nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-018-0103-9

Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap

Jüri Reimand[1,2,8], Ruth Isserlin[3,8], Veronique Voisin[3], Mike Kucera[3], Christian Tannus-Lopes[3], Asha Rostamianfar[3], Lina Wadi[1], Mona Meyer[1], Jeff Wong[3], Changjiang Xu[3], Daniele Merico[4,5] and Gary D. Bader[3,6,7*]

Pathway enrichment analysis helps researchers gain mechanistic insight into gene lists generated from genome-scale (omics) experiments. This method identifies biological pathways that are enriched in a gene list more than would be expected by chance. We explain the procedures of pathway enrichment analysis and present a practical step-by-step guide to help interpret gene lists resulting from RNA-seq and genome-sequencing experiments. The protocol comprises three major steps: definition of a gene list from omics data, determination of statistically enriched pathways, and visualization and interpretation of the results. We describe how to use this protocol with published examples of differentially expressed genes and mutated cancer genes; however, the principles can be applied to diverse types of omics data. The protocol describes innovative visualization techniques, provides comprehensive background and troubleshooting guidelines, and uses freely available and frequently updated software, including g:Profiler, Gene Set Enrichment Analysis (GSEA), Cytoscape and EnrichmentMap. The complete protocol can be performed in ~4.5 h and is designed for use by biologists with no prior bioinformatics training.

Introduction

Comprehensive quantification of DNA, RNA and proteins in biological samples is now routine. The resulting data are growing exponentially, and their analysis helps researchers discover novel biological functions, genotype–phenotype relationships and disease mechanisms[1,2]. However, analysis and interpretation of these data represent a major challenge for many researchers. Analyses often result in long lists of genes that require an impractically large amount of manual literature searching to interpret. A standard approach to addressing this problem is pathway enrichment analysis, which summarizes the large gene list as a smaller list of more easily interpretable pathways. Pathways are statistically tested for over-representation in the experimental gene list relative to what is expected by chance, using several common statistical tests that consider the number of genes detected in the experiment, their relative ranking and the number of genes annotated to a pathway of interest. For instance, experimental data containing 40% cell cycle genes are surprisingly enriched, given that only 8% of human protein-coding genes are involved in this process.

In a recent example, we used pathway enrichment analysis to help identify histone and DNA methylation by the polycomb repressive complex (PRC2) as the first rational therapeutic target for ependymoma, one of the most prevalent childhood brain cancers[3]. This pathway is targetable by available drugs such as 5-azacytidine, which was used on a compassionate basis in a terminally ill patient and stopped rapid metastatic tumor growth[3]. In another example, we analyzed rare copy-number variants (CNVs) in autism and identified several significant pathways affected by gene deletions, whereas few significant hits were identified with case–control association tests of single genes or loci[4,5]. These examples illustrate the useful insights into biological mechanisms that can be achieved using pathway enrichment analysis.

[1]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada. [3]The Donnelly Centre, University of Toronto, Toronto, ON, Canada. [4]Deep Genomics Inc., Toronto, ON, Canada. [5]The Centre for Applied Genomics (TCAG), The Hospital for Sick Children, Toronto, ON, Canada. [6]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [7]Department of Computer Science, University of Toronto, Toronto, ON, Canada. [8]These authors contributed equally: Jüri Reimand, Ruth Isserlin. *e-mail: gary.bader@utoronto.ca

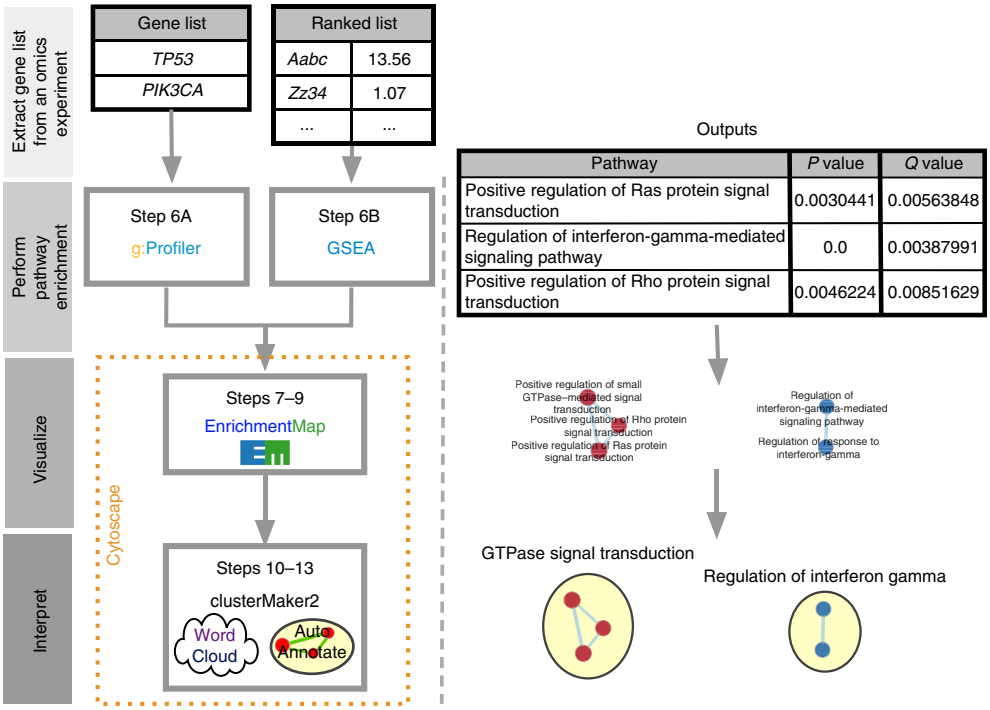


Fig. 1 | Protocol overview. Gene lists derived from diverse omics data undergo pathway enrichment analysis, using g:Profiler or GSEA, to identify pathways that are enriched in the experiment. Pathway enrichment analysis results are visualized and interpreted in Cytoscape using its EnrichmentMap, AutoAnnotate, WordCloud and clusterMaker2 applications. Protocol overview is shown on the left, starting from gene list input, and example outputs at each stage are shown on the right.

Development of the protocol

This protocol covers pathway enrichment analysis of large gene lists typically derived from genome-scale (omics) technology. The protocol is intended for experimental biologists who are interested in interpreting their omics data. It requires only an ability to learn and use 'point-and-click' computer software, although advanced users can benefit from the automatic analysis scripts we provide as Supplementary Protocols 1–4. We analyze previously published human gene expression and somatic mutation data as examples[6–8]; however, our conceptual framework is applicable to analysis of lists of genes or biomolecules from any organism derived from large-scale data, including proteomics, genomics, epigenomics and gene-regulation studies. We extensively use pathway enrichment analysis for many projects and have evaluated numerous available tools[9–12]. The software packages we cover here have been selected for their ease of use, free access, advanced features, extensive documentation and up-to-date databases, and they are ones we use daily in our research and recommend to collaborators and students. In addition, we have provided feedback to the developers of these tools, allowing them to implement features we have needed in published analyses. These tools are g:Profiler[13], GSEA[14], Cytoscape[15] and EnrichmentMap[16], all freely available online:

- g:Profiler (https://biit.cs.ut.ee/gprofiler/)
- GSEA (http://software.broadinstitute.org/gsea/)
- Cytoscape (http://www.cytoscape.org/)
- EnrichmentMap (http://www.baderlab.org/Software/EnrichmentMap)

Overview of the procedure

This section outlines the major stages of pathway enrichment analysis. A detailed step-by-step protocol is provided in the Procedure below. Pathway enrichment analysis involves three major stages (Fig. 1; see Box 1 for basic definitions).

1 *Definition of a gene list of interest using omics data.* An omics experiment comprehensively measures the activity of genes in an experimental context. The resulting raw dataset generally requires computational processing, such as normalization and scoring, to identify genes of interest,

Box 1 | Definitions

Pathway. Genes that work together to carry out a biological process.

Gene set. A set of related genes. A 'pathway gene set' includes all genes in a pathway. Gene sets can be based on various relationships between genes, such as cellular localization (e.g., nuclear genes) or enzymatic function (e.g., protein kinases). Details such as protein interactions are not included.

Gene list of interest. The list of genes derived from an omics experiment that is input to pathway enrichment analysis.

Ranked gene list. In many omics data (e.g., that from RNA-seq for gene expression), genes can be ranked according to some score (e.g., level of differential expression) to provide more information for pathway enrichment analysis. Pathways enriched in genes clustered at the top of a ranked list would score higher than if the pathway genes are randomly scattered across the ranked list.

Pathway enrichment analysis. A statistical technique to identify pathways that are significantly represented in a gene list or ranked gene list of interest.

Multiple testing correction. Thousands of pathways may be individually tested for enrichment, and this could lead to significant enrichment *P* values appearing by chance alone. Multiple testing correction is a statistical technique to correct the *P* values from individual enrichment tests to address this problem and reduce the chance of false-positive enrichment (Box 3).

Leading-edge gene. A subset of genes found in the ranking at or just before the maximal ES in a GSEA analysis. This subset of genes often accounts for a pathway being defined as enriched.

considering the experimental design. For example, a list of genes differentially expressed between two groups of samples can be derived from RNA-seq data[17]. Gene lists derived from other types of omics experiments, such as gene expression microarrays[18], quantitative proteomics[19,20], germline and somatic genome sequencing[21–23], and global DNA methylation assays[24,25], can be used in this protocol; however, each type of data may require specific pre-processing steps (see 'Comparison to alternative methods' section).

2 *Pathway enrichment analysis*. A statistical method is used to identify pathways enriched in the gene list from stage 1, relative to what is expected by chance. All pathways in a given database are tested for enrichment in the gene list (see Box 2 for a list of pathway databases). Several established pathway enrichment analysis methods are available, and the choice of which to use depends on the type of gene list (see 'Comparison to alternative methods' section).

3 *Visualization and interpretation of pathway enrichment analysis results*. Many enriched pathways can be identified in stage 2, often including related versions of the same pathway. Visualization can help identify the main biological themes and their relationships for in-depth study and experimental evaluation.

Stage 1: definition of a gene list of interest using omics data

Genome-scale experiments generate raw data that must be processed to obtain gene-level information suitable for pathway enrichment analysis (Supplementary Protocols 1 and 2). The specific processing steps are particular to the omics experiment type and may be standard, and thus usually straightforward to implement, or not, in which case advanced computational skills may be needed for data processing. Standard processing methods are available for established omics technologies and are most conveniently performed by the core facility that generates the data.

There are two major ways to define a gene list from omics data: list or ranked list. Certain omics data naturally produce a gene list, such as all somatically mutated genes in a tumor identified by exome sequencing, or all proteins that interact with a bait in a proteomics experiment. Such a list is suitable for direct input into pathway enrichment analysis using g:Profiler (Step 6A). Other omics data naturally produce ranked lists. For example, a list of genes can be ranked by differential gene expression score or sensitivity in a genome-wide CRISPR screen. Some pathway enrichment analysis approaches analyze a ranked gene list filtered by a particular threshold (e.g., FDR-adjusted *P* value <0.05 and fold-change >2). Alternative approaches, such as GSEA, are designed to analyze ranked lists of all available genes and do not require a threshold. A whole-genome ranked list is suitable for input into pathway enrichment analysis using GSEA (Step 6B). A partial (non-whole-genome) ranked gene list should be analyzed using g:Profiler.

As an example, we describe the analysis of raw RNA-seq data from ovarian cancer samples to define a ranked gene list[7]. DNA sequence reads are quality filtered (e.g., by trimming to remove low-quality bases) and mapped to a genome-wide reference set of transcripts to enable counting of reads

Box 2 | Pathway enrichment analysis resources

Pathway databases
We list a selection of large, open-access and conveniently accessible pathway databases that offer the maximal value for pathway enrichment analysis. Hundreds of pathway databases are available for many purposes[82].

Gene set databases

- Gene Ontology (GO)[57]: GO provides a hierarchically organized set of thousands of standardized terms for biological processes, molecular functions and cellular components, as well as curated and predicted gene annotations based on these terms for multiple species. Biological process GO annotations are the most commonly used resource for pathway enrichment analysis.
- Molecular Signatures Database (MSigDB)[80,81]: MSigDB is a database of gene sets based on GO, pathways, curation, individual omics studies, sequence motifs, chromosomal position, oncogenic and immunological expression signatures, and various computational analyses maintained by the GSEA team (http://www.msigdb.org). A relatively non-redundant collection of 'hallmark' gene sets is available. The data can be used with many pathway enrichment methods.

Detailed biochemical pathway databases
These databases are maintained by a team of curators who manually collect detailed pathway information, including biochemical reactions, gene regulatory events and other gene interactions. The information can be exported or converted to gene set format.

- Reactome[58]: The most actively updated general-purpose public database of human pathways (http://www.reactome.org).
- Panther[38]: Human signaling pathways (http://pantherdb.org/pathway).
- NetPath[60]: Human signaling pathways with a focus on cancer and immunology (http://www.netpath.org/).
- HumanCyc[59]: Human metabolic pathways (http://humancyc.org/).
- National Cancer Institute (NCI) Pathway Interaction Database (PID): Human cancer-related signaling pathways; this database is no longer updated.
- KEGG[83]: The KEGG database is most useful for its intuitive pathway diagrams. It contains multiple types of pathways, some of which are not normal pathways but are rather disease-associated gene sets, such as 'pathways in cancer' (http://www.genome.jp/kegg/). Up-to-date GMT files for KEGG pathways are currently not freely available because of data licensing restrictions.

Pathway meta-databases
These databases collect detailed pathway descriptions from multiple originating pathway databases.

- Pathway Commons[45]: Collects information from other pathway databases and provides it in a standardized format (http://www.pathwaycommons.org).
- WikiPathways[48]: A community-driven collection of pathways that also includes pathways from other databases (http://www.wikipathways.org/).

per transcript. Read counts are aggregated at the gene level (counts per gene). Typically, RNA-seq data for multiple biological replicates (three or more) for each of multiple experimental conditions (two or more, e.g., treatment versus control) are available. Read counts per gene are normalized across all samples to remove unwanted technical variation between samples, for example, due to differences in sequencing lane or total read number per sequencing run[26–28]. Next, read counts per gene are tested for differential expression across sample groups (e.g., treatment versus control) (Supplementary Protocols 1 and 2 for RNA-seq and microarray data, respectively). Software packages such as edgeR[29], DESeq[30], limma/voom[31,32] and Cufflinks[33] implement procedures for RNA-seq data normalization and differential expression analysis. Differential gene expression analysis results include: (i) the *P* value of the significance of differential expression; (ii) the related *Q* value (a.k.a. adjusted *P* value) that has been corrected for multiple testing across all genes, for example, by using the Benjamini–Hochberg false-discovery rate (BH-FDR) procedure[34] (Box 3); (iii) effect size and direction of expression change so that upregulated genes are positive and at the top of the list and downregulated genes are negative and at the bottom of the list, often expressed as log-transformed fold-change. The gene list is then ranked by one or more of these values (e.g., −log10 *P* value multiplied by the sign of log-transformed fold-change) and studied using pathway enrichment analysis.

Stage 2A: pathway enrichment analysis of a gene list using g:Profiler (Step 6A)

The default analysis implemented in g:Profiler and similar web-based tools[35–38] searches for pathways whose genes are significantly enriched (i.e., over-represented) in the fixed list of genes of interest, as

Box 3 | Multiple testing correction

Repeated statistical testing used in a typical pathway enrichment analysis will result in some apparently significant *P* values by chance alone. To correct this, multiple-testing correction methods systematically reduce the significance of each *P* value derived from a series of tests. In this protocol, g:Profiler and GSEA automatically apply multiple-testing correction to *P* values. The most commonly used method is the BH-FDR (or often simply FDR)[34]. It is based on a step-down procedure that estimates the fraction of falsely enriched pathways over enriched pathways, using the uncorrected *P*-value threshold and the number of tests. For instance, given that 100 pathways have enrichment *P* value <0.05 and an FDR of 5% at *P* value <0.05 means that five of those pathways are expected to be falsely enriched. As an alternative, the classic Bonferroni multiple-testing correction adjusts the significance threshold by dividing it by the number of tests. Practically, the method multiplies each uncorrected *P* value by the number of conducted tests and applies a significance cutoff (e.g., a *P* value of 0.001 will become an insignificant *Q* value 0.1 if 100 pathways have been tested). This technique ensures that the probability of selecting at least one falsely enriched pathway is below the corrected *P* value threshold. Bonferroni correction is typically considered overly conservative for differential gene expression and pathway enrichment analysis because some fraction of false-positive findings can be tolerated. Importantly, both Bonferroni and BH-FDR assume tests are independent, whereas pathways are typically not independent because of overlapping genes and crosstalk. Therefore, BH-FDR estimates for pathway analysis can be inaccurate, although, practically, they are still useful for filtering and hypothesis generation and thus are routinely used.

compared to all genes in the genome (Step 6A)(Box 4). The *P* value of the enrichment of a pathway is computed using a Fisher's exact test and multiple-test correction is applied (Box 3).

g:Profiler also includes an ordered enrichment test, which is suitable for lists of up to a few thousand genes that are ordered by a score, whereas the rest of the genes in the genome lack meaningful signal for ranking. For example, significantly mutated genes may be ranked by a score from a cancer driver prediction method[6]. This analysis repeats a modified Fisher's exact test on incrementally larger sub-lists of the input genes and reports the sub-list with the strongest enrichment *P* value for each pathway[39]. g:Profiler searches a collection of gene sets representing Gene Ontology (GO) terms, pathways, networks, regulatory motifs, and disease phenotypes. Major categories of gene sets can be selected to customize the search.

