS, RA, sarcoid, and SLE (Figure 2). The assignments of teams and topics were based on the participants' self-declared research backgrounds; additionally, each group had at least one participant who felt proficient using OMiCC after the half-day orientation. Half of the groups were assigned to focus on humans with one group per disease and similarly, the other half of the groups were assigned to the corresponding mouse models. The participants were asked to use OMiCC (<https://omicc.niaid.nih.gov>) to annotate sample groups and create CGPs between disease and control samples in the studies they identified. They were also encouraged to consult the primary publications describing the studies to help ensure the accuracy of their annotations. Although Sjögren's syndrome was not originally assigned to any group, the sarcoidosis groups were not able to find sufficient studies from which to construct CGPs and thus was subsequently assigned to focus on Sjögren's syndrome. Compendia of CGPs created by the teams can be accessed and reused within OMiCC (see Data and Software Availability).

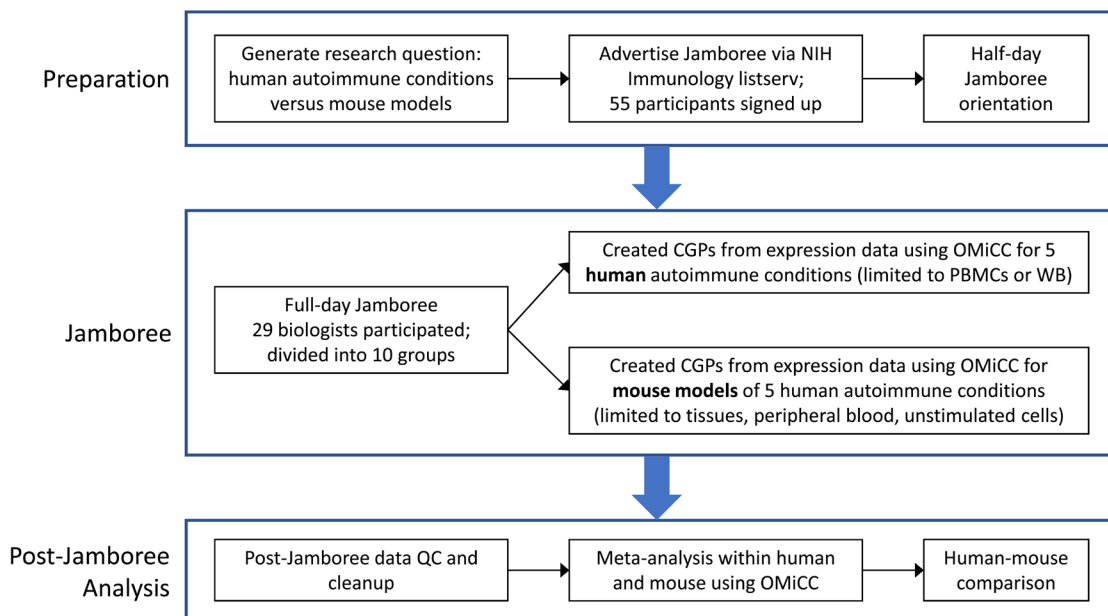


Figure 2. NIH OMiCC Jamboree workflow/timeline. Workflow of the NIH Jamboree detailing steps taken prior to, during, and after the actual Jamboree event.

Data curation and quality control (QC) for downstream analyses

A total of 86 human CGPs were collected from the Jamboree, spreading across the six diseases. Participants were instructed to identify public microarray datasets in OMiCC that contained data derived from whole blood (WB) or peripheral blood mononuclear cells (PBMCs) of both healthy controls and affected patients; they were asked to avoid studies of stimulated cells. Post-Jamboree CGP QC was required in order to correct misplaced annotations or to standardize annotations created with free text. Only 54 of the 86 CGPs were created with samples annotated as PBMC or WB. We removed an additional 15 CGPs for the following reasons: 1) incorrect sample annotations; 2) the CGP did not contain sample groups from both cases and controls; and 3) the samples in the CGP significantly overlapped with those in another CGP (Jaccard index > 66%). As a result, 39 human CGPs representing five diseases (note that no WB or PBMC samples passed QC for Sjögren's syndrome) were included in the downstream analyses (Table 1).

Participants of the mouse teams created a total of 94 CGPs from mouse models of the aforementioned six diseases. Participants were instructed to identify public microarray datasets in OMiCC that contained data derived from non-blood tissues, WB, or PBMCs of both healthy and diseased mice; they were asked to avoid studies of stimulated cells. Due to the complexities of the mouse models and studies, the overall quality of the CGPs was comparatively lower than that of the human CGPs. For example, a substantial fraction of CGPs contained data from stimulated cells despite our explicit call for avoiding such studies; these CGPs were excluded.

Four CGPs were excluded because they were duplicates of other CGPs. Some CGPs had young, clinically unaffected mice as controls and older, clinically ill mice as cases (e.g., age-related disease progression models), while others were obtained from purified cell subsets (e.g., CD4+ T cells and B cells). We still included these CGPs in our final set with the goal of identifying conserved signals through meta-analysis. After this curation process, 34 CGPs remained across four diseases because no samples from sarcoidosis or Sjögren's syndrome passed QC (Table 1).

In addition to the individual disease datasets (i.e., a collection of CGPs), all the CGPs for each species were combined to create a pan-disease compendium—one for human and one for mouse.

Meta-analysis

Meta-analysis was conducted in OMiCC to derive differential expression signatures for each dataset (note that OMiCC uses a rank-based meta-analysis R package called RankProd¹¹, version 2.36.0). The results were reported at the gene level, based on internal OMiCC mappings between platform-specific probe identifiers and standard HUGO gene names. For each gene, this method reports the false prediction rate (PFP - similar to false discovery rate (FDR)) for both increased and decreased expression (herein referred to as the UP and DOWN genes, or differentially expressed (DE) genes when they are combined). Using PFP ≤ 0.05 as a threshold, we identified UP and DOWN genes for each disease (meta-analysis per species) and for each species (meta-analysis across all CGPs within a species to derive pan-disease gene signatures). Genes with conflicting indications (which is possible with the RankProd method used

Table 1. Overview of datasets generated in the Jamboree. Each dataset is comprised of a set of comparison group pairs (CGPs), which in turn contain a number of case and control microarray samples. Since the same sample may be selected in more than one CGP, the number of unique samples in each group is listed. Common genes are those measured across all platforms in a dataset. These genes were considered in the ranked-based meta-analyses, some of which were identified as having significantly (PFP ≤ 0.05) increased (UP) or decreased (DOWN) expression. Genes in both UP and DOWN lists were removed. The datasets 'human_pan-disease' and 'mouse_pan-disease' were created by combining all CGPs constructed for each species.

