TECHNICAL ADVANCE   **Open Access**

CrossMark

integRATE: a desirability-based data integration framework for the prioritization of candidate genes across heterogeneous omics and its application to preterm birth

Haley R. Eidem[1], Jacob L. Steenwyk[1], Jennifer H. Wisecaver[1,2], John A. Capra[1,3,4], Patrick Abbot[1] and Antonis Rokas[1,3,4*]

Abstract

Background: The integration of high-quality, genome-wide analyses offers a robust approach to elucidating genetic factors involved in complex human diseases. Even though several methods exist to integrate heterogeneous omics data, most biologists still manually select candidate genes by examining the intersection of lists of candidates stemming from analyses of different types of omics data that have been generated by imposing hard (strict) thresholds on quantitative variables, such as *P*-values and fold changes, increasing the chance of missing potentially important candidates.

Methods: To better facilitate the unbiased integration of heterogeneous omics data collected from diverse platforms and samples, we propose a desirability function framework for identifying candidate genes with strong evidence across data types as targets for follow-up functional analysis. Our approach is targeted towards disease systems with sparse, heterogeneous omics data, so we tested it on one such pathology: spontaneous preterm birth (sPTB).

Results: We developed the software integRATE, which uses desirability functions to rank genes both within and across studies, identifying well-supported candidate genes according to the cumulative weight of biological evidence rather than based on imposition of hard thresholds of key variables. Integrating 10 sPTB omics studies identified both genes in pathways previously suspected to be involved in sPTB as well as novel genes never before linked to this syndrome. integRATE is available as an R package on GitHub (https://github.com/haleyeidem/integRATE).

Conclusions: Desirability-based data integration is a solution most applicable in biological research areas where omics data is especially heterogeneous and sparse, allowing for the prioritization of candidate genes that can be used to inform more targeted downstream functional analyses.

Keywords: Prematurity, Integrative genomics, Complex disease, Candidate gene ranking, Venn diagram

Background

Biological processes underlying disease pathogenesis typically involve a complex, dynamic, and interconnected system of molecular and environmental factors [1]. Advances in high-throughput omics technologies have allowed for the collection of data corresponding to the genomic, transcriptomic, epigenomic, proteomic, and metabolomic elements that contribute to variation in these biological processes [2]. However, each of these omics approaches, when employed in isolation, can only capture variation within a single layer of a much more complicated biological system [3, 4]. For example, even though the thousands of single nucleotide polymorphisms (SNPs) that have been linked to complex diseases or traits via genome-wide association studies (GWAS) have greatly contributed to our understanding of complex disease, these SNPs may only be tagging the

* Correspondence: antonis.rokas@vanderbilt.edu
[1]Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA
[3]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
Full list of author information is available at the end of the article

causal genetic element(s), and we still lack in depth knowledge of the molecular mechanisms underlying the vast majority of these associations [5, 6]. Similarly, transcriptomics studies routinely identify hundreds to thousands of differentially expressed genes between diseased and healthy tissue samples, but disentangling the disease-causing changes in gene expression from its byproducts can be far more challenging [7]. Given the limitations of each omics approach and their focuses on different layers of the biological system, integration of different types of omics data to identify the key biological pathways involved in disease has emerged as a promising avenue for research [4].

One integrative study design is to obtain diverse types of omics data from the same tissue samples or patient cohorts. The resulting data can then be vertically integrated (Fig. 1a, top left) to identify candidate genes and pathways involved in complex disease. Alternatively, a single type of omics data can be collected from a variety of tissue samples or patient cohorts, facilitating their horizontal integration across many samples, which can substantially increase the experiment's power (Fig. 1a, top right). In both vertical and horizontal integration study designs, the availability of diverse types of omics data from the same samples enables the use of a variety of statistical integration approaches (Fig. 1a, bottom) [8]. For example, multi-staged integration uses multiple steps to first identify associations between different data types and then identify associations between data types and the phenotype of interest [9], whereas meta-dimensional integration combines data simultaneously based on concatenation, transformation, or model building [10].

Although multi-omics data sets generated using vertical and horizontal study designs are becoming increasingly common, such data sets are lacking for many complex diseases [11–15]. Often, heterogeneous omics data are collected study by study, for a limited set of tissue samples and across only one or two omics data types at a time (Fig. 1b, top). For each study, a long list of genes or genomic regions with associated data is produced and sorted based on effect size (e.g., fold change), significance (e.g., *P*-value), or some other criterion. Hard thresholds can then be imposed on *P*-values, for example, to bin the genes or genomic regions and identify significant candidates for further analysis; this type of approach can then be applied across multiple, heterogeneous omics studies.

Several problems exist with the imposition of hard thresholds, however. Including (or excluding) genes or genomic regions as candidates based on *P*-value, fold change, expression level, and/or odds ratio cutoffs introduces biases and removes information, especially when combining multiple cutoffs from several criteria [16–18]. These cutoffs can sometimes even be arbitrary, like selecting the top n or n% from each data set. Additionally,

statistical significance is not always equivalent to biological significance, meaning that non-statistically significant genes may still be involved in disease pathogenesis, or vice versa. Moreover, while selecting the top n genes might limit the scope of further functional analysis, the alternative approach of selecting all significant hits could mean that thousands of genes are identified as candidates. A final consideration in analyzing heterogeneous omics data is that we sometimes do not know any genes, pathways, or networks that have already been shown to be involved in complex disease. Some integration methods, especially those based on prediction (e.g., machine learning, network analysis), depend on the availability of such knowledge for algorithm training and cannot be performed in their absence [8, 9, 19–22].

Desirability functions provide a way to integrate heterogeneous omics data in systems where gold standards (i.e., genes known to be involved in the complex disease under investigation) are not yet known (Fig. 1b, bottom). Originally developed for industrial quality control, desirability functions have been successfully used in chemoinformatics to rank compounds for drug discovery and have been proposed as a way to integrate multiple selection criteria in functional genomics experiments [23–27]. In the context of integrating diverse but heterogeneous omics data, desirability functions allow for the ranking and prioritizing of candidate genes based on cumulative evidence across data types and their variables, rather than within-study separation of significant and non-significant genes based on single variables in single studies. For example, a 2015 study initially proposed the use of desirability functions to integrate multiple selection criteria for ranking, selecting, and prioritizing genes across heterogeneous biological analyses and demonstrated its use by analyzing a set of microarray-generated gene expression data [23].

To facilitate data integration in the presence of heterogeneous multi-omics data and when prior biological knowledge is limited, we propose a desirability-based framework to prioritize candidate genes for functional analysis. To facilitate application of our framework, we built a user-friendly software package called integRATE, which takes as input data sets from any omics experiment and generates a single desirability score based on all available information. This approach is targeted towards biological processes or diseases with particularly sparse or heterogeneous data, so we test integRATE on a set of 10 omics data sets related to spontaneous preterm birth (sPTB), a complex disease where heterogeneous multi-omics data are the only omics data currently available.

Design
Variable transformation
First, relevant studies need to be identified for integration; this selection can be based on any number of

Fig. 1 Selecting a data integration strategy depends on the structure of accessible multi-omics data. (**a**, left) If multiple types of omics data ('multi-omics') are available for the same cohort of patients, vertical integrative analysis can be performed to combine information across data types. This integration can be achieved using a variety of multi-staged and meta-dimensional statistical approaches that identify disease subtypes, regulatory networks, and driver genes. (**a**, right) If the opposite is true and a specific type of omics data is available across a number of different patient cohorts, horizontal meta-analysis can be performed to increase statistical power and identify disease-associated perturbations. **b** In some cases, however, experimental data are only available for different omics data types from different cohorts of patients and neither vertical nor horizontal data integration can be performed. In these situations, integration relies on mapping data to common units (e.g., genes or pathways) and then either integrating transformed data or simply overlapping candidate sets. The software approach presented here (integRATE) utilizes desirability functions to transform and integrate heterogeneous data allowing for the prioritization of candidate genes for functional analysis

characteristics including tissue(s) sampled, disease subtype, or experimental designs (Fig. 2, step 1). The data in each of these studies (e.g., gene expression data, proteomic data, GWAS data, etc.) are typically specific to or can be mapped to individual genetic elements (e.g., genes) in the genome. Furthermore, each study's data contain genetic element-specific values for many different variables (e.g., P-value, odds ratio, fold change, etc.). Then desirability functions are fit to the observations for each variable within a study (e.g., P-value, odds ratio, fold change, etc.) according to whether low values are most desirable (d_{low}, e.g., P-value), high values are most desirable (d_{high}, e.g., odds ratio), or extreme values are most desirable ($d_{extreme}$, e.g., fold change) (Fig. 2, step 2).

