4

Functional Interpretation of Omics Data by Profiling Genes and Diseases Using MeSH–Controlled Vocabulary

Takeru Nakazato, Hidemasa Bono and Toshihisa Takagi
Database Center for Life Science, Research Organization of Information and Systems
Japan

1. Introduction

One of the major aims of molecular biology and medical science is to understand disease mechanisms. A genetic disorder is a disease caused by abnormalities in genes and chromosomes, and researchers often report the identification of disease-relevant genes and correlations between phenotypes and genotypes (Butte & Kohane 2006; Lamb 2007; Perez-Iratxeta *et al.* 2002, 2005, 2007).

Omics analysis using microarray, new generation sequencing (NGS) technology, and mass spectrometry is widely employed for determining genome sequences and profiling gene expression. Changes in gene expression on a genome-wide scale can be detected by omics analysis, which provides various types of huge datasets. These data are often archived in public databases; nucleotide sequences in the DDBJ/EMBL/GenBank International Nucleotide Sequence Database (INSD) (Cochrane *et al.* 2011), gene expression in Gene Expression Omnibus (GEO) (Barrett *et al.* 2011), and journal articles in MEDLINE. Currently, research cannot continue without the use of these databases. In Japan, the Database Center for Life Science (DBCLS) has developed infrastructure for researchers to access and easily reuse these data by providing index sites such as INSD and GEO yellow pages and by constructing a portal site for life science databases and tools. Researchers can easily analyze public data in conjunction with their own omics data.

Here we present an analytical method to clarify the associations between genes and diseases. We characterized genes and diseases by assigning a MeSH-controlled vocabulary (Nakazato *et al.* 2008, 2009). Our objective was to help interpret omics data from molecular and clinical aspects by comparing these feature profiles.

2. Omics databases

2.1 Microarray data and its repository sites

Microarray technology is one of the most widely employed tools for genome-wide analyses of changes in gene expression under various conditions including diseases and drug treatments. This approach is also useful for examining genome copy number variations, methylation status, and transcription factor binding.

Microarray data are archived in public databases such as GEO at the National Center for Biotechnology Information (NCBI) (Barrett *et al.* 2011) and ArrayExpress at the European

Bioinformatics Institute (EBI) (Parkinson *et al.* 2011). GEO is a public functional genomics data repository that accepts array- and sequence-based data. It has been developed and maintained by NCBI since 2000. GEO archives three types of data: datasets derived from research projects, samples such as species and cell lines used, and platforms to produce data (i.e., chipsets and massively parallel sequencers). GEO contains approximately 22,000 series of experiments, approximately 8500 platforms, and approximately 540,000 samples as of March 2011.

GEO data are freely downloadable (http://www.ncbi.nlm.nih.gov/geo/); therefore, researchers can utilize the data to perform further analyses and compare their own data as omics analysis. However, it is extremely difficult to grasp the GEO archived data because experimental conditions referred in each GEO entry are complicated and partially described in plain English.