Dataset	CGPs	Unique Cases	Unique Controls	Common Genes	UP Genes	DOWN Genes
human_dm1	3	168	146	10681	762	766
human_ms	9	171	99	7909	971	1106
human_pan-disease	39	1101	795	7808	1021	976
human_ra	15	335	188	11254	1316	1550
human_sarcoid	4	141	160	16012	2513	2381
human_sle	8	286	202	10689	1183	500
mouse_dm1	13	62	26	5753	739	670
mouse_ms	4	17	14	7009	326	403
mouse_pan-disease	34	141	84	5644	849	682
mouse_ra	9	32	20	12801	1123	662
mouse_sle	8	30	24	13040	626	524

by OMiCC), i.e. those suggested to have increased and decreased expression for the same disease, were removed. The resulting gene lists and meta-analysis output were exported as text files for further processing. Prior to any downstream analyses, mouse genes were mapped to human genes using NCBI's homology maps (ftp://ftp.ncbi.nlm.nih.gov/pub/homology_maps/human/, version 12/27/15) and those with either no or non-unique mappings were discarded. The robustness of the RankProd (rank based) results was evaluated using another effect-size metric called Cohen's d , which was calculated in R as

$$d = t \sqrt{\frac{n_D + n_C}{n_D n_C} \left(\frac{n_D + n_C}{n_D + n_C - 2} \right)},$$

where t is the t statistic reported by OMiCC, and n_D and n_C are the number of samples in the disease and control groups, respectively.

Gene set enrichment analysis

Gene set based enrichment (or over-representation) analyses were carried out separately for the UP and DOWN genes from each of the four diseases in human and mouse (i.e., DM1, MS, RA, and SLE) against terms in KEGG (<http://www.genome.jp/kegg/>) or Reactome (<http://www.reactome.org/>) containing 3 to 500 genes, using the R clusterProfiler¹² (version 3.0.5) and ReactomePA¹³ (version 1.16.2) packages, respectively. In addition, to illustrate how similar analyses can be performed without any programming, enrichment analyses were also carried out using the web-based Toppgene tool¹⁴ (<https://toppgene.cchmc.org/enrichment.jsp>; using default settings and discarding any input gene that mapped to multiple entries). Pan-disease signatures were generated by meta-analyzing each of the two pan-disease compendia (one for human and one for mouse)—a pan-disease compendium contains the CGPs from all diseases within a species. The method implemented by the above software determines enrichment by evaluating the statistical significance of the overlap between the input DE gene list and target gene sets using the hypergeometric test, and we considered gene sets and pathways with an adjusted p -value of ≤ 0.05 to be significantly enriched. Conserved signatures between human and mouse were determined simply by finding the gene sets and pathways that were significantly enriched in both human and mouse.

Ethics

This work did not require ethics approval, as per NIH guidelines.

Results

Dataset 1. R data file

<http://dx.doi.org/10.5256/f1000research.10465.d146994>

Contains: 1) detailed information about the CGPs included in our analyses; 2) gene-by-compendium matrices of PFP values (can be interpreted as FDR) outputted by OMiCC (one matrix for UP genes; another for DOWN genes)—starting with this data matrix the user can elect to use any PFP cutoff to define DE genes (note that we have one compendium per disease per species, and a pan-disease compendium per species); and 3) gene-set over-representation analysis results generated in R.

Dataset 2. R markdown script to generate the data analysis report

<http://dx.doi.org/10.5256/f1000research.10465.d146995>

The script can be used with Dataset 1 as the data source to generate the main data figures and associated descriptions.

Dataset 3. Meta-analysis output files exported from OMiCC

<http://dx.doi.org/10.5256/f1000research.10465.d146996>

Dataset 4. Results of Toppgene analyses against KEGG, Reactome, and Gene Ontology (GO) Biological Process terms using the DE genes listed in Table S1 as input

<http://dx.doi.org/10.5256/f1000research.10465.d146997>

Disease gene signatures

Using the 39 human and 34 mouse CGPs created by the Jamboree participants (after QC), for each disease we ran meta-analysis across the CGPs in each disease within OMiCC. The number of DE genes varies substantially across diseases, possibly driven in part by differences in sample sizes and in the number of common genes shared among profiling platforms in each disease/CGP collection (Table 1 and Figure 3A; a list of DE genes for each disease is in Table S1). Comparison of the DE gene sets among diseases, separately for UP and DOWN genes, reveals strong signature overlaps among some diseases. Figures 3B–C show the odds ratios (OR) between pairs of diseases and those with OR > 1 have higher than the expected number of overlapping genes. Interestingly, there tended to be stronger overlap between pairs of diseases within a species than that between the same disease across human and mouse.

Effect size comparison

Given that meta-analysis results can be method dependent¹⁵, we next assessed the robustness of the rank-based meta-analysis method used by OMiCC by an independent analysis using a standardized effect-size metric known as Cohen's d , which is the mean difference of expression values between the case and control groups normalized by the joint standard deviation. For each CGP, we ranked the genes according to their Cohen's d value. Then for each collection of CGPs by which an OMiCC meta-analysis was performed (e.g., RA in humans), we calculated the median rank of each gene among the CGPs. The genes with large effect sizes according to Cohen's d should be enriched for those identified as having increased expression by the rank-based method in OMiCC, and conversely for the decreased expression genes. The comparison indicates that for most diseases, the OMiCC rank-based results are largely consistent with the effect-size approach, although there were a number of genes discordant between the two methods (Figure S1).