Pathway enrichment methods that use the Fisher's exact test or related tests require the definition of background genes for comparison. All annotated protein-coding genes are often used as default. This leads to inappropriate inflation of *P* values and false-positive results if the experiment can directly measure only a subset of all genes. For example, setting a custom background is important in analyzing data from targeted sequencing or phosphoproteomics experiments. The appropriate custom background would include all genes in the sequencing panel or all known phosphoproteins, respectively.

Stage 2B: pathway enrichment analysis of a ranked gene list using GSEA (Step 6B)

Pathway enrichment analysis of a ranked gene list is implemented in the GSEA software[14] (Step 6B) (Box 4). GSEA is a threshold-free method that analyzes all genes on the basis of their differential expression rank, or other score, without prior gene filtering. GSEA is particularly suitable and is recommended when ranks are available for all or most of the genes in the genome (e.g., for RNA-seq data). However, it is not suitable when only a small portion of genes have ranks available, for example, in an experiment that identifies significantly mutated cancer genes (Stage 2A; Step 6A).

GSEA searches for pathways whose genes are enriched at the top or bottom of the ranked gene list, more so than expected by chance alone. For instance, if the topmost differentially expressed genes are involved in the cell cycle, this suggests that the cell cycle pathway is regulated in the experiment. By contrast, the cell cycle pathway is probably not significantly regulated if the cell cycle genes appear randomly scattered through the whole ranked list. To calculate an enrichment score (ES) for a pathway, GSEA progressively examines genes from the top to the bottom of the ranked list, increasing the ES if a gene is part of the pathway and decreasing the score otherwise. These running sum values are weighted, so that enrichment in the very top- (and bottom-) ranking genes is amplified, whereas enrichment in genes with more moderate ranks are not amplified. The ES score is calculated as the maximum value of the running sum and normalized relative to pathway size, resulting in a normalized enrichment score (NES) that reflects the enrichment of the pathway in the list. Positive and negative NES values represent enrichment at the top and bottom of the list, respectively. Finally, a permutation-based *P* value is computed and corrected for multiple testing to produce a permutation-based false-discovery rate (FDR) *Q* value that ranges from 0 (highly significant) to 1 (not significant)

Box 4 | Statistical tests in pathway enrichment analysis

A common statistical test used for pathway enrichment analysis of a gene list is a Fisher's exact test based on the hypergeometric distribution. It determines whether the fraction of genes of interest in the pathway is higher compared to the fraction of genes outside the pathway (i.e., background set). Since this test was first introduced[84], many improved tests have been developed[85] that take advantage of continuous experimental scores and avoid applying arbitrary thresholds. We categorize types of statistical enrichment tests as follows:

1 *Ranked versus non-ranked tests.* Ranked tests take as input a ranked gene list, whereas non-ranked tests such as Fisher's exact test take as input a gene list of interest. Ranked tests are preferable for experiments that produce meaningful ranks such as differential gene expression, because arbitrary thresholds can be avoided. Non-ranked tests are preferable for experiments that naturally generate a gene list of interest (e.g., somatic mutations in cancer, proteins that interact with a bait protein). Examples of ranked tests include the modified Fisher's exact test implemented in the g:Profiler 'ordered query' option, and the modified Kolmogorov–Smirnov test implemented in GSEA.

2 *Exact versus permutation-based tests.* Exact tests use a mathematical model (e.g., a distribution) to directly compute an exact *P* value. Permutation-based tests utilize data resampling to estimate an empirical *P* value, typically expressed as the number of permutations with results as good as or better than the ones observed for real data, divided by the number of permutations. For example, in a case–control study, we can randomize the case and control labels 1,000 times, each time repeating the pathway enrichment analysis to see how frequently we observe an equal or stronger pathway enrichment signal. Permutation tests can be customized to consider specific data properties and biases. Exact tests, if applicable, are preferable, because these can quickly compute accurate *P* values. However, devising the right exact test for a specific application may be challenging, in which case a custom permutation test is often a preferred option.

3 *Competitive versus self-contained tests.* Competitive tests determine whether the gene list of interest is enriched in pathways relative to all genes in the background set. Thus, each pathway 'competes' for enrichment in the gene list against genes of the background set. By contrast, self-contained tests calculate statistics uniquely at the pathway level, ignoring genes of the background set. For instance, a self-contained test can evaluate whether the gene expression within a given pathway is different in case samples compared to control samples[85]. Competitive pathway enrichment analysis is most popular and is usually appropriate for gene expression data. However, self-contained tests must be used if single gene differences are not significant and need to be pooled at the pathway gene set level to identify signal, for example, when analyzing rare gene mutations or other data with low per-gene counts[86]. Hybrid approaches may be preferable to self-contained tests in specific circumstances. For instance, for rare CNV data, correcting a self-contained test for global CNV burden leads to more specific biological results[68]. Finally, competitive enrichment tests such as Fisher's exact test ignore correlation among genes, whereas modified competitive tests such as Camera[71] consider these and thus typically produce more rigorous results (see Supplementary Protocol 3, for example). Self-contained tests do not present this issue.

In summary, if genes in your data can be ranked, a ranked test should be used. Fisher's exact test is generally chosen for non-ranked gene lists, and a modified version of the test is available for ranked lists. A competitive test is adequate in most cases, unless the signal at the gene level is weak.

(Box 3). The same analysis is performed starting from the bottom of the ranked gene list to identify pathways enriched in the bottom of the list. Resulting pathways are selected using the FDR *Q* value threshold (e.g., $Q < 0.05$) and ranked using NES. In addition, the 'leading edge' aspect of the GSEA analysis identifies specific genes that most strongly contribute to the detected enrichment signal of a pathway.

GSEA has two methods for determining the statistical significance (*P* value) of the ES: gene set permutation and phenotype permutation. The gene set permutation test requires a ranked list, and GSEA compares the observed pathway ES to a distribution of scores obtained by repeating the analysis with randomly sampled gene sets of matching sizes (e.g., 1,000 times). The phenotype permutation test requires expression data for all samples (e.g., biological replicates), along with a definition of sample groups called 'phenotypes' that are compared with each other (e.g., cases versus controls; tumor versus. normal samples). The observed pathway ES is compared to a distribution of scores obtained by randomly shuffling the samples among phenotype categories and repeating the analysis (e.g., 1,000 times), including computation of the ranked gene list and resulting pathway ES. Gene set permutation is recommended for studies with limited variability and biological replicates (i.e., two to five per condition). In this case, differential gene expression values should be computed outside of GSEA, using methods that include variance stabilization (such as edgeR[29], DESeq[30] and limma/voom[31,32]) and imported into the GSEA software before pathway analysis. Phenotype permutation should be used with a larger number of replicates (e.g., at least ten per condition). The main advantage of the phenotype permutation approach is that it maintains the structure of gene sets with biologically important gene correlations during permutation, in contrast to the gene set permutation approach. This protocol covers only gene set permutation because it is appropriate for the most common use case of pathway enrichment analysis. Phenotype permutation is computationally

expensive and, for the current version of GSEA, requires custom programming to compute ESs and differential expression statistics separately for thousands of phenotype randomizations. For advanced users, we provide a supplementary protocol for this procedure (Supplementary Protocol 4).

By default, the GSEA desktop software searches the MSigDB gene set database, which includes pathways, published gene signatures, microRNA target genes and other gene set types (Box 2). The user can also provide a custom database as a text-based GMT (Gene Matrix Transposed) file in which each line defines a pathway, with its name, identifier and a list of genes it contains. Gene identifiers in the GMT file must match those in the input gene list.

Stage 3: visualization and interpretation of pathway enrichment analysis results (Steps 7–13)

Pathway information is inherently redundant, as genes often participate in multiple pathways, and databases may organize pathways hierarchically by including general and specific pathways with many shared genes (e.g., 'cell cycle' and 'M-phase of cell cycle'). Consequently, pathway enrichment analysis often highlights several versions of the same pathway. Collapsing redundant pathways into a single biological theme simplifies interpretation. We recommend addressing such redundancy with visualization methods such as EnrichmentMap[16], ClueGO[40] and others[41–43]. An 'enrichment map' is a network visualization that represents overlaps among enriched pathways (Fig. 1), whereas 'EnrichmentMap' refers to the Cytoscape application that creates the visualization. Pathways are shown as circles (nodes) that are connected with lines (edges) if the pathways share many genes. Nodes are colored by ES, and edges are sized on the basis of the number of genes shared by the connected pathways. Network layout and clustering algorithms automatically group similar pathways into major biological themes. The EnrichmentMap software takes as input a text file containing pathway enrichment analysis results and another text file containing the pathway gene sets used in the original enrichment analysis. Interactive exploration of pathway ES (filtering nodes) and connections between pathways (filtering edges) is possible (Step 9A(xii and xiii) and 9B(xiii and xiv)). Multiple enrichment analysis results can be simultaneously visualized in a single enrichment map, in which case different colors are used on the nodes for each enrichment. If the gene expression data are optionally loaded, clicking on a pathway node will display a gene expression heat map of all genes in the pathway.

An enrichment map helps identify interesting pathways and themes. First, expected themes should be identified to help validate the pathway enrichment analysis results (positive controls). For instance, growth-related pathways and other hallmarks of cancer[44] are expected to be identified in analyses of cancer genomics datasets. Second, pathways not previously associated with the experimental context are evaluated more carefully as potential discoveries. Pathways and themes with the strongest ESs should be studied first, followed by progressively weaker signals (Step 12). Third, interesting pathways are examined in more detail, examining genes within the pathways (e.g., expression heat maps and the GSEA leading edge genes). Further, gene expression values can be overlaid on a pathway diagram, if available, from databases such as Pathway Commons[45], Reactome[46], KEGG[47] or WikiPathways[48], using tools such as PathVisio[49]. If a diagram is not available, tools such as STRING[50] or GeneMANIA[51] can be used with Cytoscape[15] to define an interaction network among pathway genes for expression overlay. This helps in visual identification of the pathway components (e.g., single genes or entire signaling cascades) that are the most altered (e.g., differentially expressed) in the experiment. In addition, master regulators for enriched pathways can be searched for by integrating gene sets of miRNA[52] or transcription factor[53] targets using the EnrichmentMap post-analysis tool. Finally, pathway enrichment analysis results can be published to support a scientific conclusion (e.g., functional differences of two cancer subtypes), or used for hypothesis generation or planning of experiments to support the identification of novel pathways. More pathway enrichment analysis examples and deeper explanation of core concepts is provided at http://www.pathwaycommons.org/guide/.

Advantages and limitations

Pathway enrichment analysis of omics data has several advantages as compared to analysis of single genes, transcripts or proteins. First, it improves statistical power in two ways: (i) it aggregates counts of mutations across all the genes and genomic regions involved in the given cell mechanism, providing a higher number of counts, which makes statistical analyses more reliable; and (ii) it reduces the dimensionality from tens of thousands of genes or millions of genomic regions (e.g., SNPs) to a

much smaller number of 'systems' or 'pathways', thereby reducing the cost of multiple hypothesis testing. Second, results are often easier to interpret because the analysis is phrased at the level of familiar concepts such as 'cell cycle'. Third, the approach can help identify potential causal mechanisms and drug targets. Fourth, results obtained from related, but different, data may be more comparable because results are projected onto a smaller, shared feature space (i.e., a limited number of pathways). Fifth, the approach facilitates integration of diverse data types, such as genomics, transcriptomics and proteomics, which can all be mapped to the same pathways. Thus, projecting disease data onto known mechanisms increases statistical and interpretative power.

The following limitations are important to consider when interpreting pathway enrichment analysis results, in general, including those covered by this protocol. Additional limitations apply, depending on the omics data type (see 'Application to diverse omics data' section). Advantages and disadvantages of specific and alternative pathway enrichment analysis methods are presented in the 'Comparison to alternative methods' section.

- Enrichment analysis is more effective for pathways in which multiple genes have strong biological signals (e.g., differential expression). For instance, in a transcriptomics experiment, we assume that evolution has optimized a cell to express a pathway only when needed, and that pathway activation or deactivation can be identified as coordinated activity of many genes in a pathway. Pathways in which activity is controlled by only a few genes or not controlled by gene expression (e.g., by post-translational regulation) will never be observed as enriched. Some pathway analysis methods address this by using activating and inhibiting gene interactions to construct quantitative models of pathway activity that include genes that are not differentially expressed yet are still important regulators. However, these methods require pathway models with detailed biochemical and regulatory gene interactions that are obtained through focused experiments and thus are in limited supply (Box 2).
- Pathway boundaries tend to be arbitrary, and different databases will disagree about which genes are involved in a given pathway. By using multiple databases, multiple pathway definitions can be analyzed, and some may be better than others at explaining the experimental data.
- Some pathway enrichment methods, such as those based on the Fisher's exact test, are statistically more likely to identify larger pathways as significant. Users can address this limitation by selecting an upper limit for the size of the gene sets considered in the analysis.
- Multi-functional genes that are highly ranked in the gene list may lead to enrichment of many different pathways, some of which are not relevant to the experiment[54]. Repeating the analysis after excluding such genes may reveal pathways whose enrichment is overly dependent on their presence or confirm the robustness of pathway enrichment.
- Pathway databases, and therefore enrichment results, are biased toward well-known pathways. In fact, pathway enrichment analysis ignores genes with no pathway annotations, sometimes called the 'dark matter of the genome', and these genes should be studied separately. For example, non-coding RNA genes currently lack systematic annotations and are not directly usable in pathway enrichment analysis.
- Most enrichment analysis methods make unrealistic assumptions of statistical independence among genes as well as pathways. Some genes may always be co-expressed (e.g., genes within a protein complex), and some pathways have genes in common. Thus, standard FDRs, which assume statistical independence between tests, are often either more or less conservative than ideal. Nonetheless, they should still be used to adjust for multiple testing and rank enriched pathways for exploratory analysis and hypothesis generation. Custom permutation tests may lead to better estimates of false discovery (see 'Comparison to alternative methods' section).

Experimental design

Pathway enrichment analysis benefits greatly from careful experimental design. Otherwise, the analysis may reveal apparently meaningful results caused by experimental biases or other confounders. This section covers a range of experimental aspects that must be considered before performing this protocol.

Experimental conditions

The experimental conditions must be well defined such that the major variations observed are responses that the experimenter would like to monitor and that are related to the biological question of interest (e.g., tumor versus normal, treated versus untreated, comparison of four disease subtypes, time series).

Number of replicates

Biological replicates are independently processed samples obtained from distinct organisms or cell lines that are required for measuring variability across samples and to compute statistical significance (*P* values). Lack of replication (i.e., one sample per group) will not permit robust estimation of the significance of the signal. Insufficient replication may result in lack of signal in the data (e.g., no significantly differentially expressed genes). The larger the variation in the set of samples, the more biological replicates are needed to accurately measure the signal. For systems with lower variability (i.e., model organisms with the same genetic background in controlled laboratory conditions, or stable cell lines derived from the same clone), at least three to four biological replicates are recommended per condition for differential analysis with variance stabilization normalization. Variance stabilization uses a global statistical model to 'stabilize' gene-wise variance estimates to reduce inaccuracies resulting from few replicates. For experiments with higher variability (e.g., tumor samples), more replicates are required; ideally, a pilot experiment followed by formal statistical power calculations[55] (sometimes called sensitivity testing) should be used to determine the minimal number of replicates required to identify the signal of differentially expressed genes or enriched pathways. Technical replicates comprising repeated experiments of the same samples are usually not needed for well-established experimental techniques, such as RNA-seq, that have low technical variability, but can be helpful for novel techniques.

Confounding factors

Differences in factors not related to the experimental question should be avoided or at least balanced across conditions so that statistical techniques such as generalized linear models can correct for each factor. Common factors include sequencing batch, nucleic acid extraction protocol, subject age and many others. Otherwise, it may be impossible to accurately separate the experimental signals coming from the experimental response from the confounding factors. Knowing important factors in advance supports correct experimental design. Statistical exploratory analyses such as clustering or principal component analysis (PCA) can help identify unknown factors. For example, cases and controls are expected to cluster separately and not by processing batch.

Outliers

Outlier samples may considerably differ from others because of major experimental or technical problems, such as contamination or sample mix-up. Alternatively, they may present extreme biological features, such as tumor samples with exceptionally aggressive phenotypes. Unbiased identification of outlier samples is possible using statistical techniques such as PCA or clustering. Pathway enrichment analysis should be performed with and without outliers to ensure robust results. Systematic removal of outliers may be justified to reduce variability in the experiment.