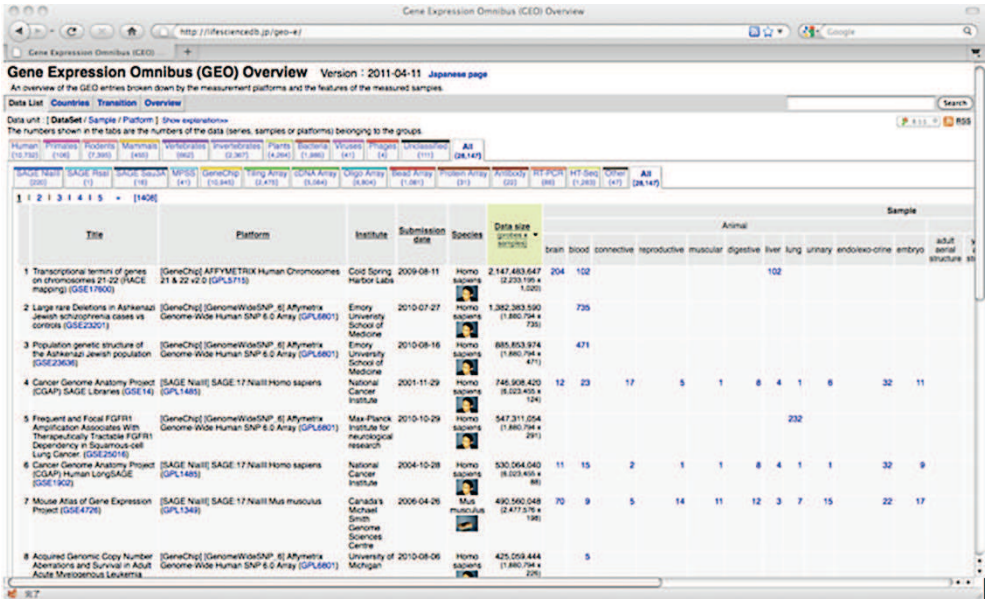


Fig. 1. Snapshot of the GEO Overview

To ease this situation, Dr. Okubo and his colleagues at the National Institute of Genetics (Japan) have developed a web service of an index site as a yellow page called the GEO Overview (http://lifesciencedb.jp/geo/), which is maintained by DBCLS (Fig. 1). It shows a list of project titles with their platforms and data provider names. In the GEO Overview, the datasets archived in GEO are categorized and organized by taxonomy (species type) and platform (methods or instruments). Researchers can easily refine the results by clicking the tabs corresponding to the taxonomy and platform of interest. In addition, the datasets can be searched using the search box at the top of the page. Hit data are categorized by histology with a hyperlink to the original GEO entry, and total data size is also provided. The GEO Overview should be helpful in outlining the abundant amount of gene expression data available from GEO. A tutorial movie for the GEO Overview is available on the TogoTV site (http://togotv.dbcls.jp/20100816.html).

2.2 NGS and its repository sites

Microarray technology has been widely employed to detect genome-wide gene expression. More recently, NGS, also called next-generation sequencing, has been performed for the same purpose. NGS is an ultra-high throughput nucleotide sequencing technology that drastically reduces the cost and time than previously possible (Shendure & Ji 2008). NGS technology has rapidly spread to approach whole-genome sequencing, metagenomics, and transcriptomics, and it also applies to epigenetics and genome-wide association study (GWAS) (Kahvejian *et al.* 2008).

NGS provides a tremendous of captured images and numerous sequence reads (Nat. Biotechnol. Editorial Board 2008), and the in-process files require huge amounts of disk space. However, NGS data are important for researchers and should be shared as well as microarray data in GEO. Thus, the NGS data are also archived in public databases; the Sequence Read Archive (SRA) (Leinonen *et al.* 2011 b) at NCBI, European Nucleotide Archive (ENA) (Leinonen *et al.* 2011 a) at EBI, and DNA Data Bank of Japan (DDBJ) Sequence Read Archive (DRA) (Kaminuma *et al.* 2010) at DDBJ. These databases are an archive databank for raw data from NGS, and the data are collaboratively synchronized. Researchers can search and download the archived data from the DDBJ site (http://trace.ddbj.nig.ac.jp/dra/).

Downloaded data from the SRA/ENA/DRA sites can be used for genome mapping, assembly, and annotation (Kaminuma *et al.* 2010). DBCLS has developed an index site for NGS data, called the Survey of Read Archives (http://sra.dbcls.jp/) as well as the GEO Overview site, to make this data more searchable and usable. The deposited NGS data contain not only sequence reads but also experimental conditions including project titles, species or cell lines, sample names, and sequencing platforms as metadata. The metadata consist of six files in XML format: submission, study, experiment, run, sample, and analysis. However, each submission does not contain all of these metadata because additional experiments or runs to be assigned to a previous project are often performed and archived as a new submission. Therefore, we determined the connections among each type of corresponding metadata and developed a project list as an index site. We attempted to curate the metadata by correcting misspellings and disambiguating spelling variations.

The Survey of Read Archives site provides a list with project titles, sample names, and a hyperlink to corresponding experiments and run data. It categorizes the data by study type including whole genome sequencing, transcriptome analysis, and metagenomics. Furthermore, the archived data are divided by platform and sample taxonomy. Thus, researchers can easily obtain final results with corresponding features of interest. The Survey of Read Archives site provides NGS statistical data such as the number of projects assigned to each study, platforms, and sample taxonomy. Table 1 shows the top ten list of NGS statistical data as of March 2011.

In addition, the Survey of Read Archives site offers a publication list that refers to NGS data. We obtained PubMed IDs (PMIDs) cited in the SRA database as reference articles. We also extracted hyperlinks and descriptions of SRA IDs from PubMed articles. The publication list provides article titles, journals, and project titles. Using this publication list, researchers can retrieve NGS data of sufficiently high quality for analysis. Users can narrow down the publication list by referring to NGS's study types, platforms, and sample species.

Study type	
Study type	The number of projects
Whole genome sequencing	2799
Transcriptome analysis	778
Metagenomics	564
Epigenetics	365
Resequencing	276
Other	259
Population genomics	110
RNASeq	104
Gene regulation study	37
Cancer genomics	18

Platforms	
Platform	The number of projects
Illumina Genome Analyzer II	1628
454 GS FLX	1209
454 Titanium	1121
Illumina Genome Analyzer	753
454 GS FLX Titanium	274
GS FLX	150
GS 20	135
Unspecified	109
Illumina HiSeq 2000	99
454 GS 20	61

Species of samples	
Species	The number of projects
Unidentified	460
Homo sapiens	414
Mus musculus	230
Metagenome sequence	170
Drosophila melanogaster	149
Marine metagenome	87
Caenorhabiditis elegans	75
Escherichia coli str. K-12 substr. MG1655	66
Arabidopsis thaliana	55
Saccharomyces cerevisiae	51

Table 1. List of top ten study types, platforms, and sample species of NGS data archived in SRA databases