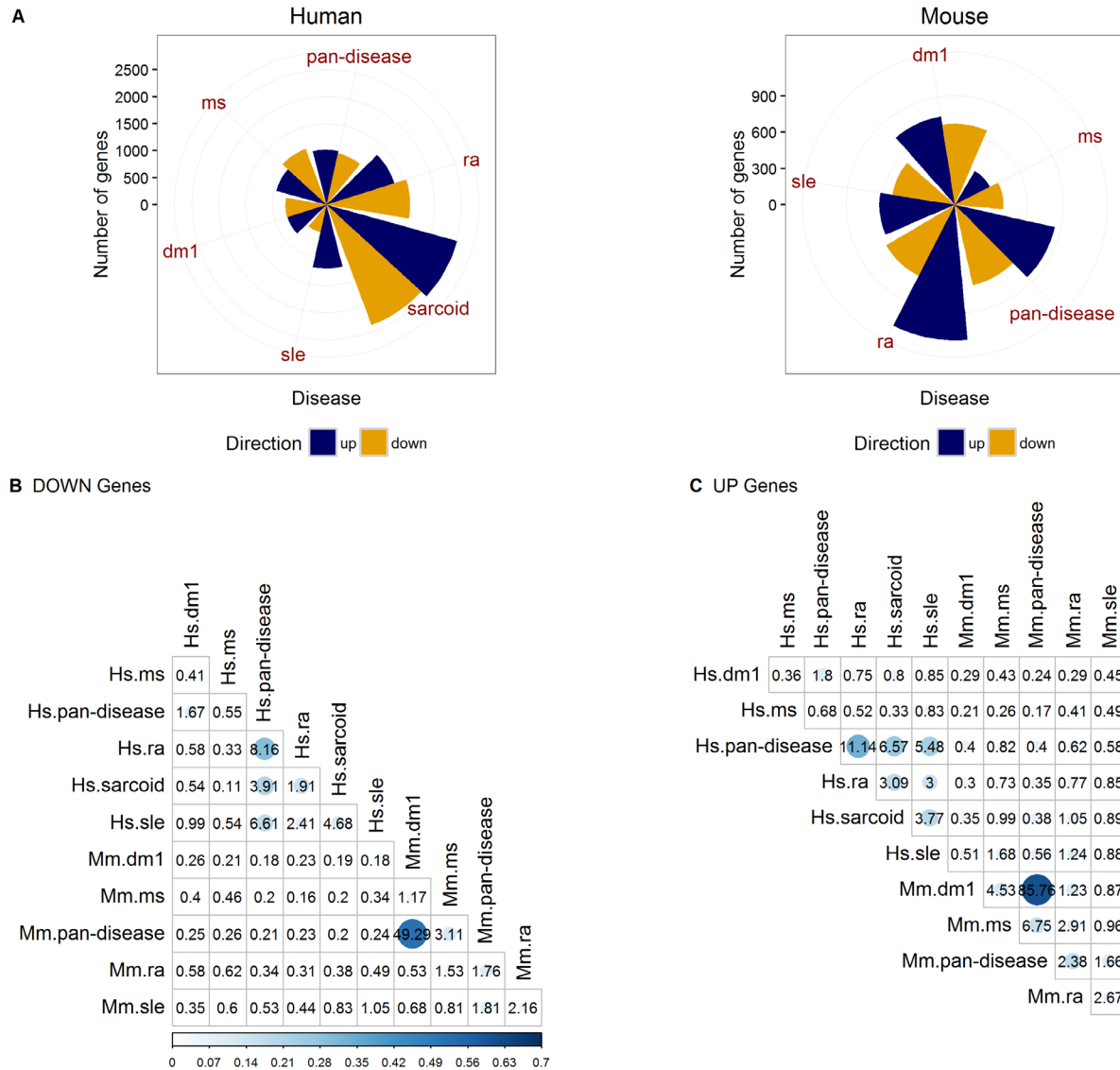


Figure 3. Comparison of differentially expressed genes across diseases in human and mouse. (A) Number of differentially expressed genes and **(B–C)** the proportion of genes that overlap (i.e. Jaccard index) between UP and DOWN genes (PFF <= 0.05) for pairs of diseases, as indicated by the size and color intensity of the circles. The number in each cell denotes the odds ratio, which is a measure of statistical association between the two groups based on the degree of gene overlap. An odds ratio of 1 suggests no association. Hs = human; Mm = Mouse.

Enriched biological processes/pathways

To gain higher level insights (e.g., pathway and biological processes) into the gene signatures identified, we assessed whether the UP and DOWN genes identified in the previous steps (Figure S2 and Table S2) are enriched for gene sets and pathways annotated in KEGG and Reactome. The analyses were conducted in R (version 3.3.1) and also with Toppgene (a web_based tool). Note that the differences between the R and Toppgene analyses can be partially explained by the fact that Toppgene assumes that all genes in the genome have been measured (i.e., the “background” set), which is not true in this analysis because we only assessed

genes common among gene-expression profiling platforms used to generate the data in the compendium (Table 1).

To generate pan-disease signatures, we next attempted to extract common enriched pathways across all diseases within each species. One simple approach is to identify overlapping signatures from the significantly enriched pathways of individual diseases, but its statistical power could be limited. Indeed, using this strategy the only globally enriched pathway is the Reactome term “Chemokine receptors bind chemokines” from the UP genes of the mouse datasets. Thus, we also tested an alternative approach where all

CGPs from each species across diseases were pooled together to form a single OMiCC compendium for meta-analysis (i.e., “human_pan-disease” and “mouse_pan-disease”; Figure 4). In this manner, the large number of samples increased the statistical power of the meta-analysis, thus resulting in the larger number of pan-disease enrichment signatures, including those reflecting broad immune activation and the well-appreciated interferon signature in human⁸ (Figure 4). However, this approach can potentially be confounded by variation in sample sizes across diseases, e.g., diseases with larger numbers of samples may dominate the signal.