Experimental sensitivity

Some experimental methods can be tuned to be more or less sensitive. For instance, the number of reads in RNA-seq experiments influences downstream analysis. For quantifying gene expression in a biological system with modest variability and testing differential expression with variance stabilization, at least three to five replicates and 10 million mapped reads are required[56]. Substantially greater sequencing depth, such as 50–100 million mapped reads, is required to investigate splice isoforms, to detect poorly expressed genes or for samples with complex cellular mixtures such as surgical resection specimens.

Choice of pathway gene set database

We recommend searching enrichment of pathway gene sets only at first, as these capture familiar normal cellular processes that are easy to interpret. GO[57] biological process terms and manually curated molecular pathways from Reactome[58], Panther[38], HumanCyc[59] and NetPath[60] are good resources for human pathways (Box 2). GO biological process annotations include a mix of manually curated and electronically inferred sources.

Filtering GO pathways by evidence code

A large fraction of gene annotations in GO originate from automatic data analysis and are not verified by human curators. These have the evidence code 'inferred from electronic annotation' (IEA). Earlier literature has cautioned against analyzing and interpreting IEA-labeled annotations[61], whereas more recent studies suggest that these are often as reliable as annotations assigned by human curators[62].

For high-confidence analyses of data from human and common model organisms that have many manually curated annotations, we generally recommend comparing versions of the analysis with and without filtering of IEA annotations to verify robustness. However, IEA annotations make up the majority of information in less-well-studied species and should be used by default in these cases. Removing IEA-coded annotations may bias the analysis toward well-studied biological processes.

Use of non-pathway gene sets

Different types of gene sets help answer a variety of questions. For instance, non-pathway gene sets corresponding to microRNA and transcription factor targets can be used to discover important regulators[52,53]. However, simultaneously analyzing all available types of gene sets reduces data interpretability. It may also lead to false negatives, as the increased number of conducted tests increases the effect of multiple testing correction and reduces the multiple-test adjusted significance of individual pathways. We therefore recommend performing the analysis of non-pathway and pathway gene sets separately.

Gene set size considerations

It is often beneficial to exclude numerous small pathways because they are redundant with larger pathways and complicate interpretation, and their abundance makes multiple-testing correction more stringent. Large pathways should be also excluded, as these are overly general (e.g., 'metabolism'), they do not contribute to interpretability of results, and their statistical significance can be inflated when using certain statistical enrichment methods (e.g., Fisher's exact test). For analyzing human gene expression data, we often recommend excluding pathway gene sets with <10–15 genes and >200–500 genes, although upper bounds of 200–2,000 genes can be found in the literature. However, for non-human organisms and other types of gene sets, which may have different gene set size distributions, larger sets may need to be included. Filtering of pathways depends on the experimental context, as different areas of biology have variable coverage in pathway databases. One can determine the lower and upper bounds of pathway size by examining the sizes of several pathways of interest that are expected to be relevant to the experiment.

Importance of using updated pathway gene sets

Pathway enrichment analysis depends on gene sets and databases used in the analysis, and many recent studies using pathway enrichment analysis are strongly impacted by outdated resources[11]. For improved reproducibility and transparency of research, investigators should report in publications the analysis date and versions of pathway enrichment analysis software and gene set databases used, as well as all analysis parameters. In addition to enrichment maps, authors should consider adding their studied gene lists and complete tables of enriched pathways as supplementary information.

Choice of gene identifier

Genes are associated with many diverse database identifiers (IDs). We recommend using unambiguous, unique and stable IDs, as some IDs become obsolete over time. For human genes, we recommend using the Entrez Gene database IDs (e.g., 4193 corresponds to MDM2) or gene symbols (*MDM2* is the official symbol recommended by the HUGO Gene Nomenclature Committee). As gene symbols change over time, we recommend maintaining both gene symbols and Entrez Gene IDs. The g:Profiler and related g:Convert tools support automatic conversion of multiple ID types to standard IDs.

Unexpected pathway results and experimental design

Unexpected biological themes revealed in a pathway analysis may indicate problems with experimental design, data generation or analysis. For example, enrichment of the apoptosis pathway may indicate a problem with the experimental protocol that led to increased cell death during sample preparation. In these cases, the experimental design and data generation should be carefully reviewed before further data interpretation.

Application to diverse omics data

This protocol uses RNA-seq data[7] and somatic mutation data[6] as examples because these data types are frequently encountered. However, the general concepts of pathway enrichment analysis that we present are applicable to many types of experiments that can generate lists of genes, such as single-cell transcriptomics, CNVs[5], proteomics[63], phosphoproteomics[64], DNA methylation[65] and

metabolomics[66]. Most data types require protocol modifications, which we only briefly discuss here. With certain data types, specialized computational methods are required to produce a gene list that is appropriate for pathway enrichment analysis, whereas with other data types, a specialized pathway enrichment analysis technique is required. Issues specific to data types and experimental methods must be considered, including:

- Different gene identifiers are recommended for certain data types. We recommend UniProt accession numbers for proteins (e.g., Q00987 for MDM2) and Human Metabolome Database IDs for metabolites (e.g., ATP is denoted as HMDB00538).
- Certain types of omics experiments by design capture only a subset of genes or proteins. To address this limited coverage, pathway enrichment analysis must define a custom background gene set of the genes that can be measured in the experiment. For example, phosphoproteomics experiments measure only proteins with one or more phosphorylation sites and thus must use the set of genes encoding phosphoproteins as the custom background gene set. Otherwise, pathway enrichment analysis would reveal inflated *P* values for general processes such as kinase signaling and protein phosphorylation.
- Pathway enrichment analysis of short non-coding genomic regions such as transcription factor binding sites from ChIP-seq experiments need additional consideration. Genomic regions must be mapped to protein-coding genes and corrected for biases such as increased signal in longer genes. Tools such as GREAT[67] automatically perform both tasks.
- Large genomic intervals that span multiple genes (e.g., from genome-wide associations, CNV and differentially methylated regions) require specialized enrichment tests such as the PLINK CNV gene set burden test[68] or INRICH[69]. Standard enrichment tests often reveal genes clustered in the genome whose signals are strongly statistically inflated because each gene is incorrectly counted as an independent signal. Gene types that are correlated with genomic position include olfactory receptors, histones, major histocompatibility complex (MHC) members and homeobox transcription factors. A simple solution to address genomic clustering of genes in a pathway involves selecting only one representative gene from each functionally homogeneous genomic cluster before enrichment analysis.
- For rare genetic variants, case–control pathway 'burden' tests are the most appropriate pathway enrichment analysis method (see 'Comparison to alternative' methods section).

Comparison to alternative methods

Pathway enrichment analysis methods

This protocol recommends the use of g:Profiler and GSEA software for pathway enrichment analysis. g:Profiler[13,39] analyzes gene lists using Fisher's exact test and ordered gene lists using a modified Fisher's test. It provides a graphical web interface and access via R and Python programming languages. The software is frequently updated, and the gene set database can be downloaded as a GMT file (http://biit.cs.ut.ee/gprofiler). GSEA[14] analyzes ranked gene lists using a permutation-based test. The software typically runs as a desktop application (http://software.broadinstitute.org/gsea). Hundreds of pathway enrichment analysis tools exist (reviewed in ref. [70]), although many rely on out-of-date pathway databases or lack unique features as compared to the most commonly used tools; as such, we do not cover them here. The following are alternative free pathway enrichment analysis software tools. Although we do not cover these tools in our protocol, we recommend the following, on the basis of their ease of use, unique features or advanced programming features.

- Enrichr[37]: This is a web-based enrichment analysis tool for non-ranked gene lists that is based on Fisher's exact test. It is easy to use, has rich interactive reporting features, and includes >100 gene set databases (called libraries), including >180,000 gene sets in multiple categories. Functionality is similar to that of the g:Profiler web server described in this protocol.
- Camera[71]: This R Bioconductor package analyzes gene lists and corrects for inter-gene correlations such as those apparent in gene co-expression data. The software is available as part of the limma package in Bioconductor (https://bioconductor.org/packages/release/bioc/html/limma.html; this is an advanced tool that requires programming expertise; Supplementary Protocol 3).
- GOseq[72]: This R Bioconductor package analyzes gene lists from RNA-seq experiments by correcting for user-selected covariates such as gene length (https://bioconductor.org/packages/release/bioc/html/goseq.html; this is an advanced tool that requires programming expertise).
- Genomic Regions Enrichment of Annotations Tool (GREAT)[67]: In contrast to common methods that analyze gene lists, GREAT analyzes genomic regions such as DNA binding sites and links these to nearby genes for pathway enrichment analysis (http://bejerano.stanford.edu/great/public/html/). See 'Application to diverse omics data' section.

Visualization tools

This protocol recommends the use of EnrichmentMap for pathway enrichment analysis visualization to aid interpretation. EnrichmentMap[16] is a Cytoscape[15] application that visualizes the results from pathway enrichment analysis and eases interpretation by displaying pathways as a network in which overlapping pathways are clustered together to identify major biological themes in the results (http://www.baderlab.org/Software/EnrichmentMap). Two alternative useful visualization tools are:

- ClueGO[40]: This Cytoscape application is conceptually similar to EnrichmentMap and provides a network-based visualization to reduce redundancy of results from pathway enrichment analysis. It also includes a pathway enrichment analysis feature for analysis of GO annotations using Fisher's exact tests. However, it currently supports only GO gene sets.
- PathVisio[49]: This desktop application provides a complementary visualization approach to those of EnrichmentMap and ClueGO. PathVisio enables the user to visually interpret omics data in the context of gene and protein interactions in a pathway of interest. PathVisio colors pathway genes according to user-provided omics data (https://www.pathvisio.org). This is the main advantage of PathVisio as compared to EnrichmentMap and ClueGO.

Topology-aware pathway analysis methods

Most pathway enrichment analysis methods treat all genes in a pathway uniformly and ignore gene interactions. By contrast, topology-aware methods explicitly model the interactions between genes. CePa[73], GANPA[74] and THINK-Back[75] use physical gene interactions or co-expression networks to assign a weight to each gene in each pathway. Weights can be derived from measures of the gene importance in the network such as degree, the number of gene connections and betweenness centrality, and can be integrated into a traditional pathway enrichment analysis method such as GSEA. Methods such as SPIA[76], Pathway-Express[77] and EnrichNet[78] generate an ES for the entire pathway that considers pathway regulatory interactions such as activation and inhibition. Although useful and potentially more accurate, regulatory and biochemical gene interactions are available for fewer genes and pathways as compared to physical interactions networks and co-expression. We anticipate that these methods will become more useful as more gene interactions in pathways are characterized in detailed molecular experiments. However collecting and curating high-quality and biochemically detailed pathway data from the literature is currently complex and expensive. Therefore, pathway enrichment analysis methods described in this protocol will probably remain the most widely used approaches for the foreseeable future.

Future perspectives

Current pathway enrichment analysis methods provide a useful high-level overview of the pathways active in a genomics experiment. However, these methods consider a simplified pathway view that involves only gene sets. Next-generation pathway analysis methods will integrate more biological pathway details, build pathway models based on multiple types of genomics data measured across many samples, and consider positive and negative regulatory relationships in the data. For instance, qualitative mathematical modeling parameterized with single-cell RNA-seq data may one day enable accurate predictions of drug combinations capable of treating a given disease under study.

Overview of the protocol

This step-by-step protocol explains how to complete pathway enrichment analysis using g:Profiler (filtered gene list) and GSEA (unfiltered, whole-genome, ranked gene list), followed by visualization and interpretation using EnrichmentMap. The example data provided for the g:Profiler analysis is a list of genes with frequent somatic single nucleotide variants (SNVs) identified in The Cancer Genome Atlas (TCGA) exome-sequencing data of 3,200 tumors of 12 types[6]. The example data provided for the GSEA analysis is a list of differentially expressed genes in two subtypes of ovarian cancer defined by TCGA[8].

Materials

Equipment

Hardware

- A personal computer with Internet access and ≥8 GB of RAM. 1 GB of RAM is sufficient to run GSEA analysis; however, Cytoscape (required to run EnrichmentMap software) requires ≥8 GB of RAM.

Software

- A contemporary web browser (e.g., Chrome) for pathway enrichment analysis with g:Profiler (Step 6A).
- g:Profiler (https://biit.cs.ut.ee/gprofiler/)
- Java Standard Edition (http://www.oracle.com/technetwork/java/javase/downloads/index.html) is required to run GSEA and Cytoscape.
- GSEA desktop application (http://software.broadinstitute.org/gsea/downloads.jsp) is used for pathway enrichment analysis (Step 6B).
- The Cytoscape desktop application (http://www.cytoscape.org/download.php), as well as the following Cytoscape applications, is required for enrichment map visualization: EnrichmentMap, v.3.1 or higher; clusterMaker2, v.0.9.5 or higher; WordCloud, v.3.1.0 or higher; AutoAnnotate, v,1.2.0 or higher. These can be conveniently downloaded and installed together by installing the 'EnrichmentMap Pipeline Collection' (http://apps.cytoscape.org/apps/enrichmentmappipelinecollection) from the Cytoscape App Store.

Input data

▲ **CRITICAL** We provide downloadable example files that are referred to throughout the protocol (Supplementary Tables 1–13). We recommend saving all these files in a personal *project data folder* before starting. We also recommend creating an additional *result data folder* to save the files generated while performing the protocol.

- A gene list or ranked gene list of interest
  - *Example data for Step 6A*. g:Profiler requires a list of genes, one per line, in a text file or spreadsheet, ready to copy and paste into a web page: for this, we use genes with frequent somatic SNVs identified in TCGA exome sequencing data of 3,200 tumors of 12 types[6]. The MuSiC cancer driver mutation detection software was used to find 127 cancer driver genes that displayed higher than expected mutation frequencies in cancer samples (Supplementary Table 1, which is derived from column B of Supplementary Table 4 in ref. [6]). Genes are ranked in decreasing order of significance (FDR *Q* value) and mutation frequency (not shown).
  - *Example data for Step 6B*. GSEA requires an RNK file with gene scores. An RNK file is a two-column text file with gene IDs in the first column and gene scores in the second column. All (or most) genes in the genome need to have a score, and the gene IDs need to match those used in the GMT file. We provide a ranked list of differentially expressed genes in ovarian cancer from TCGA (Supplementary Table 2). This cohort was previously stratified into four molecular subtypes on the basis of gene expression data, defined as differentiated, immunoreactive, mesenchymal and proliferative[7,8]. We compared the immunoreactive and mesenchymal subtypes to demonstrate the protocol. Step 5 of Supplementary Protocol 1 shows how this file was created.
- Pathway gene set database
  - In Step 6A, g:Profiler maintains an up-to-date set of pathway gene sets from multiple sources and no further input from the user is required, but a database of pathway gene sets is required for Step 6B (GSEA). Supplementary Table 3 contains a database of pathway gene sets used for pathway enrichment analysis in the standard GMT format, downloaded from http://baderlab.org/GeneSets. This file contains pathways downloaded on 1 July 2017 from eight data sources: GO[57], Reactome[58], Panther[38], NetPath[60], NCI[79], MSigDB curated gene sets (C2 collection, excluding Reactome and KEGG)[80], MSigDB Hallmark (H collection)[81] and HumanCyc[59]. The gene sets available from http://baderlab.org/GeneSets are updated monthly. A GMT file is a text file in which each line represents a gene set for a single pathway. Each line includes a pathway ID, a name and the list of associated genes in a tab-separated format.

Procedure

Software installation ● Timing 5 min

1 Download the required input and output files from the Supplementary Materials of the protocol.
   - Create two directories, *project data folder* and *results data folder*.
   - Place all downloaded input and example output files into the project data folder.
   - As you progress through the protocol, place any newly generated files into the results data folder.

