A comprehensive and centralized database for exploring omics data in Autoimmune Diseases

Jordi Martorell-Marugán [1,2], Raúl López-Domínguez [1], Adrián García-Moreno [1], Daniel Toro-Domínguez [1,3], Juan Antonio Villatoro-García [1,4], Guillermo Barturen [3], Adoración Martín-Gómez [5], Kevin Troule [6], Gonzalo Gómez-López [6], Fátima Al-Shahrour [6], Víctor González-Rumayor [2], María Peña-Chilet [7,8], Joaquín Dopazo [7,8,9], Julio Sáez-Rodríguez [10,11,12], Marta E. Alarcón-Riquelme [3,13] and Pedro Carmona-Sáez [1,4,*]

[1] Bioinformatics Unit. GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS Granada, 18016, Granada, Spain. [2] Atrys Health S.A., Barcelona, Spain. [3] Genetics of Complex Diseases. GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS Granada, 18016, Granada, Spain. [4] Department of Statistics. University of Granada, 18071, Granada, Spain. [5] Nephrology Units. AADEA: Asociación Andaluza de Enfermedades Autoinmunes. Hospital de Poniente, 04700. Almería, Spain. [6] Bioinformatics Unit. Spanish National Cancer Center, CNIO, Madrid, Spain. [7] Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), CDCA, Hospital Virgen del Rocío, 41013, Sevilla, Spain. [8] Bioinformatics in Rare Diseases (BiER). Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), FPS, Hospital Virgen del Rocio. 41013. Sevilla, Spain. [9] INB-ELIXIR-es, FPS, Hospital Virgen del Rocío, Sevilla, 42013, Spain. [10] Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Faculty of Medicine, 52074 Aachen, Germany. [11] European Molecular Biology Laboratory-The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [12] Institute for Computational Biomedicine, Heidelberg University Hospital and Heidelberg University, Faculty of Medicine, Bioquant Heidelberg, Hedelberg 69120, Germany. [13] Unit of Chronic Inflammatory Diseases, Institute of Environmental Medicine, Karolinska Institutet, 17177, Stockholm, Sweden.

*corresponding author: Pedro Carmona-Sáez (pedro.carmona@genyo.es)

Abstract

Autoimmune diseases are heterogeneous pathologies with difficult diagnosis and few therapeutic options. In the last decade, several omics studies have provided significant insights into the molecular mechanisms of these diseases. Nevertheless, data from different cohorts and pathologies are stored independently in public repositories and a unified resource is imperative to assist researchers in this field. Here, we present ADEx (https://adex.genyo.es), a database that integrates 82 curated transcriptomics and methylation studies covering 5609 samples for some of the most common autoimmune diseases. The database provides, in an easy-to-use environment, advanced data analysis and statistical methods for exploring omics datasets, including meta-analysis, differential expression or pathway analysis.

Keywords: Autoimmune disease, database, GEO, transcriptomics, epigenomics, curation, dataset, interferon signature, gene expression, meta-analysis

Background

Autoimmune diseases (ADs) are a group of complex and heterogeneous disorders characterized by immune responses to self-antigens leading to tissue damage and dysfunction in several organs. The pathogenesis of ADs is not fully understood, but both environmental and genetic factors have been linked to their development [1]. Although these disorders cause damage to different organs and their clinical outcomes vary between them, they share many risk factors and molecular mechanisms [2]. Some examples of ADs are systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), Sjögren's syndrome (SjS), systemic sclerosis (SSc), considered systemic autoimmune diseases (SADs) and type 1 diabetes (T1D), which is considered an organ-specific autoimmune disease. Most of these diseases are classified as rare given their prevalence, but altogether ADs affect up to 3 % of the population considering conservative estimates [3].

In ADs patients, the pathology is developed during several years but it is only detected when tissue damage is significant. For that reason, early diagnosis is important and complicated. Additionally, some ADs often show a non-linear outcome that alternates between active and remission stages thus making their study even more difficult. Despite huge efforts have been made to develop ADs biomarkers and therapies, these do not fit for every patient and their clinical responses differ greatly [4].

During the past decade, the use of omics technologies has provided new insights into the molecular mechanisms associated with the development of ADs, opening new scenarios for biomarkers and treatments discovery [5]. In this context, it is remarkable the characterization of the type I interferon (IFN) gene expression signature as a key factor in the pathology of some SADs, especially in SLE and SjS [6], which has improved our knowledge of the underlying molecular mechanisms and has opened new therapeutic strategies based on blocking the pathways related to this signature.

Regardless of the large amount of omics studies describing new biomarkers and therapeutic strategies in ADs [7–10], in most cases these biomarkers are not consistent across different studies or have not fully accomplished their diagnostic goals. Indeed, the widely studied IFN signature is highly variable between patients [11] and it is associated with differences in response to treatments which target it, as has been reported for example in the phase-II results of Sifalimumab clinical trial for SLE patients [12]. In

addition, in most of the cases, biomarkers are defined from the analysis of a single type of omic data (commonly gene expression), but multi-omics data integration can provide a more complete understanding of molecular mechanisms and more robust and biologically relevant biomarkers.