3. Disease database

3.1 Online Mendelian Inheritance in Man (OMIM)

OMIM is one of the most widely referred disease databases by biological researchers (Amberger *et al.* 2009; Hamosh *et al.* 2002, 2005). It contains more than 21,000 detailed entries

of genetic diseases and disease-relevant human genes as a knowledge bank. Each disease entry provides a full-text overview in some categories including clinical features, diagnoses, and pathogenesis, and gene entry consists of sections such as cloning, gene function, and allelic variants sections.
OMIM was originally created as a printed version called the Mendelian Inheritance in Man (MIM), which has been annotated by Dr. McKusick and his colleagues for over 40 years (McKusick 2007). OMIM, which is the online version of MIM, is accessible through the internet from the NCBI site (http://www.ncbi.nlm.nih.gov/omim/). Its contents are the copyright of Johns Hopkins University. Data are updated daily on this database, and approximately 70 new entries are added per month (http://www.ncbi.nlm.nih.gov/Omim/dispupdates.html). The OMIM content is derived from the peer-reviewed biomedical literature.

3.2 Previous work using OMIM data

OMIM is an excellent resource to obtain information on genetic diseases and disease-relevant genes and for researchers attempting to understand disease features. Using Entrez Gene or Ensembl as a gene database, gene features including gene names, genomic locations, and gene ontology (GO) terms can be obtained. However, OMIM is not completely exploited for omics analysis because of its bibliographic data structure; it is written in plain English (Bajdik *et al.* 2005). To overcome these difficulties, previous studies attempted to extract knowledge described in OMIM and make it easier to use that knowledge for biological research, including omics analysis.
Some groups focused on terms referred in the clinical synopsis (CS) section of OMIM (Cantor & Lussier 2004; Freudenberg & Propping 2002; Hishiki *et al.* 2004; Masseroli *et al.* 2005 a; van Driel *et al.* 2006). The OMIM CS section contains keywords and key phrases for the mode of inheritance, symptoms, and phenotypes such as eye color, pain sensitivity, height, and weight.
Table 2 shows a partial list of terms referred in the CS section for Prader–Willi syndrome (OMIM ID: 176270) as an example.
This section describes clinical features of disorders and their modes of inheritance such as autosomal dominant, body system such as almond-shaped eyes, and endocrine features such as growth hormone deficiency.
As a previous study, categorization of each OMIM disease entry using particular criteria such as episodes, etiology, tissue, onset, and inheritance has been attempted (Freudenberg & Propping 2002). They also calculated correlations between OMIM entries on the basis of profile similarities.
Masseroli *et al*. normalized various descriptions such as Neuro and Neurologic in the CS section and characterized OMIM disease entries (Masseroli *et al.* 2005 a). They developed a web service called GFINDer to analyze phenotypes of inherited disorders (Masseroli *et al.* 2005 b).
Cohen *et al*. also developed a web service to search the OMIM CS section called CSI-OMIM (Cohen *et al*. 2011).
Using CS terms, researchers can retrieve disease information from OMIM without using text-mining techniques. Although the OMIM full-text content includes detailed biological and genetic descriptions, the CS terms are mainly clinical and diagnostic, and therefore, it is difficult to decipher disease information in conjunction with biological process data such as gene expression data. Furthermore, the CS terms such as Cardiac and Cardiovascular are

ambiguous because the assigned terms are often defined by the author's original description in the cited articles. We therefore utilized the medical subject headings (MeSH)-controlled vocabulary to characterize OMIM entries.

Headings	Subheadings	Feature
Inheritance	-	Isolated cases
Growth	Height	Mean adult male height, 155 cm
		Mean adult female height, 147 cm
		Steady childhood growth
	Weight	Onset of obesity from 6 months to 6 years
		Central obesity
Respiratory	-	Hypoventilation
		Hypoxia
Skeletal	-	Osteoporosis
		Osteopenia
Endocrine features	-	Hyperinsulinemia
		Growth hormone deficiency
		Hypogonadotropic hypogonadism
Miscellaneous	-	Food related behavioral problems include excessive appetite and obsession with eating
		Temperature instability
		High pain threshold
Molecular basis	-	Microdeletion of 15q11 in 70% of patients confirmed by fluorescent in situ hybridization
		Remainder of cases secondary to maternal disomy
		Rare cases secondary to chromosome translocation

Table 2. Keywords and key phrases referred in the OMIM CS section for Prader–Willi syndrome (partial)

4. Feature profiling of OMIM data using MeSH keywords

Many methods such as noise reduction (Li & Wong 2001 a, 2001 b), hierarchical clustering (Eisen *et al.* 1998), and self-organization maps (Tamayo *et al.* 1999) have been proposed for analyzing omics data including microarray. However, these methods are statistical approaches, and molecular biology and medicine researchers often need to grasp their microarray data from a biological viewpoint.