2 Install Java v.8 or higher. Follow the download and installation instructions for Java JRE at http://www.oracle.com/technetwork/java/javase/downloads/index.html.
? TROUBLESHOOTING

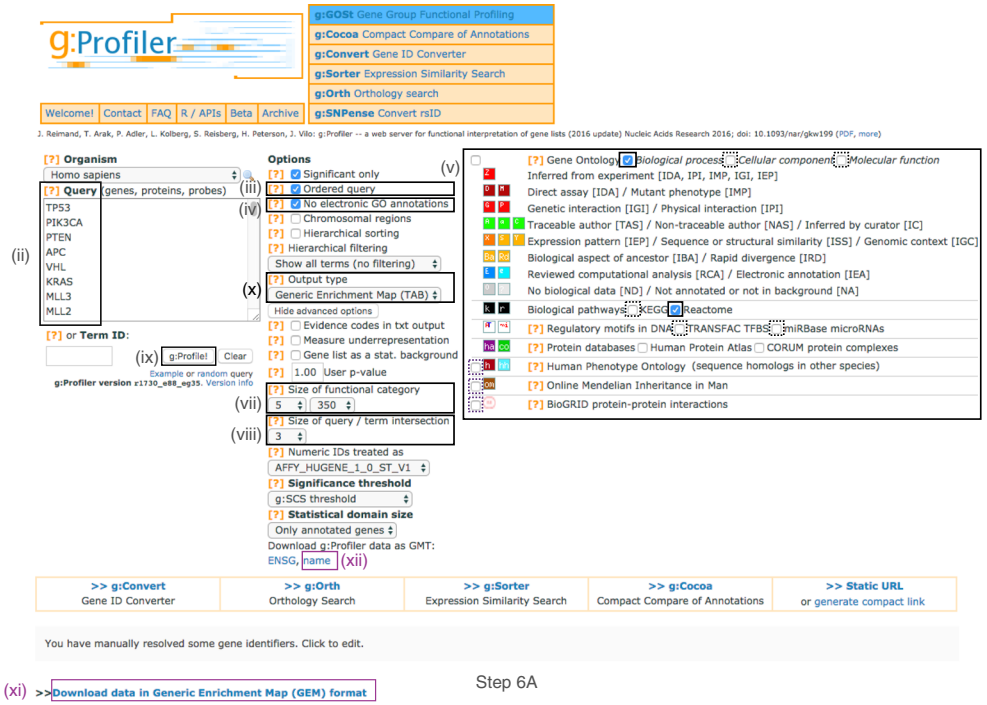


Fig. 2 | Screenshot of g:Profiler user interface. Protocol Step 6A involves populating the g:Profiler interface. Procedural steps are highlighted with rectangles and roman numerals (refer to Step 6A(i–xii)). Purple boxes highlight files that must be downloaded for subsequent protocol steps. The remaining boxes indicate parameters for the analysis.

3 Download the latest version of GSEA. We recommend the javaGSEA desktop application available at http://www.broadinstitute.org/gsea/downloads.jsp. Free registration is required.

4 Download the latest version of Cytoscape from http://www.cytoscape.org. Cytoscape v.3.6.0 or higher is required.

5 Install the required Cytoscape applications.
- Launch Cytoscape.
- Go to *Apps* → *App Manager* (i.e., open the *Apps* menu and select the item *App Manager*).
- In the *Install Apps* tab search bar, search for *EnrichmentMap*.
- Click on *EnrichmentMap Pipeline Collection* in the center panel. Verify that it is v.1.0.0 or higher.
- Click on the *Install* button.
- Go to the *Currently Installed* tab and verify that the applications (EnrichmentMap, clusterMaker2, WordCloud and AutoAnnotate) have been installed.

Pathway enrichment analysis ● Timing 3–20 min

6 Two major types of gene lists are used in pathway enrichment analysis of omics data. Flat (unranked) gene lists of dozens to thousands of genes can be analyzed using g:Profiler (option A). A statistical threshold is required to compile a gene list from omics data. By contrast, ranked, whole-genome gene lists are suitable for pathway enrichment analysis using GSEA (option B). Gene lists analyzed with GSEA do not require prior filtering using statistical thresholds. Partial, filtered ranked gene lists can also be analyzed with g:Profiler. Select Step 6A or 6B, depending on the type of gene list you have.

(A) **Pathway enrichment analysis of a gene list using g:Profiler ● Timing 3 min**

(i) Open the g:Profiler website at http://biit.cs.ut.ee/gprofiler/ (Fig. 2).

(ii) Paste the gene list (Supplementary Table 1) into the *Query* field in top-left corner of the screen. The gene list can be space-separated or one per line. The organism for the analysis, *Homo sapiens*, is selected by default. The input list can contain a mix of gene and protein IDs, symbols and accession numbers. Duplicated and unrecognized IDs will be removed automatically, and ambiguous symbols can be refined in an interactive dialogue after submitting the query.

(iii) Check the box next to *Ordered query*. This option treats the input as an ordered gene list and prioritizes genes with higher mutation ESs at the beginning of the list.

(iv) (Optional) Check the box next to *No electronic GO annotations*. This option will discard less reliable GO annotations (IEAs) that are not manually reviewed.

(v) Set filters on gene annotation data using the menu on the right. We recommend that initial pathway enrichment analyses includes only biological processes (BPs) of GO and molecular pathways of Reactome. Keep the two checkboxes checked and uncheck all other boxes in the menu.

(vi) Click on *Show Advanced Options* to set additional parameters.

(vii) Set the values of *Size of functional category* in the dropdown menu to 5 ('min') and 350 ('max'). Large pathways are of limited interpretative value, whereas numerous small pathways decrease the statistical power because of excessive multiple testing.

(viii) Set the *Size of query/term intersection* in the dropdown menu to 3. The analysis will consider only more reliable pathways that have three or more genes in the input gene list.

(ix) Click *g:Profile!* to run the analysis. A graphical heat map image will be shown, with detected pathways shown along the *y* axis (left) and associated genes of the input list shown along the *x* axis (top). Resulting pathways are organized hierarchically into related groups. g:Profiler uses graphical output by default and switches to textual output when a large number of pathways is found. g:Profiler returns only statistically significant pathways with *P* values adjusted for multiple testing correction (called *Q* values). By default, results with *Q* values <0.05 are reported. g:Profiler reports unrecognized and ambiguous gene IDs that can be resolved manually.

(x) Use the dropdown menu *Output type* and select the option *Generic Enrichment Map (TAB)*. This file is required for visualizing pathway results with Cytoscape and EnrichmentMap.

(xi) Click *g:Profile!* again to run the analysis with the updated parameters. The required link *Download data in Generic Enrichment Map (GEM) format* will appear under the g:Profiler interface. Download the file from the link and save it on your computer in your *result data folder* created in Step 1. Example results are provided in Supplementary Table 4.

(xii) Download the required GMT file by clicking on the link *name* at the bottom of the *Advanced Options* form. The GMT file is a compressed ZIP archive that contains all gene sets used by g:Profiler (e.g., gprofiler_hsapiens.NAME.gmt.zip). The gene set files are divided by data source. Download and uncompress the ZIP archive to your project folder. All required gene sets for this analysis will be in the file hsapiens.pathways.Name.gmt (Supplementary_Table5_hsapiens.pathways.NAME.gmt). Place the saved file in your *result data folder* created in Step 1.

(B) **Pathway enrichment analysis of a ranked gene list using GSEA** ● **Timing ~20 min**

(i) Launch GSEA by opening the downloaded GSEA file (gsea.jnlp) (Fig. 3).
? TROUBLESHOOTING

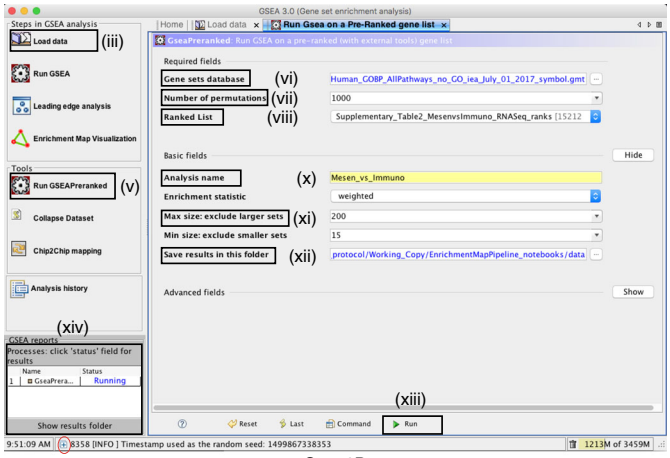
Loading the required data files into GSEA

(ii) Click on *Load Data* in the top left corner of the *Steps in GSEA Analysis* section.
? TROUBLESHOOTING

(iii) In the *Load Data* tab, click on *Browse for files …*

(iv) Find your *project data folder* and select the Supplementary_Table2_MesenvsImmuno_RNASeq_ranks.rnk file. Also select the pathway gene set definition (GMT) file using a multiple-select method such as shift-click (Supplementary Table 3). Click the *Choose* button to continue. A message box indicates that the files were loaded successfully. Click the *OK* button to continue.
▲ **CRITICAL STEP** GSEA also supplies its own gene set files, which are accessible directly through the GSEA interface from the MSigDB resource[80,81]. These files do not need to be imported into GSEA. To define the GMT file, find the MSigDB gene set files in the first tab, *Gene Matrix (from website)*, of the *Select one or more genesets* dialog. The latest versions of the MSigDB gene set files are shown in bold, but the earlier versions can also be accessed. To select multiple gene set files, click on the desired files while holding the control key in Windows or the command key in macOS.

(v) Click on *Run GSEAPreranked* in the side bar under *Tools*. The *Run GSEA on a Pre-Ranked gene list* tab will appear.
? TROUBLESHOOTING



Step 6B

Fig. 3 | Screenshot of GSEA user interface. Step 6B involves populating the GSEA interface (v.3.0). Procedural steps are highlighted with rectangles and roman numerals (refer to Step 6B(i–xiv)). At the bottom left corner of the screen there is a '+' sign (circled in red at the bottom of the figure). Click on the '+' to see progress messages such as 'shuffleGeneSet for GeneSet 4661/4715 nperm: 1000'. This message indicates that GSEA is shuffling 4,715 gene sets 1,000 times each, 4,661 of which are complete.

Specification of the following parameters:

(vi) *Gene sets database*. Click on the button '(…)' located to the right and wait a few seconds for the gene set selection window to appear. Go to the *Gene matrix (local gmx/gmt)* tab by using the top right arrow. Click on the downloaded local GMT file, 'Supplementary_Table3_Human_GOBP_AllPathways_no_GO_iea_July_01_2017_symbol.gmt', and click on *OK* at the bottom of the window.

(vii) *Number of permutations*. This specifies the number of times that the gene sets will be randomized to create the null distribution to calculate the *P* value and FDR *Q* value. Use the default value of 1,000 permutations.

▲ **CRITICAL STEP** Higher numbers of permutations require longer computation times. To calculate the FDR *Q* value for each gene set, the dataset is randomized by permuting the genes in each gene set and recalculating the *P* values for the randomized set. This parameter specifies how many times this randomization is done. The more randomizations are performed, the more precise the FDR *Q* value estimation will be (to a limit, as eventually the FDR *Q* value will stabilize at the actual value). On a Windows machine with 16 G of RAM and an i7 3.4-GHz processor, an analysis with 10, 100, 500 or 1,000 randomizations on our example set with the above defined parameters takes 155, 224, 544, and 1,012 s, respectively.

(viii) *Ranked List*. Select the ranked gene list by clicking on the right-most arrows and highlighting the rank file (Supplementary Table 2).

(ix) Click the *Show* button next to *Basic fields* to display additional options.

(x) *Analysis name*. Change the default 'my_analysis' to a specific name, for example, 'Mesen_vs_Immuno'.

(xi) *Max size: exclude larger sets*. By default, GSEA sets the upper limit to 500. Set this to 200 to remove the larger sets from the analysis.

(xii) *Save results in this folder*. Navigate to the folder where GSEA should save the results. We recommend you choose the *result data folder* created in Step 1. Otherwise, GSEA will use the default location 'gsea_home/output/[date]' in your home directory.

Running GSEA

(xiii) Run GSEA by clicking on the *Run* button located at the bottom of the window. Expand the window if the button is not visible. The *GSEA reports* pane at the bottom left of the window will show the status 'Running'. It will be updated to 'Success' upon completion. This is expected to be a long-running process, depending on the speed of your computer.

? TROUBLESHOOTING

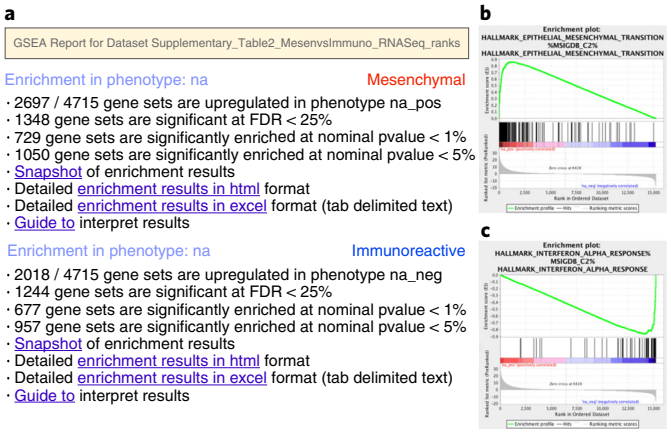


Fig. 4 | GSEA output overview. a, Web page summary of GSEA results showing pathways enriched in the top or bottom of the ranked list, with the 'na_pos' and 'na_neg' phenotypes corresponding to enrichment in upregulated and downregulated genes, respectively. These have been manually labeled here as mesenchymal and immunoreactive, respectively. Clicking on 'Snapshot' under either of the phenotypes will show the top 20 enrichment plots for that phenotype. **b**, An example enrichment plot for the top pathway in the mesenchymal set. **c**, An example enrichment plot for the top pathway in the immunoreactive set.

Examination of GSEA results

(xiv) Once the GSEA analysis is complete, a green notification 'Success' will appear in the bottom left section of the screen. All GSEA output files will be automatically saved and be available in the folder you specified in the GseaPreranked interface (Step 6B(xii)). Click on *Success* to open the results in your web browser. Pathways enriched in top-ranking genes (i.e., upregulated) are shown in the first set ('na_pos'; 'mesenchymal' in this protocol) and pathways enriched in bottom-ranked genes (i.e., downregulated) are shown in the second set ('na_neg'; 'immunoreactive') (Fig. 4).

? TROUBLESHOOTING

(xv) In the web browser results summary, click on the *Snapshot* link under the results to get an overview of the top 20 findings. The most significant pathways for the first phenotype ('na_pos') should clearly display enrichment in top-ranking (i.e., upregulated) genes (left side of the plot). Conversely, the most significant pathways for the second phenotype ('na_neg') should clearly display enrichment in bottom-ranked (i.e., downregulated) genes (right side of the plot) (Fig. 4).

▲ CRITICAL STEP When running GSEA with expression data as input (instead of a pre-calculated rank file), a phenotype label (i.e., biological condition or sample class) is provided as input for each sample and specified in the GSEA 'cls' file. When running GSEA, the two phenotypes to compare for differential gene expression analysis are specified and these phenotypes are used in the pathway enrichment result files. By contrast, in a GSEA preranked analysis (i.e., when a ranked gene list is provided by the user), GSEA automatically labels one phenotype 'na_pos' (corresponding to enrichment in the genes at the top of the ranked list, where 'na' means the phenotype label is 'not available') and the other 'na_neg' (corresponding to enrichment in the genes at the bottom of the ranked list). This convention is also used by the EnrichmentMap software, designating the first phenotype as 'positive' and the second phenotype as 'negative'.

? TROUBLESHOOTING

(xvi) In the web browser results summary, click on *Detailed enrichment results in HTML format* and use the row numbering to check the number of pathways that have FDR Q values <0.05 to determine appropriate thresholds for EnrichmentMap in the next step of the protocol. If no pathways are reported at $Q < 0.05$, more lenient thresholds such as $Q < 0.1$ or $Q < 0.25$ could be used (Fig. 5). The threshold $Q < 0.25$ provides very lenient filtering, and it is not uncommon to find thousands of enriched pathways at this level. Robust analyses should use a cutoff of $Q < 0.05$ or lower. Filtering only by uncorrected P values is inappropriate and not recommended.

? TROUBLESHOOTING

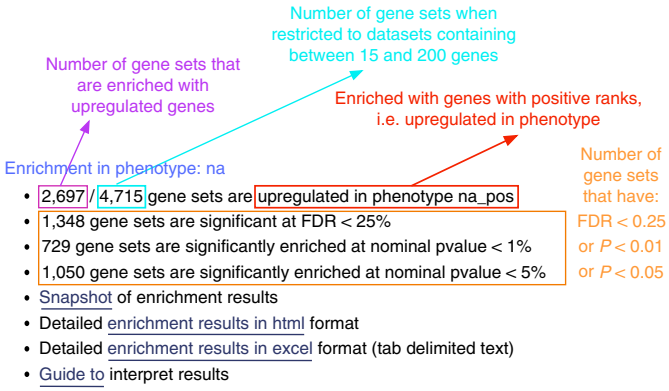


Fig. 5 | Class/phenotype-specific GSEA output. Class/phenotype-specific GSEA output in the web page summary shows how many gene sets were found enriched in upregulated genes, regardless of significance (purple), the total number of gene sets used after size filtering (cyan), the phenotype name (red) and the number of gene sets that pass different thresholds (orange).

