UNIVERSIDAD DE GRANADA

PROGRAMA DE DOCTORADO EN BIOMEDICINA

Centro Pfizer – Universidad de Granada – Junta de Andalucía de Genómica e Investigación Oncológica (GENYO)

Atrys Health S.A.

Doctoral thesis

Computational and statistical methods for integrated analysis of biomedical data

Jordi Martorell Marugán

Directed by

Pedro Carmona Sáez

Víctor González Rumayor

Granada, 2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Jordi Martorell Marugán
ISBN: 978-84-1306-853-4
URI: http://hdl.handle.net/10481/68192

“Your focus determines your reality”

Qui-Gon Jinn

AGRADECIMIENTOS

En primer lugar, es bien sabido que los científicos ya no somos personas aisladas en nuestros laboratorios capaces de hacer grandes descubrimientos por sí solos. La investigación ahora se hace en equipo y es multidisciplinar. En consecuencia, las tesis doctorales ya no son solo del doctorando y de su director, sino que son de un equipo y de sus colaboradores. Por ello, aunque en la portada solo aparezcan 3 nombres, esta tesis es el resultado del trabajo de muchas personas, incluyendo a todos los autores de las publicaciones que han dado lugar esta tesis. Gracias a todos.

Pero para desarrollar una tesis no solamente hace falta el trabajo, sino que es necesario un buen ambiente en el que crecer, no solo científicamente, sino como persona. En esta tarea han colaborado y siguen colaborando Dani, Raúl, Adrián, Alba, Juan Antonio e Iván. No solo sois grandes profesionales, sois amigos. Hacéis que el trabajo no parezca trabajo y tenéis siempre la mano tendida para ayudar a los demás. Muchas gracias y seguid siendo así.

I also wanted to thank Marco and Giuseppe for their supervision during my stay in the MPBA group. Despite the bad situation due to the COVID-19 pandemic, you have been very supportive and I learnt a lot with you. Thank you.

También agradezco a Víctor y al equipo de Atrys su apoyo, que ha sido determinante para que esta tesis se haya podido realizar. Gracias, Víctor.

Finalmente, merece una especial mención Pedro, quien confió en mí pese a no tener experiencia, quien me ha enseñado tanto, quien me introdujo en el mundo de la investigación y quien siempre me ha apoyado. Él ha sido mi director, mi guía y mi amigo. Muchas gracias y espero que cumplas pronto con tus objetivos, te lo mereces.

QUALITY CRITERIA TO APPLY FOR THE DEGREE OF "INTERNATIONAL PHD" BY THE UNIVERSITY OF GRANADA

This doctoral thesis has been prepared according to the University of Granada requirements to apply for an International PhD.

SCIENTIFIC PUBLICATIONS

The following articles containing the thesis results were published in high impact journals in the research field of the thesis:

- **Martorell-Marugan J**, Toro-Dominguez D, Alarcon-Riquelme ME & Carmona-Saez P (2017) MetaGenyo: a web tool for meta-analysis of genetic association studies. *BMC Bioinformatics* 18, 563. Impact factor: 2.213, Category: Mathematical & Computational Biology, Position: 14/59 (**Q1**).

- Toro-Domínguez D*, **Martorell-Marugán J***, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME & Carmona-Sáez P (2019) ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics* 35, 880–882. Impact factor: 5.610, Category: Mathematical & Computational Biology, Position: 3/59 (**Q1**, **D1**). *Shared first authorship.

- **Martorell-Marugán J**, González-Rumayor V & Carmona-Sáez P (2019) mCSEA: detecting subtle differentially methylated regions. *Bioinformatics* 35, 3257–3262. Impact factor: 5.610, Category: Mathematical & Computational Biology, Position: 3/59 (**Q1**, **D1**).

Given that the criteria to write the thesis as a compendium of publications is met, this document have been prepared following this modality.

INTERNATIONAL RESEARCH STAY

During the period September 2020 – December 2020, an international internship was performed at the Predictive Models for Biomedicine & Environment (MPBA) department at the Fondazione Bruno Kessler (FBK) in Trento, Italy. During this 3 months visit, Dr. Giuseppe Jurman supervised the project titled "Artificial Intelligence for precision medicine in Systemic Autoimmune Diseases" funded by an EMBO Short-Term Fellowships grant.

LANGUAGE OF THE THESIS AND PRESENTATION

This doctoral thesis has been written in English and the conclusions will be defended in English. Following the University of Granada regulation, the abstract and conclusions sections have been written in Spanish too.

GRANTS AND FUNDING

This doctoral thesis has been possible thanks to the following grants awarded to the candidate:

- Ayudas para contratos para la formación de investigadores en empresas (Doctorados Industriales) 2016. Ministerio de Economía, Industria y Competitividad.

- Short-Term Fellowship. European Molecular Biology Organization (EMBO).

In addition, this thesis received support from the following research projects:

- Identificación de biomarcadores en lupus eritematoso sistémico mediante análisis integrado de transcriptoma y metiloma. Consejería de Salud de la Junta de Andalucía. Reference PI-0173-2017.

- Molecular reclassification to find clinically useful biomarkers for systemic autoimmune diseases (PRECISESADS). EU-Innovative Medicines Initiative (IMI). Reference 115565.

INDEX

ABSTRACT

During recent years, the new omics technologies have revolutionized the biomedical research paradigm, changing from studying few specific elements based on previous hypotheses to studying complete systems like the genome or the transcriptome, generating hypotheses from the data. This change has created the necessity of a new profile in the biomedical research, the bioinformatician or computational biologist, who combines knowledge about biology, informatics and statistics in order to analyse these huge amounts of data and to develop new analytical methods.

In this context of massive data generation, different public repositories were created where researchers can submit the data generated in their studies with the aim of guaranteeing the reproducibility of their results and of doing the data usable in other retrospective studies. For the last years, the amount of stored data in public repositories has grown exponentially thanks to the lowering costs of the necessary technologies to generate them. One of the most used repositories is the Gene Expression Omnibus (GEO), maintained by the NCBI. GEO contain the data generated in all types of omics projects, including gene expression, methylation or DNA sequencing, among others.

The availability of all these amounts of information offers an invaluable resource to generate and test hypotheses through the use and integration of these data. However, for that aim, proper statistical and computational methods for integrating information are necessary. Among the strategies to reanalyse public data is the meta-analysis, consisting on the combination of the results from different studies using proper statistical techniques with the aim of increasing the statistical power and resolving discrepancies between studies, among other applications.

The main objective of this doctoral thesis has been the development of computational methods for the integration of heterogeneous data sets with the aim of analysing them in conjunction using meta-analysis and integrated analysis methods.

Firstly, MetaGenyo was developed, which is a web tool for performing meta-analysis of genetic association studies. MetaGenyo implements all the necessary steps to conduct this kind of studies guaranteeing the use of proper statistical techniques for each case.

Once the potential of meta-analysis to integrate data from different origins to analyse common effects and detect anomalies was demonstrated, we used the framework developed in MetaGenyo to develop ImaGEO, another interactive tool that implements the complete analysis workflow to integrate gene expression studies and to analyse them jointly to find differentially expressed genes. ImaGEO permits the download and analysis of transcriptomics data from the GEO repository using the studies identifiers as input.

Gene expression regulation is a complex process in which DNA methylation has an important role. Hence, studying methylation alterations and integrating them with gene expression data is a key for unrevealing the molecular mechanisms of several diseases. In this context, mCSEA was developed, which is a new method for detecting differentially methylated regions. This algorithm was integrated in an R package published in the Bioconductor repository. Among other features, mCSEA includes a module to integrate differentially methylated regions with the expression of close genes, addressing a common problem in multi-omics methylation-expression studies.

Finally, databases that compile public information from specific pathologies have been established as a new alternative for the exploit of those data by the research community of that area. ADEx was developed following that premise, which is a database that compiles the available transcriptomics and methylation studies from autoimmune diseases and includes a layer of visualization and analysis, including meta-analysis. Using ADEx, the interferon signature in different pathologies was studied and several potential biomarkers with altered gene expression patterns consistent in each diseased were identified.

In conclusion, in this thesis four tools for exploiting and integrating biomedical data were developed, using known meta-analysis techniques as well as a new algorithm for differential methylation detection.

RESUMEN

Durante los últimos años, las nuevas tecnologías ómicas han revolucionado el paradigma de la investigación biomédica, pasando de estudiar unos pocos elementos concretos basándose en hipótesis previas a estudiar sistemas completos como el genoma o el transcriptoma, generando hipótesis a partir de los datos. Este cambio ha creado la necesidad de un nuevo perfil en la investigación biomédica, el del bioinformático o biólogo computacional, que combina conocimientos de biología, informática y estadística para analizar estas grandes cantidades de datos y desarrollar nuevos métodos analíticos.

En este contexto de generación de datos masivos, se crearon distintos repositorios públicos en los que los investigadores pueden subir los datos que generan en sus estudios con el fin de garantizar la reproducibilidad de sus resultados y de que puedan ser usados en otros estudios retrospectivos. Durante los últimos años, la cantidad de datos almacenados en repositorios públicos ha crecido exponencialmente gracias al abaratamiento de las tecnologías necesarias para generarlos. Uno de los repositorios más usados es el Gene Expression Omnibus (GEO), mantenido por el NCBI. GEO contiene los datos generados en todo tipo de proyectos ómicos, incluyendo datos de expresión, metilación o secuenciación de ADN, entre otros.

La disponibilidad de toda esta gran cantidad de información ofrece un recurso inestimable para generar y contrastar hipótesis mediante el uso o integración de estos datos. No obstante, para ello se requieren metodologías estadísticas y computaciones apropiadas que puedan ser aplicadas para la integración de información. Entre las estrategias para reanalizar datos públicos se encuentra el meta-análisis, que es la combinación de los resultados de distintos estudios mediante técnicas estadísticas apropiadas con el fin de aumentar el poder estadístico y de resolver discrepancias entre estudios, entre otras aplicaciones.

El objetivo principal de esta tesis doctoral ha sido el desarrollo de métodos computacionales para la integración de conjuntos de datos heterogéneos y de distinto origen, con el objeto de analizarlos conjuntamente mediante metodologías de meta-análisis y análisis integrado de datos.

En primer lugar, se desarrolló MetaGenyo, una herramienta web para llevar a cabo meta-análisis de estudios de asociación genética. MetaGenyo implementa todos los pasos

necesarios para llevar a cabo este tipo de estudios garantizando el uso de las técnicas estadísticas adecuadas en cada caso.

Visto el potencial de los meta-análisis para integrar datos de diferente origen con el objetivo de analizar efectos comunes o detectar anomalías, utilizamos el *framework* desarrollado en MetaGenyo para desarrollar ImaGEO, otra herramienta interactiva que implementa un flujo completo de análisis para integrar estudios de expresión génica y analizarlos conjuntamente para la búsqueda de genes diferencialmente expresados. ImaGEO permite la descarga y procesamiento de datos desde el repositorio GEO usando los identificadores de los estudios como entrada.

La regulación de la expresión génica es un proceso complejo en el cual la metilación del ADN tiene un papel importante. Por lo tanto, el estudio de las alteraciones en la metilación y su integración con datos de expresión génica es clave para revelar los mecanismos moleculares de muchas enfermedades. En este contexto, se desarrolló mCSEA, un nuevo método de detección de regiones diferencialmente metiladas. Este algoritmo se integró en un paquete de R publicado en el repositorio Bioconductor. Entre otras funcionalidades, mCSEA incluye un módulo de integración de las regiones diferencialmente metiladas con la expresión de genes cercanos, abordando un problema común de estudios multi-ómicos de metilación-expresión.

Finalmente, las bases de datos que recopilan información pública sobre patologías concretas se han establecido como una nueva alternativa para la explotación de dichos datos por parte de la comunidad investigadora de dicha área. Bajo esta premisa se preparó ADEx, una base de datos que recopila los estudios disponibles de transcriptómica y metilación sobre enfermedades autoinmunes e incluye una capa de exploración y análisis, incluyendo meta-análisis. Con ADEx se estudió el comportamiento de la firma del interferón en distintas patologías y se identificaron potenciales biomarcadores con alteraciones en la expresión consistentes dentro de las distintas enfermedades.

En conclusión, en esta tesis se han desarrollado cuatro herramientas para la explotación e integración de datos biomédicos mediante la aplicación de técnicas conocidas de meta-análisis y un nuevo algoritmo de detección de metilación diferencial.

ABBREVIATIONS

AD: autoimmune disease

BMIQ: Beta-Mixture Quantile

ChIP: chromatin immunoprecipitation

CI: confidence intervals

CGI: CpG islands

COHCAP: City of Hope CpG Island Analysis Pipeline

CpG: cytosine-phosphate-guanine

DMP: differentially methylated position

DMR: differentially methylated region

DNMT: DNA methyltransferase

DNA: deoxyribonucleic acid

EBI: European Bioinformatics Institute

EGA: European Genome-phenome Archive

ENA: European Nucleotide Archive

ENCODE: Encyclopedia of DNA Elements

eQTL: expression quantitative trait loci

ES: Enrichment Score

EWAS: epigenome-wide association study

FEM: fixed effect model

FGSEA: fast gene set enrichment analysis

GAS: genetic association study

GDC: Genomic Data Commons

GEO: Gene Expression Omnibus

GSA: Gene-set analysis

GSEA: Gene Set Enrichment Analysis

GWAS: genome-wide association study

HDAC: histone deacetylase

HWE: Hardy-Weinberg equilibrium

ICGC: International Cancer Genome Consortium

IFN: interferon

IMA: Illumina Methylation Analyzer

INDEL: small-scale insertion/deletion

JCR: Journal Citation Reports

mRNA: messenger RNA

NCBI: National Center for Biotechnology Information

NGS: next generation sequencing

NHGRI: National Human Genome Research Institute

Noob: normal-exponential out-of-band

OR: odds ratio

PCAWG: pan-cancer analysis of whole genomes

qPCR: quantitative polymerase chain reaction

REM: random effect model

RFLP: Restriction Fragment Length Polymorphism

RNA: ribonucleic acid

SE: standard error

SNP: single nucleotide polymorphism

SWAN: Subset-quantile Within Array Normalization

TCGA: The Cancer Genome Atlas

WGBS: whole-genome bisulfite sequencing

1 INTRODUCTION

1.1 OMICS DATA: A REVOLUTION IN BIOMEDICINE

During the 20th century, the development of molecular biology techniques permitted the generation of biological data like the sequences of amino acids in proteins or nucleotide sequences in deoxyribonucleic acid (DNA). As early as 1965, Margaret Dayhoff, who is considered one of the first computational biologist, published *Atlas of Protein Sequence and Structure*, a book where she compiled all the known protein sequences at that time (Dayhoff, 1965). Some years later, on 1982, GenBank database (Clark et al., 2016) was created to store the DNA sequences submitted by many laboratories around the world. These initial biological databases were the beginning of the exponential growth of data in life sciences.

However, the amount of time and resources to generate biological data during those times was huge. Maybe the best example of that was the Human Genome Project (Lander et al., 2001), started in 1990 and officially finished in 2003 when most of the human genome was sequenced and assembled (International Human Genome Sequencing Consortium, 2004). This milestone was achieved after 13 years of work and a budget of 4.8 billion dollars (Lewin et al., 2018).

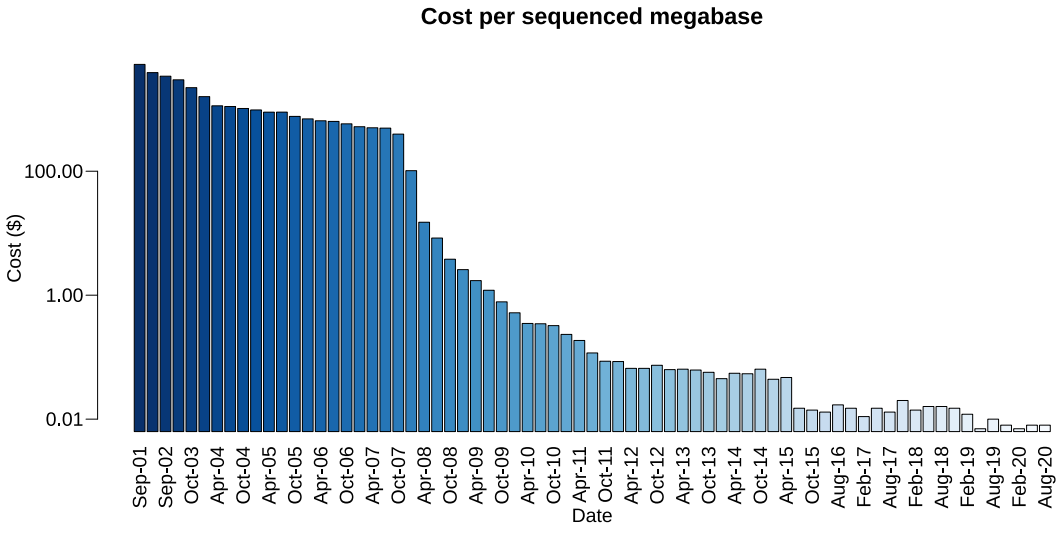


Figure 1. Cost of sequencing 1 megabase of DNA along time. Y axis is on logarithmic scale. This figure was generated from the data published by the National Human Genome Research Institute (NHGRI) (Wetterstrand, 2020).

Nevertheless, a few years later after the Human Genome Project completion, the DNA sequencing technologies advanced towards high-throughput sequencing or next generation sequencing (NGS). Since then, the necessary time and money to sequence DNA has dropped drastically (Figure 1), passing from $5292 per sequenced megabase (Mb) in 2001 to $0.008 in 2020 (Wetterstrand, 2020). In fact, the sequencing of a human genome costs currently less than $700.

These innovations permitted the onset of **omics** disciplines, a set of areas that study biological systems from a global perspective, like genome sequences (genomics), gene expression (transcriptomics) or epigenetic regulation (epigenomics). Microarrays technologies have also contributed to the initial increase of omics studies, although its use is decreasing in accordance with the drop of NGS prices. Other omics, like proteomics and metabolomics, do not depend on DNA sequencing, but other technological advances facilitated their growth (e.g., mass spectrometry).

Omics data have changed the perspective of the biomedicine research. For instance, instead of studying a few candidate mutations in one or few genes suspected to be linked to a disease, the whole genome can be compared between cases and controls to identify mutations associated to the risk of having the disease. This is a hypothesis-free approach, opposite to the classical hypothesis-driven one used in science during centuries. The number of biomedical advances thanks to this change in perspective is countless. For instance, transcriptomics-derived biomarkers permit the stratification of breast cancer patients depending on their risk of recurrence (Cronin et al., 2007; Prat et al., 2012). Using omics-based biomarkers it is possible to perform a deep characterization of a disease in order to stratify patients and to prescribe them the best available treatments based on their specific molecular background. This constitutes the bases of precision medicine, so omics data are fueling precision medicine through the discovery and use of biomarkers (Quezada et al., 2017).

1.2 PUBLIC OMICS REPOSITORIES

In parallel to the popularization of the technologies and, consequently, to the increasing number of omics studies, public omics data repositories emerged. These repositories allow scientists to store their data in common platforms and to make them available to

the scientific community. Omics data sharing is key to ensure the transparency and reproducibility of the biomedical studies that employ high-throughput technologies.

One of the most popular repositories is **Gene Expression Omnibus** (GEO) (Edgar et al., 2002), developed by the National Center for Biotechnology Information (NCBI). The initial purpose of GEO was to store microarrays data before NGS development, but it has been adapted to share any type of high-throughput data, including ribonucleic acid (RNA)-Seq, chromatin immunoprecipitation (ChIP)-Seq or methylation data, among others.

In order to organize the large amounts of data stored in GEO, each dataset, sample and platform is associated to a unique GSE, GSM and GPL code respectively. In addition, authors are requested to follow some minimal guidelines to organize their data in order to facilitate the usage of those data by the users. During the recent years, the amount of data available in GEO has risen drastically (Figure 2). At the end of 2020, 140,758 datasets and 4,082,707 samples generated with 21,646 annotated platforms were publicly available in GEO.

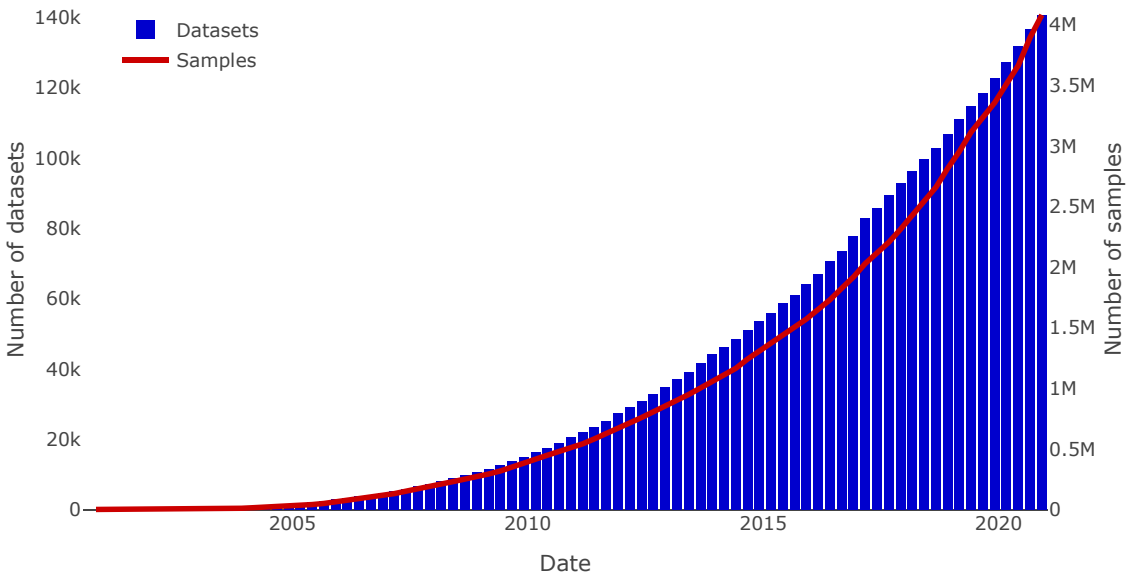


Figure 2. Trend of the number of available datasets and samples in GEO since the third quarter of 2000 to the end of 2020. Blue bars indicate the number of datasets and the red line indicates the number of samples. The data used to generate this plot was obtained from the GEO summary webpage (https://www.ncbi.nlm.nih.gov/geo/summary/?type=history).

The Sequence Read Archive (SRA) (Leinonen et al., 2011), also maintained by the NCBI, is another public repository dedicated to store raw sequence and alignment data from NGS experiments. Among other data, SRA stores the raw data from high-throughput GEO datasets. Similarly to GEO, SRA has experienced a huge growth during the recent years and, at the end of 2020, SRA stored more than 46 x 10^{15} sequenced bases (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi).

The European analogous databases to SRA and GEO are the European Nucleotide Archive (ENA) (Harrison et al., 2021) and ArrayExpress (Athar et al., 2019) respectively, both developed by the European Bioinformatics Institute (EBI). The amount of available data in these databases can be exemplified by the 119 trillion sequences available in ENA and the more than 59 Tb of archived data in ArrayExpress. Another relevant European database is the European Genome-phenome Archive (EGA) (Lappalainen et al., 2015), which has the particularity of containing identifiable personal biomedical data from patients who consented the use of their data for research.

In addition to the previous general data repositories, there are also several databases dedicated to share data from specific projects which generated large amounts of omics data. For instance, The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) project generated over 2.5 Pb of multi-omics data from 33 cancer types which can be accessed through the Genomic Data Commons (GDC) Data Portal (Grossman et al., 2016). The International Cancer Genome Consortium (ICGC) also compiles large amounts of information in their data portal, including the TCGA data but also other projects like the recent pan-cancer analysis of whole genomes (PCAWG) (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

Another important project that share large amounts of data is the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium, 2012), which is focused on human functional genomics data. Data from RNA-Seq, ChIP-Seq, ATAC-Seq or DNase-Seq, among others, from over 13,000 datasets is available in the ENCODE data portal (Davis et al., 2018).

1.3 EXPLOITING PUBLIC OMICS DATA: META-ANALYSIS

In the previous section, some of the most popular public data repositories have been reviewed. However, there are a lot of other projects and databases with available data that, using the proper statistical and computational techniques, can be an invaluable resource for hypothesis testing and knowledge generation.

Perez-Riverol *et al.* recently evaluated the impact of public omics datasets in research, reaching the conclusion that many of the public datasets are re-analysed at least once and, in April 2019, 58,054 of the datasets followed up by their OmicsDI resource has at least one citation, specially transcriptomics datasets (Perez-Riverol et al., 2019). Some concrete examples of relevant works where public data were used are articles published in the Cell journal (Bailey et al., 2018; Chen et al., 2018; Huang et al., 2018; Sanchez-Vega et al., 2018). These metrics make evident the current importance of public omics data reutilization.

Public data download and analysis is not always easy and commonly requires advanced computational skills. In addition, even for expert users, these tasks are usually time consuming. For those reasons, several software has been developed in order to facilitate these tasks. Some examples are the GEOquery (Davis and Meltzer, 2007) and TCGAbiolinks (Colaprico et al., 2016) R packages, which permit the easy data downloading from GEO and TCGA data respectively.

In addition to individual reanalyses, a common approach to exploit public data is to perform **meta-analyses**. A meta-analysis is a general approach to combine the data and/or results from several studies of the same problem in order to solve discrepancies across these studies and to increase the statistical reliability of the results raising the sample size. Meta-analysis has been applied to all types of data and, although there are some common steps for all types of meta-analyses (e.g., data selection or quality control), each type of meta-analysis has particularities depending on the kind of data analysed.

There are three main approaches to conduct a meta-analysis: methods based on effect sizes combination, methods based on P-values combination and methods based on ranks combination (Toro-Domínguez et al., 2020), described in Section 3.1.

1.4 META-ANALYSIS OF GENETIC ASSOCIATION STUDIES

Genetic epidemiology discipline uses clinical data from families and populations to find phenotypic traits (e.g. diseases) that are heritable and, therefore, are caused by variations in the DNA (The Psychiatric GWAS Consortium, 2009). Once a phenotype is determined to be heritable, different molecular biology techniques are used in order to discover which genetic traits are linked to these kinds of heritable phenotypes. Before the development of DNA sequencing techniques, strategies like Restriction Fragment Lengths Polymorphisms (RFLPs) allowed to determine genomic regions associated with the studied phenotype. A major achievement in this context was the identification of the Huntington's disease gene (Gusella et al., 1983). However, after the emergence of DNA sequencing techniques and, specially, NGS, researchers have been able to discover associations between variations in the DNA and phenotypes.