The desirability score for each genetic element can be calculated by applying one of the following equations to a given variable:

$$d_{low} = \begin{cases} 0 & Y > B \\ \left[\dfrac{Y-B}{A-B}\right]^s & A \leq Y \leq B \\ 1 & Y < A \end{cases} \tag{1}$$

$$d_{high} = \begin{cases} 0 & Y < A \\ \left[\dfrac{Y-A}{B-A}\right]^s & A \leq Y \leq B \\ 1 & Y > B \end{cases} \tag{2}$$
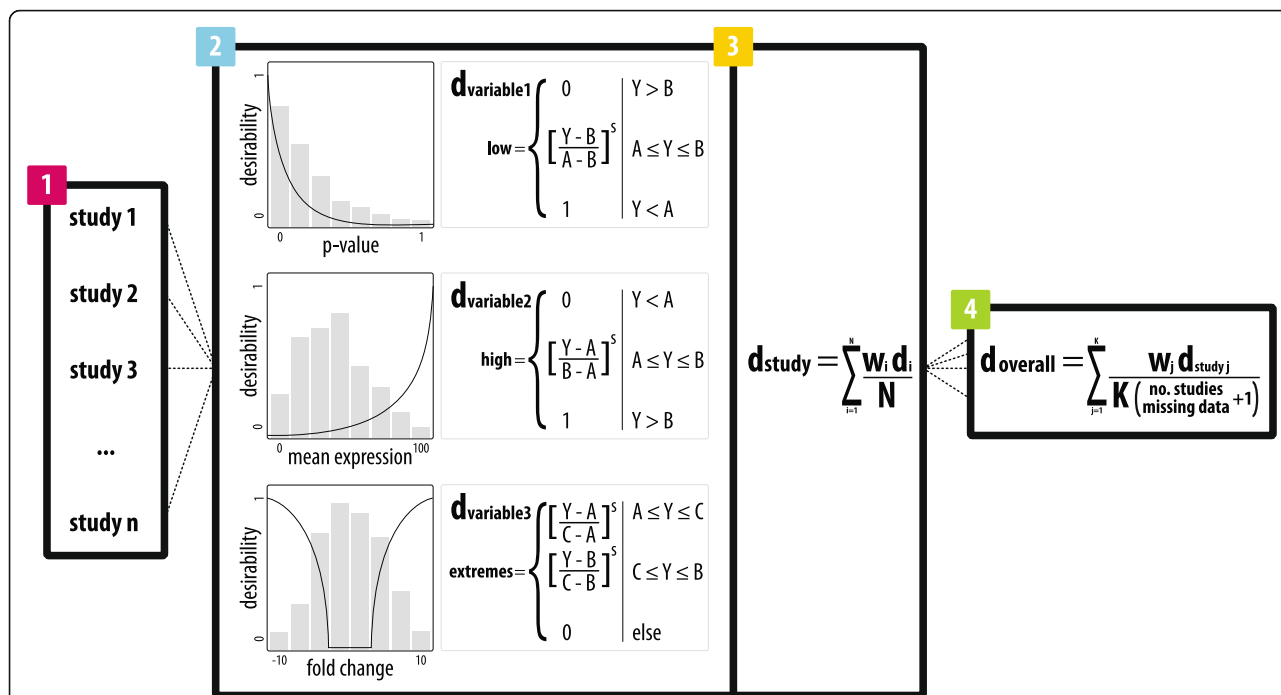
Fig. 2 integRATE relies on three main steps to identify studies, integrate data, and rank candidate genes. (1) Relevant studies must first be identified for integration based on any number of features including, but not limited to: phenotype, experimental design, and data availability. (2) Data corresponding to all variables in each study are then transformed according to the appropriate desirability function. In this step, the user assigns a function based on whether low values are most desirable (d_{low}), high values are most desirable (d_{high}), or extreme values are most desirable ($d_{extreme}$) and can customize the shape of the function with other variables like cut points (A, B, C), scales (s), and weights (w) to better reflect the data distributions or to align with user opinion regarding data quality and relevance. (3) These variable-based scores are integrated (d_{study}) with a straightforward arithmetic mean (where weights can also be applied) to produce a single desirability score for each gene in each study containing information from all variables simultaneously. (4) Finally, study-based desirability scores are integrated to produce a single desirability score for each gene ($d_{overall}$) that includes information from all variables in all studies and reflects its cumulative weight of evidence from each data set identified in step 1. These scores are normalized by the number of studies containing data for each gene and can be used to rank and prioritize candidate genes for follow up computational and, most importantly, functional analyses

$$d_{extreme} = \begin{cases} \left[\dfrac{Y-A}{C-A}\right]^s & A \le Y \le C \\ \left[\dfrac{Y-B}{C-B}\right]^s & C \le Y \le B \\ 0 & else \end{cases} \quad (3)$$

In these equations, Y is the variable value and s is the scale coefficient affecting the function's rate of change that can be customized according to user preference. Alternatively, the equations could be used without any scaling by setting the scale coefficient to 1. For d_{low} and d_{high}, A is the low cut point and B is the high cut point where the function changes. For $d_{extreme}$, A is the low cut point, C is the intermediate cut point, and B is the high cut point where the function changes. The user can customize these cut points based on numerical values (e.g., *P*-value < 0.05) or percentile values (e.g., top 10%). The resulting values, ranging from 0 to 1 (or the minimum and maximum values specified) are transformed desirability scores based on information from each variable.

Variable integration

Next, desirability scores for each of the N variables within a study (e.g., *P*-value, odds ratio, fold change, etc.) are combined using an arithmetic mean so that genetic elements (e.g., genes) with desirability scores of zero for any given variable remain in the analysis (Fig. 2, step 3). Desirability for genetic elements within a study can be calculated by:

$$d_{study} = \sum_{i=1}^{N} \frac{w_i d_i}{N} \quad (4)$$

In this equation, w_i is the weight parameter (assigned to each variable), d_i is desirability score for each genetic element based on the values of each variable derived from Eqs. (1), (2) or (3), and N is the total number of transformed variables. This step produces a single desirability score (d_{study}) for each genetic element in the study containing information from all transformed variables. Here, the user is also able to include variable

Eidem *et al. BMC Medical Genomics*   (2018) 11:107

Page 5 of 13

weights (w_i) when integrating their desirability scores, which can be useful in cases where certain variables are considered more informative or accurate than others.

Study integration

Finally, the d_{study} values can be integrated using the arithmetic mean to produce a single desirability score ($d_{overall}$) for each genetic element, representing its desirability as a candidate according to the weight of evidence from all variables in all K studies that were integrated (Fig. 2, step 4). The overall score used to prioritize candidates can be calculated by:

$$d_{overall} = \sum_{j=1}^{K} \frac{w_j d_{study\ j}}{K(no.studies\ missing\ data + 1)} \qquad (5)$$

In this equation, w_j is the weight parameter (assigned to each study), $d_{study\ j}$ is the desirability score for each study, and K is the total number of studies integrated. Importantly, the overall desirability score $d_{overall}$ is normalized by the number of studies missing data for each genetic element to account for the number of values contributing to each overall desirability score. This normalization factor can be used to calculate a soft cut-off for the most desirable candidates that is equivalent or higher than the desirability score that would be achieved by a genetic element with a perfect desirability score of 1 in a single study but missing from all other studies. We call genetic elements achieving desirability scores equal to or above this cutoff 'desirable.'