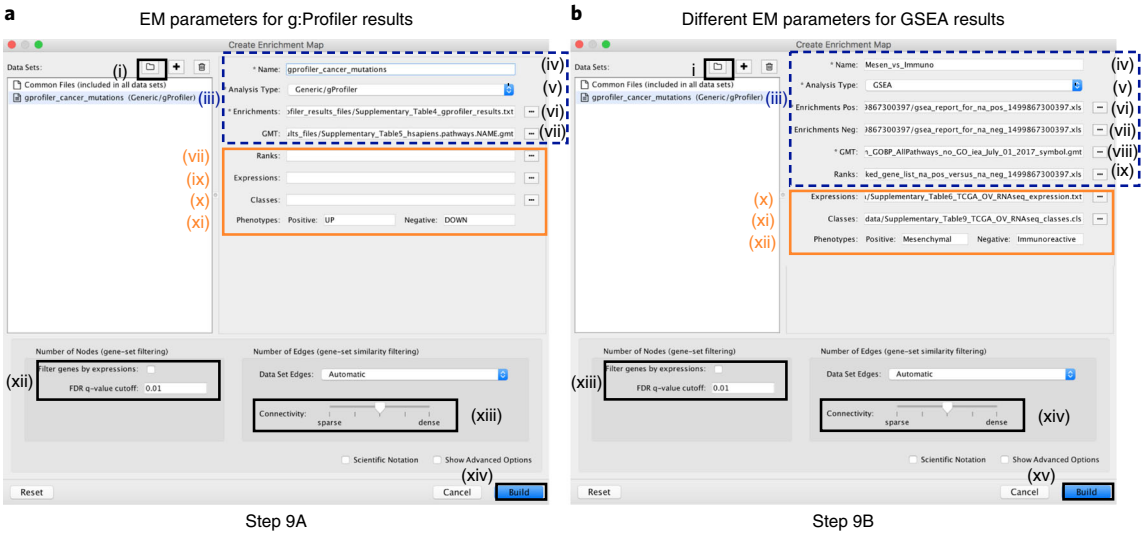


Fig. 6 | Screenshot of the EnrichmentMap software user interface. a,**b**, Input fields in the EnrichmentMap interface for g:Profiler (**a**) and GSEA (**b**) results. Procedural steps are shown for Step 9A and 9B. Other than the specific input files, the parameters are the same for the two analysis types. Attributes surrounded by a dashed box should be filled out automatically if the user selects an appropriate folder with the required files. Missing file names indicate that EnrichmentMap was unable to find the specified file. Orange boxes indicate optional files. For the examples presented in the protocol, optional files are used for the GSEA analysis but not for the g:Profiler analysis to demonstrate the two distinct use cases. EM, EnrichmentMap.

Visualization of enrichment results with EnrichmentMap ● Timing ~5 min

7 Launch the Cytoscape software. Introductory Cytoscape tutorials can be found at http://tutorials.cytoscape.org.

8 In the menu, click *Apps* → *EnrichmentMap*.

9 The *Create Enrichment Map* panel will appear (Fig. 6). Creation of enrichment maps with g:Profiler (option A) and GSEA (option B) requires different input files.

(A) **Creation of enrichment maps for g:Profiler results generated in Step 6A**

(i) In the *Create Enrichment Map* panel, click on folder icon (Fig. 6a).

(ii) Locate and select your *result data folder* containing the g:Profiler results and click on *Open*. **? TROUBLESHOOTING**

(iii) In the right-hand pane, g:Profiler output files will be automatically populated into their specified fields. Alternatively, users can click on the '+' symbol to specify each of the required files manually.

(iv) In the right-hand pane, modify the *Name* of the created dataset if desired. By default, EnrichmentMap will use the name of the g:Profiler enrichment results file (e.g., 'Supplementary_Table4_gprofiler_results.txt').

(v) Verify that the *Analysis Type* is set to *Generic/gProfiler*.

(vi) Verify the *Enrichments* results file is the g:Profiler file downloaded in Step 6A(xi) (or alternatively, manually specify 'Supplementary_Table4_gprofiler_results.txt').

(vii) Verify the *GMT* specified is the file retrieved from the g:Profiler website in Step 6A(xii). Use the file 'hsapiens.pathways.NAME.gmt' (or alternatively manually specify 'Supplementary_Table5_hsapiens.pathways.NAME.gmt') that contains the gene sets corresponding to GO biological processes and Reactome pathways.

Specification of additional files:

(viii) *Expressions*. (Optional) Upload an expression matrix for the genes analyzed in g:Profiler or, alternatively, upload an expression dataset of all genes. If the expression dataset contains additional genes not used for the g:Profiler search, their expression values will still appear in the heat map of the enrichment map (for an example file, see Supplementary Table 6).

(ix) *Ranks*. (Optional) Ranks for the gene list or the expression data can be specified (for an example, see Supplementary Table 2).

(x) *Classes*. (Optional) This is a GSEA CLS file defining the phenotype (i.e., biological conditions) of each sample in the expression file; for an example, see Supplementary Table 7. This file is required only for phenotype randomization in GSEA; however, providing it to EnrichmentMap will label the columns of the expression file in the EnrichmentMap heat map viewer by phenotype.

(xi) *Phenotypes*. (Optional) If there are two different phenotypes in the expression data, update the phenotype labels so that 'positive' represents the phenotype associated with positive values (mesenchymal in this example) and 'negative' represents that associated with negative values (immunoreactive in this example).

? TROUBLESHOOTING

Tuning parameters

(xii) *Number of Nodes*. By default, g:Profiler returns only statistically significant results ($Q < 0.05$), so the *FDR q-value cutoff* parameter can be set to 1 in the EnrichmentMap Input panel, unless a more stringent filtering is desired. For this protocol, set the FDR Q value to 0.01. (Optional) Select *Filter genes by expressions* to exclude any genes in the gene set definition file (i.e., the GMT file) that are not found in the supplied expression file. If *Filter genes by expressions* is not selected, any gene that is not found in the expression file will be retained and will be presented in the expression heat map viewer with all of its associated expression values grayed out.

(xiii) *Number of Edges*. Keep the connectivity slider in the center. If the network is too cluttered because of too many connections (edges), move the slider to the left to make the network sparser. Alternatively, if the network is too sparse (i.e., there are too many disconnected pathways), move the slider to the right to obtain a more densely connected network.

▲CRITICAL STEP Moving the slider to the left (or right) will adjust the underlying similarity statistic threshold to make the resulting network sparser (or denser). The slider is set with predefined defaults, but users can fine-tune the similarity metric by selecting *Show advanced options* at the bottom of the *Create Enrichment Map* panel. Predefined values appear as tick marks on the slider and include Jaccard > 0.35, Jaccard > 0.25, combined > 0.375, overlap > 0.5, and overlap > 0.25.

(xiv) Click the *Build* button at the bottom of the EnrichmentMap *Input* panel. A *Building EnrichmentMap* box appears and indicates the progress status. This box will disappear once the map has been created successfully. See Fig. 7a for the resulting enrichment maps from the g:Profiler analysis.

(B) **Creation of enrichment maps from GSEA results generated in Step 6B**:

(i) In the *Create Enrichment Map* panel, click on the folder icon and locate the GSEA results folder created in Step 6B (Fig. 6).

(ii) Click on the GSEA folder to select it. Click on *Open*.

▲CRITICAL STEP If you specify a directory that contains multiple GSEA results, rather than an individual GSEA results folder, EnrichmentMap will treat each GSEA results folder as a separate dataset. This enables easy multi-dataset analyses. If you want only one dataset but inadvertently selected the directory containing multiple GSEA results instead of

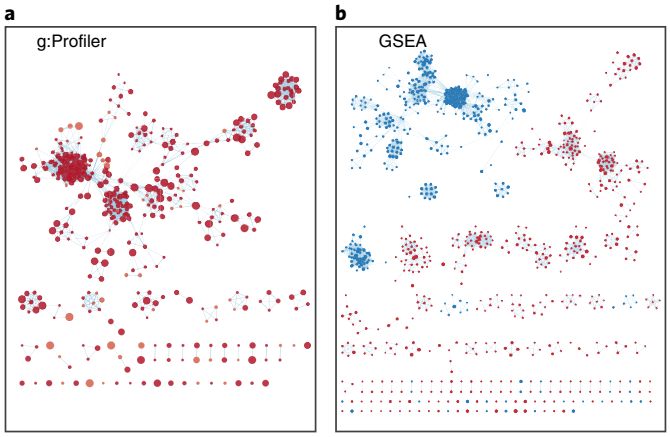


Fig. 7 | Resulting enrichment maps (no manual formatting). a,**b**, Unformatted enrichment maps generated from Steps 6A (**a**) and 6B (**b**). Each node (circle) represents a distinct pathway, and edges (blue lines) represent the number of genes overlapping between two pathways, determined using the similarity coefficient. **a**, Enrichment map of significantly mutated cancer driver genes generated using the g:Profiler analysis in Step 6A. **b**, Enrichment map of pathways enriched in upregulated genes in mesenchymal (red) and immunoreactive (blue) ovarian cancer samples using the GSEA analysis in Step 6B.

selecting an individual folder, simply select the datasets you do not want to use and click on the trash can at the top of the EnrichmentMap input panel to remove them.

(iii) In the right-hand pane, GSEA output files will be auto-populated into their specified fields. Alternatively the '+' symbol can be clicked to specify each of the required files manually.

(iv) In the right-hand pane, modify the *Name* of the created dataset if desired. By default, EnrichmentMap will use the first part of the GSEA folder name before the last dot (.) to create the dataset name. For example, if the directory is called Mesen_vs_Immuno.GseaPreranked.12345, the name will be populated as Mesen_vs_Immuno.GseaPreranked.

(v) Verify that the *Analysis Type* is set to *GSEA*.

(vi) *Enrichments Pos*. Verify that the file name is set to '[your_path_to_gsea_dir]/Mesen_vs_Immuno.GseaPreranked.12345/gsea_report_for_na_pos_12345.xls', where '12345' is a unique number generated by GSEA. Alternatively navigate to the 'Supplementary_Table8_gsea_report_for_na_pos.xls' file.

? TROUBLESHOOTING

(vii) *Enrichments Neg*. Verify that the file name is set to '[your_path_to_gsea_dir]/gsea_report_for_na_neg_12345.xls', where '12345' is a unique number generated by GSEA. Alternatively navigate to the file 'Supplementary_Table9_gsea_report_for_na_neg.xls'.

? TROUBLESHOOTING

(viii) *GMT*. Verify that the file name is set to 'Supplementary_Table3_Human_GOBP_AllPathways_no_GO_iea_July_01_2017_symbol.gmt'. Alternatively, navigate to the file 'Supplementary_Table3_Human_GOBP_AllPathways_no_GO_iea_July_01_2017_symbol.gmt'.

? TROUBLESHOOTING

(ix) *Ranks*. Verify that the file name is set to 'ranked_gene_list_na_pos_versus_na_neg_12345.xls', where '12345' is a unique number generated by GSEA. Alternatively, navigate to the 'Supplementary_Table2_MesenvsImmuno_RNASeq_ranks.rnk' file.

? TROUBLESHOOTING

Specification of additional files

(x) *Expressions*. (Optional) Upload an expression matrix for the genes analyzed in GSEA. For an example file, see 'Supplementary_Table6_TCGA_OV_RNAseq_expression.txt'.

(xi) *Classes*. (Optional) This is a GSEA CLS file defining the phenotype (i.e., biological conditions) of each sample in the expression file. For an example, see 'Supplementary_Table7_TCGA_OV_RNAseq_classes.cls.' This file is required only for phenotype randomization in GSEA; however, providing it to EnrichmentMap will label the columns of the expression file in the EnrichmentMap heat map viewer by phenotype.

(xii) *Phenotypes*. (Optional) In the text boxes replace, 'na_pos' with 'Mesenchymal' and 'na_neg' with 'Immunoreactive'. 'Mesenchymal' will be associated with red nodes because it

corresponds to the positive phenotype, whereas 'Immunoreactive' phenotypes will be labeled blue.

▲ **CRITICAL STEP** If you load the CLS file before specifying the phenotypes, EnrichmentMap will automatically guess the phenotypes from the class file. If your class file specifies more than two phenotypes, EnrichmentMap will choose the first two phenotypes defined in the file. To annotate the phenotypes in the EnrichmentMap heat map, the specified phenotype labels need to exactly match the GSEA CLS file.

Tuning parameters

(xiii) *Number of Nodes*. Set the FDR *Q* value cutoff to 0.01. (Optional) Select *Filter genes by expressions* to exclude any genes in the gene set definition file (i.e., the GMT file) that are not found in the supplied expression file. If *Filter genes by expressions* is not selected, any gene that is not found in the expression file will be retained and will be presented in the expression viewer with all of its associated expression values grayed out.

? TROUBLESHOOTING

(xiv) *Number of Edges*. Keep the connectivity slider in the center. To create networks with fewer edges,(a sparser network), move the slider to the left. Alternatively, to create networks with more edges (a denser network), move the slider to the right.

▲ **CRITICAL STEP** Moving the slider to the left (or right) will adjust the underlying similarity statistic threshold to make the resulting network sparser (or denser). The slider is set with predefined defaults, but users can fine-tune the similarity metric by selecting *Show advanced options* at the bottom of the *Create Enrichment Map* panel. Predefined values appear as tick marks on the slider and include Jaccard > 0.35, Jaccard > 0.25, combined > 0.375, overlap > 0.5, and overlap > 0.25.

(xv) Click the *Build* button at the bottom of the EnrichmentMap Input panel. A *Building EnrichmentMap* box appears and indicates the progress status. This box will disappear once the map has been created successfully. See Fig. 7b for the resulting enrichment map from the GSEA analysis.

Navigation and interpretation of the enrichment map ● Timing ~4 h

▲ **CRITICAL** An enrichment map must be interpreted to discover novel information about a dataset and must be manually refined to create a publication-quality figure.

10 To explore the enrichment map, select the network of interest in the *Control Panel* located at the left side of the Cytoscape window (Fig. 8). The *Network Panel* can be selected using the leftmost tab of the *Control Panel*. The selected network will appear in the main window; navigate to it (zoom and pan) using Cytoscape controls (Fig. 8 (**i**)), and explore the pathways by reading the gene set labels. Pathways with many common genes often represent similar biological processes and are grouped together as sub-networks or themes in the network. Click on a node to display the corresponding genes in the table below the network view (Fig. 8 (**ii**)).

? TROUBLESHOOTING

11 To find a gene or pathway of interest, type its name in the search bar located in the top right corner (Fig. 8 (**iii**)). All pathways containing that gene will be highlighted. For example, *TP53* and *BGN* are the top genes in g:Profiler and GSEA analyses, respectively.

? TROUBLESHOOTING

12 To find the most enriched pathways, look at the *Table Panel* located below the network view. Select the *Node Table* tab (Fig. 8 (**iv**)), and select and sort the column named 'EM#_fdr_qvalue' (for g:Profiler) or 'EM#_NES' (for GSEA) by clicking on the column header. To highlight a subset of these pathways in the network, select rows corresponding to the pathways of interest, right-click on any selected row in the table and select *Select nodes from selected rows*.

? TROUBLESHOOTING

13 Depending on the focus of the analysis, there are different actions that can be taken on the resulting enrichment map. Follow option A to explore the *Table Panel* heat map; option B to organize and clarify the network; option C to define major biological themes; option D to create a simplified network view; option E to manually arrange the network nodes; and option F to create a subnetwork that highlights a specific theme subset. Skip ahead to Step 14 to save the image and to generate legends.

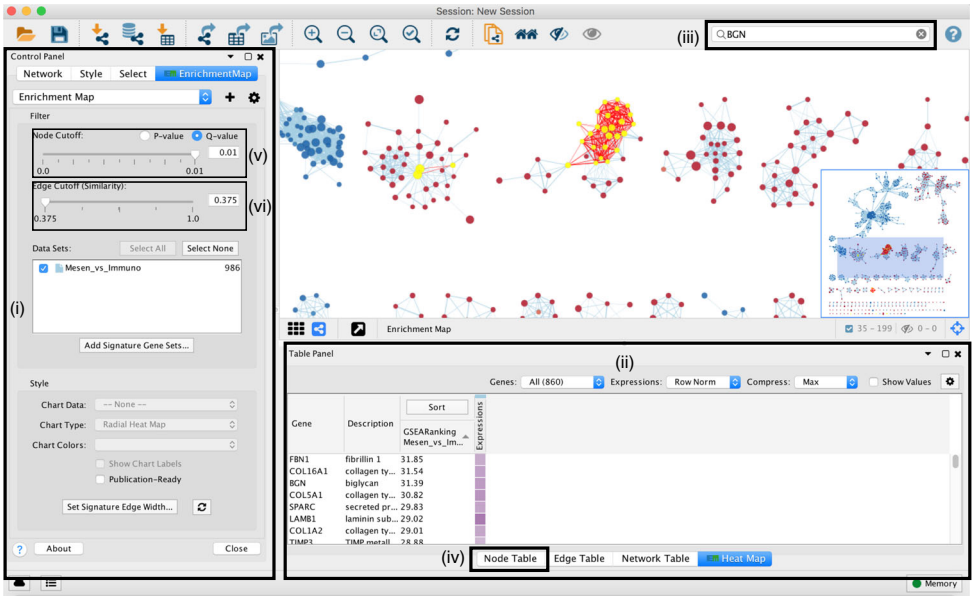


Fig. 8 | Overview of EnrichmentMap panels in Cytoscape. (**i**) Cytoscape 'Control Panel', which contains 'Networks', 'Styles' and 'Select' tabs as well as the 'EnrichmentMap' main panel. (**ii**) The 'Table Panel' contains tables with node, edge and network attributes, as well as an enrichment map 'Heat Map' panel displaying expression for genes associated with selected nodes and edges. (**iii**) Cytoscape search bar, which can be used to search for genes in the enrichment map. (**iv**) 'Node Table' containing values for all variables associated with each node in the network. (**v**) *Q*-value or *P*-value slider bar. By default, the slider is set to *Q* value if the data contains *Q* values but can be changed to use *P* values by selecting the 'P-value' radio button. All nodes that pass the initial *Q*-value threshold are displayed in the enrichment map. By moving the slider to the left, the *Q*-value threshold is adjusted to a lower value, removing any nodes that do not pass the *Q*-value threshold. The currently set threshold will be displayed in the accompanying text box. Thresholds can be manually adjusted by modifying the text box value directly. (**vi**) 'Edge Cutoff (Similarity)' slider bar. The slider bar modifies the similarity threshold. The similarity threshold can only be increased; i.e., edges are required to have more genes in common in order to remain visible, which will remove edges from the network that do not satisfy the threshold. One can also manually change the threshold by modifying the text box value directly.