Researchers often use a controlled vocabulary, called ontology, to annotate biological features including genes. The most popular ontology for biologists is GO (Ashburner *et al.* 2000). GO consists of three categories: biological process, molecular function, and cellular component. For omics analysis, GO terms are often utilized by assigning corresponding terms to genes of interest (Khatri & Draghici 2005; Zeeberg *et al.* 2003, 2005). However, GO cannot be applied to annotate OMIM diseases because it focuses on features at the molecular level, and no term corresponding to specific diseases or chemical substance exists.

Here we introduce MeSH terms to characterize genes and diseases.

4.1 MeSH

MeSH (http://www.nlm.nih.gov/mesh/) is a controlled vocabulary and contains more than 23,000 keywords (Nelson *et al.* 2004). These keywords are hierarchically categorized into 15 concepts such as disease, chemicals and drugs, and anatomy. MeSH was originally curated for indexing MEDLINE articles by the National Library of Medicine (NLM). Researchers can view MeSH keywords assigned to each MEDLINE article in PubMed results. In a PubMed search, some queries are automatically added by corresponding MeSH terms, and PubMed is searched by a converted query. PubMed also accepts MeSH keywords as an input query. MeSH has over 177,000 entry terms that assist in finding the most appropriate MeSH heading. For example, vitamin C is an entry term for ascorbic acid. In addition, another approximately 200,000 terms of chemical compounds and proteins are available as the Substance Names. MeSH is freely on the NLM site in XML and ASCII formats and is updated annually.

4.2 Feature profiling of OMIM data

4.2.1 Data collection

We retrieved OMIM data available as of January 2010 by downloading them from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/repository/OMIM/) and by using the web service with Entrez Programming Utilities (eUtils, http://eutils.ncbi.nlm.nih.gov/). We obtained MeSH terms (2010 release) from the NLM web site (http://www.nlm.nih.gov/mesh/meshhome.html). MEDLINE article data were also obtained from NLM.

4.2.2 Article extraction related to each OMIM entry

As previously described, MeSH is originally used for keywords to index MEDLINE articles and is not directly linked to OMIM entries. Thus, we developed a method to retrieve articles referred in each OMIM entry. A schematic view of the pipeline for generating OMIM–PMID associations is shown in Fig. 2. The pipeline consists of three steps.

First, we retrieved PMIDs cited in the OMIM reference section (Fig. 2 a). Alzheimer Disease, AD (OMIM ID: 104300) was used as an example, and 191 articles were referred in the OMIM reference section as of March 2011. Previous studies also extracted hyperlinks to external databases to utilize MeSH terms for interpreting microarray data (Djebbari *et al.* 2005; Masys *et al.* 2001).

Next, we retrieved OMIM IDs referred in the Secondary Source ID section of MEDLINE articles (Fig. 2 b). We also collected the OMIM ID descriptions from full-text articles by searching PubMed Central. IDs of external databases including GenBank and GEO referred in full-text articles are often assigned to MEDLINE articles as a Secondary Source ID. As of April 2010, 5463 OMIM IDs were assigned to MEDLINE articles as a Secondary Source ID. These Secondary Source IDs are assigned by NLM, but not all IDs are extracted.

In the last step, we obtained PMIDs of articles assigned with MeSH terms corresponding to each OMIM entry (Fig. 2 c). As described above, MeSH contains disease category terms; therefore, there is often a MeSH keyword corresponding to each OMIM entry. For example, the OMIM entry for Alzheimer Disease, AD corresponded to the MeSH term Alzheimer Disease.