Conserved signatures between human and mouse

We next used a conservative approach to assess shared gene set/pathway signatures between human and mouse by requiring that enriched terms be statistically significant in both human and mouse (after multiple-testing correction). Interestingly, using this criterion, all pan-disease enrichments conserved between human and mouse were derived from the UP genes (Figure 5), which may partially reflect that increases in immune cell frequencies (e.g., increases in monocytes in blood and/or tissues) were potential underlying drivers of these species-conserved, pan-disease signatures.

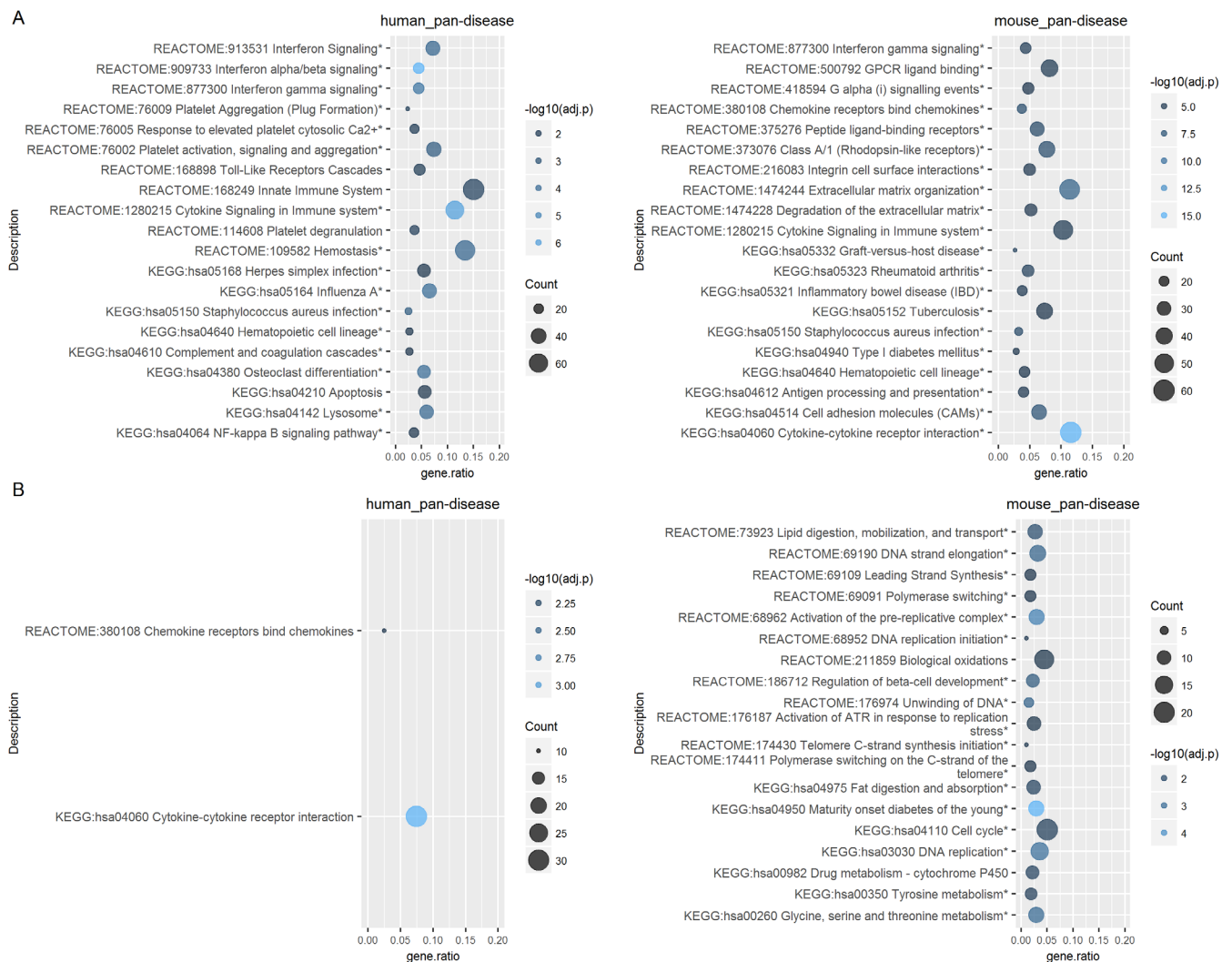


Figure 4. Pan-disease enrichment signatures. Over-representation analyses of the (A) UP genes and (B) DOWN genes (PFF <= 0.05) identified by using all CGPs from each species in the meta-analysis. The analyses were performed in both R and ToppGene; the top 20 enriched terms identified in R are shown. Terms found also in ToppGene are indicated by an asterisk (*). P-values are adjusted by Benjamini and Hochberg (BH) FDR correction (shown as 'adj.p'). Counts (indicated by circle size) and gene ratios (x-axis) respectively denote the number and proportion of genes in the UP or DOWN signature that also appear in the target gene set.