In this context, there are two main approaches to discover the genetic cause of a phenotype: genetic association studies (GAS) and genome-wide association studies (GWAS). The main difference between these approaches is the scale: while GAS test the association in one or few candidate regions, GWAS test for associations along the entire genome. The use of one or other approach depends on several variables like the complexity of the studied treat or the previous knowledge about the studied phenotype.

There are various kinds of mutations that can be studied in these analyses. The most common genetic variation are single nucleotide polymorphisms (SNPs), which are variations on a single nucleotide in the DNA. Other common studied variations are small-scale insertions/deletions (INDELs), which are small insertions or deletions in the genome.

GAS present some limitations. For instance, these studies are usually performed with a small sample size, they are performed with patients belonging to one single ethnicity or have other methodological issues (Li and Meyre, 2013). These problems result in low reproducibility and insufficient statistical power of each individual study.

To address these limitations, meta-analyses of GAS may be used. Meta-analysis has been used in a wide range of disciplines, and GAS are no exception. In fact, during the last years, these kind of meta-analyses have experienced an exponential increase (Ioannidis et al., 2013). However, GAS meta-analyses require some important steps and

considerations that are frequently overlooked by the authors of these studies (Ioannidis et al., 2013). These common errors are committed, in part, because these meta-analyses are sometimes performed by researchers without the necessary statistical knowledge. In addition, there is not any dedicated software to perform GAS meta-analyses, so authors use general meta-analysis or statistical programs which do not guide through the necessary steps of a GAS meta-analysis.

1.5 META-ANALYSIS OF GENE EXPRESSION DATA

According to the central dogma of molecular biology (Crick, 1970), RNA polymerases synthesize RNA molecules from DNA templates (e.g., genes) during the process known as transcription. Then, ribosomes use those RNA molecules, known as messenger RNA (mRNA) to produce proteins during the translation process. Latter research revealed that this scenario is more complex, occurring phenomena like reverse transcription (when RNA is transcribed to DNA though reverse transcriptases).

Gene expression is defined as the process by which the information encoded in genes is used to produce their corresponding gene product, which usually are functional proteins, although they can also be functional non-coding RNAs. Given the major importance of gene expression in the basis of biology and medicine, measuring it has been a very active research field in biomedicine. Gene expression can be estimated measuring the proteins in cells with molecular biology techniques like Western blots or high-throughput approaches (proteomics). However, proteins are not easy to identify and quantify, and non-coding RNA abundances can not be measured given that they are not translated into proteins. For these reasons, it is common to estimate gene expression from the levels of mRNA in the samples, which is a simpler molecule easier to work with. Techniques like Northern blot or quantitative polymerase chain reaction (qPCR) may be used to quantify one or few mRNAs. However, new high-throughput technologies permitted the onset of **transcriptomics**, which is the study of whole sets of transcripts produced by the genome (known as transcriptome).

Among the first high-throughput technologies for gene expression quantification, hybridization arrays, or **microarrays**, had been a very popular choice before NGS became affordable. This technology is based on the hybridization between transcripts and probes, DNA sequences fixed on a solid surface (array) (Schena et al., 1995). The mRNA

has to be converted to DNA and marked with fluorescent dyes. After hybridization and washing steps, the arrays are scanned in order to measure the fluorescence signal in each spot. A lot of different microarrays technologies were used with differences in the probes employed (complementary DNA or oligonucleotides) or the number of detected fluorescent signals (1 color or 2 colors arrays) (Schulze and Downward, 2001). Furthermore, many companies commercialized microarrays with different technologies, and even custom arrays were designed by research groups for specific tasks. For these reasons, the gene expression data generated with microarrays is very heterogeneous and specific analytical pipelines must be applied for each platform.

Together with the emergence of NGS technologies, a new approach for gene expression quantification, **RNA-Seq**, started to replace microarrays. RNA-Seq consists on the application of some of the available NGS technologies to sequence RNA molecules. Despite, as happened with microarrays, there are different RNA-Seq protocols with technical variations among them, the raw data produced in these experiments is much more homogeneous, consisting generally on the sequences and the associated bases qualities in *fastq* files.

Both microarray and RNA-Seq data analysis share some essential steps to pass from raw data to expression matrices ready to perform statistical analyses. The specific methods depends on both the kind of data and the aim of the objectives and they were comprehensively reviewed for microarrays (Slonim and Yanai, 2009) and RNA-Seq (Conesa et al., 2016).

Gene expression meta-analysis, as other types of meta-analysis, may be used to integrate different studies focused on the same phenotype, increasing the statistical power and resolving discordances between individual studies. This approach has been applied, for instance, in the context of cancer (Bell et al., 2017; O'Mara et al., 2016) and autoimmune diseases (Afroz et al., 2017; Song et al., 2014).

However, gene expression meta-analyses have other applications more specific to this kind of data. For instance, a meta-analysis can be performed to integrate studies of different phenotypes in order to find common differentially expressed genes among them. This approach may be useful to discover novel molecular similarities among conditions which may be the basis to propose known therapies for some disorders as novel therapies

for other molecularly similar disorders (drug repurposing). Shared gene expression signatures for autoimmune diseases (Toro-Domínguez et al., 2014) or neurological diseases (Kelly et al., 2019) have been discovered following this strategy.

Furthermore, another application is the identification of opposed gene signatures among different phenotypes. For instance, it was known that cancer and central nervous system diseases (like Alzheimer's disease, Parkinson's disease and schizophrenia) have an inverse comorbidity. For that reason, a meta-analysis was performed searching for genes underexpressed in one disorder and overexpressed in other ones, finding inverse patterns between both groups of pathologies (Ibáñez et al., 2014).

1.6 DNA METHYLATION

Epigenetics study heritable changes in gene expression occurring without changes in the DNA sequence (Berger et al., 2009). There are multiple epigenetic marks involving consequences for gene expression, like chromatin remodeling, non-coding RNAs expression, histone modifications or DNA methylation (Ibeagha-Awemu and Zhao, 2015). Most of epigenetic epidemiology studies centers on DNA methylation because it is relatively stable and there are many available platforms to quantify it (Bakulski and Fallin, 2014).

DNA methylation is the addition of a methyl group (CH_3) at the C5 position of a cytosine, generally located in a cytosine-phosphate-guanine (CpG) dinucleotide (Figure 3) (Bird, 2002). DNA methyltransferases (DNMTs) are the enzymes which fix and preserve DNA methylation. CpG islands (CGI) are groups of CpG dinucleotides in a genomic region. There are more than 40,000 CGI in a mammalian genome and they are usually located in the promoters or in the first exon of the genes (Orlando et al., 2012), but there are also many CGI distant from annotated promoters (Deaton and Bird, 2011). Around 70 % of human gene promoters are linked to a CGI, including almost all housekeeping genes and a significant proportion of development regulators and tissue-specific genes (Saxonov et al., 2006). All these evidences point that CGIs are strongly correlated with transcription initiation.

The details of how DNA methylation affects gene expression are still under research. An increase in the methylation levels in CpG regions reduce the physical accessibility of DNA to transcription factors, resulting in a decrease in gene expression (Suzuki and Bird,

2008). On the other hand, methylation in CpG regions is recognized by a group of proteins called MDBs. MDBs are associated with chromatin remodeling proteins enzymes like histone deacetylases (HDACs). So, when a MDB recognize a methylated CpG region, it recruits HDACs which remove acetyl groups from the histones, compacting the chromatin and impeding transcription factors and transcription machinery to interact with DNA, decreasing the gene expression (Jones and Laird, 1999).

It is known that DNA methylation has an important role in many cellular processes, so it is currently being studied by many researchers in order to get a better understanding of human development and diseases (Robertson, 2005). In this context, epigenome-wide association studies (EWAS) search for associations between lifestyle, environmental factors or diseases and epigenetic changes, mostly DNA methylation (Flanagan, 2015).

There are different technologies to determine DNA methylation status, being whole-genome bisulfite sequencing (WGBS) one of the most accurate and with highest coverage. However, Illumina's BeadChip arrays (Infinium HumanMethylation27, Infinium HumanMethylation450 and Infinium MethylationEPIC) are much more affordable and simpler to analyse, and they are currently the most used platforms in human EWAS (Teh et al., 2016).

Illumina BeadChip microarrays contain probes to interrogate different amounts of CpG sites (e.g., 485,577 CpGs in the Infinium HumanMethylation450) (Morris and Beck, 2015). While the Infinium HumanMethylation27 microarray only contained one type of probes, the Infinium HumanMethylation450 and Infinium MethylationEPIC genotype bisulfite-converted DNA using two types of methylation assays (Figure 4).

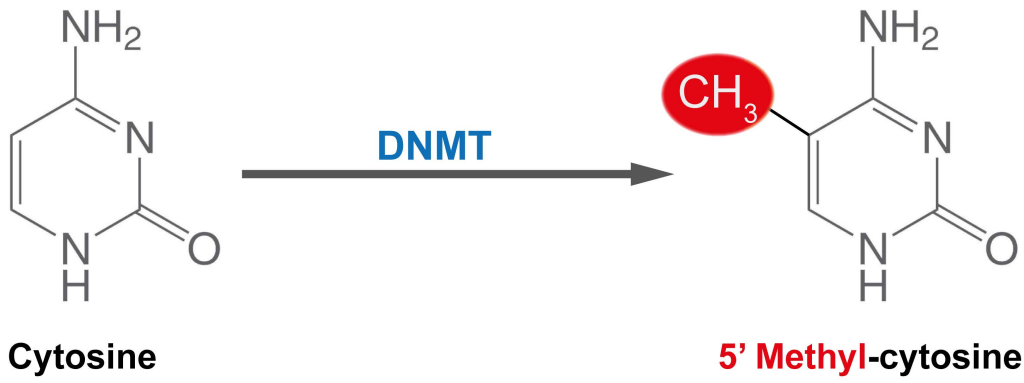


Figure 3. Diagram of the methylation of a cytosine in the DNA.

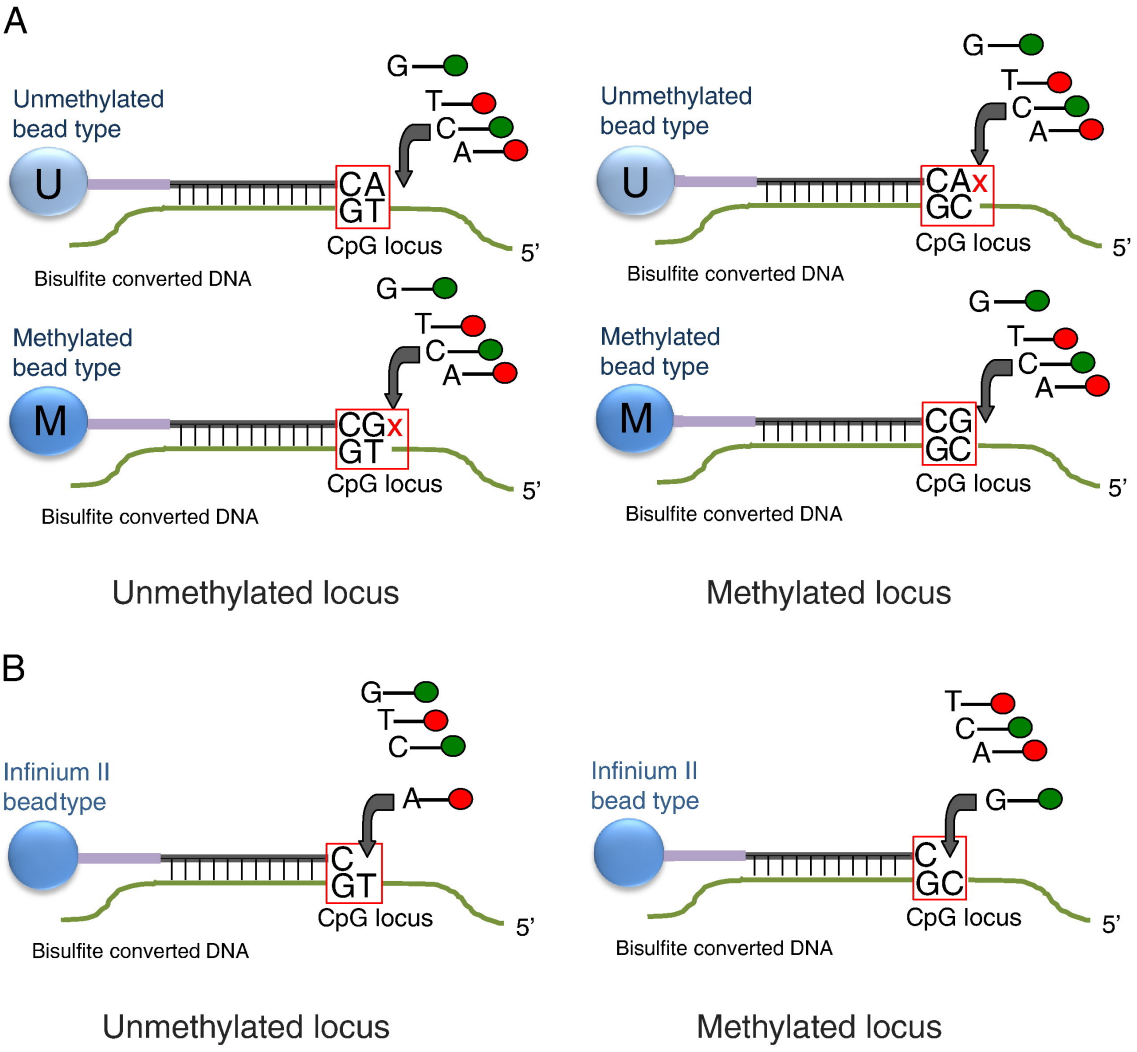


Figure 4. Infinium I (A) and Infinium II (B) methylation assays designs. 3' extreme of Infinium I probes matches either a GT dinucleotide coming from an unmethylated CpG site or a GC dinucleotide coming from a methylated CpG site. Infinium II probes hybridize until the G from the GT/GC dinucleotides, Methylation is determined by single-base extension: A and G bases are labeled with different colors, and they are incorporated to unmethylated and methylated sites respectively. Figure extracted from Bibikova *et al.*, 2011 © Elsevier, license number 4991800130743.

Infinium I assay has 2 probes per CpG site: one probe hybridizes to the methylated cytosine and the other one hybridizes to the unmethylated one, which has been transformed to thymine after bisulfite treatment and genomic amplification. On the other hand, Infinium II assay consists of one single probe per CpG site capable to hybridize to both methylated (C) or unmethylated (T) sites, each of one previously marked with different colors (Bibikova et al., 2011).

1.7 OMICS DATA IN AUTOIMMUNE DISEASES

The **immune system** is a complex network of organs, cells and biological processes that act coordinately as a defense against external agents such as bacteria or viruses (Nicholson, 2016). In mammals, there are 5 main types of immune cell types: neutrophils, lymphocytes, monocytes, basophils and eosinophils. Nevertheless, for each cell type there are different subpopulations depending on the receptors (i.e., proteins that bind to ligands or cytokines) present in their membranes.

The immune system is divided into the **innate** immune system and the **adaptative** immune system, although both are very connected. The innate immune system is the first one acting in case of infection and it triggers a relatively unspecific response based on inflammation, which is the recruitment of some immune cell types like macrophages and neutrophils. If the infection persists, the adaptative immune system is activated by the recruitment of lymphocytes (T-cells and B-cells), which make specific responses and remember the pathogen so, in case of a future infection from the same pathogen, a more efficient and quick response is activated.

There are several diseases due to misfunctions in the immune system, classified in groups like immunodeficiencies (loss in the ability to activate immune responses), allergies (hypersensitivity reactions) or **autoimmune diseases** (ADs). ADs are originated when the adaptative immune system mistakenly recognizes healthy own tissues of the organism as foreign and activate an immune response against them. There are organ-specific ADs (e.g., type I diabetes) with an autoimmune response against one organ, and systemic ADs (e.g., systemic lupus erythematosus) which affect several organs. Commonly, ADs are chronic and affects significantly the lifespan of the patients, being at the top 10 causes of death for women and being risk factors for other severe illnesses like cardiovascular diseases and cancer (Lens-Pechakova, 2016).

Recent omics studies discovered that, in ADs, there is an interplay between environmental factors, genetic predisposition, epigenomics alterations and gene expression (Teruel et al., 2017). These studies also put in evidence the high molecular variability existing not only between different ADs, but also among patients of the same disease and even for the same patient along time (Toro-Domínguez et al., 2018). Such heterogeneity makes very challenging to advance in precision medicine for these diseases, and large amounts of

data from several patients is needed to classify patients and prescribe them the most appropriate treatments.

A major research topic in this context is the activity of the interferon (IFN) signature. IFN is a group of cytokines that trigger the immune response. Molecularly, IFN binds to specific cell receptors, initiating a downstream signaling that results in the gene expression regulation of several genes. The alteration of such genes, named IFN signature, has been observed in several ADs. However, there are many ADs patients that do not present this signature, evidencing again the heterogeneity of these diseases.

As for other diseases, many omics studies have been performed to get insight into ADs, especially transcriptomics and epigenomics ones (Teruel et al., 2017). These studies normally focus on specific biological questions, so the generated data is used to solve these questions and it is normally deposited on public repositories like GEO. These datasets are especially valuable in the context of ADs, given the necessity of collecting large amounts of data to overcome with the heterogeneity of these diseases and to advance towards precision medicine by, for instance, finding shared molecular patterns among diseases (Barturen et al., 2020).

However, for several reasons it is not easy to integrate these data. The first challenge is to identify the specific studies that can be interesting to reanalyse, given that repositories like GEO employ unstructured text to store the metadata of each dataset (Chen et al., 2019). In addition, many different platforms have been used to generate such data (e.g., different gene expression microarrays). Furthermore, there is not a standard procedure to process omics data, resulting in heterogeneous pipelines that introduce systematic biases among studies.

Despite these inconveniences, some works integrated different ADs datasets. For instance, Toro-Domínguez *et al*. found a shared molecular pattern between systemic lupus erythematosus, Sjögren's syndrome and rheumatoid arthritis performing a gene expression meta-analysis of public data (Toro-Domínguez et al., 2014). Furthermore, the same authors integrated several lupus datasets to perform a drug repurposing study, proposing new treatments for this disease (Toro-Domínguez et al., 2017). These are examples of the potential of reusing public ADs data. However, these works implicated a remarkable effort to overcome with the challenges already mentioned.

1.8 MULTI-OMICS DATA INTEGRATION

Single-omics biomedical studies have increased drastically our knowledge about the molecular mechanisms of many diseases. However, during the recent years, several articles are evidencing that, although single-omics studies may be very useful, the integrated analysis of more than one omics data type (multi-omics studies) may provide a deeper knowledge of the complex molecular pathways and interactions in the context of complex diseases (Civelek and Lusis, 2014). These diseases can not be easily explained by a single layer of information, but by the interplay between several biological layers. For this reason, regardless the proved utility of single-omics studies, they usually provide limited information regarding the molecular causes of diseases (Huang et al., 2017). To solve this limitation, some projects like TCGA (Weinstein et al., 2013) generated multi-omics data covering multiple biological layers like DNA sequence, gene expression, DNA methylation and so on.

Multi-omics studies present an opportunity to acquire a better understanding of diseases and their cause-effect relationships, resulting in important advances like novel therapies developments (Hasin et al., 2017) and it is expected that, during the next years, multi-omics studies play an essential role in the understanding of complex diseases. However, these approaches also present many challenges. Novel statistical methods are necessary to perform proper analyses and to extract new knowledge from multi-omics data.

A typical multi-omics study approach is the integration of gene expression and the methylation status, given that the relationship between both is well studied. Methylation in the promoter region of a gene is usually correlated with repression of that gene expression through the inhibition of transcription factors binding and the recruitment of proteins involved in gene repression (Long et al., 2017). A recent example of a successful expression and methylation integration is a work where genes with coherent expression and methylation changes were proposed as colorectal cancer biomarkers (Kerachian et al., 2020).

Another very common multi-omics integration are expression quantitative trait loci (eQTL) studies. In this approach, genomics and transcriptomics data are integrated in order to find genetic variations correlated with changes in gene expression. This technique has been very useful to identify causal variants in genomics studies (Claussnitzer et al., 2015).

Many methods have been proposed to integrate multi-omics data not from only 2 or few technologies, but from many biological layers (Gomez-Cabrero et al., 2019). If the integrated layers have a direct relationship, like methylation and gene expression, it is frequent to perform correlation analyses to find associations between them. More advanced statistical models can be also applied. For instance, mediation analysis is based on regression models and considers one layer as a cause of a disease and another one as a mediator (Sun and Hu, 2016). Genomics and transcriptomics data have been integrated by mediation analysis, considering the gene expression as the mediator between SNPs and the pathology (Yen-Tsung Huang et al., 2014).

Another common strategy is to use prior knowledge of molecular relationships to map and model multi-omics data. These networks can be based, for instance, on metabolic pathways (Jeong et al., 2000), physical interaction between proteins (Behrends et al., 2010), genomic interactions (Vidal et al., 2011) and other physical interactions between biological layers.

Furthermore, there is a group of unsupervised methods that do not rely on previous knowledge. Some popular unsupervised data integration methods are based on matrix factorization, like NMF (Zhang et al., 2011) and iCluster (Shen et al., 2009). There are also Bayesian unsupervised methods, such as Patient-Specific Data Fusion (Yuan et al., 2011), Bayesian Consensus Clustering (Lock and Dunson, 2013) and Multiple Dataset Integration (Kirk et al., 2012). Finally, SNF (Wang et al., 2014) is a popular unsupervised network-based method for clustering patients into subgroups.

2 OBJECTIVES

The main objective of this doctoral thesis is the development and application of novel bioinformatics and statistical methods to address important challenges in the context of biomedical data analysis and data integration. For that aim, the following specific objectives were proposed.

1. Development of a bioinformatics workflow for **genetic association meta-analysis** that implements the required analytical steps. This workflow includes data preparation, Hardy-Weinberg equilibrium testing, heterogeneity tests, genetic models construction, publication bias assessment, subgroup analysis and sensitivity analysis. Implementation of the entire workflow in a web-based application.

2. Implementation of a pipeline to perform **gene expression meta-analysis** with public data. The pipeline includes data download and preparation, quality controls, meta-analysis based on the effect sizes and P-values combination methods and enrichment analysis. Implementation of a web-based interface to launch the meta-analyses and to generate interactive reports.

3. Implementation of a novel algorithm based on Gene Set Enrichment Analysis to detect subtle **differentially methylated regions** overlooked by the current methods. Integration of the differentially expressed regions and the expression of nearby genes to find **multi-omics** coherent signals. Preparation of an R package with the entire workflow.

4. Compilation and curation of the expression and methylation data available in public repositories from five autoimmune diseases to find novel biomarkers and to explore common and unique molecular profiles. Development of a web-based database to share, analyse and integrate the compiled data, including features like networks analysis and meta-analysis.

3 METHODS

The specific methodology employed to accomplish each of the thesis objectives are described in the corresponding articles provided in the Appendix. However, in this section, we explain in more detail some general methods shared by different articles, as well as specific methods that are briefly summarised in the publications.

3.1 META-ANALYSIS

In this thesis, we implemented several meta-analysis methods in order to integrate different types of biomedical data. Concretely, we applied methods based on effect sizes combination to GAS and gene expression meta-analysis. In addition, we also implemented methods based on P-values combination to gene expression data. Finally, we integrated meta-analysis based on ranks combination in our autoimmune diseases database in order to combine the results from several transcriptomics studies.

3.1.1 Meta-analysis based on effect sizes combination

In this approach, the effect (defined as the intensity of a phenomenon) among different studies is compared. The effect size may be calculated differently depending on the type of data analysed. For instance, in genetic association studies, it is common to use the odds-ratio to calculate the effect size (Stringer et al., 2011). On the other hand, in gene expression studies, the effect sizes are calculated from the differential expression between experimental groups, commonly using the Hedge's estimator (Hedges, 1982). Independently of the method used to calculate the effect sizes, the following step is to calculate a combined effect of all the studies for each element we are analyzing (e.g., mutations for genetic studies). The combined effect allows to estimate if the phenomenon is statistically significant or not considering the included studies. It may be calculated using a fixed effect model (FEM) or a random effect model (REM).

FEM is a very conservative method as it assumes a high level of homogeneity among studies (Nakagawa et al., 2017). For this reason, it should be only applied after checking this assumption with methods like the Cochran's Q test or the I^2 statistic (Higgins and Thompson, 2002). The combined effect ($\overline{T}$) in a meta-analysis assuming a FEM is calculated following Equation 1 (Borenstein et al., 2009):

$$\bar{T} = \frac{\sum_{i=1}^{k} \omega_i T_i}{\sum_{i=1}^{k} \omega_i} \quad (1)$$

Where T_i are the individual effects from each study and ω_i are the weights of each study, calculated in Equation 2:

$$\omega_i = \frac{1}{V(T_i)} \quad (2)$$

Being $V(T_i)$ the variance of the effect of each study.

On the other hand, **REM** assumes that the effect sizes is different among the studies, so the combined effect indicates the average of the effect sizes rather than an equal effect size for all the studies like in FEM (Cohn and Becker, 2003). Consequently, REM is a less conservative method than FEM and it commonly represents better the biological reality, where heterogeneity is very common (Waldron and Riester, 2016). The combined effect in REM ($\overline{T}^*$) is calculated likewise FEM's $\overline{T}$ (Equation 3), but the weights in REM (ω_i^*) are calculated from both within-study variance, $V(T_i)$, and the between-study variance, τ^2 (Equation 4).

$$\bar{T} = \frac{\sum_{i=1}^{k} \omega_i^* T_i}{\sum_{i=1}^{k} \omega_i^*} \quad (3)$$

$$\omega_i^* = \frac{1}{V(T_i) + \tau^2} \quad (4)$$

τ^2 may be calculated following Equation 5:

$$\tau^2 \begin{cases} \frac{Q - df}{C} \; if \; Q > df \\ 0 \; if \; Q \; \leq df \end{cases} \tag{5}$$

Where Q is calculated with Equation 6, C is calculated with Equation 7 and df are the degrees of freedom (number of studies – 1).

$$Q = \sum_{i=1}^{k} \omega_i (T_i - \bar{T}) \tag{6}$$

$$C = \sum_{i=1}^{k} \omega_i - \frac{\sum_{i=1}^{k} {\omega_i}^2}{\sum_{i=1}^{k} \omega_i} \tag{7}$$

3.1.2 Meta-analysis based on P-values combination

In this approach, the P-values from the individual studies are combined into a new integrated P-value. In this way, very heterogeneous studies (e.g. with different experimental conditions er analytical platform) can be integrated easily (Sutton et al., 2000). There are different methods to achieve this. The simplest ones are using the maximum (**Wilkinson's method**) or the minimum (**Tippet's method**) P-values among the studies.