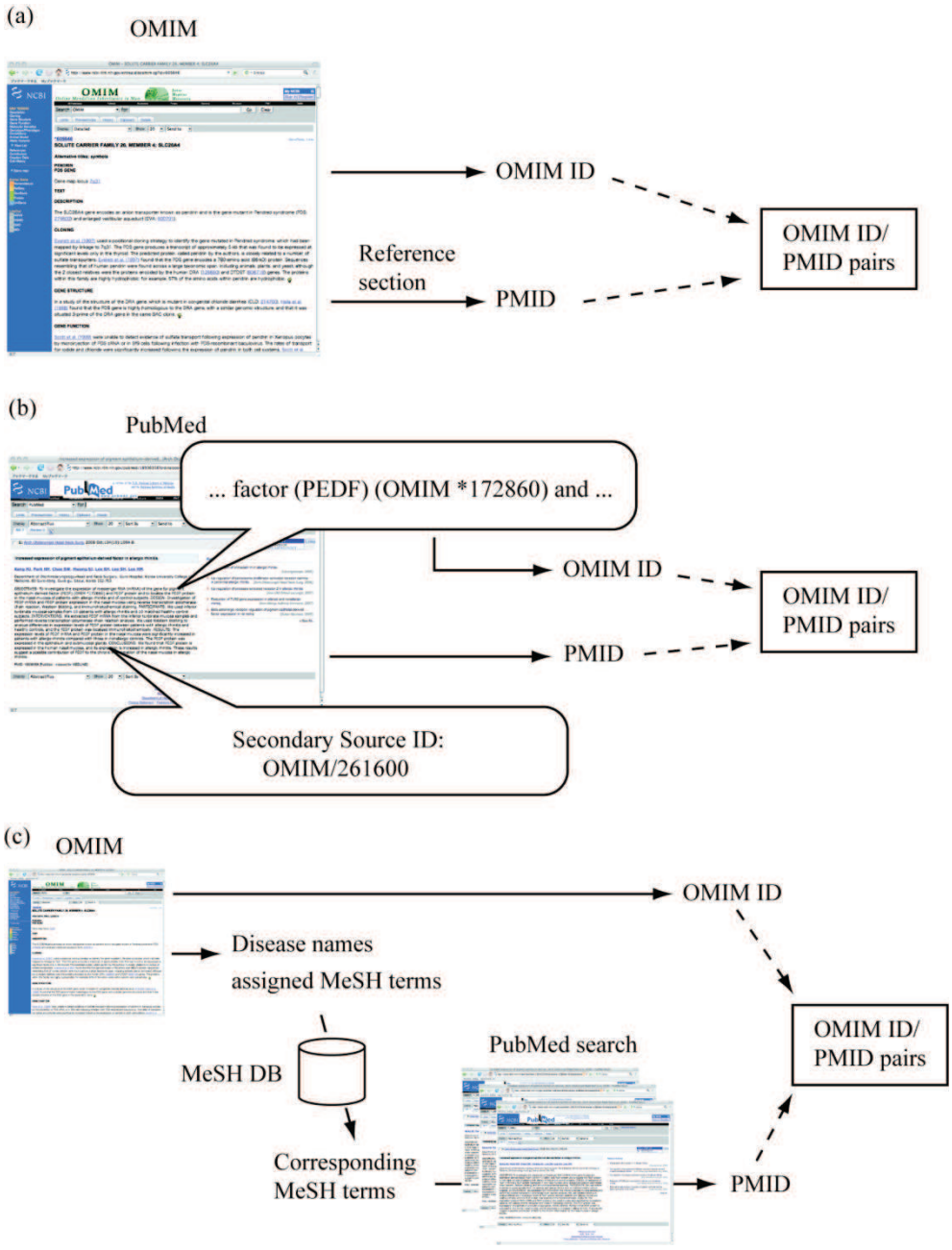


Fig. 2. Schematic view of the pipeline for generating pairs of OMIM entries and relevant PMIDs

We also obtained articles referring to human genes. OMIM contains entries describing not only genetic diseases but also disease-relevant human genes. By obtaining articles related to each OMIM entry using these steps, we also obtained articles on human genes. To complement articles on human genes, we obtained articles using Entrez Gene as a gene database. Using the process described above, we obtained PMIDs referred in Entrez Gene, describing Entrez Gene ID in the abstract, and assigned them to corresponding MeSH terms.

Accordingly, we retrieved approximately 500,000 unique pairs of OMIM IDs and PMIDs and generated approximately 2,000,000 OMIM–MeSH pairs.

In a previous version (Nakazato *et al.* 2008, 2009), we retrieved PMIDs by searching PubMed using disease names. To identify contexts indicating genes and diseases from articles is a major theme, and many approaches using text mining, such as named entity recognition (NER), have been reported (Gaudan *et al.* 2005; Hirschman *et al.* 2005; Jensen *et al.* 2006; Shatkay 2005). One of the difficulties is that a single disease often has many names, e.g., type 2 diabetes, non-insulin dependent diabetes, and NIDDM. Another problem is that the same abbreviation may refer to several diseases, genes, or drugs; e.g., EVA refers to enlarged vestibular aqueduct (disease), epithelial V-like antigen (gene), and ethylene vinyl acetate (chemical). Thus, we attempted to overcome this by creating abbreviations/long-form pairs for disease names, such as PWS and Prader–Willi syndrome, and searched MEDLINE for articles co-occurring with both names. However, this text-mining approach is noisy, and therefore, we discontinued applying this step in this version of data creation.

4.2.3 Scoring associations between OMIM entries and MeSH terms

OMIM contains gene entries as molecular mechanisms and disease entries as their phenotypes (Amberger *et al.* 2009). We calculated the scores of diseases and genes separately. These types are indicated by symbols prefixed to the OMIM ID such as #143100 (Huntington Disease; HD) and *613004 (Huntingtin; HTT). We divided the OMIM entries into three groups according to these types: sequence known (*, +), locus known (%), and phenotype (#, none). We then calculated p values as scores of OMIM–MeSH pairs in each group. The p value is the probability of the actual or a more extreme outcome under the null hypothesis. A lower p value means a larger significance of association. We used R language to calculate the p values.