Software

The methodology described above is implemented in our software, integRATE, available on GitHub as an R package (https://github.com/haleyeidem/integRATE). Although we focus on using desirability functions to integrate heterogeneous omics data corresponding to complex human diseases, integRATE can be applied to data sets from any phenotype, species, and data type (provided that the units can all be mapped to a common set of elements, such as genes). Functionality is provided for the application of customizable desirability functions as well as data visualization.

Implementation

One human complex genetic disease where the omics data available are heterogeneous is preterm birth (PTB). Defined as birth before 37 weeks of completed gestation, PTB is the leading cause of newborn death worldwide [28]. Although 30% of preterm births are medically indicated due to complications including preeclampsia (PE) or intrauterine growth restriction (IUGR), the remaining 70% occur spontaneously either due to the preterm

premature rupture of membranes (PPROM) or idiopathically (sPTB). Further complicating factors are that multiple maternal and fetal tissues are involved (e.g., placenta, fetal membranes, umbilical cord, myometrium, decidua, etc.) as well as multiple genomes (maternal, paternal, and fetal) [29]. Evidence from family, twin, and case-control studies suggests that genetics plays a role in determining birth timing and a recent GWAS identified a handful of genes linked to prematurity [30]. Nevertheless, the pathogenesis of PTB and its many subtypes remains poorly understood [31–33].

The publicly available data for sPTB consist of several different independently conducted omics analyses that would be challenging to analyze with statistical approaches developed for vertical and horizontal integration [30, 34, 35]. Although these omics data have been analyzed in isolation, integration of their information using the desirability-based platform implemented in integRATE may provide unique insights into the complex mechanisms involved in regulating birth timing and, thus, allow for the identification and prioritization of novel candidate genes for further functional and targeted analyses.

Study identification

Studies were initially identified based on the PubMed searches (up to 10/19/2017) using combinations of terms, including "Pregnancy", "Humans", "Preterm birth", "Placenta", "Decidua", "Myometrium", "Cervix Uteri", "Extra-embryonic Membranes", "Blood", "Plasma" and "Umbilical Cord". Studies that reported conducting a genome-wide omics analysis of sPTB from a preliminary scan of the abstract were downloaded for full-text assessment. Furthermore, a thorough investigation was conducted of their associated reference lists to identify studies not captured via PubMed. Additionally, each study had to meet the following inclusion criteria:

1) Experimental group consisted of sPTB cases only and was not confounded by other pregnancy phenotypes (e.g., preeclampsia),
2) Analysis was genome-wide and not targeted to any specific subset of genes or pathways, and
3) Full data set was publicly available (not just top n%).

We identified 54 studies through the first phase of our literature search, but only 10 data sets that met all inclusion criteria. All excluded studies are listed in Additional file 1 with reasons for exclusion and the 10 data sets used in our pilot analysis are outlined in Table 1 [30, 34–46].

Data transformation

Each of the 10 data sets was mapped to a gene-based format. This step was necessary because integRATE

Eidem *et al. BMC Medical Genomics*   (2018) 11:107

Page 6 of 13

Table 1 The 10 sPTB omics data sets identified for desirability-based integration

First author	Year	Experiment	Control	Tissue	omics Type
Zhang	2017	sPTB	term	maternal blood	genomics (GWAS)
Ackerman	2015	sPTB	term	placenta	transcriptomics (RNA-seq)
Heng	2014	sPTB	term	maternal blood	transcriptomics (microarray)
Chim	2012	sPTB	term	maternal blood	transcriptomics (microarray)
Mayor-Lynn	2011	sPTB	term	placenta	transcriptomics (microarray)
de Goede	2017	sPTB	term	cord blood	epigenomics (microarray)
Fernando	2015	sPTB	term	cord blood	epigenomics (microarray)
Parets	2015	sPTB	term	maternal blood	epigenomics (microarray)
Cruickshank	2013	sPTB	term	fetal blood	epigenomics (microarray)
Heng	2015	sPTB	term	maternal blood	proteomics (mass spec)

applies desirability functions both within and across studies and, in order for that integration to be possible, the genetic elements of each study have to match.

Genomics
SNP-based data containing *P*-values and effect sizes were mapped to genes with MAGMA, as outlined in the Zhang et al. supplementary methods (http://ctg.cncr.nl/software/magma) [39, 47, 48].

Transcriptomics
Gene expression data from microarray experiments were accessed via GEO (https://www.ncbi.nlm.nih.gov/geo/) and re-analyzed using the GEO2R plugin (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html) [40–43]. Raw RNA-seq data from Ackerman et al. were analyzed in-house with custom scripts [34].

Epigenomics
Methylation data were accessed via GEO (https://www.ncbi.nlm.nih.gov/geo/) and re-analyzed using the GEO2R plugin (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html) [36, 37, 44–46].

Proteomics
Protein expression data were downloaded from supplementary files associated with each publication and the protein IDs were mapped to genes using Ensemble's Bio-Mart tool (https://www.ensembl.org/info/data/biomart/index.html) [35, 38].

Application of integRATE
After mapping results from all 10 omics studies to genes, we used integRATE to calculate desirabilities for all genes across all variables within studies. We ran four different sPTB analyses:

1) In the first analysis (**iR-none**), we ran integRATE with no added customizations (e.g., no cut points, no scales (i.e., scale coefficient = 1), no minimum or maximum desirabilities, etc.) (Figs. 3, 4 and 5, Additional file 2).

2) In the second analysis (**iR-num**), we ran integRATE using numerical cut points ($P = 0.0001$, 0.1 and fold change = 1.5, 0.5, − 0.5, − 1.5) and no scales (Additional files 3, 4, 5, 6).

3) In the third analysis (**iR-per**), we ran integRATE using percentile cut points ($P = 5$, 95%, and fold change = 5, 50, 95%) and no scales (Additional files 7, 8, 9, 10).

4) In the fourth analysis (**HardThresh**), we considered statistically significant genes from each study to represent the results that would have been obtained if the typical approach based on hard thresholds and intersection of significant genes across studies outlined earlier was applied (Additional files 11, 12). All genes with adjusted *P*-values < 0.1 or unadjusted *P*-values < 0.05 were deemed significant in each study and intersected to compare with the results from integRATE [49].

To test whether the analyses described above produced results different from what might occur at random, we performed a permutation test shuffling desirabilities for all 26,868 genes 1000 times.