(A) **Exploring the *Table Panel* heat map** ● Timing **45 min**

(i) When a gene expression matrix is provided as input to EnrichmentMap, we can study the expression pattern of the genes included in enriched pathways. Click on either an individual node or a group of nodes to generate a gene expression heat map that will appear in the *Heat Map* tab of the *Table Panel* (Fig. 9). If the analysis is based on GSEA results and a rank file is supplied, the 'leading edge' genes will be highlighted in yellow for individual node selections. Several options for heat map visualization are available.

▲ **CRITICAL** If no expression file is provided to EnrichmentMap as input, it will automatically create a dummy expression file in which any gene found in the enrichment file will be given a placeholder expression value of 0.25, and any gene found in a pathway but not found in the enrichment results file will be assigned a placeholder expression value of 'NA' (not applicable). Therefore, clicking on any node in the enrichment map will show the genes used for the analysis as well as genes in the pathway that are not part of the query set.

? TROUBLESHOOTING

(ii) Adjust the *Sort* option. Sorting options include hierarchical clustering, ranks or no sorting. To change the sorting option, click on the *Sort* button visible in the top left corner of the heat map table (Fig. 9 (**i**)). By default, the heat map is sorted by ranks if a rank file is supplied. In the absence of a rank file, no sort is applied. Additional rank files can be uploaded for comparison through the *Settings* menu located at the top right corner in the *Heat Map* panel (Fig. 9 (**vi**)). Choosing which rank file to sort from can be done by clicking on the *Sort* button and selecting the rank file by name. Click the arrow next to the currently sorted column to invert the sort order. Click any of the column names to sort the selected column.

? TROUBLESHOOTING

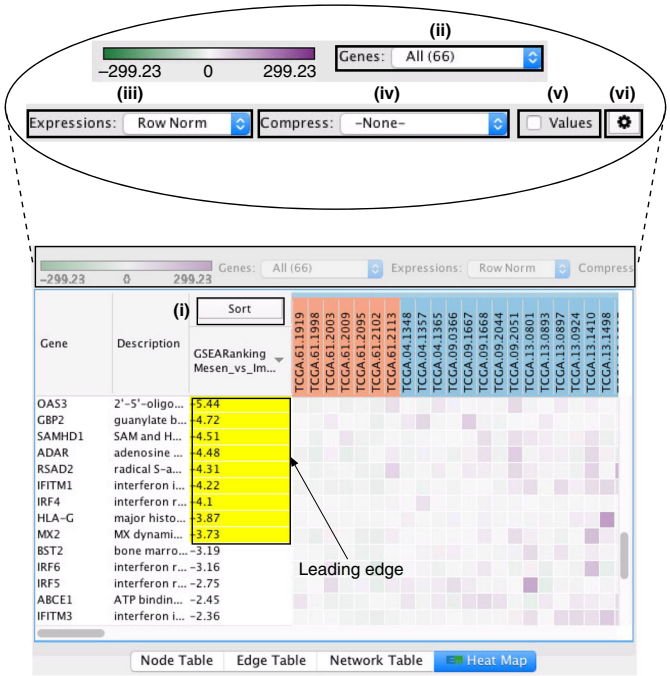


Fig. 9 | Example heat map in EnrichmentMap. Heat map created by selecting the immunoreactive pathway interferon alpha beta signaling pathway from Reactome. The heat map is useful for visualization of detailed gene expression patterns for a pathway of interest. Magenta corresponds to high expression, and green corresponds to low expression. This heat map is for GSEA results, thus the 'leading edge' genes are highlighted in yellow; these genes have the largest contribution to the enrichment signal. (**i**–**vi**) Additional controls in the *Heat Map* panel include sorting options (**i**), selection of genes to include (**ii**), expression data visualization options (**iii**), data compression options (**iv**), the option to show values (**v**) and heat map settings (**vi**).

(iii) Define *Genes* you wish to visualize in the heat map (Fig. 9 (**ii**)). Data can be viewed for all genes contained in the selected nodes (union of nodes) or just for the genes common to selected nodes (intersection of nodes). By default, all genes are shown.

(iv) Change the *Expressions* value visualization depending on your data type (Fig. 9 (**iii**)). Data can be viewed as they were loaded (*Values*), as row-normalized, in which case the row mean is subtracted from each value and then divided by the row's standard deviation (*Row Norm*), or as log-transformed (*Log*).

(v) *Compress* heat map columns (Fig. 9 (**iv**)). By default, all expression values are visible as individual columns in the heat map for expression sets with <50 samples. It is possible to compress the data into a single column by selecting one of the aggregation methods—*Median*, *Max* (maximum) or *Min* (minimum)—listed under *Compress*. If a CLS file has been uploaded, the expression set can be compressed using one column per defined sample group using the *Class* option. If the expression matrix contains ≥50 samples, EnrichmentMap will automatically compress the values to their median value by default.

(vi) Check *Values* (Fig. 9 (**v**)) to show the expression numerical values in addition to the heat map color scale.

(vii) Perform additional fine-tuning of the heat map using the *Settings* panel, accessed by clicking on the cog icon (Fig. 9 (**vi**)). This includes functionality to add new rank files, export the heat map data as a tab-delimited text file or PDF image, change the distance metric for hierarchical clustering, or turn on the node table heat map autofocus. The resulting heat map can be seen in Fig. 9. In this figure, genes are sorted using the GSEA rank file, highlighting the leading edge in yellow. All genes contained in selected nodes are shown, expression values are row-normalized, no compression is applied and individual expression numerical values are not shown. Column headings are colored according to sample phenotype. Red color refers to the first phenotype (mesenchymal), and blue to the second phenotype (immunoreactive).

? TROUBLESHOOTING

(viii) The heat map can be exported to a text file for further analysis: click on the *Settings* icon of the heat map (Fig. 9 (**vi**)) and select *Export as TXT*.

(ix) If only an individual node is selected, a dialog will offer to export the *Leading edge only* for GSEA analysis. If selected, only the highlighted genes will be exported; otherwise, the entire set of genes is saved.

? TROUBLESHOOTING

(x) Specify the file name and location and click *Save*.

(B) **Organization and clarification of the network** ● **Timing 30 min**

(i) If the network has too many nodes, go to the EnrichmentMap tab in the *Control Panel* and use the *Node Cutoff Q-value* threshold slider. Adjusting to a numerical value closer to 0 will remove less significant nodes (Fig. 8 (**v**)).

(ii) If the network is too interconnected, go to the *EnrichmentMap* tab in the *Control Panel* and increase the *Edge Cutoff (Similarity)* threshold; this will remove connections between less related nodes (Fig. 8 (**vi**)).

(iii) Apply the network layout again after adjusting the cutoffs (see the *Layout* menu in Cytoscape). The default layout algorithm is the unweighted *Prefuse Force Directed* layout. We also recommend that the prefuse force-directed layouts be weighted using the gene set similarity coefficient. Alternative layout algorithms are available and we encourage experimentation with them.

▲ **CRITICAL STEP** There are many different layout algorithms available in Cytoscape that can be used for EnrichmentMap. We recommend using an edge-weighted layout, which considers the overlap score between pathways. Most layouts (except yFiles) offer the ability to organize just the selected nodes. Experiment with different layouts to see which works best with your data. If you do not like the resulting layout, press command-z on macOS or Ctrl-z on Windows or click on *Edit* → *Undo* to revert to the previous view.

(iv) To restore nodes or edges, adjust the threshold sliders to their original positions.

(v) *Separate two different phenotypes*. It can be helpful to separate two different phenotypes (i.e., place all the red nodes to one side and all the blue nodes to the other). To do this, go to the *Select* tab in the *Control Panel* (Fig. 8 (**i**)).

(vi) Click on the '+' symbol and select *Column filter*.

(vii) Click on *Choose column…* and select 'EM1_NES (Mesem_vs_Immuno)'.

(viii) Click on the box next to *between* and change the value to 0. Click *Apply* at the bottom of the panel.

(ix) All red nodes should now be selected. Click and hold on any selected node and drag selection to the left until it does not overlap any blue nodes.

(x) Select *Layouts* from the *Cytoscape* menu and apply *Prefuse Force Directed Layout* → *Selected Nodes Only* → *(none)*.

(xi) Go back to the *Control Panel Select* tab and adjust the slider to select all negative values. Click on *Apply* at the bottom of the *Select* tab.

(xii) All blue nodes should now be selected. Click and hold on any selected node and drag selection to the right until it does not overlap any red nodes.

(xiii) Select *Layouts* from the Cytoscape menu bar and apply *Prefuse Force Directed Layout* → *Selected Nodes Only* → *(none)*.

(C) **Defining major biological themes** ● **Timing 2.5 h**

▲ **CRITICAL** Enrichment maps typically include clusters of similar pathways representing major biological themes. Clusters can be automatically defined and summarized using the AutoAnnotate Cytoscape application. AutoAnnotate first clusters the network using the clusterMaker2 application and then summarizes each cluster on the basis of word frequency within the pathway names via the WordCloud app.

(i) From the Cytoscape menu bar, launch AutoAnnotate by selecting *Apps* → *AutoAnnotate* → *New Annotation Set…*

The *AutoAnnotate: Create Annotation Set* panel will appear.

(ii) In the *Quick Start* tab, click on *Create Annotations*.

? TROUBLESHOOTING

(iii) Each cluster in the network will have a circle annotation drawn around it and will be associated with a set of words (by default three) that correspond to the most frequent node labels in the cluster. These words are automatically selected and often must be manually renamed (Step 13E(iii)). Moving individual nodes within a cluster will automatically resize

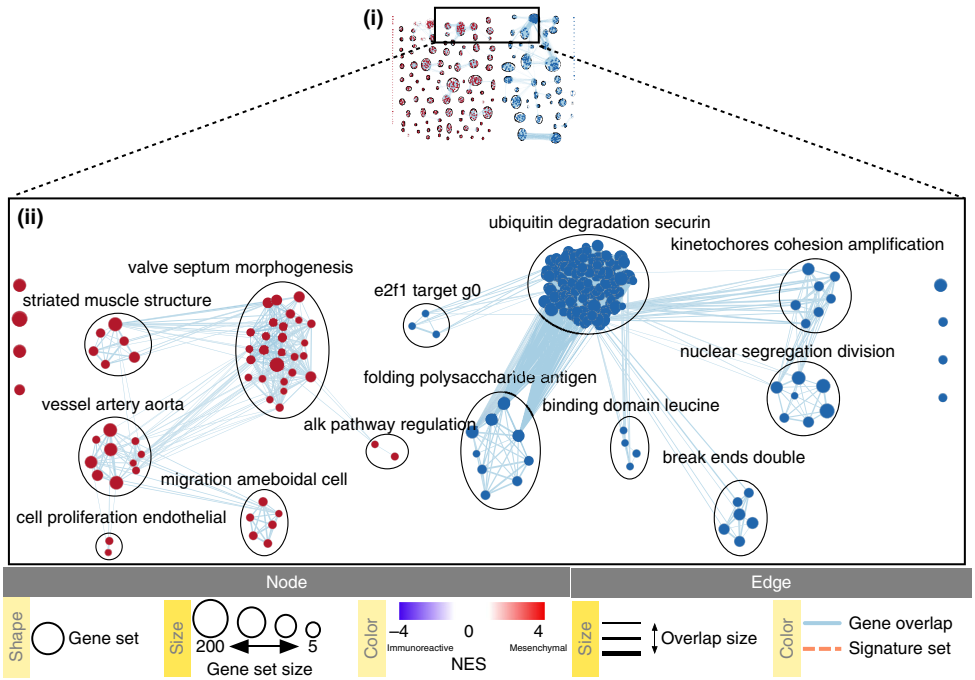


Fig. 10 | Resulting publication-ready enrichment map. (**i**) Overall thumbnail view of the publication-ready enrichment map created with parameters FDR *Q* value < 0.01, and combined coefficient >0.375 with combined constant = 0.5. (**ii**) Zoomed-in section of publication-ready enrichment map, in which red and blue nodes represent mesenchymal and immunoreactive phenotype pathways, respectively. Nodes were manually laid out to form a clearer picture. Clusters of nodes were labeled using the AutoAnnotate Cytoscape application. Individual node labels were removed for clarity using the publication-ready button in EnrichmentMap and exported as PNG and PDF files. A legend was manually added at the bottom of the figure.

the surrounding circle and moving an entire cluster will redraw the surrounding circle in the new cluster location.

? TROUBLESHOOTING

(iv) Manually arrange clusters to clean up the figure. Move nodes to reduce node and label overlap. Figure 10 shows the results of this process.

? TROUBLESHOOTING

(D) **Creation of a simplified network view** ● **Timing 15 min**

▲ **CRITICAL** This creates a single group node for each cluster with a summarized name and provides an overview of the enrichment result themes that is useful for enrichment maps containing many nodes (Fig. 11).

(i) In the *Control Panel*, select the *AutoAnnotate* tab.

(ii) Click on the *Menu* icon in the upper right corner.

(iii) Select *Collapse All*.

? TROUBLESHOOTING

(iv) *Scale the collapsed network for better viewing*. In the Cytoscape menu bar, select *View* → *Show Tool Panel*.

(v) Go to the *Node Layout Tools* panel located at the bottom of the *Control Panel*.

(vi) Locate the *Scale* slider bar and use it on unselected nodes (uncheck *Selected only*).

(vii) Move the slider left to tighten the node spacing. Close the *Node Layout Tools* panel when finished.

(E) **Manual arrangement of network nodes and updating of theme labels** ● **Timing 45 min**

▲ **CRITICAL** This section is required for the clearest network view and for a publication-quality figure. For instance, it is useful to bring together similar themes, such as signaling or metabolic pathways, even if they are not connected in the map. Use of space should be optimized so that large amounts of white space are not present. This is a time-consuming step, but the more effort spent, the higher the quality of the resulting figure will be (Fig. 10).

(i) If the focus of the figure is only on a subset of the network, it can be easier to work with just the subset. To create this, select the nodes of interest, then in the Cytoscape menu bar select

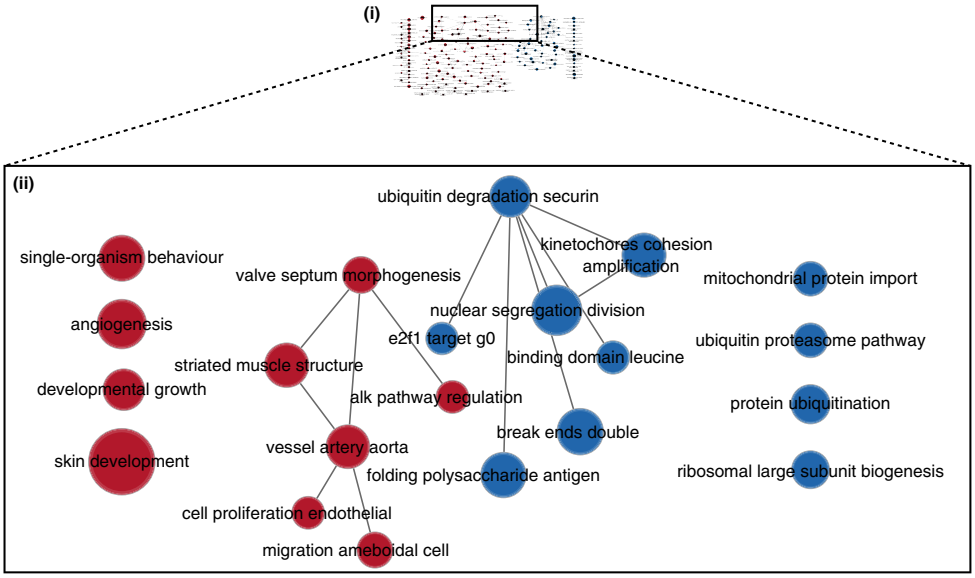


Fig. 11 | Collapsed enrichment map. The enrichment map was summarized by collapsing node clusters using the AutoAnnotate application. Each cluster of nodes in Fig. 10 is now represented as a single node. The network was scaled for better node distribution and manually adjusted to reduce node and label overlap. (**i**) Overall thumbnail view of the entire collapsed enrichment map. (**ii**) Zoomed-in section of the publication-ready collapsed enrichment map that corresponds to the zoomed-in network in Fig. 10 (**ii**).