4.2.4 Data visualization

To visualize retrieved features of OMIM disease entries with relevant MeSH terms, we developed a web-based software application called the gene disease features ontology-based overview system (Gendoo) (Nakazato *et al.* 2008, 2009). Gendoo accepts OMIM IDs, OMIM titles, Entrez Gene IDs, gene names, and MeSH terms as input queries. For disease names, Gendoo currently uses descriptions of title, alternative titles, and symbols referred in OMIM, and therefore, not all synonyms are included in the disease name dictionary. We will increase the number of synonyms by adding the canonical name and synonyms (entry terms) for corresponding MeSH terms and by extracting disease names from MEDLINE and OMIM resources using text mining. Gendoo generates high-scoring lists that display relevant MeSH terms for diseases, drugs, biological phenomena, and anatomy together with their scores. These MeSH terms are sorted according to their scores. The background color

of each association indicates its p value. Gendoo also provides a hierarchical tree view of MeSH terms associated with diseases of interest using JavaScript and cascading style sheet (CSS) resources from the Yahoo! User Interface (YUI) library (http://developer.yahoo.com/yui/).

Gendoo can be openly accessed at http://gendoo.dbcls.jp/. Every association file including Entrez Gene/OMIM IDs, MeSH, and their scores is available from the web site. Dictionary files including gene/disease names, synonyms, and IDs are also downloadable. These web services and files are freely available under a Creative Commons Attribution 2.1 Japan license (http://creativecommons.org/licenses/by/2.1/jp/deed.en).

4.3 Obtained feature profiles

4.3.1 Example 1: Positive control

Table 3 shows a list of scores and MeSH terms closely associated with Alzheimer Disease, AD (OMIM ID: 104300) and Amyloid Beta A4 Precursor Protein, APP (OMIM ID: 104760) as a positive control.

Alzheimer Disease; AD		
MeSH terms	Category	p value
Alzheimer Disease	Disease	0
Amyloid Beta Protein	Chemicals and Drugs	0
Brain	Anatomy	0

Amyloid Beta A4 Precursor Protein; APP		
MeSH terms	Category	p value
Alzheimer Disease	Disease	4.20×10^{-231}
Amyloid Beta Protein	Chemicals and drugs	5.12×10^{-214}
Brain	Anatomy	7.32×10^{-35}

Table 3. Lists of scores and keywords related to Alzheimer Disease and the Amyloid Beta A4 Precursor Protein

Alzheimer disease is a neurodegenerative disorder caused by accumulation of amyloid plaques in the brain. Here we used three MeSH terms as keywords to describe features of Alzheimer disease: Alzheimer Disease, Amyloid Beta Protein, and Brain. The scores among the keywords and the OMIM entry of Alzheimer Disease were small; thus, the retrieved associations seemed to properly illustrate features of the disease. The entry of the Amyloid Beta A4 Precursor Protein as an example of a gene was also strongly associated with these keywords.

4.3.2 Example 2: Retrieved profile

Table 4 lists the top three keywords related to Prader–Willi syndrome (OMIM ID: 176270) for the features of the fields, such as Diseases, Chemicals and Drugs, Biological Phenomena, and Anatomy, as examples of retrieved profiles.

Prader–Willi syndrome results from deletion of paternal copies of the imprinted small nuclear ribonucleoprotein polypeptide N (SNRPN) and necdin genes within chromosome 15 (Horsthemke & Wagstaff 2008). The results showed that the keyword clearly reflected the features of Prader–Willi syndrome, including Chromosomes, Human, Pair 15, Genomic Imprinting, and Ribonucleoproteins, Small Nuclear. For the features of Prader–Willi syndrome, the OMIM CS section presents morphologies and clinical and diagnostic fields such as mean height and temperature instability (Table 2). This approach illustrates the disease features from a clinical and biological perspective. To retrieve more clinical and diagnostic features with MeSH, we increased the number of novel associations using terms from the MeSH category Analytical, Diagnostic, and Therapeutic Techniques and Equipment.

MeSH category	MeSH terms	p value
Diseases	Prader–Willi syndrome	0
	Angelman syndrome	4.05×10^{-140}
	Obesity	6.94×10^{-128}
Chemicals and Drugs	Human growth hormone	5.86×10^{-68}
	Ribonucleoproteins, small nuclear	4.29×10^{-62}
	Ghrelin	1.58×10^{-50}
Biological Phenomena	Chromosomes, human, pair 15	0
	Genomic imprinting	2.47×10^{-131}
	Obesity	1.69×10^{-121}
Anatomy	Chromosomes, human, pair 15	0
	Chromosomes, human, 13–15	1.25×10^{-30}
	Adipose tissue	3.93×10^{-13}

Table 4. List of top three keywords related to Prader–Willi syndrome

4.3.3 Example 3: Comparison of profiles between diseases

We applied this analysis to types 1 and 2 diabetes mellitus (OMIM IDs 222100 and 125853, respectively). Figure 3 shows a summary of typical features and their scores for type 1 and 2 diabetes mellitus. Each cell color on the heat map reflects the p value of the association.

