(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 November 2017 (16.11.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/196872 A1

(51) International Patent Classification:
G06F 19/18 (2011.01)

(21) International Application Number:
PCT/US2017/031799

(22) International Filing Date:
09 May 2017 (09.05.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/333,407 09 May 2016 (09.05.2016) US

(71) Applicant: **WAYNE STATE UNIVERSITY** [US/US]; 656 W. Kirby, 4249 FAB, Detroit, Michigan 48202 (US).

(72) Inventors: **DRAGHICI, Sorin**; c/o Wayne State University, Dept. of Computer Science, 431 State Hall, 5143 Cass Avenue, Detroit, Michigan 48202 (US). **NGUYEN, Tin Chi**; c/o Wayne State University, Dept. of Computer Science, 431 State Hall, 5143 Cass Avenue, Detroit, Michigan 48202 (US).

(74) Agent: **KOTSIS, Damian H.** et al.; Harness, Dickey & Pierce, P.L.C., P.O. Box 828, Bloomfield Hills, Michigan 48303 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) Title: ORTHOGONAL APPROACH TO INTEGRATE INDEPENDENT OMIC DATA

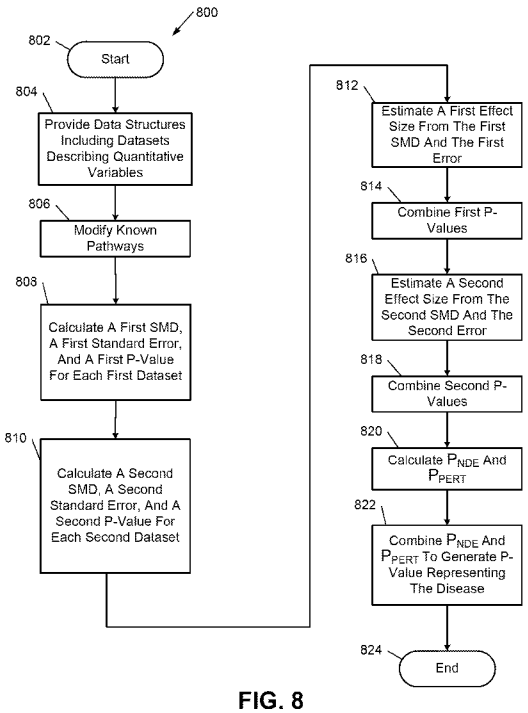

FIG. 8

(57) Abstract: Methods and devices for integrating a plurality of data types are provided. The methods include obtaining, via a processor, a plurality of datasets of a given type including measurements of one or more quantitative variables related to a phenotype comparison, and a plurality of datasets of a different type including measurements of one or more quantitative variables related to the same phenotype comparison; calculating, via the processor, effect sizes of the variables of the first type, effect sizes of the variables of the second type, and global p-values for the first and second data types; and combining, via the processor, the effect sizes and/or the global p-values to identify the variables of either type that are relevant in the given phenotype comparison.

WO 2017/196872 A1

ORTHOGONAL APPROACH TO INTEGRATE INDEPENDENT OMIC DATA

GOVERNMENT RIGHTS

[0001] This invention was made with U.S. Government support under NIH R01 DK089167, R42 GM087013 and NSF DBI-0965741. The Government has certain rights in the invention.

CROSS-REFERENCE TO RELATED APPLICATIONS

[0002] This application claims the benefit of U.S. Provisional Application No. 62/333,407, filed on May 9, 2016. The entire disclosure of the above application is incorporated herein by reference.

FIELD

[0003] The present disclosure relates to two-dimensional data integration that combines data obtained from many independent experiments.

BACKGROUND

[0004] This section provides background information related to the present disclosure which is not necessarily prior art.

[0005] High-throughput technologies for gene expression profiling, such as DNA microarray or RNA-Seq, have transformed biomedical research by allowing for comprehensive monitoring of biological processes. A typical comparative analysis of expression data, *e.g.*, patients ("unhealthy condition," *i.e.*, disease) versus control samples ("healthy condition"), generally yields a set of genes that are differentially expressed (DE) between the conditions. These sets of DE genes contain the genes that are likely to be involved in the biological processes responsible for the disease. However, such sets of genes are often insufficient to reveal the underlying biological mechanisms. In addition, due to inherent bias and batch effects present in individual studies, independent experiments studying the same disease often yield completely different lists of DE genes, making interpretation extremely difficult.

[0006] In order to translate these lists of DE genes into a better understanding of biological phenomena, a variety of knowledge bases have been developed that map genes to functional modules. Depending on the amount of information that one wishes to include, these modules can be described as simple gene sets based on a function,

process or component (*e.g.*, the Molecular Signatures Database MSigDB), organized in a hierarchical structure that contains information about the relationship between the various modules or organized into pathways that describe in detail all known interactions between various genes that are involved in a certain phenomenon. Exemplary pathway databases include: the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and Biocarta.

[0007] Analysis techniques have been developed to help interpret such sets of DE genes. The earliest approaches use Over-Representation Analysis (ORA) to identify gene sets that have more DE genes than expected by chance. The drawbacks of this type of approach include that: (i) it only considers the number of DE genes and completely ignores expression changes; (ii) it assumes that genes are independent, which they are not; and (iii) it ignores the interactions between various modules. Functional Class Scoring (FCS) approaches, such as Gene Set Enrichment Analysis (GSEA) and Gene Set Analysis (GSA), have been developed to address some of the issues raised by ORA approaches. The main improvement of FCS is the observation that small but coordinated changes in expression of functionally related genes can have significant impacts on pathways. Both FCS and ORA approaches can be used with gene sets, ontologies, or pathways. However, these approaches do not account for the hierarchical structure of pathways or interactions between genes. Topology-based approaches, which fully exploit all the knowledge about how genes interact as described by pathways, have been developed more recently. The first such techniques were ScorePAGE for metabolic pathways and the Impact Analysis for signaling pathways.

[0008] Non-coding RNAs, especially microRNAs (miRNAs) have come into the spotlight more recently. Data describing observed and predicted interactions between miRNA and mRNA is accumulating rapidly in several databases, such as, for example, miRTarBase, miRWalk, starBase, and TargetScan. In addition, miRNA expression platforms, datasets and analysis tools have become more and more prevalent.

[0009] Two of the most widely used approaches to include miRNA expression data for the purpose of pathway analysis are Micrographite and PARADIGM. Micrographite is a topology-aware pathway analysis approach that is able to integrate sample-matched miRNA and mRNA expression. PARADIGM uses a probabilistic graphical model (PGM) to integrate information of different data types, which may include mRNA and miRNA.

[0010] One drawback of these tools for integrating miRNA and mRNA is that they need sample-matched data. In other words, these tools require both data types to be available for each individual patient. This requirement reduces their practical availability because sample-matched data is relatively rare and difficult or expensive to obtain. Therefore, the vast amount of available expression data, both mRNA and miRNA, is not fully utilized.

[0011] Another drawback is that these methods are unable to exploit heterogeneous information available across independent studies. Therefore, they are not able to address the inevitable bias inherent in individual studies. It would be tremendously beneficial if all datasets associated with a given condition could be analyzed together because of the increased power expected to be associated with the much larger number of measurements in the combined dataset. Large public repositories such as Gene Expression Omnibus, The Cancer Genome Atlas (cancergenome.nih.gov), ArrayExpress, and Therapeutically Applicable Research to Generate Effective Treatments (ocg.cancer.gov/programs/target) store thousands of datasets, within which there are independent experimental series with similar patient cohorts and experiment design. Expression data, mRNA as well as miRNA, are particularly prevalent in public databases, such that some disease conditions are represented by half a dozen studies or more.

[0012] The process of combining sample-matched data of different types is referred to as "vertical" integrative analysis, while that of combining multiple unmatched studies using the same data type is referred as "horizontal" meta-analysis. Thus, the vertical and horizontal analyses are considered "orthogonal" classes of data integration. For microarray data, one of the earliest horizontal approaches for combining multiple microarray datasets included the use of Fisher's method. Since then, other sophisticated approaches have been proposed for the integration of multiple gene expression datasets, on both gene and pathway levels. The majority of these meta-analysis approaches work by combining p-values obtained from individual gene expression datasets. However, the approaches typically do not try to account for data heterogeneity, attributed to batch effects, patient heterogeneity, and disease complexity, responsible for expression changes across different sources. Accordingly, there remains a need for a framework that is able to integrate unmatched miRNA and mRNA data obtained from many independent laboratories.

SUMMARY

[0013] This section provides a general summary of the disclosure, and is not a comprehensive disclosure of its full scope or all of its features.

[0014] The current technology provides a method of integrating a plurality of data types. The method includes obtaining, via a processor, a plurality of datasets of a given type including measurements of one or more quantitative variables related to a phenotype comparison, and a plurality of datasets of a different type including measurements of one or more quantitative variables related to the same phenotype comparison; calculating, via the processor, a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the variables and for each dataset present in the plurality of datasets of the first type; calculating, via the processor, a second SMD, a second standard error, and a second p-value for each of the variables and for each data set present in the plurality of datasets of the second type; combining, via the processor, all the effect sizes in each individual dataset to calculate an effect size for each of the variables of the first data type, from the first SMD and the first standard error; combining, via the processor, all p-values in each individual dataset to calculate a global p-value for this first data type; combining, via the processor, all the effect sizes in each individual dataset to calculate an effect size for each of the variables of the second data type, from the second SMD and the second standard error; combining, via the processor, all p-values in each individual dataset to calculate a global p-value for the second data type; and combining, via the processor, the effect sizes of the variables of the first type with the effect sizes of the variables of the second type and/or combining the p-values of the variables of the first type with the p-values of the variables of the second type to identify the variables of either type that are relevant in the given phenotype comparison.

[0015] In various embodiments, there are more than two data types.

[0016] The current technology also provides a method of identifying a pathway associated with a disease. The method includes obtaining, via a processor, a plurality of first datasets describing a first quantitative variable related to the disease and a plurality of second datasets describing a second quantitative variable related to the disease, the plurality of first datasets and the plurality of second datasets being provided from independent studies, wherein each of the plurality of first datasets and each of the plurality of second datasets includes data regarding disease samples and healthy control samples; modifying, via the processor, known pathways related to the

disease with information provided in both the plurality of first datasets and the plurality of second datasets to generate augmented pathways including a plurality of first nodes associated with the first quantitative variable and a plurality of second nodes associated with the second quantitative variable, wherein the first nodes and second nodes are individually interconnected; calculating, via the processor, a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the plurality of first datasets; calculating, via the processor, a second SMD, a second standard error, and a second p-value for each of the plurality of second datasets; estimating, via the processor, a first effect size from the first SMD and the first standard error; combining, via the processor, the first p-values; estimating, via the processor, a second effect size from the second SMD and the second standard error; combining, via the processor, the second p-values; calculating, via the processor, a probability of obtaining at least an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values; and combining, via the processor, P_{NDE} and P_{PERT} to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease.

[0017] In various embodiments, the estimating a first effect size and the estimating a second effect size are performed by using a Restricted Maximum Likelihood (REML) algorithm.

[0018] In various embodiments, the combining the first p-values and the combining the second p-values is performed by add-CLT.

[0019] In various embodiments, the first quantitative variable and the second quantitative variable individually include one of molecular data and clinical data.

[0020] In various embodiments, the molecular data describes assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation; and the clinical data describes patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing score.

[0021] In various embodiments, the method further includes generating a plurality of single p-values corresponding to a plurality of pathways and generating a

graphical representation of the pathways ranked according to their corresponding single p-values.

[0022] The current technology also provides an apparatus for identifying a pathway associated with a disease. The apparatus includes a memory configured to store one or more applications; a processor communicatively coupled to memory, the processor, upon executing the one or more applications, is configured to: obtain a plurality of first datasets describing a first quantitative variable related to the disease and a plurality of second datasets describing a second quantitative variable related to the disease, the plurality of first datasets and the plurality of second datasets being provided from independent studies, wherein each of the plurality of first datasets and each of the plurality of second datasets includes data regarding disease samples and healthy control samples; modify known pathways related to the disease with information provided in both the plurality of first datasets and the plurality of second datasets to generate augmented pathways including a plurality of first nodes associated with the first quantitative variable and a plurality of second nodes associated with the second quantitative variable, wherein the first nodes and second nodes are individually interconnected; calculate a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the plurality of first datasets; calculate a second SMD, a second standard error, and a second p-value for each of the plurality of second datasets; estimate a first effect size from the first SMD and the first standard error; combine the first p-values; estimate a second effect size from the second SMD and the second standard error combine the second p-values; calculate a probability of obtaining at least an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values; and combine P_{NDE} and P_{PERT} to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease.

[0023] In various embodiments the processor is configured to estimate a first effect size and estimate a second effect size using a Restricted Maximum Likelihood (REML) algorithm.

[0024] In various embodiments the processor is configured to combine the first p-values and to combine the second p-values by add-CLT.

[0025] In various embodiments the first quantitative variable and the second quantitative variable individually include one of molecular data and clinical data.

[0026] In various embodiments the molecular data describes assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation; and the clinical data describes patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing score.

[0027] In various embodiments the processor is configured to generate a plurality of single p-values corresponding to a plurality of pathways and generate a graphical representation of the pathways ranked according to their corresponding single p-values.

[0028] In various embodiments, the processor is further configured to cause the graphical representation to be displayed at a display.

[0029] Additionally, the current technology provides a distributed computing system for identifying a pathway associated with a disease. The distributed computing system includes a first server configured to store a plurality of first datasets; a second server configured to store a plurality of second datasets, the second server different from the first server; a third server communicatively coupled to the first server and the second server via a distributed communication network, the third server including: a memory configured to store one or more applications; processor communicatively coupled to the memory, the processor, upon executing the one or more applications, is configured to: obtain the plurality of first datasets describing a first quantitative variable related to the disease and the plurality of second datasets describing a second quantitative variable related to the disease, the plurality of first datasets and the plurality of second datasets being provided from independent studies, wherein each of the plurality of first datasets and each of the plurality of second datasets includes data regarding disease samples and healthy control samples; modify known pathways related to the disease with information provided in both the plurality of first datasets and the plurality of second datasets to generate augmented pathways including a plurality of first nodes associated with the first quantitative variable and a plurality of second nodes associated with the second quantitative variable, wherein the first nodes and second nodes are individually interconnected; calculate a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the plurality of first datasets; calculate a second SMD, a second standard error, and a second p-value for each of the plurality of second datasets; estimate a first effect size from the first SMD and the first

standard error; combine the first p-values; estimate a second effect size from the second SMD and the second standard error; combine the second p-values; calculate a probability of obtaining at least an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values; and combine P_{NDE} and P_{PERT} to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease.

