Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and “Big Data” Biology

Juan C. Vivar,¹ Priscilla Pemu,² Ruth McPherson,³ and Sujoy Ghosh^{1,4}

Abstract

Unparalleled technological advances have fueled an explosive growth in the scope and scale of biological data and have propelled life sciences into the realm of “Big Data” that cannot be managed or analyzed by conventional approaches. Big Data in the life sciences are driven primarily via a diverse collection of ‘omics’-based technologies, including genomics, proteomics, metabolomics, transcriptomics, metagenomics, and lipidomics. Gene-set enrichment analysis is a powerful approach for interrogating large ‘omics’ datasets, leading to the identification of biological mechanisms associated with observed outcomes. While several factors influence the results from such analysis, the impact from the contents of pathway databases is often under-appreciated. Pathway databases often contain variously named pathways that overlap with one another to varying degrees. Ignoring such redundancies during pathway analysis can lead to the designation of several pathways as being significant due to high content-similarity, rather than truly independent biological mechanisms. Statistically, such dependencies also result in correlated *p* values and overdispersion, leading to biased results. We investigated the level of redundancies in multiple pathway databases and observed large discrepancies in the nature and extent of pathway overlap. This prompted us to develop the application, ReCiPa (Redundancy Control in Pathway Databases), to control redundancies in pathway databases based on user-defined thresholds. Analysis of genomic and genetic datasets, using ReCiPa-generated overlap-controlled versions of KEGG and Reactome pathways, led to a reduction in redundancy among the top-scoring gene-sets and allowed for the inclusion of additional gene-sets representing possibly novel biological mechanisms. Using obesity as an example, bioinformatic analysis further demonstrated that gene-sets identified from overlap-controlled pathway databases show stronger evidence of prior association to obesity compared to pathways identified from the original databases.

Introduction

RAPID ADVANCES IN GENOMIC TECHNOLOGIES and current improvements in cost-efficiencies have paved the way for the acquisition of high-content genetic and genomic information on a large number of samples, leading to the “Big Data” revolution in biology (Dumbbill, 2013). Microarrays for the genome-wide detection of gene expression or single nucleotide polymorphisms (SNPs), large-scale epigenetic profiling, and next-generation DNA and RNA sequencing are examples of some of the technologies that generate complex, high-content datasets. Analysis and interpretation of these datasets present some unique computational and interpretive challenges. A strategy commonly employed to address the

“Big Data” analysis challenge is to reduce the dimensionality of the data by collapsing individual data points into larger blocks of biologically meaningful groups and subsequently analyzing the groups, instead of the individual components. For example, in the case of whole-genome gene expression profiling, one could first categorize the genes into coherent biologically meaningful gene-sets (pathways) and then interrogate changes in expression at the pathway, instead of the gene level (Dinu et al., 2008). Commonly referred to as ‘pathway analysis,’ this approach has also been successfully applied to the analysis of genome-wide association data that query genome scale association of single-nucleotide polymorphisms (SNPs) to disease and other outcomes (Wang et al., 2007). By collapsing genes and SNPs into pathways, one

¹Biomedical Biotechnology Research Institute, North Carolina Central University, Durham, North Carolina.

²Clinical Research Center, Morehouse School of Medicine, Atlanta, Georgia.

³University of Ottawa Heart Institute, Ottawa, Ontario, Canada.

⁴Dulce-NUS Graduate Medical School, Singapore.

reduces the number of observations from tens of thousands or millions to only a few hundreds. This leads to two distinct benefits. First, it facilitates statistical analysis by significantly reducing the multiple-testing burden. Second, it focuses the analysis *de facto* on biologically relevant mechanisms that are defined by the pathways. Thus interrogating gene function at the level of pathways facilitates biological interpretation of genetic data, which is often missed if only individual polymorphisms or individual gene expressions are studied (Song and Black, 2008; Wang et al., 2007).

Gene-set enrichment analysis is usually carried out through the use of several public or user-developed gene annotation databases that store information on gene-sets, or pathways. Some of the well-known public pathway repositories include the Kyoto Encyclopedia of Genes and Genomes or KEGG (Ogata et al., 1999), Biocarta, Reactome (Vastrik et al., 2007) and Gene Ontology or GO databases (Harris et al., 2004). Additionally there exist database suites, such as the Molecular Signatures Database (MSigDB), that contain collections of pathway databases organized into categories (Subramanian et al., 2005b). Generally, pathway databases are organized around differing biological themes. For example, KEGG is a collection of pathway maps representing our current understanding of the molecular interaction and reaction networks for metabolism, cellular information processing, organismal systems, and human diseases. The Chemical and Genomic Perturbation database (or CGP, a component of MSigDB), on the other hand, is a collection of gene-sets that represent gene expression signatures experimentally observed in chemical and genetic perturbation studies. A common goal of pathway analysis is to employ one or more of these pathway databases to detect statistically significant changes in the expression or association signals of gene-sets, related to the outcome of interest. The development of statistical and bioinformatics tools for the identification of statistically significant pathways is a highly active area of research, and a substantial number of statistical procedures for assessing pathway enrichment has been introduced over the last few years (Ackermann and Strimmer, 2009; Bankhead et al., 2009; Dutta et al., 2012; Gu et al., 2012; Huang da et al., 2009; Ibrahim et al., 2012; Tarca et al., 2009). A recent review (Khatri et al., 2012) on the last 10 years of pathway analysis further identified a gradual progression of endeavor; a first generation of ‘over-representation analysis (ORA)’ that statistically evaluates the fraction of genes in a particular pathway from among the set of genes satisfying a user-defined threshold for changes in expression; a second generation of ‘functional class scoring (FCS)’ methodologies based on a ‘no-cutoff’ strategy that includes all genes from an experiment for pathway enrichment analysis; and a third generation of pathway “topology-based” approaches that, while methodologically similar to FCS, uses additional information on gene–gene interactions to compute gene-level statistics and pathway enrichment (Draghici et al., 2007; Shojaie and Michailidis, 2009).

While pathway analysis provides a number of benefits over single gene or SNP based approaches, the dependence of the obtained results on the content of pathway databases remains an important but underexplored area of study. Based on how such databases are constructed and assembled, a repository may include differently named pathways containing sets of genes that overlap with one another to varying degrees. This leads to redundancies among several pathways. Such overlap

or redundancy can either be a reflection of true biological parsimony where one gene may function in multiple biological processes [e.g., dual roles of bacterial isocitrate lyase 1 in the glyoxalate and methylcitrate cycles or parallel roles for iron regulatory protein 1 as a cytoplasmic aconitase and a regulatory RNA-binding protein (Frishman and Hentze, 1996; Gould et al., 2006)], or a consequence of database organization (e.g., hierarchical structures such as Gene Ontology). Regardless of the reason for such overlap, the inclusion of a large number of highly overlapping gene-sets in pathway analysis could lead to the designation of several pathways as being significant and highly ranked simply due to their high content-similarity, rather than reflecting truly diverse biological mechanisms. On the other hand, most of the statistical tests underlying pathway analysis are based on the null hypothesis under which pathways are considered to be independent of each other (Gatti et al., 2010; Goeman and Buhlmann, 2007; Shi et al., 2008). However, when there is overlap or redundancy among the gene constituents of two or more pathways, the pathways are not truly ‘independent’. Such dependencies between pathways could result in correlated *p* values and overdispersion of the number of significant pathways, leading to biased results. Additionally, as pathway ranks are often the basis for pathway selection, there remains the risk of selecting redundant higher-ranking pathways at the expense of lower ranked pathways that could possibly have informed on novel biological mechanisms.

In this report, we have investigated the nature and degree of pathway redundancy among several pathway databases commonly used for genomic and genetic studies. Our results show that pathway databases differ widely with respect to the extent of overlap among their constituent gene-sets. To mitigate the effects from gene-set redundancies, we have developed an application in the R programming language named ReCiPa (Redundancy Control in Pathway Databases). ReCiPa allows for user-specified control of overlap in pathway databases. Such control improves gene-set analysis by reducing the overlap among highly ranked pathways and allowing for additional biological mechanisms to be included for further analysis.

Material and Methods

Datasets and databases

Datasets from whole-genome expression profiling and genome-wide association analysis were generated in the authors’ laboratories. Pathway databases were downloaded from Pathway Commons (www.pathwaycommons.org) and Molecular Signatures Database (MSigDB: <http://www.broadinstitute.org/gsea/msigdb/index.jsp>). The following pathway databases were downloaded from MSigDB: Kyoto Encyclopedia of Genes and Genomes (KEGG), Biocarta, Reactome, Chemical and Genomics Perturbations (CGP). The NCI-Nature Pathway Interaction database (NCI-PID) was downloaded from Pathway Commons.

Algorithm for pathway overlap estimation

Let M be a square matrix with elements $M[i, j], i, j = 1, \dots, N$, where N is the total number of pathways in the database. Each pathway is a set of genes, then let i be the list of

genes that belong to pathway i . The total number of elements in a set A is represented by $|A|$, then the pathway i has size $|g_i|$. The elements $M[i, j]$ are defined as follows:

$$M[i, j] = \frac{|g_i \cap g_j|}{|g_j|}, \quad i, j = 1, \dots, N,$$

where the numerator is the total number of *common* genes between pathways i and j and the denominator is the size of pathway j . Therefore, each element of matrix M is the proportion representing this overlap with respect to pathway i . It is clear that M is not symmetric and the diagonal elements $M[i, i]$ are equal to 1. Since we are not interested in these elements, they are set to zero. Using the maximum and minimum overlap variables ($max_overlap$ and $min_overlap$), all the elements that meet the condition $M[i, j] > max_overlap$ are selected. From this smaller list, we keep only those that meet $M[j, i] > min_overlap$ for later use. These overlap variables will be described below. It is possible that both $M[i, j]$ and $M[j, i]$ meet the condition, but since ReCiPa's goal is to merge or combine redundant pathways, we need only one pair $M[i, j]$ or $M[j, i]$. The selected will be the pair with the greater overlap proportion, the lower is removed. With all these elements, a table with 6 columns is created: *Column 1*: Pathway i ; *Column 2*: Pathway j ; *Column 3*: Size of pathway i , $|g_i|$; *Column 4*: Size of pathway j , $|g_j|$; *Column 5*: Overlap between pathway i and pathway j $M[i, j]$; *Column 6*: Number of common genes, $|g_i \cap g_j|$.

With this table ready, there are several cases that could occur. We will describe them briefly:

Case 1: Overlap = 1 and pathway j appears more than once in Column 2. This means that more than one pathway is completely contained in pathway j . Merging all these pathways with pathway j will result in the same pathway j and its new name will add the names of the other pathways.

Case 2: Overlap = 1 and pathway i appears more than once in Column 1. This means that pathway i is entirely contained within more than one pathway. Then pathway i is merged with one of those pathways in Column 2, that with the greater number of genes in Column 4. Names are added accordingly.

Case 3: Overlap = 1 and pathways i and j are not repeated in the table. Both pathways are merged and the new name is built as described under Case 1.

Case 4: Overlap < 1 and pathway i appears more than once in Column 1. In this case, pathway i is partially contained within more than one pathway. Pathway i is merged with the pathway with the greater overlap (Column 5). The new pathway is named as described in Case 1. The genes unique to pathway i and pathway j are included in the list of genes in the new pathway.

$$g_j^* = \{g_i \cap g_j\} + \{g_i \setminus \{g_i \cap g_j\}\} + \{g_j \setminus \{g_i \cap g_j\}\}.$$

Case 5: Overlap < 1 and pathway j is present one or more times in Column 2. This is processed in a similar way as Case 4. Pathways showing greatest overlap are merged and the genes unique to both pathways are added to the list of genes in the newly merged pathway.

After one round of screening, the built “superpathways” are added to the database and pathways that were merged are removed. The updated database is used to generate a new overlap matrix M and the method of merging pathways is continued. The process is repeated until all entries of the matrix M are lower than the maximum overlap specified by the user.

Using the ReCiPa algorithm

ReCiPa is available as a package in CRAN (<http://cran.r-project.org/web/packages/ReCiPa/>). After installing and loading the package ‘ReCiPa’ in any R distribution, the function ‘ReCiPa’ requires four arguments:

ReCiPa(*path_db_name*, *new_path_db_name*, *max_overlap*, *min_overlap*)

path_db_name: the full name of the pathway database file. ReCiPa has been tested with gene set data extracted from the MSigDB. Pathway files in MSigDB are tab-delimited text files with the extension “gmt” (gene matrix transposed file) and are imported into ReCiPa without further manipulation. In the “gmt” format, each row represents a gene set and tabs are used as column separators. The first column contains pathway or gene-set names and the second column contains a pathway descriptor (e.g., a weblink to a pathway descriptor website). Columns 3 and beyond each contain a gene identifier (e.g., HUGO names or Entrez IDs) for the genes included in the pathway described in Column 1. Databases from other sources should be formatted similarly (.gmt) for analysis in ReCiPa.

new_path_db_name: the name of the new file that will be created. At the end of the process, the full name of the new file will also contain the following, separated by underscores: the maximum and minimum redundancy allowed, the number of loops or repetitions, the number of pathways in the new database and the maximum number of genes in a pathway. Example: “new kegg” will become “new_kegg_0.75_0.10_8_157_399.txt”, if we set “max_overlap” to 0.75 and “min_overlap” to 0.10.

max_overlap: the maximum overlap or redundancy allowed by the user between any two pathways. The values range from 0 to 1, where 0 indicates total independence between any pair of pathways with no gene repetitions (this, in certain cases can generate very large pathway ensembles or “megapathways” that are not suited for pathway analysis). If the overlap is set to 1, it will return the original input file. Any value between 0 and 1 will generate a new pathway list containing pathways with the desired maximum redundancy among them.

min_overlap: the minimum overlap allowed by the user between any pathways j and i that already met the previous condition ($M[i, j]$ being greater than $max_overlap$). The pathways corresponding to overlaps below this threshold are not considered for the merging. For example, on Reactome database, the pathway Early phase of HIV life cycle (13 genes) is completely contained in the pathway HIV infection (183 genes), then $M[i, j] = 1$ and $M[j, i] = 0.071$. If $max_overlap$ is set to 0.75 and $min_overlap$ is set to 0.10, these two pathways will not be combined.

Determining biological relevance via Biograph

The relevance of candidate genes and pathways to the biological phenotype of interest (e.g., ‘obesity’) was estimated

in silico via the Biograph knowledge-mining software (Liekens et al., 2011). Biograph assembles and analyzes information from heterogeneous databases via unsupervised data mining techniques and stochastic random walks to generate a map of relations linking ‘source concepts’ (e.g., phenotypes such as ‘obesity’) to ‘targets’ (e.g., genes from the pathways of interest). This network of relationships is analyzed by the Biograph algorithm to score and rank the different ‘paths’ linking concepts to targets, resulting in an automated formulation of one or more functional hypotheses. The relative strength of each hypothesis is computed into an aggregate score for each gene, which allows one to assess the ‘proximity’ of the association between the gene and the phenotype being studied. The mean of the gene scores are utilized to derive a pathway proximity score from the group of genes belonging to the pathway. The pathway-based scores are then used to estimate the relative relevance or proximity of a pathway to the phenotype of ‘obesity’.

Results

Characterizing redundancies in pathway databases

The extent of pathway redundancies and the distribution of percent overlap are shown in Figure 1a, b. In this context, ‘redundancy’ is defined as the proportion of common genes between two gene-sets, with respect to each of the gene sets (thus two redundancy values are obtained for each pairwise gene set comparison, represented by $M[i, j]$ and $M[j, i]$ in the above equations). We interrogated five different pathway databases (KEGG, Biocarta, CGP, NCI-PID, and Reactome) and quantified the number of pathways at different levels of redundancies. Figure 1a plots the proportion of total pathways that exhibit redundancies greater than pre-specified thresholds (x-axis). The trajectories are color-coded to represent the five pathway databases. Since all databases exhibit $\geq 0\%$ redundancies among its pathways, all trajectories begin at a value of 1 on the y-axis. Subsequently, however, the trajectories for the different databases diverge, based on the proportion of their pathways that meet each redundancy threshold. In Figure 1b, we investigated the distribution of pathway redundancy in each database by considering pathways that show at least 25% overlap with another pathway. By setting the lower redundancy bound at 25%, we were better able to visualize the differences in redundancy distributions among the databases. Boxplot analysis showed the KEGG, Biocarta, and CGP databases to have a similar distribution of the percent redundancies in the 25%–100% range. Within the same range, however, the median percent redundancy was approximately 2-fold higher in the Reactome database, and even greater in the NCI-PID database. Additional summaries of pathway redundancies among the five pathway databases are provided in Supplementary Table S1; supplementary data are available online at www.liebertpub.com/omi.

Effect of overlap control on redundancies among top scoring pathways

We next investigated the effects of redundancy reduction on pathway analysis results (Fig. 2). A whole-genome expression profiling dataset comparing blood gene expression profiles from obese and lean subjects (Ghosh et al., 2010) was

analyzed by GSEA using the Reactome pathway database. GSEA was conducted either on the original Reactome pathways or on overlap-controlled versions allowing for 75% or 95% maximum overlap among pathways. The top five statistically significant obese-upregulated pathways and ‘superpathways’ (ranked by nominal p values) identified from each version of the database are shown, along with a heatmap depicting the degree of overlap among them (smaller overlaps in yellow and larger overlaps in red). The exact content of the five ‘superpathways’ are listed in Supplementary Table S2. The pair-wise percent overlap between any two pathways is also marked in each cell. As is evident from Figure 2, a high degree of inter-pathway overlap was observed among the top five pathways in the original version of Reactome, whereas the overlap was significantly diminished in overlap-controlled versions of the database. We did not notice a qualitative change in the size of the top-ranked pathways in the original or overlap-controlled versions of the database (indicated by the ‘size’ column in Fig. 2).

Effect of overlap control on pathway ranks

The effect of overlap-control on pathway ranks was investigated across three different studies—a whole-genome expression profiling study in cultured progenitor cells from obese and lean subjects conducted on Agilent Human Whole-Genome microarrays (Study 1; (Pemu et al., 2012)); a genome-wide association study (GWAS) on morbid obesity, generated on Affymetrix SNP chip 6.0 (Study 2); and the previously mentioned blood gene expression study on obese and lean subjects, conducted on Affymetrix U133 Plus 2 microarrays (Study 3; (Ghosh et al., 2010)). Studies 1 and 3 were analyzed via the GSEA algorithm (Subramanian et al., 2005; 2005a), whereas Study 2 was analyzed by the GSASNP software (Nam et al., 2010). To compare effects across pathway databases, studies 1 and 2 were analyzed using KEGG and study 3 was analyzed using Reactome. For Studies 1 and 3, the obese-downregulated gene-sets were considered. Gene-sets were ranked by their nominal p values. In all studies, the ranks of the top 20 pathways (identified from the original database) were compared between the original or overlap-controlled versions of pathway databases. Overall, there was a net gain in ranks for several pathways when the analysis was conducted with the overlap-controlled databases, due to the merging of two or more highly ranked but highly redundant pathways from the original version. This is schematically represented for Study 1 in Figure 3 where the ranks of 13 pathways move upward in the overlap-controlled database compared to the original version and only one pathway show a loss in rank. Four of the original pathways are collapsed into two superpathways (SP4 and SP10) upon overlap-control. This trend was observed for all three studies, irrespective of the biological samples (blood or progenitor cells), the type of study (gene-expression or GWAS), or the pathway databases used (KEGG or Reactome). Table 1 provides a summary of the comparative analysis of *pathway ranks* in each of the three studies. The number of pathways that merged into superpathways was greater in Reactome compared to KEGG and in all cases the number of pathways that gained in rank exceeded the number of pathways with loss in ranks in the overlap-controlled versions. Further details on the pathways and analysis of their relative ranks are provided in Supplementary Table S3.

FIG. 1. Distribution of redundancies in pathway databases. Results are shown for five pathway databases (Biocarta, KEGG, Reactome, NCI-PID, and CGP). **(a)** Comparison of the proportion of redundancies (y-axis) at defined overlap thresholds (x-axis) for the different pathway databases. In this context, 'redundancy' is defined as the proportion of common genes between two gene-sets, with respect to each of the gene-sets. The 25% and 75% overlap thresholds are indicated by *dashed vertical lines*. **(b)** Boxplots depicting the distribution of 25% or greater pathway redundancies among the 5 pathway databases. The total number of pairwise comparisons at $\geq 25\%$ redundancies are shown for each database.

					Five top pathways by Nominal p-value	Size	p-value	Database	
1		0.812	0.812	1.000	0.812	REACTOME_REGULATION_OF_GENE_EXPRESS SION IN BETA CELLS	75	0.000	ORIGINAL
2	0.976		0.976	0.976	0.976	REACTOME_VIRAL_MRNA_TRANSLATION	74	0.000	
3	0.976	0.976		0.976	0.976	REACTOME_PEPTIDE_CHAIN_ELONGATION	74	0.001	
4	0.886	0.719	0.719		0.719	REACTOME_REGULATION_OF_BETA_CELL_ DEVELOPMENT	76	0.001	
5	0.863	0.863	0.863	0.863		REACTOME_FORMATION_OF_A_POOL_OF_ FREE 40S SUBUNITS	73	0.001	
1		0.719	0.719	0.000	0.009	SUPERPATHWAY_24_095_0_2	76	0.000	0.95
2	0.381		0.381	0.028	0.000	SUPERPATHWAY_18_095_0_16	133	0.003	
3	0.599	0.599		0.197	0.015	SUPERPATHWAY_10_095_0_6	112	0.032	
4	0.000	0.062	0.281		0.000	SUPERPATHWAY_40_095_0_6	66	0.127	
5	0.008	0.000	0.015	0.000		SUPERPATHWAY_42_095_0_10	84	0.156	
1		0.000	0.028	0.000	0.000	SUPERPATHWAY_9_075_0_16	133	0.005	0.75
2	0.000		0.000	0.000	0.324	SUPERPATHWAY_32_075_0_13	96	0.039	
3	0.061	0.000		0.276	0.000	SUPERPATHWAY_22_075_0_7	67	0.153	
4	0.000	0.000	0.124		0.142	SUPERPATHWAY_17_075_0_15	66	0.231	
5	0.000	0.290	0.000	0.191		SUPERPATHWAY_7_075_0_5	75	0.318	

FIG. 2. Comparison of redundancy among top scoring pathways for original and overlap-controlled versions of pathways. An obesity gene expression dataset was analyzed by gene-set enrichment analysis using Reactome database with different degrees of allowed overlap (100%, 95%, and 75%) between pathways. The degree of pairwise overlap among the top five pathways identified as differentially regulated between obese and lean subjects were calculated and represented as a heatmap. The extent of redundancy between any two pathways is indicated as a fraction of 1 in each colored cell where the two pathways intersect. The column 'Size' refers to the size (gene number) of each pathway. For modified versions of Reactome (Database column values of 0.95 and 0.75), pathways are merged according to overlap criteria to create superpathways. Superpathways are named as 'Superpathway_a_b_c_d', where *a* is an arbitrary numeric assignment of the new pathway, *b* and *c* are the percent maximum and minimum overlap allowed, and *d* is the number of original pathways merged to create the superpathway. For this application, the minimum overlap was set to 0.

Relevance of pathways to obesity

In order to determine if the top-ranked pathways identified from the overlap-controlled databases were biologically more relevant compared to those identified from the original databases, we used the Biograph knowledge-mining tool to obtain a 'proximity' score for each pathway to the 'obesity' phenotype. Taking the top 20 pathways (ranked by nominal gene-set enrichment *p* values), we hypothesized that the pathways identified in Study 3 from the 95_50 overlap-controlled version of Reactome would be significantly more proximal to 'obesity' compared to the top 20 pathways identified from the unaltered version of Reactome. To test this hypothesis, individual proximity scores were generated for each gene member of the top 20 pathways (from both database versions). Pathway proximity scores were derived by taking the average of the proximity scores for all gene members in a pathway. The pathways from both unaltered and overlap-controlled Reactome databases were then ranked by their proximity scores and the ranks compared via the Wilcoxon rank sum test. The difference in ranks was found to be statistically significant, with better ranks observed for the gene-sets derived from the 0.95_0.50 version of Reactome

compared to the original version ($p=0.0489$). This suggests that the pathways identified from the overlap-controlled Reactome are enriched for genes for which there is greater prior evidence of association to the phenotype of obesity. Detailed results of the Biograph analysis and the Wilcoxon rank sum test are provided in Supplementary Table S4.

Discussion

The results of the analysis conducted in this study leads to some significant observations on the influence of pathway redundancies in gene-set enrichment analysis. First, we found that pathway databases can differ dramatically in the extent of redundancies contained in them. The trajectories from Figure 1a show that for the five databases compared, the NCI-PID database is the most divergent, displaying a high proportion of pathways at all levels of redundancies considered. The other pathway databases all display a greater reduction in pathway proportions as the redundancy thresholds are increased, with the greatest reduction observed for the CGP pathway. An analysis of the distribution of percent redundancies among the different databases also demonstrated database-specific effects, such as the high degree of

FIG. 3. Comparison of pathway ranks from gene-set enrichment analysis on original and overlap-controlled pathway databases. The top 20 gene-sets identified by GSEA pathway analysis software are listed for a gene-expression study (Study 1 in text) using either the original or an overlap-controlled version of KEGG (at least 75% overlap between gene-sets). The original pathway ranks are shown under the 'original' column and their corresponding new ranks are shown in the 'overlap-controlled (OC)' column. The movement of ranks between the two analyses is indicated by arrows. The names of the original 20 pathways are listed on the right. Pathways from the original database that were merged into superpathways (SP4 and SP10) are also indicated. The absence of an arrow indicates that the corresponding rank for that pathway did not change between the original and overlap-controlled databases.

inter-pathway overlap in the Reactome and NCI-PID databases compared to others (Fig. 1b). This discrepancy in the degree of overlap among pathways in the different databases is an important finding, since it suggests that the results from pathway analysis can be significantly impacted by the choice of the pathway database analyzed. For example, databases such as NCI-PID or Reactome can give rise to scenarios where the majority of the highly ranked pathways are also highly

redundant. In contrast, databases such as KEGG, CGP, or Biocarta are less likely to be influenced by large redundancies since the majority of pathways involve a smaller degree of overlap.

A comparative gene-set enrichment analysis using original or an overlap-controlled version of the Reactome database demonstrated clear reductions in redundancy among top-scoring pathways when the overlap-controlled database was

TABLE 1. EFFECT OF OVERLAP-CONTROL IN PATHWAY DATABASES ON PATHWAY RANKS

Study	Type	Pathway database	# Pathways considered	P value range in top 20 pathways		# Pathways merged	# Pathways with improved rank	# Pathways with reduced rank in OC
				Original	Overlap controlled (OC)			
1	Gene expression	KEGG	20	0.00–0.008	0.00–0.09	4	14	2
2	GWAS	KEGG	20	0.00–0.03	0.00–0.05	6	8	2
3	Gene expression	Reactome	20	0.00–0.02	0.00–0.19	12	14	3

The changes in ranks for the top 20 pathways are compared between analysis involving an original, or an overlap-controlled (OC) version of pathway databases for three separate studies. Column 1, study id; column 2, type of study; column 3, pathway database used; column 4, number of top pathways considered; column 5, range of nominal *p* values for the top 20 pathways from the original pathway database; column 6, range of nominal *p* values for the top 20 pathways from the OC database; column 7, number of top 20 pathways that were merged into superpathways in the OC database; column 8, number of top 20 pathways with improved ranks in analysis involving the OC database compared to the original database; column 9, number of top 20 pathways with reduced ranks in the OC database analysis. For KEGG, the OC version contained a maximum overlap of 75% and minimum overlap of 0%. For Reactome, the corresponding maximum and minimum overlaps were set at 95% and 50% respectively.

used (Fig. 2). Furthermore, analysis across three separate studies (encompassing both genomic and transcriptomic datasets) consistently demonstrated a general gain in pathway ranks with the overlap-controlled databases (Fig. 3). Generally, we also observed a degradation of p values in the overlap-controlled versions, compared to the original version (' p value' column). However, the small pathway p values observed in the original version of Reactome are likely to be overestimates due to similar sets of genes being repeatedly considered for the top-ranked pathways (preventing the p values from being truly independent of one another). The p values observed in the overlap-controlled versions are more aligned to the assumption of independence among the pathways and are possibly more reflective of the true degree of association of pathways to the phenotypes.

The results from Figure 2 and Table 1 imply some important advantages of overlap control. First, the level of redundancy between the top-scoring pathways can be significantly reduced, especially when using pathway databases with high redundancies such as Reactome or NCI-PID. Second, the overlap-controlled databases allow for more hypotheses (additional pathways) to be tested. In analysis involving the original pathway databases, several of these hypotheses could be missed if pathway selection is based on thresholding of the ranks. For example, if an investigator selected the top 15 pathways from a pathway analysis, then pathways 16–20 from the original version of KEGG would be excluded (Fig. 3, ranks 16–20 under 'original'). However, three of these five pathways would be included in results obtained from the overlap-controlled version of KEGG due to their new ranks being less than 15 (ranks 13–15 under 'overlap-controlled'). As a specific example of how this might affect results, we observed that in Study 2, the 'glycerolipid metabolism' pathway was ranked 19th in the original version of KEGG, whereas the rank improved to 14th in the overlap-controlled version. Previous reports have repeatedly implicated this pathway in several metabolic disorders, including insulin resistance, obesity, type 2 diabetes, and metabolic syndrome (Prentki and Madiraju, 2008). Altered glycerolipid metabolism has also been shown to impact beta-cell compensation and normoglycemia in nondiabetic obese subjects (Nolan et al., 2006), contribute to insulin resistance (Unger et al., 1999), and show impairments in obese Zucker diabetic fatty rats (Oltman et al., 2006). Thus, although 'glycerolipid metabolism' is a very relevant hypothesis for genetic association to obesity in Study 3, it would be missed if only the top 15 pathways were considered for further study following an analysis using the original version of KEGG.

Conclusion

Gene-set analysis is an exploratory and hypotheses-generating tool and it is important to maximize the discovery of diverse hypotheses in the exploratory phase for follow-up and validation in downstream studies. Our studies demonstrate that this goal may not be fully achieved if the analysis is restricted only to the original versions of pathway databases, especially when highly redundant gene-sets occupy higher ranks at the cost of excluding lower ranked pathways that could possibly represent truly novel hypotheses. Overlap control in pathway databases reduces redundancies among top-ranking gene-sets and allows for the expanded discovery of candidate biological mechanisms in omics studies.

Gene-set analysis is widely applied today for the elucidation of biological mechanisms from high-dimensional genomic and genetic datasets. Most of the current methodological research in this area has focused on the development of statistical and bioinformatics tools for the selection and prioritization of pathways of interest. The presence of redundancies in gene-sets is, however, a well-recognized fact and earlier attempts have been made to reduce such redundancies by either limiting the information content [as in GO Slim versions of Gene Ontology, which contain only a subset of the terms of the full ontology (McCarthy et al., 2007)], or by grouping and displaying statistically validated, overlapping functional annotations together [as in the DAVID analysis tool (Huang et al., 2009)]. In this article, we present a complementary approach that preserves all available information in a pathway database but allows a restructuring of the database contents to reduce overlap *prior* to conducting gene-set enrichment analysis.

A potential limitation of the current study is that it is focused on analyzing gene-set contents and has not taken into account the interactions among the genes in such gene-sets. Thus the methods developed are applicable to ORA or FCS based approaches, but their applicability to pathway topology based tools is unknown. The effects of overlap-control on pathway topology clearly remain a future direction of research. Our choice of method was largely governed by the wide availability and application of gene-set-based enrichment tools today compared to pathway topology-based applications. Additionally, pathway topology approaches currently suffer from some of their own limitations including the lack of cell- and treatment-specific topology information and the inability to model dynamic changes in network structure.

In summary, our results demonstrate that the control of overlap in pathway databases improves the results of gene-set analysis by (i) reducing redundancies among the top-scoring pathways, and (ii) allowing for the inclusion of more diverse biological mechanisms as pathways of potential interest. Taken together, this should lead to notable improvements in gene-set analysis outputs from genomic and genetic datasets. The availability of ReCiPa further provides end-users with a simple tool to enhance gene-set enrichment studies.

Acknowledgments

This study was supported by grants from the National Institutes of Health [DK088319, MD000175, HL059868] and American Heart Association [SDG4230068] to S.G; National Institutes of Health [RR017694-06A1, RR022814 and RR11104] to PP; and the Canadian Institutes of Health Research MOP2390941 to RM. The authors also thank Prof. David Banks (Dept. of Statistical Science, Duke University) for helpful discussions during the development of ReCiPa.

Author Disclosure Statement

No competing financial interests exist for any author.

References

- Ackermann M, and Strimmer K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10, 47.

- Bankhead A, 3rd, Sach I, Ni C, et al. (2009). Knowledge based identification of essential signaling from genome-scale siRNA experiments. *BMC Syst Biol* 3, 80.
- Dinu I, Liu Q, Potter JD, et al. (2008). A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data. *Cancer Inform* 6, 357–368.
- Draghici S, Khatri P, Tarca AL, et al. (2007). A systems biology approach for pathway level analysis. *Genome Res* 17, 1537–1545.
- Dumbbill E. (2013). Making sense of Big Data. *Big Data January Preview Issue*, BD1-BD2 (DOI: 10.1089/big.2012.1503).
- Dutta B, Wallqvist A, and Reifman J. (2012). PathNet: A tool for pathway analysis using topological information. *Source Code Biol Med* 7, 10.
- Frishman D, and Hentze MW. (1996). Conservation of aconitase residues revealed by multiple sequence analysis. Implications for structure/function relationships. *Eur J Biochem/EBS* 239, 197–200.
- Gatti D, Barry W, Nobel A, Rusyn I, and Wright F. (2010). Heading down the wrong pathway: On the influence of correlation within gene sets. *BMC Genom* 11, 574.
- Ghosh S, Dent R, Harper ME, Gorman SA, Stuart JS, and McPherson R. (2010). Gene expression profiling in whole blood identifies distinct biological pathways associated with obesity. *BMC Med Genom* 3, 56.
- Goeman JJ, and Buhlmann P. (2007). Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* 23, 980–987.
- Gould TA, Van De Langemheen H, Munoz-Elias EJ, McKinney JD, and Sacchettini JC. (2006). Dual role of isocitrate lyase 1 in the glyoxylate and methylcitrate cycles in *Mycobacterium tuberculosis*. *Mol Microbiol* 61, 940–947.
- Gu Z, Liu J, Cao K, Zhang J, and Wang J. (2012). Centrality-based pathway enrichment: A systematic approach for finding significant pathways dominated by key genes. *BMC Syst Biol* 6, 56.
- Harris MA, Clark J, Ireland A, et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32, D258–261.
- Huang Da W, Sherman BT, and Lempicki RA. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37, 1–13.
- Huang DW, Sherman BT, and Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44–57.
- Ibrahim MA, Jassim S, Cawthorne MA, and Langlands K. (2012). A topology-based score for pathway enrichment. *J Comput Biol* 19, 563–573.
- Khatri P, Sirota M, and Butte AJ. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol* 8, e1002375.
- Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, and Del-Favero J. (2011). BioGraph: Unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol* 12, R57.
- McCarthy FM, Bridges SM, Wang N, et al. (2007). AgBase: A unified resource for functional analysis in agriculture. *Nucleic Acids Res* 35, D599–603.
- Nam D, Kim J, Kim SY, and Kim S. (2010). GSA-SNP: A general approach for gene set analysis of polymorphisms. *Nucleic Acids Res* 38, W749–754.
- Nolan CJ, Leahy JL, Delghingaro-Augusto V, et al. (2006). Beta cell compensation for insulin resistance in Zucker fatty rats: Increased lipolysis and fatty acid signaling. *Diabetologia* 49, 2120–2130.
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, and Kanehisa M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27, 29–34.
- Oltman CL, Richou LL, Davidson EP, Coppey LJ, Lund DD, and Yorek MA. (2006). Progression of coronary and mesenteric vascular dysfunction in Zucker obese and Zucker diabetic fatty rats. *Am J Physiol Heart Circ Physiol* 291, H1780–1787.
- Pemu PE, Anderson L, Gee BE, Ofili EO, and Ghosh S. (2012). Early alterations of the immune transcriptome in cultured progenitor cells from obese African-American women. *Obesity* 20, 1481–1490.
- Prentki M, and Madiraju SR. (2008). Glycerolipid metabolism and signaling in health and disease. *Endocr Rev* 29, 647–676.
- Shi J, Levinson DF, and Whittemore AS. (2008). Significance levels for studies with correlated test statistics. *Biostatistics* 9, 458–466.
- Shojaie A, and Michailidis G. (2009). Analysis of gene sets based on the underlying regulatory network. *J Comput Biol* 16, 407–426.
- Song S, and Black MA. (2008). Microarray-based gene set analysis: A comparison of current methods. *BMC Bioinforma* 9, 502.
- Subramanian A, Tamayo P, Mootha V, et al. (2005a). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102, 15545–15550.
- Subramanian A, Tamayo P, Mootha VK, et al. (2005b). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102, 15545–15550.
- Tarca AL, Draghici S, Khatri P, et al. (2009). A novel signaling pathway impact analysis. *Bioinformatics* 25, 75–82.
- Unger RH, Zhou YT, and Orci L. (1999). Regulation of fatty acid homeostasis in cells: Novel role of leptin. *Proc Natl Acad Sci USA* 96, 2327–2332.
- Vastrik I, D'Eustachio P, Schmidt E, et al. (2007). Reactome: A knowledge base of biologic pathways and processes. *Genome Biol* 8, R39.
- Wang K, Li M, and Bucan M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 81, 1278–1283.

Address correspondence to:

Dr. Sujoy Ghosh

Center for Computational Biology

Dulce-NUS Graduate Medical School

8 College Road

Singapore 169857

E-mail: sujoy.ghosh@duke-nus.edu.sg