Results
In total, our sPTB analyses integrated gene-based results from 10 omics studies (1 genomics, 4 transcriptomics, 4 epigenomics, and 1 proteomics; Table 1) and included data sets ranging from 422 genes [35] to 20,841 genes [42]. The null distribution generated by our random permutation test had mean desirability range from 0.056 to 0.062, with an average of 0.059 (95% CI [0.058, 0.061]) (Fig. 3).
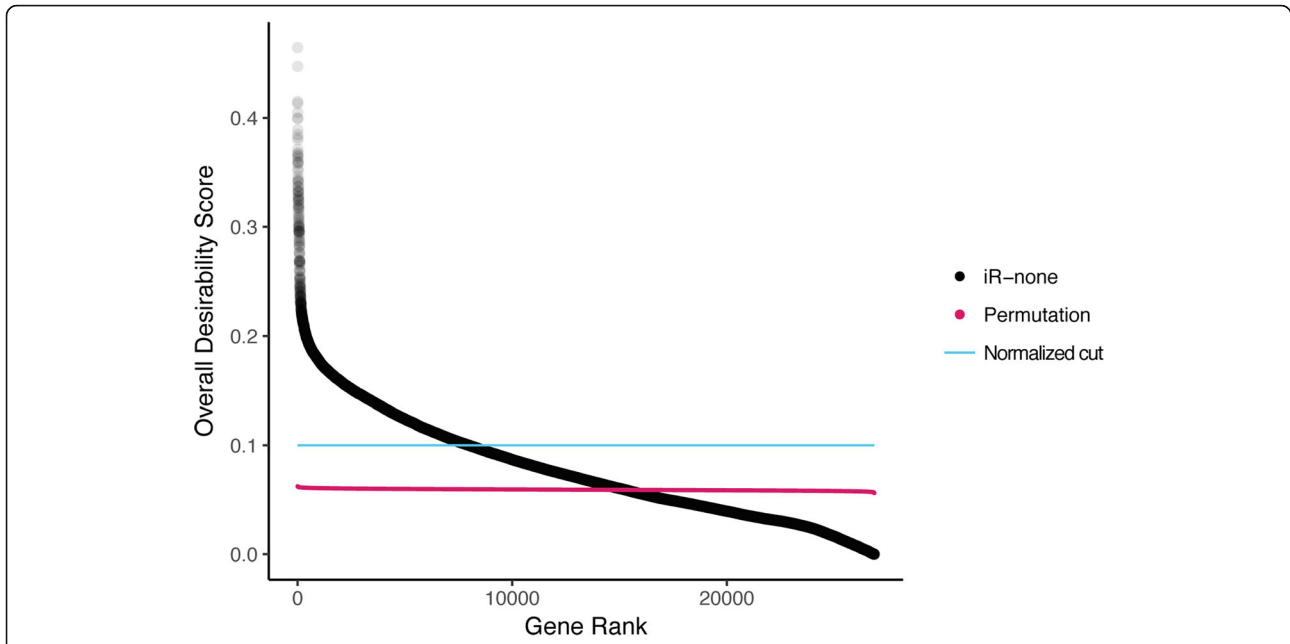
Fig. 3 After integration, 7977/26,868 genes were identified as highly desirable. All genes in the iR-none analysis were sorted from most desirable (rank = 1) to least desirable (rank = 26,868) and plotted according to their overall desirability scores, ranging from 8.04E-16 to 0.46. Because this analysis included 10 omics studies, the normalized lower bound for our set of 'desirable' candidate genes is 0.1 (blue line) and 7977 genes achieved scores greater than or equal to that value. Furthermore, the results of our permutation test are plotted in pink, with a mean of 0.059 (95% CI [0.058, 0.061]). All desirability scores for the entire data set are available in Additional file 2 (and in Additional files 3 and 7 for iR-num and iR-per, respectively)
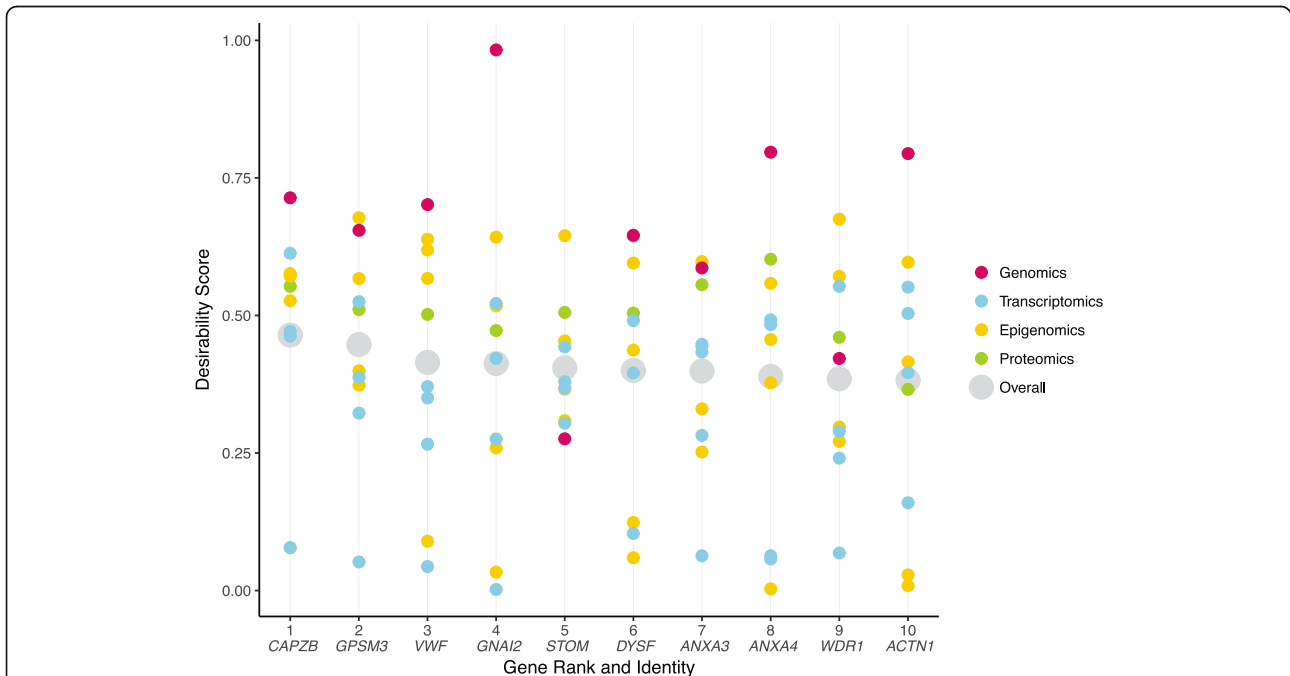
Fig. 4 The top 10 most desirable genes have a wide range of desirabilities across data types. The top 10 genes from our iR-none analysis have overall desirabilities ranging from 0.38 (*ACTN1*) to 0.46 (*CAPZB*), but the d_{study} values range, even when organized by data type. Some genes, like *STOM*, appear to be highly ranked not because of any extremely high d_{study} value, but rather due to a lack of low d_{study} values in any data type. In other words, this gene is likely not identified as particularly important in any individual study but shows a consensus of relatively strong evidence across all 10 studies. Contrastingly, other genes, like *CAPZB*, appear to be highly ranked due to one very high desirability score in a single data type (GWAS) that overpowers underwhelming evidence in other studies
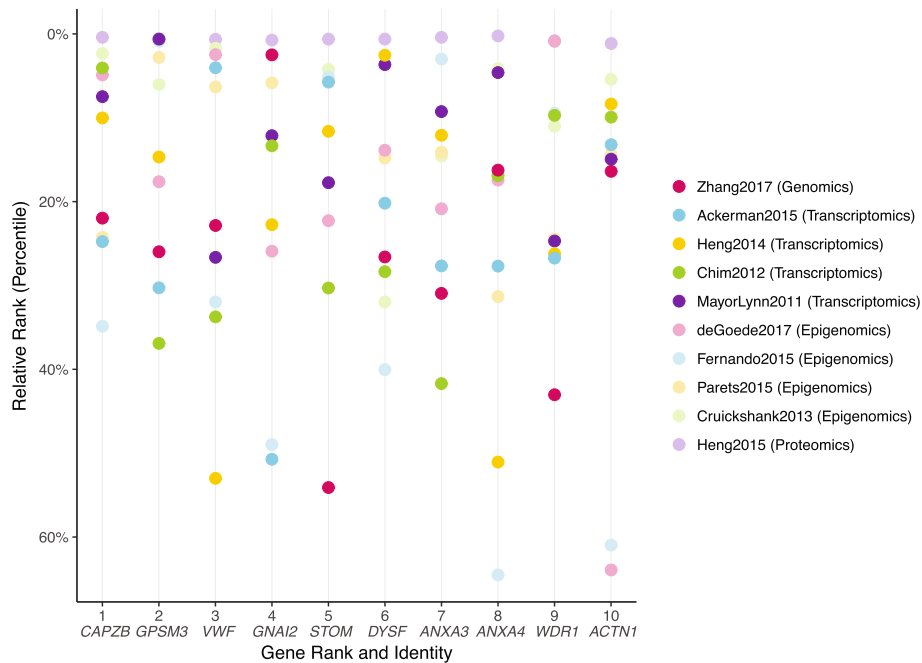
Eidem *et al. BMC Medical Genomics* (2018) 11:107