[0030] In various embodiments, the processor is configured to estimate a first effect size and estimate a second effect size using a Restricted Maximum Likelihood (REML) algorithm.

[0031] In various embodiments, the processor is configured to combine the first p-values and to combine the second p-values by add-CLT.

[0032] In various embodiments, the first quantitative variable and the second quantitative variable individually include one of molecular data and clinical data.

[0033] In various embodiments, the molecular data describes assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation; and the clinical data describes patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing score.

[0034] In various embodiments, the processor is configured to generate a plurality of single p-values corresponding to a plurality of pathways and generate a graphical representation of the pathways ranked according to their corresponding single p-values.

[0035] In various embodiments, the distributed computing system further includes a display, wherein the processor is further configured to cause display of the graphical representation at the display.

[0036] Further areas of applicability will become apparent from the description provided herein. The description and specific examples in this summary are intended for purposes of illustration only and are not intended to limit the scope of the present disclosure.

DRAWINGS

[0037] The drawings described herein are for illustrative purposes only of selected embodiments and not all possible implementations, and are not intended to limit the scope of the present disclosure.

[0038] FIG. 1 is a simplified block diagram of an example distributed computing system.

[0039] FIG. 2 is a functional block diagram of an example implementation of a client device.

[0040] FIG. 3 is a functional block diagram of an example implementation of a server.

[0041] FIG. 4 is a functional block diagram of an example database in accordance with an example implementation of the present disclosure.

[0042] FIG. 5 shows a graphical representation of a framework according to various aspects of the current technology. The input includes: (i) a pathway database and a miRNA database including known targets (panel a), (ii) multiple mRNA expression datasets (panel b), and (iii) multiple miRNA expression datasets (panel c). Each expression dataset includes two groups of samples, *e.g.*, disease versus control. The framework first augments the signaling pathways with miRNA molecules and their interactions with coding mRNA genes (panel d). It then calculates the standardized mean difference and its standard error in each expression dataset. The summary size effect across multiple datasets for each data type are then estimated using the REstricted Maximum Likelihood (REML) algorithm (panels e,f). Similarly, the p-value for differential expression is calculated for each dataset and then combined using the additive method (add-CLT). The augmented pathways, the combined p-values, and the estimated size effects then serve as input for ImpactAnalysis, which is a topology-aware pathway analysis method (panel g).

[0043] FIG. 6 shows a graphical representation of an augmented pathway regarding colorectal cancer. The green rectangle nodes (light shaded rectangles) and black arrows show the KEGG genes and their interactions while the blue nodes (dark shaded rectangles) and bar-headed lines show the miRNAs and their interactions with the genes, respectively. In each miRNA node added, the total number of miRNAs (circles) that are known to target the gene, and the names of the miRNA (blue (dark shaded) rectangles) that were actually measured in the 8 colorectal miRNA datasets,

are shown. This is a subset of the total set of miRNAs known to target genes on this pathway.

[0044] FIG. 7 shows a graphical representation of an augmented pathway regarding pancreatic cancer. The green rectangle nodes (dark shaded rectanges) and black arrows show the KEGG genes and their interactions while the blue nodes (dark shaded rectangles) and bar-headed lines show the miRNAs and their interactions with the genes. In each miRNA node added, the total number of miRNAs (circles) that are known to target the gene, and the names of the miRNA (blue (dark shaded) rectangles) that were actually measured in the 6 pancreatic miRNA datasets, are shown. This is a subset of the total set of miRNAs known to target genes on this pathway.

[0045] FIG. 8 is a flow chart illustrating an example method for identifying a pathway associated with a disease in accordance with an example embodiment of the present disclosure.

[0046] Corresponding reference numerals indicate corresponding parts throughout the several views of the drawings.

DETAILED DESCRIPTION

[0047] Example embodiments will now be described more fully with reference to the accompanying drawings.

[0048] The current technology provides a framework that is able to integrate unmatched miRNA and mRNA data obtained from many independent laboratories. While validated in the context of pathway analysis, the framework can be modified to adapt to other domains or applications. This framework is not meant to compete with any existing approach, but to serve as a bridge between "horizontal" and "vertical" data integration. Each building block or technique of the framework can be easily substituted for by any other similar technique to suit the purpose of future analysis.

[0049] The framework is illustrated using 15 mRNA and 14 miRNA datasets related to two human diseases (also referred to as "conditions"), colorectal cancer and pancreatic cancer. The datasets were generated by independent labs, for different sets of patients. For both conditions, the framework is able to identify pathways relevant to the phenotypes. Accuracy is obtained only by integrating the data in both directions (horizontal and vertical). However, it is understood that the framework can be applied to other diseases, conditions, or characteristics as well.

[0050] The framework provides an orthogonal meta-analysis. Orthogonal classes of integrative techniques can be further combined to unravel underlying mechanisms of complex diseases. With vast databases of various data types being made available, this framework is widely applicable because of its relaxed restrictions on the data being integrated.

[0051] Below are simplistic examples of a distributed computing environment in which the systems and methods of the present disclosure can be implemented. Throughout the description, references to terms such as servers, client devices, applications and so on are for illustrative purposes only. The terms server and client device are to be understood broadly as representing computing devices with one or more processors and memory configured to execute machine readable instructions. The terms application and computer program are to be understood broadly as representing machine readable instructions executable by the computing devices.

[0052] FIG. 1 shows a simplified example of an example computing system 100. The computing system 100 includes a distributed communications system 110, one or more client devices 120-1, 120-2, ..., and 120-M (collectively, client devices 120), and one or more servers 130-1, 130-2, ..., and 130-M (collectively, servers 130). N and M are integers greater than or equal to one. The distributed communications system 110 may include a local area network (LAN), a wide area network (WAN) such as the Internet, or other type of network. For example, the servers 130 may be located at different geographical locations. The client devices 120 and the servers 130 communicate with each other via the distributed communications system 110. The client devices 120 and the servers 130 connect to the distributed communications system 110 using wireless and/or wired connections.

[0053] The client devices 120 may include smartphones, personal digital assistants (PDAs), laptop computers, personal computers (PCs), etc. The servers 130 may provide multiple services to the client devices 120. For example, the servers 130 may execute software applications developed by one or more vendors. The server 130 may host multiple databases that are relied on by the software applications in providing services to users of the client devices 120.

[0054] FIG. 2 shows a simplified example of the client device 120-1. The client device 120-1 may typically include a central processing unit (CPU) or processor 150, one or more input devices 152 (e.g., a keypad, touchpad, mouse, touchscreen, etc.), a

display subsystem 154 including a display 156, a network interface 158, memory 160, and bulk storage 162.

[0055] The network interface 158 connects the client device 120-1 to the distributed computing system 100 via the distributed communications system 110. For example, the network interface 158 may include a wired interface (for example, an Ethernet interface) and/or a wireless interface (for example, a Wi-Fi, Bluetooth, near field communication (NFC), or other wireless interface). The memory 160 may include volatile or nonvolatile memory, cache, or other type of memory. The bulk storage 162 may include flash memory, a magnetic hard disk drive (HDD), and other bulk storage devices.

[0056] The processor 150 of the client device 120-1 executes an operating system (OS) 164 and one or more client applications 166. The client applications 166 include an application that accesses the servers 130 via the distributed communications system 110.

[0057] FIG. 3 shows a simplified example of the server 130-1. The server 130-1 typically includes one or more CPUs or processors 170, a network interface 178, memory 180, and bulk storage 182. In some implementations, the server 130-1 may be a general-purpose server and include one or more input devices 172 (e.g., a keypad, touchpad, mouse, and so on) and a display subsystem 174 including a display 176.

[0058] The network interface 178 connects the server 130-1 to the distributed communications system 110. For example, the network interface 178 may include a wired interface (e.g., an Ethernet interface) and/or a wireless interface (e.g., a Wi-Fi, Bluetooth, near field communication (NFC), or other wireless interface). The memory 180 may include volatile or nonvolatile memory, cache, or other type of memory. The bulk storage 182 may include flash memory, one or more magnetic hard disk drives (HDDs), or other bulk storage devices.

[0059] The processor 170 of the server 130-1 executes an operating system (OS) 184 and one or more server applications 186, which may be housed in a virtual machine hypervisor or containerized architecture. The bulk storage 182 may store one or more databases 188 that store data structures used by the server applications 186 to perform respective functions.

[0060] As shown in FIG. 4, the databases 188 store various data structures for storing multiple datasets. For example, a first database 202 may store a first dataset that describes a first quantitative variable related to the disease. A second database

204 may store a second dataset that describes a second quantitative variable related to the disease. While FIG. 4 illustrates a first database 202 and a second database 204, the distributed computing system 100 can include any number of databases without departing from the spirit of the disclosure. The databases 202, 204 store quantitative variables that can include molecular data and/or clinical data. For example, the molecular data can include assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation. The clinical data can include patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing scores. It is understood that the databases 202, 204 includes a plurality of datasets of a given type that include measurements of one or more quantitative variables related to a phenotype comparison. Additionally, the databases 202, 204 include a plurality of datasets of a different type that measurements of one or more quantitative variables related to the same phenotype comparison. The datasets can represent data pertaining to financial, health, business, social, geography, geology, and the like.

[0061] The server 130-1 receives and stores data to the corresponding data structures. The data can be received from the client devices 120-1 through 120-M and/or servers 130-2 through 130-N. The data can be provided by or obtained from disparate entities. In an example embodiment, the computing system 100 employs an edge computing architecture, a fog computing architecture, a centralized computing architecture, and the like. Thus, due to the quantity of data within the respective datasets 202, 204, data can be stored in databases 188 proximate to the server 130-1 allowing for resource pooling, latency reduction, and increased processing power.

[0062] As described herein, the processor 170 executes the one or more server applications 186 to perform the functionality described herein. For example, in one or more embodiments, the processor 170 accesses data within the various data structures to perform the functionality described herein.

Summary

[0063] MicroRNAs (miRNAs) are small non-coding RNA molecules whose primary function is to regulate the expression of gene products via hybridization to mRNA transcripts, resulting in suppression of translation or mRNA degradation. Although miRNAs have been implicated in complex diseases, including cancer, their impact on distinct biological pathways and phenotypes is largely unknown. Current integration approaches require sample-matched miRNA/mRNA datasets, resulting in

limited applicability in practice. Because these approaches cannot integrate heterogeneous information available across independent experiments, they neither account for bias inherent in individual studies, nor do they benefit from increased sample size. The current technology provides a novel framework able to integrate miRNA and mRNA data (vertical data integration) available in independent studies (horizontal meta-analysis) allowing for a comprehensive analysis of the given phenotypes. To demonstrate the utility of the framework, a meta-analysis of pancreatic and colorectal cancer, using 1,471 samples from 15 mRNA and 14 miRNA expression datasets, is conducted. The current two-dimensional data integration approach greatly increases the power of statistical analysis relative to conventional approaches and correctly identifies pathways known to be implicated in the phenotypes. The framework is general and can be used to integrate other types of data obtained from high-throughput assays.

Methods

[0064] The classical pathway analysis begins by considering a comparison between two conditions, *e.g.*, disease versus healthy. Evidence for differential gene expression can be provided by any technique such as fold change, t-statistic, Kolmogorov-Smirnov statistic, or perturbation factor. These statistics are then compared against a null distribution to determine how unlikely it is for the observed differences between the two conditions to occur by chance, thereby producing a ranked list of DE genes. After this hypothesis testing is done at the gene level, the next step is hypothesis testing at the pathway level producing a ranked list of impacted pathways. In summary, the input of a classical pathway analysis method includes: (i) a pathway database, and (ii) a gene expression dataset. The output is a list of pathways ranked according to their p-values.

[0065] Similarly, the input of the new approach includes: (i) a pathway database, (ii) a database of miRNA-mRNA interactions, (iii) multiple gene expression datasets, and (iv) multiple miRNA expression datasets. Each dataset is obtained from an independent study of the same disease. A framework that transforms the new problem into the classical pathway analysis problem is now provided.

[0066] Fig. 5 illustrates a pipeline of the framework for the case of colorectal cancer. Panel (a) represents biological knowledge obtained from databases: pathway information (i.e., database 204) and miRNA targets (i.e., database 202). Panel (b) shows a set of gene expression datasets obtained from independent studies coming

from different laboratories. Seven datasets (GSE4107, GSE9348, GSE15781, GSE21510, GSE23878, GSE41657, and GSE62322), related to the same disease, colorectal cancer, are used for this example. Each dataset has two groups of samples: disease (group D) and control/healthy (group C). Panel (c) represents a set of miRNA expression datasets (GSE33125, GSE35834, GSE39814, GSE39833, GSE41655, GSE49246, GSE54632, and GSE73487), also from colorectal cancer. Similar to gene expression datasets, each miRNA dataset consists of disease and control samples. The data provided in panels (a,b,c) serve as input for the framework.

[0067] Pathways in databases are typically described as graphs, where nodes are genes and edges are interactions between genes. In a first step, existing pathways are extended with additional interactions between miRNAs and mRNAs. Panel (d) shows a part of the pathway *Colorectal cancer*, where blue (circular) nodes are genes and red nodes (beginning with "mi") are miRNAs. Arrow-headed lines represent activation while bar-headed lines represent inhibition. For example, hsa-miR-483-5p is known to suppress the expression of MAPK3 and therefore an inhibition relationship is added between the two nodes in the pathway. All pathways are extended to include the known miRNA-mRNA interactions. Estimating expression changes of each node (gene, miRNA) under the effects of the disease is then performed.