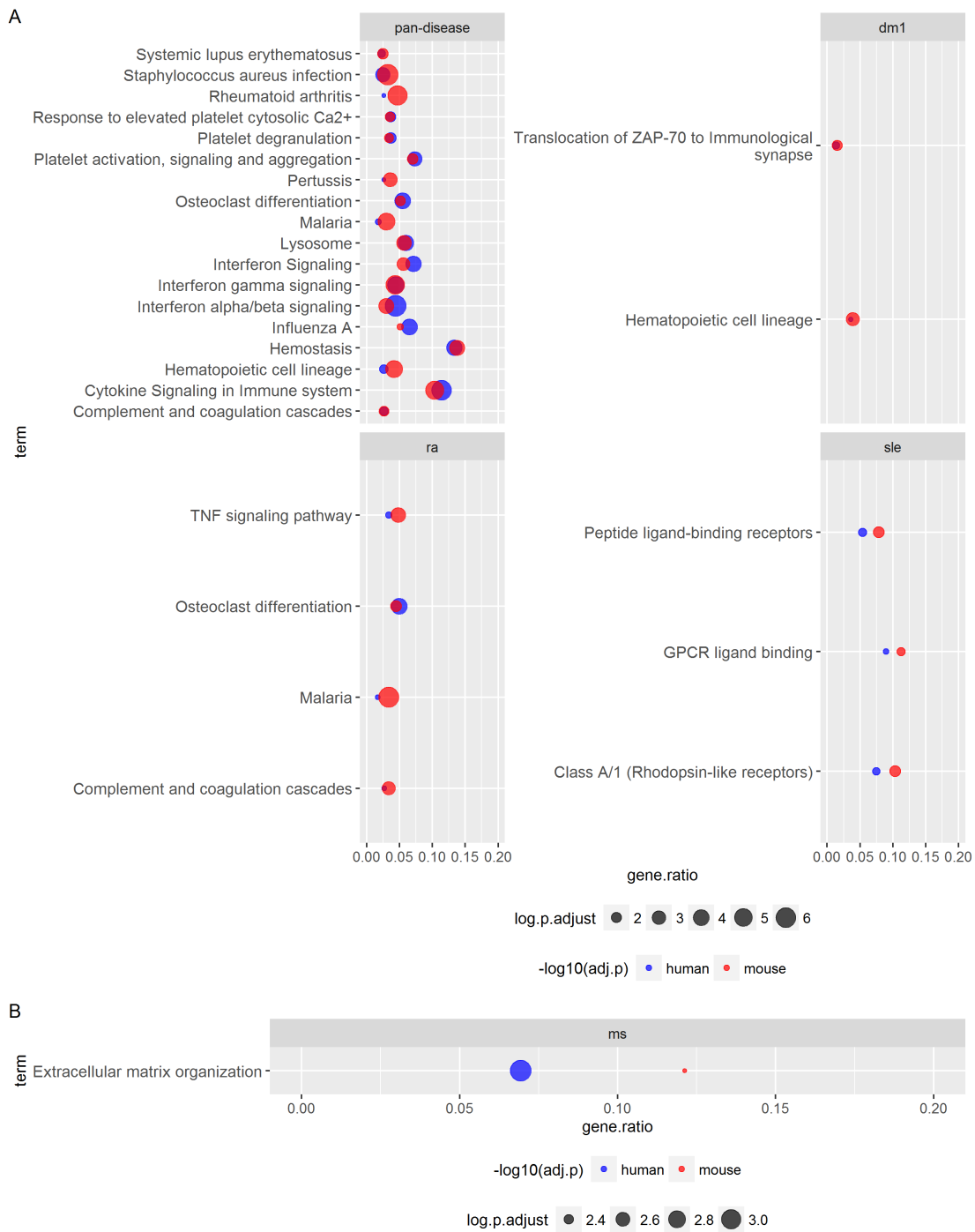


Figure 5. Pan-species, pan-disease signatures: Overlap of enriched biological processes between human and mouse. Over-representation analyses of the **(A)** UP genes and **(B)** DOWN genes (PFP ≤ 0.05) identified within OMiCC were carried out against KEGG and Reactome terms (see also Figures 4 and Figure S2 and Table S2). For each individual CGP compendium (disease or pan-disease), gene sets or terms with adjusted p-value ≤ 0.05 , as defined by the hypergeometric test after adjustment by the Benjamini and Hochberg (BH) FDR procedure, in both human and mouse are listed. These overlapping terms highlight signatures conserved between human and mouse. Gene ratios (x-axis) denote the fraction of genes in the respective signature (human and mouse as denote by blue and red, respectively) that are in the target gene set (y-axis).

Discussion

Our crowdsourcing exercise illustrates that a group of biologists without formal bioinformatics training can use OMiCC, a programming-free web-based platform, to generate a sizable number of CGPs during a day-long group exercise with a shared scientific goal. This is encouraging because CGP construction can be time consuming, requires biological expertise, and is often required for public data reuse and meta-analysis. Our observation suggests that other groups should be able to replicate our experience in their own institutions to pursue other scientific questions. However, there are some caveats: substantial QC was required to remove improperly constructed CGPs, such as those created from data obtained using stimulated cells (which was an exclusion criteria we specified, but nonetheless, compliance was less than perfect). Additionally, CGPs were more difficult to construct for the biologically more complex mouse models, and thus more were removed in the QC process. It is likely that early participant feedback on CGP quality during the Jamboree would help ensure higher quality CGPs, thereby reducing some of the required post-Jamboree QC. This also suggests that extending the Jamboree to two days, for example, with another day to review and QC the CGPs by the participants, could be valuable.

Following QC, meta-analysis performed within OMiCC led to several interesting observations: firstly, evaluation of DE genes showed substantial signature overlaps among diseases within species, and to a lesser extent, between the two species. Secondly, these findings were largely consistent when evaluated using an effect-size based approach. However, caution needs to be exercised in interpreting the results as the identification of DE genes can be influenced by a number of variables that cannot be controlled in this type of analysis. For example, as more CGPs from independent studies using different platforms are included in the analysis, the number of common genes among the platforms typically decreases, thus reducing the number of genes for which differential expression can be evaluated. Meta-analysis of CGPs containing overlapping samples can also give a false sense of robustness because the true PFP (or FDR) can be higher than what is reported. Other potential confounding factors include unequal distributions of age and race (or strain for mice) between sample groups within CGPs. However, these can also increase the heterogeneity across CGPs, so any conserved signals that emerge from the meta-analyses of the CGPs are likely relatively robust¹⁶. Barring differences in meta-analysis methodologies, our analysis identified a larger number of pan-disease DE genes in human compared to an earlier, similar meta-analysis effort⁸ (1021 versus 210 UP and 976 versus 202 DOWN genes), likely in part because our analysis included a larger number of CGPs/studies curated by the Jamboree participants. This highlights the potential benefit of using crowdsourcing to amass a large multi-study dataset within a relative short amount of time.