On the other hand, **Fisher's method** (Equation 8) can be used to give importance to very significant P-values, due to one single small P-value may result in a significant combined P-value (Heard and Rubin-Delanchy, 2018). **Pearson's method** (Equation 9) is similar, but it is more sensible to large P-values.

$$P = -2 \times \sum_{i=1}^{k} ln(p_i) \quad (8)$$

$$P = -2 \times \sum_{i=1}^{k} ln(1 - p_i) \quad (9)$$

Finally, **Stouffer's method** (Equation 10) may be used.

$$P = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \quad (10)$$

Being Z_i defined in Equation 11.

$$Z_i = \Phi^{-1}(1 - p_i) \quad (11)$$

Where Φ is the normal standard normal cumulative distribution function.

The major advantage of Stouffer's method is that weights can be included (Equation 12) in order to assign different importance to the different studies (e.g., studies with more samples can have higher weights).

$$P = \frac{\sum_{i=1}^{k} \omega_1 Z_i}{\sqrt{\sum_{i=1}^{k} \omega_i^2}} \quad (12)$$

3.1.3 Meta-analysis based on ranks combination

In these methods, a statistic (e.g., fold-change) is used to order the features of each individual dataset obtaining the ranks for each study. Then, the ranks are combined in order to find features with consistent high and low ranks. Ranks combination methods are nonparametric, meaning that none probability distribution is assumed, and they permit to integrate very heterogeneous data. In addition, these methods are robust against outliers. However, if there is diversity of variance among the datasets, the accuracy can be affected (Hong and Breitling, 2008).

Rank Product method consists on multiplying the ranks of each feature among the studies (Equation 13). The P-value is obtained permutating with random ranks, what makes this method computationally demanding (Breitling and Herzyk, 2005) and, therefore, it is only recommended when the amount of included studies is low.

$$RP_g = \prod_i^k r_{ig} \tag{13}$$

A similar method is **Rank Sum** (Equation 14), which is more efficient at the cost of providing less accurate results than Rank Product:

$$RS_g = \sum_i^k r_{ig} \tag{14}$$

3.2 META-ANALYSIS OF GENETIC ASSOCIATION STUDIES

One of the objectives of this thesis was to implement a complete workflow to perform GAS meta-analyses. For that aim, it was first necessary to understand the methodology employed to analyse the individual studies and, then, to implement all the required steps to integrate the results from several individual studies.

3.2.1 Individual GAS methodology

A GAS typically follows a case-control design which compares the frequencies of reference and candidate alleles among the case and control subjects. A simple, yet very common example of a GAS focuses on SNPs. In the simplest scenario, there are two alleles (e.g., A allele, suspected to be linked to a disease and B allele, the reference in the population). Therefore, each subject may be homozygote for the candidate allele (AA), heterozygote (AB) or homozygote for the reference allele (BB). These three possibilities are often summarized in two variables comparisons following a specific genetic model.

For instance, for the dominant genetic model, it is assumed that having one A allele is enough to have the disease. Therefore, the tested comparison is *AA + AB vs BB*. Such comparison may be used to prepare a two-by-two contingency table (see example at Table 1). From the contingency table, four variables can be extracted (*a*, *b*, *c* and *d*) that may be used to calculate the Odds Ratio (OR) following Equation 15.

$$OR = \frac{odds\ disease\ with\ AA\ or\ AB}{odds\ disease\ with\ BB} = \frac{(a/b)}{(c/d)} = \frac{a\ x\ d}{b\ x\ c} \qquad (15)$$

As can be deduced, OR is the odds of having the disease if the subject has a candidate genotype (AA or AB for dominant model) divided by the odds of having the disease if the subject has a reference genotype (BB for dominant model). An OR > 1 indicates A is linked to the disease following dominant model. If OR is close to 1, there is no linkage between the SNP and the disease. An OR < 1 indicates A allele has a protective effect against the disease.

Table 1. Contingency table for a case-control GAS using a dominant genetic model.

	Cases	**Controls**
AA + AB	a	b
BB	c	d

The contingency table may be also used to assess the statistical significance of the association. For that aim, confidence intervals (CI) and P-values may be calculated. Typically, a Fisher's exact test (Fisher, 1934) is used to calculate the P-value under the null hypothesis that OR is 1.

Genetic association to a phenotype may follow different genetic models. Depending on the model, the contingency table will be different and therefore the resulting OR. In addition to the dominant model, the comparisons may follow a recessive model (AA vs AB + BB) or an over-dominant model (AB vs AA + BB). Pairwise comparisons considering only two of the possible genotypes may also be performed (AA vs BB, AA vs AB or AB vs BB). In addition, an allelic model may be used (A vs B), where the linkage of each allele is tested.

It is common that researchers calculate the ORs for different genetic models and choose the one with the highest statistical confidence (Attia et al., 2003; Thakkinstian et al., 2005). However, this approach has been criticized due to the increase of risk of false positives and some authors recommend to choose a genetic model before OR calculation (Bagos, 2013). In any case, if different models are tested, a P-value adjustment for multiple testing should be applied in order to mitigate this risk.

3.2.2 GAS meta-analysis workflow

As previously commented, it is common that published GAS meta-analyses contain some methodological errors. In order to mitigate this, we implemented a complete workflow that guides the users through all the analytical steps. Next, we describe each one of these steps that were implemented in our pipeline.

One common error in this type of meta-analysis is not testing if the controls from each included study are in **Hardy-Weinberg equilibrium** (HWE). HWE is a model used in population genetics which demonstrates mathematically that genotype frequencies are constant among generations unless some event (e.g., natural selection, genetic drift, etc.) alters such frequencies (Edwards, 2008). HWE establish that, for a locus with A and B alleles with frequencies f(A) = p and f(B) = q, the expected frequency for the AA and BB homozygotes and AB heterozygotes are $f(AA) = p^2$, $f(BB) = q^2$ and $f(AB) = 2pq$ respectively.

HWE assumptions are useful in GAS meta-analyses because, theoretically, control subjects should accomplish them (Khoury et al., 2005). Therefore, if HWE is violated in a study, it is very likely that there is some technical problem in that study (e.g., errors during genotyping). That is why HWE in the controls from all the included studies in a GAS meta-analysis should always be tested as a quality control, and studies with controls deviated from HWE should be removed from the analysis. HWE is usually tested with a χ^2 test, a general statistical method to assess if there is a significant difference between the expected and the observed frequencies. A P-value for the χ^2 test may be obtained for each study. A P-value lower than a threshold (commonly 0.05) indicates that there is a deviation from HWE and that study should be discarded from further analyses.

GAS meta-analyses normally apply effect sizes combination methods (FEM or REM models), explained in the previous section. In this case, the effects are the ORs from the different studies. FEM are indicated when there is homogeneity in the analysed data and REM otherwise. However, heterogeneity tests are not always performed and FEM or REM models are applied arbitrarily, what is another major error in a GAS meta-analysis. There are several methods to assess heterogeneity, but the most common in this context are the I^2 estimator and the Cochran's Q test (Whitehead and Whitehead, 1991), as well as a χ^2 test to get a P-value.

As said in the previous section, different genetic models may be used to calculate the ORs of a GAS. This is also the case when several GASs are meta-analysed, so the same genetic models have to be applied to all the studies in order to get comparable ORs.

A very useful representation in these meta-analyses is a **Forest plot**. In this plot, the effect sizes with the CI are shown for each study, as well as the weights assigned and the number of subjects in each group of the contingency table for cases and controls. Heterogeneity metrics may also be included in these plots. Therefore, this is a very convenient plot to summarize the results of a GAS meta-analysis. An example of a Forest plot can be observed in Figure 2 of the MetaGenyo article (Section 7.1).

Another important step of these meta-analyses is to test for **publication bias** in the selected studies. Publication bias may occur because significant positive results have more chances to be published and cited than negative results (Hopewell et al., 2009). This

bias can be very problematic for a meta-analysis or systematic review due to they can be skewed towards positive results.

The main way to check for the presence of publication bias in the studies is a visual inspection of a **funnel plot**. In a funnel plot, the estimated effects (e.g., OR) of the individual studies are plotted against some precision or size measure of the studies, often the standard error (SE). (Sterne et al., 2011). The SE is plotted on the Y axis with a reverse scale, so those studies with lower SE are placed at the top of the graphic. If there is not bias, the points should shape an inverse funnel, with the studies with higher SE and varying ORs dispersed at the bottom, and the studies with lower SE and less disperse ORs concentrated at the top. An asymmetric funnel plot may indicate publication bias, as well as other methodological problems like inadequate studies design, errors in the analyses or artefacts (Egger et al., 1997). In order to avoid relying on a subjective interpretation of the funnel plots, Egger *et al*. proposed a test for funnel plots asymmetry based on regression analysis (Egger et al., 1997; Sterne et al., 2000).

If there are covariates in the data like race, country and so on, it is a good practice to perform not only a general meta-analysis with all the studies, but also **subgroup meta-analyses** including only the studies from specific groups. This approach may be useful to discover associations specific to a population that may be overlooked in a general analysis.

Finally, it is very recommended to perform a **sensitivity analysis** in order to verify that the meta-analysis results are not excessively influenced by one or few very significant studies. A common strategy to assess this is the leave-one-out method, consisting of repeating the meta-analysis n times (being n the number of included studies), but excluding one study in each repeat (Viechtbauer and Cheung, 2010). If one of the studies is influencing too much on the results, it becomes evident through this method because when that study is excluded, the results will be much less significant (e.g., higher P-value or ORs closer to 1).

3.3 META-ANALYSIS OF GENE EXPRESSION DATA

A gene expression meta-analysis also has some particularities compared to other types of data. In order to implement a workflow to perform gene expression meta-analyses, we took into account these particularities and included analytical steps to address them.

On the first place, a **quality control** of each included study should be performed in order to detect outliers and other technical problems (Toro-Domínguez et al., 2020). Outlier detection is usually performed checking the similarity between samples with techniques like principal component analysis, correlation or clustering (Shieh and Hung, 2009). In addition, missing values are a common problem in this kind of data, given that a reliable signal is not always obtained for all the genes and samples and some microarrays contain probes only for a fraction of the genes. This problem can be handled imputing the missing values using the mean expression of the gene in the non-missing samples or using similarity models like K-nearest neighbor (Aittokallio, 2010; Liew et al., 2011).

The next issue in gene expression meta-analysis is the **gene annotation**. Given the diversity of platforms and annotation databases available, it is common that the different studies do not share the same gene annotation. Therefore, it is necessary to collapse the gene names to common identifiers (e.g., gene symbol). Furthermore, it is common that different rows of the data matrices are measures of the same gene (for instance, some probes in microarrays target different regions of the same gene). For that reason, the different values referring to the same genes should be summarized with some strategies like obtaining the mean or median value, or choosing the row with the highest mean value (Miller et al., 2011). Another common problem, also due to the heterogeneity in platforms and gene annotations, is that some genes may be present in some of the included studies and not in others. Typically, the solution is to select the genes shared by all the datasets, although some new approaches are being proposed in order to avoid the loss of data caused by such strategy (Bobak et al., 2020).

Once the meta-analysis is performed using one of the methods described in Section 3.1, a list of differentially expressed genes across the studies is obtained. The most common representation of the results is a **heatmap** with the expression values of significant genes. Furthermore, a functional analysis may be used with the aim of identifying pathways, cellular functions, etc. enriched with significant genes from the meta-analysis.

3.4 DIFFERENTIAL METHYLATION ANALYSIS

With the aim of developing a novel algorithm for differential methylation analysis with data generated with Illumina BeadChip platforms, it was important to understand how these data are generated and processed and the current strategies to perform these

analyses. Furthermore, in this section we also introduce Gene Set Enrichment Analysis, which is the main method we employed to develop our new algorithm.

3.4.1 BeadChip data processing

In order to convert the fluorescence signals measured by the Illumina microarrays to proper methylation data for statistical analyses, there are a set of preprocessing steps that should be performed. Most of the available software for these tasks are R packages available in Bioconductor, like *minfi* (Aryee et al., 2014), *RnBeads* (Assenov et al., 2014) and *wateRmelon* (Pidsley et al., 2013).

Raw data of Illumina BeadChips platforms are *idat* files containing the fluorescence intensities at each microarray spot. These raw *idat* files may be processed to obtain a β-value for each interrogated site using Equation 16.

$$\beta = \frac{M}{M + U + 100} \tag{16}$$

Being M the methylated probe intensity and U the unmethylated probe intensity. As can be deduced from the previous formula, β is a number between 0 and 1. 0 indicates that all interrogated copies for a CpG site are unmethylated, while 1 indicates that all copies are methylated.

On the other hand, methylation can be also quantified with the M-values, obtained with Equation 17.

$$Mv = \log_2\left(\frac{M}{U}\right) = \text{logit}(\beta) \tag{17}$$

β-values are normally more useful to do graphical representations due to they are very easy to interpret. However, M-values are generally more suitable to perform statistical analyses given that they better accomplish with the statistical assumptions of the EWAS

parametric methods and it has been demonstrated that they provide a better detection ratio of differentially methylated sites and more true positives than β-values (Du et al., 2010).

It is essential to **normalize** the data in order to correct it from systematic bias introduced by the technology itself. Normalization allows the data to represent with maximum fidelity the biological features which generated it. Some common normalization methods are functional normalization (Fortin et al., 2014), normal-exponential out-of-band (Noob) correction (Triche et al., 2013), Subset-quantile Within Array Normalization (SWAN) (Maksimovic et al., 2012) and Beta-Mixture Quantile (BMIQ) (Teschendorff et al., 2013).

In order to improve sensitivity and specificity, it is important to **filter** the methylation data. A useful first filtering step is to use the detection P-values calculated by some packages, like *minfi*, for each CpG site. This detection P-value is obtained comparing the methylated and unmethylated DNA signals at each site to the background signal measured from control probes. Probes with detection P-value > 0.01 are not trusty and it is recommended to remove them from the analyses. In addition, it is recommended to discard those CpGs located in the sex chromosomes if both male and female samples are included in the study (Maksimovic et al., 2017), as well as probes containing SNPs (Naeem et al., 2014) and those which are cross-reactive (Chen et al., 2013).

Once the methylation data is processed, a typical analysis in an EWAS is to obtain the **differentially methylated positions** (DMPs) and the **differentially methylated regions** (DMRs) between the experimental groups.

3.4.2 DMPs detection with linear models

The most common strategy to calculate the DMPs is to apply linear models with packages like *limma* (Ritchie et al., 2015). As previously commented, it is recommended to use normalized M-values rather than β-values for this analysis. The models can be fitted, in addition to the explanatory variable (e.g., experimental group), with different covariates that could act as potential confounding factors in the data, both from technical (e.g., sample plate) and biological (e.g., age or gender) origin.

3.4.3 DMRs analysis

With the DMPs identification, the probability of each CpG site to be differentially methylated between different groups is calculated. However, methylation patterns are not usually found in isolated CpGs. Instead of that, clusters of proximal CpGs are

hypermethylated or hypomethylated (Peters et al., 2015). That is the reason why several methods have been designed to detect DMRs instead of individual differentially methylated CpGs. There are two main approaches to analyse DMRs: to search DMRs in **predefined** regions (e.g., promoters or CGIs) or to search DMRs ***de novo***, without relying on previous annotations of the genome.

Among the predefined methods, some of the most used tools are Illumina Methylation Analyzer (*IMA*) (Wang et al., 2012), *RnBeads* (Assenov et al., 2014) and City of Hope CpG Island Analysis Pipeline (*COHCAP*) (Warden et al., 2013) R packages. *IMA* and *COHCAP* methods average the methylation values in the predefined regions and tests for differential methylation using these values. On the other hand, *RnBeads* rely on *limma* results, aggregating the P-values obtained by the linear modelling by the predefined regions.

On the other hand, *de novo* methods like *DMRcate* (Peters et al., 2015), *bumphunter* (Jaffe et al., 2012) or *Probe Lasso* (Butcher and Beck, 2015) use different algorithms to inspect the whole genome searching for DMRs.

3.4.4 Gene Set Enrichment Analysis

Gene-set analysis (GSA) was developed in the context of expression data analysis with the aim of overcoming methodological limitations in standard analyses, like the lack of reproducibility or the fact that several phenotypes cause a limited change in a set of genes rather than big alterations in one or few genes (Ein-Dor et al., 2006). There are many GSA algorithms, each one with different biological and statistical assumptions.

Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) is one of the first developed GSA methods and still nowadays it is one of the most popular approaches. It uses a given metric (e.g., fold-change or t statistic) to rank all the measured genes and applies a weighted Kolmogorov–Smirnov statistic (Hollander and Wolfe, 1999) to calculate an Enrichment Score (ES). Basically, ES for each set is calculated running through the entire ranked list increasing the score when a gene in the set is encountered and decreasing the score when the gene encountered is not in the set analysed. ES of this set is the maximum difference from 0. Significance of each ES is calculated permuting the sets and recomputing ES, getting a null distribution for the ES. Nominal P-value is calculated relative to the computed null distribution. Once the nominal P-values for all

sets have been computed, GSEA adjust them for multiple hypothesis testing by a very conservative Family-Wise Error Rate method and a less restrictive FDR method.

Although GSEA algorithm has been widely used, it has some limitations regarding its efficiency, especially due to the calculation of null distributions step. For this reason, fast gene set enrichment analysis (FGSEA) method was proposed as a faster implementation of GSEA algorithm (Korotkevich et al., 2019). This new algorithm was implemented in the Bioconductor package *fgsea* (Korotkevich et al., 2019).

4 RESULTS

4.1 METAGENYO: GENETIC ASSOCIATION STUDIES META-ANALYSIS

As previously commented in Section 1.4, although the amount of GAS meta-analysis is constantly growing, many of them contain methodological errors that detract from their usefulness. That is, in part, because of the absence of a dedicated software for this specific type of meta-analysis. Researchers, commonly without deep statistical knowledge or programming skills, use software intended for general meta-analysis. However, these software do not guide users through all the necessary steps of a GAS meta-analysis. As a consequence, one or more steps described in the previous section are frequently skipped.

With the aim of providing a tool to improve this situation, we developed **MetaGenyo** (Martorell-Marugan et al., 2017), a web application that guides users through all the meta-analysis steps and permits to perform these analyses in an intuitive and easy way. MetaGenyo is available at http://bioinfo.genyo.es/metagenyo and, until February 2021, it has been accessed almost 8,000 times and the paper has 49 accumulated citations according to Google Scholar. One of the most recent applications of MetaGenyo is a work which associated the T allele of the SNP rs35705950 to a protective effect against COVID-19 (Moorsel et al., 2020).

In this article (provided in Section 7.1), in addition to present MetaGenyo, we reviewed previously published software packages to perform GAS meta-analysis, providing a comparative summary of all of them. Furthermore, we carried out a meta-analysis of A23G SNP of XPA gene and digestive cancers proposed in (He et al., 2015) using MetaGenyo. We found a new association between this SNP and a protective effect of the heterozygous genotype against esophageal cancer under the codominant genetic model. This significant association was overlooked by the authors of the original meta-analysis given that they did not test the overdominant genetic model.

This first article of the doctoral thesis was the first step towards the accomplishment of the thesis objectives since the published software permits to exploit public biomedical data (in this case, data from genetic studies) with an interface usable by any user without advanced statistical or bioinformatics knowledge.

4.2 IMAGEO: GENE EXPRESSION META-ANALYSIS

The amount of gene expression meta-analysis publications has been growing during the recent years thanks to centralized repositories like GEO that facilitates the obtention of transcriptomics data. Although there were several tools to perform gene expression meta-analysis, many of them require advanced programming skills to download and preprocess the data and to apply the proper meta-analysis methods. There were some interactive tools that facilitated these analyses. However, these tools lacked some useful features, like the possibility to input GEO identifiers to automatically download and analyse the corresponding datasets.

For those reasons, we developed **ImaGEO**, a web tool that permits to introduce GEO datasets identifiers to, automatically, download and perform a meta-analysis with the corresponding datasets. Furthermore, users can also upload expression data not available in GEO in order to integrate them with other studies. The article describing ImaGEO was published at Bioinformatics journal (Toro-Domínguez, Martorell-Marugán et al., 2019). In this work, in addition to show the ImaGEO features, we showed the ImaGEO potential performing a meta-analysis of lung cancer and Alzheimer's disease searching for opposite gene expression patterns. We found several genes deregulated in opposite directions, as was expected from these disorders given their inverse comorbidity.

This second article fits into the scope of the thesis given that it presents an interactive tool to exploit the large amount of transcriptomics data available in the public repository GEO.

4.3 MCSEA: DMRS ANALYSIS AND INTEGRATION WITH GENE EXPRESSION

Although there are many available tools to detect DMRs from methylation arrays data, all of them share the characteristic that they are optimized to detect DMRs where the differential methylation is very marked between experimental conditions. This condition is true in diseases with severe dysregulation of methylation status, being the better example cancer (De Smet et al., 1999). However, this is not always the case, especially in the context of complex diseases, where the differential methylation may be much more subtle, but still have a role in the pathogenesis of those disorders. Some examples are found in hypertension or schizophrenia patients (Guerrero-Bosagna et al., 2014).

Given that available EWAS methodologies have focused on detecting large methylation differences between phenotypes, there is a lack of bioinformatics tools designed to find small methylation changes in complex phenotypes, so, when this kind of study is performed, there are usually a lot of false negatives in the results or even non-significant results at all (Bohlin et al., 2015; Chiavaroli et al., 2015; Gervin et al., 2012). This lack of proper tools to perform such kind of EWAS motivated us to develop a new approach based on GSEA. We called our approach **mCSEA** (for methylated CpGs Set Enrichment Analysis) and it is capable to detect subtle but consistent methylation differences in predefined genomic regions. In addition, we developed a module to integrate the mCSEA results with gene expression data in order to find DMRs correlated with expression alterations in surrounding genes.

We implemented mCSEA in an R package which we submitted to the Bioconductor repository. An article presenting this tool was published in the Bioinformatics journal (Martorell-Marugán et al., 2019). In this article, we simulated methylation data with differential methylation in some regions. We simulated different levels of differential methylation, from very marked to very subtle difference. We compared the performance of mCSEA with other available tools for these simulated data, obtaining better results for mCSEA not only when the methylation difference is small, but also when it is large, being a suitable method for both scenarios. In addition, we reanalysed previously published data from children exposed to maternal diabetes (Kim et al., 2017), finding differentially methylated promoters in genes related to metabolic disorders, as was expected from this experimental design.

This work fulfills the thesis scope given that, although a minimum of R language knowledge is required to use mCSEA, it still does not require advanced programming skills in order to run their functions. Furthermore, the module of methylation and expression integration allows to perform multi-omics analyses, what is another of the thesis objectives.

4.4 ADEx: Autoimmune diseases database

The lack of a data repository dedicated to ADs motivated us to develop **ADEx** (for Autoimmune Diseases Explorer), a database with public omics data processed homogeneously that facilitates the reuse of those data in retrospective studies. The

database is accessible through a web tool with visualization and analysis options, including differential gene expression analysis, networks analysis and meta-analysis, among others. We benefitted from the experience acquired during the development of MetaGenyo and ImaGEO to build the ADEx application in the same framework. Furthermore, the meta-analysis methods developed for ImaGEO and the multi-omics integration included in mCSEA were incorporated in ADEx.

Using ADEx, we explored the IFN signature in all the expression studies, confirming its heterogeneity between diseases and tissues. Furthermore, we proposed a set of candidate biomarkers for each disease using the meta-analysis integrated in ADEx. The article describing ADEx is currently under review in BMC Bioinformatics journal, and a preprint is available at the bioRxiv repository (Martorell-Marugán et al., 2020).

This work fits into all the main purposes of this thesis, given that ADEx integrates public biomedical data in an easy-to-use database, and it also allows to integrate expression and methylation data from the same samples.

5 CONCLUSIONS

In this doctoral thesis, new computational and statistical methods for the analysis and integration of biomedical data have been developed. Specifically, the main conclusions of this thesis are the following:

1. MetaGenyo web tool allows to perform meta-analysis of genetic association studies rigorously. It is the first web-based application for this specific type of meta-analysis. The utility of MetaGenyo was demonstrated finding a novel association between the polymorphism rs1800975 of XPA gene and the esophageal cancer incidence.

2. The developed application ImaGEO includes state-of-the-art methods to integrate heterogeneous gene expression studies. Its main novelty lies in the automatic data acquisition and processing directly from the GEO repository.

3. mCSEA is a new algorithm for detecting differentially methylated regions based on the Gene Set Enrichment Analysis method. This algorithm outperforms other available tools both when the methylation differences are large and when they are very subtle. Using mCSEA, differentially methylated regions in siblings discordant to maternal diabetes exposure during their gestation were detected. It is possible to integrate the methylation analyses results with gene expression data with a function included in the mCSEA package.

4. Expression and methylation data available from 5 autoimmune diseases have been compiled in the ADEx database. This is the first omics database dedicated to autoimmune diseases. ADEx permits exploring the genes profiles along many studies. It also includes functionalities for multi-omics integration. Interferon signature was studied with ADEx, uncovering a differential behavior depending on the disease and the tissue. Using the ADEx integrated meta-analyses, it is possible to obtain lists of genes with similar gene expression alterations across studies. These genes are potential biomarkers for these diseases.

6 CONCLUSIONES

En la presente tesis doctoral, se han desarrollado nuevos métodos computacionales y estadísticos para el análisis e integración de datos biomédicos. Concretamente, las principales conclusiones de esta tesis son las siguientes:

1. La herramienta web MetaGenyo permite realizar meta-análisis de estudios de asociación genética rigurosamente. Es la primera aplicación web para este tipo específico de meta-análisis. La utilidad de MetaGenyo se demostró encontrando una nueva asociación entre el polimorfismo rs1800975 del gen XPA y la incidencia del cáncer de esófago.

2. La aplicación desarrollada ImaGEO incluye métodos punteros para integrar estudios de expresión génica heterogéneos. Su principal novedad radica en la adquisición y procesamiento automáticos de datos directamente desde el repositorio GEO.

3. mCSEA es un nuevo algoritmo para la detección de regiones diferencialmente metiladas basado en el método Gene Set Enrichment Analysis. Este algoritmo supera a otras herramientas disponibles tanto cuando las diferencias de metilación son grandes como cuando son muy sutiles. Usando mCSEA, se detectaron regiones diferencialmente metiladas en hermanos discordantes a exposición de diabetes materna durante su gestación. Es posible integrar los resultados de los análisis de metilación con datos de expresión génica con una función incluida en el paquete de mCSEA.