Figure 3 summarizes the feature profiles; only type 1 diabetes mellitus was closely related to Autoimmune Diseases and Spleen (their p values were 4.55×10^{-5} and 5.53×10^{-7}, respectively), whereas type 2 diabetes was associated with Obesity (p value = 1.18×10^{-15}) and Adipocytes (p value = 5.17×10^{-5}). Type 1 diabetes mellitus involves the immune system, whereas type 2 diabetes mellitus is a metabolic disorder (Rother 2007). These retrieved profiles reflect the biological features of the diseases. This result suggests that MeSH profiles can clarify the differences and similarities in features between OMIM entries.

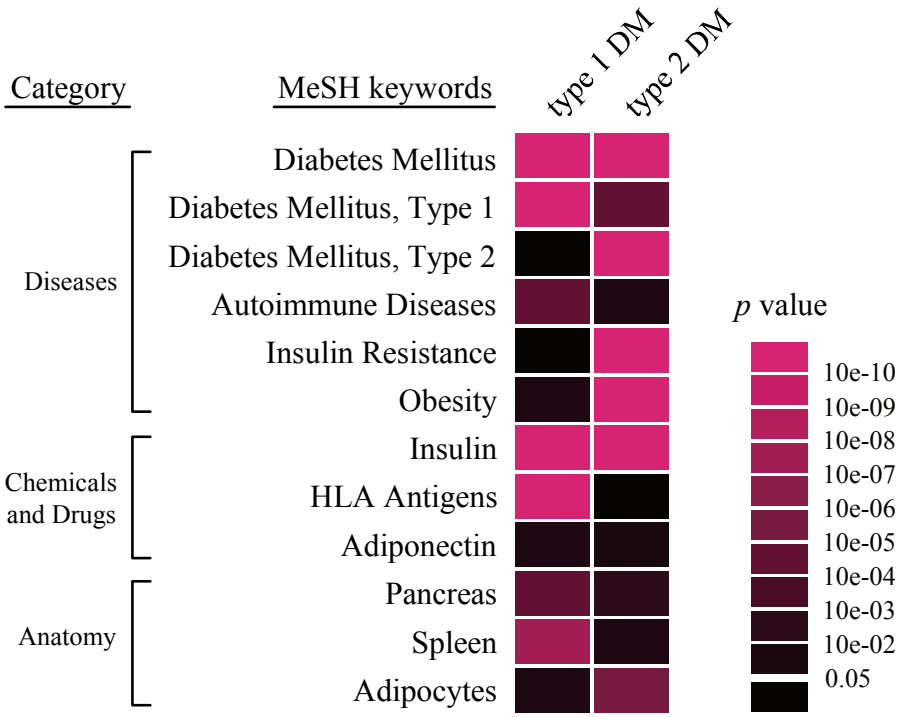


Fig. 3. Differences and similarities between feature profiles of type 1 and 2 diabetes mellitus. DM, diabetes mellitus

5. Discussion

Diverse types of life science data are available including nucleotide sequences at the molecular level and clinical records at the individual level. Omics analysis makes it easy to detect genome-wide upregulation and downregulation of gene expression under various conditions. We analyzed these raw data using several approaches such as statistical clustering and pathway analysis. In addition, to decipher phenotype information in conjunction with molecular data, we often relate genes that drastically change their expression levels to diseases as a result of omics analysis. However, molecular biologists only understand mechanisms for specific diseases. Although OMIM is an excellent knowledge bank for various diseases, it is not completely exploited for omics analysis because of the bibliographic data structure of OMIM. Moreover, drug information is not linked to elements associated with diseases and genes of interest. To alleviate this problem, we comprehensively characterized diseases and genes referred in OMIM with MeSH-controlled vocabulary. MeSH profiles allow disease features to be shared and compared.

Using GO terms, researchers can decipher their omics data from a molecular viewpoint. The developed feature profiles illustrate related diseases and drugs. We could obtain more clinical and medical data using these MeSH profiles.

Furthermore, the profiles can be applied to analyses of disease-relevant genes by comparing the similarities among profiles of OMIM entries and groups of genes such as those found in

the gene expression clustering results. Researchers can also obtain overviews of the features of unfamiliar diseases.

Genetic disease subtype entries are available in OMIM. For example, diabetes mellitus, non-insulin dependent, 3 (NIDDM3, OMIM ID: 603694) is a genetic subtype of diabetes mellitus, non-insulin dependent (NIDDM, OMIM ID: 125853, i.e., type 2 diabetes mellitus), which is linked to chromosome 20q12-q13.1. Another genetic type of NIDDM, NIDDM1 (OMIM ID: 601283), is reportedly linked significantly to chromosome 2q37.3. The differences in the clinical features between these two NIDDM genetic types seem to be unclear but the genetic mechanisms are probably different. Omics analysis emphasizes these genetic differences.

The diabetes mellitus entry was missing in OMIM, although the entries diabetes mellitus, insulin-dependent (type 1 diabetes mellitus) and diabetes mellitus, non-insulin dependent (type 2 diabetes mellitus) were present. We plan to create a dictionary of diseases using not only OMIM but also MeSH disease category and ICD-10 terms.