[0068] Panel (e) shows expression changes and p-values for one gene in the mRNA data, across several datasets. Here, the MAPK3 gene is used as an example. In the forest plot shown in this panel, each horizontal line represents the expression change in each study. The small black box in each line shows a standardized mean difference (SMD) and the segment shows the confidence interval of SMD. Standardized mean difference is used instead of raw difference because the independent studies measure the expression in a variety of ways (different platforms, sample preparation, etc.). The number on the right side of each line is the p-value of the test for differential expression, using the modified t-test provided in the limma package.

[0069] As shown in Fig. 5, the SMD and p-value of a gene vary from study to study. REstricted Maximum Likelihood (REML) algorithm is used to estimate the central tendency of SMD. The add-CLT method is used to combine the independent p-values. Likewise, estimated SMDs and p-values for miRNA datasets (panel f) are computed.

[0070] The augmented pathways, the combined p-value, together with the estimated size effect then serve as input for classical pathway analysis. Here, Impact

Analysis, which is a topology-aware pathway analysis method, is used to calculate a p-value for each augmented pathway (panel g).

[0071] Standardized mean difference for each gene

[0072] As an example, a study composed of two independent groups is considered, and it is desired to compare their means for a given gene. Here, $\overline{x_1}$ and $\overline{x_2}$ represent the sample means for that gene in the two groups, n_1 and n_2 the number of samples in each group, and S_{pooled} the pooled standard deviation of the two groups. The pooled standard deviation and the standardized mean difference (SMD) can be estimated as follows.

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \quad (1)$$

$$\mathrm{d} = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}} \quad (2)$$

[0073] The estimation of the standardized mean difference described in Equation (2) may be called Cohen's *d*. The variance of Cohen's *d* is given as follows.

$$\mathrm{V}_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{\mathrm{d}^2}{2(n_1 + n_2)} \quad (3)$$

[0074] In the above equation, the first term reflects uncertainty in the estimate of the mean difference, and the second term reflects uncertainty in the estimate of S_{pooled}. The standard error of *d* is the square root of V_d. Cohen's *d*, which is based on sample averages, tends to overestimate the population effect size for small samples. *n* represents the degrees of freedom used to estimate S_{pooled}, *i.e.*, $n = n_1 + n_2 - 2$. The corrected effect size, or Hedges' *g*, is computed as follows:

$$J = \frac{\Gamma(\frac{n}{2})}{\sqrt{\frac{n}{2}}\Gamma\left(\frac{n-1}{2}\right)} \quad (4)$$

$$g = J \cdot d \quad (5)$$

where Γ is a gamma function. Here, Hedge' *g* is used as the standardized mean difference (SMD) between disease and control groups for each gene/miRNA.

[0075] Random-effects model and REML

[0076] A collection of *m* studies is considered, where the effect size estimates, $y_1, \ldots, y_m$ have been derived from a set of studies, each of them modeled as in Equation (5). A fixed-effects model would assume that there is one true effect size which underlies all of the studies in the analysis, such that all differences in observed

effects are due to sampling error. However, this assumption is implausible because it cannot account for heterogeneity between studies.

[0077] In contrast, the random-effects model allows for variability of the true effect. For example, the effect size might be higher (or lower) in studies where the participants are older, or have a healthier lifestyle compared to others. The random-effects model assumes that each effect size estimate can be decomposed into two variance components by a two-stage hierarchical process. The first variance represents variability of the effect size across studies, and the second variance represents sampling error within each study. The random-effects model may be:

$$y_i = \mu + N(0, \sigma^2) + N\left(0, \sigma_{\varepsilon_i}^2\right) \quad (6),$$

where μ is the central tendency of the effect size, $N(0,\sigma^2)$ represents the error term by which the effect size in the i^{th} study differs from the central tendency μ, and $N(0,\sigma_{\varepsilon_i}^2)$ represents the sampling error.

[0078] The derivation and formulation of the REstricted Maximum Likelihood (REML) algorithm is known in the art. The log-likelihood function for Equation (6) is given by Equation (7).

$$l(\mu, \sigma^2; y) = -\frac{1}{2}\sum_{i=1}^{m} \ln\left(\sigma^2 + \sigma_{\varepsilon_i}^2\right) - \frac{1}{2}\ln \sum_{i=1}^{m} \frac{1}{\sigma^2 + \sigma_{\varepsilon_i}^2} - \frac{1}{2}\sum_{i=1}^{m} \frac{(y_i - \mu)^2}{\sigma^2 + \sigma_{\varepsilon_i}^2} \quad (7)$$

[0079] The REML estimators of $\hat{\mu}$ and $\hat{\sigma}^2$ are then computed by iteratively maximizing the log-likelihood. In the current framework, $\hat{\mu}$ is calculated for each node (mRNA and miRNA) of the extended pathways. The estimated *overall effect size* $\hat{\mu}$ and the *combined p-value* of individual genes and miRNAs serve as input for Impact Analysis.

[0080] Combining independent p-values

[0081] Here is a summary of some classical methods for combining independent p-values. The additive method that is used to combine p-values for each mRNA and miRNA molecule in the current framework is then described.

[0082] Fisher's method is the most widely used method for combining independent p-values. Considering a set of m independent significance tests, the resulting p-values $P_1, P_2, \ldots, P_m$ are independent and uniformly distributed on the interval [0, 1] under the null hypothesis. The random variables $X_i = -2\ln P_i$ ($i \in \{1, 2, \ldots, m\}$) follow a chi-squared distribution with two degrees of freedom (χ_{2m}^2). Consequently, the log product of m independent p-values follows a chi-squared

distribution with $2m$ degrees of freedom. If one of the individual p-values approaches zero, which is often the case for empirical p-values, then the combined p-value approaches zero as well, regardless of other individual p-values. For example, if $P_1 \rightarrow 0$, then $X \rightarrow \infty$ and therefore, $Pr(X) \rightarrow 0$ regardless of P_2, P_3, ..., P_m.

[0083] Stouffer's method is another classical method that is closely related to Fisher's. The test statistic of Stouffer's method is the sum of p-values transformed into standard normal variables, divided by the square root of m. Denoting ϕ as the standard normal cumulative distribution function, and p_i ($i \in [1..m]$) the individual p-values that are independently and uniformly distributed under the null, the z-scores are calculated as $z_i = \phi^{-1}(1 - p_i)$. By definition, these z-scores follow the standard normal distribution. The summary statistic of Stouffer's method ($\frac{\sum_{i=1}^{m} z_i}{\sqrt{m}}$) also follows the standard normal distribution under the null hypothesis. Similar to Fisher's method, the combined p-values approach zero when one of the individual p-values approaches zero.

[0084] The additive method uses the sum of the p-values as the test statistic, instead of the log product. Consider the p-values resulting from m independent significance tests, P_1, P_2, ..., P_m. Let the sum of these p-values, $X = \sum_{i=1}^{m} P_i$ ($X \in [0, m]$), be the new random variable. X follows the Irwin-Hall distribution with the following probability density function (pdf):

$$f(x) = \frac{1}{(m-1)!} + \sum_{i=0}^{\lfloor x \rfloor} (-1)^i \binom{m}{i} (x-i)^{m-1} \quad (8)$$

when m is large, some addends will be too small or too large to be stored in the memory. This leads to a totally inaccurate calculation when m passes a certain threshold, depending on the number of bits used to store numbers on the computer. For this reason, a modified version of the additive method, named add-CLT, was proposed.

[0085] Let Y represent the average of p-values: $Y = \frac{\sum_{i=1}^{m} P_i}{m}$ ($Y \in [0,1]$). Since $Y = \frac{x}{m}$ the probability density function (pdf) and the corresponding cumulative distribution function (cdf) of Y can be derived using a linear transformation of X as follows:

$$g(y) = \frac{\mathrm{m}}{(m-1)!} + \sum_{i=0}^{\lfloor m \cdot y \rfloor} (-1)^i \binom{m}{i} (m \cdot y - i)^{m-1}$$

$$G(y) = \frac{1}{m!} + \sum_{i=0}^{\lfloor m \cdot y \rfloor} (-1)^i \binom{m}{i} (m \cdot y - i)^m \qquad (9)$$

The variable Y is the mean of m independent and identically distributed (i.i.d.) random variables (the p-values from each individual experiment), that follow a uniform distribution with a mean of $\frac{1}{2}$ and a variance of $\frac{1}{12}$. From the Central Limit Theorem, the average of such m i.i.d. variables follows a normal distribution with mean $\mu = \frac{1}{2}$ and variance $\sigma^2 = \frac{1}{12m}$, *i.e.*, $Y \sim N(\frac{1}{2}, \frac{1}{12m})$ for sufficiently large values of m. The transition from the additive method to the Central Limit Theorem takes place at the $m \geq 20$ threshold

[0086] Here, the add-CLT method described above is used to combine the p-values calculated from the modified t-test (limma package).

[0087] Graphical representation of augmented pathways

[0088] A formal description of the pathway augmentation process is provided. Let $P = (V, E)$ be the graphical representation of the pathway to be extended with miRNA-mRNA interactions. V is the set of vertices (genes) while the directed edges in E represent the interactions between genes in the pathway. Each interaction includes an ordered pair of vertices and the type of interaction between the pair, *i.e.*, $E = \{(x_i, y_i), r_i\}$ where $x_i, y_i \in G$ (gene set) and r_i is the type of relation between x_i and y_i, such as *activation, repression, phosphorylation, etc.* Topology-based pathway analysis methods, such as Impact Analysis, use interaction types to weigh the edges or to set the strength of signal propagation along the paths in a pathway.

[0089] From the miRNA database, a set of miRNAs and their targets is provided. Denote Z as the set of known miRNAs, $\zeta \in Z$ is one miRNA, and $t(\zeta)$ is the set of known targets for the miRNA ζ. The augmented pathway of $P = (V, E)$ is denoted as $P^* = (V^*, E^*)$ and is constructed as follows.

$$\mathrm{V}^* = \mathrm{V} \cup \{\zeta \in \mathrm{Z} : \mathrm{t}(\zeta) \cap \mathrm{V} \neq \emptyset\}$$
$$\mathrm{E}^* = \mathrm{E} \cup \{(\zeta, \mathrm{g}, repression) : \zeta \in \mathrm{Z}, \mathrm{g} \in \mathrm{t}(\zeta) \cap \mathrm{V}\} \qquad (10)$$

[0090] In other words, if a miRNA ζ targets a gene g that belongs to the pathway, ζ is added to the pathway and ζ is then connected with its targets in the pathway. By default, the interaction type of new edges is *repression*, which represents the translation blockage of miRNAs to mRNA. The interaction type can be changed to suit the interaction between the miRNA molecule and its targets. All pathways in the pathway database are extended using the formulation described in Equation (10). The

R package mirIntegrator for pathway augmentation is available on Bioconductor website (world wide web. bioconductor.org).

[0091] Impact analysis of augmented pathways

[0092] The Impact Analysis method combines two types of evidence: (i) the over-representation of DE genes in a given pathway, and (ii) the perturbation of that pathway, caused by disease, as measured by propagating expression changes through the pathway topology. These two aspects are captured, respectively, by the independent probability values, P_{NDE} and P_{PERT}. Impact Analysis formulation is summarized.

[0093] The first p-value, P_{NDE}, is obtained using the hypergeometric model, which is the probability of obtaining at least the observed number of differentially expressed genes. The second p-value, P_{PERT}, depends on the identity of the specific genes that are differentially expressed as well as on the interactions described by the pathway. It is calculated based on the perturbation factor in each pathway. The perturbation factor of a gene, $PF(g)$, is calculated as follows.

$$PF(g) = \Delta \mathrm{E(g)} + \sum_{\mu \epsilon US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \quad (11)$$

The first term represents the signed normalized expression change of the gene *g*, *i.e.*, log standardized mean difference as shown in panels (e,f) of Fig. 5. The second term is the sum of perturbation factors of upstream genes, normalized by the number of downstream genes of each such upstream gene. The value of β_{ug} quantifies the strength of interaction between *u* and *g*. Here, $\beta_{ug} = 1$ for *activation* and $\beta_{ug} = -1$ for *repression*.

[0094] The above equation essentially describes the perturbation factor PF for a gene as a linear function of the perturbation factors of all genes in a given pathway. In the stable state of the system, all relationships must hold, so the set of all equations defining the impact factors for all genes form a system of simultaneous equations whose solution will provide the values for the gene perturbation factors PF_g. The net perturbation accumulation at the level of each gene, $Acc(g)$, is calculated by subtracting the observed expression change from the perturbation factor.

$$\mathrm{Acc(g)} = \mathrm{PF(g)} - \Delta \mathrm{E(g)} \quad (12)$$

[0095] The total accumulated perturbation in the pathway is then computed as follows.

$$\mathrm{Acc}(P_i) = \sum_{g \epsilon P_i} Acc(g) \quad (13)$$

[0096] The null distribution of $Acc(P_i)$ is built by permutation of expression change. The p-value, P_{PERT}, is then calculated by the probability of having values more extreme than the actually observed $Acc(P_i)$.

[0097] To compute P_{NDE} and P_{PERT}, the following input is required: the graphical representation of the pathway, the combined p-value of each node of the graph, and the estimated overall standardized mean difference. In short, the graphical representation of the augmented pathways is provided in Equation (10), the p-value for each node of the augmented pathways is computed using Equation (9), and the expression change, $\Delta E(g)$, is estimated by iteratively maximizing the log-likelihood function in Equation (7). These two p-values, P_{NDE} and P_{PERT}, are then combined to get a single p-value that represents how likely the pathway is impacted under the effect of the disease. In one or more embodiments, the processor 170 causes the display 176 to generate a graphical representation of the single p-value. Additionally, the processor 170 causes the display 176 to generate a graphical representation of the impact analysis representing the disease and/or the augmented pathways (see panel (g) of FIG. 5).

Experimental Results

[0098] A total of 1,471 samples from 29 public datasets for two human diseases, colorectal and pancreatic cancer, were analyzed. The datasets were generated in independent laboratories, from different individual tissue samples, and were run on different high-throughput platforms. The diseases were selected based on two criteria: (i) there are many publicly available miRNA and mRNA datasets, and (ii) there is a pathway specific to the disease (target pathway). The colorectal data consists of 7 mRNA and 8 miRNA datasets while the pancreatic data consists of 8 mRNA and 6 miRNA datasets. The processed data sets were downloaded directly from the Gene Expression Omnibus using the GEOquery package. The data were rescaled using a log transformation if they were not already in log scale (base 2). The details of each dataset, such as the number of samples, tissues, and platforms, are reported in Table 1.

Table 1. Description of miRNA and mRNA expression datasets used in the experimental studies. All of the data were downloaded from Gene Expression Omnibus.

Cancer	Data	Accession ID	Control	Disease	Tissue	Platform
Colorectal	mRNA	GSE4107	10	12	Colonic mucosa	Affymetrix HG U133 Plus 2.0
		GSE9348	12	70	Colonic mucosa	Affymetrix HG U133 Plus 2.0
		GSE15781	10	13	Colon	ABI HG Survey 2
		GSE21510	25	123	Colon	Affymetrix HG U133 Plus 2.0
		GSE23878	24	35	Colon	Affymetrix HG U133 Plus 2.0
		GSE41657	12	25	Colonic mucosa, epithelial neoplasm	Agilent-014850 HG 4×44K G4112F
		GSE62322	18	20	Colon	Affymetrix HG U133A
	miRNA	GSE33125	9	9	Colon	Illumina Human v2 MicroRNA
		GSE35834	23	55	Colon & rectum	Affymetrix miRNA 1.0
		GSE39814	9	10	FHC, HCT116, & SW480 cells	Agilent-021827 Human miRNA
		GSE39833	11	88	Peripheral blood serum	Agilent-021827 Human miRNA
		GSE41655	15	33	Colonic mucosa, & epithelial neoplasm	Agilent-021827 Human miRNA
		GSE49246	40	40	Colon	Sun Yat-Sen Human microRNA
		GSE54632	5	5	Colonic and rectal mucosa	Affymetrix miRNA 1.0
		GSE73487	23	90	Colon	Affymetrix miRNA 1.0
Pancreatic	mRNA	GSE15471	39	39	Pancreas	Affymetrix HG U133 Plus 2.0
		GSE19279	3	4	Pancreas, pancreatic duct	Affymetrix HG U133A
		GSE27890	4	4	Pancreas, ductal epithelia	Affymetrix HG U133 Plus 2.0
		GSE32676	7	25	Pancreas	Affymetrix HG U133 Plus 2.0
		GSE36076	10	3	Peripheral blood mononuclear cells	Affymetrix HG U133 Plus 2.0
		GSE43288	3	4	Pancreas	Affymetrix HG U133A
		GSE45757	9	132	Pancreatic epithelial & cancer cells	Affymetrix HG U133A
		GSE60601	3	9	CD14++ & CD16- cells	Affymetrix HG U133 Plus 2.0
	miRNA	GSE24279	22	136	Pancreas	Febit human miRBase v11
		GSE25820	4	5	Pancreatic duct	Agilent-019118 Human miRNA
		GSE32678	7	25	Pancreas	miRCURY LNA microRNA, v.11.0
		GSE34052	6	6	Pancreas	Agilent-029297 Human miRNA
		GSE43796	5	26	Pancreas	Agilent-031181 Human miRNA V16
		GSE60978	6	51	Pancreatic duct	Agilent-031181 Human miRNA V16

[0099] The databases used in this analysis are KEGG for pathways, and miRTarBase for miRNAs. 182 signaling pathways are downloaded from KEGG version 76 (Dec-04-2015) by means of the R package ROntoTools. These pathways are augmented with known miRNAs and their target interactions, downloaded from miRTarBase. For each mRNA/miRNA, the *modified t-test*, available in the limma package, is used to test for differential expression of mRNA/miRNAs. a*dd-CLT* is used as the method to combine independent p-values. The combined p-values are then adjusted for multiple comparisons using False Discovery Rate (FDR). For expression change, *Hedges' g* is used as effect size, and the *REML* method is used to estimate the central tendency of effect sizes. Following convention, only mRNA/miRNAs having FDR-corrected combined p-values less than 5% are taken into consideration. Among these significant genes, mRNA/miRNAs are chosen that have the highest estimated

SMD as differentially expressed, up to 10% of total measured mRNA/miRNAs. All the R scripts used for data processing, pathway augmentation, and analysis are available.

[0100] For both diseases, the orthogonal approach (ImpactAnalysis_I) is compared with 5 other approaches: pathway-level meta-analysis (ImpactAnalysis_P), gene-level meta-analysis (ImpactAnalysis_G), plus the 3 meta-analysis approaches available in MetaPath package. Because the input data sets include multiple studies, none of which are sample-matched, pathway analysis using approaches that integrate matched mRNA and miRNA expression cannot be performed.

[0101] For pathway-level meta-analysis (ImpactAnalysis_P), Impact Analysis is performed on each mRNA expression dataset and then the independent p-values for each pathway are combined. For example, if there are 7 mRNA datasets, there are 7 nominal p-values per pathway–one for each study. These 7 p-values are independent and thus can be combined using the add-CLT method to get one combined p-value. The final result is a list of 182 p-values for 182 signaling pathways. The combined p-values for multiple comparisons are then adjusted using FDR.

[0102] For gene-level meta-analysis (ImpactAnalysis_G), the modified t-test for each mRNA dataset were performed and then the p-values were combined. With 7 mRNA datasets, for example, each gene will have 7 independent p-values, which will be combined into one p-value. We also calculate the SMD and standard error of each gene in each study, then use the REML algorithm to calculate the overall effect size across the 7 studies. Finally, pathway analysis is performed on 182 KEGG pathways using the combined p-values and the estimated effect sizes, resulting in a graphical representation, *i.e.*, a list, of pathways ranked according to their p-values. The p-values of pathways for multiple comparisons are adjusted using FDR.

[0103] The integrative approach (ImpactAnalysis_I) is similar to ImpactAnalysis_G, with the exception that ImpactAnalysis_I uses both mRNA and miRNA data. The meta-analysis is done on the mRNA/miRNA level and then the combined p-values and estimated effect sizes of mRNA/miRNAs serve as the input to the ImpactAnalysis.

[0104] MetaPath is a dedicated approach that performs meta-analysis at both gene (MetaPath_G) and pathway levels (MetaPath_P) with a GSEA-like approach, and then combines the results (MetaPath_I) to give the final p-value and ranking of pathways. MetaPath first calculates the t-statistic for each gene in each study. In MetaPath_G, these statistics are combined for each gene using maxP. The combined

statistics are then used to calculate enrichment scores for each pathway using a Kolmogorov-Smirnov test. In MetaPath_P, the pathway enrichment analysis is done first before meta-analysis. In MetaPath_I, the p-values of MetaPath_G and MetaPath_P are combined using minP.

[0105] For each of the two diseases, there is a *target* KEGG pathway, which is the pathway created to describe the main phenomena involved in the respective disease. The augmented pathway for *Colorectal cancer* is displayed in Fig. 6. The green rectangle nodes (light shaded rectangles) show the KEGG genes and the black arrows show the interactions between the genes. The blue nodes (dark shaded rectanges) and the bar-headed lines show the miRNA molecules and their interactions with the genes, where the bar-headed lines represents the "repression" activity. In each augmented node, two types of information are displayed: i) the total number of miRNAs that are known to target the corresponding gene, and ii) the miRNAs that were actually measured in the 8 miRNA colorectal datasets. The former is displayed in circles while the latter is listed in blue rectangles (dark shaded rectangles). For example, the gene TGFβ (in the far left of the figure) has 9 miRNAs that are known to target the gene but only two miRNAs (hsa:miR-375 and hsa:miR-633) were included in the miRNA data. Similarly, the augmented pathway for *Pancreatic cancer* is displayed in Fig. 7. The graphs show that both *target pathways* are heavily regulated by miRNA molecules.

[0106] In this experimental study, it is expected that a good pathway analysis approach would be able to identify the very pathway that describes the disease phenomena as the most significant in each particular disease. Hence, the various methods based on this criterion are compared.

[0107] Colorectal cancer

[0108] 8 miRNA (GSE33125, GSE35834, GSE39814, GSE39833, GSE41655, GSE49246, GSE54632, and GSE73487) and 7 mRNA (GSE4107, GSE9348, GSE15781, GSE21510, GSE23878, GSE41657, and GSE62322datasets are obtained from the Gene Expression Omnibus (GEO), as shown in Table 1.

[0109] Table 2 shows the results of the 6 approaches. The horizontal line across each list marks the cutoff $FDR = 0.01$. The pathway highlighted in green is the target pathway *Colorectal cancer*. MetaPath_P (pathway-level meta-analysis) identifies no significant pathway at the 1% cutoff, and ranks the target pathway at position 16^{th}. Similarly, MetaPath_G (gene-level meta-analysis) and MetaPath_I (combination of

gene- and pathway-level) identify no significant pathways. They rank the target pathway at positions 9^{th} and 15^{th}, respectively.

Table 2. The 16 top ranked pathways and FDR-corrected p-values obtained by combining colorectal data using 6 approaches: MetaPath_P, MetaPath_G, MetaPath_I, ImpactAnalysis_P, ImpactAnalysis_G, and ImpactAnalysis_I. The horizontal lines show the 1% significance threshold. The target pathway is colorectal cancer. All other approaches, MetaPath_P, MetaPath_G, MetaPath_I, ImpactAnalysis_P, ImpactAnalysis_G fail to identify the target pathway as significant, and rank it at the positions 16th, 9th, 15th, 61st, and 10th, respectively. On the contrary, the integrative approach, ImpactAnalysis_I, identifies the target pathway as significant and ranks it on top.

	MetaPath_P (mRNA, pathway-level)		MetaPath_G (mRNA, gene-level)		MetaPath_I (mRNA, both-level)	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Aldosterone-regulated sodium reabsoption	0.0941	Thyroid cancer	0.1460	Thyroid cancer	0.1460
2	Peroxisome	0.2319	Dorso-ventral axis formation	0.1533	Aldosterone-regulated sodium reabsorption	0.1880
3	Pancreatic cancer	0.2402	Mineral absorption	0.1550	Endocrine and other factor-regulated calcium reabsorption	0.2006
4	Small cell lung cancer	0.2500	PPAR signaling pathway	0.1575	Mineral absorption	0.2047
5	Endocrine and other factor-regulated calcium reabsorption	0.2540	Ribosome biogenesis in eukaryotes	0.2376	PPAR signaling pathway	0.2065
6	Epithelial cell signaling in Helicobacter pylori infection	0.2630	Renin-angiotensin system	0.2609	Dorso-ventral axis formation	0.227
7	Mineral absorption	0.2727	Vibrio cholerae infenction	0.3002	Small cell lung cancer	0.2713
8	Glioma	0.3234	Aldosterone-regulated sodium reabsorption	0.3478	Renin-angiotensin system	0.2731
9	Dorso-ventral axis formation	4.4665	*Colorectal cancer*	0.3514	Pancreatic cancer	0.2811
10	Epstein-Barr virus infection	0.4683	Bile secretion	0.4286	Peroxisome	0.2870
11	NOD-like receptor signaling pathway	0.4772	Pancreatic secretion	0.4361	Ribosome biogensis in eukaryotes	0.2906
12	Legionellosis	0.4772	Epithelial cell signaling in Helicobacter pylori infection	0.4427	Vibrio cholerae infection	0.2918
13	GmRH signaling pathway	0.4778	Intestinal immune netwrok for IgA production	0.4519	Epithelial cell signaling in Helicobacter	0.2951
14	Progesterone-mediated oocyte maturation	0.4946	Type I diabetes mellitus	0.4576	Glioma	0.3561
15	TNF signaling pathway	0.5135	Cardiac muscle contraction	0.4607	*Colorectal cancer*	0.4047
16	*Colorectal cancer*	0.5178	Allograft rejection	0.4616	NOD-like receptor signaling pathway	0.4693

	ImpactAnalysis_P (mRNA, pathway-level)		ImpactAnalysis_G (mRNA, gene-level)		ImpactAnalysis_I (mRNA, both-level)	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	PPAR signaling pathway	$<10^{-4}$	Ribosome biogenesis in eukaryotes	0.0008	*Colorectal cancer*	0.0002
2	Rheumatoid arthritis	$<10^{-4}$	Cell cycle	0.0008	Ribosome biogensis in eukaryotes	0.0002
3	Cytokine-cytokine receptor interaction	$<10^{-4}$	Mineral absorption	0.0185	PPAR signaling pathway	0.0002
4	Chemokine signaling pathway	$<10^{-4}$	p53 signaling pathway	0.0292	Cell cycle	0.0006
5	Bile secretion	$<10^{-4}$	Progesterone-mediated oocyte maturation	0.0347	Progesterone-mediated oocyte maturation	0.0077
6	MicroRNAs in cancer	0.0005	Oocyte Meiosis	0.0348	Oocyte meiosis	0.0130
7	Malaria	0.0007	Bile secretion	0.0364	TGF-beta signaling pathway	0.0130
8	Mineral absorption	0.0012	PPAR signaling pathway	0.0915	Parkinson's disease	0.0130
9	Pancreatic secretion	0.0046	Small cell lung cancer	0.1014	Peroxisome	0.0139
10	ECM-receptor interaction	0.0047	*Colorectal cancer*	0.1036	MicroRNAs in cancer	0.0140
11	Insulin secretion	0.0047	RNA transport	0.1059	Thyroid cancer	0.0214
12	Amoebiasis	0.0056	RNA degradation	0.1720	RNA transport	0.0214
13	Complement and coagulation cascades	0.0111	MicroRNAs in cancer	0.2051	AGE-RANGE signaling pathway in diabetic complications	0.0214
14	P13K-Akt signaling pathway	0.0131	Peroxisome	0.2051	NOD-like receptor signaling pathway	0.0304
15	TNF signaling pathway	0.0194	Pathways in cancer	0.2080	Endometrial cancer	0.0309
16	Transcriptional misregulation in cancer	0.0267	Parkinson's disease	0.3194	Pancreatic cancer	0.0309

[0110] The ImpactAnalysis_P approach identifies 12 pathways, among which there are many pathways that are related to cancer. However, the target pathway *Colorectal cancer* is not significant and is ranked 61^{st} with adjusted $p = 0.99$. The gene-level meta-analysis (ImpactAnalysis_G) offers some improvement over ImpactAnalysis_P by improving the ranking (10^{th}) and adjusted p-value ($p = 0.1$) of the target pathway *Colorectal cancer*. However, the target pathway is still not significant with the given threshold. The orthogonal meta-analysis, ImpactAnalysis_I, is able to further boost the power of the gene-level meta-analysis. It identifies 5 significant pathways, with the target pathway *Colorectal cancer* ranked at the very top. This is very likely due to the additional information provided by miRNA expression and prior knowledge accumulated in miRTarBase.