Most of the omics datasets generated from different cohorts and studies in ADs published to date have been deposited and are available in public repositories such as Gene Expression Omnibus (GEO) [13] or ArrayExpress [14]. Although all these valuable data can be used in retrospective analyses in order to generate new knowledge and accelerate drug discovery and diagnosis, it is not easy to compare neither to integrate available data because they are generated from different platforms and/or processed with different analytic pipelines. In this context, there are great efforts from the bioinformatics community to develop standardized data analysis workflows and resources that facilitate data integration and reproducible analysis. For example, Lachmann et al. [15] have recently reprocessed a large collection of raw human and mouse RNA-Seq data from GEO and Sequence Read Archive (SRA) using a unified pipeline and they have developed the ARCHS4 as a resource to provide direct access to these data through a web-based user interface. Other singular projects such as The Cancer Genome Atlas (TCGA) [16] or the Genotype-Tissue Expression project (GTEx) [17] provide also large and homogeneously processed datasets for tumor samples and human tissues respectively. These unprecedented resources motivate the development of applications and data portals to help researchers gather information with the aim of improving diagnosis and treatment in multiple diseases, most notably in cancer research, where such information is actually being used in the clinical practice [18].

Despite such enormous potential, in the context of ADs there is a lack of a centralized and dedicated resource that facilitates the exploration, comparison and integration of available omics datasets. This is indeed an area in which this type of application would be tremendously beneficial, given that the low prevalence of each individual disease makes difficult the recruitment of large patients cohorts [4].

To bridge this gap, in this work we have compiled and curated most of the publicly available gene expression and methylation datasets for five ADs: SLE, RA, SjS, SSc and T1D. To this end, we have developed and applied homogeneous pipelines from raw data and we developed ADEx (Autoimmune Disease Explorer), a data portal where these

processed data can be downloaded and exploited through multiple exploratory and statistical analyses. ADEx facilitates data integration and analysis to potentially improve diagnosis and treatment of ADs.

In order to demonstrate the potential, we queried the database to explore the expression pattern of IFN regulated genes across all autoimmune diseases. This analysis revealed that the IFN signature is consistent in SLE and SjS but it shows heterogeneity in RA samples. In a second analysis, we integrated all datasets in order to define a set of consistent biomarkers for each disease considering the expression data from multiple studies.

Construction and content

We have prepared five different pipelines to process data for each platform (RNA-Seq, Affymetrix and Illumina gene expression microarrays, and Illumina methylation arrays 27K and 450K). All these workflows are written in R language and are publicly available in GENyO Bioinformatics Unit GitHub (https://github.com/GENyO-BioInformatics/ADEx_public). Figure 1 contains an overview of the different steps performed to prepare the data for ADEx application.

Data collection

Collection of the datasets included in ADEx was carried out by searching in the GEO web page with ADs names as key terms. We filtered the results by study type (expression profiling by array, expression profiling by high throughput sequencing and methylation profiling by array), organism (Homo sapiens) and platform manufacturer (Affymetrix or Illumina).

We downloaded the metadata for these initial datasets with GEOquery [19] R package in order to apply our inclusion criteria and exclude those studies and samples that do not meet them. We only included case-control studies from samples, which were not treated with drugs in vitro. Exclusively datasets with available raw data were considered. Studies whose controls and cases belong to different tissues were discarded. We only selected datasets with 10 samples at least. We divided the datasets containing samples from

different diseases, platforms, tissues or cell types in subgroups so that these are constant and avoid possible batch effects.

82 datasets containing 5609 samples passed our filtering criteria (see Additional file 1 for complete information about all included datasets). Then we downloaded their raw data with GEOquery [19]. For expression microarrays, we downloaded CEL files and raw text files for Affymetrix and Illumina platforms respectively. For RNA-Seq, we downloaded the fastq files from the European Nucleotide Archive. For methylation microarrays, we downloaded raw methylation tables if they were available and idat files otherwise.

Metadata curation

GEO does not require submitters to use either a fixed structure or standard vocabulary to describe the samples of an experiment. For that reason, it was necessary to manually homogenize the information provided within all the selected datasets using standardized terms. There are some methods for automatic curation of GEO metadata, but manual curation is still necessary to get high-quality metadata [20]. This metadata curation was an essential step for the following analyses and permits an easy datasets information exploration.

Platforms curation

We have used a total of 12 different gene expression platforms from microarray and RNA-Seq technologies. Microarray platforms quantify expression levels in probes. In order to match probe identifiers to gene names, platforms annotation files are available from GEO. However, we found that some of these annotation files match probes to inappropriate gene names. On the one hand, some platforms save gene names with errors due to the conversion of gene names such as MARCH1 or SEPT1 into dates, a common error that has been reported previously [21]. In these cases, we fixed manually these genes in the annotation files. On the other hand, some platforms use obsolete or different aliases to refer to the same genes. We used human genes' information from NCBI repository in order to match aliases with actual official gene symbols and substituted them in the platform annotations.