4. Se han compilado datos de expresión y metilación disponibles para 5 enfermedades autoinmunes en la base de datos ADEx. Esta es la primera base de datos ómicos dedicada a enfermedades autoinmunes. ADEx permite explorar los perfiles de genes a lo largo de muchos estudios. También incluye funciones para la integración de multi-ómicas. Con ADEx se estudió la firma del interferón, revelando un comportamiento diferente dependiendo de la enfermedad y el tejido. Usando el meta-análisis integrado de ADEx, es posible obtener listas de genes con alteraciones en su expresión similares entre estudios. Estos genes son biomarcadores potenciales para estas enfermedades.

7 APPENDIX. ARTICLES

7.1 METAGENYO: A WEB TOOL FOR META-ANALYSIS OF GENETIC ASSOCIATION STUDIES

This article was published at BMC Bioinformatics journal Volume 18, 563, December 16, 2017. (https://doi.org/10.1186/s12859-017-1990-4) under an Open Access license. This is the accepted version of the article. According to the publisher (BMC), this article version has unrestricted reuse permission.

MetaGenyo: A web tool for meta-analysis of genetic association studies

Jordi Martorell-Marugan[1], Daniel Toro-Dominguez[1,2], Marta E. Alarcon-Riquelme[2,3] and Pedro Carmona-Saez[1*]

[1] Bioinformatics Unit and [2] Medical Genomics. Centre for Genomics and Oncological Research (GENYO), Granada, Spain. [3] Institute for Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

* Correspondence: pedro.carmona@genyo.es

Abstract

Background: Genetic association studies (GAS) aims to evaluate the association between genetic variants and phenotypes. In the last few years, the number of this type of study has increased exponentially, but the results are not always reproducible due to experimental designs, low sample sizes and other methodological errors. In this field, meta-analysis techniques are becoming very popular tools to combine results across studies to increase statistical power and to resolve discrepancies in genetic association studies. A meta-analysis summarizes research findings, increases statistical power and enables the identification of genuine associations between genotypes and phenotypes. Meta-analysis techniques are increasingly used in GAS, but it is also increasing the amount of published meta-analysis containing different errors. Although there are several software packages that implement meta-analysis, none of them are specifically designed for genetic association studies and in most cases their use requires advanced programming or scripting expertise.

Results: We have developed MetaGenyo, a web tool for meta-analysis in GAS. MetaGenyo implements a complete and comprehensive workflow that can be executed in an easy-to-use environment without programming knowledge. MetaGenyo has been developed to guide users through the main steps of a GAS meta-analysis, covering Hardy-Weinberg test, statistical association for different genetic models, analysis of heterogeneity, testing for publication bias, subgroup analysis and robustness testing of the results.

Conclusions: MetaGenyo is a useful tool to conduct comprehensive genetic association meta-analysis. The application is freely available at http://bioinfo.genyo.es/metagenyo/.

Keywords: Genetic association study, Meta-analysis, Web tool, Shiny

Background

Genetic association studies (GAS) estimate the statistical association between genetic variants and a given phenotype, usually complex diseases [1]. In the last few years, the number of genetic association studies has increased exponentially, but the results are not consistently reproducible. This lack of reproducibility may be influenced by several factors, including the analysis of non-heritable phenotype, inappropriate quality control, wrong statistical analysis, low sample size, population stratification, incorrect multiple-testing correction or technical biases [2].

Meta-analysis is a statistical technique for combining results across studies and it is becoming very popular as a method for resolving discrepancies in GAS. It summarizes research findings, increases statistical power and enables the identification of genuine associations [3]. In this context, in 2011 there was a 64-fold increase in genetics-related meta-analysis compared to 1995 [4].

Despite the increasing number of publications in this field there is a lack of dedicated software tools to perform a complete GAS meta-analysis in a friendly environment. In this context, most published works in the field have used commercial software suites such as STATA [5] or SPSS [6]. These are statistical software packages that include general functions for meta-analysis in their configuration. In addition, freely available R packages such as meta [7] or metafor [8] are also widely used but all these solutions share common limitations: do not provide all required steps for a GAS meta-analysis (e.g. evaluating Hardy Weinberg equilibrium (HWE) or genetic models) and require advanced statistical or bioinformatics knowledge to be properly used.

In this context, Park et al. have reported several analytical errors in published GAS meta-analysis [9], many of them could be avoided using a dedicated software for GAS meta-analysis with predefined functions and automatic computations of the required statistics.

Here we present MetaGenyo, an easy-to-use web application which implements a complete meta-analysis workflow for GAS. Once the data has been loaded, it provides a guided and complete workflow that comprises the main steps in GAS meta-analysis, including HWE test, checking heterogeneity, publication bias indicators, statistical association testing for different genetics models, subgroup analysis and robustness testing. The use of MetaGenyo does not require advanced statistical or bioinformatics

knowledge and we hope it will be a useful application for researchers working in the field of genetic association studies.

Implementation

MetaGenyo has been implemented as a web tool using shiny [10], a web application framework for RStudio [11]. Backend computations are carried out in R using available packages and custom scripts. MetaGenyo provides the following functionalities:

Testing HWE

Departures from HWE can occur due to genotyping errors, selection bias and stratification [12]. Therefore, goodness-of-fit of HWE should be checked in each study before pooling data. HardyWeinberg package [13,14] is used to compute a P-value for each study in the control population in order to identify low-quality studies. As we test for HWE in several studies, the obtained P-values are corrected by Benjamini and Hochberg false discovery rate (FDR) [15].

Genetic Models

Given two alleles (A, a) the three possible genotypes (AA, Aa, aa) can be dichotomized in different ways yielding different genetic models. GAS can be carried out assuming a specific genetic model based on biological criteria but in most of the cases different models are simultaneously evaluated. MetaGenyo performs meta-analysis in several ways [16], including allele contrast (A vs. a), recessive (AA vs. Aa + aa), dominant (AA + Aa vs. aa) and overdominant (Aa vs. AA + aa) genetic models as well as pairwise comparisons (AA vs. aa, AA vs Aa and Aa vs aa). All P-values are adjusted for multiple testing with the Bonferroni method [17].

Statistical analysis and Heterogeneity

To perform meta-analysis, MetaGenyo combines the effect sizes of the included studies by weighting the data according to the amount of information in each study. Association values are calculated based on two different statistic models: Fixed Effects Model (FEM) and Random Effects Model (REM). The choosing between both models depends on the

amount of heterogeneity in the data, which is also evaluated with heterogeneity indicators such as I^2 and Cochran's Q test (see on-line help of the program). Meta package (7) is used to get such heterogeneity indicators and association results. Finally, this same package is used to generate Forest plots to summarize information for effect size and the corresponding 95% confidence interval (CI) of each study and the pooled effect. Forest plots can be generated for FEM, REM or both, and can be downloaded with very high resolution.

Publication Bias

Publication bias occurs because of meta-analysis are performed using published studies, which usually report only significant associations, while studies showing no significant results tend to remain unpublished. This may therefore give a falsely skewed positive result. To test for publication bias, MetaGenyo provides funnel plots and Egger's test [16] for each genetic model. Funnel plots are generated with meta package [7] and Egger's test is performed using the metafor package [8].

Subgroup Analysis

MetaGenyo provides a subgroup analysis in order to evaluate associations in a subset of studies based on the user defined criteria (e.g. studies from the same country). Many genetic associations are population-specific and can be undiscovered in a general meta-analysis, but discovered when studies are split. For each group, a meta-analysis is performed with FEM or REM, depending on the heterogeneity test: If heterogeneity P-value < 0.1, REM will be used. Otherwise, FEM will be used instead. These results are downloadable in Excel and text formats.

Sensitivity Analysis

In order to test the robustness of the meta-analysis performed, MetaGenyo performs a leave-one-out influence analysis using meta package [7]. Briefly, the meta-analysis is repeated several times, each time excluding one of the studies, in order to determine how each individual study affects the overall statistics [18]. A forest plot with these results is generated for the selected genetic model.

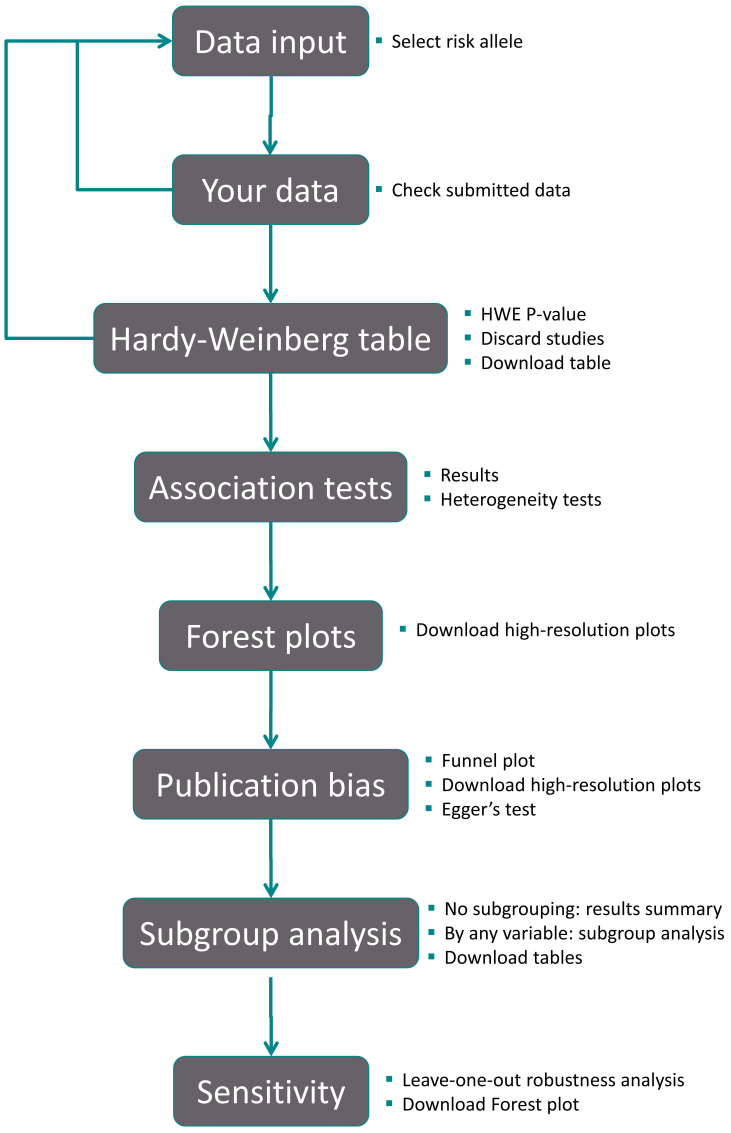


Figure 1. Overview of MetaGenyo. The scheme represents the tool's workflow. First, data is uploaded by the user and it can be reviewed. Secondly, HWE P-values are calculated, so users can decide to exclude some bad-quality samples and reupload their data. In Association tests, Forest plots, Publication bias and Subgroup analysis tabs users can download the meta-analysis results. Finally, users can check the sensitivity analysis.

Software Usage

An overview of MetaGenyo is provided in the on-line help of the application and Figure 1. First, the user loads the collected data from individual studies as a text or excel file with some specifications on the file format. Once the data has been loaded, a complete analysis is performed providing results and visualizations in different tabs: (1) The data tab, where the user can check if the data has been correctly submitted. (2) Hardy-Weinberg tab, where a HWE P-value column is added to the data. (3) Association values tab. This contains different association values and heterogeneity indicators for each

genetic model. (4) Forest plot tab contains forest plot visualizations in high-quality image format for each genetic model. (5) Publication bias tab, where the user can see the funnel plot and Egger's test results. (6) Subgroup analysis tab to obtain a summary of the analysis or to evaluate the association and heterogeneity results taking into account stratification based on user-defined variables and, finally, (7) Sensitivity tab to perform a robustness analysis.

Results and discussion

Despite there are many programs designed to perform GWAS meta-analysis (reviewed in [19]), there is a lack of tools specially designed to perform GAS meta-analysis, so researchers use general statistical or meta-analysis software, adapting it to the particular purposes in such type of meta-analysis. This lack of dedicated software increases the required resources to perform a GAS meta-analysis, facilitates the inclusion of methodological errors and requires advance bioinformatics expertise.

Table 1. Characteristics of available meta-analysis software.

	STATA	SPSS	MIX	MetaEasy	meta	rmeta	metafor	MetaGenyo
USABILITY								
Availability	Commercial	Commercial	Commercial[a]	Free[b]	Free	Free	Free	Free
Web-based	No	No	No	No	No	No	No	Yes
Operating system	Windows, Mac OS, Linux	Windows, Mac OS, Linux	Windows	Windows	Windows, Mac OS, Linux	Windows, Mac OS, Linux	Windows, Mac OS, Linux	Any[c]
Guided workflow	No	No	No	No	No	No	No	Yes
Programming knowledge	Yes[d]	Yes[d]	No	No	R language	R language	R language	No
FUNCTIONALITIES								
Specific for GAS meta-analysis	No	No	No	No	No	No	No	Yes
HWE testing	Yes	No	No	No	No	No	No	Yes
Heterogeneity assessment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Random/Fixed effect models	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Forest plot	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Automatic testing of genetic models	No	No	No	No	No	No	No	Yes
Publication bias	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Subgroup analysis	Yes	No	Yes	No	Yes	No	Yes	Yes
Robustness analysis	Yes	No	Yes	No	Yes	No	Yes	Yes
P-value correction for multiple testing	Yes	Yes	No	No	No	No	No	Yes

Notes: (a) There is a MIX free version with reduced capabilities. **(b)** MetaEasy is free, but it depends on the proprietary software Microsoft Excel. **(c)** MetaGenyo is accessed through an internet browser, so there are no limitations regarding the operating system used to access it. **(d)** Although STATA and SPSS are command-based software, there are Graphical user interfaces (GUIs) available which permits replacing scripting by user-friendly interactive commands.

Among the most widely used software solutions in this field are STATA [5], SPSS [6] and SAS [20]. These are popular software suites that provide a set of statistical functions that can be used in a broad range of applications and data analysis problems, but they are proprietary software and are not specialized in GAS meta-analysis. These limitations are partially overcome by R packages such as meta [7], rmeta [21] and metafor [8]. These are freely-available software libraries to perform a complete meta-analysis in a flexible way. However, their use requires R programming skills, they do not provide a guided workflow and they are not specifically designed to perform GAS meta-analysis. In addition, there are some Excel extensions such as MIX [22] and MetaEasy [23]. These extensions are easy to use, but they require the usage of the proprietary software Microsoft Excel.

In this context, MetaGenyo is a user-friendly web application that implements a complete meta-analysis following a guided workflow, which does not require programming knowledge. Table 1 contains a summary of the main advantages and disadvantages of some reviewed GAS meta-analysis software.

To demonstrate the functionality of MetaGenyo we have used data from a published GAS meta-analysis [24]. In this study, the authors performed a meta-analysis to study the association between the A23G SNP of XPA gene (rs1800975) and digestive cancers. They collected genotype information from 18 case-control studies including 4170 patients and 6929 controls in total. In this polymorphism, the G allele was considered the reference, so the A allele was the risk allele (this parameter must be specified in MetaGenyo). Results from the complete analysis and a comparison with results reported in the original article can be found in Additional file 1.

Briefly, both sets of results are highly concordant, but in the original publication the authors did not correct the P-values for multiple testing or evaluated different genetic models as provided by MetaGenyo. In this context, we also found some discrepancies between both sets of results due to use of inappropriate statistical tests or labeling mistakes, specially at the subgroup analysis step (see Additional File 1). Because MetaGenyo automatically performs all meta-analysis steps in a guided analysis we reduced these potential sources of errors. All these similarities and differences are detailed in Additional file 1.

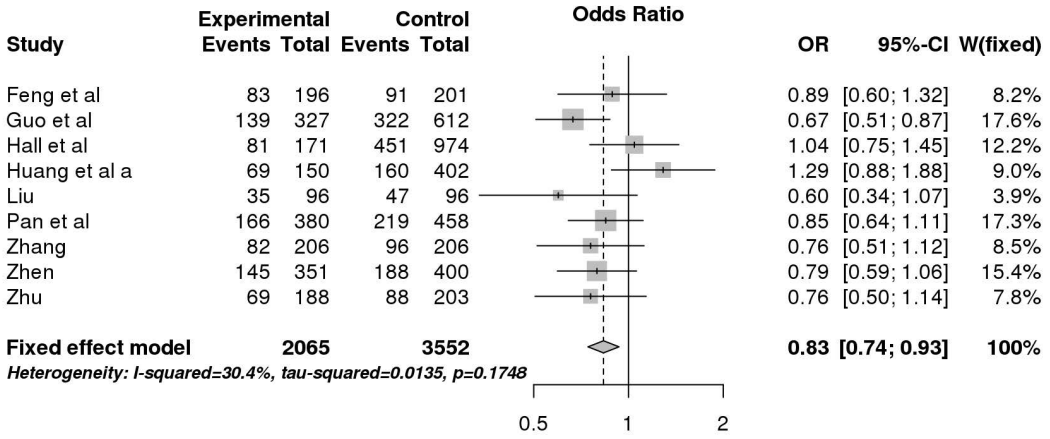


Figure 2. Forest plot with overdominant model and FEM statistics of esophageal cancer data generated with MetaGenyo. The tested comparison is AG vs. AA + AG (overdominant model) and FEM was used.

The application generated results for all possible genetic models and allowed us to easily evaluate results for different subgroups in a unified framework. In this context, using the tumor type feature to stratify the data revealed a significant association for the overdominant model in esophageal cancer studies not previously reported (OR=0.84, 95% CI=0.73-0.96, P-value=0.0016, Bonferroni-adjusted P-value=0.0448) [Figure 2]. Although the original work reported no significant association between this polymorphism and the risk of any type of digestive cancer for the studied models, there may be a protective effect of AG genotype against the risk of esophageal tumors overlooked at the original article because the authors did not test this genetic model. Indeed, a similar association has been found in another GAS meta-analysis with lung cancer samples [25].

Conclusions

In this work, we present MetaGenyo, a free easy-to-use web tool to perform GAS meta-analysis. It provides a guided workflow through the most important steps of a meta-analysis.

We demonstrated MetaGenyo's functionality replicating a previously published meta-analysis [24]. In addition, thanks to the automatic testing of several genetic models and subgroup analysis we found a significant association between rs1800975 SNP in XPA

gene and esophageal cancer under the overdominant genetic model that may be interesting enough for further testing.

Surprisingly, there is a large heterogeneity in statistical methods, lack of quality control steps or misleading reporting and interpretation of results in many published meta-analysis [9]. Therefore, an application such as MetaGenyo will be a very useful tool for the research community providing a guided and solid workflow.

Availability

Project name: MetaGenyo

Availability: MetaGenyo web tool, example datasets and help are accessible at http://bioinfo.genyo.es/metagenyo/.

Any restrictions on use by academics: none

List of abbreviations

CI: Confidence intervals; FDR: False discovery rate; FEM: Fixed effect model; GAS: Genetic association study; GUI: Graphical user interface; GWAS: Genome-wide association study; HWE: Hardy-Weinberg equilibrium; OR: Odds-ratio; REM: Random effect model; χ^2: Goodness-of-fit chi-square.

Declarations

Acknowledgements

We thank Alberto Ramirez and Manuel Martínez for helpful technical assistance.

Funding

JMM is partially supported by Ministerio de Economía y Competitividad [grant number PEJ 2014-A-95230].

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Author's contributions

PCS conceived the project and directed the software development. JMM designed the software and performed the analysis. DTD and MAR tested the software, provided improvements and test cases. PCS and JMM wrote the manuscripts. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

References

1. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet. 2001 Feb;2(2):91–9.

2. Li A, Meyre D. Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. Int J Obes 2005. 2013 Apr;37(4):559–67.

3. Trikalinos TA, Salanti G, Zintzaras E, Ioannidis JPA. Meta-analysis methods. Adv Genet. 2008;60:311–34.

4. Ioannidis JPA, Chang CQ, Lam TK, Schully SD, Khoury MJ. The geometric increase in meta-analyses from China in the genomic era. PloS One. 2013;8(6):e65602.

5. StataCorp. Stata Statistical Software. College Station, TX: StataCorp LP; 2015.

6. IBM Corp. SPSS Statistics for Windows. Armonk, NY: IBM Corp; 2016.

7. Schwarzer G. meta: An R Package for Meta-Analysis. R News. 2007;7(3):40–6.

8. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1–48.

9. Park JH, Eisenhut M, van der Vliet HJ, Shin JI. Statistical controversies in clinical research: overlap and errors in the meta-analyses of microRNA genetic association studies in cancers. Ann Oncol Off J Eur Soc Med Oncol. 2017 Jun 1;28(6):1169–82.

10. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R [Internet]. 2017. Available from: https://CRAN.R-project.org/package=shiny

11. RStudio Team. RStudio: Integrated Development for R [Internet]. Boston, MA: RStudio Inc; 2015. Available from: http://www.rstudio.com/

12. Salanti G, Amountza G, Ntzani EE, Ioannidis JPA. Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. Eur J Hum Genet EJHG. 2005 Jul;13(7):840–8.

13. Graffelman J. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. J Stat Softw. 2015;64:1–22.

14. Graffelman J, Camarena JM. Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum Hered. 2008;65(2):77–84.

15. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc. 1995;57(1):289–300.

16. Attia J, Thakkinstian A, D'Este C. Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. J Clin Epidemiol. 2003 Apr;56(4):297–303.

17. Bonferronin C. E. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze; 1936.

18. Viechtbauer W, Cheung MW-L. Outlier and influence diagnostics for meta-analysis. Res Synth Methods. 2010 Apr;1(2):112–25.

19. Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res. 2012 May;40(9):3777–84.

20. SAS Institute Inc. SAS. Cary, NC: SAS Institute Inc; 2011.

21. Lumley T. rmeta: Meta-analysis [Internet]. 2012. Available from: https://CRAN.R-project.org/package=rmeta

22. Bax L, Yu L-M, Ikeda N, Tsuruta H, Moons KG. Development and validation of MIX: comprehensive free software for meta-analysis of causal research data. BMC Med Res Methodol. 2006 Oct 13;6:50.

23. Kontopantelis E, Reeves D. MetaEasy: A Meta-Analysis Add-In for Microsoft Excel. J Stat Softw. 2009;30(7).

24. He L, Deng T, Luo H. XPA A23G polymorphism and risk of digestive system cancers: a meta-analysis. OncoTargets Ther. 2015;8:385–94.

25. Liu X, Lin Q, Fu C, Liu C, Zhu F, Liu Z, et al. Association between XPA gene rs1800975 polymorphism and susceptibility to lung cancer: a meta-analysis. Clin Respir J. 2016 Jul 27;

Additional files

File name: Additional file 1

File format: .pdf

Title of data: MetaGenyo's use case Description of data: Document showing the results of analyzing the data provided by [24] using MetaGenyo and comparison with the original results

Additional file 1

Here we provide a use case and a detailed analysis of results from the reanalysis of data reported by He et al. [1].

The input file for MetaGenyo was a text file containing the original data of the meta-analysis and reformatted to accomplish with the MetaGenyo input format [Supplementary table 1].

The application guides the user across the statistical functions that should be used in the analysis, allowing non-expert users to perform a complete meta-analysis covering all the required steps.

P-values for HWE were calculated in controls and adjusted P-values greater than 0.05 indicated that the study fits with HWE conditions [Supplementary table 2]. The unadjusted P-values are the same that those calculated by the original authors [Table 1 in the original paper]. However, as the analysis comprises several studies, it is important to adjust P-values for multiple testing in order to reduce false positives, which were not calculated by the original authors. MetaGenyo corrects P-values by FDR method.

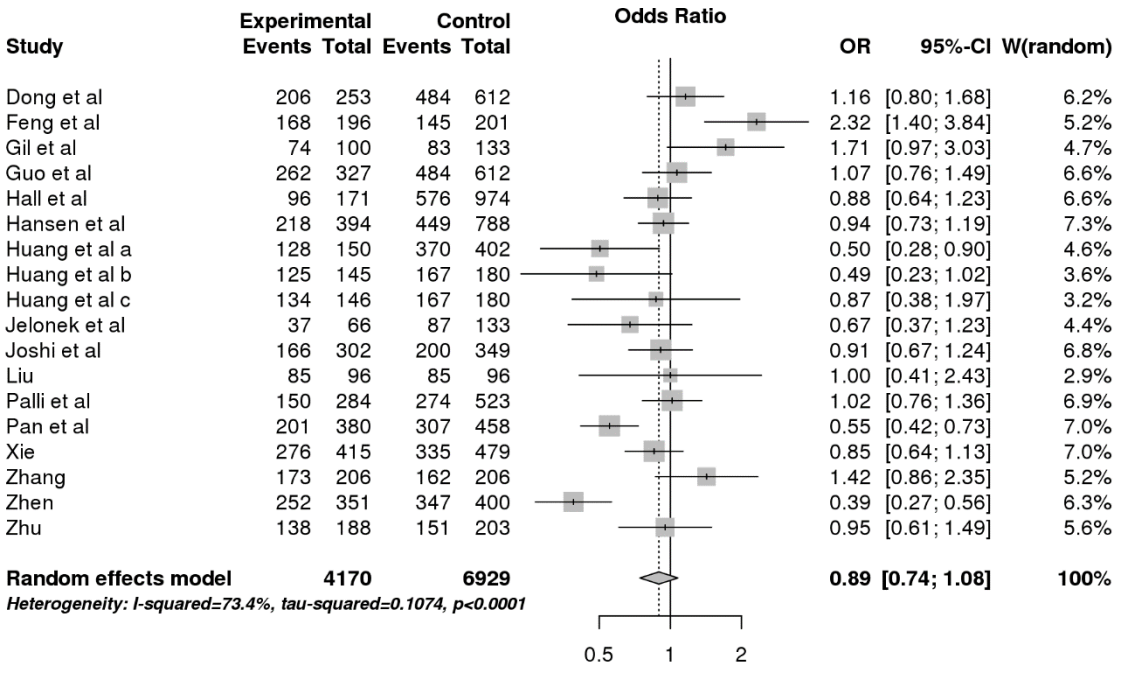
Supplementary Table 1. Input table for the example meta-analysis.