[0111] Three of the other 4 pathways that are identified by ImpactAnalysis_I appear to be true positives. The *Cell Cycle* and *Ribosome Biogenesis* pathways are implicated in the proliferation aspect of cancer tissue. *PPAR signaling* has a role in colorectal cancer, although it is not fully understood. *Progesterone-mediated oocyte maturation* is clearly a false positive which may have appeared due to the presence of several cell cycle genes in that pathway.

[0112] Pancreatic cancer

[0113] 8 mRNA (GSE15471, GSE19279, GSE27890, GSE32676, GSE36076, GSE43288, GSE45757, and GSE60601) and 6 miRNA datasets (GSE24279, GSE25820, GSE32678, GSE34052, GSE43796, and GSE60978) are obtained from Gene Expression Omnibus (GEO), as shown in Table 1. Again, the current approach (ImpactAnalysis_I) is compared with 5 other approaches: pathway-level meta-analysis, gene-level meta-analysis using only mRNA data, plus 3 meta-analysis approaches available in the MetaPath package as shown in Table 3.

Table 3. The 10 top ranked pathways and FDR-corrected ***p***-values obtained by combining colorectal data using 6 approaches: MetaPath_P, MetaPath_G, MetaPath_I, ImpactAnalysis_P, ImpactAnalysis_G, and ImpactAnalysis_I. The horizontal lines show the 1% significance threshold. The target pathway is *pancreatic cancer*. All other approaches, MetaPath_P, MetaPath_G, MetaPath_I, ImpactAnalysis_P, ImpactAnalysis_G fail to identify the target pathway as significant, and rank it at the positions 17^{th}, 91^{st}, 91^{st}, 32^{nd}, and 8^{th}, respectively. On the contrary, the integrative approach, ImpactAnalysis_I, identifies the target pathway as significant and ranks it on top.

	MetaPath_P (mRNA, pathway-level)		MetaPath_G (mRNA, gene-level)		MetaPath_I (mRNA, both-level)	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	Graft-versus-host-disease	0.4782	Autoimmune thyroid disease	0.0020	Type I diabetes mellitus	0.0040
2	Small cell lung cancer	0.5440	Allograft rejection	0.0020	Autoimmune thyroid disease	0.0040
3	SNARE interactions in vesicular transport	0.553	Type I diabetes mellitus	0.003	Allograft rejection	0.004
4	Leishmaniasis	0.6404	Graft-versus-host disease	0.0040	Graft-versus-host-disease	0.0080
5	Bladder cancer	0.7010	GABAergic synapse	0.0050	GABAergic synapse	0.0100
6	MicroRNAs in cancer	0.7244	Asthma	0.0073	Asthma	0.0147
7	Phagosome	0.7330	Morphine addiction	0.0074	Morphine addiction	0.0149
8	Type I diabetes mellitus	0.7515	ECM-receptor interaction	0.0104	ECM-receptor interaction	0.0208
9	Pertussis	0.7682	Maturity onset diabetes of the young	0.0139	Maturity onset diabetes of the young	0.0278
10	Dorso-ventral axis formation	0.7941	Renin-angiotensin system	0.0153	Renin-angiotensin system	0.0307

	ImpactAnalysis_P (mRNA, pathway-level)		ImpactAnalysis_G (mRNA, gene-level)		**ImpactAnalysis_I (mRNA + miRNA)**	
	Pathway	p.fdr	Pathway	p.fdr	Pathway	p.fdr
1	PI3K-Akt signaling pathway	0.0019	Small cell lung cancer	0.0217	Pancreatic cancer	0.0017
2	MicroRNAS in cancer	0.0076	Pathways in cancer	0.0217	Small cell lung cancer	0.0017
3	Small cell lung cancer	0.0276	Viral carcinogenesis	0.0217	Pathways in cancer	0.0017
4	Pathways in cancer	0.0962	ECM-receptor interaction	0.0480	Proteoglycans in cancer	0.0017
5	TNF signaling pathway	0.1106	Hepatitis B	0.0480	Amoebiasis	0.0031
6	PPAR signaling pathway	0.1216	HRLV-I infection	0.0623	AGE-RANGE signaling pathway in diabetic complications	0.0040
7	NF-kappa B signaling pathway	0.1502	Chronic myeloid leukemia	0.0623	Focal adhesion	0.0040
8	Shigellosis	0.2491	*Pancreatic cancer*	0.0623	HTLV-I infection	0.0119
9	Chemokine signaling pathway	0.2742	Amoebiasis	0.0639	Chronic myeloid leukemia	0.0125
10	T cell receptor signaling pathway	0.3200	Pathogenic Escherichia coli infection	0.0639	ECM-receptor interaction	0.0142

[0114] MetaPath_P identifies no significant pathway and *Graft-versus-host disease* is ranked on top with adjusted p-value 0.4782. The target pathway *Pancreatic cancer* is ranked 17^{th} with adjusted $p = 0.89$. MetaPath_G identifies 7 significant pathways. The target pathway is not significant (adjusted $p = 0.22$) and is ranked 91^{st}. In consequence, the combination of these two methods, MetaPath_I, also fails to identify the target pathway as significant (adjusted $p = 0.34$ with ranking 91^{st}).

[0115] The pathway-level meta-analysis (ImpactAnalysis_P) identifies the *PI3K-Akt signaling pathway* and *MicroRNAs in cancer* as significant. The significance of *MicroRNAs in cancer* may indicate the importance of miRNA in pancreatic cancer, and *PI3K-Akt signaling* alteration is known to be involved in many cancers. However, the target pathway is not significant (adjusted $p = 0.95$ with ranking 32^{nd}). The gene-level meta-analysis (ImpactAnalysis_G) improves the ranking of the target pathway (8^{th}) but the p-value of the target pathway is still not significant. The orthogonal approach, ImpactAnalysis_I, identifies 7 pathways as significant. The target pathway *Pancreatic cancer* is ranked on top with FDR-corrected p-value 0.0017.

[0116] Of the 6 significant non-target pathways found by ImpactAnalysis_I, three are cancer-related by name (*Small cell lung cancer, Pathways in cancer, Proteoglycans in cancer*). The breakdown of cell matrix adhesions, such as *Focal Adhesion* is an

important property of metastasis - most pancreatic cancers are discovered when they are already high grade.

[0117] In contrast to the 3 variations of the existing method, MetaPath, the proposed method ImpactAnalysis_I was able to effectively combine both independent datasets, as well as the two different types of data (mRNA and miRNA), and correctly report the target pathway as the most significantly impacted pathway in both meta-analysis studies. The results demonstrate that the correct pathways are identified only when the data are integrated both horizontally (combining multiple studies using the same data type) and vertically (combining miRNA with mRNA expression). This orthogonal meta-analysis uses three different kinds of data integration: integration of mRNA and miRNA, combining p-values and combining SMDs for genes and miRNA molecules.

[0118] Time complexity

[0119] The data analysis was done on a personal MacBook Pro that has 8 GB 1600 MHz DDR3 RAM, 2.9 GHz Intel Core i7. Because MetaPath cannot exploit multiple processors, all the analysis were run using a single core. The time needed to run MetaPath was 39 minutes for Colorectal cancer and 47 minutes for Pancreatic cancer.

[0120] For ImpactAnalysis_I, the p-value for each gene/miRNA in each dataset is first calculated using the limma package. The p-values are then combined to get one combined p-value per gene/miRNA. Next, the standardized mean difference (SMD) is calculated for each dataset and then the REML algorithm is applied to estimate to overall SMD, using the metafor package. The estimated SMDs and the combined p-values are processed by ROntoTools to produce the p-value for each pathway. ImpactAnalysis_I performes the analysis using the pathways augmented with the relevant miRNAs. The running time for ImpactAnalysis_I is 4 minutes for each of Colorectal and Pancreatic. The running time of each approach is reported in Table 4.

Table 4. Running time of each pathway analysis in minutes (m).

Method	Input	Colorectal	Pancreatic
ImpactAnalysis_I	mRNA & miRNA	4 m	4 m
MetaPath	mRNA	39 m	47 m

Discussion

[0121] One straightforward *horizontal* integration is to combine individual p-values provided by each study. In this way, any pathway analysis approach (such as GSEA or GSA) can be applied to the collected mRNA datasets in order to calculate a p-value for each pathway in each study, and then combine these independent p-values. An advantage of this approach is its flexibility. MetaPath combines p-values in this way, but with the slight difference that the p-values are combined on both gene and pathway levels. The drawback is that each of these methods is designed to work with one single matrix of expression values, *i.e.*, one data type. This matrix can be forcefully extended to include other data types as well, but in order to do this, the data must be sample-matched. In other words, all types of assays must be performed on every single sample. In addition, because different data types are assayed on different platforms, the data need to be normalized together, for these approaches to function properly. However, the correct way to do such a cross-platform normalization is still an open problem. The same limitations apply to analysis tools dedicated to miRNA and mRNA integration. For meta-analysis, these approaches would require multiple sets of sample-matched data. Performing different assays on one set of samples is already expensive; asking for many sets of matched samples for the same disease is even more impractical.

[0122] Although primarily designed to overcome the matched-sample bottleneck discussed above, the current framework also aims to address a well-known limitation of p-value-based meta-analyses. Classical approaches often rely on hypothesis testing to identify differential expression. This results in critical information loss. While the p-value is partly a function of effect size, it is also partly a function of sample size. For example, with large sample size, a statistical test will tend to find differences as significant, unless the effect size is exactly zero. In reality, any individual study will include some degree of batch effects, such as sampling/study bias, noise, and measurement errors. Simply combining individual p-values would not correct such problems. On the contrary, meta-analysis of effect sizes across all studies would definitely compensate for and eliminate such random effects. This point is illustrated in the results included herein, in particular in the difference between ImpactAnalysis_P and ImpactAnalysis_G for both colorectal and pancreatic cancer (Tables 2 and 3). The former simply combines the p-values, while the latter takes into consideration both p-values and effect sizes across different

studies. ImpactAnalysis_G offers a great improvement over ImpactAnalysis_P using the same sets of mRNA data.

[0123] The current framework contemplates the computational complexity at both gene and pathway levels. For individual genes and miRNA molecules, the framework not only calculates p-values, but also iteratively estimates the effect sizes and variances. In principle, the iterative algorithm requires more computation than meta-analyses that use closed-form expressions. At pathway-level, Impact Analysis is a non-parametric approach that constructs an empirical distribution of all measured values for each pathway. This requires more computation and storage than parametric approaches, such as the hypergeometric test or Fisher's exact test. However, this is mitigated by the power of modern computers which are able to perform all needed computations in less than 10 minutes, even for datasets with more than 1,000 samples (Table 4). In addition, the current framework allows for parallel computing at the gene-level to reduce the time complexity. However, the time values described here (see, for example, Table 4) do not take advantage of the ability to parallelize the computation in order to be comparable with the results obtained with MetaPath. All values reported in this table are obtained on a single core for both approaches.

[0124] The biological results presented here could be further validated by investigating the other pathways reported as significant, and identifying the putative mechanisms that could explain all measured changes. A tool such as iPathway-Guide, could be used to provide more in depth functional analysis, including identification of drugs that are known to act on the observed signaling cascades. Follow-up experiments in which tumor cell lines, or samples from xenografts, are treated with those drugs would validate (or not) both the putative mechanisms investigated, as well as the other significant pathways. If many or all significant pathways were mechanistically implicated in the respective conditions, the proposed orthogonal meta-analysis approach would be further validated.

[0125] Another direct application of the orthogonal framework is to infer condition-specific miRNA activity. The proposed gene-level meta-analysis basically identifies genes and miRNAs that are differentially expressed (DE) under the studied condition. This list of DE genes/miRNAs is obtained from a large number of studies and therefore it is expected to be more reliable than any individual study taken alone. From the list of DE genes/miRNAs and the computed statistics (effect sizes and variances), new putative targets of miRNAs can be identified using casual inference techniques.

The predicted interactions between miRNA and mRNA can be further verified by established gene-specific experimental validation, such as qRT-PCR, luciferase reporter assays, and western blot.

Summary

[0126] A two-dimensional data integration that is able to combine mRNA and miRNA expression data obtained from many independent experiments is provided herein. The framework first augments pathway knowledge available in pathway databases with miRNA-mRNA interactions from miRNA knowledge bases. It then computes the statistics that are essential for pathway analysis, *i.e.*, the standardized mean difference (SMD) and p-value for differential expression. For each entity, these p-values and the SMDs are computed by combining multiple studies using robust horizontal meta-analysis techniques. Finally, the framework performs a topology-based pathway analysis to identify pathways that are likely to be impacted under the given condition.

[0127] To evaluate the framework, 1,471 samples from 15 mRNA and 14 miRNA expression datasets related to two human cancers were examined using 6 different meta-analysis approaches (3 MetaPath approaches and 3 meta-analysis approaches that utilize Impact Analysis). It was demonstrated that the correct pathways are identified only when the data are integrated both horizontally (combining multiple studies using the same data type) and vertically (combining miRNA with mRNA expression).

[0128] This technology serves as a bridge between the two orthogonal types of data integration. The result is to unblock the sample-matched data bottleneck, by successfully integrating mRNA and miRNA datasets measured from independent laboratories for different sets of patients. Furthermore, it increases the power of statistical approaches because it allows many studies to be analyzed together. With vast databases of various data types being made available, this framework is widely applicable because of its relaxed restrictions on the data being integrated. The framework is flexible enough to integrate data types other than mRNA and miRNA, which was described herein as an example. It can also be modified to suit other purposes besides pathway analysis.