File → *New* → *Network* → *From selected nodes* → *all edges*, or alternatively use the corresponding icon in the Cytoscape tools menu bar (*New network from selection* → *all edges*).

(ii) When the purpose of the figure is to show a large network and highlight only the main themes, click on *Publication ready*, located at the bottom of the *EnrichmentMap* in the *Control Panel* to remove node labels. To revert to the original network, click on the *Publication ready* button again.

? TROUBLESHOOTING

(iii) Rename the theme names generated by AutoAnnotate to better explain groups of pathways. Automatic WordCloud-generated theme names are useful for quickly exploring an enrichment map, but frequently need to be renamed for publication-quality figures. Naming should carefully consider all pathways and genes within each theme. Themes can be renamed in AutoAnnotate by right-clicking the name in the *AutoAnnotate* panel in the 'Cluster' column, and selecting *Rename…* Alternatively, labels can be changed in external drawing software (follow Steps 14–16 to export a file to use in external drawing software). Note that re-clustering the network will reset the theme names.

(F) **Creation of a subnetwork that highlights a specific theme subset** ● **Timing 10 min**

▲ **CRITICAL** Enrichment maps of rich omics datasets are often large and complicated, and it is often useful to emphasize specific themes or relevant pathways in a final figure. For example, we will select the top mesenchymal and immunoreactive pathways and create a subnetwork for detailed visualization.

(i) Click on the *Select* tab in the *Control Panel* (Fig. 8 (**i**)).

(ii) Click on the '+' symbol and select *Column filter*.

(iii) Click on *Choose column…* and select *EM1_NES (Mesem_vs_Immuno)*.

(iv) Click on the box next to *between* and replace the negative value with '2.5'. Do not change the positive value next to *inclusive*. Click *Enter*.

(v) Click on the '+' symbol and select *Column filter*.

(vi) Click on *Choose column…* and select *EM1_NES (Mesem_vs_Immuno)*.

(vii) Click on the box next to *inclusive* and change the value to '-2.5'. Do not change the negative value next to *between*. Click *Enter*.

(viii) Above the two column filters you just added, change the dropdown option from *Match all (AND)* to *Match any (OR)*.

(ix) Click on *Apply*. Under the *Apply* button, it should display the number of nodes and edges selected. In this example, 32 nodes should be selected.

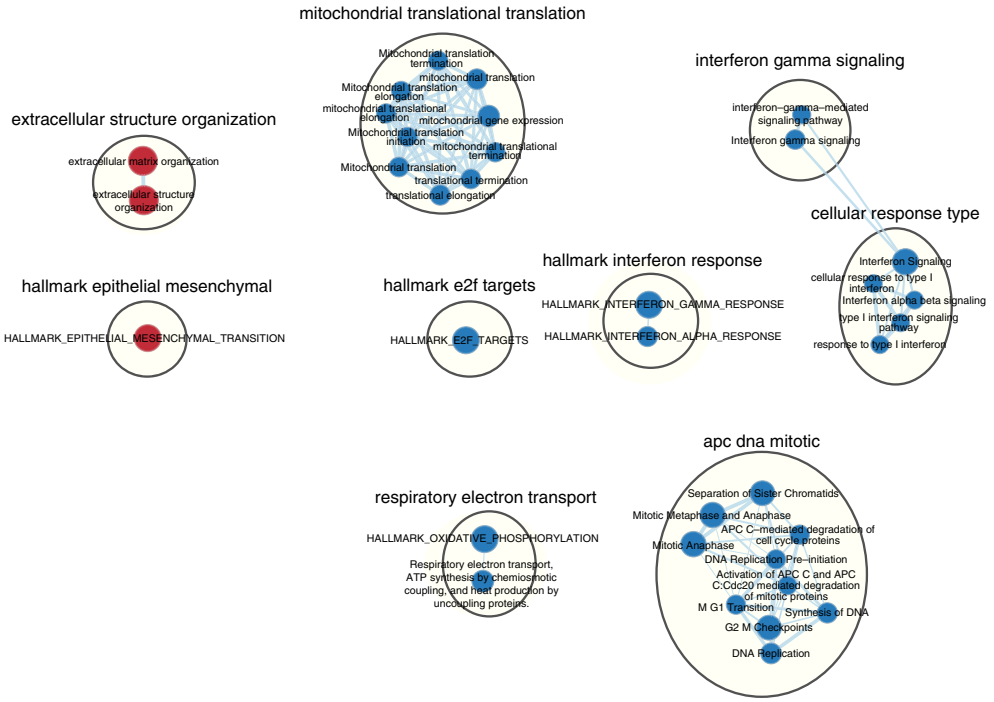


Fig. 12 | Subnetwork example. Subnetwork of the main enrichment map (Fig. 10) was manually created by selecting pathways with the top NES values and creating a new network from the selection. Red and blue nodes are mesenchymal and immunoreactive phenotypes, respectively. Clusters of nodes were automatically labeled using the AutoAnnotate application. Annotations in the subnetwork may differ slightly from those in the main network, as word counts were normalized on a network basis.

(x) From the Cytoscape menu, choose *Select File → New → Network → From selected nodes, all edges.*
? TROUBLESHOOTING

(xi) A new, smaller network should appear. Manually move nodes around to optimize the layout.

(xii) Annotate the network as described in Step 13C (Fig. 12).

Exporting figures, creating legends and saving work ● Timing 15 min

14 Export the image. In the Cytoscape menu bar, select *File → Export as Image…* Set the *Export File Format* to PDF (*.pdf).
▲CRITICAL STEP Vector-based PDF and SVG formats are recommended for publication-quality figures because they can be zoomed without losing quality. Either file type can be edited using software packages such as Adobe Illustrator or Inkscape. The PNG file format is recommended for high-quality online images, whereas the JPG format is not recommended because it may lead to visual artifacts due to lossy compression.

15 Click on *Browse…* to specify file name and location.

16 Click on *Save* to close the browser window and then on *OK*.
? TROUBLESHOOTING

17 Identify network creation parameters. In the previous step, we exported the network as an image, but the parameters used to create the map are needed to interpret and reproduce the image. In the *Control Panel*, locate the *EnrichmentMap* input panel and click on the cog (*Settings*) icon in the top right-hand corner.

18 Click on *Show Creation Parameters*. In the displayed panel, you will find the FDR *Q* value, similarity metric and threshold parameters to be added to the text legend of figure. In this example: 'Enrichment map was created with parameters $q < 0.01$, and Jaccard Overlap combined coefficient > 0.375 with combined constant $= 0.5$'.
▲CRITICAL STEP The *EnrichmentMap Creation Parameters* panel shows only the parameters that were used at network creation. If you modified the network using filters or the EnrichmentMap slider bars, you will have to update the thresholds accordingly.

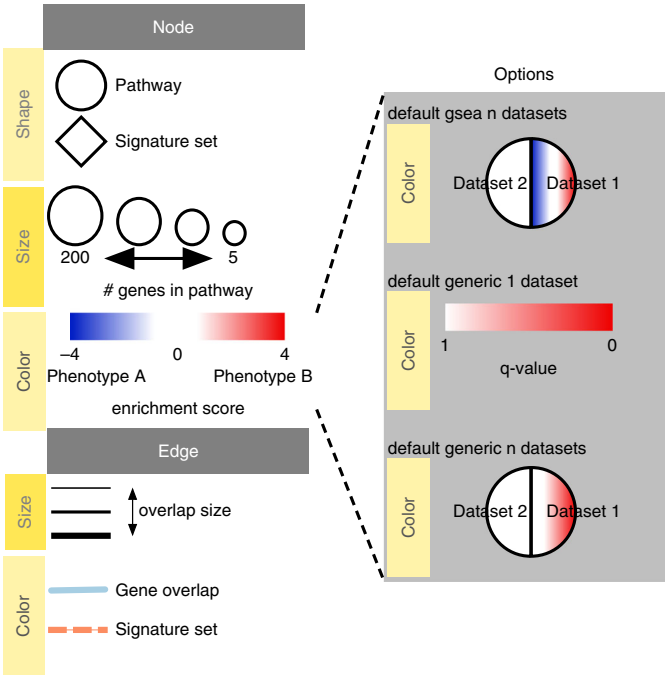


Fig. 13 | Generic enrichment map legend. Enrichment map attributes can be copied for use in a custom figure legend. Only components relevant to the analysis should be copied. Post-analysis 'Signature set' nodes are included in the generic legend (not covered in this protocol). Post-analysis nodes highlight pathways in the enrichment map that contain specific genes of interest such as targets of drugs or microRNAs.

19 Show and create a legend. In the *Control Panel*, locate the *EnrichmentMap* input panel and click on the cog icon in the top right corner. Click on *Show Legend*. The example shows a simplified legend; however, many different node and edge attributes, such as size, shape and color, can be used in the enrichment map to represent different aspects of the data. It is important to describe them in the text associated with the figure or in the figure itself as a legend. Figure 13 shows the basic legend components (available as SVG and PDF images at http://baderlab.org/Software/EnrichmentMap#Legends) that can be used for an enrichment map figure. You can manually select the components relevant to your analysis. See the bottom of Fig. 10 for the legend components used for current analysis.

20 Save all work as a Cytoscape session. In the Cytoscape menu, select *File* → *Save As…* Navigate to the directory in which you wish to save the session and specify the desired file name.

Troubleshooting

Troubleshooting advice can be found in Table 1.

Table 1 | Troubleshooting table

Step	Problem	Possible reason	Solution
2	User does not know whether Java is installed on the computer or which version is installed		The Java website (https://www.java.com) has a help page called 'How to find Java version in Windows or Mac – Manual method' (https://www.java.com/en/download/help/version_manual.xml) to find which version of Java is already installed
6B(i)	Unable to launch GSEA	Unable to associate .jar file with Java application	When launching GSEA on macOS for the first time, you may get the error 'gsea.jnlp cannot be opened because it is from an unidentified developer'. Click on 'Ok'. Instead of double-clicking on the gsea.jnlp icon/file, right-click and select *Open*. The same error 'gsea.jnlp can't be opened because it is from an unidentified developer' will appear, but this time it will give you

Table continued

Table 1 (continued)

Step	Problem	Possible reason	Solution
			the option to *Open* or *Cancel*. Click on *Open*. After this initial opening, subsequent double-clicks on *gsea.jnlp* will launch GSEA without errors or warnings. If GSEA still fails to launch, GSEA can alternatively be launched from the command line. Go to the GSEA download site (http://www.broadinstitute.org/gsea/downloads.jsp) and download the javaGSEA JAR file (the second option on the download site). Open a command-line terminal. On macOS, the terminal can be found in *Applications* → *Utilities* → *Terminal*. On Windows, type `cmd` in the Windows program files search bar. Then navigate to the directory where the file javaGSEA.jar was downloaded, using the command `cd`. For example, on macOS run `cd ~/Downloads` if you downloaded the GSEA.jar file to your *Downloads* folder. Run the command `java -Xmx4G -jar gsea-3.0.jar`, where `-Xmx` specifies how much memory is given to GSEA
6B(ii)	User needs more information about GSEA file formats (GMT, RNK, CLS, GCT)	To be able to format his/her own files for GSEA	GSEA has a useful help documentation on file formats available at https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats
6B(v)	GSEA seems non-responsive	A large GMT file is being loaded	It may take 5–10 s for GSEA to load input files. The files are loaded successfully once a message appears on the screen, e.g., 'Files loaded successfully: 2/2. There were no errors'
6B(xiii)	GSEA looks non-responsive, but it is actually computing enrichments	No progress bar	GSEA has no progress bar to indicate estimated time to completion. A run can take a few minutes or hours, depending on the size of the data and the computer speed. Click on the '+' in the bottom left corner of the screen to see messages such as 'shuffleGeneSet for GeneSet 4661/4715 nperm: 1000' (circled in red at the bottom of Fig. 3). This message indicates that GSEA is shuffling 4,715 gene sets 1,000 times each, 4,661 of which are complete. Once the permutations are complete, GSEA generates the report
6B(xiv)	Error message 'Java Heap space'	GSEA was launched with insufficient memory	The error message 'Java Heap space' indicates that the software has run out of memory. Another version of GSEA is needed if you are running the GSEA desktop application. There are multiple options available for download from the GSEA website. You can download a webstart application that launches GSEA with 1, 2, 4 or 8 GB of RAM. Upgrade to a webstart that launches with more memory. If you are already using the webstart that launches with 8 GB, then you require a GSEA JAVA .jar file, which can be executed from the command line with increased memory (see Troubleshooting for Step 6B(i) for details)
6B(xv)	User needs to access previous results but cannot find them	GSEA application was closed since running the analysis	If the GSEA software is closed, you can access previous results by opening the working folder and opening the 'index.html' file. Alternatively, you can re-launch GSEA and click on *Analysis history*, then *History* and then navigate to date of your analysis. Although all analyses, regardless of where the results files were saved, are listed under history, they are organized by the date the analyses were run. If you cannot remember when you ran a specific analysis, then you may have to manually search through a few directories to find the desired analysis
6B(xvi)	Few or no results returned by GSEA	Possible identifier mapping issue	Check the number of gene sets that were analyzed. If the number is low (e.g., low hundreds), it could indicate gene ID mapping problems
9A(ii)	Autoload of g:Profiler results creates many datasets with incorrect file specifications	There are too many text files within the directory specified	To simplify loading g:Profiler results into EnrichmentMap and populating the correct fields in the EnrichmentMap interface, place the g:Profiler results file and gene set file (i.e., Supplementary_Table4_gprofiler_results.txt and Supplementary_Table5_hsapiens.pathways.NAME.gmt) into a directory together by themselves
9A(xi)	User cannot create a g:Profiler map with more than one phenotype		Although an individual g:Profiler analysis has only one phenotype, it is possible to modify a single results file to contain two analyses. This is relevant when the phenotypes are mutually exclusive. For the analysis you want to associate with the additional phenotype (which would correspond to

Table continued

Table 1 (continued)

Step	Problem	Possible reason	Solution
			downregulated genes in GSEA PreRanked, thus called 'negative') open the g:Profiler results file (preferably in a spreadsheet, so you can easily modify a single column). The fifth column specifies the phenotype. Update the column to have the value of '−1' for each result in the file. Open the second analysis file. Copy all the results from the second file and paste them into the updated negative g:Profiler file. Save the file and use it as the g:Profiler enrichment results file in the EnrichmentMap interface instead of the original results files. Pathways corresponding to two phenotypes will be colored red and blue in the resulting enrichment map. One limitation with this approach is that a pathway cannot be included in both the positive and the negative sets
9B(vi), (vii), (ix)	A random number is appended to the GSEA directory name		Each GSEA analysis generates a random number that is appended to the names of the files and directories. The number will be different for each new analysis
9B(viii)	EnrichmentMap uses a GMT file that was not the original file input to GSEA	The original GMT file was moved or no longer exists in the location in which GSEA saved it	If EnrichmentMap cannot find the original GMT file used in the GSEA analysis, it will use a filtered GMT file found in the GSEA 'edb' results directory. EnrichmentMap will not be able to find your original GMT file if you have moved it since running GSEA analysis. Although it is a GMT file, it has been filtered to contain only genes found in the expression file. If you use this filtered file, you will obtain different pathway connectivity depending on the expression data being used. You should always use the original GMT file used for the GSEA analysis and not the filtered one in the results directory
9B(xiii)	User cannot provide a sufficiently precise *Q* value	Scientific notation is not enabled	To set the threshold to a small number, select *Scientific Notation* and set a *Q* value cutoff such as 1E−04
10	Few or no pathways are present in EnrichmentMap	Input dataset may not contain enough signal to find enriched pathways	A pathway enrichment analysis resulting in few or no enriched pathways may be caused by suboptimal statistical processing used to define the original gene list. Enriched pathways are unlikely to be found if the gene list ranks are too noisy and the most important genes are not at the top of the list, no genes are highly significant, or a large fraction of genes are highly significant. If the gene list has been correctly defined, analyzing further databases of pathways and gene sets or setting more liberal filters may improve results
11	User cannot find any pathways with search gene	Check that the gene identifier type used for the search matches the identifier type used in the analysis	Multiple genes separated by spaces can be entered into the search bar. Any pathway that contains the gene will be selected and highlighted in the network. Adding keywords with 'AND' into the search bar will show only pathways that contain all genes in the search query (e.g., 'geneA AND geneB'). If the original analysis did not use gene symbols, then you will not be able to search by gene symbols. Instead, use the identifier type that the analysis was based on, for example Entrez Gene ID or Ensembl gene ID
12	There are very few entries in the node table, although the network contains many nodes	Some nodes in the network are already selected	If there are very few records in the node table, make sure that no nodes are selected in the network. Or click on the gear icon and change the setting from *Auto* to *Show all*
13A(i)	Leading-edge genes are not highlighted when clicking on a pathway node	Analysis was not done with GSEA, or GSEA rank file or enrichment results were not supplied when the enrichment map was built	The leading edge can be displayed only if the rank file is provided when the network is built. The rank file supplied needs to be identical to the one used for the GSEA analysis for the leading-edge calculation to function
13A(ii)	User does not know which sort option to choose		In the case of multiple conditions or conditions with variable expression profiles (e.g., cancer patient samples), hierarchical clustering tends to generate a more informative visualization
13A(vii)	Heat map column names are not colored by dataset	No CLS file was loaded or there is a mismatch between the CLS file and the phenotype definition	If the heat map columns are not colored for a GSEA analysis, make sure the phenotype names specified in the EnrichmentMap input panel match the class names specified in the class file (MesenchymalvsImmunoreactive_RNA-Seq_classes.cls). Also see Troubleshooting for Step 9B(xiii)