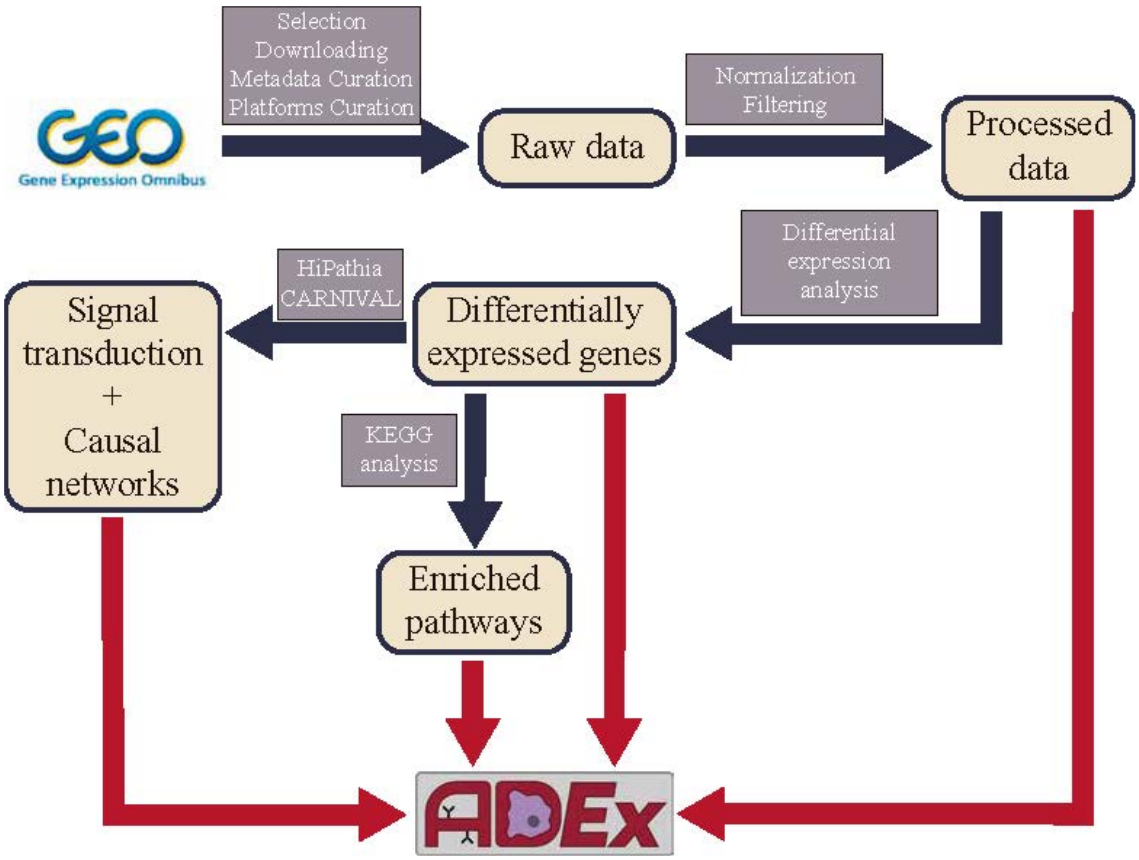

Figure 1. Processing pipeline for ADEx data. Black arrows indicate intermediate processing steps. Red arrows indicate the inputs to ADEx application.

Data processing

Raw data from Illumina expression microarrays were loaded by reading the plain text files. In order to remove background noise, we kept only the probes that had a Detection P-value lower than 0.05 in 10 % of the samples. Then we performed a background correction and quantile normalization [22] using neqc function from limma package [23].

CEL files from Affymetrix expression microarrays platforms were loaded to R environment with affy package [24]. To filter low intensity probes, we removed all probes with an intensity lower than 100 in at least 10 % of the samples. Normalization was carried out computing Robust Multichip Average (RMA) normalization [25] with affy package [24].

For RNA-Seq datasets, fastq files were aligned to human transcriptome reference hg38 using STAR 2.4 [26] and raw counts were obtained with RSEM v1.2.31[27] with default parameters. Raw counts were filtered using NOISeq R package [28], removing those features that have an average expression per condition lower than 0,5 counts per million

(CPM) and a coefficient of variation (CV) higher than 100 in all conditions. Counts normalization was carried out with TMM method [29].

We translated microarrays probes identifiers to gene symbols using our curated annotation tables. For those genes targeted by two or more microarray probes, we calculated the median expression values of all their targeting probes. For RNA-Seq, we translated ENSEMBL identifiers to gene symbols using biomaRt package [30, 31].

Methylation raw data are available in GEO as idat or text files depending on the dataset. Idat files were read with minfi package [32], while text files were read in the R environment. In both cases, poorly performing probes with a detection P-value above 0.05 in more than 10 % of samples were removed. Probes adjacent to SNPs, located in sexual chromosomes or reported to be cross-reactive [33] were also removed. We normalized the methylation signals using quantile normalization with lumi package [34]. Finally, for datasets generated with 450k platform, we applied BMIQ normalization [35] using wateRmelon package [36] in order to correct for the two types of probes contained in this platform.

Differential expression analysis

We performed a differential expression analysis in all datasets independently towards the identification of differential patterns among disease samples and healthy controls. These analyses were performed in different ways depending on the source of data. Gene expression profiles from microarray platforms were carried out by the standard pipeline of limma package [23]. We used lmFit function to fit a linear model to the gene expression values followed by the execution of a t-test by the empirical Bayes method for differential activity (eBayes function). On the other hand, gene expression profiles from RNA-Seq platforms were analyzed by the standard pipeline of DESeq2 package [37]. In both cases, differential expression analysis provided P-values, adjusted P-values by False Discovery Rate (FDR) and log2 Fold-Change (FC).

Pathway analysis

Pathway enrichment analysis was precomputed for each expression dataset using differential expression analysis results. We considered DEGs those genes with a FDR lower than 0.05 and we performed hypergeometric tests to check if each pathway contains

more DEGs as expected by chance. We used KEGGprofile 1.24.0 R package to perform this analysis but beforehand we manually updated its dependency, KEGG.db, the database used to perform the statistical test. The pathways were plotted using the KEGG mapper tool Search&Color Pathway, with the genes colored by their FC between case and control samples.

Signaling network analysis

We integrated signaling network analysis applying HiPathia software [38] to gene expression data so that changes in the activity of the network from different pathways can be detected. We precomputed this analysis for each gene expression dataset. Firstly, we translated the gene expression matrix and scaled it. Then, we calculated the transduction signal and compared among conditions, cases and controls.

Causal networks inference

We used the CARNIVAL [39] R package pipeline to analyze the causal networks architectures from gene expression data. For that aim, we followed the instructions published by their creators at https://github.com/saezlab/transcriptutorial. Briefly, differential expression analyses were performed with limma [23] and the results were used to calculate the transcription factor activities with DoRothEA [40] and the pathways activities with PROGENy [41]. These results were the input of CARNIVAL to calculate the upstream regulatory signaling pathways for each expression dataset. Finally, the results were stored in interactive html reports.

Database architecture

Pursuing an optimal data organization and quick access to all the data in ADEx, we have enabled an internal database with PostgreSQL. We chose this technology since it is open source and it is best suited to the huge dimensionality of omics datasets.

Webtool

ADEx user interface was designed with RStudio Shiny package. The application uses a set of external packages to perform analysis and graphics on demand. Most of the plots are generated with ggplot2 [42]. All the computations in the Meta-Analysis section are performed whenever users request them. Biomarkers analysis is performed with the Rank

Products algorithm integrated in RankProd R package [43]. The tool runs on our own server with CentOS 7.0 operating system, 16 processors and 32 Gb of RAM memory.

Utility and discussion

Data collection and processing

ADEx contains data from 5609 samples. We have processed 82 expression and methylation datasets from case-control studies for SLE, RA, SjS, SSc and T1D diseases (see Table 1 for a summary and Additional file 1 for complete information about all included datasets). We have manually curated all metadata in order to standardize the nomenclature of phenotypes, cell types, etc. from different studies and discard samples or datasets that do not meet the selection criteria (see Construction and content section). The processed datasets are available from the Download Data section in the application.

The ADEx application

ADEx data portal can be used to download and analyze the processed data. ADEx is freely available at https://adex.genyo.es. The tool is divided in 6 different sections arranged in different tabs (Figure 2a).

Table 1. Summary of accessible studies and samples by disease and data type in ADEx.

	Expression	**Methylation**	**Total**
Disease	**Datasets - Samples**	**Datasets - Samples**	**Datasets - Samples**
SLE	20 - 2053	13 - 628	33 - 2681
RA	17 - 1122	3 - 835	20 - 1957
SjS	9 - 400	1 - 29	10 - 429
SSc	5 - 229	1 - 37	6 - 266
T1D	11 - 176	2 - 100	13 - 276

Section 1: Data overview

Information about the available datasets can be found in both table or pie plot formats in this section. In tables, information about the sample phenotype and their data origin is provided. In pie plots quantitative information is provided regarding the clinical and phenotype information. All this information has been extracted from GEO or from the associated published articles whenever supplied. This information can be presented individually for each dataset or grouped by disease. While a single dataset is being explored, the experiment summary is shown. Users can use this section to identify datasets of their interest to be analyzed in the following sections.

Section 2: Gene Query

This section was created in order to explore the expression and methylation of a specific gene, or the correlation between them, within a single dataset. Users can explore the different gene expression values for each dataset comparing case and control samples with a boxplot. Meanwhile, methylation data is presented at CpG level, so that users can select a region of the gene (e.g. promoter) and the mean methylation value for cases and controls is plotted for every CpG probe contained in the selected region.

It has been demonstrated the strong relationship of gene expression and methylation levels [44]. That is why, in this section, users can also integrate both expression and methylation values to search for direct or inverse correlations. Finally, gene expression correlation analysis can be performed in order to get insight into the relationship between different genes and to find groups of coexpressed genes.

Section 3: Gene Set Query

Here users can select several datasets and genes in order to explore the FC between patients and controls across studies. All datasets from a disease can be automatically selected by clicking the right buttons, or individual studies can be selected by clicking directly on the table. Users can introduce a list of genes to explore their expression, although there are several preloaded gene lists covering the coexpression modules reported by Chaussabel et al. [45]. These modules consist of sets of coexpressed genes among hundreds of samples from different diseases. Each transcriptional module is associated with different pathways and cell types, most of them related to the immune

system [45]. See our use case 1 for an example of this type of analysis (Figures 2b and 2c).

Section 4: Analyze Dataset

In this section, we focus the analysis on whole datasets instead of individual genes. By default, a heatmap with the expression of the top 50 differentially expressed genes (DEGs) sorted by FDR is displayed. It is also possible to sort them by FC and cutoffs can be applied to both statistics. Additionally, differential expression analysis results can be downloaded as an excel table.

Furthermore, users can also study the KEGG [46] enriched pathways associated with the dataset selected. These results are precomputed using all the DEGs that have an FDR value below 0.05. A table gathers the significantly enriched KEGG pathways along with their associated hypergeometric test statistics and an interactive plot shows detailed information of the participant genes in the pathway colored according to their FC.

Beyond conventional pathway enrichment methods, we have implemented more sophisticated mechanistic models of cell signaling activity which have demonstrated to be very sensitive in deciphering disease mechanisms [38, 47] as well as the mechanisms of action of drugs [48, 49]. To offer this functionality we have applied HiPathia software [38] to gene expression data. This method estimates changes in the activity of signaling circuits defined into different pathways. With this approach, it becomes possible to study in detail the specific signaling circuits altered in ADs within the different signaling pathways. We precomputed this analysis for each dataset and the results are available as tables and interactive reports.

Finally, in this section the results of causal pathways analyses are available. We used CARNIVAL [39] software to construct the network topologies from the gene expression datasets in order to identify upstream alterations propagated through signaling networks in autoimmune diseases.

Section 5: Meta-Analysis

ADEx also implements meta-analysis functionalities based on gene expression data to integrate and jointly analyze different and heterogeneous datasets. We implemented a meta-analysis approach to search for biomarkers and common gene signatures across

different datasets from the same or different pathologies [50] based on the FCs of each dataset and gene. Datasets have to be selected similarly to Section 3 to launch the meta-analysis. See our use case 2 for examples of this type of analysis (Figure 3).