Author	Ethnicity	Tumor type	Source of control	GG cases	GA cases	AA cases	GG control	GA control	AA control
Dong et al	Asian	Gastric	PB	47	120	86	128	322	162
Feng et al	Asian	Esophageal	HB	28	83	85	56	91	54
Gil et al	Caucasian	Colorectal	HB	26	58	16	50	67	16
Guo et al	Asian	Esophageal	PB	65	139	123	128	322	162
Hall et al	Caucasian	Esophageal	HB	75	81	15	398	451	125
Hansen et al	Caucasian	Colorectal	PB	176	187	31	339	359	90
Huang et al a	Asian	Esophageal	PB	22	69	59	32	160	210
Huang et al b	Asian	Cardiac	PB	20	60	65	13	55	112
Huang et al c	Asian	Gastric	PB	12	57	77	13	55	112
Jelonek et al	Caucasian	Colorectal	HB	29	33	4	46	70	17
Joshi et al	Caucasian	Colorectal	PB	136	133	33	149	170	30
Liu	Asian	Esophageal	PB	11	35	50	11	47	38
Palli et al	Caucasian	Gastric	PB	134	115	35	249	215	59
Pan et al	Caucasian	Esophageal	HB	179	166	35	151	219	88
Xie	Asian	Hepatocellular	PB	139	203	73	144	219	116
Zhang	Asian	Esophageal	HB	33	82	91	44	96	66
Zhen	Asian	Esophageal	PB	99	145	107	53	188	159
Zhu	Asian	Esophageal	PB	50	69	69	52	88	63

Abbreviations: PB=population-based; HB=hospital-based.

Supplementary Table 2. P-value and adjusted P-value by FDR of χ^2 test for HWE in control samples of each study.

Author	HWE P-value	HWE adjusted P-value
Dong et al	0.1694	0.4064
Feng et al	0.1806	0.4064
Gil et al	0.3686	0.6032
Guo et al	0.1694	0.4064
Hall et al	0.8751	0.8751
Hansen et al	0.7312	0.8751
Huang et al a	0.8434	0.8751
Huang et al b	0.0966	0.3478
Huang et al c	0.0966	0.3478
Jelonek et al	0.2252	0.4088
Joshi et al	0.0558	0.3478
Liu	0.5353	0.7412
Palli et al	0.2271	0.4088
Pan et al	0.5893	0.7577
Xie	0.0711	0.3478
Zhang	0.4116	0.6174
Zhen	0.8259	0.8751
Zhu	0.0631	0.3478



Supplementary figure 1. Forest plot with dominant genetic model and REM statistics generated with MetaGenyo.

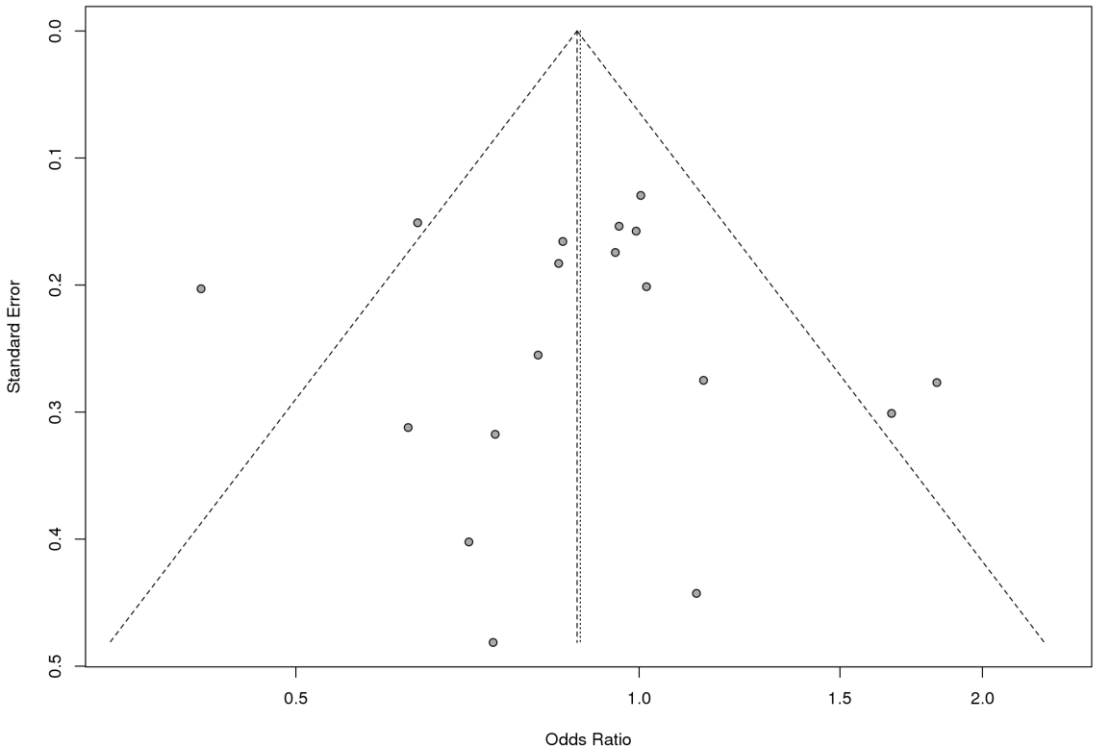
In the next step, statistical associations were evaluated for different genetics models. In the original work, the authors evaluated four different genetic models: dominant model (AA + AG vs. GG), recessive model (AA vs. AG + GG), heterozygote comparison (AG vs. GG) and homozygote comparison (AA vs. GG). All those comparisons can be performed with MetaGenyo, in addition to allele contrast (A vs. G), overdominant model (AG vs. AA + GG) and (AA vs. AG) comparison. A forest plot was obtained to summarize the results applying REM statistics with dominant genetic model [Supplementary Figure 1], as the original authors did [Figure 2 in the original paper]. We noticed that some individual statistics are slightly different between the original forest plot and those reported by MetaGenyo. After comparing the forest plot reported in the original publication and the data that was used to generate it, we realized that some labels were exchanged in this plot (e.g. Guo et al. label actually contains the data from Gil et al. study). This mislabeling caused the discrepancies between both forest plots.

Supplementary table 3. MetaGenyo's subgroup analysis results with dominant genetic model and splitting the samples by ethnicity.

Ethnicity	Test of association			Test of heterogeneity		
	OR	95 % CI	P-value	Model	I^2	P-value
Overall	0.8940	[0.7426; 1.0762]	0.2365	Random	0.7339	0.0001
Asian	0.8968	[0.6625; 1.2139]	0.4807	Random	0.7823	0.0001
Caucasian	0.8809	[0.7059; 1.0993]	0.2619	Random	0.6626	0.0068

Supplementary table 4. MetaGenyo's subgroup analysis results with dominant genetic model and splitting the samples by tumor type.

Tumor type	Test of association			Test of heterogeneity		
	OR	95 % CI	P-value	Model	I^2	P-value
Overall	0.8940	[0.7426; 1.0762]	0.2365	Random	0.7339	0.0001
Colorectal	0.9549	[0.8025; 1.1362]	0.6029	Fixed	0.4521	0.1401
Esophageal	0.8668	[0.6102; 1.2314]	0.4249	Random	0.8395	0.0001
Gastric	1.0527	[0.8449; 1.3117]	0.6471	Fixed	0.0000	0.7701

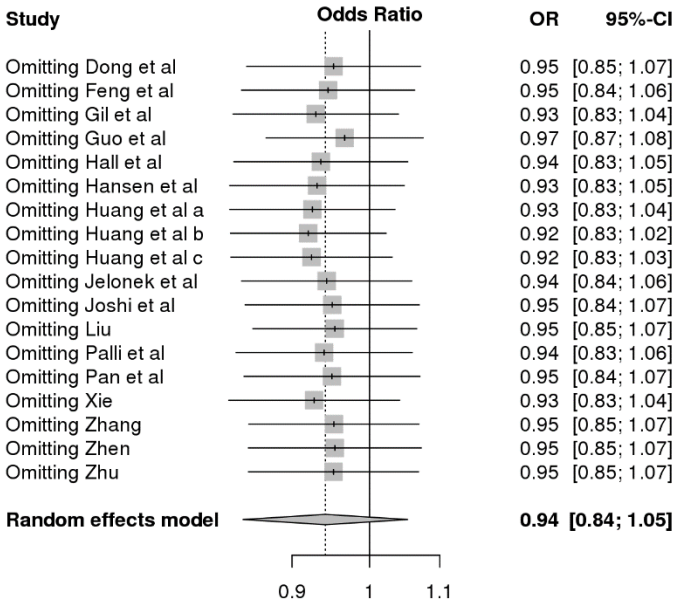


Supplementary figure 2. Funnel plot of AG vs. GG comparison generated by MetaGenyo.

However, in gastric tumor samples in which we did not found agreement, we observed that, for some reason, original authors included samples from cardiac cancer in the gastric cancer group, causing these discrepancies with MetaGenyo results.

A funnel plot was also generated, revealing that there was not publication bias in the data (see Supplementary figure 2). This MetaGenyo output is very similar to the previously published one [Figure 5 in the original paper], except the x and y axes are the opposite between both figures and MetaGenyo use the OR as the x axis, while the original plot uses the log(OR).

Finally, a sensitivity analysis was performed with MetaGenyo generating a forest plot of the results excluding one of the studies in each step [Supplementary figure 3] revealing that the results were not biased by any single study from those originally included in the work. The original authors performed the same sensitivity analysis and reached the same conclusions, but they did not include a forest plot of such analysis.



Supplementary figure 3. Forest plot of sensitivity analysis under overdominant model and REM statistics.

References

1. He L, Deng T, Luo H. XPA A23G polymorphism and risk of digestive system cancers: a meta-analysis. Onco Targets Ther. 2015;8:385–94.

2. Ried K. Interpreting and understanding meta-analysis graphs--a practical guide. Aust Fam Physician. 2006;35(8):635–8.

7.2 ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database

This article was published at Bioinformatics journal Volume 35, Issue 5, 01 March 2019, Pages 880–882 (https://doi.org/10.1093/bioinformatics/bty721). This is the accepted version of the article. According to the publisher (Oxford University Press), this article version has Non-Commercial reuse permission after a 12 months embargo finished on March 1, 2020.

ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database

Daniel Toro-Domínguez[1,2,†], Jordi Martorell-Marugán[1,†], Raúl López-Domínguez[1], Adrián García-Moreno[1], Víctor González-Rumayor[3], Marta E. Alarcón-Riquelme[2,4,*] and Pedro Carmona-Sáez[1,*]

[1] Bioinformatics Unit, [2] Area of Medical Genomics, GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, PTS, 18016, Granada, Spain. [3] Atrys Health, Barcelona, Spain [4] Institute of Environmental Medicine, Karolinska Institute, Stockholm, Sweden.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Abstract

Summary: The Gene Expression Omnibus (GEO) database provides an invaluable resource of publicly available gene expression data that can be integrated and analyzed to derive new hypothesis and knowledge. In this context, gene expression meta-analysis is increasingly used in several fields to improve study reproducibility and discovering robust biomarkers. Nevertheless, integrating data is not straightforward without bioinformatics expertise. Here, we present ImaGEO, a web tool for gene expression meta-analysis that implements a complete and comprehensive meta-analysis workflow starting from GEO dataset identifiers. The application integrates GEO datasets, applies different meta-analysis techniques and provides functional analysis results in an easy-to-use environment. ImaGEO is a powerful and useful resource that allows researchers to integrate and perform meta-analysis of GEO datasets to lead robust findings for biomarker discovery studies.

Availability: ImaGEO is accessible at http://bioinfo.genyo.es/imageo/

Contact: marta.alarcon@genyo.es or pedro.carmona@genyo.es

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Due to the increasing use of high-throughput techniques, the amount of information available in biomedical databases is growing exponentially. In particular, Gene Expression Omnibus (GEO) (Barrett et al., 2013) is a public gene expression repository that contains more than 94000 datasets and over 2 million of samples. This is an invaluable resource that, with the appropriate methods and tools, can be exploited to integrate gene expression data for applications such as biomarker discovery (Toro-Domínguez et al., 2014), disease classification or phenotype comparisons (Carmona-Sáez et al., 2017), among others. Several software tools have been developed to take advantage of this information. GEO2R was originally available in GEO portal to allow researchers without computational skills to perform differential expression analysis in individual datasets. In the last few years, tools such as ShinyGEO (Dumas et al., 2016) or ScanGEO (Koeppen et al., 2017) have extended some of the GEO2R functionalities to explore, retrieve and analyze gene expression data in an easy-to-use environment.

However, this amount of data offers new possibilities beyond the analysis of individual datasets. In this context, there is an increasing number of studies that integrate different datasets to perform gene expression meta-analysis (geMAs). This technique is usually applied to increase the sample size but it can be also used to integrate datasets from different phenotypes in order to discover common biomarkers (Toro-Domínguez et al., 2014). In this context, there are different tools for geMAs such as INMEX (Xia et al., 2013) or ExAtlas (Sharov et al., 2015), but they lack from a complete workflow starting from GEO identifiers that, at the same time, requires a minimal user interaction in terms of data processing. A detailed comparative analysis of available tools is provided in the additional material.

In this work, we present ImaGEO, a web-based application to perform a complete geMAs starting from GEO identifiers. The application provides a step-by-step workflow that guides the user through the entire analysis accelerating the re-use of publicly available gene expression data for biomarker discovery purposes. The application currently supports a curated set of platforms from Illumina, Affymetrix and Agilent for human and model organisms including Saccharomyces cerevisiae, Drosophila melanogaster, Danio rerio and Mus musculus and others such as Rattus norvegicus and Pseudomonas aeruginosa (see the online help with complete list of supported platforms).

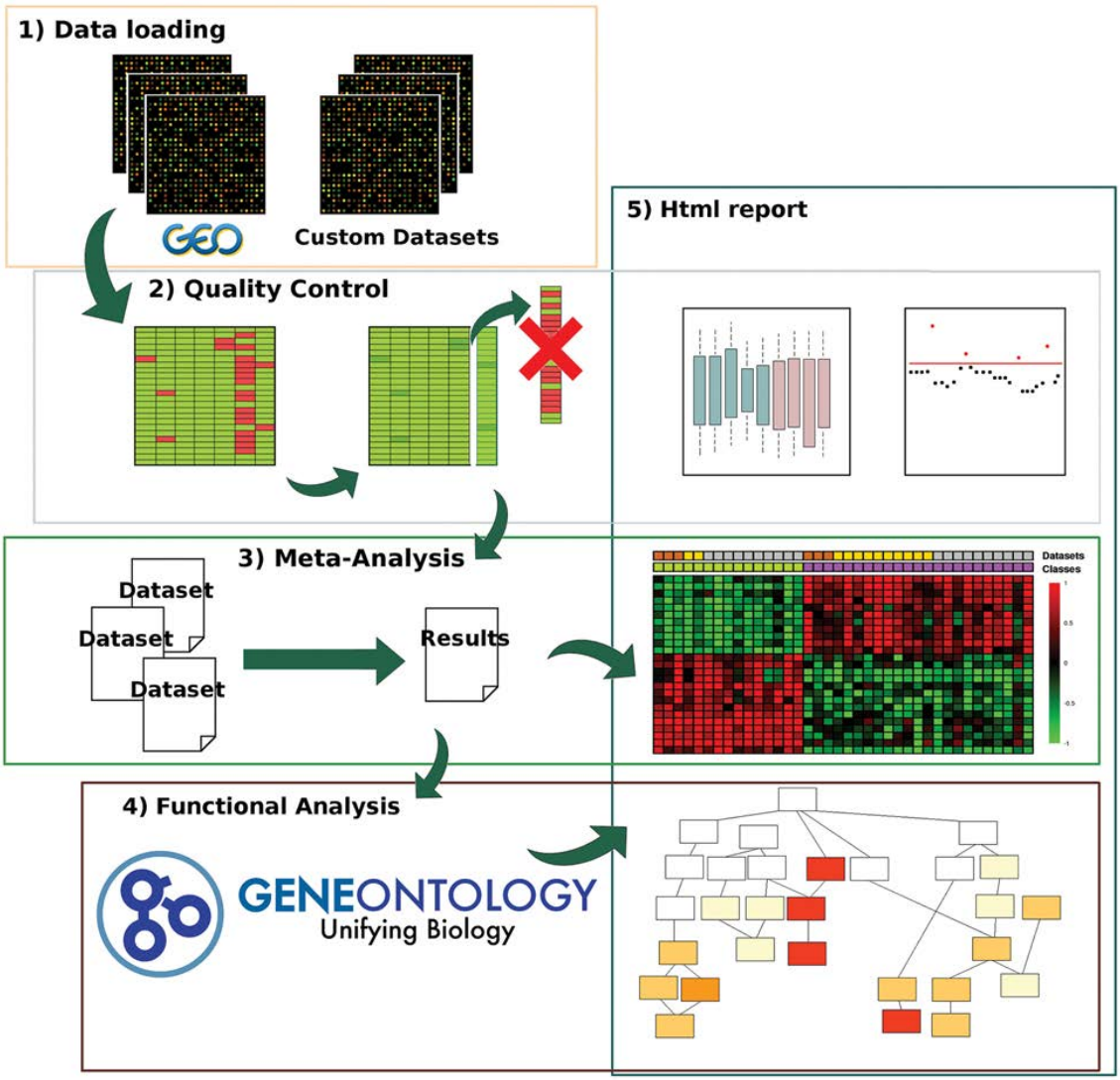


Figure 1: Workflow of ImaGEO. The image summarizes and orders the five modules of ImaGEO. **1)** Data input from GEO or custom data. **2)** Quality Control is performed for each dataset followed by sample/gene filtering. **3)** Gene expression Meta-analysis. **4)** Functional analysis 5) Results in html report.

2 Methods

ImaGEO has been developed in Shiny, a web application framework for R. Internally, it is divided in 5 modules (Figure 1): (1) Data loading and processing module where users can enter the GEO IDs or upload custom datasets and establish the parameters of the analysis. GEO data is retrieved and processed using GEOquery package (Davis and Meltzer, 2007). Expression values are transformed to logarithmic unless they already are and probe identifiers are annotated to unique gene identifiers. (2) Quality control module that shows data metrics and quality control checking. (3) Meta-analysis module contains

a total of 9 different meta-analysis methods adapted from functions contained in MetaDE R package, which are effect size (ES), Fisher's, Stouffer's, adaptively-weighted, sum or product of ranks and the selection of minimum or maximum p-value across results. (4) Functional analysis module. Enrichment analysis of Gene Ontology terms is performed in the list of over- and under-expressed genes obtained in the meta-analysis. (5) Report Module: where a html report is generated to explore all results using Nozzle.R1 R package. The report is divided in four sections that summarize the results of each step. First, a summary section contains an overview of the data and the analysis parameters used. Secondly, the quality control study shows the distribution of the expression values in boxplots and the missing values for each dataset along with a comparison pre and post quality control. Thirdly, the results section displays an interactive table with significant genes annotated with Gene Symbol identifiers, gene names, p-values, corrected p-values, fold-change values. In addition, heatmaps of top 100 and all significant genes are available. Finally, if the user chooses the enrichment analysis its results are provided in table format. A detailed documentation of methods and results can be found in the application web site.

3 Case study

As a working example we provide a use case that identify genes deregulated in opposite directions among lung cancer (LC) and Alzheimer (AD), two diseases that display inverse co-morbidity according to epidemiological data (Sánchez-Valle et al., 2017). This is another type of application of geMAs that can be easily conducted in ImaGEO and can be applied to analyze inverse gene expression patterns among phenotypes, for example for drug repurposing analysis. As input we used Alzheimer (GEO IDs: GSE5281, GSE48350 and GSE4757) and lung cancer datasets (GEO IDs: GSE33532, GSE19188, GSE19804, GSE7670 and GSE10072). To detect genes that were deregulated in opposite directions we simply switched group labels (cases/controls) in both diseases. Therefore, selecting effects sizes (ES) and Random Effects Model as meta-analysis options we obtained 997 genes that were over-expressed in AD and under-expressed in LC (AD+/LC-) and 1220 genes with opposite patterns (AD-/LC+). Similarly to the results reported by Sanchez-Valle et al, functional analysis of AD+/LC- genes yielded biological pathways related to inflammatory responses and processes associated to AD-/LC+ genes

were related to synaptic transmission and mitochondrial activities, that the authors stated could be implicated in the inverse co-morbidity between these diseases. This analysis was executed in a few minutes and is a good example of the potential of ImaGEO to perform a comprehensive geMAs. We are confident that it will be a useful application for the research community to exploit and re-use GEO data for deriving new biological knowledge and hypothesis generation.

Acknowledgements

This work is part of the thesis of JMM and DTD in the doctorate program of Biomedicine of the University of Granada.

Funding

This work has been partially supported by Junta de Andalucía (PI-0173-2017).

Conflict of Interest: none declared.

References

Barrett,T. et al. (2013) NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res., 41, D991-995.

Carmona-Sáez,P. et al. (2017) Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. Bioinforma. Oxf. Engl., 33, 3691–3695.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics, 23, 1846–1847.

Dumas,J. et al. (2016) shinyGEO: a web-based application for analyzing gene expression omnibus datasets. Bioinformatics, 32, 3679–3681.

Koeppen,K. et al. (2017) ScanGEO: parallel mining of high-throughput gene expression data. Bioinforma. Oxf. Engl., 33, 3500–3501.

Sánchez-Valle,J. et al. (2017) A molecular hypothesis to explain direct and inverse co-morbidities between Alzheimer's Disease, Glioblastoma and Lung cancer. Sci. Rep., 7.

Sharov,A.A. et al. (2015) ExAtlas: An interactive online tool for meta-analysis of gene expression data. J. Bioinform. Comput. Biol., 13, 1550019.

Toro-Domínguez,D. et al. (2014) Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. Arthritis Res. Ther., 16, 489.

Xia,J. et al. (2013) INMEX—a web-based tool for integrative meta-analysis of expression data. Nucleic Acids Res., 41, W63–W70.

Additional file

Additional Table 1 contains a summary of the main features of available web tools for Gene Expression Meta-analysis.

	NetworkAnalyst (INMEX)	**Exatlas**	**Gemma**	**CancerMA**	**ImaGEO**
Ref	(Xia, Gill, & Hancock, 2015)	(Sharov, Schlessinger, & Ko, 2015)	(Zoubarev et al., 2012)	(Feichtinger, McFarlane, & Larcombe, 2012)	
Guided workflow	Yes	No	No	No	Yes
Programming/Bioinformatics knowledge	Partial (data formatting and annotating)	No	No	No	No
Meta-analysis Techniques	Effect sizes, P-values, Rank products and Vote counts	Effect sizes, z-score, and Fisher's methods(Pairwise comparison)	P-value	P-value	Effect sizes, Fisher's, Stouffer's, adaptively-weighted, sum or product of ranks, minimum or maximum p-value
Group assignation	Manually or by programming	Manually, one by one	Automatically	No	Manually
Data input	Dataset formatted	GEO IDs/ Custom data	GEO IDs and set of genes or GO terms	Set of genes	GEO IDs/Custom data
Platform supported	47 built-in microarray probe ID libraries for human, mouse, and rat. For other IDs or organisms, make sure all datasets have the same ID type.	All GEO platforms that contain gene symbol references into GEO files	413 platforms	HG-U133 Plus 2	94 platforms
Species	16	43	10	Human	Human, mouse, rat, *Escherichia coli*, *Pseudomona aeruginosa*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Danio rerio* and *Saccharomyces cerevisiae*
QC	No	Inner dataset normalization, filtering genes and samples	Gene assignment, detect outliers and inner dataset normalization	Inner dataset normalization	Inner dataset normalization, outliers detection, preprocessing and filtering by genes and samples
Functional Analysis	Protein-Protein interactions, Gene-miRNA interactions, TF-gene interactions, Protein-Drug interactions, Protein-chemical interactions, KEGG, Reactome, GO	KEGG, phenotypes, GO	Annotation by gene	GO	GO

References

Feichtinger, J., McFarlane, R. J., & Larcombe, L. D. (2012). CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data. *Database : The Journal of Biological Databases and Curation*, *2012*(February), bas055. https://doi.org/10.1093/database/bas055

Sharov, A. A., Schlessinger, D., & Ko, M. S. H. (2015). ExAtlas: An interactive online tool for meta-analysis of gene expression data. *Journal of Bioinformatics and Computational Biology*, *13*(6), 1550019. https://doi.org/10.1142/S0219720015500195

Xia, J., Gill, E. E., & Hancock, R. E. W. (2015). NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature Protocols*, *10*(6), 823–844. https://doi.org/10.1038/nprot.2015.052

Zoubarev, A., Hamer, K. M., Keshav, K. D., Luke Mccarthy, E., Santos, J. R. C., Van rossum, T., … Pavlidis, P. (2012). Gemma: A resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, *28*(17), 2272–2273. https://doi.org/10.1093/bioinformatics/bts430

7.3 mCSEA: Detecting subtle differentially methylated regions

This article was published at Bioinformatics journal Volume 35, Issue 18, 15 September 2019, Pages 3257–3262 (https://doi.org/10.1093/bioinformatics/btz096). This is the accepted version of the article. According to the publisher (Oxford University Press), this article version has Non-Commercial reuse permission after a 12 months embargo finished on September 15, 2020.

mCSEA: Detecting subtle differentially methylated regions

Jordi Martorell-Marugán[1,2], Víctor González-Rumayor[2] and Pedro Carmona-Sáez[1,*]

[1] Bioinformatics Unit. GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, Granada, Spain. [2] Atrys Health, Barcelona, Spain.

*To whom correspondence should be addressed.

Abstract

Motivation: The identification of differentially methylated regions (DMRs) among phenotypes is one of the main goals of epigenetic analysis. Although there are several methods developed to detect DMRs, most of them are focused on detecting relatively large differences in methylation levels and fail to detect moderate, but consistent, methylation changes that might be associated to complex disorders.

Results: We present mCSEA, an R package that implements a Gene Set Enrichment Analysis method to identify differentially methylated regions from Illumina 450K and EPIC array data. It is especially useful for detecting subtle, but consistent, methylation differences in complex phenotypes. mCSEA also implements functions to integrate gene expression data and to detect genes with significant correlations among methylation and gene expression patterns. Using simulated datasets we show that mCSEA outperforms other tools in detecting DMRs. In addition, we applied mCSEA to a previously published dataset of sibling pairs discordant for intrauterine hyperglycemia exposure. We found several differentially methylated promoters in genes related to metabolic disorders like obesity and diabetes, demonstrating the potential of mCSEA to identify differentially methylated regions not detected by other methods.

Availability: mCSEA is freely available from the Bioconductor repository.

Contact: pedro.carmona@genyo.es

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

DNA methylation is by far the most studied epigenetic mark. It affects gene expression and has an important role in several disorders. Epigenome-wide association studies (EWAS) are performed to find associations between DNA methylation alterations and a given phenotype (Flanagan, 2015).

There are several methodologies to determine DNA methylation status, including high-throughput techniques such as whole-genome bisulfite sequencing (WGBS) or methylation arrays. WGBS is the one with the highest coverage but Illumina's BeadChip arrays (Infinium HumanMethylation450 and Infinium MethylationEPIC) are still much more affordable and simpler to analyze, and they are currently the most used platforms in human EWAS (Teh et al., 2016).