Page 8 of 13

Fig. 5 The top 10 most desirable genes show a large discrepancy in their percentile ranks across studies. After ranking the genes in each study by desirability (using the iR-none analysis) and calculating their percentiles based on the number of unique ranks, the top 10 most desirable genes appear to show even greater variability in relative ranking across not just data type, but individual studies. All 10 genes are in the top 25% of the (smaller) proteomics study, but their relative rankings vary significantly in all other studies. Furthermore, while none of the genes are in the top 25% of the GWAS study [30], other studies, like one of the transcriptomics analyses [43], show a large range in relative rankings, with certain highly desirable genes ranked very high and others ranked very low

iR-none

First, the software was run without any added cuts, weights, or scales, resulting in a list of 26,868 genes with data from one or more of the 10 omics studies (Additional file 2). Normalized desirabilities for these 26,868 genes ranged from 8.04E-16 to 0.46 (mean = 0.08 ± 0.05) (Fig. 3). Furthermore, 7977 genes (29.7%) had desirabilities ≥0.1 corresponding to values equal to or higher than what would be achieved if a given gene achieved maximal desirability in one study but was absent from all others. These top 7977 genes were enriched for 70 unique GO-Slim Biological Process categories, including pathways involved in metabolic processes, immunity, and signal transduction (Additional file 13) [50]. Additionally, 15,285/26,868 (56.9%) genes achieved desirabilities greater than the permutation mean of 0.059. The top 10 genes (Figs. 4 and 5) had desirabilities ranging from 0.46 (*CAPZB*) to 0.38 (*ACTN1*) and were all represented in each of the 10 omics data sets analyzed. This analysis applied inte-gRATE without cut points, allowing for a straightforward, linear transformation of data across all variables and studies.

iR-num

We next applied cut points based on numerical values (Additional file 3). *P*-values such that values smaller than

0.0001 received the maximum desirability score of 1 and values larger than 0.1 received the minimum desirability score of 0. All *P*-values between 0.0001 and 0.1 were transformed according to the d_{low} function. For $d_{extreme}$ functions, 4 cut points were assigned and we chose commonly used values of 0.5 and 1.5 (or their equivalents if the values were log transformed). Therefore, fold changes below − 1.5 or above − 1.5 (or below log2(1/3) or above log2(3)) received the maximum desirability score of 1 and fold changes between − 0.5 and 0.5 (or between log2(1/1.5) and log2(1.5)) received the minimum desirability score of 0. Intermediate values were transformed according to the $d_{extreme}$ function. This approach mirrors what was applied in a previous implementation of the desirability framework for omics data, and takes into account prior knowledge of typical *P*-value and fold change distributions [23]. While the top most desirable genes in iR-num appeared to be better candidates in each individual study (Additional file 6), using these cut points corresponding to standard significant *P*-value and fold change cut offs greatly reduced the number of desirable genes identified (Additional file 3). Specifically, only 1386/26,868 (5.1%) genes achieved desirabilities greater than the permutation mean of 0.059 and the top 10 most desirable genes were analyzed by only 4 or 5 studies instead of all 10 (Additional file 5).

Eidem *et al. BMC Medical Genomics*   (2018) 11:107

Page 9 of 13

iR-per

Finally, we applied cut points based on percentiles (Additional file 7). *P*-values were cut such that those in the top 5% received the maximum desirability score of 1 and those in the bottom 5% received the minimum desirability score of 0, with all values in between transformed according to the d_{low} function. Fold changes were cut such that those in the top 5% and bottom 5% received the maximum desirability score of 1 and those in the middle 50% received the minimum desirability score of 0, with all other values transformed according to the $d_{extreme}$ function. In this analysis, 16,604/26,868 (61.8%) genes achieved desirabilities greater than the permutation mean of 0.059.

HardThresh

For comparison, we also manually selected candidate genes by imposing a hard threshold on *P*-value (*P*-value < 0.05 if unadjusted and *P*-value < 0.1 if adjusted) (Additional file 11). After binning data into 'significant' gene lists, we intersected these lists to pull out genes that would have been identified simply by selecting the intersection of all significant genes. Although 18,727 genes were considered 'significant' in at least one study, no genes were identified as significant in all 10 studies. The top candidate gene (*KIAA0040*) was significant in 6/10 studies and 15 other were identified in 5/10 studies (Additional file 12). Interestingly, none of these 16 genes appear in the top 10 of our most desirable candidates after integration and, even more generally, none are specifically discussed in any of the studies, either.

Using integRATE to identify the most desirable sPTB genes

In our sPTB pilot analyses, members of the annexin family (*ANXA3*, *ANXA4* and *ANXA9*) appear in the top 10 most desirable candidate gene sets regardless of analysis approach (e.g., without cut points as well as with numerical *and* percentile cut points). This family is involved in calcium-dependent phospholipid binding and membrane-related exocytotic and endocytotic events, including endosome aggregation mediation (*ANXA6*). In a previous proteomic analysis, *ANXA3* was found to be differentially expressed in cervicovaginal fluid 26–30 days before the eventual onset of sPTB as compared to before healthy, term deliveries [51]. Furthermore, members of the annexin family are known to be involved in coagulation (*ANXA3*, *ANXA4*). Coagulation has been previously suggested to be involved in PTB and, even though the mechanism of such involvement is still a mystery, it is interesting that several genes involved in coagulation or blood disorders appear in our top candidate lists [52]. In addition to *ANXA3* and *ANXA4*, *VWF* (or Von Willebrand Factor) is a gene encoding a glycoprotein involved in coagulation that has been found to be expressed significantly more in preterm infant serum

as compared to term [53, 54]. Finally, another highly desirable candidate, *STOM*, encodes an integral membrane protein that localizes to red blood cells, the loss of which has been linked to anemia [55].

In addition to coagulation, another biological process represented across our results is actin regulation and muscle activity. The most notable gene associated with this biological process is *CAPZB*, which encodes part of an actin binding protein that regulates actin filament dynamics and stabilization and is present in the top 10 most desirable candidate gene list in all three analyses. Although *CAPZB* has never been linked to sPTB or other pregnancy pathologies, its role in muscle function could be linked to myometrial and uterine contractions that, when they occur prematurely, might be directly involved in the development of sPTB [56, 57]. Another one of our top candidates, *ACTN1*, is also involved in actin regulation and, even more interestingly, has also been linked to blood and bleeding disorders [58, 59]. Finally, several other highly desirable genes identified in one or more of our integrative analyses include *GPSM3*, *WDR1*, and *DYSF*, are all involved in the development and regulation of muscle or in the pathogenesis of muscle-related diseases [60–62].

Even outside the top 10 most desirable genes across our integrative analyses, we found genes both previously identified as being involved in pregnancy or sPTB pathology as well as involved in pathways potentially relevant to sPTB (Additional file 2). For example, one gene falling just outside the top 10 most desirable candidates in all analyses is *MMP9*, a matrix metalloproteinase. Interestingly, *MMP9* has been linked not only to sPTB, but also to PPROM and PE across a number of fetal and maternal tissues and at a variety of time points during pregnancy [63–67]. *MMP9* gene expression has been observed as significantly higher during preterm labor than during term labor in maternal serum, placenta, and fetal membranes [68–70]. Even in the first trimester, levels of *MMP9* in maternal serum were higher in PE cases than in healthy controls, suggesting that increased *MMP9* protein expression is linked to the underlying inflammatory processes governing PE pathogenesis [66]. Finally, fetal plasma *MMP9* concentration has been found to be significantly higher in fetuses with PPROM than in early and term deliveries with intact membranes, implicating *MMP9* in the membrane rupture mechanism controlling early delivery due to membrane rupture [67]. We see similar evidence of *MMP9* as a desirable sPTB candidate maintained across omics and tissue types in our integRATE analyses, raising the hypothesis that its role in inflammation and extracellular matrix organization relates to sPTB even in the absence of PPROM or PE.