Using tools outside of OMiCC, gene set/pathway enrichment analysis revealed that, as expected, a higher level of conservation across diseases than that at the gene level. Some of the enriched KEGG and Reactome terms were consistent with previous reports, e.g., “cytokine signaling” was enriched in genes with increased expression in human SLE. It is well-established that SLE patients exhibit increased expression of IFN-inducible genes in blood compared to healthy controls^{17,18}. The term “cytokine signaling”

was also enriched (albeit to a lesser magnitude) in RA, as well as in the human and mouse pan-disease signatures, and it was furthermore conserved between human and mouse; these results are again consistent with previous reports^{8,19–21}. Our pathway enrichment analysis also identified some less well-established, but potentially biologically interesting associations. For example, the KEGG term “Malaria” is enriched in the UP genes in RA due to genes such as CR1, GYPA, ICAM1, PECAM1, and TLR4. It is not clear whether this is related to the fact that anti-malarial drugs, such as hydroxychloroquine, have been used as a secondary treatment for RA for many years²², and it has been suggested that hydroxychloroquine interferes with Toll-like receptor signaling²³ to reduce immune cell activation and proliferation, although its exact mechanism of action in ameliorating RA is still not well understood. Another potentially interesting observation is the enrichment of platelet-related pathways in a number of signatures. The involvement of platelets has been implicated in various autoimmune diseases²⁴, particularly in RA²⁵, and has been proposed as a potential therapeutic target for some of the autoimmune diseases assessed here²⁶.

In reflection, there are several ways in which our Jamboree could have been improved, such as offering more extensive training using OMiCC prior to data exploration, providing early feedback on the construction of CGPs, and creating independent discovery and validation cohorts to strengthen the robustness of our preliminary observations. Despite some of the caveats associated with our analyses and results, overall we provided evidence that user-friendly crowdsourcing and analysis platforms, such as OMiCC, can potentially accelerate the pace by which public data can be utilized to generate and test hypotheses.

Data and software availability

Gene expression and sample group data

The comparison group pairs (CGPs, e.g., RA versus healthy) created by the Jamboree participants and used in the post-Jamboree analyses have been made public in OMiCC at: <https://omicc.niaid.nih.gov/>. They are collected in compendia whose names have the format 2016-NIH-Jamboree-Species-Disease (species can either be Human or Mouse while diseases include DM1, MS, RA, SLE, and Sarcoid). These compendia can be retrieved in OMiCC by using the compendia search function (on OMiCC homepage: Search > On Compendia) and searching for the keyword ‘2016-NIH-Jamboree’. This information can also be retrieved from **Dataset 1** listed below.

To retrieve the raw microarray data, a user can construct new compendia using selected CGPs from the Jamboree compendia collection (see the Community and Sharing Features section of the OMiCC Tutorial) and export the gene expression data from the web site.

Meta-analyses and gene enrichment analyses data

F1000Research: Dataset 1. R data file, [10.5256/f1000research.10465.d146994](https://doi.org/10.5256/f1000research.10465.d146994)²⁷

F1000Research: Dataset 2. R markdown script to generate the data analysis report, [10.5256/f1000research.10465.d146995](https://doi.org/10.5256/f1000research.10465.d146995)²⁸

F1000Research: Dataset 3. Meta-analysis output files exported from OMiCC, [10.5256/f1000research.10465.d146996](https://doi.org/10.5256/f1000research.10465.d146996)²⁹

F1000Research: Dataset 4. Results of Toppgene analyses against KEGG, Reactome, and Gene Ontology (GO) Biological Process terms using the DE genes listed in [Table S1](#) as input, [10.5256/f1000research.10465.d146997](https://doi.org/10.5256/f1000research.10465.d146997)³⁰

Author contributions

WWL helped design the Jamboree, performed post-Jamboree data curation, designed and performed post-Jamboree data analysis, and wrote the manuscript; RS designed and organized the Jamboree, performed post-Jamboree data curation, and wrote the manuscript; OJWG participated in the Jamboree; JST conceived and guided the project, designed and helped organize the Jamboree, helped design

post-Jamboree data analysis plan, helped post-Jamboree data curation, and wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This research was funded by the Intramural Programs of the National Institute of Allergy and Infectious Diseases (NIAID) and the Center for Information Technology (CIT) at the National Institutes of Health.

Acknowledgements

We thank BCBB/OCICB of NIAID for providing computing support and web hosting; NIH Facilities for providing the OMiCC Jamboree hosting venue; and members of the J.S.T. lab for discussions.

Consortium/Collective Authors

The OMiCC Jamboree Working Group

(Listed alphabetically by last name)

James Austin¹, Neha Bansal¹, Julián Candia², Ehren Dancy¹, Karen L. Elkins³, Sara Faghihi-Kashani⁴, Julio Gomez-Rodriguez⁵, Liliana Guedez⁶, Yongjian Guo¹, Maria J. Gutierrez⁷, Trung Ho⁸, Reiko Horai⁶, Sunmee Huh⁹, Chie Iwamura¹⁰, Jaimy Joy¹¹, Ju-Gyeong Kang¹², Sunil Kaul⁹, Laura B. Lewandowski¹³, Candace Liu¹, Yong Lu¹, Nathan P. Manes¹, Mary J. Mattapallil⁶, Sarfraz Memon⁹, M. Jubayer Rahman¹⁰, Kameron B. Rodrigues¹⁰, Bruno Silva¹¹, Amit Singh¹¹, Anthony J. St. Leger⁶, Jessica Tang¹², Abigail Thorpe¹, Hang Xie³, Yongge Zhao⁹, Ofer Zimmerman¹

1. National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Bethesda, MD, USA, 20892
2. Trans-NIH Center for Human Immunology, NIH
3. Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA, 20993
4. National Institute of Environmental Health Sciences, NIH
5. National Human Genome Research Institute, NIH
6. National Eye Institute, NIH
7. Johns Hopkins University School of Medicine, Baltimore, MD, USA, 21287
8. Uniformed Services University of Health Sciences, Bethesda, MD, USA, 20814
9. National Cancer Institute, NIH
10. National Institute of Diabetes and Digestive and Kidney Diseases, NIH
11. National Institute on Aging, NIH
12. National Heart, Lung and Blood Institute, NIH
13. National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH

Supplementary material

Figure S1. Meta-analysis results are largely consistent between the rank-based and effect-size-based meta-analysis methods.