EWAS are usually applied to find associations between individual CpG sites and outcomes. However, methylation patterns are not usually found in isolated CpGs, but clusters of proximal CpGs are hypermethylated or hypomethylated (Peters et al., 2015). That is the reason why several methods have been designed to detect differentially methylated regions (DMRs) instead of differentially methylated positions (DMPs). In this context, some methods use predefined regions as candidates for DMRs identification (e.g. gene promoters or CpG Islands), while others do not rely on previous annotations and search de novo DMRs.

There are two different paradigms related to DNA methylation pointed out in a recent review by Leenen et al. (Leenen et al., 2016). The first one is that, in some disorders such as cancer, regulatory regions are clearly hypermethylated or hypomethylated, with methylation differences greater than 60 % (see for example De Smet et al., 1999; Mikeska and Craig, 2014). However, there is a second paradigm in which complex disorders are associated with very subtle differences in CpGs methylation, with methylation differences of 1-10 % between phenotypes. As remarked by Leenen et al. these subtle methylation differences are relevant hallmarks associated to the diversity of many complex non-malignant diseases, such as type 2 diabetes, major depression, schizophrenia, hypertension, and cardiovascular diseases (see for example Levenson, 2010; Guerrero-Bosagna et al., 2014).

Nevertheless, most of the available DMR methods have focused on detecting large methylation differences between phenotypes. In this context, they have worked properly and they have allowed the discovery of many epigenetic causes of several diseases (Lappalainen and Greally, 2017). However, these tools may fail to detect significant DMRs in complex diseases or heterogeneous phenotypes, where there might be small differences among methylation signals but consistent across the analyzed regions and samples. Therefore, no individual CpGs or regions may meet the threshold for statistical significance in many published studies, although there may be biologically meaningful differences (see for example Bohlin et al., 2015; Chiavaroli et al., 2015; van Dongen et al., 2015; Gervin et al., 2012; Kim et al., 2017).

In addition, some of these tools average all sites in a given region, but if a significant pattern is associated to a subset of sites it may be underestimated if all sites are analyzed as a block.

This scenario motivated us to develop a new approach based on Gene-Set Enrichment analysis (GSEA) (Subramanian et al., 2005), a popular methodology for functional analysis that was specifically designed to avoid some related drawbacks in the field of gene expression. GSEA is able to detect significant gene sets that exhibit strong cross-correlation when differential expression of individual genes is modest from the statistical point of view. GSEA uses a given statistical metric to rank all genes of a genome and applies a weighted Kolmogorov–Smirnov (KS) statistic (Hollander and Wolfe, 1999) to calculate an Enrichment Score (ES). Basically, ES for each set is calculated running through the entire ranked list increasing the score when a gene in the set is encountered and decreasing the score when the gene encountered is not in the analyzed set. ES of this set is the maximum difference from 0. The significance of each ES is calculated permuting the sets and recomputing ES, getting a null distribution for the ES.

We have developed a new R package in which we have implemented a GSEA-based differential methylation analysis where gene sets are defined as sets of CpG sites in predefined regions. This new tool, named mCSEA (methylated CpGs Set Enrichment Analysis), is capable to detect subtle but consistent methylation differences in predefined genomic regions from 450K and EPIC microarrays data. The R package is freely available in Bioconductor repository.

2 Materials and methods

2.1 mCSEA workflow

mCSEA R package consists of five main functions (Figure 1). The first step is to rank all the CpG probes by differential methylation. As input, a presorted list can be used, but if a matrix of β-values or M-values is provided the rankProbes() function applies limma (Ritchie et al., 2015) to fit a linear model and return the t-statistic assigned to each CpG site.

The main mCSEA function, mCSEATest(), evaluates the enrichment of CpG sites belonging to the same region in the top positions of the ranked list by applying the GSEA implementation of the fgsea package (Sergushichev, 2016). Regions whose CpG sites are over-represented in the top or bottom of the list can be detected as differentially methylated. As predefined regions, mCSEA allows users to perform analysis based on promoters, gene bodies and CpG-islands (CGIs). These predefined regions were defined based on R annotation packages IlluminaHumanMethylation450kanno.ilmn12.hg19 and IlluminaHumanMethylationEPICanno.ilm10b2.hg19 for 450K and EPIC arrays respectively. We defined each region as shown in Table 1, following previous works (Sandoval et al., 2011). In addition, researchers can provide a set of defined regions in the analysis by providing a file with genomic positions.

mCSEATest() function provides different statistics for each analyzed region, including a P-value of the regions to be differentially methylated, a P-value adjusted by false discovery rate (FDR) and the ES. In addition, a Normalized Enrichment Score (NES) is calculated in order to correct the bias for the different region sizes. The necessity and implementation of NES was explained in the original GSEA's paper (Subramanian et al., 2005).

mCSEA package include two functions to visualize the results: mCSEAPlot() and mCSEAPlotGSEA(). The former represents methylation values of a given region in its genomic context (see Figure 3 (A) for an example). The latter generates GSEA's enrichment plot (see Figure 3 (C) for an example), showing the positions of the CpG in a determined region along the entire ranked list.

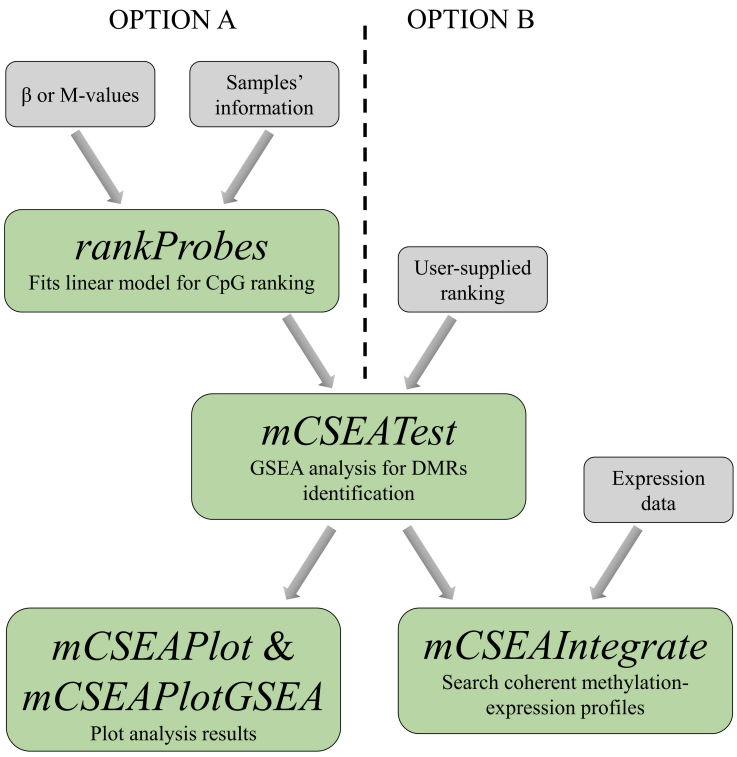


Figure 1. mCSEA workflow. Grey boxes are input data and green boxes are mCSEA's functions. The scheme also shows the order in which functions should be executed.

Table 1. Terms from annotation data used for defining each type of region in mCSEA.

Region type	Column from annotation data	Terms
Promoters	UCSC_RefGene_Group	TSS1500, TSS200, 5'UTR, 1stExon
Gene bodies	UCSC_RefGene_Group	Body
CpG Islands	Relation_to_Island	Island, N_Shore, S_Shore, N_Shelf, S_Shelf

Finally, the package implements a function, mCSEAIntegrate(), which integrates gene expression data in the analysis. For that purpose, the leading edge CpGs of each region is first defined. The leading edge is the set of CpGs that contributes to the ES of the region, so these CpGs are the most differentially methylated ones. These sites are averaged for each region in each sample. Then, Pearson's correlation coefficient is calculated between each region's methylation and the proximal gene(s)' expression (i.e. genes within 1500 base pairs upstream and downstream from the region). If the integration is performed with promoters, significant negative correlations are returned, due to it has been observed an inverse correlation between promoters' methylation and gene expression (Jones and Baylin, 2002). On the contrary, if the integration is performed in gene bodies, significant positive correlations are returned instead, due to a positive correlation between gene body

methylation and expression has been observed (Aran et al., 2011). If the integration is performed in CGIs, both positive and negative significant correlations are returned, due to CGIs can be located in both promoters and gene bodies.

2.2 Methods comparison

In order to test our method, we used both simulated and real data. We simulated 450K β-values for 20 samples using the same approach as Peters et al. (Peters et al., 2015). We randomly selected 714 promoters to be hypermethylated and another 714 promoters to be hypomethylated in 10 samples (cases) compared to the other 10 (controls). Only promoters with at least 5 associated CpGs were selected. We simulated datasets with a β-value mode differences among phenotypes (Δβ) ranging from 0.9 to 0.05 across promoter CpG sites. We compared mCSEA's performance with state-of-the-art solutions, both predefined (IMA (Wang et al., 2012) and RnBeads (Assenov et al., 2014)) and de novo (DMRcate (Peters et al., 2015), bumphunter (Jaffe et al., 2012), and Probe Lasso (Butcher and Beck, 2015)) algorithms. IMA package uses as input raw idat files and not a β-values matrix. Therefore, to compare its approach using the simulated data we implemented the method that is applied by IMA, that is to calculate the median of the methylation values for each predefined region and to apply limma to these averaged values. We did not included COHCAP package (Warden et al., 2013) due to it restricts the analysis to CGIs. For all methods we used default parameters with the exceptions compiled in Supplementary Table 1.

All results were considered significant using FDR < 0.05 threshold. For IMA and RnBeads, we searched for DMRs in promoter regions and we considered as true positives (TP) those promoters annotated with the actual differentially methylated promoters, and as false positives (FP) the called regions not annotated with the actual DMRs. For the rest of the methods, due to they return de novo DMRs, we considered as TP those actual DMRs overlapping at least one called region, and as FP the called regions not overlapping any actual DMR. For all methods we considered as false negatives (FN) the actual DMRs not called by the corresponding method.

For each method and Δβ we calculated the sensitivity (Equation 1) and the precision or positive predictive value (PPV) (Equation 2).

$$sensitivity = \frac{TP}{TP+FN} \times 100 \quad (1)$$

$$precision = PPV = \frac{TP}{TP+FP} \times 100 \quad (2)$$

We also tested the performance of the proposed method in the methylation datasets previously published by Kim et al., 2017. This dataset contains Illumina 450K methylation data from 18 sibling pairs discordant for intrauterine exposure to maternal gestational diabetes mellitus (GDM). This data is publicly available from GEO database (GEO ID: GSE102177). We reanalyzed the data with IMA, RnBeads, DMRcate, Probe Lasso, bumphunter and mCSEA. We selected these methods because all of them are popular tools for DMRs analysis and allow complex experimental designs with paired samples and covariates, as was our case. Probe Lasso does not directly allow paired–analysis but we adapted its functions to include it in the comparison.

3 Results

3.1 Comparison of DMRs analysis packages

We performed a functional comparison of mCSEA and the most popular R packages used to DMRs analysis from Illumina microarrays data (Table 2). An essential function of this kind of software is the capability to analyze data from complex experimental designs, due to methylation data is very sensitive to environmental factors (Marsit, 2015) and it is important to take into account sex, age, ethnicity and other confounding factors. In addition, some experiments require a paired analysis (e.g. when normal and cancer cells are extracted from the same patient). mCSEA can handle with both, covariates adjusting and paired analysis. Other important features compared were the type of regions that can be included in the analysis and the capacity of integrating gene expression data. Our method and COHCAP are the only tools capable to integrate gene expression data in the analysis to define genes that show strong correlation in gene expression and methylation data, which is a very relevant feature.

Table 2. Comparison of available R packages for DMRs analysis using Illumina's microarray data.

	IMA	**RnBeads**	**DMRcate**	**Bumphunter**	**COHCAP**	**Probe Lasso[1]**	**mCSEA**
Reference	(Wang *et al.*, 2012)	(Assenov *et al.*, 2014)	(Peters *et al.*, 2015)	(Jaffe *et al.*, 2012)	(Warden *et al.*, 2013)	(Butcher and Beck, 2015)	-
DMRs analyzed	Predefined	Predefined	*De novo*	*De novo*	Predefined	*De novo*	Predefined
Platforms	27K and 450K	27K and 450K	450K and EPIC	27K, 450K and EPIC	27K and 450K[2]	450K and EPIC	450K and EPIC[2]
Statistical test	Wilcoxon rank-sum, t-test and empirical Bayes	CpG-level P-values aggregation with Fisher's method	Kernel Smoothing	Bumphunter algorithm	ANOVA	Probe Lasso algorithm	GSEA
Accepts methylation matrix as input	No	Yes	Yes	Yes	Yes	Yes	Yes
Adjusting for covariates	Yes	Yes	Yes	Yes	Only one[3]	No	Yes
Paired analysis	Yes	Yes	Yes	Yes	Yes[3]	No	Yes
Implemented parallelization	No	Yes	Yes	Yes	No	Yes	Yes
Integration of Gene Expression Data	No	No	No	No	Yes	No	Yes
Predefined Regions	UCSC-defined regions (TSS1500, 5' UTR, gene body…)	Promoters, gene bodies, CGIs, tilling regions, user-defined regions	-	-	CGIs	-	Promoters, gene bodies, CGIs, user-defined regions

[1]Implemented in ChAMP package (Morris *et al.*, 2014). [2]Other platforms can be analyzed introducing custom annotations. [3]It is only possible to adjust for one covariate or to perform a paired analysis, but not both.

3.2 Simulated data results

We calculated the number of TP, FP and FN returned by each tested method for each Δβ interval, in addition to sensitivity and PPV (Supplementary Table 2). As can be noted in Figure 2, mCSEA yielded a 100 % of sensitivity detecting methylation differences ranging from Δβ=0.9 to Δβ=0.2 and it outperformed the rest of methods when the methylation differences were especially small (0.05). In addition, mCSEA returns a low number of FP, resulting in a high PPV for all Δβ (Supplementary Table 2). Only DMRcate and Probe Lasso overcome mCSEA in PPV, but at the cost of having a significantly lower sensitivity for all Δβ.

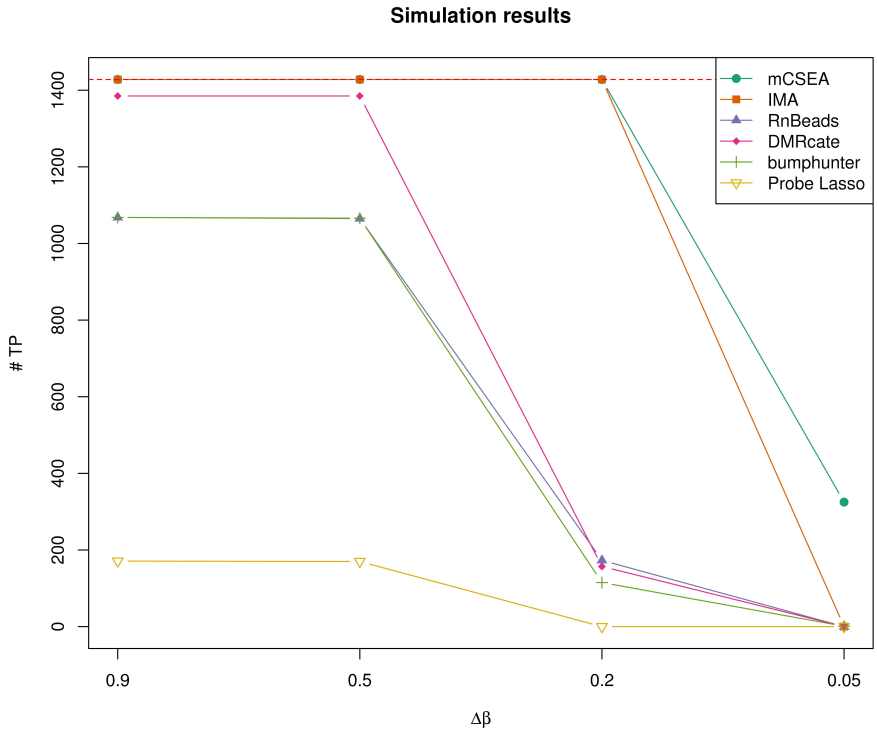


Figure 2. Performance with simulated data. Each line represents results from different methods. The Y-axis represents the number of TP for each Δβ. Red line represents the total number of TP included in the dataset (1428).

3.3 DMRs in maternal diabetes exposure discordant siblings

To demonstrate the mCSEA's functionality, we analyzed the data reported by Kim et al. (Kim et al., 2017). This is a methylation dataset from child sibling pairs: one of the siblings was exposed to maternal diabetes during their gestation, while the other was not. This intrauterine hyperglycemia exposure is associated with an increased risk of obesity and diabetes. Authors collected data from discordant siblings for maternal diabetes exposure in order to get insight into possible epigenetic aberrations in the exposed sibling. Methylation differences in such type of experiment were expected to be very subtle and, in fact, the authors did not report any significant result from the statistical point of view (FDR < 0.05), but they focused in the most differentially methylated genes and discussed their biological relevance.

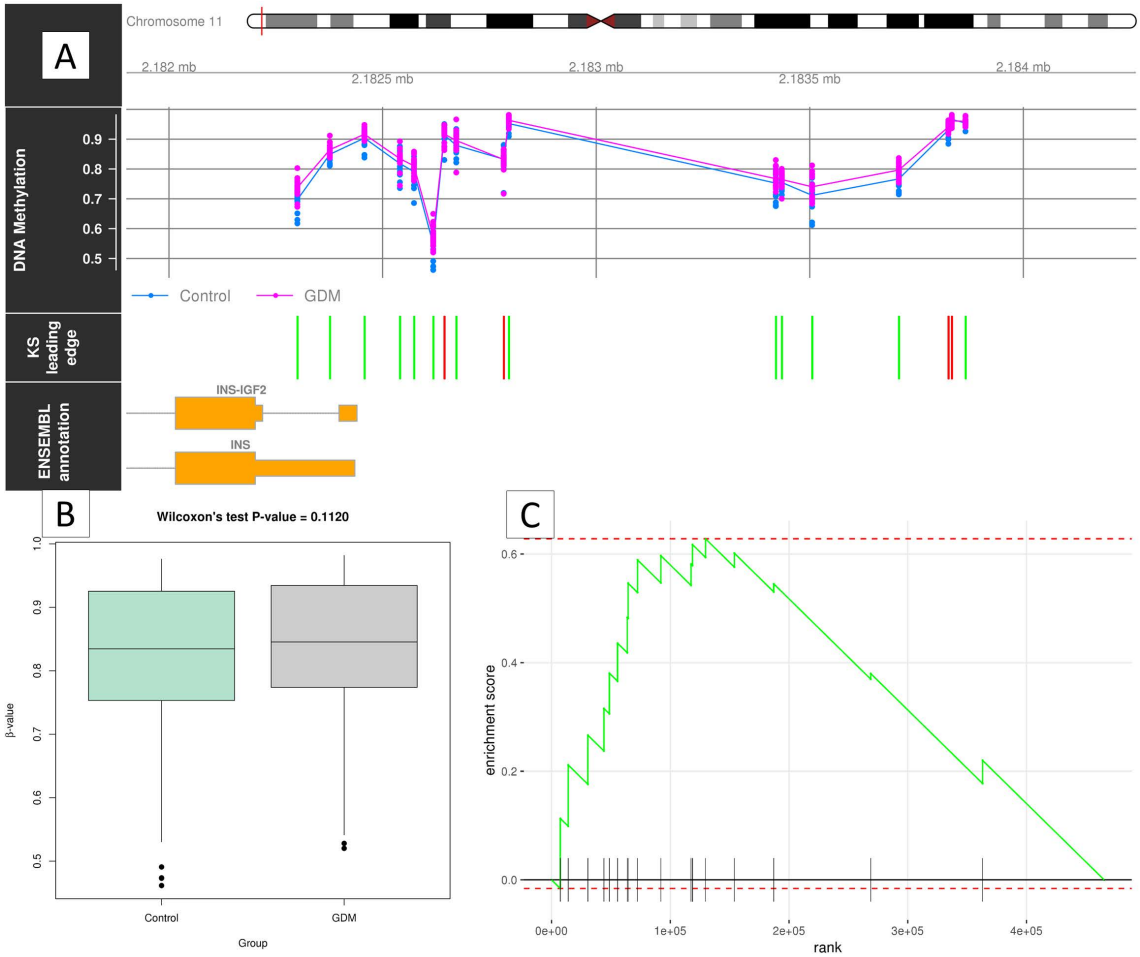


Figure 3. INS gene promoter methylation in GDM and control samples. Methylation is quantified with β-values. **A)** Genomic context of INS promoter. Each point represents the methylation of each sample. Lines link the mean methylation of each group. KS leading edge panel marks with green bars those CpGs contributing to the ES and with red bars the rest of them. This plot was obtained with mCSEAPlot() function, implemented in mCSEA package. **B)** Boxplot showing the subtle difference in INS promoter methylation status between controls and GDM samples. **C)** GSEA plot for INS promoter. Vertical lines mark the location of INS-associated CpGs along the entire ranked list of analyzed CpGs (horizontal black line). Red lines represent the maximum and minimum ES. This plot was obtained with mCSEAPlotGSEA() function, implemented in mCSEA package.

In this dataset, DMRcate and Probe Lasso did not return any significant DMR. These methods applied limma to detect significant DMPs and call DMRs based on them. Although they work properly when methylation differences are high, they did not reveal any significant result for slight methylation differences.

RnBeads is also based on limma for detecting DMRs, but it combines the results by region types (promoters, CGI, and so on) aggregating the P-values obtained by the linear modeling, so, even if there are not any significant DMPs, RnBeads is potentially capable to find significant DMRs. However, this was not the case. This method did not return any

significant DMR (FDR < 0.05). IMA approach did not return any significant DMR neither.

Bumphunter yielded one significant DMR (FDR = 0.03, FWER = 0.01) located at the promoter of SDHAP3 pseudogene. Up to our knowledge, there is not any known relationship between SDHAP3 and development or metabolic disorders.

mCSEA yielded 1055 significant DMRs (FDR < 0.05) in gene promoters: 228 hypermethylated and 827 hypomethylated promoters in cases compared to controls (Supplementary Table 3).

To assess the biological significance of these results, we performed an enrichment analysis using Enrichr (Chen et al., 2013). The most significant enriched pathway in KEGG database (Kanehisa and Goto, 2000) is "Maturity onset diabetes of the young" (hsa04950) pathway (adjusted P-value = 0.0011) (Supplementary Table 4). This pathway is related with a type of diabetes characterized to appear in patients younger than 25 years old and to be non-insulin dependent. Promoter regions of nine out of the twenty-six genes associated to this pathway were identified as significantly differential methylated regions, including PDX1, FOXA2, PAX6 or INS. INS gene, which we found to be hypermethylated in cases, is an important gene that has been previously associated to diabetes in several works and it has been reported as a silenced gene with a fully methylated promoter associated to diabetes development (Yang et al., 2011). In addition, it has been observed that high levels of glucose increase the INS methylation level (Yang et al., 2011), so this hypermethylation could be induced during gestation. Methylation differences in INS promoter between children exposed and non-exposed to intrauterine hyperglycemia are subtle, but consistent across all CpG sites of the promoter (Figure 3). Such small methylation difference is the cause why this DMR remains undetected by all the other tested methods. The same may be occurring in many other genomic regions.

On the other hand, the most significant enriched pathway from OMIM Disease database is obesity (adjusted P-value = 0.0085) (Supplementary Table 5). Eight out of fifteen genes related to this disease contained significant DMRs, including UCP1, UCP3, GHRL or PCSK1. So, we found methylation alterations in genes related to diabetes and obesity, the two main diseases associated to intrauterine GDM exposure.

4 Conclusions

Here we present mCSEA, a novel R package for predefined DMRs detection based on GSEA method. We compared mCSEA with the most widely used methods to detect DMRs. Our method outperformed the rest of solutions for detecting small methylation differences in the simulated dataset. It is especially remarkable the capability of mCSEA to find DMRs even with the methylation difference as small as 0.05 between groups, but consistent along a relatively large region. We reanalyzed a previously published dataset, obtaining barely no significant results with other methods. However, mCSEA yielded several significant DMRs in promoters for genes associated to relevant biological pathways.

We think that mCSEA will provide researchers with a useful tool to detect DMRs in datasets from complex diseases in which the methylation differences among phenotypes are small but consistent.

Acknowledgements

This work is part of the JMM's PhD results. JMM is enrolled in the PhD program in Biomedicine at the University of Granada, Spain.

Funding

JMM is partially funded by Ministerio de Economía, Industria y Competitividad. This work is partially funded by Consejería de Salud, Junta de Andalucía (Grant PI-0152-2017).

Conflict of Interest: none declared.

References

Aran,D. et al. (2011) Replication timing-related and gene body-specific methylation of active human genes. Hum. Mol. Genet., 20, 670–680.

Assenov,Y. et al. (2014) Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods, 11, 1138–1140.

Bohlin,J. et al. (2015) Effect of maternal gestational weight gain on offspring DNA methylation: a follow-up to the ALSPAC cohort study. BMC Res Notes, 8, 321.

Butcher,L.M. and Beck,S. (2015) Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. Methods, 72, 21–28.

Chen,E.Y. et al. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics, 14, 128.

Chiavaroli,V. et al. (2015) Infants born large-for-gestational-age display slower growth in early infancy, but no epigenetic changes at birth. Sci Rep, 5, 14540.

De Smet,C. et al. (1999) DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. Mol. Cell. Biol., 19, 7327–7335.

van Dongen,J. et al. (2015) Epigenome-Wide Association Study of Aggressive Behavior. Twin Res Hum Genet, 18, 686–698.

Flanagan,J.M. (2015) Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol. Biol., 1238, 51–63.

Gervin,K. et al. (2012) DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. PLoS Genet., 8, e1002454.

Guerrero-Bosagna,C. et al. (2014) Identification of genomic features in environmentally induced epigenetic transgenerational inherited sperm epimutations. PLoS ONE, 9, e100194.

Hollander,M. and Wolfe,D.A. (1999) Nonparametric Statistical Methods Wiley, New York.

Jaffe,A.E. et al. (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol, 41, 200–209.

Jones,P.A. and Baylin,S.B. (2002) The fundamental role of epigenetic events in cancer. Nature Reviews Genetics, 3, 415–428.

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

Kim,E. et al. (2017) DNA methylation profiles in sibling pairs discordant for intrauterine exposure to maternal gestational diabetes. Epigenetics, 12, 825–832.

Lappalainen,T. and Greally,J.M. (2017) Associating cellular epigenetic models with human phenotypes. Nat. Rev. Genet., 18, 441–451.

Leenen,F.A.D. et al. (2016) DNA methylation: conducting the orchestra from exposure to phenotype? Clin Epigenetics, 8, 92.

Levenson,V.V. (2010) DNA methylation as a universal biomarker. Expert Rev. Mol. Diagn., 10, 481–488.

Marsit,C.J. (2015) Influence of environmental exposure on human epigenetic regulation. J. Exp. Biol., 218, 71–79.

Mikeska,T. and Craig,J.M. (2014) DNA methylation biomarkers: cancer and beyond. Genes (Basel), 5, 821–864.

Peters,T.J. et al. (2015) De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin, 8, 6.

Ritchie,M.E. et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res., 43, e47.

Sandoval,J. et al. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics, 6, 692–702.

Sergushichev,A. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv, 060012.

Subramanian,A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A., 102, 15545–15550.

Teh,A.L. et al. (2016) Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics, 11, 36–48.

Wang,D. et al. (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics, 28, 729–730.

Warden,C.D. et al. (2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res., 41, e117.

Yang,B.T. et al. (2011) Insulin promoter DNA methylation correlates negatively with insulin gene expression and positively with HbA1c levels in human pancreatic islets. Diabetologia, 54, 360–367.

Supplementary data

Supplementary Table 1. Parameters used for methods comparison with simulated data.

Method	Version	Function	Non-default parameters
mCSEA	1.0.1	mCSEATest	regionsTypes = "promoters"
RnBeads	1.10.0	rnb.execute.computeDiffMeth	adjust.sva = FALSE
			region.types = "promoters"
			adjust.celltype = FALSE
DMRcate	1.14.0	dmrcate	min.cpgs = 5
			C = 2
Bumphunter[1]	2.9.9	champ.DMR	minProbes = 5
			maxGap = 1000
Probe Lasso[1]	2.9.9	champ.DMR	method = "ProbeLasso"
			minProbes = 5

[1]Bumphunter and Probe Lasso were executed through ChAMP package

Supplementary Table 2. Simulated data analysis results for each compared method.

				Δβ = 0.9	
Method	TP	FP	FN	Sensitivity (%)	PPV (%)
mCSEA	1428	1	0	100.00	99.93
Averaged regions	1428	136	0	100.00	91.30
RnBeads	1068	486	360	74.79	68.73
DMRcate	1385	0	43	96.99	100.00
Bumphunter	1068	68	360	74.79	94.01
Probe Lasso	171	0	1257	11.97	100.00
				Δβ = 0.5	
Method	TP	FP	FN	Sensitivity (%)	PPV (%)
mCSEA	1428	5	0	100.00	99.65
Averaged regions	1428	145	0	100.00	90.78
RnBeads	1065	474	363	74.58	69.20
DMRcate	1385	2	43	96.99	99.86
Bumphunter	1066	79	362	74.65	93.10
Probe Lasso	170	0	1258	11.30	100.00
				Δβ = 0.2	
Method	TP	FP	FN	Sensitivity (%)	PPV (%)
mCSEA	1428	25	0	100.00	98.28
Averaged regions	1428	133	0	100.00	91.48
RnBeads	173	65	1255	12.11	72.69
DMRcate	157	0	1271	10.99	100.00
Bumphunter	115	77	1313	8.05	59.90
Probe Lasso	0	0	1428	0.00	NA
				Δβ = 0.1	
Method	TP	FP	FN	Sensitivity (%)	PPV (%)
mCSEA	325	31	1103	22.76	91.29
Averaged regions	0	2	1428	0.00	0.00
RnBeads	0	0	1428	0.00	NA
DMRcate	0	0	1428	0.00	NA
Bumphunter	1	68	1427	0.07	1.45
Probe Lasso	0	0	1428	0.00	NA

Supplementary Tables 3, 4 and 5 are Excel spreadsheets that can be accessed at the Bioinformatics journal site.

7.4 A COMPREHENSIVE AND CENTRALIZED DATABASE FOR EXPLORING OMICS DATA IN AUTOIMMUNE DISEASES

This article is currently under review at BMC Bioinformatics journal. This is a non peer-reviewed preprint published at the bioRxiv repository (2020.06.10.144972; doi: https://doi.org/10.1101/2020.06.10.144972).

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.

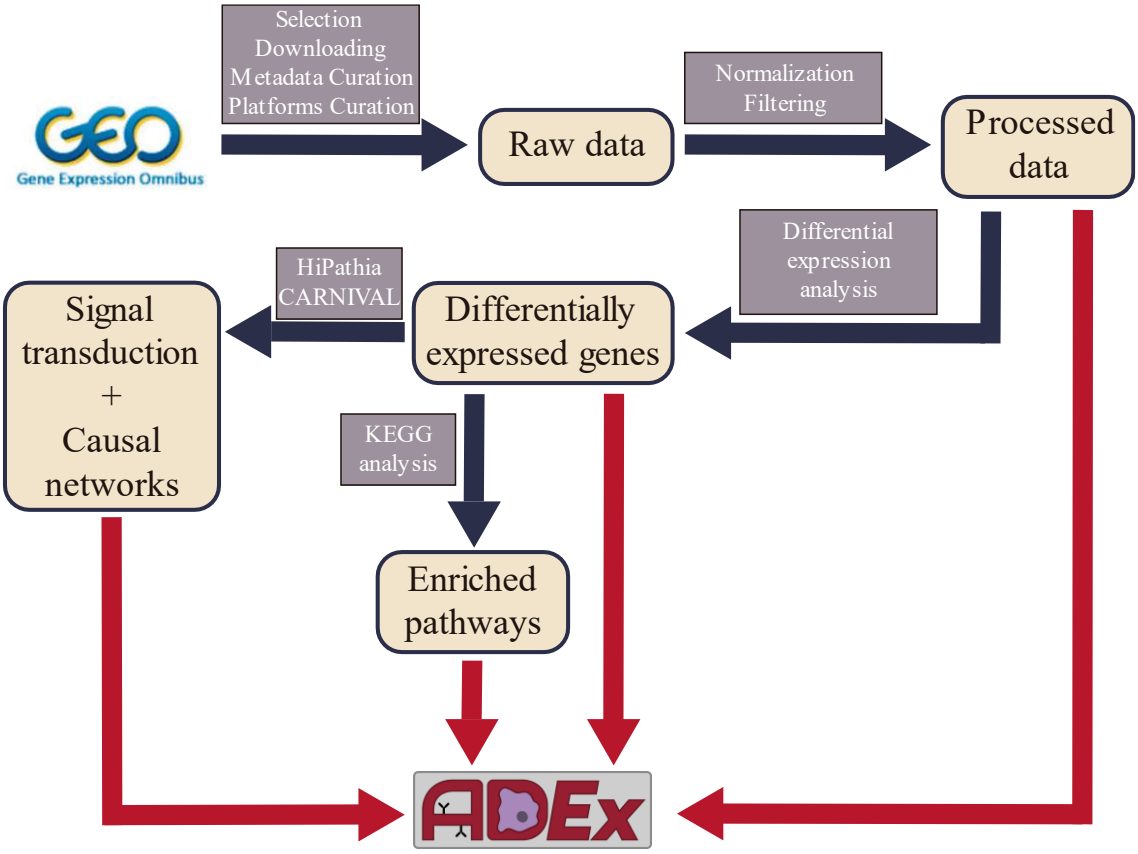


Figure 1. Processing pipeline for ADEx data. Black arrows indicate intermediate processing steps. Red arrows indicate the inputs to ADEx application.

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).

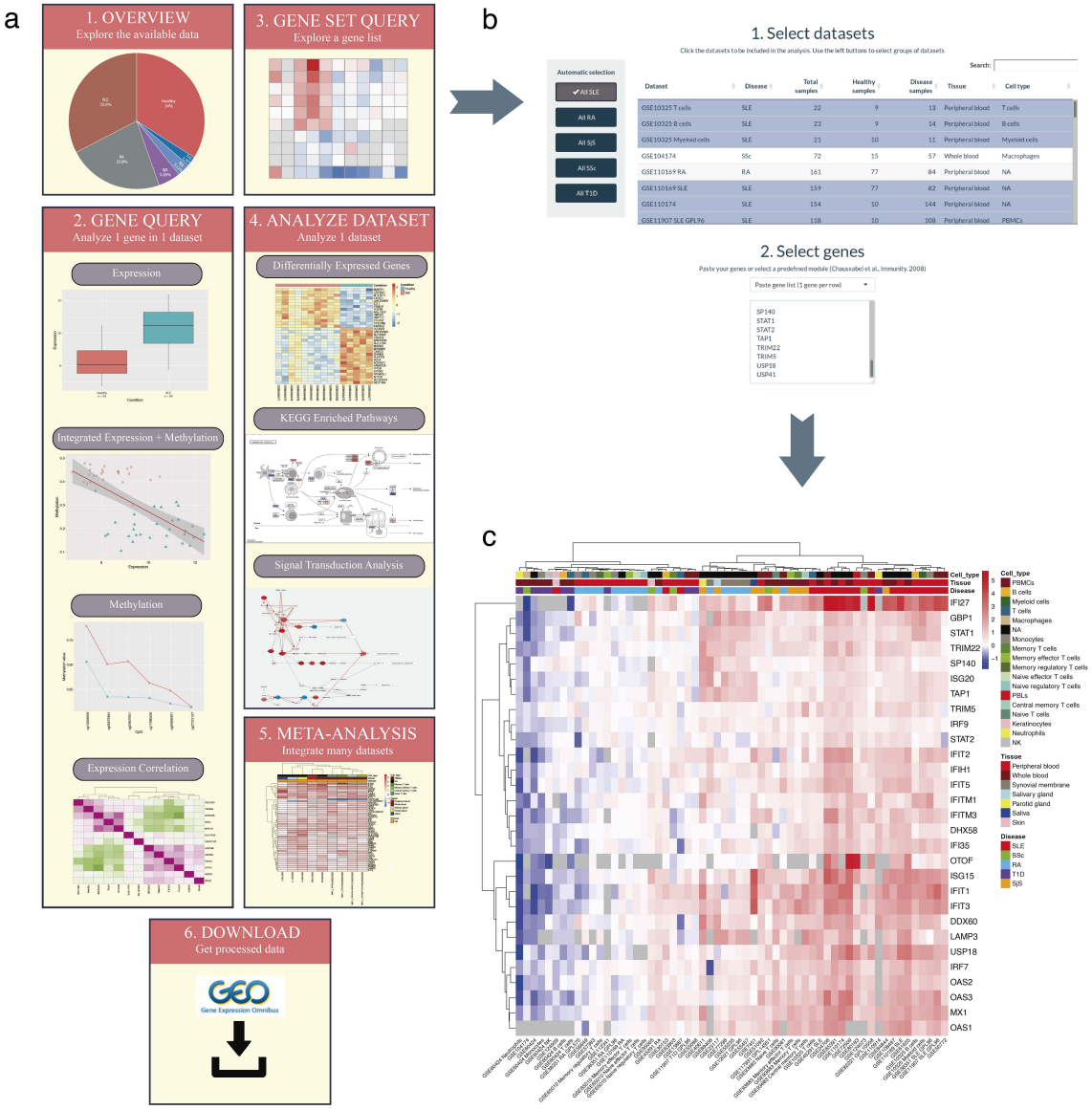


Figure 2. Overview of ADEx application and analysis of IFN signature across diseases. a) ADEx has six main sections. Section 1 provides information about available datasets. In section 2, users can explore expression and methylation for individual genes. Section 3 implements a module to explore data 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.

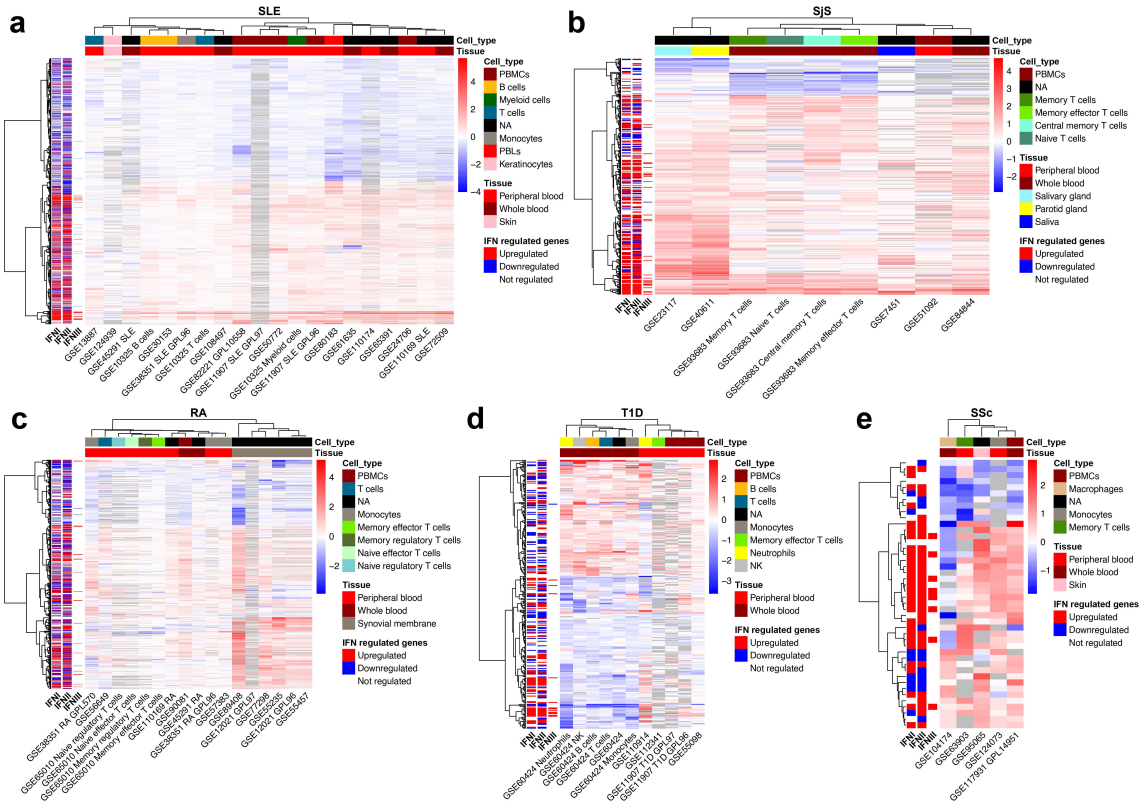


Figure 3. Integration of multiple datasets reveal candidate biomarkers for each disease. The observed effect of IFN I, II and III on gene expression is annotated at the left of each heatmap. Color represents the log2 FC. Heatmaps contains the significant biomarkers for **a)** SLE, **b)** SjS, **c)**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

List of abbreviations

AD: autoimmune disease; CPM: counts per million; CV: coefficient of variation; DEG: differentially expressed gene; FC: Fold-Change; FDR: False Discovery Rate; GEO: Gene Expression Omnibus; GTEx: Genotype-Tissue Expression project; IFN: interferon; IFN I: type I IFN; IFN II: type II IFN; RA: rheumatoid arthritis; RMA: Robust Multichip Average; SAD: systemic autoimmune disease; SjS: Sjögren's syndrome; SLE: systemic lupus erythematosus; SRA: Sequence Read Archive; SSc: systemic sclerosis; T1D: type 1 diabetes; TCGA: The Cancer Genome Atlas

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The original datasets analysed during the current study are available in the GEO repository. The processed datasets generated during the current study are available in the ADEx database, https://adex.genyo.es.

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.

Additional files

Additional file 1

File format: .pdf

Title of data: Description of the datasets included in the ADEx database.

Description of data: This table contains information about each study included in ADEx, with disease, platform, sample size and reference (if available).

Additional file 2

File format: .xlsx

Title of data: Significant biomarkers for each disease.

Description of data: Excel spreadsheet with the significant biomarkers found in the use case 2 for each disease, including the mean log2 FC between case and control samples for each gene.

Additional file 1

Additional file 1. Description of the datasets included in the ADEx database. This table contains information about each study included in ADEx, with disease, platform, sample size and reference (if available).

Dataset	Studied disease	Experimental strategy	Platform	Sample size	Reference
GSE10325	SLE	Expression profiling by array	[HG-U133A] Affymetrix Human Genome U133A Array	67	[1]
GSE104174	SSc	Expression profiling by high throughput sequencing	Illumina HiSeq 2500 (Homo sapiens)	72	[2]
GSE108497	SLE	Expression profiling by array	Illumina HumanHT-12 V4.0 expression beadchip	512	NA
GSE110007	SjS	Methylation profiling by array	Illumina HumanMethylation450 BeadChip (HumanMethylation450_ 15017482)	31	[3]
GSE110169	SLE, RA	Expression profiling by array	[HG-U219] Affymetrix Human Genome U219 Array	234	[4]
GSE110174	SLE	Expression profiling by array	[HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate	154	[4]
GSE110607	SLE	Methylation profiling by genome tiling array	Illumina HumanMethylation450 BeadChip (HumanMethylation450_ 15017482)	104	[5]
GSE110914	T1D	Expression profiling by high throughput sequencing	Illumina HiSeq 2500 (Homo sapiens)	42	[6]
GSE112341	T1D	Expression profiling by high throughput sequencing	Illumina HiSeq 2500 (Homo sapiens)	22	[7]
GSE117931	SSc	Expression profiling by array, Methylation profiling by genome tiling array	llumina HumanHT-12 WG-DASL V4.0 R2 expression beadchip, Illumina HumanMethylation450 BeadChip (HumanMethylation450_ 15017482)	74	NA
GSE11907	SLE	Expression profiling by array	[HG-U133A] Affymetrix Human Genome U133A Array [HG-U133B] Affymetrix Human Genome U133B Array	546	[8]

Dataset	Studied disease	Experimental strategy	Platform	Sample size	Reference
GSE12021	RA	Expression profiling by array	[HG-U133A] Affymetrix Human Genome U133A Array [HG-U133B] Affymetrix Human Genome U133B Array	57	[9]
GSE124073	SSc	Expression profiling by high throughput sequencing	Illumina HiSeq 2000 (Homo sapiens)	73	[10]
GSE124939	SLE	Expression profiling by high throughput sequencing	Illumina HiSeq 4000 (Homo sapiens)	72	[11]
GSE13887	SLE	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	27	[12]
GSE23117	SjS	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	15	[13]
GSE24706	SLE	Expression profiling by array	Illumina HumanWG-6 v3.0 expression beadchip	48	[14]
GSE27895	SLE	Methylation profiling by array	Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)	23	[15]
GSE30153	SLE	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	26	[16]
GSE38351	SLE,RA	Expression profiling by array	[HG-U133A] Affymetrix Human Genome U133A Array [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	74	[17]
GSE40611	SjS	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	49	[18]
GSE42861	RA	Methylation profiling by array	Illumina HumanMethylation450 BeadChip (HumanMethylation450_15017482)	689	[19]
GSE45291	SLE,RA	Expression profiling by array	[HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate	805	[20]
GSE50772	SLE	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	81	[21]
GSE51092	SjS	Expression profiling by array	Illumina HumanWG-6 v3.0 expression beadchip	222	[22]

Dataset	Studied disease	Experimental strategy	Platform	Sample size	Reference
GSE55098	T1D	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	22	[23]
GSE55235	RA	Expression profiling by array	[HG-U133A] Affymetrix Human Genome U133A Array	30	[24]
GSE55457	RA	Expression profiling by array	[HG-U133A] Affymetrix Human Genome U133A Array	33	[24]
GSE56606	T1D	Methylation profiling by array	Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)	100	[25]
GSE56649	RA	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	22	[26]
GSE57383	RA	Expression profiling by array	[HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate	112	[27]
GSE57869	SLE	Methylation profiling by array	Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2)	12	[28]
GSE59250	SLE	Methylation profiling by array	Illumina HumanMethylation450 BeadChip (HumanMethylation450_15017482)	434	[29]
GSE60424	T1D	Expression profiling by high throughput sequencing	Illumina HiScanSQ (Homo sapiens)	134	[30]
GSE61635	SLE	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	129	NA
GSE63903	SSc	Expression profiling by array	Illumina HumanHT-12 V4.0 expression beadchip	14	[31]
GSE65010	RA	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	48	[32]
GSE65391	SLE	Expression profiling by array	Illumina HumanHT-12 V4.0 expression beadchip	996	[33]
GSE71841	RA	Methylation profiling by array	Illumina HumanMethylation450 BeadChip (HumanMethylation450_15017482)	24	NA
GSE72509	SLE	Expression profiling by high throughput sequencing	Illumina HiSeq 2500 (Homo sapiens)	117	[34]

Dataset	Studied disease	Experimental strategy	Platform	Sample size	Reference
GSE7451	SjS	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	20	[35]
GSE77298	RA	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	23	[36]
GSE80183	SLE	Expression profiling by high throughput sequencing	Illumina HiSeq 2000 (Homo sapiens)	16	[37]
GSE82221	SLE	Expression profiling by array, Methylation profiling by genome tiling array	Illumina HumanHT-12 V4.0 expression beadchip, Illumina HumanMethylation450 BeadChip (HumanMethylation450_ 15017482)	110	[38]
GSE84844	SjS	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	60	[39]
GSE87095	RA	Methylation profiling by array	Illumina HumanMethylation450 BeadChip (HumanMethylation450_ 15017482)	122	[40]
GSE89408	RA	Expression profiling by high throughput sequencing	Illumina HiSeq 2000 (Homo sapiens)	218	[41]
GSE90081	RA	Expression profiling by high throughput sequencing	Illumina HiSeq 2000 (Homo sapiens)	24	[42]
GSE93683	SjS	Expression profiling by array	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	48	[39]
GSE95065	SSc	Expression profiling by array	[HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array (HGU133A2 Hs ENTREZG 19.0.0)	33	NA

References

1. Hutcheson J, Scatizzi JC, Siddiqui AM, Haines GK, Wu T, Li Q-Z, et al. Combined deficiency of proapoptotic regulators Bim and Fas results in the early onset of systemic autoimmunity. Immunity. 2008;28:206–17.

2. Moreno-Moral A, Bagnati M, Koturan S, Ko J-H, Fonseca C, Harmston N, et al. Changes in macrophage transcriptome associate with systemic sclerosis and mediate GSDMA contribution to disease risk. Ann Rheum Dis. 2018;77:596–601.

3. Cole MB, Quach H, Quach D, Baker A, Taylor KE, Barcellos LF, et al. Epigenetic Signatures of Salivary Gland Inflammation in Sjögren's Syndrome. Arthritis Rheumatol Hoboken NJ. 2016;68:2936–44.

4. Hu Y, Carman JA, Holloway D, Kansal S, Fan L, Goldstine C, et al. Development of a Molecular Signature to Monitor Pharmacodynamic Responses Mediated by In Vivo Administration of Glucocorticoids. Arthritis Rheumatol Hoboken NJ. 2018;70:1331–42.

5. Ulff-Møller CJ, Asmar F, Liu Y, Svendsen AJ, Busato F, Grønbaek K, et al. Twin DNA Methylation Profiling Reveals Flare-Dependent Interferon Signature and B Cell Promoter Hypermethylation in Systemic Lupus Erythematosus. Arthritis Rheumatol Hoboken NJ. 2018;70:878–90.

6. Vecchio F, Lo Buono N, Stabilini A, Nigi L, Dufort MJ, Geyer S, et al. Abnormal neutrophil signature in the blood and pancreas of presymptomatic and symptomatic type 1 diabetes. JCI Insight. 2018;3.

7. Gao P, Uzun Y, He B, Salamati SE, Coffey JKM, Tsalikian E, et al. Risk variants disrupting enhancers of TH1 and TREG cells in type 1 diabetes. Proc Natl Acad Sci U S A. 2019;116:7581–90.

8. 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.

9. Huber R, Hummert C, Gausmann U, Pohlers D, Koczan D, Guthke R, et al. Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane. Arthritis Res Ther. 2008;10:R98.

10. Mariotti B, Servaas NH, Rossato M, Tamassia N, Cassatella MA, Cossu M, et al. The Long Non-coding RNA NRIR Drives IFN-Response in Monocytes: Implication for Systemic Sclerosis. Front Immunol. 2019;10:100.

11. Tsoi LC, Hile GA, Berthier CC, Sarkar MK, Reed TJ, Liu J, et al. Hypersensitive IFN Responses in Lupus Keratinocytes Reveal Key Mechanistic Determinants in Cutaneous Lupus. J Immunol Baltim Md 1950. 2019;202:2121–30.

12. Fernandez DR, Telarico T, Bonilla E, Li Q, Banerjee S, Middleton FA, et al. Activation of mammalian target of rapamycin controls the loss of TCRzeta in lupus T cells through HRES-1/Rab4-regulated lysosomal degradation. J Immunol Baltim Md 1950. 2009;182:2063–73.

13. Greenwell-Wild T, Moutsopoulos NM, Gliozzi M, Kapsogeorgou E, Rangel Z, Munson PJ, et al. Chitinases in the salivary glands and circulation of patients with Sjögren's syndrome: macrophage harbingers of disease severity. Arthritis Rheum. 2011;63:3103–15.

14. Li Q-Z, Karp DR, Quan J, Branch VK, Zhou J, Lian Y, et al. Risk factors for ANA positivity in healthy persons. Arthritis Res Ther. 2011;13:R38.

15. Jeffries MA, Dozmorov M, Tang Y, Merrill JT, Wren JD, Sawalha AH. Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. Epigenetics. 2011;6:593–601.

16. Garaud J-C, Schickel J-N, Blaison G, Knapp A-M, Dembele D, Ruer-Laventie J, et al. B cell signature during inactive systemic lupus is heterogeneous: toward a biological dissection of lupus. PloS One. 2011;6:e23900.

17. Smiljanovic B, Grün JR, Biesen R, Schulte-Wrede U, Baumgrass R, Stuhlmüller B, et al. The multifaceted balance of TNF-α and type I/II interferon responses in SLE and RA: how monocytes manage the impact of cytokines. J Mol Med Berl Ger. 2012;90:1295–309.

18. Horvath S, Nazmul-Hossain ANM, Pollard RPE, Kroese FGM, Vissink A, Kallenberg CGM, et al. Systems analysis of primary Sjögren's syndrome pathogenesis in salivary glands identifies shared pathways in human and a mouse model. Arthritis Res Ther. 2012;14:R238.

19. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.

20. Bienkowska J, Allaire N, Thai A, Goyal J, Plavina T, Nirula A, et al. Lymphotoxin-LIGHT pathway regulates the interferon signature in rheumatoid arthritis. PloS One. 2014;9:e112545.

21. Kennedy WP, Maciuca R, Wolslegel K, Tew W, Abbas AR, Chaivorapol C, et al. Association of the interferon signature metric with serological disease manifestations but not global activity scores in multiple cohorts of patients with SLE. Lupus Sci Med. 2015;2:e000080.