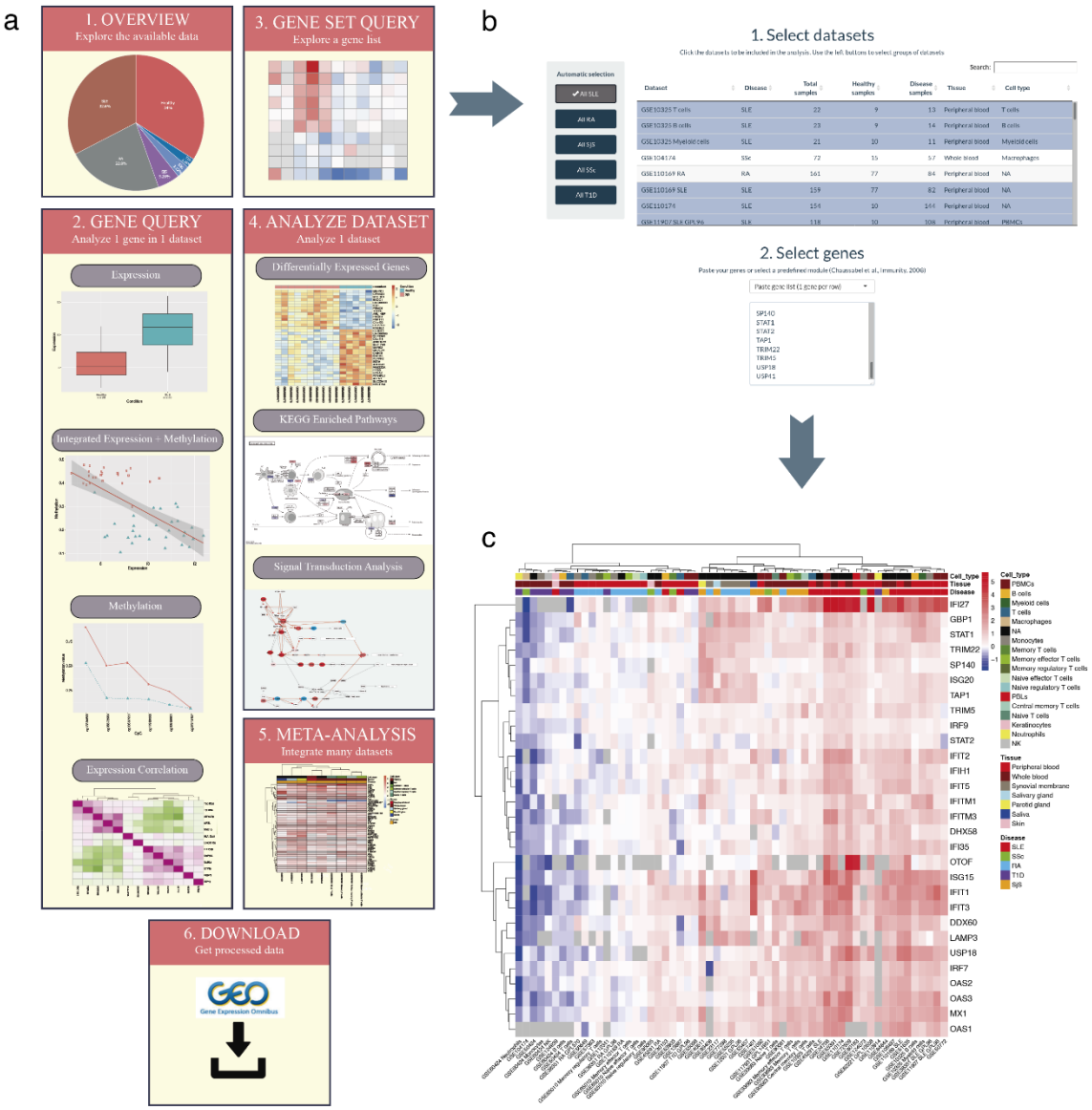

Figure 2. Overview of ADEx application and analysis of IFN signature across diseases. a) ADEx has six main sections. Section 1 provides information about available datasets. In section 2, users can explore expression and methylation for individual genes. Section 3 implements a module to explore data for a gene list, such as gene module or genes from a biological pathway, across several datasets. Section 4 allows researchers to perform analysis on individual datasets retrieving differential expression signatures and pathways and cell signaling enrichment analyses. Section 5 implements meta-analysis methods to integrate multiple datasets in order to define common biomarkers. Section 6 is for data download. **b) Gene Set Query section screenshot.** Datasets and gene set input is shown. Users select data there to plot a heatmap. **c) IFN signature expression generally separates SLE and SjS from other ADs.** Heatmap with the IFN genes generated in ADEx. Color represents the log2 FC of disease versus healthy samples (red for overexpression and blue for underexpression).

Section 6: Download data

In this section, users can select one or several datasets and download them. Curated data is obtained with the aim of performing additional analyses externally to ADEx application.

Use case 1: Exploring the IFN signature across diseases

Using as a query a set of genes (a gene expression signature, genes from the same pathway, etc.), it becomes straightforward to explore how the signature is expressed across different datasets or diseases. In order to demonstrate the potential of ADEx, we explored the IFN signature expression status in different diseases given its importance in the autoimmune disorders [11]. To address this goal, we evaluated the expression level across all datasets of IFN signature previously defined [51] (Figure 2b). We observed that IFN signature is strongly overexpressed in SLE and SjS patients (Figure 2c), as previously described [52, 53]. These two diseases are clearly separated from the other pathologies based on these IFN-regulated modules. RA IFN signature is highly heterogeneous, which is coherent with previous studies [54]. Interestingly, IFN modules are overexpressed in most of the RA studies that used synovial membrane tissue, while this effect is absent or very subtle in most of the RA blood studies. This is expected because the primary inflammation sites in this disease are the synovial joints [55].

Use case 2: Biomarker discovery in ADs

To show the functionality of ADEx for biomarker discovery, we also performed a disease-centered meta-analysis with all the datasets included in the database in order to define candidate biomarkers for each disease. We removed those genes with NA values in more than 75 % of the samples and we used RankProd package [43] to calculate the Rank Product statistics and the adjusted P-value. We considered significant those genes with adjusted P-value < 0.05. Since there are datasets from different cell types, tissues, platforms and so on, our aim was to find global biomarkers independently of all those variables. We discovered 1703 consistently deregulated genes in SLE, 367 in SjS, 743 in RA, 45 in SSc and 294 in T1D (Figure 3 and Additional file 2). We used the information from Interferome database [56] to annotate each gene depending on how each type of IFN affects its expression (upregulation or downregulation). For that aim, we queried the Interferome database, searching for genes with an absolute log2 FC > 2 after IFN addition.

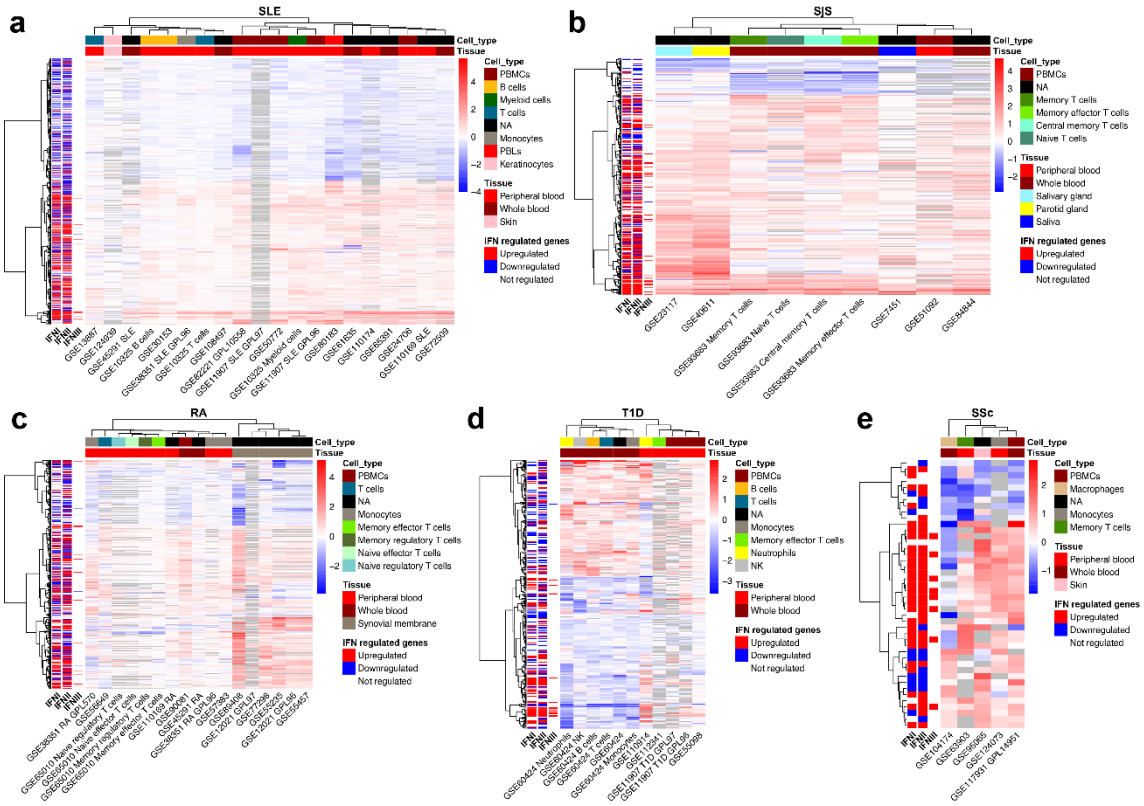

Figure 3. Integration of multiple datasets reveal candidate biomarkers for each disease. The observed effect of IFN I, II and III on gene expression is annotated at the left of each heatmap. Color represents the log2 FC. Heatmaps contains the significant biomarkers for **a)** SLE, **b)** SjS, **c)** RA, **d)** T1D and **e)** SSc.

Given that this database contains different experimental conditions, we averaged the log2 FC and considered as genes upregulated by IFN those with an average log2 FC > 0 and as downregulated those with an average log2 FC < 0. As can be observed in Figure 3, most of SLE, SjS and RA biomarkers are expressed according to the observed IFN effect on them, supporting the major role of IFN action in these diseases. It is notable the contribution of type II IFN (IFN II) to the observed expression changes. IFN II role in ADs is frequently underestimated in favour of type I IFN (IFN I) and, in fact, IFN signature definitions commonly focus on genes regulated by IFN I [6, 10, 52]. However, it has been demonstrated that Type II IFN has a key role in ADs pathogenesis [57]. Our findings support such importance and the need to focus the attention on IFN II regulation pathways to design new therapeutic strategies.

In RA, the strongest biomarker signals come from synovial tissue studies, and these datasets are perfectly separated from the blood studies. This is coherent with the IFN signature expression results (Figure 2c).

Conclusions

Despite that the heterogeneity of ADs is evident, there are common molecular mechanisms involved in the activation of immune responses. In this context, integrative analyses of multiple studies are crucial to discover shared and differential molecular signatures [58]. Nowadays there are many ADs datasets publicly available, but a strong computational knowledge is necessary in order to analyze them properly. With the aim of filling this gap between experimental research and computational biology, interactive easy-to-use software are valuable tools to perform exploratory and statistical analysis without strong computational expertise. This type of tool has been developed for other diseases and has helped to reuse public data and generate new knowledge and hypotheses [59–61].

A resource of this type is urged in the field of ADs to: 1) Compile available ADs' public data in a single data portal, 2) Access to integrable data processed with uniform pipelines, and 3) Perform both individual and integrated analysis interactively. We developed ADEx database to accomplish all those objectives. Then, we used ADEx data and functions to illustrate our tool potential exploring the IFN signature in different diseases and revealing genes consistently over- and underexpressed which could be good biomarkers for these diseases.

As far as we know, ADEx is the first ADs omics database and we expect it to be a reference in this area. During the coming years, ADEx will be expanded including data from more ADs and other omics.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is partially funded by FEDER/Junta de Andalucía-Consejería de Economía y Conocimiento (Grant CV20-36723), Consejería de Salud (Grant PI-0173-2017) and Innovative Medicines Initiative (grant GA-115565).

Authors' contributions

PCS conceived and directed the project. JMM designed the web functionality and interface and prepared the processing pipelines. RLD processed the data. AGM designed and implemented the SQL database and its communication with the website. DTD, KT, GGL and FA contributed to the use cases. JD, AMG and MPC implemented HiPathia analysis. JSR implemented CARNIVAL analysis. VGR, MAR, JAVG and GB tested the software and provided improvements. PCS, JMM, RLD and AGM wrote the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

We would like to thank all the authors of the datasets included in ADEx. We also would like to thank Alberto Ramírez for his technical support during the implementation of ADEx in our server. This work is part of the JMM's PhD thesis. JMM is enrolled in the PhD program in Biomedicine at the University of Granada, Spain.

References

1. Salaman MR. A two-step hypothesis for the appearance of autoimmune disease. Autoimmunity. 2003;36:57–61.

2. Jörg S, Grohme DA, Erzler M, Binsfeld M, Haghikia A, Müller DN, et al. Environmental factors in autoimmune diseases and their role in multiple sclerosis. Cell Mol Life Sci. 2016;73:4611–22.

3. Cooper GS, Stroehla BC. The epidemiology of autoimmune diseases. Autoimmun Rev. 2003;2:119–25.

4. Barturen G, Beretta L, Cervera R, Van Vollenhoven R, Alarcón-Riquelme ME. Moving towards a molecular taxonomy of autoimmune rheumatic diseases. Nat Rev Rheumatol. 2018;14:75–93.

5. Kim H-Y, Kim H-R, Lee S-H. Advances in systems biology approaches for autoimmune diseases. Immune Netw. 2014;14:73–80.

6. Thorlacius GE, Wahren-Herlenius M, Rönnblom L. An update on the role of type I interferons in systemic lupus erythematosus and Sjögren's syndrome. Curr Opin Rheumatol. 2018;30:471–81.