Discussion

By using desirability functions to rank genes within studies and combine results across studies, integRATE allows

Eidem *et al. BMC Medical Genomics*   (2018) 11:107

Page 10 of 13

for the identification of candidate genes supported across experimental conditions and omics data types. This is especially important when heterogeneous sets of omics data, like those available for sPTB, where the statistical approaches developed for vertical or horizontal integration are challenging to apply. We have shown that integRATE can map any omics data to a common [0, 1] scale for linear integration and produce a list of the most desirable candidates according to their weight of evidence across available studies. These candidates then become promising targets for follow-up functional testing depending on where in the data their desirability signals come from. Analysis of 10 heterogeneous omics data sets on sPTB showed that the gene candidates identified using desirability functions appear to be much more broadly supported than those identified by the intersection of all significant genes across all studies and contain both genes that have been previously associated with sPTB as well as novel ones (Figs. 4 and 5, Additional file 12).

integRATE identifies both known and novel candidate genes associated with a complex disease, including ones that are not among the top candidates in any single omics study but are consistently (i.e., across studies) recovered as significantly (or nearly significantly) associated. For example, genes that are significantly differentially expressed at an intermediate to high level across *many* studies will have high desirability scores. Furthermore, integRATE can identify such genes across omics types, tissues, patient groups, and any other variable condition. Although integRATE allows for this kind of synergistic, desirability-based analysis, it is important to note that integRATE is not a statistical tool nor is it intended to be the end point of any analysis. Rather, it is a straightforward framework for the identification of well-supported candidate genes in any phenotype where true multi-omics data are unavailable and can also serve as a springboard for future functional analysis, an essential next-step in testing whether the candidates are actually involved in the biology of the disease or phenotype at hand.

In our analyses, the genomics data set was typically the one with the highest desirability scores for each of the top 10 genes (Fig. 4) and the proteomics data set was typically the one in which the relative rank of the top 10 genes was the highest (Fig. 5). Both of these trends may appear surprising considering that our analyses contained just one genomics and one proteomics data set compared to four transcriptomics and four epigenomics ones (Table 1). There are three reasons for these two trends. First, there is substantial heterogeneity among the top genes identified by the four transcriptomic studies (see also [71]), as well as among the top genes identified by the four epigenomic studies. As a consequence, there is no common signature of the four

transcriptomic studies or the four epigenomic studies (see Fig. 4). Second, there are many more genes with high desirability scores in the genomics data set than in the other nine data sets (Additional file 1). However, we note that the ranking of the top 10 genes is not driven by the genomics data set; as we discuss below (see last paragraph of the discussion section), only one of the top 10 genes (*EBF1*) is among the candidate genes identified to be significantly associated with preterm birth and gestation length in the genomics data set [47]. Third, the number of differentially expressed proteins (mapped to genes) in the proteomics data set, and as a consequence the number of genes with desirability scores in this data set, was substantially lower than that of all other studies (and included hundreds of genes vs tens of thousands of genes). As a result, the percentile rank of the top 10 genes for the proteomics data set (Fig. 5) was much higher than their percentile rank in other data sets. However, as shown in Fig. 4, the desirability scores of the top 10 genes in the proteomics data set were typically neither very high nor very low, and did not appear to exert a disproportionate influence on the ranking of our top 10 genes.

Importantly, there is no single principled strategy for the selection of cut points. In our sPTB analyses (iR-none, iR-num, and iR-per), we observed that the imposition of cut points corresponding to generally agreed upon values (e.g., *P*-value < 0.0001) has the potential to greatly affect the resulting gene prioritization. On this basis, we propose that desirability functions are best used to integrate highly heterogeneous omics data without imposed numerical cut points for *P*-values, fold changes, and other variables. Implemented this way, one can maximize the information from the analysis of each omics data set used in prioritizing candidate genes. But users may also have reasons to want to put more weight on data sets that are of higher quality or on data types that may be more informative. In such instances, the weight parameter can be used to reflect study quality instead of imposing cut points (e.g., studies that fail to achieve *P*-values as low as others in the integrative analysis can be weighted less to reflect potentially lower experimental quality).

A recent GWAS analysis, the largest of its kind across pregnancy research, identified several candidate genes with SNPs linked to PTB [47]. This study linked *EBF1*, *EEFSEC*, and *AGTR2* to preterm birth and *EBF1*, *EEF-SEC*, *AGTR2*, and *WNT4* to gestational duration (with *ADCY5* and *RAP2C* linked suggestively). By analyzing 43,568 women of European ancestry, this large study is the first to identify variants and genes that are statistically associated with sPTB. Interestingly, our integrative analysis identified *EBF1* as a desirable candidate ($d_{overall}$ = 0.15 [top 3%] in iR-none and $d_{overall}$ = 0.23 [top 1%] in

iR-per), suggesting that this gene, in addition to GWAS, might also be functionally linked to sPTB pathogenesis across transcriptomics, epigenomics, and proteomics studies. Even when analyzing the 9 other omics studies without this GWAS data set, *EBF1* still achieved a $d_{overall}$ score of 0.17, placing it in the top 2% of all genes (Additional file 14). While our integrative analysis supports the identification of *EBF1* as an interesting candidate gene for follow up, the lack of signal for any of the other GWAS-identified hits also reinforces the need to approach complex phenotypes like sPTB from a variety of omics perspectives, since sequenced-based changes may impact the phenotype in indirect and complicated functional ways.

Conclusions

Desirability-based data integration (and our integRATE software) is a solution most applicable in biological research areas where omics data is especially heterogeneous and sparse. Our approach combines information from all variables across all related studies to calculate the total weight of evidence for any given gene as a candidate involved in disease pathogenesis, for example. Although not a statistical approach, this method of data integration allows for the prioritization of candidate genes based on information from heterogeneous omics data even without known 'gold standard' genes to test against and can be used to inform more targeted downstream functional analyses.

Additional files

Additional file 1: Results of meta-analysis to identify studies for integration. We outline the 10 studies meeting all inclusion criteria for integrative analysis. Furthermore, we list the other 44 studies that we identified through our literature search but we excluded from the data analysis as well as reasons for their exclusion. (XLSX 42 kb)

Additional file 2: All results from iR-none. All desirability scores across all variables in all studies as well as overall desirabilities and normalized overall desirabilities are presented. (XLSX 3762 kb)

Additional file 3: All results from iR-num. All desirability scores across all variables in all studies as well as overall desirabilities and normalized overall desirabilities are presented. (XLSX 1988 kb)

Additional file 4: Results from iR-num. All genes in the analysis including numerical cut points were sorted from most desirable (rank = 1) to least desirable (rank = 26,869) and plotted according to their overall desirability scores. (EPS 611 kb)

Additional file 5: Top 10 genes from iR-num by data type. Desirability scores for the top 10 most desirable genes are plotted according to the type of omics analysis. (EPS 761 kb)

Additional file 6: Top 10 genes from iR-num by study. Desirability scores for the top 10 most desirable genes are plotted according to individual study. (EPS 784 kb)

Additional file 7: All results from iR-per. All desirability scores across all variables in all studies as well as overall desirabilities and normalized overall desirabilities are presented. (XLSX 3512 kb)

Additional file 8: Results from iR-per. All genes in the iR-per analysis were sorted from most desirable (rank = 1) to least desirable (rank = 26,869) and plotted according to their overall desirability scores. (EPS 618 kb)

Additional file 9: Top 10 genes from iR-per by data type. Desirability scores for the top 10 most desirable genes are plotted according to the type of omics analysis. (EPS 764 kb)

Additional file 10: Top 10 genes from iR-per by study. Desirability scores for the top 10 most desirable genes are plotted according to individual study. (EPS 789 kb)

Additional file 11: Raw data for manual overlap based on significance dichotomization. All 18,727 genes identified as significant in at least 1 study and overlap across the entire data set. (XLSX 769 kb)

Additional file 12: Genes binned as significant in 4 or more omics studies. Upset plot showing intersections of significant genes across all 10 omics studies. (EPS 12 kb)

Additional file 13: GO-Slim gene set enrichment results. The PANTHER output for gene set functional enrichment is provided, including 37 statistically enriched biological pathways. (XLSX 13 kb)

Additional file 14: All results *without* including the Zhang 2017 data set [30]. All desirability scores across all variables in all studies as well as overall desirabilities and normalized overall desirabilities are presented. (XLSX 2908 kb)

Abbreviations
GWAS: Genome-wide association study; IUGR: Intrauterine growth restriction; PE: Preeclampsia; PPROM: Premature rupture of membranes; PTB: Preterm birth; SNP: Single nucleotide polymorphism; sPTB: Spontaneous preterm birth

Availability of data and materials
All the data associated with and supporting the findings of this study are included in the manuscript and its supplementary files.