For each dataset, bar-code plots are used to illustrate the positions of the (A) UP and (B) DOWN genes identified by the rank-based meta-analysis method in the gene list sorted by the median rank (across CGPs) of the effect size indicated by the Cohen's d statistic in each CGP. The gene lists are sorted in ascending order with higher statistics indicating larger effect sizes. The red and blue areas represent genes with positive and negative median Cohen's d values, respectively.

[Click here to access the data.](#)

Figure S2. Top 20 enriched biological processes for individual diseases.

Over-representation analyses of the (A) UP genes and (B) DOWN genes (PFP ≤ 0.05) identified by a rank-based meta-analysis method (RankProd within OMiCC) were carried out for each disease individually from human and mouse against KEGG and Reactome terms. The analyses were performed in both R and ToppGene; the top 20 enriched terms identified in R are shown. Terms found also in ToppGene are indicated by an asterisk (*). P-values are adjusted by Benjamini and Hochberg (BH) FDR correction. Counts (indicated by circle size) and gene ratios (x-axis) respectively denote the number and proportion of genes in the UP signature that also appear in the target gene set.

[Click here to access the data.](#)

Table S1. Differentially expressed genes in each dataset.

A spreadsheet listing all of the genes identified by RankProd (in OMiCC) to exhibit increased (UP) and decreased (DOWN) expression (PFP ≤ 0.05) in disease versus control comparisons in each of the datasets.

[Click here to access the data.](#)

Table S2. Enriched pathways for the differentially expressed genes in each dataset.

A spreadsheet listing all of the enriched KEGG and Reactome terms (p.adj ≤ 0.05 ; adjusted by Benjamini and Hochberg correction) derived from UP and DOWN genes identified in each dataset. Terms also found by ToppGene analysis are indicated by an asterisk (*).

[Click here to access the data.](#)

References

- Rung J, Brazma A: **Reuse of public genome-wide gene expression data.** *Nat Rev Genet.* 2013; **14**(2): 89–99.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Saez-Rodriguez J, Costello JC, Friend SH, et al.: **Crowdsourcing biomedical research: leveraging communities as innovation engines.** *Nat Rev Genet.* 2016; **17**(8): 470–86.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Celi LA, Lokhandwala S, Montgomery R, et al.: **Datathons and Software to Promote Reproducible Research.** *J Med Internet Res.* 2016; **18**(8): e230.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shah N, Guo Y, Wendelsdorf KV, et al.: **A crowdsourcing approach for reusing and meta-analyzing gene expression data.** *Nat Biotechnol.* 2016; **34**(8): 803–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sparks R, Lau WW, Tsang JS: **Expanding the immunology toolbox: embracing public-data reuse and crowdsourcing.** *Immunity.* 2016.
[Publisher Full Text](#)
- Seok J, Warren HS, Cuenca AG, et al.: **Genomic responses in mouse models poorly mimic human inflammatory diseases.** *Proc Natl Acad Sci U S A.* 2013; **110**(9): 3507–12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Webb DR: **Animal models of human disease: inflammation.** *Biochem Pharmacol.* 2014; **87**(1): 121–30.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Toro-Domínguez D, Carmona-Sáez P, Alarcón-Riquelme ME: **Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis.** *Arthritis Res Ther.* 2014; **16**(6): 489.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Godec J, Tan Y, Liberzon A, et al.: **Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation.** *Immunity.* 2016; **44**(1): 194–206.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Takao K, Miyakawa T: **Genomic responses in mouse models greatly mimic human inflammatory diseases.** *Proc Natl Acad Sci U S A.* 2015; **112**(4): 1167–72.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hong F, Breitling R, McEntee CW, et al.: **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.** *Bioinformatics.* 2006; **22**(22): 2825–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Yu G, Wang LG, Han Y, et al.: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *OMICS.* 2012; **16**(5): 284–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yu G, He QY: **ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.** *Mol Biosyst.* 2016; **12**(2): 477–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Chen J, Bardes EE, Aronow BJ, et al.: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res.* 2009; **37**(Web Server issue): W305–11.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tseng GC, Ghosh D, Feingold E: **Comprehensive literature review and statistical considerations for microarray meta-analysis.** *Nucleic Acids Res.* 2012; **40**(9): 3785–99.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sweeney TE, Haynes WA, Vallania F, et al.: **Methods to increase reproducibility in differential gene expression via meta-analysis.** *Nucleic Acids Res.* 2016; pii: gkw797.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Baechler EC, Battilwalla FM, Karypis G, et al.: **Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus.** *Proc Natl Acad Sci U S A.* 2003; **100**(5): 2610–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bennett L, Palucka AK, Arce E, et al.: **Interferon and granulopoiesis signatures in systemic lupus erythematosus blood.** *J Exp Med.* 2003; **197**(6): 711–23.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Higgs BW, Liu Z, White B, et al.: **Patients with systemic lupus erythematosus, myositis, rheumatoid arthritis and scleroderma share activation of a common type I interferon pathway.** *Ann Rheum Dis.* 2011; **70**(11): 2029–36.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Liu Z, Bethunaickan R, Huang W, et al.: **Interferon- α accelerates murine systemic lupus erythematosus in a T cell-dependent manner.** *Arthritis Rheum.* 2011;