[0129] FIG. 8 illustrates an example method 800 for identifying a pathway associated with a disease in accordance with an example embodiment of the present disclosure. Method 800 begins at 802. At 804, multiple data structures, such as databases 202,

204 that provide a first dataset describing a first quantitative variable related to the disease and a second dataset describing a second quantitative variable related to the disease is provided.

[0130] At 806, known pathways are modified that are related to the disease with information provided in both the first datasets and the second datasets to generate augmented pathways including a plurality of first nodes associated with the first quantitative variable and a plurality of second nodes associated with the second quantitative variable. At 808, a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the first datasets is calculated.

[0131] At 810, a second standardized mean difference (SMD), a second standard error, and a second p-value for each of the second datasets is calculated. At 812, a first effect size from the first SMD and the first standard error is estimated. At 814, the first p-values are combined. At 816, a second effect size from the second SMD and the second standard error is estimated. At 818, the second p-values are combined.

[0132] At 820, a probability of obtaining an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values. At 822, the P_{NDE} and the P_{PERT} are combined to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease. At 824, the method 800 ends.

Conclusion

[0133] Spatial and functional relationships between elements (for example, between modules) are described using various terms, including "connected," "engaged," "interfaced," and "coupled." Unless explicitly described as being "direct," when a relationship between first and second elements is described in the above disclosure, that relationship encompasses a direct relationship where no other intervening elements are present between the first and second elements, and also an indirect relationship where one or more intervening elements are present (either spatially or functionally) between the first and second elements. As used herein, the phrase at least one of A, B, and C should be construed to mean a logical (A OR B OR C), using a non-

exclusive logical OR, and should not be construed to mean "at least one of A, at least one of B, and at least one of C."

[0134] In the figures, the direction of an arrow, as indicated by the arrowhead, generally demonstrates the flow of information (such as data or instructions) that is of interest to the illustration. For example, when element A and element B exchange a variety of information but information transmitted from element A to element B is relevant to the illustration, the arrow may point from element A to element B. This unidirectional arrow does not imply that no other information is transmitted from element B to element A. Further, for information sent from element A to element B, element B may send requests for, or receipt acknowledgements of, the information to element A.

[0135] In this application, including the definitions below, the term 'module' or the term 'controller' may be replaced with the term 'circuit.' The term 'module' may refer to, be part of, or include processor hardware (shared, dedicated, or group) that executes code and memory hardware (shared, dedicated, or group) that stores code executed by the processor hardware.

[0136] The module may include one or more interface circuits. In some examples, the interface circuits may include wired or wireless interfaces that are connected to a local area network (LAN), the Internet, a wide area network (WAN), or combinations thereof. The functionality of any given module of the present disclosure may be distributed among multiple modules that are connected via interface circuits. For example, multiple modules may allow load balancing. In a further example, a server (also known as remote, or cloud) module may accomplish some functionality on behalf of a client module.

[0137] The term code, as used above, may include software, firmware, and/or microcode, and may refer to programs, routines, functions, classes, data structures, and/or objects. Shared processor hardware encompasses a single microprocessor that executes some or all code from multiple modules. Group processor hardware encompasses a microprocessor that, in combination with additional microprocessors, executes some or all code from one or more modules. References to multiple microprocessors encompass multiple microprocessors on discrete dies, multiple microprocessors on a single die, multiple cores of a single microprocessor, multiple threads of a single microprocessor, or a combination of the above.

[0138] Shared memory hardware encompasses a single memory device that stores some or all code from multiple modules. Group memory hardware encompasses a memory device that, in combination with other memory devices, stores some or all code from one or more modules.

[0139] The term memory hardware is a subset of the term computer-readable medium. The term computer-readable medium, as used herein, does not encompass transitory electrical or electromagnetic signals propagating through a medium (such as on a carrier wave); the term computer-readable medium is therefore considered tangible and non-transitory. Non-limiting examples of a non-transitory computer-readable medium are nonvolatile memory devices (such as a flash memory device, an erasable programmable read-only memory device, or a mask read-only memory device), volatile memory devices (such as a static random access memory device or a dynamic random access memory device), magnetic storage media (such as an analog or digital magnetic tape or a hard disk drive), and optical storage media (such as a CD, a DVD, or a Blu-ray Disc).

[0140] The apparatuses and methods described in this application may be partially or fully implemented by a special purpose computer created by configuring a general purpose computer to execute one or more particular functions embodied in computer programs. The functional blocks and flowchart elements described above serve as software specifications, which can be translated into the computer programs by the routine work of a skilled technician or programmer.

[0141] The computer programs include processor-executable instructions that are stored on at least one non-transitory computer-readable medium. The computer programs may also include or rely on stored data. The computer programs may encompass a basic input/output system (BIOS) that interacts with hardware of the special purpose computer, device drivers that interact with particular devices of the special purpose computer, one or more operating systems, user applications, background services, background applications, etc.

[0142] The computer programs may include: (i) descriptive text to be parsed, such as HTML (hypertext markup language), XML (extensible markup language), or JSON (JavaScript Object Notation) (ii) assembly code, (iii) object code generated from source code by a compiler, (iv) source code for execution by an interpreter, (v) source code for compilation and execution by a just-in-time compiler, etc. As examples only, source

code may be written using syntax from languages including C, C++, C#, Objective-C, Swift, Haskell, Go, SQL, R, Lisp, Java®, Fortran, Perl, Pascal, Curl, OCaml, Javascript®, HTML5 (Hypertext Markup Language 5th revision), Ada, ASP (Active Server Pages), PHP (PHP: Hypertext Preprocessor), Scala, Eiffel, Smalltalk, Erlang, Ruby, Flash®, Visual Basic®, Lua, MATLAB, SIMULINK, and Python®.

[0143] None of the elements recited in the claims are intended to be a means-plus-function element within the meaning of 35 U.S.C. §112(f) unless an element is expressly recited using the phrase “means for” or, in the case of a method claim, using the phrases “operation for” or “step for.”

CLAIMS

What is claimed is:

1. A method of integrating a plurality of data types, the method comprising:

obtaining, via a processor, a plurality of datasets of a given type comprising measurements of one or more quantitative variables related to a phenotype comparison, and a plurality of datasets of a different type comprising measurements of one or more quantitative variables related to the same phenotype comparison;

calculating, via the processor, a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the variables and for each dataset present in the plurality of datasets of the first type;

calculating, via the processor, a second SMD, a second standard error, and a second p-value for each of the variables and for each data set present in the plurality of datasets of the second type;

combining, via the processor, all the effect sizes in each individual dataset to calculate an effect size for each of the variables of the first data type, from the first SMD and the first standard error;

combining, via the processor, all p-values in each individual dataset to calculate a global p-value for this first data type;

combining, via the processor, all the effect sizes in each individual dataset to calculate an effect size for each of the variables of the second data type, from the second SMD and the second standard error;

combining, via the processor, all p-values in each individual dataset to calculate a global p-value for the second data type; and

combining, via the processor, the effect sizes of the variables of the first type with the effect sizes of the variables of the second type and/or combining the p-values of the variables of the first type with the p-values of the variables of the second type to identify the variables of either type that are relevant in the given phenotype comparison.

2. The method according to Claim 1, wherein there are more than two data types.

3. A method of identifying a pathway associated with a disease, the method comprising:

obtaining, via a processor, a plurality of first datasets describing a first quantitative variable related to the disease and a plurality of second datasets describing a second quantitative variable related to the disease, the plurality of first datasets and the plurality of second datasets being provided from independent studies, wherein each of the plurality of first datasets and each of the plurality of second datasets comprises data regarding disease samples and healthy control samples;

modifying, via the processor, known pathways related to the disease with information provided in both the plurality of first datasets and the plurality of second datasets to generate augmented pathways comprising a plurality of first nodes associated with the first quantitative variable and a plurality of second nodes associated with the second quantitative variable, wherein the first nodes and second nodes are individually interconnected;

calculating, via the processor, a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the plurality of first datasets;

calculating, via the processor, a second SMD, a second standard error, and a second p-value for each of the plurality of second datasets;

estimating, via the processor, a first effect size from the first SMD and the first standard error;

combining, via the processor, the first p-values;

estimating, via the processor, a second effect size from the second SMD and the second standard error;

combining, via the processor, the second p-values;

calculating, via the processor, a probability of obtaining at least an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values; and

combining, via the processor, P_{NDE} and P_{PERT} to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease.

4. The method according to Claim 3, wherein the estimating a first effect size and the estimating a second effect size are performed by using a Restricted Maximum Likelihood (REML) algorithm.

5. The method according to Claim 3, wherein the combining the first p-values and the combining the second p-values is performed by add-CLT.

6. The method according to Claim 3, wherein the first quantitative variable and the second quantitative variable individually comprise one of molecular data and clinical data.

7. The method according to Claim 6, wherein:

the molecular data describes assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation; and

the clinical data describes patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing score.

8. The method according to Claim 3, further comprising:

generating a plurality of single p-values corresponding to a plurality of pathways and generating a graphical representation of the pathways ranked according to their corresponding single p-values.

9. An apparatus for identifying a pathway associated with a disease, the apparatus comprising:

a memory configured to store one or more applications;

a processor communicatively coupled to memory, the processor, upon executing the one or more applications, is configured to:

obtain a plurality of first datasets describing a first quantitative variable related to the disease and a plurality of second datasets describing a second quantitative variable related to the disease, the plurality of first datasets and the plurality of second datasets being provided from independent studies, wherein each of the plurality of first datasets and each of the plurality of second datasets comprises data regarding disease samples and healthy control samples;

modify known pathways related to the disease with information provided in both the plurality of first datasets and the plurality of second datasets to generate augmented pathways comprising a plurality of first nodes associated with the first quantitative variable and a plurality of second nodes associated with

the second quantitative variable, wherein the first nodes and second nodes are individually interconnected;

calculate a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the plurality of first datasets;

calculate a second SMD, a second standard error, and a second p-value for each of the plurality of second datasets;

estimate a first effect size from the first SMD and the first standard error;

combine the first p-values;

estimate a second effect size from the second SMD and the second standard error;

combine the second p-values;

calculate a probability of obtaining at least an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values; and

combine P_{NDE} and P_{PERT} to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease.

10. The apparatus according to Claim 9, wherein the processor is configured to estimate a first effect size and estimate a second effect size using a Restricted Maximum Likelihood (REML) algorithm.

11. The apparatus according to Claim 9, wherein the processor is configured to combine the first p-values and to combine the second p-values by add-CLT.

12. The apparatus according to Claim 9, wherein the first quantitative variable and the second quantitative variable individually comprise one of molecular data and clinical data.

13. The apparatus according to Claim 12, wherein:

the molecular data describes assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation; and

the clinical data describes patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing score

14. The apparatus according to Claim 9, wherein the processor is configured to generate a plurality of single p-values corresponding to a plurality of pathways and generate a graphical representation of the pathways ranked according to their corresponding single p-values.

15. The apparatus according to Claim 14, wherein the processor is further configured to cause the graphical representation to be displayed at a display.

16. A distributed computing system for identifying a pathway associated with a disease, the distributed computing system comprising:

a first server configured to store a plurality of first datasets;

a second server configured to store a plurality of second datasets, the second server different from the first server;

a third server communicatively coupled to the first server and the second server via a distributed communication network, the third server comprising:

a memory configured to store one or more applications;

a processor communicatively coupled to the memory, the processor, upon executing the one or more applications, is configured to:

obtain the plurality of first datasets describing a first quantitative variable related to the disease and the plurality of second datasets describing a second quantitative variable related to the disease, the plurality of first datasets and the plurality of second datasets being provided from independent studies, wherein each of the plurality of first datasets and each of the plurality of second datasets comprises data regarding disease samples and healthy control samples;

modify known pathways related to the disease with information provided in both the plurality of first datasets and the plurality of second datasets to generate augmented pathways comprising a plurality of first nodes associated

with the first quantitative variable and a plurality of second nodes associated with the second quantitative variable, wherein the first nodes and second nodes are individually interconnected;

calculate a first standardized mean difference (SMD), a first standard error, and a first p-value for each of the plurality of first datasets;

calculate a second SMD, a second standard error, and a second p-value for each of the plurality of second datasets;

estimate a first effect size from the first SMD and the first standard error;

combine the first p-values;

estimate a second effect size from the second SMD and the second standard error;

combine the second p-values;

calculate a probability of obtaining at least an observed relationship between the first and second quantitative variables associated with the disease (P_{NDE}) and a p-value that depends on identities of first or second quantitative variables that are differentially related and described by the pathway (P_{PERT}) from the augmented pathways, the estimated first effect size, the combined first p-values, the estimated second effect size, and the combined second p-values; and

combine P_{NDE} and P_{PERT} to generate a single p-value that represents how likely a pathway is impacted under the effect of the disease.

17. The distributed computing system according to Claim 16, wherein the processor is configured to estimate a first effect size and estimate a second effect size using a Restricted Maximum Likelihood (REML) algorithm.

18. The distributed computing system according to Claim 16, wherein the processor is configured to combine the first p-values and to combine the second p-values by add-CLT.

19. The distributed computing system according to Claim 16, wherein the first quantitative variable and the second quantitative variable individually comprise one of molecular data and clinical data.

20. The distributed computing system according to Claim 19, wherein:

the molecular data describes assay results related to at least one of mRNA, miRNA, protein abundance, metabolite abundance, and methylation; and

the clinical data describes patient information related to at least one of weight, blood pressure, blood metabolite level, blood sugar, heart rate, vision score, and hearing score.

21. The distributed computing system according to Claim 16, wherein the processor is configured to generate a plurality of single p-values corresponding to a plurality of pathways and generate a graphical representation of the pathways ranked according to their corresponding single p-values.

22. The distributed computing system according to Claim 21, further comprising a display, wherein the processor is further configured to cause display of the graphical representation at the display.