Authors' contributions
Conceived and designed experiments: HRE, AR. Performed experiments: HRE. Developed scripts: HRE, JS. Analyzed data: HRE. Wrote paper: HRE. Assisted with project development: AR, JHW, PA, JAC. Provided feedback: JHW, PA, JAC, AR. All authors have read and approved the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

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

Author details
[1]Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA. [2]Department of Biochemistry, Purdue University, West Lafayette, IN, USA.

Eidem *et al. BMC Medical Genomics*   (2018) 11:107

Page 12 of 13

³Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. ⁴Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA.

References

1. Gohlke JM, Thomas R, Zhang Y, Rosenstein MC, Davis AP, Murphy C, et al. Genetic and environmental pathways to complex diseases. BMC Syst Biol. 2009;3:46.
2. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83.
3. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148:1293–307. https://doi.org/10.1016/j.cell.2012.02.009.
4. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19:299–310. https://doi.org/10.1038/nrg.2018.4.
5. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013;93: 779–97.
6. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
7. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome profiling in human diseases: new advances and perspectives. Int J Mol Sci. 2017;18:1652. https://doi.org/10.3390/ijms18081652.
8. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16:85–97. https://doi.org/10.1038/nrg3868.
9. Holzinger ER, Ritchie MD. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. Pharmacogenomics. 2012;13:213–22. https://doi.org/10.2217/pgs.11.145.
10. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. J Biomed Inform. 2012;45:1191–8. https://doi.org/10.1016/j.jbi.2012.07.008.
11. Peng C, Li A, Wang M. Discovery of bladder cancer-related genes using integrative heterogeneous network modeling of multi-omics data. Sci Rep. 2017;7:15639. https://doi.org/10.1038/s41598-017-15890-9.
12. Pavel AB, Sonkin D, Reddy A. Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. BMC Syst Biol. 2016;10:16. https://doi.org/10.1186/s12918-016-0260-9.
13. Zhu J, Shi Z, Wang J, Zhang B. Empowering biologists with multi-omics data: colorectal cancer as a paradigm. Bioinformatics. 2015;31:1436–43. https://doi.org/10.1093/bioinformatics/btu834.
14. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8. https://doi.org/10.1038/nature07385.
15. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17:98–110. https://doi.org/10.1016/j.ccr.2009.12.020.
16. Cohen J. The cost of dichotomization. Appl Psychol Meas. 1983;7:249–53. https://doi.org/10.1177/014662168300700301.
17. Streiner DL. Breaking up is hard to do: the heartbreak of dichotomizing continuous data. Can J Psychiatr. 2002;47:262–6. https://doi.org/10.1177/070674370204700307.
18. Barnwell-Ménard J-L, Li Q, Cohen AA. Effects of categorization method, regression type, and variable distribution on the inflation of type-I error rate when categorizing a confounding variable. Stat Med. 2015;34:936–49. https://doi.org/10.1002/sim.6387.
19. Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. Expert Rev Proteomics. 2004;1:67–75. https://doi.org/10.1586/14789450.1.1.67.
20. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CMT, Beyene J. Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics. 2009;2009:1–13. https://doi.org/10.4061/2009/869093.
21. Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. Mamm Genome. 2007;18:389–401. https://doi.org/10.1007/s00335-007-9040-6.
22. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010;11:476–86. https://doi.org/10.1038/nrg2795.
23. Lazic SE. Ranking, selecting, and prioritising genes with desirability functions. PeerJ. 2015;3:e1444. https://doi.org/10.7717/peerj.1444.
24. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012;4:90–8. https://doi.org/10.1038/nchem.1243.
25. Harrington E. The desirability function. Ind Qual Control. 1965;21:494–8.
26. Derringer G, Suich R. Simultaneous optimization of several response variables. J Qual Technol. 1980;12:214–9.
27. Derringer G. A balancing act: optimizing a products properties. Qual Prog. 1994;27:51.
28. Romero R, Dey SK, Fisher SJ. Preterm labor: one syndrome, many causes. Science. 2014;345:760–5. https://doi.org/10.1126/science.1251816.
29. Eidem HR, McGary KL, Capra JA, Abbot P, Rokas A. The transformative potential of an integrative approach to pregnancy. Placenta. 2017;57: 204–15.
30. Zhang G, Jacobsson B, Muglia LJ. Genetic associations with spontaneous preterm birth. N Engl J Med. 2017;377:2401–2. https://doi.org/10.1056/NEJMc1713902.
31. Plunkett J, Muglia LJ. Genetic contributions to preterm birth: implications from epidemiological and genetic association studies. Ann Med. 2008;40: 167–79. https://doi.org/10.1080/07853890701806181.
32. Muglia LJ, Katz M. The enigma of spontaneous preterm birth. N Engl J Med. 2010;362:529–35. https://doi.org/10.1056/NEJMra0904308.
33. Lengyel C, Muglia LJ, Pavlíčev M. Genetics of Preterm Birth. In: eLS. Chichester: Wiley; 2014. https://doi.org/10.1002/9780470015902.a0025448.
34. Ackerman WE, Buhimschi IA, Eidem HR, Rinker DC, Rokas A, Rood K, et al. Comprehensive RNA profiling of villous trophoblast and decidua basalis in pregnancies complicated by preterm birth following intra-amniotic infection. Placenta. 2016;44:23–33. https://doi.org/10.1016/j.placenta.2016.05.010.
35. Heng YJ, Taylor L, Larsen BG, Chua HN, Pung SM, Lee MWF, et al. Albumin decrease is associated with spontaneous preterm delivery within 48 h in women with threatened preterm labor. J Proteome Res. 2015;14:457–66. https://doi.org/10.1021/pr500852p.
36. Parets SE, Conneely KN, Kilaru V, Fortunato SJ, Syed TA, Saade G, et al. Fetal DNA methylation associates with early spontaneous preterm birth and gestational age. PLoS One. 2013;8:e67489. https://doi.org/10.1371/journal.pone.0067489.
37. Cruickshank MN, Oshlack A, Theda C, Davis PG, Martino D, Sheehan P, et al. Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. Genome Med. 2013;5:96. https://doi.org/10.1186/gm500.
38. Saade GR, Boggess KA, Sullivan SA, Markenson GR, Iams JD, Coonrod DV, et al. Development and validation of a spontaneous preterm delivery predictor in asymptomatic women. Am J Obstet Gynecol. 2016;214:633.e1–633.e24. https://doi.org/10.1016/j.ajog.2016.02.001.
39. Zhang G, Bacelis J, Lengyel C, Teramo K, Hallman M, Helgeland Ø, et al. Assessing the causal relationship of maternal height on birth size and gestational age at birth: a Mendelian randomization analysis. PLoS Med. 2015;12:e1001865. https://doi.org/10.1371/journal.pmed.1001865.
40. Makieva S, Dubicke A, Rinaldi SF, Fransson E, Ekman-Ordeberg G, Norman JE. The preterm cervix reveals a transcriptomic signature in the presence of premature prelabor rupture of membranes. Am J Obstet Gynecol. 2017;216: 602.e1–602.e21. https://doi.org/10.1016/j.ajog.2017.02.009.
41. Heng YJ, Pennell CE, Chua HN, Perkins JE, Lye SJ. Whole blood gene expression profile associated with spontaneous preterm birth in women with threatened preterm labor. PLoS One. 2014;9:e96901. https://doi.org/10.1371/journal.pone.0096901.
42. Chim SSC, Lee WS, Ting YH, Chan OK, Lee SWY, Leung TY. Systematic identification of spontaneous preterm birth-associated RNA transcripts in maternal plasma. PLoS One. 2012;7:e34328. https://doi.org/10.1371/journal.pone.0034328.
43. Mayor-Lynn K, Toloubeydokhti T, Cruz AC, Chegini N. Expression profile of MicroRNAs and mRNAs in human placentas from pregnancies complicated by preeclampsia and preterm labor. Reprod Sci. 2011;18:46–56. https://doi.org/10.1177/1933719110374115.
44. de Goede OM, Lavoie PM, Robinson WP. Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. Clin Epigenetics. 2017;9:39. https://doi.org/10.1186/s13148-017-0339-1.