22. Lessard CJ, Li H, Adrianto I, Ice JA, Rasmussen A, Grundahl KM, et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. Nat Genet. 2013;45:1284–92.

23. Yang M, Ye L, Wang B, Gao J, Liu R, Hong J, et al. Decreased miR-146 expression in peripheral blood mononuclear cells is correlated with ongoing islet autoimmunity in type 1 diabetes patients 1miR-146. J Diabetes. 2015;7:158–65.

24. Woetzel D, Huber R, Kupfer P, Pohlers D, Pfaff M, Driesch D, et al. Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. Arthritis Res Ther. 2014;16:R84.

25. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011;7:e1002300.

26. Ye H, Zhang J, Wang J, Gao Y, Du Y, Li C, et al. CD4 T-cell transcriptome analysis reveals aberrant regulation of STAT3 and Wnt signaling pathways in rheumatoid arthritis: evidence from a case-control study. Arthritis Res Ther. 2015;17:76.

27. Rosenberg A, Fan H, Chiu YG, Bolce R, Tabechian D, Barrett R, et al. Divergent gene activation in peripheral blood and tissues of patients with rheumatoid arthritis, psoriatic arthritis and psoriasis following infliximab therapy. PloS One. 2014;9:e110657.

28. Hong K-M, Kim H-K, Park S-Y, Poojan S, Kim M-K, Sung J, et al. CD3Z hypermethylation is associated with severe clinical manifestations in systemic lupus erythematosus and reduces CD3ζ-chain expression in T cells. Rheumatol Oxf Engl. 2017;56:467–76.

29. Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, et al. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. PLoS Genet. 2013;9:e1003678.

30. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. PloS One. 2014;9:e109760.

31. Ayano M, Tsukamoto H, Kohno K, Ueda N, Tanaka A, Mitoma H, et al. Increased CD226 Expression on CD8+ T Cells Is Associated with Upregulated Cytokine Production and Endothelial Cell Injury in Patients with Systemic Sclerosis. J Immunol Baltim Md 1950. 2015;195:892–900.

32. Walter GJ, Fleskens V, Frederiksen KS, Rajasekhar M, Menon B, Gerwien JG, et al. Phenotypic, Functional, and Gene Expression Profiling of Peripheral CD45RA+ and CD45RO+ CD4+CD25+CD127(low) Treg Cells in Patients With Chronic Rheumatoid Arthritis. Arthritis Rheumatol Hoboken NJ. 2016;68:103–16.

33. 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.

34. Hung T, Pratt GA, Sundararaman B, Townsend MJ, Chaivorapol C, Bhangale T, et al. The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. Science. 2015;350:455–9.

35. Hu S, Wang J, Meijer J, Ieong S, Xie Y, Yu T, et al. Salivary proteomic and genomic biomarkers for primary Sjögren's syndrome. Arthritis Rheum. 2007;56:3588–600.

36. Broeren MGA, de Vries M, Bennink MB, Arntz OJ, Blom AB, Koenders MI, et al. Disease-Regulated Gene Therapy with Anti-Inflammatory Interleukin-10 Under the

Control of the CXCL10 Promoter for the Treatment of Rheumatoid Arthritis. Hum Gene Ther. 2016;27:244–54.

37. Rai R, Chauhan SK, Singh VV, Rai M, Rai G. RNA-seq Analysis Reveals Unique Transcriptome Signatures in Systemic Lupus Erythematosus Patients with Distinct Autoantibody Specificities. PloS One. 2016;11:e0166312.

38. Zhu H, Mi W, Luo H, Chen T, Liu S, Raman I, et al. Whole-genome transcription and DNA methylation analysis of peripheral blood mononuclear cells identified aberrant gene regulation pathways in systemic lupus erythematosus. Arthritis Res Ther. 2016;18:162.

39. Tasaki S, Suzuki K, Nishikawa A, Kassai Y, Takiguchi M, Kurisu R, et al. Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjögren's syndrome. Ann Rheum Dis. 2017;76:1458–66.

40. Julià A, Absher D, López-Lasanta M, Palau N, Pluma A, Waite Jones L, et al. Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. Hum Mol Genet. 2017;26:2803–11.

41. Guo Y, Walsh AM, Fearon U, Smith MD, Wechalekar MD, Yin X, et al. CD40L-Dependent Pathway Is Active at Various Stages of Rheumatoid Arthritis Disease Progression. J Immunol Baltim Md 1950. 2017;198:4490–501.

42. Shchetynsky K, Diaz-Gallo L-M, Folkersen L, Hensvold AH, Catrina AI, Berg L, et al. Discovery of new candidate genes for rheumatoid arthritis through integration of genetic association data with expression pathway analysis. Arthritis Res Ther. 2017;19:19.

Additional file 2

Additional file 2 is an Excel spreadsheet that can be downloaded from the bioRxiv repository.

8 SCIENTIFIC PRODUCTION

This section contains a summary of the doctoral candidate's scientific production during the development of the thesis, including works and collaborations out of the focus of the thesis. For articles published in JCR journals, the impact factor (IF) and position (Q = quartile, D = decile) are included.

8.1 ARTICLES WITH THESIS RESULTS

1. MetaGenyo: a web tool for meta-analysis of genetic association studies. *BMC Bioinformatics*. IF: 2.213 **(Q1)**.

2. ImaGEO: integrative gene expression meta-analysis from GEO database. *Bioinformatics*. IF: 5.610 **(D1)**.

3. mCSEA: detecting subtle differentially methylated regions. *Bioinformatics*. IF: 5.610 **(D1)**.

4. A comprehensive and centralized database for exploring omics data in Autoimmune Diseases. *Preprint in bioRxiv*.

8.2 OTHER WORKS AS FIRST AUTHOR

5. Deep Learning in Omics Data Analysis and Precision Medicine. *Book chapter in Computational Biology*. Codon Publications, Brisbane, Australia.

6. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. *Science of The Total Environment*. IF: 6.551 **(D1)**.

7. Detecting Differentially Methylated Promoters in Genes Related to Diseased Phenotypes Using R Sofware. *Bio-protocol*. Accepted, in press.

8.3 CO-AUTHORSHIP IN COLLABORATIONS

8. Metagene projection characterizes GEN2.2 and CAL-1 as relevant human plasmacytoid dendritic cell models. *Bioinformatics*. IF: 5.481 **(D1)**.

9. Identification and visualization of differential isoform expression in RNA-seq time series. *Bioinformatics*. IF: 4.531 **(D1)**.

10. Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. *Arthritis & Rheumatology*. IF: 9.002 **(D1)**.

11. Exosomal miRNA profile as complementary tool in the diagnostic and prediction of treatment response in localized breast cancer under neoadjuvant chemotherapy. *Breast Cancer Research*. IF: 4.998 **(Q2)**.

12. NOMePlot: analysis of DNA methylation and nucleosome occupancy at the single molecule. *Scientific Reports*. IF: 3.998 **(Q1)**.

13. Differential Treatments Based on Drug-induced Gene Expression Signatures and Longitudinal Systemic Lupus Erythematosus Stratification. *Scientific Reports*. IF: 3.998 **(Q1)**.

14. Direct detection of miR-122 in hepatotoxicity using dynamic chemical labelling overcomes stability and isomiR challenges. *Analytical Chemistry*. IF: 6.350 **(D1)**.

15. A survey of gene expression meta-analysis: methods and applications. *Briefings in Bioinformatics*. IF: 9.101 **(D1)**.

16. Polycomb regulation is coupled to cell cycle transition in pluripotent stem cells. *Science Advances*. IF: 12.804 **(D1)**.

17. The molecular clock protein Bmal1 regulates cell differentiation in mouse embryonic stem cells. *Life Science Alliance*. IF: 2.622 **(Q2)**.

18. Analysis of Menstrual Blood Stromal Cells Reveals SOX15 Triggers Oocyte-Based Human Cell Reprogramming. *iScience*. IF: 4.447 **(Q1)**.

19. iPS-Derived Early Oligodendrocyte Progenitor Cells from SPMS Patients Reveal Deficient In Vitro Cell Migration Stimulation. *Cells*. IF: 4.366 **(Q2)**.

20. DREIMT: a drug repositioning database and prioritization tool for immunomodulation. *Bioinformatics*. IF: 5.610 **(Q1)**.

21. Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. *Arthritis & Rheumatology*. IF: 9.586 **(D1)**.

22. Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth. *Nature Communications*, IF: 12.121 (**D1)**. Accepted, in press.

9 REFERENCES

Afroz, S., Giddaluru, J., Vishwakarma, S., Naz, S., Khan, A.A., and Khan, N. (2017). A Comprehensive Gene Expression Meta-analysis Identifies Novel Immune Signatures in Rheumatoid Arthritis Patients. Front. Immunol. *8*, 74.

Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief. Bioinform. *11*, 253–264.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods *11*, 1138–1140.

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N.A., Petryszak, R., Papatheodorou, I., et al. (2019). ArrayExpress update – from bulk to single-cell expression data. Nucleic Acids Res. *47*, D711–D715.

Attia, J., Thakkinstian, A., and D'Este, C. (2003). Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. J. Clin. Epidemiol. *56*, 297–303.

Bagos, P.G. (2013). Genetic model selection in genome-wide association studies: robust methods and the use of meta-analysis. Stat. Appl. Genet. Mol. Biol. *12*, 285–308.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385.e18.

Bakulski, K.M., and Fallin, M.D. (2014). Epigenetic epidemiology: promises for public health research. Environ. Mol. Mutagen. *55*, 171–183.

Barturen, G., Babaei, S., Català-Moll, F., Martínez-Bueno, M., Makowska, Z., Martorell-Marugán, J., Carmona-Sáez, P., Toro-Domínguez, D., Carnero-Montoro, E., Teruel, M., et al. (2020). Integrative Analysis Reveals a Molecular Stratification of Systemic Autoimmune Diseases. Arthritis Rheumatol. *n/a*.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68–76.

Bell, R., Barraclough, R., and Vasieva, O. (2017). Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. Curr. Mol. Med. *17*, 200–210.

Berger, S.L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. Genes Dev. *23*, 781–783.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288–295.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Bobak, C.A., McDonnell, L., Nemesure, M.D., Lin, J., and Hill, J.E. (2020). Assessment of Imputation Methods for Missing Gene Expression Data in Meta-Analysis of Distinct Cohorts of Tuberculosis Patients. Pac. Symp. Biocomput. Pac. Symp. Biocomput. *25*, 307–318.

Bohlin, J., Andreassen, B.K., Joubert, B.R., Magnus, M.C., Wu, M.C., Parr, C.L., Håberg, S.E., Magnus, P., Reese, S.E., Stoltenberg, C., et al. (2015). Effect of maternal gestational weight gain on offspring DNA methylation: a follow-up to the ALSPAC cohort study. BMC Res. Notes *8*, 321.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H. (2009). Introduction to Meta-Analysis (Wiley).

Breitling, R., and Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. J. Bioinform. Comput. Biol. *3*, 1171–1189.

Butcher, L.M., and Beck, S. (2015). Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data. Methods San Diego Calif *72*, 21–28.

Chen, G., Ramírez, J.C., Deng, N., Qiu, X., Wu, C., Zheng, W.J., and Wu, H. (2019). Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database *2019*.

Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., Cancer Genome Atlas Research Network, and Liang, H. (2018). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell *173*, 386-399.e12.

Chen, Y., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., and Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics *8*, 203–209.

Chiavaroli, V., Cutfield, W.S., Derraik, J.G.B., Pan, Z., Ngo, S., Sheppard, A., Craigie, S., Stone, P., Sadler, L., and Ahlsson, F. (2015). Infants born large-for-gestational-age display slower growth in early infancy, but no epigenetic changes at birth. Sci. Rep. *5*, 14540.

Civelek, M., and Lusis, A.J. (2014). Systems genetics approaches to understand complex traits. Nat. Rev. Genet. *15*, 34–48.

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2016). GenBank. Nucleic Acids Res. *44*, D67–D72.

Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.

Cohn, L.D., and Becker, B.J. (2003). How meta-analysis increases statistical power. Psychol. Methods *8*, 243–253.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. *44*, e71.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

Cronin, M., Sangli, C., Liu, M.-L., Pho, M., Dutta, D., Nguyen, A., Jeong, J., Wu, J., Langone, K.C., and Watson, D. (2007). Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. Clin. Chem. *53*, 1084–1091.

Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinforma. Oxf. Engl. *23*, 1846–1847.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Dayhoff, M.O. (1965). Atlas of protein sequence and structure (National Biomedical Research Foundation.).

De Smet, C., Lurquin, C., Lethé, B., Martelange, V., and Boon, T. (1999). DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. Mol. Cell. Biol. *19*, 7327–7335.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L., and Lin, S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics *11*, 587.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Edwards, A.W.F. (2008). G. H. Hardy (1908) and Hardy–Weinberg Equilibrium. Genetics *179*, 1143–1150.

Egger, M., Smith, G.D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ *315*, 629.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. Natl. Acad. Sci. U. S. A. *103*, 5923–5928.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fisher, R.A. (1934). Statistical methods for research workers, 5th ed. (Oliver and Boyd: Edinburgh).

Flanagan, J.M. (2015). Epigenome-wide association studies (EWAS): past, present, and future. Methods Mol. Biol. Clifton NJ *1238*, 51–63.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M., and Hansen, K.D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. *15*, 503.

Gervin, K., Vigeland, M.D., Mattingsdal, M., Hammerø, M., Nygård, H., Olsen, A.O., Brandt, I., Harris, J.R., Undlien, D.E., and Lyle, R. (2012). DNA methylation and gene expression changes in monozygotic twins discordant for psoriasis: identification of epigenetically dysregulated genes. PLoS Genet. *8*, e1002454.

Gomez-Cabrero, D., Tarazona, S., Ferreirós-Vidal, I., Ramirez, R.N., Company, C., Schmidt, A., Reijmers, T., Paul, V. von S., Marabita, F., Rodríguez-Ubreva, J., et al. (2019). STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. Sci. Data *6*, 256.

Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Med. *375*, 1109–1112.

Guerrero-Bosagna, C., Weeks, S., and Skinner, M.K. (2014). Identification of genomic features in environmentally induced epigenetic transgenerational inherited sperm epimutations. PloS One *9*, e100194.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., and Sakaguchi, A.Y. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234–238.

Harrison, P.W., Ahamed, A., Aslam, R., Alako, B.T.F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., et al. (2021). The European Nucleotide Archive in 2020. Nucleic Acids Res. *49*, D82–D85.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. Genome Biol. *18*, 83.

He, L., Deng, T., and Luo, H. (2015). XPA A23G polymorphism and risk of digestive system cancers: a meta-analysis. OncoTargets Ther. *8*, 385–394.

Heard, N., and Rubin-Delanchy, P. (2018). Choosing Between Methods of Combining p-values. Biometrika *105*, 239–246.

Hedges, L.V. (1982). Fitting Categorical Models to Effect Sizes from a Series of Experiments. J. Educ. Stat. *7*, 119–137.

Higgins, J.P.T., and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. Stat. Med. *21*, 1539–1558.

Hollander, M., and Wolfe, D.A. (1999). Nonparametric Statistical Methods (New York: Wiley).

Hong, F., and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinforma. Oxf. Engl. *24*, 374–382.

Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D., and Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst. Rev. MR000006.

Huang, K.-L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355-370.e14.

Huang, S., Chaudhary, K., and Garmire, L.X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front. Genet. *8*, 84.

Ibáñez, K., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A., and Valencia, A. (2014). Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. PLOS Genet. *10*, e1004173.

Ibeagha-Awemu, E.M., and Zhao, X. (2015). Epigenetic marks: regulators of livestock phenotypes and conceivable sources of missing variation in livestock improvement programs. Front. Genet. *6*, 302.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Ioannidis, J.P.A., Chang, C.Q., Lam, T.K., Schully, S.D., and Khoury, M.J. (2013). The geometric increase in meta-analyses from China in the genomic era. PloS One *8*, e65602.

Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P., and Irizarry, R.A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int. J. Epidemiol. *41*, 200–209.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651–654.

Jones, P.A., and Laird, P.W. (1999). Cancer epigenetics comes of age. Nat. Genet. *21*, 163–167.

Kelly, J., Moyeed, R., Carroll, C., Albani, D., and Li, X. (2019). Gene expression meta-analysis of Parkinson's disease and its relationship with Alzheimer's disease. Mol. Brain *12*, 16.

Kerachian, M.A., Javadmanesh, A., Azghandi, M., Mojtabanezhad Shariatpanahi, A., Yassi, M., Shams Davodly, E., Talebi, A., Khadangi, F., Soltani, G., Hayatbakhsh, A., et al. (2020). Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor. Sci. Rep. *10*, 2813.

Khoury, M.J., Little, J., and Burke, W. (2005). Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. Prev. Chronic. Dis. *2*, A29.

Kim, E., Kwak, S.H., Chung, H.R., Ohn, J.H., Bae, J.H., Choi, S.H., Park, K.S., Hong, J.-S., Sung, J., and Jang, H.C. (2017). DNA methylation profiles in sibling pairs discordant for intrauterine exposure to maternal gestational diabetes. Epigenetics *12*, 825–832.

Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z., and Wild, D.L. (2012). Bayesian correlated clustering to integrate multiple datasets. Bioinformatics *28*, 3290–3297.

Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. BioRxiv 060012.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., et al. (2015). The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. *47*, 692–695.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21.

Lens-Pechakova, L.S. (2016). Centenarian Rates and Life Expectancy Related to the Death Rates of Multiple Sclerosis, Asthma, and Rheumatoid Arthritis and the Incidence of Type 1 Diabetes in Children. Rejuvenation Res. *19*, 53–58.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. Proc. Natl. Acad. Sci. U. S. A. *115*, 4325–4333.

Li, A., and Meyre, D. (2013). Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. Int. J. Obes. 2005 *37*, 559–567.

Liew, A.W.-C., Law, N.-F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief. Bioinform. *12*, 498–513.

Lock, E.F., and Dunson, D.B. (2013). Bayesian consensus clustering. Bioinformatics *29*, 2610–2616.

Long, M.D., Smiraglia, D.J., and Campbell, M.J. (2017). The Genomic Impact of DNA CpG Methylation on Gene Expression; Relationships in Prostate Cancer. Biomolecules *7*.

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. *13*, R44.

Maksimovic, J., Phipson, B., and Oshlack, A. (2017). A cross-package Bioconductor workflow for analysing methylation array data. F1000Research *5*.

Martorell-Marugan, J., Toro-Dominguez, D., Alarcon-Riquelme, M.E., and Carmona-Saez, P. (2017). MetaGenyo: a web tool for meta-analysis of genetic association studies. BMC Bioinformatics *18*, 563.

Martorell-Marugán, J., González-Rumayor, V., and Carmona-Sáez, P. (2019). mCSEA: Detecting subtle differentially methylated regions. Bioinforma. Oxf. Engl.

Martorell-Marugán, J., López-Domínguez, R., García-Moreno, A., Toro-Domínguez, D., Villatoro-García, J.A., Barturen, G., Martín-Gómez, A., Troule, K., Gómez-López, G., Al-Shahrour, F., et al. (2020). A comprehensive and centralized database for exploring omics data in Autoimmune Diseases. BioRxiv 2020.06.10.144972.

Miller, J.A., Cai, C., Langfelder, P., Geschwind, D.H., Kurian, S.M., Salomon, D.R., and Horvath, S. (2011). Strategies for aggregating gene expression data: The collapseRows R function. BMC Bioinformatics *12*, 322.

Moorsel, C.H.M. van, Vis, J. van der, Benschop, C., Ellinghaus, D., Duckworth, A., Scotton, C., Ruven, H.J.T., Quanjel, M.J.R., and Grutters, J. (2020). The MUC5B Promotor Polymorphism Associates with Severe COVID-19 in the European Population. SRRN.

Morris, T.J., and Beck, S. (2015). Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. Methods San Diego Calif *72*, 3–8.

Naeem, H., Wong, N.C., Chatterton, Z., Hong, M.K.H., Pedersen, J.S., Corcoran, N.M., Hovens, C.M., and Macintyre, G. (2014). Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. BMC Genomics *15*, 51.

Nakagawa, S., Noble, D.W.A., Senior, A.M., and Lagisz, M. (2017). Meta-evaluation of meta-analysis: ten appraisal questions for biologists. BMC Biol. *15*, 18.

Nicholson, L.B. (2016). The immune system. Essays Biochem. *60*, 275–301.

O'Mara, T.A., Zhao, M., and Spurdle, A.B. (2016). Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome. Sci. Rep. *6*, 36677.

Orlando, D.A., Guenther, M.G., Frampton, G.M., and Young, R.A. (2012). CpG Island Structure and Trithorax/Polycomb Chromatin Domains in Human Cells. Genomics *100*, 320–326.

Perez-Riverol, Y., Zorin, A., Dass, G., Vu, M.-T., Xu, P., Glont, M., Vizcaíno, J.A., Jarnuczak, A.F., Petryszak, R., Ping, P., et al. (2019). Quantifying the impact of public omics data. Nat. Commun. *10*, 3512.

Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., V Lord, R., Clark, S.J., and Molloy, P.L. (2015). De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin *8*, 6.

Pidsley, R., Y Wong, C.C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L.C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics *14*, 293.

Prat, A., Parker, J.S., Fan, C., and Perou, C.M. (2012). PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. Breast Cancer Res. Treat. *135*, 301–306.

Quezada, H., Guzmán-Ortiz, A.L., Díaz-Sánchez, H., Valle-Rios, R., and Aguirre-Hernández, J. (2017). Omics-based biomarkers: current status and potential use in the clinic. Bol. Med. Hosp. Infant. Mex. *74*, 219–226.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Robertson, K.D. (2005). DNA methylation and human disease. Nat. Rev. Genet. *6*, 597–610.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell *173*, 321-337.e10.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. *103*, 1412–1417.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467–470.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays — a technology review. Nat. Cell Biol. *3*, E190–E195.

Shen, R., Olshen, A.B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics *25*, 2906–2912.

Shieh, A.D., and Hung, Y.S. (2009). Detecting outlier samples in microarray data. Stat. Appl. Genet. Mol. Biol. *8*, Article 13.

Slonim, D.K., and Yanai, I. (2009). Getting Started in Gene Expression Microarray Analysis. PLoS Comput. Biol. *5*.

Song, G.G., Kim, J.-H., Seo, Y.H., Choi, S.J., Ji, J.D., and Lee, Y.H. (2014). Meta-analysis of differentially expressed genes in primary Sjogren's syndrome by using microarray. Hum. Immunol. *75*, 98–104.

Sterne, J.A., Gavaghan, D., and Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J. Clin. Epidemiol. *53*, 1119–1129.

Sterne, J.A.C., Sutton, A.J., Ioannidis, J.P.A., Terrin, N., Jones, D.R., Lau, J., Carpenter, J., Rücker, G., Harbord, R.M., Schmid, C.H., et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ *343*, d4002.

Stringer, S., Wray, N.R., Kahn, R.S., and Derks, E.M. (2011). Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PloS One *6*, e27964.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Sun, Y.V., and Hu, Y.-J. (2016). Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. Adv. Genet. *93*, 147–190.

Sutton, A.J., Jones, D.R., Sheldon, T., Abrams, K., and Song, F. (2000). Methods for Meta-analysis in Medical Research.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. *9*, 465–476.

Teh, A.L., Pan, H., Lin, X., Lim, Y.I., Patro, C.P.K., Cheong, C.Y., Gong, M., MacIsaac, J.L., Kwoh, C.-K., Meaney, M.J., et al. (2016). Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics *11*, 36–48.

Teruel, M., Chamberlain, C., and Alarcón-Riquelme, M.E. (2017). Omics studies: their use in diagnosis and reclassification of SLE and other systemic autoimmune diseases. Rheumatol. Oxf. Engl. *56*, i78–i87.

Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinforma. Oxf. Engl. *29*, 189–196.

Thakkinstian, A., McElduff, P., D'Este, C., Duffy, D., and Attia, J. (2005). A method for meta-analysis of molecular association studies. Stat. Med. *24*, 1291–1306.

The Psychiatric GWAS Consortium (2009). Genomewide association studies: History, rationale and prospects for psychiatric disorders. Am. J. Psychiatry *166*, 540–556.

Toro-Domínguez, D., Carmona-Sáez, P., and Alarcón-Riquelme, M.E. (2014). Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. Arthritis Res. Ther. *16*, 489.

Toro-Domínguez, D., Carmona-Sáez, P., and Alarcón-Riquelme, M.E. (2017). Support for phosphoinositol 3 kinase and mTOR inhibitors as treatment for lupus using in-silico drug-repurposing analysis. Arthritis Res. Ther. *19*.

Toro-Domínguez, D., Martorell-Marugán, J., Goldman, D., Petri, M., Carmona-Sáez, P., and Alarcón-Riquelme, M.E. (2018). Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. Arthritis Rheumatol. Hoboken NJ *70*, 2025–2035.

Toro-Domínguez, D., Martorell-Marugán, J., López-Domínguez, R., García-Moreno, A., González-Rumayor, V., Alarcón-Riquelme, M.E., and Carmona-Sáez, P. (2019). ImaGEO: integrative gene expression meta-analysis from GEO database. Bioinforma. Oxf. Engl. *35*, 880–882.

Toro-Domínguez, D., Villatoro-García, J.A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M.E., and Carmona-Sáez, P. (2020). A survey of gene expression meta-analysis: methods and applications. Brief. Bioinform.

Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., and Siegmund, K.D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. *41*, e90–e90.

Vidal, M., Cusick, M.E., and Barabási, A.-L. (2011). Interactome networks and human disease. Cell *144*, 986–998.

Viechtbauer, W., and Cheung, M.W.-L. (2010). Outlier and influence diagnostics for meta-analysis. Res. Synth. Methods *1*, 112–125.

Waldron, L., and Riester, M. (2016). Meta-Analysis in Gene Expression Studies. In Statistical Genomics, E. Mathé, and S. Davis, eds. (New York, NY: Springer New York), pp. 161–176.

Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods *11*, 333–337.

Wang, D., Yan, L., Hu, Q., Sucheston, L.E., Higgins, M.J., Ambrosone, C.B., Johnson, C.S., Smiraglia, D.J., and Liu, S. (2012). IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics *28*, 729–730.

Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. *41*, e117–e117.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. Nat. Genet. *45*, 1113–1120.

Wetterstrand, K.A. (2020). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Whitehead, A., and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. Stat. Med. *10*, 1665–1677.

Yen-Tsung Huang, Tyler J. VanderWeele, and Xihong Lin (2014). Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. Ann. Appl. Stat. *8*, 352–376.

Yuan, Y., Savage, R.S., and Markowetz, F. (2011). Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. PLOS Comput. Biol. *7*, e1002227.

Zhang, S., Li, Q., Liu, J., and Zhou, X.J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics *27*, i401–i409.