7. Xie X, Li F, Li S, Tian J, Chen J-W, Du J-F, et al. Application of omics in predicting anti-TNF efficacy in rheumatoid arthritis. Clin Rheumatol. 2018;37:13–23.

8. Arriens C, Mohan C. Systemic lupus erythematosus diagnostics in the 'omics' era. Int J Clin Rheumatol. 2013;8:671–87.

9. Teruel M, Chamberlain C, Alarcón-Riquelme ME. Omics studies: their use in diagnosis and reclassification of SLE and other systemic autoimmune diseases. Rheumatol Oxf Engl. 2017;56 suppl_1:i78–87.

10. Ferreira RC, Guo H, Coulson RMR, Smyth DJ, Pekalski ML, Burren OS, et al. A type I interferon transcriptional signature precedes autoimmunity in children genetically at risk for type 1 diabetes. Diabetes. 2014;63:2538–50.

11. Rönnblom L, Eloranta M-L. The interferon signature in autoimmune diseases. Curr Opin Rheumatol. 2013;25:248–53.

12. Khamashta M, Merrill JT, Werth VP, Furie R, Kalunian K, Illei GG, et al. Sifalimumab, an anti-interferon-α monoclonal antibody, in moderate to severe systemic lupus erythematosus: a randomised, double-blind, placebo-controlled study. Ann Rheum Dis. 2016;75:1909–16.

13. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.

14. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update--simplifying data submissions. Nucleic Acids Res. 2015;43 Database issue:D1113-1116.

15. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018;9:1366.

16. Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat Genet. 2013;45:1113–20.

17. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.

18. Jang Y, Choi T, Kim J, Park J, Seo J, Kim S, et al. An integrated clinical and genomic information system for cancer precision medicine. BMC Med Genomics. 2018;11 Suppl 2. doi:10.1186/s12920-018-0347-9.

19. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinforma Oxf Engl. 2007;23:1846–7.

20. Wang Z, Lachmann A, Ma'ayan A. Mining data and metadata from the gene expression omnibus. Biophys Rev. 2019;11:103–10.

21. Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. Genome Biol. 2016;17:177.

22. Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Res. 2010;38:e204.

23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

24. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinforma Oxf Engl. 2004;20:307–15.

25. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat Oxf Engl. 2003;4:249–64.

26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29:15–21.

27. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

28. Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43:e140.

29. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.

30. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma Oxf Engl. 2005;21:3439–40.

31. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.

32. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.

33. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8:203–9.

34. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinforma Oxf Engl. 2008;24:1547–8.

35. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinforma Oxf Engl. 2013;29:189–96.

36. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293.

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

38. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget. 2017;8:5160–78.

39. Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. Npj Syst Biol Appl. 2019;5:40.

40. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29:1363–75.

41. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018;9:20.

42. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2009. https://www.springer.com/us/book/9780387981413. Accessed 30 Apr 2019.

43. Del Carratore F, Jankevics A, Eisinga R, Heskes T, Hong F, Breitling R. RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. Bioinforma Oxf Engl. 2017;33:2774–5.

44. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9:465–76.

45. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2008;29:150–64.

46. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

47. Cubuk C, Hidalgo MR, Amadoz A, Pujana MA, Mateo F, Herranz C, et al. Gene Expression Integration into Pathway Modules Reveals a Pan-Cancer Metabolic Landscape. Cancer Res. 2018;78:6059–72.

48. Amadoz A, Sebastian-Leon P, Vidal E, Salavert F, Dopazo J. Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. Sci Rep. 2015;5:18494.

49. Esteban-Medina M, Peña-Chilet M, Loucera C, Dopazo J. Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. BMC Bioinformatics. 2019;20:370.

50. 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;16:489.

51. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, Edens M, et al. Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. Cell. 2016;165:551–65.

52. Crow MK. Type I Interferon in the Pathogenesis of Lupus. J Immunol Baltim Md 1950. 2014;192:5459–68.

53. Nguyen CQ, Peck AB. The Interferon-Signature of Sjögren's Syndrome: How Unique Biomarkers Can Identify Underlying Inflammatory and Immunopathological Mechanisms of Specific Diseases. Front Immunol. 2013;4:142.

54. Rodríguez-Carrio J, Alperi-López M, López P, Ballina-García FJ, Suárez A. Heterogeneity of the Type I Interferon Signature in Rheumatoid Arthritis: A Potential Limitation for Its Use As a Clinical Biomarker. Front Immunol. 2017;8:2007.

55. Guo Q, Wang Y, Xu D, Nossent J, Pavlos NJ, Xu J. Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. Bone Res. 2018;6:15.

56. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, et al. INTERFEROME v2.0: an updated database of annotated interferon-regulated genes. Nucleic Acids Res. 2013;41 Database issue:D1040–6.

57. Pollard KM, Cauvi DM, Toomey CB, Morris KV, Kono DH. Interferon-γ and Systemic Autoimmunity. Discov Med. 2013;16:123–31.

58. 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. doi:10.1186/s13075-014-0489-x.

59. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME, et al. ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinforma Oxf Engl. 2019;35:880–2.

60. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–4.

61. Díez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. Epigenetics Chromatin. 2015;8. doi:10.1186/s13072-015-0014-8.