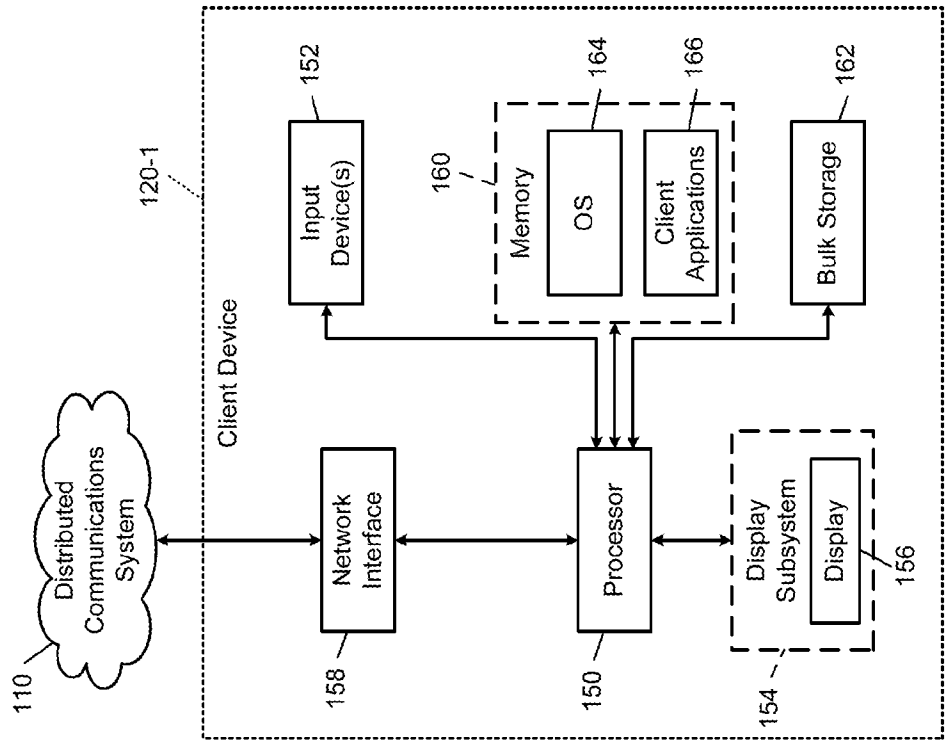
120-1
Client Device
Input Device(s)
152
160
Memory
OS
164
Client Applications
166
Bulk Storage
162
Distributed Communications System
110
Network Interface
158
Processor
150
Display Subsystem
154
Display
156

FIG. 2

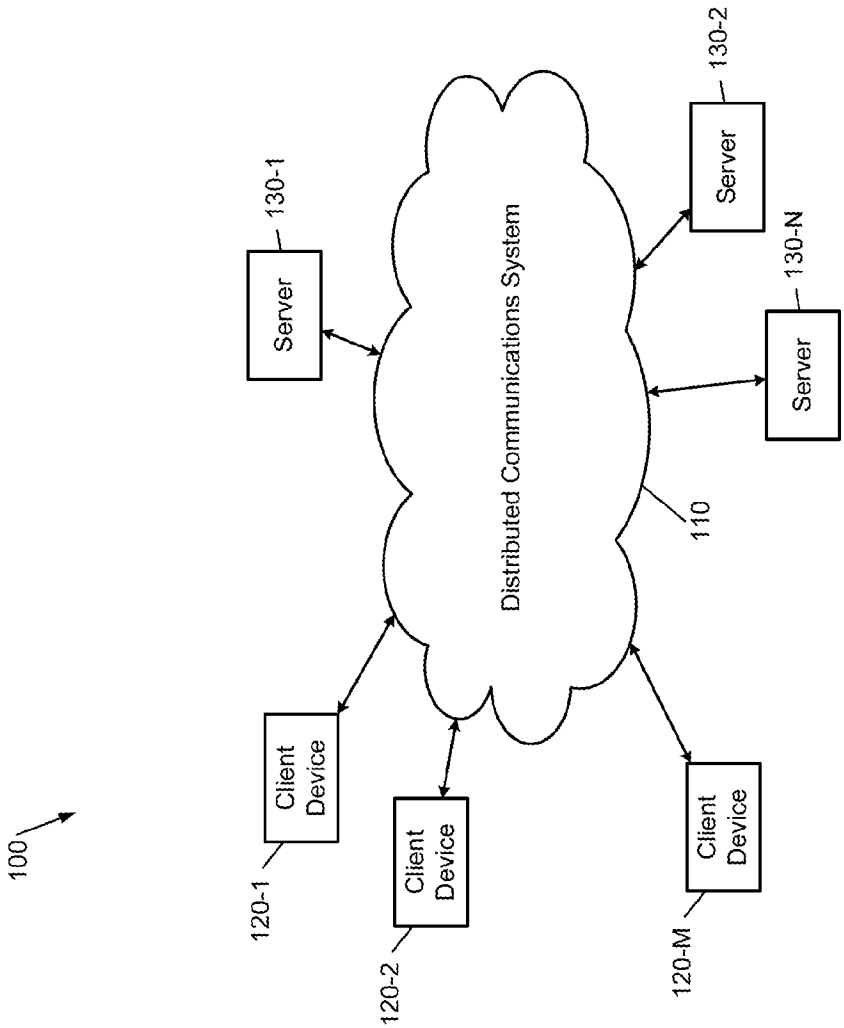
100
Distributed Communications System
110
Server
130-1
Server
130-2
Server
130-N
Client Device
120-1
Client Device
120-2
Client Device
120-M

FIG. 1

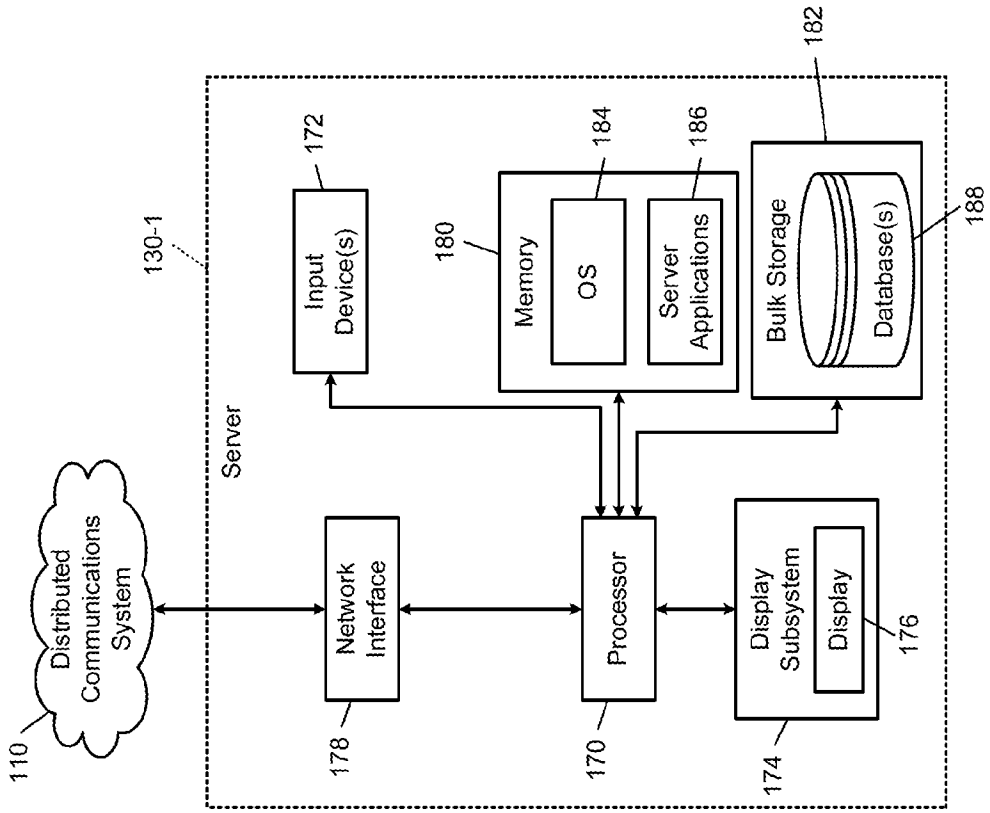
110
Distributed Communications System
130-1
Server
178
Network Interface
170
Processor
174
Display Subsystem
176
Display
172
Input Device(s)
180
Memory
184
OS
186
Server Applications
182
Bulk Storage
188
Database(s)

FIG. 3

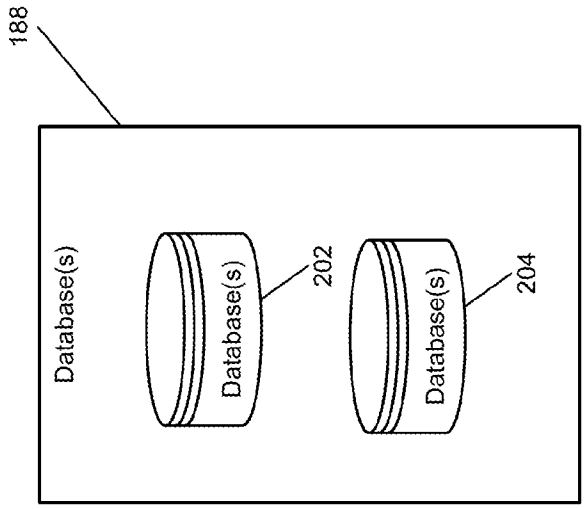
188
Database(s)
202
Database(s)
204
Database(s)

FIG. 4

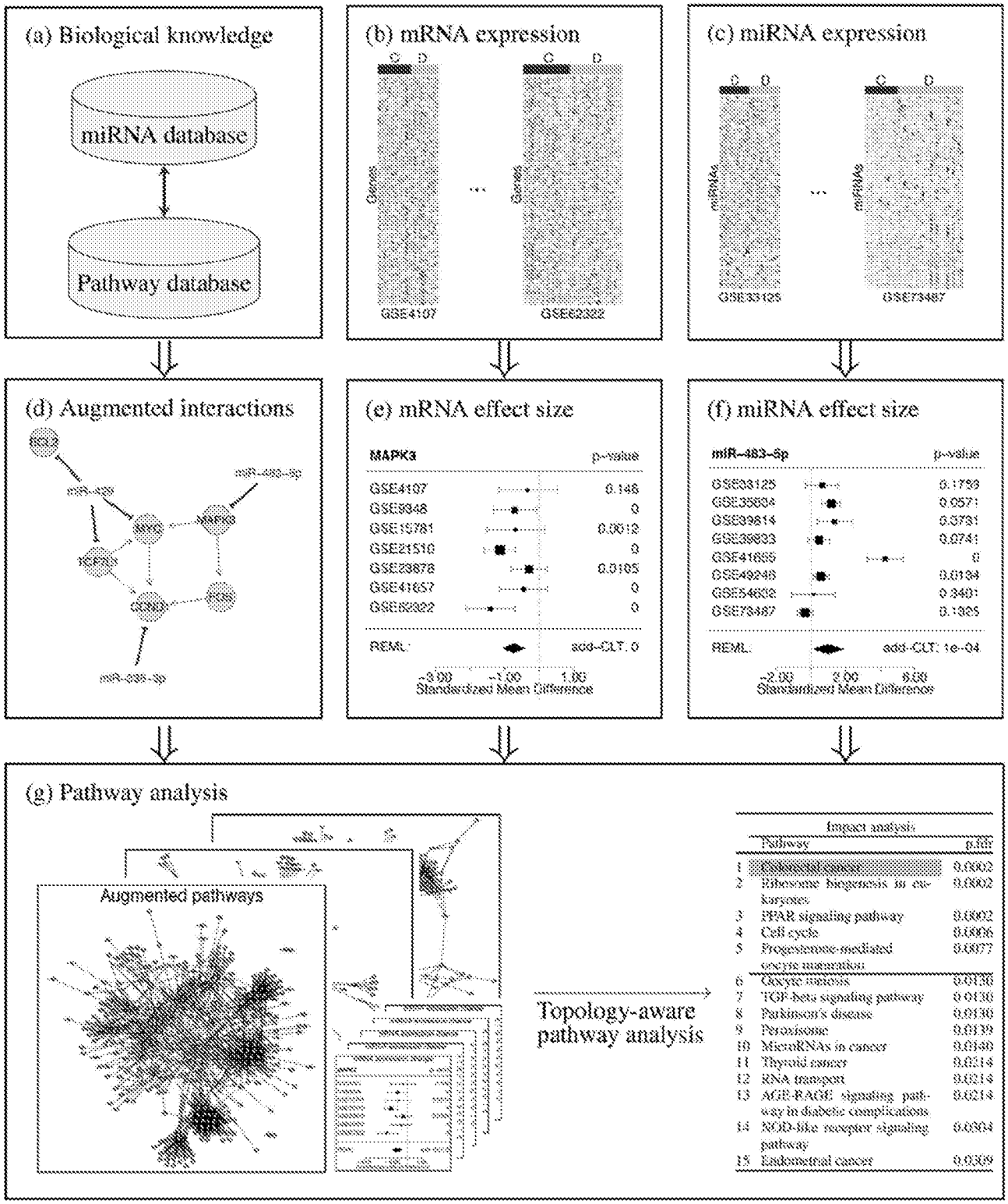
(a) Biological knowledge
miRNA database
Pathway database
(b) mRNA expression
C D
Genes
GSE4107
GSE62322
(c) miRNA expression
miRNAs
GSE33126
GSE73467
(d) Augmented interactions
(e) mRNA effect size
MAPK3
p-value
GSE4107 0.146
GSE9348 0
GSE15781 0.0012
GSE21510 0
GSE28878 0.0105
GSE41657 0
GSE62322 0
REML add-CLT: 0
Standardized Mean Difference
(f) miRNA effect size
miR-483-5p
p-value
GSE33126 0.1799
GSE35634 0.0571
GSE39614 0.3731
GSE39833 0.0741
GSE41655 0
GSE49249 0.0134
GSE54002 0.3401
GSE73467 0.1325
REML add-CLT: 1e-04
Standardized Mean Difference
(g) Pathway analysis
Augmented pathways
Topology-aware pathway analysis
Impact analysis
Pathway p.fdr
1 Cell cycle 0.0002
2 Ribosome biogenesis in eukaryotes 0.0002
3 PPAR signaling pathway 0.0002
4 Cell cycle 0.0006
5 Progesterone-mediated oocyte maturation 0.0077
6 Oocyte meiosis 0.0130
7 TGF-beta signaling pathway 0.0130
8 Parkinson's disease 0.0130
9 Peroxisome 0.0139
10 MicroRNAs in cancer 0.0140
11 Thyroid cancer 0.0214
12 RNA transport 0.0214
13 AGE-RAGE signaling pathway in diabetic complications 0.0214
14 NOD-like receptor signaling pathway 0.0304
15 Endometrial cancer 0.0309