- 63(1): 219–29.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Tsokos GC, Lo MS, Reis PC, *et al.*: **New insights into the immunopathogenesis of systemic lupus erythematosus.** *Nat Rev Rheumatol.* 2016; **12**(12): 716–30.
[PubMed Abstract](#) | [Publisher Full Text](#)
 22. van Vollenhoven RF: **Treatment of rheumatoid arthritis: state of the art 2009.** *Nat Rev Rheumatol.* 2009; **5**(10): 531–41.
[PubMed Abstract](#) | [Publisher Full Text](#)
 23. Kyburz D, Brentano F, Gay S: **Mode of action of hydroxychloroquine in RA-evidence of an inhibitory effect on toll-like receptor signaling.** *Nat Clin Pract Rheumatol.* 2006; **2**(9): 458–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
 24. Habets KL, Huizinga TW, Toes RE: **Platelets and autoimmunity.** *Eur J Clin Invest.* 2013; **43**(7): 746–57.
[PubMed Abstract](#) | [Publisher Full Text](#)
 25. Boilard E, Nigrovic PA, Larabee K, *et al.*: **Platelets amplify inflammation in arthritis via collagen-dependent microparticle production.** *Science.* 2010; **327**(5965): 580–3.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 26. Boilard E, Blanco P, Nigrovic PA: **Platelets: active players in the pathogenesis of arthritis and SLE.** *Nat Rev Rheumatol.* 2012; **8**(9): 534–42.
[PubMed Abstract](#) | [Publisher Full Text](#)
 27. Lau WW, Sparks R; OMiCC Jamboree Working Group, *et al.*: **Dataset 1 in: Meta-analysis of crowdsourced data compendia suggests pan-disease transcriptional signatures of autoimmunity.** *F1000Research.* 2016.
[Data Source](#)
 28. Lau WW, Sparks R; OMiCC Jamboree Working Group, *et al.*: **Dataset 2 in: Meta-analysis of crowdsourced data compendia suggests pan-disease transcriptional signatures of autoimmunity.** *F1000Research.* 2016.
[Data Source](#)
 29. Lau WW, Sparks R; OMiCC Jamboree Working Group, *et al.*: **Dataset 3 in: Meta-analysis of crowdsourced data compendia suggests pan-disease transcriptional signatures of autoimmunity.** *F1000Research.* 2016.
[Data Source](#)
 30. Lau WW, Sparks R; OMiCC Jamboree Working Group, *et al.*: **Dataset 4 in: Meta-analysis of crowdsourced data compendia suggests pan-disease transcriptional signatures of autoimmunity.** *F1000Research.* 2016.
[Data Source](#)

Open Peer Review

Current Referee Status:  

Version 1

Referee Report 18 April 2017

doi:10.5256/f1000research.11275.r21277



Markus Riester 

Novartis Institutes for BioMedical Research, Cambridge, MA, USA

This paper describes a thorough case study using the author's recently published OMiCC web service. This service provides re-processed expression data and allows the curation and selection of datasets by disease experts without requiring bioinformatic expertise. Performing gene expression meta-analyses is challenging and time consuming for precisely the reasons this tool addresses and tools like OMiCC are therefore a welcome addition to the field.

The paper is clearly written and both design and implementation are in general solid.

A shortcoming of the design is that the curation teams were all assigned different tasks. It would have been interesting to see the overlap of curations obtained by independent teams.

In addition, I have a few minor comments and optional suggestions regarding the analyses:

1. A brief literature review of existing solutions (for example InsilicoDB) appears to be missing in both this manuscript and the main paper.
2. A challenge of comparing array data from different platforms is that some genes might be captured with varying quality across platforms. It is unclear what was done to identify problematic probe sets or genes. Various R packages (e.g. metaArray) for example calculate Integrative Correlation scores. These scores identify probe sets which behave differently across platforms in terms of co-expressed genes.
3. Another challenge is the extensive reuse of specimens and data in public datasets. The authors write that duplicates were identified and removed. As a completely optional suggestion, we recently published the doppelgangR package that automates the identification of duplicates.
4. It is unclear if the software can generate more classical meta-analysis visualizations like forest plots.
5. The number of different platforms included in the meta-analysis and whether platform was a significant source of heterogeneity could be made clearer.
6. I probably would have performed the gene set analysis using expression data collapsed to pathways, for example by GSVA, ssGSEA or related newer methods. These methods turn a gene-by-sample matrix into a pathway-by-sample matrix; the same gene-centric methods can be

then applied to pathways. I am not aware of any existing literature comparing pathway meta-analysis methods and this is thus another optional comment. This might however be a cleaner approach than pooling the mouse and human datasets.

7. Axis and legend labels sometimes use R variable names (such as "gene.ratio") instead of proper annotation (using `xlab()`, `scale_fill_discrete()` etc.)
8. fRMA is in theory better for meta-analyses compared to standard RMA since then all datasets use the same reference pool for normalization. I am however again not aware of a systematic comparison and the impact on meta-analyses.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 16 January 2017

doi:[10.5256/f1000research.11275.r19356](https://doi.org/10.5256/f1000research.11275.r19356)



Hans Lehrach¹ , **Hans-Jörg Warnatz²**, **Thomas Risch²**

¹ Department Vertebrate Genomics, Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany

² Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany

Lau and colleagues describe an interesting effort of a group of biologists without formal bioinformatics training to use a programming-free web-based platform to generate a sizable number of comparison group pairs (CGPs) during a day-long group exercise, using gene expression data from humans and mouse models. The subsequent gene and gene set enrichment analyses – performed after quality control of the generated CGPs – yield reasonable results for a number of autoimmune diseases, resulting in

plausible enrichments identified for genes and gene sets associated with inflammation and immune processes.

The described effort is a potentially scalable method for analysis of very large data sets using the combined manpower of a large number of individuals. To produce more quantifiable data of this process, it would however be interesting to compare the results of duplicates (do individual groups working in isolation on the same question come up with the same or different results. How does the result of such a 1 day Jamboree compare with the results of a single expert working for a month? If you would rerun the exercise, how different would you expect the results to be?

It might also be interesting to systematically eliminate one dataset at a time to quantitate its influence on the final result.

One major aspect of the study worth more detailed reporting is the way quality controls are carried out on the CGPs collected by the crowd. This aspect will become even more important when a large crowd is used, and more CGPs are collected, and constitutes one of the main pillars of all subsequent analyses. Therefore, I would suggest that the authors report in more detail their strategies and the conduction of the quality controls, along also with more details on potential caveats and pitfalls.

Other comments:

Methods

Major comment:

- "we considered gene sets and pathways with an adjusted p-value of ≤ 0.05 to be significantly enriched"
Here it is not clear if p-values were adjusted for multiple testing, e.g. using Benjamini-Hochberg correction.
- A more detailed description should be given for the ToppGene analysis

Minor comment:

- Toppgene is actually spelled "ToppGene"

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