45. Hong X, Sherwood B, Ladd-Acosta C, Peng S, Ji H, Hao K, et al. Genome-wide DNA methylation associations with spontaneous preterm birth in US blacks: findings in maternal and cord blood samples. Epigenetics. 2018;13: 163–72. https://doi.org/10.1080/15592294.2017.1287654.

46. Fernando F, Keijser R, Henneman P, van der Kevie-Kersemaekers A-MF, Mannens MM, van der Post JA, et al. The idiopathic preterm delivery methylation profile in umbilical cord blood DNA. BMC Genomics. 2015;16: 736. https://doi.org/10.1186/s12864-015-1915-4.

47. Zhang G, Feenstra B, Bacelis J, Liu X, Muglia LM, Juodakis J, et al. Genetic associations with gestational duration and spontaneous preterm birth. Obstet Gynecol Surv. 2017;73:83–5. https://doi.org/10.1097/01.ogx. 0000530434.15441.45.

48. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11:e1004219. https://doi. org/10.1371/journal.pcbi.1004219.

49. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33:2938–40. https://doi.org/10.1093/bioinformatics/btx364.

50. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45: D183–9. https://doi.org/10.1093/nar/gkw1138.

51. Heng YJ, Liong S, Permezel M, Rice GE, Di Quinzio MKW, Georgiou HM. Human cervicovaginal fluid biomarkers to predict term and preterm labor. Front Physiol. 2015;6. https://doi.org/10.3389/fphys.2015.00151.

52. Velez DR, Fortunato SJ, Thorsen P, Lombardi SJ, Williams SM, Menon R. Preterm birth in Caucasians is associated with coagulation and inflammation pathway gene variants. PLoS One. 2008;3:e3283. https://doi. org/10.1371/journal.pone.0003283.

53. Cowman J, Quinn N, Geoghegan S, Müllers S, Oglesby I, Byrne B, et al. Dynamic platelet function on von Willebrand factor is different in preterm neonates and full-term neonates: changes in neonatal platelet function. J Thromb Haemost. 2016;14:2027–35. https://doi.org/10.1111/jth.13414.

54. Strauss T, Elisha N, Ravid B, Rosenberg N, Lubetsky A, Levy-Mendelovich S, et al. Activity of Von Willebrand factor and levels of VWF-cleaving protease (ADAMTS13) in preterm and full term neonates. Blood Cells Mol Dis. 2017; 67:14–7. https://doi.org/10.1016/j.bcmd.2016.12.013.

55. Zhu Y, Paszty C, Turetsky T, Tsai S, Kuypers FA, Lee G, et al. Stomatocytosis is absent in "stomatin"-deficient murine red blood cells. Blood. 1999;93:2404–10 http://www.ncbi.nlm.nih.gov/pubmed/10090952.

56. Littlefield R, Almenar-Queralt A, Fowler VM. Actin dynamics at pointed ends regulates thin filament length in striated muscle. Nat Cell Biol. 2001;3:544–51. https://doi.org/10.1038/35078517.

57. Caldwell JE, Heiss SG, Mermall V, Cooper JA. Effects of CapZ, an actin-capping protein of muscle, on the polymerization of actin. Biochemistry. 1989;28:8506–14. https://doi.org/10.1021/bi00447a036.

58. Bottega R, Marconi C, Faleschini M, Baj G, Cagioni C, Pecci A, et al. ACTN1-related thrombocytopenia: identification of novel families for phenotypic characterization. Blood. 2015;125:869–72. https://doi.org/10.1182/blood-2014-08-594531.

59. Kunishima S, Okuno Y, Yoshida K, Shiraishi Y, Sanada M, Muramatsu H, et al. ACTN1 mutations cause congenital macrothrombocytopenia. Am J Hum Genet. 2013;92:431–8. https://doi.org/10.1016/j.ajhg.2013.01.015.

60. Zhao P, Chidiac P. Regulation of RGS5 GAP activity by GPSM3. Mol Cell Biochem. 2015;405:33–40. https://doi.org/10.1007/s11010-015-2393-3.

61. Ono S. Functions of actin-interacting protein 1 (AIP1)/WD repeat protein 1 (WDR1) in actin filament dynamics and cytoskeletal regulation. Biochem Biophys Res Commun. 2017. https://doi.org/10.1016/j.bbrc.2017.10.096.

62. Liu J, Aoki M, Illa I, Wu C, Fardeau M, Angelini C, et al. Dysferlin, a novel skeletal muscle gene, is mutated in Miyoshi myopathy and limb girdle muscular dystrophy. Nat Genet. 1998;20:31–6. https://doi.org/10.1038/1682.

63. Athayde N, Romero R, Gomez R, Maymon E, Pacora P, Mazor M, et al. Matrix metalloproteinases-9 in preterm and term human parturition. J Matern Neonatal Med. 1999;8:213–9. https://doi.org/10.3109/14767059909052049.

64. Chen J, Khalil RA. Matrix metalloproteinases in normal pregnancy and preeclampsia. In: Progress in molecular biology and translational science; 2017. p. 87–165. https://doi.org/10.1016/bs.pmbts.2017.04.001.

65. Xu P, Alfaidy N, Challis JRG. Expression of matrix metalloproteinase (MMP)-2 and MMP-9 in human placenta and fetal membranes in relation to preterm and term labor. J Clin Endocrinol Metab. 2002;87:1353–61. https://doi.org/ 10.1210/jcem.87.3.8320.

66. Poon LCY, Nekrasova E, Anastassopoulos P, Livanos P, Nicolaides KH. First-trimester maternal serum matrix metalloproteinase-9 (MMP-9) and adverse pregnancy outcome. Prenat Diagn. 2009;29:553–9. https://doi. org/10.1002/pd.2234.

67. Romero R, Chaiworapongsa T, Espinoza J, Gomez R, Yoon BH, Edwin S, et al. Fetal plasma MMP-9 concentrations are elevated in preterm premature rupture of the membranes. Am J Obstet Gynecol. 2002;187:1125–30. https:// doi.org/10.1067/mob.2002.127312.

68. Tency I, Verstraelen H, Kroes I, Holtappels G, Verhasselt B, Vaneechoutte M, et al. Imbalances between matrix metalloproteinases (MMPs) and tissue inhibitor of metalloproteinases (TIMPs) in maternal serum during preterm labor. PLoS One. 2012;7:e49042. https://doi.org/10.1371/journal. pone.0049042.

69. Sundrani DP, Chavan-Gautam PM, Pisal HR, Mehendale SS, Joshi SR. Matrix metalloproteinase-1 and -9 in human placenta during spontaneous vaginal delivery and caesarean sectioning in preterm pregnancy. PLoS One. 2012;7:e29855.

70. Yonemoto H, Young CB, Ross JT, Guilbert LL, Fairclough RJ, Olson DM. Changes in matrix metalloproteinase (MMP)-2 and MMP-9 in the fetal amnion and chorion during gestation and at term and preterm labor. Placenta. 2006;27:669–77.

71. Eidem HR, Ackerman WE, McGary KL, Abbot P, Rokas A. Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. BMC Med Genet. 2015;8:27.