FIG. 5

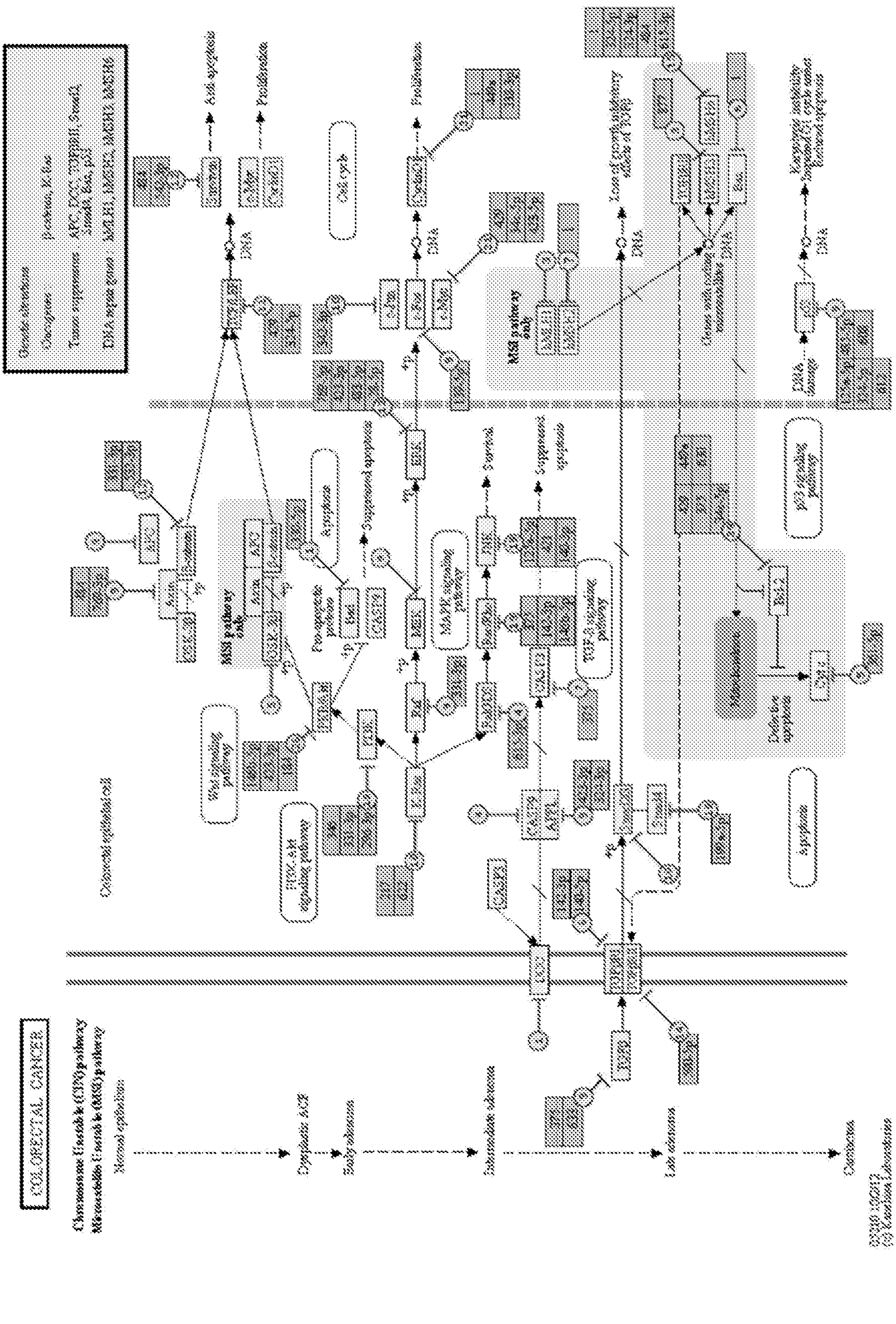

FIG. 6

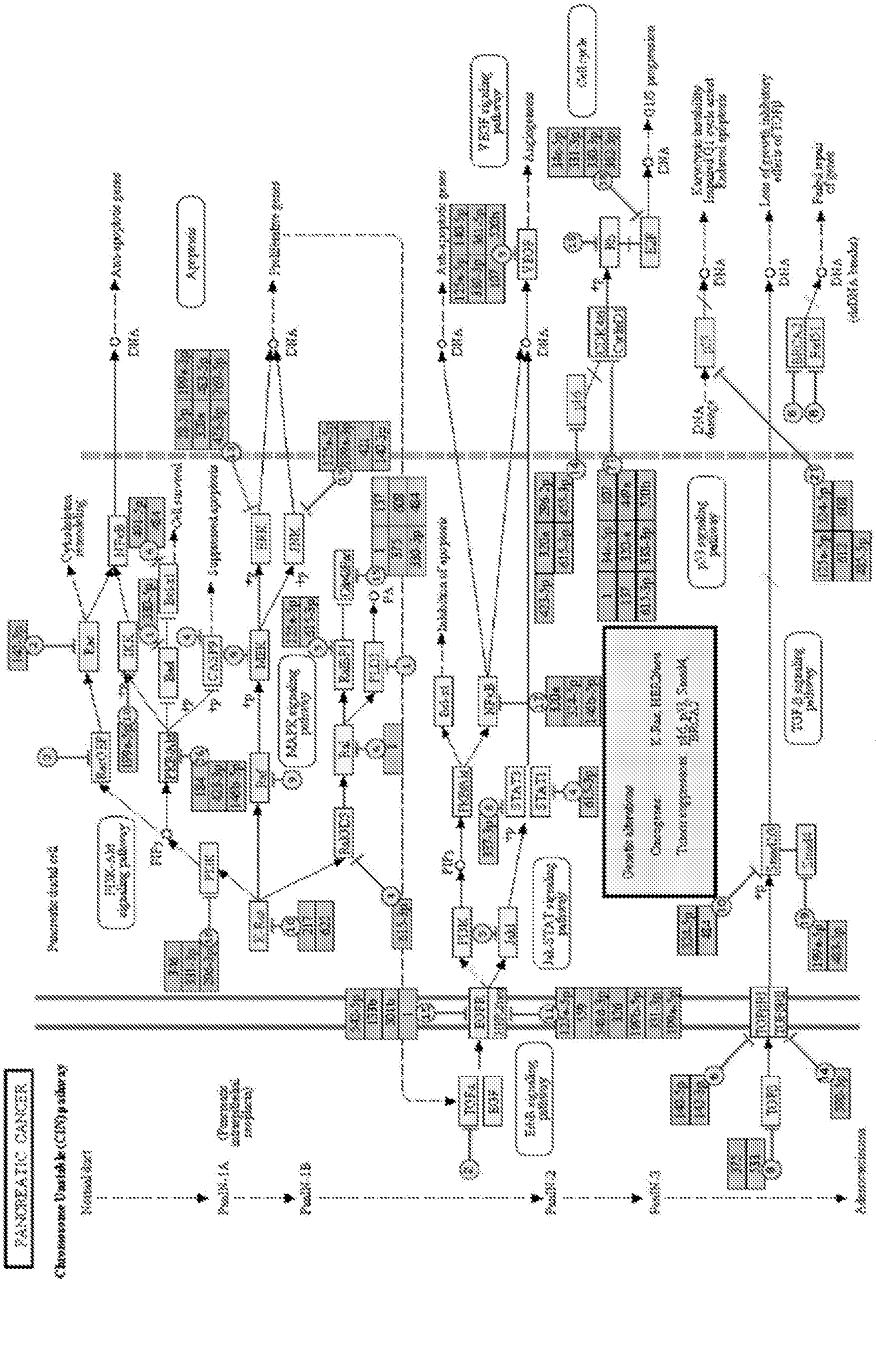

FIG. 7

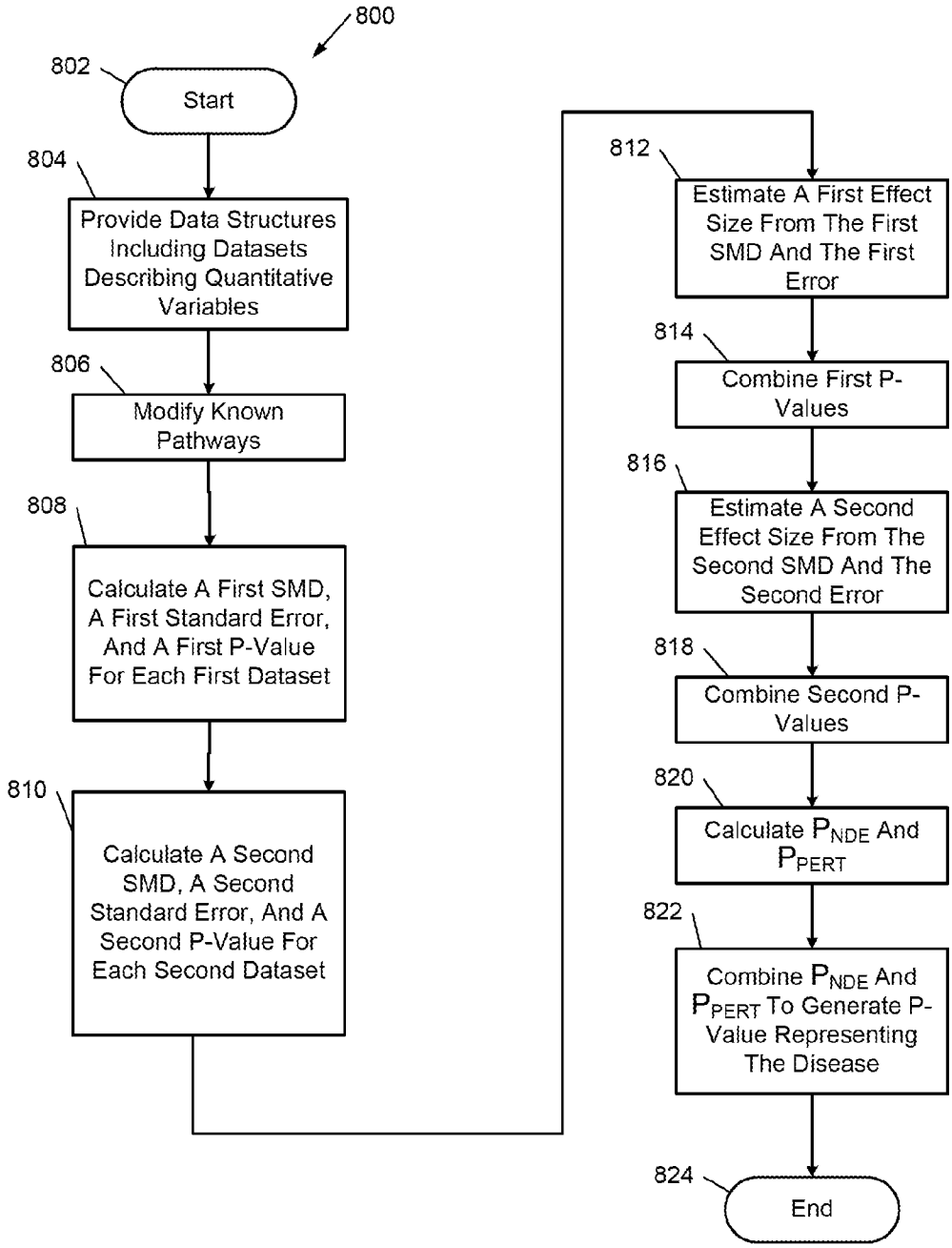
800
802
Start
804
Provide Data Structures Including Datasets Describing Quantitative Variables
806
Modify Known Pathways
808
Calculate A First SMD, A First Standard Error, And A First P-Value For Each First Dataset
810
Calculate A Second SMD, A Second Standard Error, And A Second P-Value For Each Second Dataset
812
Estimate A First Effect Size From The First SMD And The First Error
814
Combine First P-Values
816
Estimate A Second Effect Size From The Second SMD And The Second Error
818
Combine Second P-Values
820
Calculate P_{NDE} And P_{PERT}
822
Combine P_{NDE} And P_{PERT} To Generate P-Value Representing The Disease
824
End

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2017/031799

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/18
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Alexander Kaever ET AL: "Meta-Analysis of Pathway Enrichment: Combining Independent and Dependent Omics Data Sets", , 28 February 2014 (2014-02-28), XP055389972, Retrieved from the Internet: URL:http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089297 [retrieved on 2017-07-11] abstract; page 1 penultimate paragraph; page 2 2nd, 3rd and penultimate paragraph; page 3 2nd paragraph; page 4 last paragraph - page 5 1st paragraph; page 9 3rd paragraph	1-22

A	CA 2 851 280 A1 (BRIGHAM & WOMENS HOSPITAL [US]) 18 April 2013 (2013-04-18) page 115 - page 116	1-22

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 July 2017	25/07/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Bankwitz, Robert

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2017/031799

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
CA 2851280 A1	18-04-2013	AU 2012322788 A1	24-04-2014
		BR 112014008925 A2	13-06-2017
		CA 2851280 A1	18-04-2013
		CN 104011210 A	27-08-2014
		EP 2766482 A1	20-08-2014
		EP 3170899 A1	24-05-2017
		HK 1201294 A1	28-08-2015
		JP 2014534810 A	25-12-2014
		KR 20140074997 A	18-06-2014
		NZ 623459 A	27-05-2016
		US 2014235697 A1	21-08-2014
		WO 2013055865 A1	18-04-2013