Table continued

Table 1 (continued)

Step	Problem	Possible reason	Solution
13A(ix)	The option to save only leading-edge genes is not available	Selection includes more than one node or dataset contains no leading-edge information (i.e., was not built from GSEA results)	The leading edge is available only for GSEA analyses. The option will appear only if the enrichment map was built with GSEA results and a rank file was specified
13C(ii)	AutoAnnotate has many tunable parameters		The default parameters are likely to work well with EnrichmentMap; however, there are many parameters within the AutoAnnotate application that can fine-tune the results. See the AutoAnnotate user manual at https://autoannotate.readthedocs.io/en/latest/
13C(iii)	Labels contain uninformative words	Node names contain uninformative words that are not excluded by default or are not considered during network normalization	If particular non-informative words keep appearing in the labels generated by AutoAnnotate, try adjusting the WordCloud normalization factor. The significance of each word is calculated on the basis of the number of occurrences in the given cluster of pathways. This causes frequent words such as 'pathway' or 'regulation' to be prominent. By increasing the normalization factor, we reduce the priority of such recurrent words in cluster labels. If that doesn't help, you can add the non-informative words to the WordCloud word exclusion list
	Labels contain '-'		If a specific character other than a space is used to separate words (e.g., '-' or '|'), it should be added as a delimiter in the WordCloud application. Launch the WordCloud application (*Apps* → *WordCloud*). In the *WordCloud* input panel, expand *Advanced options*. Click on *Delimiters…* Add your delimiters. Click on *OK*. In the *AutoAnnotate* input panel, click on the menu button (icon with three horizontal lines). Select *Recalculate Labels…* for this change to take effect
13C(iv)	Labels are bigger for bigger clusters, but user wants all the labels to be the same size	Setting to scale labels to the size of the cluster is enabled	The number of nodes in a cluster determines label size by default. Thus, the cluster size may relate to pathway popularity instead of importance in the experiment. Annotation labels can all be set to the same size by unchecking the option *Scale font by cluster size* in the AutoAnnotate results panel
13D(iii)	Pop-up after selecting *Collapse all* shows up every time user collapses the clusters	User has not specified *Don't ask me again* option on pop-up	Once you click on *Collapse All*, a pop-up window will show the message 'Before collapsing clusters please go to the menu Edit → Preferences → Group preferences and select 'Enable attribute aggregation''. There is no need to adjust this parameter repeatedly. Click on *Don't ask me again* and *OK* if you have set this parameter previously
	Collapsing the network takes a long time	The larger the network or the more clusters in a network, the longer collapsing will take	For large networks, collapsing and expanding may take time. For a quick view of the collapsed network, you can create a summary network by selecting the *Create summary Network…* option. There are two options for the summary network: *clusters only*, which creates a summary network with just the circled clusters, or *clusters and unclustered nodes*, which creates a summary network that also includes the singleton nodes that are not part of any cluster
	Collapsed network contains gray nodes instead of colored nodes as they were in the pre-collapsed network	Attribute aggregation is not enabled	If the nodes in the resulting collapsed network are gray, then you forgot to enable attribute aggregation. Expand the clusters and, before collapsing clusters again, go to *Edit* → *Preferences* → *Group preferences* and select *Enable attribute aggregation*
13E(ii)	In the EnrichmentMap input panel, the bottom options *Publication Ready* and *Set Signature Edge Width* are not visible	The *Node Layout Tools* is open	Close the *Node Layout Tools* window using the × symbol located at the top right corner
13F(x)	The created subnetwork is empty	Nodes were not selected before the creation of the subnetwork	Make sure that the nodes that will be part of the subnetwork are selected before creation of the subnetwork
16	Exported image contains only a small subset of the network	Only what is visible in the view is exported	In *image export*, only the visible part of the map will be exported. Make sure that the entire network is visible on your screen before exporting

Timing

The time required for this protocol mainly depends on the level of manual curation and the time spent on visual organization of the enrichment map. Smaller networks are generally easier to organize. Before laying out the final network, it is worth revisiting the enrichment analysis to ensure it works as expected, exploring the network fully and selecting the parts to emphasize in a final publication-quality figure.

The above example analysis was performed on a Windows 7 machine with 8 GB of RAM and Java 8. Increased RAM and processor speed will decrease your analysis time for some of the steps, in particular the GSEA analysis (Step 6B).

Steps 1–5, software installation: 5 min
Step 6A, pathway enrichment analysis of a gene list using g:Profiler: 3 min
Step 6B, pathway enrichment analysis of a ranked gene list using GSEA: ~20 min
Steps 7–9, visualization of enrichment results with EnrichmentMap: ~5 min
Steps 10–13, navigation and interpretation of the enrichment map: ~4 h
Step 13A, exploring the *Table Panel* heat map: 45 min
Step 13B, organization and clarification of the network: 30 min
Step 13C, defining major biological themes: 2.5 h
Step 13D, creation of a simplified network view: 15 min
Step 13E, manual arrangement of network nodes and updating of theme labels: 45 min
Step 13F, creation of a subnetwork that highlights a specific theme subset: 10 min
Steps 14–20, exporting figures, creating legends and saving work: 15 min

Anticipated results

Pathway enrichment analysis of a gene list using g:Profiler

Two main example files are produced in the analysis:

- *Supplementary_Table4_gprofiler_results.txt*. This file, downloaded from the g:Profiler website in Step 6A(xi), contains the set of enriched pathways from the g:Profiler analysis as a table with six columns. The first column, 'GO.ID', contains the pathway identifier. The second column, 'description', contains the pathway description. The third column, 'p.Val', contains the enrichment *P* value of the pathway. The fourth column, 'FDR', contains the FDR-adjusted enrichment *P* value of the pathway. g:Profiler provides only FDR-adjusted *P* values and therefore values in the third and fourth columns are equal. The fifth column, 'Phenotype', contains a '1', indicating the positive phenotype the analysis belongs to. Results from g:Profiler will always specify the phenotype to be 1, but users can manually change this column in order to merge two distinct result sets together. The sixth column, 'Genes', contains a comma-separated list of genes belonging to the pathway.
- *Supplementary_Table5_hsapiens.pathways.NAME.gmt*. This file, downloaded from the g:Profiler website in Step 6A(xii), contains the set of pathways used for the g:Profiler analysis. Each row of the file represents a pathway, where the first column is the pathway name or identifier, the second column is the pathway description and any subsequent column is a gene that is part of the pathway.

Pathway enrichment analysis of a ranked gene list using GSEA

Two main example files are produced in the analysis:

- *Supplementary_Table8_gsea_report_for_na_pos.xls*. This is the main GSEA result table listing the pathways associated with the na_pos phenotype (explained in Step 6B(xv)).
- *Supplementary_Table9_gsea_report_for_na_neg.xls*. This is the main GSEA result table listing the pathways associated with the na_neg phenotype (explained in Step 6B(xv)).

Both files contain tables with 11 columns describing the degree of enrichment of the pathways associated with each phenotype. For a detailed description of all the values found in the table, see https://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html (under the heading 'Detailed Enrichment Results').

GSEA will create an entire directory of result files. Unless specified otherwise, the GSEA results folder will be placed in your home directory under gsea_home/(current date). The folder will be named with the name specified in Step 6B(xi) (e.g., 'Mesen_vs_Immuno' per this protocol), with a 13-digit random number appended to the end. To see a visual summary of the analysis, use a web browser to open the 'index.html' file found in the results directory. Step 6B(xiv, xv, and xvi), as well as Figs. 4 and 5, describes the interpretation of these results.

Visualization of pathway enrichment results as an enrichment map

Example enrichment maps resulting from analyses with g:Profiler and GSEA can be seen in Fig. 7a,b (for publication-ready GSEA results, see Fig. 10), respectively. Pathway information is inherently redundant, and enrichment analysis often highlights several versions of the same pathway as a result. Collapsing redundant pathways into a single biological theme simplifies interpretation (Fig. 11). An enrichment map is a network representing overlap among enriched pathways that are represented as circles (nodes) and are connected with lines (edges) sized based on the number of genes shared by the connected pathways. Multiple sets of results can be simultaneously visualized in a single enrichment map, in which case different colors are used to color pathways resulting from different analyses. If the gene expression data are loaded, clicking on a pathway node will display a gene expression heat map of all genes in the pathway.

Navigation and interpretation of the enrichment map

An enrichment map helps identify interesting pathways and themes characteristic of omics data. The analysis can be validated by identifying expected themes, such as growth-related pathways in a cancer genomics analysis. Expected themes are identified on the basis of expert knowledge of the studied biological system. Additional pathways are evaluated as potential discoveries. Genes in detected pathways should be interpreted using the original omics data and pathway diagrams[45,47–49,58], gene interaction networks[50,51] or regulatory networks of transcription factors[52,53]. Finally, enrichment maps can be published to support scientific conclusions or used for generating hypotheses for follow-up experiments. Pathway enrichment analysis results can change on the basis of the parameters used (e.g., minimum and maximum pathway size or selected pathway databases), thus the robustness of conclusions should be tested by varying these parameters.

Reporting Summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The protocol uses publicly available software packages (GSEA v.3.0 or higher, g:Profiler, Enrichment Map v.3.0 or higher, Cytoscape v.3.6.0 or higher) and custom R scripts that apply publicly available R packages (edgeR, Roast, Limma, Camera). Custom scripts are available in the Supplementary Protocols and at our GitHub web sites (https://github.com/BaderLab/Cytoscape_workflows/tree/master/EnrichmentMapPipeline and https://baderlab.github.io/Cytoscape_workflows/EnrichmentMapPipeline/index.html).

References

1. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
2. Stephens, Z. D. et al. Big data: astronomical or genomical? *PLoS Biol.* **13**, e1002195 (2015).
3. Mack, S. C. et al. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**, 445–450 (2014).
4. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
5. Pinto, D. et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
6. Kandoth, C. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
7. Verhaak, R. G. et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* **123**, 517–525 (2013).
8. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
9. Cline, M. S. et al. Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
10. Creixell, P. et al. Pathway and network analysis of cancer genomes. *Nat Methods* **12**, 615–621 (2015).
11. Wadi, L., Meyer, M., Weiser, J., Stein, L. D. & Reimand, J. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**, 705–706 (2016).
12. Reyna, M. A. et al. Pathway and network analysis of more than 2,500 whole cancer genomes. Preprint at https://www.biorxiv.org/content/early/2018/08/07/385294 (2018).
13. Reimand, J. et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–89 (2016).

14. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
15. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
16. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
17. Anders, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8**, 1765–1786 (2013).
18. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
19. Silva, T. S. & Richard, N. Visualization and differential analysis of protein expression data using R. *Methods Mol. Biol.* **1362**, 105–118 (2016).
20. Schubert, O. T., Rost, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat. Protoc.* **12**, 1289–1294 (2017).
21. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
22. Gonzalez-Perez, A. et al. Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10**, 723–729 (2013).
23. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10**, 1556–1566 (2015).
24. Assenov, Y. et al. Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11**, 1138–1140 (2014).
25. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
26. Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).
27. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
28. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
29. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
30. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
31. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol* **3**, Article3 (2004).
32. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
33. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
34. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* **9**, 811–818 (1990).
35. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
37. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
38. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2013).
39. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).
40. Bindea, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
41. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).
42. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
43. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).
44. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
45. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
46. Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
47. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).

48. Kelder, T. et al. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* **40**, D1301–D1307 (2012).
49. Kutmon, M. et al. PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput. Biol.* **11**, e1004085 (2015).
50. Szklarczyk, D. et al. STRINGv10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
51. Warde-Farley, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).
52. Lechman, E. R. et al. Attenuation of miR-126 activity expands HSC in vivo without exhaustion. *Cell Stem Cell* **11**, 799–811 (2012).
53. Jhas, B. et al. Metabolic adaptation to chronic inhibition of mitochondrial protein synthesis in acute myeloid leukemia cells. *PLoS ONE* **8**, e58367 (2013).
54. Ballouz, S., Pavlidis, P. & Gillis, J. Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic Acids Res.* **45**, e20 (2017).
55. Krzywinski, M. & Altman, N. Power and sample size. *Nat. Methods* **10**, 1139–1140 (2013).
56. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30**, 301–304 (2014).
57. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
58. Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).
59. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
60. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* **11**, R3 (2010).
61. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* **9**, 509–515 (2008).
62. Skunca, N., Altenhoff, A. & Dessimoz, C. Quality of computationally inferred gene ontology annotations. *PLoS Comput. Biol.* **8**, e1002533 (2012).
63. Wojtowicz, E. E. et al. Ectopic miR-125a expression induces long-term repopulating stem cell capacity in mouse and human hematopoietic progenitors. *Cell Stem Cell* **19**, 383–396 (2016).
64. Tong, J. et al. Integrated analysis of proteome, phosphotyrosine-proteome, tyrosine-kinome, and tyrosine-phosphatome in acute myeloid leukemia. *Proteomics* **17**, 1600361 (2017).
65. Kamdar, S. N. et al. Dynamic interplay between locus-specific DNA methylation and hydroxymethylation regulates distinct biological pathways in prostate carcinogenesis. *Clin. Epigenetics* **8**, 32 (2016).
66. Liu, Y. et al. Metabolomic profiling in liver of adiponectin-knockout mice uncovers lysophospholipid metabolism as an important target of adiponectin action. *Biochem. J.* **469**, 71–82 (2015).
67. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
68. Raychaudhuri, S. et al. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.* **6**, e1001097 (2010).
69. Lee, P. H., O'Dushlaine, C., Thomas, B. & Purcell, S. M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797–1799 (2012).
70. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* **8**, e1002375 (2012).
71. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
72. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
73. Gu, Z. & Wang, J. CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics* **29**, 658–660 (2013).
74. Fang, Z., Tian, W. & Ji, H. A network-based gene-weighting approach for pathway analysis. *Cell Res.* **22**, 565–580 (2012).
75. Farfan, F., Ma, J., Sartor, M. A., Michailidis, G. & Jagadish, H. V. THINK Back: KNowledge-based Interpretation of High Throughput data. *BMC Bioinformatics* **13**(Suppl. 2), S4 (2012).
76. Tarca, A. L. et al. A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82 (2009).
77. Draghici, S. et al. A systems biology approach for pathway level analysis. *Genome Res.* **17**, 1537–1545 (2007).
78. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* **28**, i451–i457 (2012).
79. Schaefer, C. F. et al. PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
80. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
81. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
82. Bader, G. D., Cary, M. P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, D504–D506 (2006).
83. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

84. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
85. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).
86. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11**, 773–785 (2010).

Acknowledgements

The authors are grateful to J. Mesirov for comments on the manuscript. This project was supported by an Investigator Award to J.R. from the Ontario Institute for Cancer Research through funding from the Government of Ontario and by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to J.R. (RGPIN-2016-06485). This work was supported by US National Institutes of Health grants P41 GM103504, R01 GM070743, U41 HG006623 and R01 CA121941 to G.D.B.

Author contributions

J.R., R.I., V.V., A.R., D.M. and G.D.B. wrote the manuscript. R.I. created the step-by-step protocols, figures, R scripts and R notebooks, except for g:Profiler (J.R.). M.K. and C.T.-L. developed EnrichmentMap 3.0 and AutoAnnotate Cytoscape applications. L.W., M.M., J.W., C.X. and V.V. tested the protocol. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-018-0103-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.D.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published online: 21 January 2019

Related links

Key references using this protocol

Pinto, D. et al. *Nature* **466**, 368–372 (2010): https://doi.org/10.1038/nature09146
Pajtler, K. W. et al. *Cancer Cell* **27**, P728–P743 (2015): https://doi.org/10.1016/j.ccell.2015.04.002
Cavalli, F. M. G. et al. *Cancer Cell* **31**, P737–P754 (2017): https://doi.org/10.1016/j.ccell.2017.05.005

nature research

Corresponding author(s): Gary Bader

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	g:Profiler (https://biit.cs.ut.ee/gprofiler/), GSEA 3.0 or higher, Cytoscape 3.6.1 or higher. Cytoscape apps: EnrichmentMap, version 3.1 or higher, clusterMaker2, version 0.9.5 or higher, WordCloud, version 3.1.0 or higher, AutoAnnotate, version 1.2.0 or higher.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data analyzed in this study is freely available as supplementary material.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Public data was used as is for protocol demonstration purposes.
Data exclusions	None
Replication	Each replication of the protocol generates essentially identical results.
Randomization	Not relevant as we use public data already allocated into groups.
Blinding	Not relevant.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging