A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data

Article *in* Current Proteomics · May 2015

DOI: 10.2174/1570164612011505062009 27

CITATIONS
5

READS
1,188

7 authors, including:

Shi Jing Lu
3 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Mohd Saberi Mohamad
University of Malaysia, Kelantan
203 PUBLICATIONS **835** CITATIONS
SEE PROFILE

Kohbalan Moorthy
Universiti Malaysia Pahang
27 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Safaai bin deris
Universiti Teknologi Malaysia
223 PUBLICATIONS **1,110** CITATIONS
SEE PROFILE

Send Orders for Reprints to reprints@benthamscience.ae

14 *Current Proteomics,* 2015, *12,* 14-27

A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data

Lu Shi Jing[1], Farah Fathiah Muzaffar Shah[1], Mohd Saberi Mohamad[1,*], Kohbalan Moorthy[1], Safaai Deris[1], Zalmiyah Zakaria[1] and Suhaimi Napis[2]

[1]*Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia;* [2]*Faculty of Biotechnology and Biomolecule Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*

M. S. Mohamad

Abstract: The emergence of high throughput data in genomic, proteomic and bioinformatics has increased the necessity to develop large amount of functional analysis tools to facilitate the inference of biologically meaningful data. A promising high throughput strategy, which is the gene-annotation enrichment analysis, can increase the elucidation ability of related biological processes. This paper reviews 35 bioinformatics enrichment tools and 5 gene set databases that are currently available in the field, which include the description of these tools and databases. This information can help tool developers and users to gain a broad view and better understanding on the bioinformatics enrichment tools and databases, enabling a better decision making in choosing tools in particular research interest.

Keywords: Bioinformatics, database, enrichment tools, functional analysis, gene set data, software.

1. INTRODUCTION

The advent of bioinformatics enrichment tools has brought great changes in the analysis of many 'omics' data analysis, including genomic, proteomic and metabolic analysis. The study of these 'omics' data typically involve scanning approaches such as microarray analysis and ChIP-on-CHIPs. These approaches often produce high throughput data or differential expressed genes as final outputs. Unfortunately, these large amounts of data are hardly interpreted and they are still a daunting and challenging task. It is difficult to determine how genes and proteins interact and regulate each other in the biological processes by the high throughput data obtained. Hence, the tools used to analyze and identify the underlying biological processes involved in these data are important and necessary to obtain useful biological meaningful data. Herein, the functional analysis of large gene list, which is derived from the emerging high-throughput 'omics' data, is important to generate and produce more comprehensive and biological relevance information that are useful in the study of disease, drug design, and Gene Ontology term identification.

Over the last few decades, many bioinformatics tools have been developed to assemble the most pertinent and enrich biological important information from these data using biological knowledge accumulated in public databases, such as Gene Ontology [1]). Large numbers of high throughput enrichment tools have been independently developed since 2002 to address the functional analyzing problem of large gene lists. The gene-annotation enrichment analysis is a promising high-throughput strategy that can increase the likelihood for investigators to identify biological processes that are most pertinent to their study. Until now, the development of publicity available tools is still rapidly growing, and this field has been very productive. However, the overwhelming of enrichment tools can cause difficulties for tool-developing group and users to identify the usefulness of all existing works that suit their study.

In 2009, 68 bioinformatics enrichment tools were collected and reviewed [2]. The authors provided a comprehensive collection and classification of enrichment tools available up to year 2008 and reviewed on the pitfall and advantages in a simpler tool-class level. In recent years, more and more enrichment tools have emerged. A review has to be carried to classify these available tools. Approximately, 35 enrichments tools were reviewed in this study. We have tried our best to provide the most accurate details and the rationales behind these tools, which have been collected from year 2008 to current year 2013, including some tools that have been left out in the survey [2]. Besides that, a review on currently available gene set databases is also provided.

2. ENRICHMENT ANALYSIS TOOLS

Enrichment analysis tools are tools that facilitate the analysis of few genes at a time to identify the biological process of organism. It is emerging as an alternative technology that allows the simultaneous measurement of the regulations and changes of genome-wide genes under certain biological condition. These tools usually produce large number of interesting gene lists as final outputs, and the biological interpretation of these gene lists is carried out. Different tools available for the annotation enrichment analysis belong to different classes, which are stand-alone tools, web start tools, web-based tools and toolbox-based tools. Table **1**

*Address correspondence to this author at the Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia; Tel: +60-7-5533153; Fax: + 60-7-5565044; E-mail: saberi@utm.my

1875-6247/15 $58.00+.00

Table 1. List of enrichment analysis tools.

No	Tool	References	Website	Omic Tools
1	ErmineJ	Lee *et al*. (2005)[3]	http://erminej.chibi.ubc.ca/	1
2	EasyGO	Zhou *et al*. (2007)[4]	http://bioinformatics.cau.edu.cn/easygo/	2
3	GeneCodis	Carmona-Saez *et al*. (2007)[5]	http://genecodis.cnb.csic.es/	1
4	CoPub	Frijters *et al*. (2008)[6]	http://services.nbic.nl/copub/portal/	2
5	FindPeaks 3.1	Fejes *et al*. (2008)[7]	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks	1
6	GeneTrailExpress	Keller *et al*. (2008)[8]	http://genetrail.bioinf.uni-sb.de/	1
7	Ontologizer 2.0	Bauer *et al*. (2008)[9]	http://compbio.charite.de/ontologizer	2
8	DIANA-mirPath	Papadopoulus *et al*. (2009)[10]	http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath	3
9	GAGE	Luo *et al*. (2009)[11]	http://www.bioconductor.org/packages/release/bioc/html/gage.html	1
10	IPA	Krämer *et al*. (2013)[12]	http://www.ingenuity.com/products/ipa	3
11	GOrilla	Eden *et al*. (2009)[13]	http://www.webcitation.org/query.php?url=http://cbl-gorilla.cs.technion.ac.il&refdoi=10.1186/1471-2105-10-48	2
12	KEA	Lachmann and Ma'ayan (2009)[14]	http://amp.pharm.mssm.edu/lib/kea.jsp	1
13	SciMiner	Hur *et al*. (2009)[15]	http://141.214.81.219/SciMiner/	2
14	ToppGene	Chen *et al*. (2009)[16]	http://toppgene.cchmc.org	1
15	WhichGenes	Glez-Pena *et al*. (2009) [17]	http://www.whichgenes.org/	1
16	geneXplain	Kolpakov *et al*. (2011) [18]	http://genexplain.com/genexplain-platform-1	3
17	ConceptGen	Sartor *et al*. (2010) [19]	http://conceptgen.ncibi.org/core/conceptGen/index.jsp	1
18	Enrichment Map	Merico *et al*. (2010) [20]	http://baderlab.org/Software/EnrichmentMap/	1
19	GSA-SNP	Nam *et al*. (2010) [21]	http://sourceforge.net/projects/gsa-snp/	1
20	HEAT	Imanishi *et al*. (2010) [22]	http://h-invitational.jp/HEAT/search.do	1
21	i-GSEA4GWAS	Zhang *et al*. (2010)[23]	http://gsea4gwas.psych.ac.cn/	1
22	MAGENTA	Segre *et al*. (2010) [24]	http://www.broadinstitute.org/mpg/magenta/	1
23	MamPhEA	Weng *et al*. (2010) [25]	http://evol.nhri.org.tw/phenome/index.jsp?platform=mmus	1
24	MetPA	Xia and Wishart (2010)[26]	http://metpa.metabolomics.ca/MetPA/faces/Home.jsp	1
25	MSEA	Xia and Wishart (2010) [27]	http://www.msea.ca	1
26	PhenoFam	Paszkowski-Rogacz *et al*. (2010) [28]	http://sourceforge.net/projects/phenofam/	1
27	WebGestalt2	Duncan *et al*. (2010) [29]	http://bioinfo.vanderbilt.edu/webgestalt/	1
28	GARNET	Rho *et al*. (2011) [30]	http://ercsb.ewha.ac.kr/garnet/	1
29	HTSanalyzeR	Wang *et al*. (2011) [31]	http://www.bioconductor.org/packages/2.12/bioc/html/HTSanalyzeR.html	1
30	MBRole	Chaogoyan and Pazos(2011) [32]	http://csbg.cnb.csic.es/mbrole	1
31	MPEA	Kankainen *et al*. (2011) [33]	http://ekhidna.biocenter.helsinki.fi/poxo/mpea/mpea/	1
32	TAFFEL	Kurki *et al*. (2011) [34]	https://bioinformatics.uef.fi/taffel/	1,2
33	GO-Elite	Zambon *et al*. (2012) [35]	http://www.genmapp.org/go_elite/	1,2
34	INRICH	Lee *et al*. (2012) [36]	http://atgu.mgh.harvard.edu/inrich/	1
35	Enrichr	Chen *et al*. (2013) [37]	http://amp.pharm.mssm.edu/Enrichr/	1

shows the list of enrichment tools evaluated in this study. Based on Table **1**, the categories for omic tools are included gene set analysis (GSA), gene ontology (GO), and small RNA-seq analysis.

2.1. ErmineJ

ErmineJ is a multiplatform stand-alone software tool introduced by Lee *et al*. (2005) [3] to analyze sets of gene that are functional-relevant derived from the microarray gene expression data. It aims to identify the underlying important biological pathway existing in the data. ErmineJ applies multiple algorithms for gene set analysis, such as resampling based method and ranked based method. These algorithms mainly focus on gene scores calculation and gene expression profile correlation identification. Besides that, it also applies over-representation analysis. For the graphical user interface, command-line interface and application programming interface are implemented for the automate analysis, which allows the user to script the use of ErmineJ using his own code. For the input of ErmineJ, it supports basic inputs, for example Gene Ontology in XML format which can be extracted from the GO consortium website. The input can also be in the form of expression profile matrix which can be used in clustering tool.

Since May 2006, ErmineJ is released under the Apache 2.0 License. Moreover, ErmineJ has been improved until version 3.0 in June 2013. In term of algorithm improvement, groups were completely redundant with others that being identified in previous versions, these groups are exhaustively identified now in ErmineJ version 3.0.

2.2. EasyGO

EasyGO is a web-based tool that was designed by Zhou *et al*. (2007) [4] to support the gene set functional interpretation of agronomical species. It contributes to the agronomical research by providing robust results visualization capabilities and user friendly interface. Besides that, it supports both farm animals and crops Affymetrix Gene Chips, which include 3 farm animals, 11 agronomical plants as well as a model plant *Arabidopsis*. EasyGO aims to facilitate the discovery of enriched biological knowledge in the field of agriculture and thus provides a better solution for biologist and agricultural scientists. For the result display, EasyGO applies two types of display style, which are text mode (fast tree traversing) and graphic mode. In functional analysis of EasyGO, the GO terms with unbalanced distribution between two groups or probe sets or genes are discovered by comparing a query list consist of all probes sets or genes on a gene chip.

Since May 2010, agriGO is the successor of EasyGO, but there will not be updated at EasyGO analysis function anymore. agriGO refers to a GO analysis toolkit that is available for agriculture community. This tool was developed by Du *et al*. (2010) [38] which used to upgrade the EasyGO tool. agriGO expands the number of supported organism and gene identifiers to 38 agricultural species. Different from EasyGO, agriGO improves the system architecture and website interface to improve its accessibility and performance. More flexible user inputs such as user-defined annotation and references are acceptable. There are four new Gene Set Enrichment Analysis (GSEA) approaches integrated in this tool to meet different demands, which are PAGE, SEA, SEACOMPARE and BLAST4ID. For the output, users can visualize and compare the different data sets in visual which is hierarchical tree graph. The inter-relationship between terms is represented in the form of direct acyclic graph.

2.3. GeneCodis

GeneCodis is a web-based tool used to discover significant concurrent annotations in gene list, developed by Carmona-Saez *et al*. (2007) [5]. It applies the *apriori* algorithm to discover the relationship among gene annotation and gene expression patterns. GeneCodis can perform modular and singular enrichment analysis at the same time, providing a fast and user friendly ways to extract biological relevant data. Besides that, it also allows the integrated analysis of annotation from KEGG pathways, GO, InterPro motifs and Swiss-Prot. In the analysis, statistical rank score for each gene is generated. The user can also combine the analysis with the annotation from sources listed above. For the functional analysis of gene list, it could outperform other available standard methods. This is because it provides significant information for the biological interpretation of high-throughput experiments within the analysis of concurrent annotations.

GeneCodis is also further improved by Nogales-Cadenas *et al.* (2009) [39], which aimed to perform functional analysis and interpretation of gene lists through integration of enrichment analysis and diverse biological information. Besides that, it also identifies the modular patterns of interrelated user defined annotation. This application provides different functional annotations, compared to the traditional annotation in which the regulatory patterns and user-defined annotation are included in the functional information. This provides the opportunity of the integration of all sources of information into a single analysis. In addition, it is one of the few tools that integrate different aspects of the biology of the genes in analyzing and offer concurrent annotation enrichment analysis. GeneCodis is accessible through SOAP web services interface, enabling users to use their own workflows and scripts to perform analysis. Furthermore, GeneCodis has been improved by Tabas-Madrid *et al*. (2012) to GeneCodis3 that able to remove noisy and redundant output from the enrichment inclusion of a reported algorithm [40].

2.4. CoPub

CoPub is a literature-based keyword enrichment analysis tool for microarray data analysis, developed by Frijters *et al*. (2008) [6]. It uses the information in Medline database to perform microarray data interpretation. There are a total of eleven thesauri that had been extracted to search Medline database information in the previous study. These searching are based on the biological items that represent the biological concept instances, such as a pathway or a gene. It describes GO terms, drugs, genes, tissues, diseases, liver pathologies and pathways. CoPub provides multiple input of human and mouse genes and generates keyword lists from several biomedical thesauri. The correlated keywords between input genes and lists link to Medline are highlighted. Besides that, CoPub supports graphical visualization of differential ex-

pressed genes and over-represented keywords in network. It gives a better insight to the correlation between genes and keywords. In addition, it provides three analysis methods such as the BioConcept search, the Gene search, as well as the microarray data analysis. CoPub has been improved to CoPub 5.0 that consists of a complete new interface to detect new hidden relations.

2.5. FindPeaks 3.1

FindPeaks 3.1 was developed by Fejes *et al*. (2008) [7], which used to identify the enrichment areas of massively parallel short-read sequencing technology. The short-read sequencing is a next-generation sequencing that discovers the protein-DNA association events and is widely applied in genomic as well as meta-genomics research. The advent of next-generation sequencing gives a better insight into genome-wide scale protein-DNA related events, and is helpful in genomics as well as meta-genomics research. FindPeaks 3.1 is useful for generating UCSC compatible custom 'WIG' track files from aligned-read files for short-read sequencing technology. It is a software application which is executable or available on any platform that is capable of running Java Runtime Environment (JRE). In general, more memory can analyze larger number of sequencing read. In practical, 40 million reads require a 4GB memory to process. Besides that, it is also known as coherent and simplified tool architecture that can manage to handle large development of new features in enrichment analysis. FindPeaks has been updated to version 3.2 under the GPL at SourceForge and being maintained the analysis functionality.

2.6. GeneTrailExpress

GeneTrailExpress is a web-based toolbox developed by Keller *et al*. (2008) [8]. It is a tool that can perform analyzing, normalizing and visualizing of gene expression data or profiles in a complex multi-step process. Besides that, the GeneTrailExpress tool can provide powerful statistical analysis and standard normalization procedures which facilitate the study of multi-variety of pathways and categories. In addition, it also provides the integration of graph visualization tool called BiNA, which enables the visualization of relevant biological pathways. This BiNA applies the cutting-edge graph-layout algorithm which guides the drawing of biological pathways. For the functional analysis, it provides all steps of the microarray evaluation pipeline, including gene set selection, normalization, data retrieval and gene scoring. The biochemical data such as metabolic and regulatory pathways can be obtained from heterogeneous databases.

2.7. Ontologizer 2.0

Ontologizer 2.0 is a Java application developed by Bauer *et al*. (2008) [9] to analyze the overrepresentation of Gene Ontology (GO) terms in gene sets statistically. Three statistical analysis algorithms are implemented, such as topology-based algorithm, one-sided Fisher's exact test and novel parent-child method. Ontologizer 2.0 also provides the multi-testing correction procedures on gene set analysis. Besides that, GraphViz (Graph visualization) tool can be integrated to explore the relationship between the data by linking individual GO terms to child terms, as well as visualize the significantly overrepresented GO terms. This Java application requires a JAVA SE 5.0 compliant Java Runtime Engine (JRE) to run. The input of the program is in OBO file format that defines the GO structure and association file. Besides that, it also supports the annotation files in Affymetrix' NetAffx Analysis Center. The annotation files are then converted into other gene names by providing a simple mapping text file.

2.8. DIANA-mirPath

DNA Intelligent Analysis (DIANA)-mirPath is a web-based tool for miRNA pathway analysis that developed by Papadopoulos *et al*. (2009) [10], which is used to analyze the human and mouse microRNAs (miRNAs) data. Strong evidences have shown that the microRNA data has the ability to modulate a molecular pathway by converting multiple target genes. Hence, this tool plays an important role in discovering the combinatorial effect of co-expressed microRNA in the modulation of pathways. The microRNA is graphically annotated (pathway map) as it is implicated in a pathway. In addition, it is used to discover the molecular pathways that may be altered by the microRNAs expression. The enrichment analysis will then be carried out on multiple sets of microRNAs targets to pathways available in Kyoto Encyclopedia of Genes and Genomes Database (KEGG). Besides that, it also provides graphical view outputs that illustrate the overview of those pathways, modulated by microRNAs. This presentation of result output helps to interpret the analysis result more efficiently. DIANA-mirPath v2.0 acts as the new version of DIANA-mirPath web server has been mentioned by Vlachos *et al.* (2012) [41]. All the statistical and analysis algorithms for DIANA-mirPath v2.0 are implemented in R language [41]. Besides, DIANA-mirPath v2.0 can also perform advanced analysis pipelines and allowed users to easily create heat maps of microRNAs against pathways interactions.

2.9. GAGE

Generally Applicable Gene-set Enrichment (GAGE) is an R-package tool that developed by Luo *et al*. (2009) [11], which used to solve the limitation of Gene Set Analysis (GSA) that analyzes gene expression data by using pathway knowledge. The GSA cannot analyze datasets with different experimental designs or sample sizes. Here, GAGE has the advantage of handling multiple microarrays datasets with different experimental designs, sample sizes, as well as profiling techniques. In this study, the GAGE outperformed the traditionally used GSA methods, which are PAGE and GSEA. The improvement was demonstrated in three aspects specificity and sensitivity, repeated experiment consistency and regulatory mechanism biological relevance. The GAGE gives a significant impact on deriving the gene sets and separating them into pathway.

2.10. IPA

Ingenuity Pathway Analysis (IPA) is a web-based functional analysis tool for comprehensive 'omics' data, such as RNAseq, small RNAseq, and so on [12], which released in 2007. Moreover, IPA acts as the tool that widely used by the

life science community. It consists of extensive library of well-characterized signaling and metabolic pathways. And also, IPA consists of extensive knowledge database on model organisms, such as human, mouse, rat, and canine. Besides, IPA also presented with direct import of gene lists from microarray analysis software, for example, Partek Genomics Suite (PGS). Krämer *et al.* (2013) presented four causal analytics algorithms that have been implemented in IPA, which are Upstream Regulator Analysis (URA), Mechanistic Networks (MN), Causal Network Analysis (CAN), and Downstream Effects Analysis (DEA) [12]. All such algorithms operate over a large scale causal graph that ensemble from Ingenuity Knowledge Base [12]. Furthermore, IPA can also create customized pathways for targets, biomarkers, biological functions, and diseases of interest.

2.11. GOrilla

GOrilla was developed by Eden *et al*. (2009) [13], which is a web-based tool that discovers and visualizes the enriched GO terms in ranked gene lists without the pre-set of background sets and explicit target. This gives advantage in situation where genomic data are naturally represented as a ranked gene list. *GOrilla* analyzes data statistically using flexible threshold approach. The p-value of observed enrichment genes are calculated in *GOrilla* using mHG, a complete characterization of underlying distribution. It enables in-depth analysis of thousands of genes and GO terms statistically in order of seconds. The outputs are visualized in hierarchical structure, enabling a clearer overview of enriched GO terms correlation. It implements a standard approach to identify enriched Gene Ontology terms, which is hyper geometric distribution, in case the ranking list inside a background set is ignored. The advantage of this tool is that it automatically removes the supplicated gene in the lists, which will cause biased result.

2.12. KEA

KEA is also known as Kinase Enrichment Analysis, which is a web-based application developed by Lachmann and Ma'ayan (2009) [14], intended to analyze kinase-substrate interactions that infer kinase list that is related with specific gene or protein list. Kinases are ranked based on likelihood of them of being functionally related to regulated cell, under different experimental conditions. This tool provides a platform for users to link lists of mammalian genes and proteins with kinases that phosphorylate them through the integration of underlying kinase-substrate databases. Kinase enrichment probability is calculated based on the distribution of kinase-substrate proportion in kinase-substrate databases to the kinase is that found related to list of genes or proteins. The input of this tool is a list of gene symbols. The input is further validated by removing the input entries that do not match a substrate in consolidated background kinase-substrate dataset.

2.13. SciMiner

SciMiner was developed by Hur *et al*. (2009) [15] for functional enrichment analysis and target identification literature mining. It is a web-based application that is used to identify genes and proteins in MEDLINE abstracts and full texts specific analysis. Functional enrichment analyses are used to identify Medical Subject Headings (MeSH) terms, highly relevant targets (genes and proteins), protein–protein interaction networks and pathways, as well as Gene Ontology (GO) terms by comparing identified targets from one search result with those from other searches or to the full HGNC [HUGO (Human Genome Organization) Gene Nomenclature Committee] gene set. Besides that, it provides a filter and manual correction for users to increase the accuracy of the tool. The filter is used to select the list to be included in the input for further analysis, while the manual correction means that the user is able to manually modify the mining result of identified target. However, the new version of SciMiner has been updated in 2010, which is available to use in Ubuntu. The analysis modules of SciMiner are including gene enrichment, Gene Ontology enrichment, MeSH term enrichment, pathway enrichment, and protein-protein interaction network of targets. In 2012, the post mining analysis module of SciMiner has been updated too, which is done by comparing target set with background set.

2.14. ToppGene

ToppGene Suite is a freely accessible web site that provides gene list enrichment analysis, candidate gene prioritization and novel disease candidate genes prioritization and identification [16]. Moreover, ToppGene, ToppFun, ToppNet, and ToppGenet have been developed by Chen *et al*. (2009) [16], but ToppGene is the main tool to be focused here. ToppGene applies fuzzy-based similarity measure on the prioritization of functional-annotation on disease candidate genes, using statistical meta-analysis. Random sampling of whole genome is used in this tool. Besides that, some of the Web network analysis algorithms, such as K-Step Markov, HITS algorithm and PageRank extended version are also implemented to perform prioritization of protein-protein interaction network based disease candidate genes. A representative profile is then generated by ToppGene and the over-representative terms are identified from the training genes. The comparison between test gen set and the representative profile of training set are then carried out. Then the expression values of average vector will be calculated.

2.15. WhichGenes

WhichGenes is a web-based application developed by Glez-Pena *et al*. (2009) [17] that performs gene set enrichment analysis. Gene sets are gathered, built, stored and exported in the enrichment analysis. Besides that, this interactive gene set building tool provides a simple interface in extracting the gene lists from unstructured biological data sources, as well as multiple databases. In addition, this tool provides four-step wizards that enable several queries to be run in parallel to help user specify new interest gene sets. Lastly, WhichGenes allows users to modify the functionalities of the tool by using a web service called Representational State Transfer. It also provides function such as generating new gene sets through the combination with existing gene sets as well as renaming, editing and deleting the existing gene sets. For the output, users are allowed to export configuring sets of the output format and choose among multiple gene identifiers.

2.16. geneXplain

The geneXplain is an online toolbox that applied in bioinformatics and system biology. Moreover, geneXplain acts as a platform of the BioUML framework that developed by Kolpakov *et al* (2011) for causal interpretation of data coming from microarray, proteomics, miRNA, and ChIP-chip/seq experiments [18]. This platform geneXplain consists of genome analysis, network analysis, and compound analysis. Besides, geneXplain also applied upstream analysis approach for implementation of machine learning and graph topological analysis algorithms [18]. The algorithms applied to identify causality key nodes in the network of gene regulation and signal transduction [18]. However, geneXplain platform developed by Valeev *et al* (2011) for system medicine studies [42]. Furthermore, geneXplain has been updated to version 3.0 with integration of numerous genome tracks that visualized by the in-built genome browser.

2.17. ConceptGen

ConceptGen was developed by Sartor *et al*. (2010) [19] to perform gene set enrichment analysis and gene set relation mapping. It is a user-friendly web-based application which elucidates biological enriched concepts in differential expressed genes. *ConceptGen* explores the relationships of predefined biological concepts from different information sources and visualizes the results using multiple perspectives. There are a total of over 20000 concepts, 14 different types of biological knowledge included in this tools. Statistical method is applied to perform a gene expression analysis. The input can be obtained from various genomic resource centers, which is then converted into NCBI Entrez Gene IDs. A gene expression analysis pipeline is developed in which human Affymetrix experiments in GEO populate the concept type. Next, the raw data undergoes pre-processing with a package and is normalized. Finally, the quality of the data is controlled using an empirical Bayes method to test for differentially expressed gene.

2.18. Enrichment Map

Enrichment Map is a network-based method to perform interpretation and visualization of enriched gene sets, developed by Merico *et al.* (2010) [20]. It aims to address the limitation of current enrichment analysis tool that hardly handles large number of redundancy gene sets. In this tool, gene sets are organized in a network; nodes are referred to gene set while edges represent the gene overlap between sets. Results are visualized in automated network layout, and related gene sets are grouped into clusters. This provides a clear and quick view on major enriched functional themes and enables an easier interpretation of the results. In particular, the results are mapped into Enrichment Map. The node size denotes the number of genes in gene sets, which are computed using overlap or Jaccard coefficient. The node is color reducing the enrichment score. The node colors range from white (which indicates no enrichment) to red (which indicates high enrichment). In 2011, Enrichment Map has been updated to v1.2 release that compatible with the Cytoscape 2.6, 2.7, and 2.8 plugin.

2.19. GSA-SNP

GSA-SNP is a gene set analysis standalone tools developed by Nam *et al*. (2010) [21]. It implements three GSA methods in this tool, which are Z-statistic method, standardized GSA and GSEA. Two types of input data marker association (p-value) and gene sets are required in this tool. The output result contains sorted gene sets with p-values and association strength in descending order, displayed in program windows as well as a CSV file. This study mostly focuses on genome-wide association (GWA) study, which aims to discover the association of genetic factors to the traits of interest. There are two types of input data required, genes sets and marker association. For the gene sets, the Gene Ontology terms are set as default analysis. The format of the input is in gmt format of MSigDB. For the marker association, the p-value represents SNP association levels with given phenotype. Besides that, it supports high level of significance from a small number of simulated analyses that are able to pool the randomized gene set scores.

2.20. HEAT

HEAT is also known as H-InvBD (H-Invitational Database) Enrichment Analysis Tool, developed by Imanishi *et al*. (2010) [22]. It performs gene set enrichment analysis by using comprehensive annotation of human genes in H-InvDB. With given a human gene set, it is able to identify specific corresponding features. It is a web based-application concerned with identification of particular features to given human gene set. It applies gene set enrichment analysis (GSEA) and Fisher's exact statistical test to perform the gene enrichment analysis. This data mining tool searches the most significant enriched user-defined gene sets and compare them to the H-InvDB representative transcript to obtain the H-InvDB annotations. The input of the analysis has to be in two or more human gen IDs. List of acceptable IDs can be checked on HEAT websites, as given in Table **1**. Next, the submitted IDs are converted into Transcripts IDs (HIT). The result output consists of feature ID, feature name, number of occurrences of genes among H-InvDB representative transcript, number of occurrences of genes in a gene set as well as P-values. The results are then sorted by the P-values. Furthermore, HEAT has been updated to version 3.1 in 2011.

2.21. *i*-GSEA4GWAS

i-GSEA4GWAS (*improved* GSEA for GWAS) is a web-based application that performs gene set enrichment analysis to genome-wide association study (GWAS). It was developed by Zhang *et al*. (2010) [23], mainly to identify the genes or pathways associated with specific traits or phenotypes especially in human disease. This tool implements SNP label permutation in Gene Set Enrichment Analysis (GSEA) which is useful in identifying the correlation between traits and genes or pathways. Besides that, this tool also focuses on the high proportion of significant genes in pathways/gene sets. It is an *improved* version of GSEA in GWAS which manages to provide new insight to complex disease study. The input data of the tool are SNP data mainly in P-values, which also accepts odds ratio and statistic. Two columns are needed in input format of SNP list, which are SNP identifier and their corresponding association value.

Besides that, the utilization of gene sets and pathway in this tool are extracted from variety of online gene set databases, such as KEGG, BioCarta, and GO terms with high confidence.

2.22. MAGENTA

MAGENTA or Meta-Analysis Gene Set Enrichment of variant Associations was developed by Segre *et al*. (2010) [24] to identify enriched functional related genes that are associated with polygenic trait or disease. It was initially founded to test the hypothesis that searched whether mitochondrial dysfunction is a primary cause of diabetes with the possibility that mitochondrial gene has modest genetic effects that influence the T2D risk. It aims to improve the performance of meta-analyses to large genome-wide association study (GWAS) statistically whose individual genotypes are not available. This can be done by combining the variant association p-values into gene scores and correcting the confounders, for example, linkage disequilibrium properties, variant number, and gene size. MAGENTA has been found successful in detecting association that is likely missed by single-marker analysis. In 2011, MAGENTA has been updated to version 2.4 that able to run on genome build 37 (hg19) or 36 (hg18).

2.23. MamPhEA

Mammalian Phenotype Enrichment Analysis (MamPhEA) was developed by Weng *et al*. (2010) [25]. It is a web-based application used to perform functional analysis of mammalian gene sets based on mouse mutant phenotypes. The study on most mouse mutants have been shown to be useful in the study of human disease progression, human genetics and physiology due to the similarities found between human gene and mouse gene. Mouse mutants have 9000 genes, which have been resembled for human clinical examinations. The knowledge on mutants strain aids the understanding on functional genes at system level. Besides that, it enables the enrichment analysis on user-defined or predefined phenotypes, giving a flexible option to specify phenotypes. In addition, it provides comprehensible results and support analysis on all mammalian species genes which are fully sequenced. In 2011, MamPhEA has been updated with MGI (Mouse Genome Informatics database resource) and its predefined phenotypes in version 4.41, and also include Ensembl version 63.

2.24. MetPA

MetPA is also known as Metabolomics Pathway Analysis, developed by Xia and Wishart (2010) [26] to perform analysis and visualization of metabolomics data in the biological context of metabolic pathways. It identifies the most relevant metabolic pathways by combining pathway topological characteristic and several advanced pathway enrichment analysis procedures. The results are the output in a network based visualization format that supports interactive and intuitive data exploration, such as dragging, point-and-click as well as lose less zooming. In addition, features such as conversion of metabolite name and automatic analysis report generation are provided by adding a comprehensive compound library. Besides that, it also implements multiple of univariate statistical procedures, comprehensive compound library for conversion of metabolite name and analysis report automatic generation. In 2012, MetPA has been updated with new pathway library for *Gallus gallus* (chicken) and *Mesorhizobium loti* (Gram negative species of bacteria).

2.25. MSEA

Metabolite Set Enrichment Analysis (MSEA) is a web-based application used to identify biologically meaningful patterns in quantitative metabolomics data. This application was developed by Xia and Wishart (2010) [27], which aimed to help researchers to discover and interpret human or mammalian metabolite concentration changes. The MSEA has been created a library which consists of approximately 1000 predefined metabolite sets. These metabolite sets cover various tissue locations, bio-fluids, diseases states and metabolic pathways. For more specific analysis, MSEA also supports custom or user defined metabolite sets. Three enrichment analysis methods are implemented in this tool, which are quantitative enrichment analysis (QEA), single sample profiling (SSP) and overrepresentation analysis (ORA). For the analysis using ORA, it only requires compound name, while the other two analysis using QEA and SSP require the compound names and its concentrations. The results are displayed in the form of tables and graphs embedded with hyperlinks to relevant pathway disease descriptors and images. In 2011, MSEA has been updated with friendly user interface that user can upload a reference metabolome in the enrichment analysis and also choose the tab.

2.26. PhenoFam

PhenoFam was developed by Paszkowski-Rogacz *et al.* (2010) [28] to carry out gene set enrichment analysis. Besides, PhenoFam is a Java web application that runs on Tomcat 5.5 server and uses MySQL database to store mappings between various proteins. It is done by utilizing structural and also functional data on the domains of protein families. It acts as annotation terms. This tool is aimed to examine the complete sets of results. It is done from quantitative high-throughput fields of study such as gene expression microarrays without any prior pre-filtering. It is also used to link a list of user-provided identifiers associated with protein features from the database of InterPro. It also evaluates whether the results connected with the domains of individual is different from the population overall and analyzed results in a genome-wide RNA interference screen. It also investigates the purpose of plexins which comprises the domain of cytoplasmic RasGAP. For the input of the tool, it accepts a wide range of identifications that are integrated in the Ensemble database. Next, the identifier mapping is carried out in order to perform analysis.

2.27. WebGestalt2

WebGestalt2 was developed by Duncan *et al*. (2010) [29], which is an updated and expanded version of Web-based Gene set enrichment analysis toolkit. Several updates have been done in this study, for example the organism used, supported input ID types, output result display, enrichment analysis statistic used and functional categories coverage.

WebGestalt2 integrates different public resources information and provides large gene sets analysis for biologists to discover important biological information. Besides that, it also integrates information from KEGG pathway and Gene Ontology term. Several methods are also added for the multiple adjustments. The original version of WebGestalt only supports human and mouse organism, but this upgraded version is expanded to cover organisms such as worm, yeast, zebra fish, rat, fly and dog. Moreover, it also supports more input ID types extracted from different databases as well as different technology platforms. In 2013, WebGestalt2 has been updated with friendly user interface that users can export enriched network modules. Other than that, WebGestalt2 has further increased the coverage of gene identifiers and function categories in various biological contexts in 2013 [43].

2.28. GARNET

Gene Annotation Relationship NEtwork Tools (GARNET) was developed by Rho *et al*. (2011) [30] to perform gene set analysis. It consists of many novel features. It includes tools for reclamation of genes from database of annotation, analysis of statistical and also the image of relationship of annotation. It also maintains the gene sets. In an attempt to permit the access to a full spectrum of knowledge on biology, it has been combined with a diversity of data annotation which consists of GO, disease, drug, domain, location of chromosomal and annotations of custom-defined. This tool also includes molecular networks such as microRNA regulations, pathways, protein-protein interaction and transcription regulations. By using kappa statistic, the relationship between annotated gene sets is then computed. For the exploration of the related annotations, a dedicated viewer is constructed for network annotation. GARNET also allowed users to expand the gene list by using molecular networks, like pathways from KEGG or BioCarta, protein-protein interaction from NCBI integrated database, and microRNA regulation.

2.29. HTSanalyzeR

HTSanalyzeR is an R/Biconductor package for integrated network analysis of high-throughput screens (HTS), developed by Wang *et al*. (2011) [31]. It is efficient software used to make combination of pipelines analysis for HTS data. Three types of analysis are implemented in this tool, which are rich sub-network identification, comparative gene set analysis, gene set enrichment analysis, and overrepresentation analysis. *HTSanalyzeR* deals with commonly used packages of pre-processing for HTS data. The results are displayed in the form of HTML pages and also network plots using enrichment map. The genes sets are extracted from various databases such as KEGG, Gene Ontology and MSigDB. Two approaches are included in gene set analysis, such as gene set enrichment analysis and hyper geometric test statistic. These two approaches are used to measure concordant trend of gene set to other stronger types and to identify overlaps between gene sets and hits respectively. For the network analysis, the rich sub networks can be identified by using *BioNet* package which is performed heuristically to produce near optimal results. Besides, *HTSanalyzeR* has been updated to version 2.10.0 in 2012.

2.30. MBRole

Metabolite Biological Role (MBRole) was developed by Chagoyen and Pazos (2011) [32], which aims to perform enrichment analysis on metabolomics data. It is a web server for the implementation of over-representation analysis of the biological and also the chemical annotations in arbitrary sets of metabolites (small chemical compounds) coming from metabolomics data of any organism or sample. It was initially proposed when there was almost no existing tool for the utilization of enrichment analysis of transcriptomic and proteomic data. However, in order to interpret them in biological terms, almost no equivalent tools exist for metabolomics data. The input of the tools is the identification of data from any of the databases or in SMILES (simplified molecular input line entry system) format. In order to perform the enrichment analysis, the user has to define the set of annotations and the background set of compounds. For statistical analysis, cumulative hyper geometric distribution is used. In addition, the false discovery rate is calculated in order to correct the result of multiple testing.

2.31. MPEA

MPEA is also known as Metabolite Pathway Enrichment Analysis, developed by Kankainen *et al*. (2011) [33]. It is used to interpret the image of biological metabolite data at the system level. The tool complies with the awareness of gene set enrichment analysis (GSEA). It also examines whether the metabolites are needed in some of the predefined pathways, which take place for the top of a query compound list. MPEA is aimed to hold the relationships of many-to-many which may happen between the compounds of query and annotations of metabolite. The metabolite profiles are examined through 14 twin pairs between different body weights. It requires a list of pre-annotated compounds or GC-MS-based MSTs and has two modes of enrichment analysis: single set and iterative. For the single set mode, it is responsible for unranked list while the iterative mode is responsible for ranked entries. For statistical analysis, hyper geometric distribution is coupled with permutation test to solve the pathway inconsistencies.

2.32. TAFFEL

Kurki *et al.* (2011) [34] proposed a method called Independent Enrichment Analysis (IEA) with software TAFFEL. Besides, TAFFEL is a Java Web Start application using Java Standard Edition 6 with NetBeans integrated development environment and MySQL database to store all the persistent data. By utilizing Gene Ontology categories and also regulators of transcription, TAFFEL makes the task easy by clustering genes into subgroups. IEA suggests transcriptional regulators which putatively controls the functions of biological in the studied condition developed method. It gives an explorative analysis on data through three main steps. First, the functional annotation from GO is used to separated differentially expressed genes into functionally homogenous gene groups. Secondly, by using the annotations from TFBS, the genes with similar *cis*-regulatory transcription factor binding side in their binding region are discovered. Thirdly, an Independent Enrichment Analysis (IEA) is applied to evaluate the TFs enrichment of GO terms in gene homoge-

nous TF annotations and vice versa. This tool gives a new approach to the differentially expressed genes analysis and can generate important hypotheses. The statistical testing of enrichment in this tool is Fisher's exact test and only the cluster whose annotation comes with occurrences is used in testing.

2.33. Go-Elite

GO-Elite pathway analysis was developed by Zambon *et al*. (2012) [35] to provide a flexible solution for pathway and ontology over representation. It is a powerful pathway analysis tool in the scope of wide array of identifiers (IDs), species pathways, ontologies and gene sets. In addition to the Gene Ontology (GO), GO-Elite allowed the user to carry out over-representation analysis (ORA) on any annotations of ontology, even in pathway database or biological IDs. The GO-Elite feats the structured nature of biological ontologies, was aimed to report a minimal set of non-overlapping terms. This tool supports various kinds of ID systems, covering disease, gene, multiple pathway databases, transcription factors, microRNA targets, biomarkers and phenotype ontologies. Besides that, by using Go-Elite, users will be able to create their own databases which include features such as ID systems, species and relationships. The results can be viewed in the networks on WikiPathways. Besides, GO-Elite release version 1.2.3 that available as stand-alone compiled versions for Windows and Mac OSX. Moreover, GO-Elite have also been updated to version 1.2.5 in 2012 with add-on the option for automated visualization of enriched pathways.

2.34. INRICH

Interval-based 22nrichment analysis (INRICH) was developed by Lee *et al*. (2012) [36] to perform interval-based enrichment analysis for genome-wide association studies. It performs the tests for signals of the association of the enriched of predefined gene-sets across independent genomic intervals. The advantage of INRICH is that it has a broad applicability, quick running time and most crucial is the robustness to possible biases of genomic and factors of confounding. Some of the factors include variation in size of gene and density of single-nucleotide polymorphism, disequilibrium of linkage within and between genes. Also, genes overlapping with similar annotations are always not described by existing methods of gene-set enrichment. In INRICH, the disease-associated genomic intervals are used as input such as GWAS SNPs. Besides that, the genes that belong to the same gene sets are merged in order to avoid biased result due to the existence of the duplicated multicounting physically clustered genes.

2.35. Enrichr

Enrichr is an integrative web-based and also application of mobile software, developed by Chen *et al*. (2013) [37]. It includes libraries of new gene-set which is an alternative ways to rank the enriched terms. Also, there are lots of approaches containing interactive visualization to show the enrichment results utilizing the library of JavaScript, Data Driven Documents (D3). This tool can be integrated into any kinds of tools that perform the analysis of gene list. Enrich had been applied to examine nine cancer cell lines. It was done by comparing their signatures of enrichment to the signatures of enrichment that matched the normal tissues. The observation was done to a common pattern of up regulation of the polycomb group PRC2. This was also done for the enrichment of the histone mark H3K27me3 in various cancer cell lines, as well as the alterations in Toll-like receptor and interleukin signaling in K562 cells. It was then compared with normal myeloid CD33+ cells. Table **2** presents the comparison table between application types and key statistical methods used in enrichment analysis tools.

3. DATABASES FOR GENE SETS

Several gene set databases are available for use to perform enrichment analysis. Information in the databases is freely accessible. Gene Set Database provides two analysis functions which are gene or gene set search and gene enrichment analysis. By using this gene set search function, users can obtain gene sets that include the gene(s) or biological term(s) they are interested in. By using the gene enrichment analysis function, users can upload their own gene list and assess the list's statistical overrepresentation in gene sets. Although genome wide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Some of the databases can be integrated into the tool available for enrichment tools, to aid the discovery of biological important knowledge in gene expressed data. Table **3** shows the list of Gene Set Databases.

3.1. KEGG Pathway

KEGG was established by Ogata *et al.* (1999) [44] under the Japanese Human Genome project. KEGG pathway is also known as Kyoto Encyclopedia of Genes and Genomes (Kanehisa *et al*., 2011) [49], which is a resource database that combines genomic, chemical and also the information of systemic functional. The gene catalogs from genomes, which are completely sequenced, are connected to higher-level systemic functions of a cell which include the organism and ecosystem. There have been major efforts done to create a knowledge base for such systemic functions, done by capturing and also coordinating experimental knowledge in the form of computable which is namely, pathway maps of KEGG, the hierarchies of BRITE functional and modules of KEGG.

Besides that, KEGG was used by Moutselos *et al*. (2009) [50], applied as KEGG converter of the KEGG Pathways database which focused in the area of in-silicon modeling of metabolic networks. Thus, in this field, the KEGG converter is applied as a web-based application. It is utilized as source of KGML files for the purpose of constructing integrated pathway SBML models which are fully functional for the simulation process.

3.2. Reactome Pathway

Reactome Pathway was established by Joshi-Tope *et al.* (2005) [45]. Reactome Pathway database was updated by Croft *et al*. (2010) [51] at the Ontario Institute for Cancer Research in the Laboratory of Cold Spring Harbor, located in the New York University in School of Medicine and The Institute of European Bioinformatics, in order to build an open source which curettes the database of bioinformatics of

Table 2. Comparison between application types and key statistical methods used in enrichment analysis tools.

No	Name	Year	Application Type	Platforms	Key Statistical Method	Visualization
1	ErmineJ	2005	Webstart	All	T-test and one way ANOVA	Yes
2	EasyGO	2007	Web-based	All	Binomial test, $\chi 2$ test, and Hypergeometric test	Yes
3	GeneCodis	2007	Web-based	All	Hypergeometric distribution and the $\chi 2$ test of independence	Yes
4	CoPub	2008	Web-based	All	R Statistics package	Yes
5	FindPeaks 3.1	2008	Standalone	Windows, Mac OS X	False discovery rate, fragment length distribution	Yes
6	GeneTrailExpress	2008	Web-based	All	Mean fold-change, Median fold-change, Unpaired t-test, paired t-test, Wilcoxon Mann-Whitney test, ANOVA, and Wilcoxon Rank-Sum test	Yes
7	Ontologizer 2.0	2008	Webstart	All	One-sided Fisher's exact test	Yes
8	DIANA-mirPath	2009	Web-based	All	Pearson's chi-squared test	Yes
9	GAGE	2009	Toolbox-based	Windows Binary, Mac OS X 10.6	T-test, Global p-value, FDR q-value	No
10	IPA	2007	Web-based	All	Right-tailed Fisher Exact test	Yes
11	Gorilla	2009	Web-based	All	mHG	Yes
12	KEA	2009	Web-based	All	Fisher Exact Test	Yes
13	SciMiner	2009	Web-based	All	Confidence scoring scheme	Yes
14	ToppGene	2009	Web-based	All	Fisher's inverse chi-square	No
15	Which Genes	2009	Web-based	All	Kolmogorov–Smirnov test	Yes
16	geneXplain	2011	Web-based	All	Hypergeometric, correlation and polynomial re-gression	Yes
17	ConceptGen	2010	Web-based	All	Modified Fisher's exact test	Yes
18	Enrichment Map	2010	Toolbox-based	All	Ratio of class means	Yes
19	GSA-SNP	2010	Standalone	All	Z-statistic	No
20	HEAT	2010	Web-based	All	Fishers' Exact test	No
21	i-GSEA4GWAS	2010	Web-based	All	Kolmogorov–Smirnov test	No
22	MAGENTA	2010	Standalone	All	One-tailed Mann-Whitney rank-sum test	No
23	MamPhEA	2010	Web-based	All	Fishers' exact test	No
24	MetPA	2010	Web-based	All	Hypergeometric test, Fishers' Exact test, Global-test, Global ANOVA	Yes
25	MSEA	2010	Web-based	All	Cumulative hypergeometric distribution and Globaltest	Yes
26	PhenoFam	2010	Web-based	All	Mann-Whitney U-test	No
27	WebGestalt2	2010	Web-based	All	Hypergeometic test	Yes
28	GARNET	2011	Integrative web-based	All	Cohen's kappa statistic	Yes
29	HTSanalyzeR	2011	Toolbox-based	Windows Binary, Mac OS X 10.6	Hypergeometric test	Yes

Table 2. Contd…….

No	Name	Year	Application Type	Platforms	Key Statistical Method	Visualization
30	MBRole	2011	Web-based	All	Cumulative hypergeometric distribution	No
31	MPEA	2011	Web-based	All	t-tests and Hypergeometric distribution coupled with permutation test	Yes
32	TAFFEL	2011	Webstart	All	Fisher's exact test	Yes
33	GO-Elite	2012	Standalone, Web-based, Toolbox-based	Windows, Mac OS X, Ubuntu	Hypergeometric distribution and Fisher's exact test	Yes
34	INRICH	2012	Standalone	Windows, Mac OS X, Linux	Hypergeometric test	No
35	Enrichr	2013	Web-based	All	Fisher's exact test	Yes

Table 3. List of gene set databases.

No	Database	References	Organisms	Website
1	KEGG Pathway	Ogata *et al.* (1999) [44]	*Homo Sapiens*	http://www.genome.jp/kegg/pathway.html
2	Reactome Pathway	Joshi-Tope *et al.* (2005) [45]	Chicken, *Drosophila* and rice	http://www.reactome.org/ReactomeGWT/entrypoint.html
3	DBTSS	Suzuki *et al*. (2002) [46]	Human transcriptional start sites	http://dbtss.hgc.jp/
4	H-InvDB	Imanishi*et al*. (2004) [47]	Human transcriptional start sites	http://swww.h-invitational.jp/
5	MsigDB	Subramanian *et al*. (2005) [48]	*Homo Sapiens*	http://www.broadinstitute.org/gsea/msigdb/index.jsp

the human pathways and reactions. Recently, a new web site has been developed with enhanced tools for pathway browsing and also analysis of data. Thus, the Pathway Browser is a Systems Biology Graphical Notation (SBGN)-based visualization system which supports zooming, scrolling and event highlighting.

This systems works with the PSIQUIC web services to cover the pathways with molecular interaction data from the Reactome Functional Interaction Network, such as for the external interaction databases including IntAct, BioGRID, ChEMBL, iRefIndex, MINT and STRING. The Pathway and the tools of Expression Analysis enable the ID mapping process. They also enable the assignment of pathway and analysis overrepresentation of user-supplied data sets. In order to support the annotation of pathway and the analysis of other species, it was important to make orthology-based inferences of pathways, especially in the species of non-human. It applies the Ensembl Compara to identify the orthologs of human proteins in 20 other species.

3.3. DBTSS

DataBase of Transcriptional Start Sites (DBTSS) was invented by Suzuki *et al*. (2002) [46]. This database contains the accurate positions of transcriptional start sites (TSSs). It is obtained through the technique proposed named TSSseq in the genomes of numerous kinds of species. In the latest version, DBTSS covers the records of the best part of human adult and also the embryonic tissues. It now comprises of 418 million TSS tag sequences from 28 tissues or cell cultures. Furthermore, it includes the integration of a series of the proposed data of transcriptomic, like the RNA-seq data of subcellular-fractionated RNAs.

This is also similar to the ChIP-seq data of histone modifications. The binding of RNA polymerase II or some of the transcription factors in cultured cell lines had also been implemented into our original TSS information. In order to get a full length of cDN's, an oligo capping method was developed. The sequence which was produced by oligo capping method was first being processed, which was due to the trimming of its vector site and low quality parts. Besides that, from the DBTSS, users can get the TSS information of a specific gene in many ways. The advantage of DBTSS is that it is able to find the structure of exon and intron and also the pattern distribution of TSS with a variety of clones.

3.4. H-InvDB

H-InvDB was developed in the "Genome Information Integration Project" (2005-2008), based upon the annotation technology established in the H-Invitational Project for annotation of human full-length cDNAs by Imanishi*et al*. (2004) [47]. H-InvDB is a complete database of human gene started in the year of 2004. Its latest version, which is H-InvDB 8.0, comprises a total of 244 709 human complementary DNA which are mapped onto the hg19 reference genome. Furthermore, 43 829 gene loci which include the non-protein coding that have been identified. Out of these loci, 35

Table 4. Comparison between gene set databases.

No.	Databases	Advantages	Disadvantages
1	KEGG Pathway	-ensures data reliability	-has fixed view of the pathway data
2	Reactome Pathway	-has huge amount of information - provides more specific information -produces quite reliable data -has very clear and overview able references	-yields sparse coverage of genome -has no possibility of combining information available in the pathways
3	DBTSS	-accurate in its TSS information	-has large variability of the number
4	H-InvDB	-gives uniquely annotated genes	-has large output of results
5	MsigDB	-able to eliminate multiple probes	-limited by grouping genes and samples simultaneously

631 have been recognized as possible protein coding genes while 22 898 of known genes are identical. In the analysis, 19 309 annotated genes have been detailed to H-InvDB and are not found in RefSeq and also Ensembl (Takeda *et al*., 2012) [52].

Thus, 233 genes of the 19 309 have protein functions in H-InvDB, which in the previous version they were annotated as unknown protein functions. Furthermore, the H-InvDB, originally developed as an integrated database of the human transcriptome based on extensive annotation of large sets of full length cDNA (FLcDNA) clones, now in the latest release H-InvDB_4.6, provides annotation for 120 558 human mRNAs, extracted from the International Nucleotide Sequence Databases (INSD), in addition to 54 978 human FLcDNAs.

3.5. MsigDB

MsigDB was invented by Subramanian *et al*. (2005) [48]. According to Liberzon *et al*. (2011) [53], MsigDB is a well-annotated gene sets. It represents the universe of the processes of biological which are critical for important and perceptive explanation of large-scale of genomic data. The Molecular Signatures Database (MsigDB) is one of the most broadly used depositaries of such sets. It is easy to use, especially with the availability of the new version of database which is MsigDB 3.0 with over than 6700 gene sets. It contains a complete modification of the collection of canonical pathways. Also, it contains the experimental signatures from the publications, annotations enhancement and the web sites upgrade. Molecular Signatures Database (MsigDB) differs from these resources from several distinguishing aspects. MsigDB is explicitly designed to provide gene sets for enrichment analysis methods.

The MsigDB web site allows users to find gene sets by searching for keywords in the annotations. Family of gene proposes a quick view of a gene set by combining its members into a small number of informative parts. The family of gene is then updated and they include tumor suppressors, translocate cancer genes, oncogenes, transcription factors, home domain proteins, protein kinases, cell differentiation markers and growth factors. As such, it is natively and seamlessly integrated with the GSEA software. Table **4** shows the comparison between gene set databases.

4. CONCLUSION

With the quick growth of genomics and successful gene set enrichment analysis, biologists have proposed lots of tools and databases to make the gene set enrichment analysis easier. Thus, the analysis of gene set enrichment has become essential in further studies, as such analysis can help us to obtain a better understanding on the tools and databases involved in biological functions for different organisms. They also enable us to use gene set enrichment to conduct researches and discover new things that we might have missed before. Successful development in such areas would not only benefit the economy, but also allow us to understand or even inflect the biological evolution.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 01-01-06-SF1234). This research is also funded by an Exploratory Research Grant Scheme (Grant number: R.J130000.7807.4L096) and a Fundamental Research Grant Scheme (Grant number: R.J130000.7807.4F190) from Malaysian Ministry of Higher Education.

LIST OF ABBREVIATIONS

DBTSS	=	Database of Transcriptional Start Sites
DIANA	=	DNA Intelligent Analysis
GAGE	=	Generally Applicable Gene-set Enrichment
GARNET	=	Gene Annotation Relationship Network Tools
GO	=	Gene Ontology
GSA	=	Gene Set Analysis

GSA-SNP	=	Gene Set Analysis – Single Nucleotide Polymorphism
GSEA	=	Gene Set Enrichment Analysis
GWAS	=	Genome Wide Association Study
HEAT	=	H-InvDB Enrichment Analysis Tool
H-InvDB	=	H-Invitation Database
HTS	=	High Throughput Screens
i-GSEA4GWAS	=	Improved GSEA for GWAS
INRICH	=	Interval-based Enrichment Analysis
IPA	=	Ingenuity Pathway Analysis
KEA	=	Kinase Enrichment Analysis
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
MAGENTA	=	Meta-analysis Gene-set Enrichment of Variant Associations
MamPhEA	=	Mammalian Phenotype Enrichment Analysis
MBRole	=	Metabolite Biological Role
MetPA	=	Metabolomics Pathway Analysis
MPEA	=	Metabolite Pathway Enrichment Analysis
MSEA	=	Metabolite Set Enrichment Analysis
MSigDB	=	Molecular Signatures Database
WebGestalt2	=	Web-based Gene-set Enrichment Analysis Toolkit

REFERENCES

[1] Camon, E.; Magrane, M.; Barrell, D.; Lee, V.; Dimmer, E.; Maslen, J.; Binns, D.; Harte, N.; Lopez, R. and Apweiler, R. The gene ontology annotation (GOA) database: Sharing knowledge in uniprot with gene ontology. *Nuclei. Acids Res.,* **2004**, *32*, D262-D266.

[2] Huang, D.W.; Sherman, B.T. and Lempicki, R.A. Bioinformatics enrichment tools: Path toward the comprehensive functional analysis of large gene lists. *Nuclei. Acids Res.,* **2009**, *37*(1), 1-13.

[3] Lee, H.K.; Braynen, W.; Keshav, K. and Pavlidis, P. Erminej: Tool for Functional Analysis of Gene Expression Data Sets. *BMC Bioinformatics,* **2005**, *6*, 269.

[4] Zhou, X. and Su, Z. EasyGO: Gene ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*, **2007**, *8*, 246.

[5] Carmona-Saez, P.; Chagoyen, M.; Tirado, F.; Carazo, J.M. and Pascual-Montano, A. GENECODIS: A Web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.,* **2007**, *8*, R3.

[6] Frijters, R.; Heupers, B.; Beek, P.V.; Bouwhuis, M.; Schaik, R.V.; Vlieg, J.D.; Polman, J. and Alkema, W. CoPub: A literature-based keyword enrichment tool for microarray data analysis. *Nucl. Acids Res.,* **2008**, *36*(2), W406-W410.

[7] Fejes, A.P.; Robertson, G.; Bilenky, M.; Varhol, R.; Bainbridge, M. and Jones, S.J.M. FindPeaks 3.1: A tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinf. App. Note,* **2008**, *24*(15), 1729-1730.

[8] Keller, A.; Backes, C.; Al-Awadhi, M.; Cerasch, A.; Kuntzer, J.; Kohlbacher, O.; Kaufmann, M. and Lenhof, H.P. GeneTrailExpress: A web-based pipeline for the statistical evaluation of microarray experiments. *BMC Bioinformatics,* **2008**, *9*, 552.

[9] Bauer, S.; Grossmann, S.; Vingron, M. and Robinson, P.N. Ontogolizer 2.0: A multifunctional tool for GO term enrichment analysis and data expression. *Bioinf. App. Note,* **2008**, *24*(14), 1650-1651.

[10] Papadopoulos, G.L.; Alexiou, P.; Maragkakis, M.; Reczko, M. and Hatzigeorgiou, A.G. DIANA-mirPath: Integrating human and mouse microRNAs in pathways. *Bioinf. App. Note,* **2009**, *25*(15), 1991-1993.

[11] Luo, W.J.; Friedman, M.S.; Shedden, K.; Hankenson, K.D. and Woolf, P.J. GAGE: Generally Applicable Gene Set Enrichment for Pathway Analysis. *BMC Bioinf.,* **2009**, *10*, 161.

[12] Krämer, A.; Green, J.; Pollard, J. and Tugendreich, S. Causal Analysis Approaches in Ingenuity Pathway Analysis (IPA). Bioinformatics, **2013**, 703.

[13] Eden, E.; Navon, R.; Steinfeld, I.; Lipson, D. and Yakhini, Z. Gorilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinf.,* **2009**, *10*, 48.

[14] Lachmann, A.; Ma'ayan, A. KEA: Kinase enrichment analysis. *Bioinf. App. Note,* **2009**, *25*(5), 684-686.

[15] Hur, J.; Schuyler, A.D.; States, D.J. and Feldman, E.L. SciMiner: Web-based literature mining tool for target identification and functional enrichment analysis. *Bioinf. App. Note,* **2009**, *25*(6), 838-840.

[16] Chen, J.; Bardes, E.E.; Aronow, B.J. and Jegga, A.G. ToppGene Suite for Gene List Enrichment Analysis and Candidate Gene Prioritization. *Nuclei Acids Res.,* **2009**, *37*, W305-W311.

[17] Glez-Pena, D.; Gomez-Lopez, G.; Pisano, D.G. and Fdez-Riverola, F. WhichGenes: A web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nuclei. Acids Res.,* **2009**, *37*, W329-W334.

[18] Kolpakov, F.; Poroikov, V.; Selivanova, G. and Kel, A. GeneXplain-Identification of causal biomarkers and drug targets in personalized cancer pathways. *J. Biomol. Techniq.,* **2011**, *22*(Suppl), S16.

[19] Sartor, M.A.; Mahavisno, V.; Keshamouni, V.G.; Cavalcoli, J.; Wright, Z.; Karnovsky, A.; Kuick, R.; Jagadish, H.V.; Mirel, B.; Weymouth, T.; Athey, B. and Omenn, G.S. ConceptGen: A gene set enrichment and gene set relation mapping tool. *Bioinformatics,* **2010**, *26*(4), 456-463.S

[20] Merico, D.; Isserlin, R.; Stueker, O.; Emili, A. and Bader, G.D. Eenrichment Map: A network-based method for gene-set enrichment visualization and interpretation. PloS ONE, **2010**, 5(11), e13984.

[21] Nam, D.G.; Kim, J.; Kim, S.Y. and Kim, S.S. GSA-SNP: A General Approach for Gene Set Analysis of Polymorphisms. *Nuclei Acids Res.*, **2010**, *38*, W749-W754.

[22] Imanishi, T.; Noda, A.O.; Sera, M. HEAT: A New Tool for Gene Set Enrichment Analysis Using Comprehensive Annotation of Human Genes in H-InvDB. *Nat. Preceding,* **2010**, doi:10.1038/npre.2010.5185.1.

[23] Zhang, K.L.; Cui, S.J.; Chang, S.H.; Zhang, L.Y. and Wang, J. i-GSEA4GWAS: A Web Server for Identification of Pathways/Gene Sets Associated With Traits by Applying an Improved Gene Set Enrichment Analysis to Genome-wide Association Study. *Nuclei Acid Res.,* **2010**, *38*(2), W90-W95.

[24] Segre, A.V.; Groop, L.; Mootha, V.K.; Daly, M.J. and Alsthuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PloS Genetics*, **2010**, *6*(8), e1001058.

[25] Weng, M.P. and Liao, B.Y. MamPhEA: A web tool for mammalian phenotype enrichment analysis. *Genome Biol.,* **2010**, *11*(1), P27.

[26] Xia, J.G. and Wishart, D.S. MetPA: A Web-based metabolomics tool for pathway analysis and visualization. *Bioinf. App. Note,* **2010**, *26*(18), 2342-2344.

[27] Xia, J.G. and Wishart, D.S. MSEA: A Web-based Tool to Identify Biologically Meaningful Patterns in Quantitative Metabolomic Data. *Nuclei Acids Res.,* **2010**, *38*, W71-W77.

[28] Paszkowski-Rogacz, M.; Slabicki, M.; Pisabarro, M.T. and Buchholz, F. PhenoFam-Gene set enrichment analysis through protein structural information. *BMC Bioinformatics,* **2010**, *11*, 254.

[29] Duncan, D.; Prodduturi, N. and Zhang, B. WebGestalt2: An updated and expanded version of the web-based gene set analysis toolkit. *BMC Bioinf.,* **2010**, *11*(4), P10.

[30] Rho, K.Y.; Kim, B.J.; Jang, G.H.; Bae, T.J.; Seo, J.H.; Seo, C.H.; Lee, J.Y.; Kang, H.J.; Lee, J.Y.; Kang, H.J.; Yu, U.S.; Kim, S.H.; Lee, S.H. and Kim, W.K. GARNET: Gene set analysis with exploration of annotation relation. *BMC Bioinformatics,* **2011**, *12*(1), S25.

[31] Wang, X.; Terfve, C.; Rose, J.C. and Markowetz, F. HTSanalyzeR: An R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinf. App. Note,* **2011**, *27*(6), 879-880.

[32] Chaogoyen, M. and Pazos, F. MBRole: Enrichment analysis of metabolomic data. *Bioinf. App. Note,* **2011**, *27*(5), 730-731.

[33] Kankainen, M.; Gopalacharyulu, P.; Holm, L. and Oresic, M. MPEA: Metabolite pathway enrichment analysis. *Bioinf. App. Note,* **2011**, *27*(13), 1878-1879.

[34] Kurki, M.I.; Paananen, J.; Storvik, M.; Yla-Herttuala, S.; Jaaskelainen, J.E.; Fraunberg, M.V.U.Z.; Wong, G. and Pehkonen, P. TAFFEL: Independent Enrichment Analysis of Gene Sets. *BMC Bioinf.,* **2011**, *12*, 171.

[35] Zambon, A.C.; Gaj, S.; Ho, I.; Hanspers, K.; Vranizan, K.; Evelo, C.T.; Conklin, B.R.; Pico, A.R. and Salomonis, N. GO-Elite: A flexible solution for pathway and ontology over-representation. *Bioinformatics,* **2012**, *28*(16), 2209-2210.

[36] Lee, P.H.; Dushlaine, C.O.; Thomas, B. and Purcell, S.M. INRICH: Interval-based enrichment analysis for genome-wide association studies. *Bioinf. App. Note,* **2012**, *28*(12), 1798-1799.

[37] Chen, E.Y.; Tan, C.M.; Kou, Y.; Duan, Q.N.; Wang, Z.C.; Meirelles, G.V.; Clark, N.R. and Ma'ayan, A. Enrichr: Interative and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.,* **2013**, *14*, 128.

[38] Du, Z.; Zhou, X.; Ling, Y.; Zhang, Z.H. and Su, Z. agriGO: A GO analysis toolkit for the agricultural community. *Nuclei Acids Res.,* **2010**, *38*(2), W64-W70.

[39] Nogales-Cadenas, R.; Carnoma-Saez, P.; Vazquez, M.; Vicante, C.; Yang, X.Y.; Tirado, F.; Carazo, J.M. and Pascual-Montano, A. GeneCodis: Interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nuclei. Acids Res.,* **2009**, *37*, W317-W322.

[40] Tabas-Madrid, D.; Nogales-Cadenas, R. and Pascual-Montano, A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic. Acids Res.,* **2012**, *40*(1), W478-W483.

[41] Vlachos, I.S.; Kostoulas, N.; Vergoulis, T.; Georgakilas, G.; Reczko, M.; Maragkakis, M.; Paraskevopoulou, M.D.; Prionidis, K.; Dalamagas, T. and Hatzigeorgiou, A. G. DIANA miRPath v. 2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res.,* **2012**, *40*(W1), W498-W504.

[42] Valeev, T.; Ryabova, A.; Tolstyh, N.; Kolpakov, F. and Kel, A. GeneXplain platform for systems medicine. Department of bioengineering and bioinformatics of MV Lomonosov Moscow State University, **2011**, 156.

[43] Wang, J.; Duncan, D.; Shi, Z. and Zhang, B. Web-based gene set analysis toolkit (WebGestalt): Update 2013. *Nucleic Acids Res.,* **2013**, *41*(W1), W77-W83.

[44] Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H. and Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.,* **1999**, *27*(1), 29-34.

[45] Joshi-Tope, G.; Gillespie, M.; Vastrik, I.; D'Eustachio, P.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G.R.; Wu, G.R.; Matthews, L.; Lewis, S.; Birney, E. and Stein, L. Reactome: A knowledgebase of biological pathways. *Nucleic Acid Res.,* **2005**, *33*, 1-10.

[46] Suzuki, Y.; Yamashita, R.; Nakai, K. and Sugano, S. DBTSS: Database of human transcriptional start sites and full-length cDNAs. *Nuclei Acids Res.,* **2002**, *30*(1), 328-331.

[47] Imanishi, T.; Itoh, T.; Suzuki, Y.; O'Donovan, C.; Fukuchi, S.; *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA Clones. *PloS Biol.,* **2004**, *2*(6), e162.

[48] Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S. and Mesirov, J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA,* **2005**, *102*, 15545-15550.

[49] Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M. and Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.,* **2011**, *40*, 109-114.

[50] Moutselos, K.; Kanaris, I.; Chatziioannou, A.; Maglogiannis, I. and Kolisis, F.N. KEGGconverter: A tool for the in-silico modelling of metabolic networks of the KEGG pathways database. *BMC Bioinform.,* **2009**, *10*, 324.

[51] Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; D'Eustachio, P. and Stein, L. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.,* **2010**, *39*, 691-697.

[52] Takeda, J.; Yamasaki, C.; Murakami, K.; Nagai, Y.; Sera, M.; Hara, Y.; Obi, N.; Habara, T.; Gojobori, T. and Imanishi1, T. H-InvDB IN 2013: An Omics study platform for human functional gene and transcript discovery. *Nucleic Acids Res.,* **2012**, *41*, 915-919.

[53] Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P. and Mesirov, J.P. Molecular signatures database (Msigdb) 3.0, **2011**, *27*, 1739-1740.

Received: December 02, 2014 Revised: February 13, 2015 Accepted: February 15, 2015