6. Conclusion

We characterized diseases and genes by generating feature profiles for associated drugs, biological phenomena, and anatomy using MeSH keywords. We developed a web service called Gendoo to visualize retrieved profiles. This approach illustrates disease features not only from a clinical but also a biological viewpoint. We also clarified the differences and similarities between disease features by comparing their profiles. Retrieved feature profiles are easy to remix such that Gendoo accelerates the process of omics analysis.

7. Acknowledgment

We thank Prof. Shoko Kawamoto and Prof. Kousaku Okubo for their helpful discussions. This work was supported by Integrated Database Project of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

8. References

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, Vol.37, Database issue, pp. D793-796. ISSN 1362-4962

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, Vol.25, No.1, pp. 25-29. ISSN 1061-4036

Bajdik, C. D., Kuo, B., Rusaw, S., Jones, S., and Brooks-Wilson, A. (2005). CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. *BMC Bioinformatics*, Vol.6, pp. 78. ISSN 1471-2105

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A., and Soboleva, A. (2011). NCBI GEO: archive for

functional genomics data sets--10 years on. *Nucleic Acids Res*, Vol.39, Database issue, pp. D1005-1010. ISSN 1362-4962

Butte, A. J., and Kohane, I. S. (2006). Creation and implications of a phenome-genome network. *Nat Biotechnol*, Vol.24, No.1, pp. 55-62. ISSN 1087-0156

Cantor, M. N., and Lussier, Y. A. (2004). Mining OMIM for insight into complex diseases. *Medinfo*, Vol.11, No.Pt 2, pp. 753-757. ISBN 1-58603-444-8

Cochrane, G., Karsch-Mizrachi, I., and Nakamura, Y. (2011). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*, Vol.39, Database issue, pp. D15-18. ISSN 1362-4962

Cohen, R., Gefen, A., Elhadad, M., and Birk, O. S. (2011). CSI-OMIM - Clinical Synopsis Search in OMIM. *BMC Bioinformatics*, Vol.12, pp. 65. ISSN 1471-2105

Djebbari, A., Karamycheva, S., Howe, E., and Quackenbush, J. (2005). MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. *Bioinformatics*, Vol.21, No.15, pp. 3324-3326. ISSN 1367-4803

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, Vol.95, No.25, pp. 14863-14868. ISSN 0027-8424

Freudenberg, J., and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, Vol.18 Suppl 2, pp. S110-115. ISSN 1367-4803

Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005). Resolving abbreviations to their senses in Medline. *Bioinformatics*, Vol.21, No.18, pp. 3658-3664. ISSN 1367-4803

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, Vol.30, No.1, pp. 52-55. ISSN 1362-4962

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, Vol.33, Database issue, pp. D514-517. ISSN 1362-4962

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, Vol.6 Suppl 1, pp. S1. ISSN 1471-2105

Hishiki, T., Ogasawara, O., Tsuruoka, Y., and Okubo, K. (2004). Indexing anatomical concepts to OMIM Clinical Synopsis using the UMLS Metathesaurus. *In Silico Biol*, Vol.4, No.1, pp. 31-54. ISSN 1386-6338

Horsthemke, B., and Wagstaff, J. (2008). Mechanisms of imprinting of the Prader-Willi/Angelman region. *Am J Med Genet A*, Vol.146A, No.16, pp. 2041-2052. ISSN 1552-4833

Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, Vol.7, No.2, pp. 119-129. ISSN 1471-0056

Kahvejian, A., Quackenbush, J., and Thompson, J. F. (2008). What would you do if you could sequence everything? *Nat Biotechnol*, Vol.26, No.10, pp. 1125-1133. ISSN 1546-1696

Kaminuma, E., Mashima, J., Kodama, Y., Gojobori, T., Ogasawara, O., Okubo, K., Takagi, T., and Nakamura, Y. (2010). DDBJ launches a new archive database with analytical

tools for next-generation sequence data. *Nucleic Acids Res*, Vol.38, Database issue, pp. D33-38. ISSN 1362-4962

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, Vol.21, No.18, pp. 3587-3595. ISSN 1367-4803

Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer*, Vol.7, No.1, pp. 54-60. ISSN 1474-175X

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V., and Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Res*, Vol.39, Database issue, pp. D28-31. ISSN 1362-4962

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Res*, Vol.39, Database issue, pp. D19-21. ISSN 1362-4962

Li, C., and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, Vol.98, No.1, pp. 31-36. ISSN 0027-8424

Li, C., and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*, Vol.2, No.8, pp. RESEARCH0032. ISSN 1465-6914

Masseroli, M., Galati, O., Manzotti, M., Gibert, K., and Pinciroli, F. (2005). Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists. *BMC Bioinformatics*, Vol.6 Suppl 4, pp. S18. ISSN 1471-2105

Masseroli, M., Galati, O., and Pinciroli, F. (2005). GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res*, Vol.33, Web Server issue, pp. W717-723. ISSN 1362-4962

Masys, D. R., Welsh, J. B., Lynn Fink, J., Gribskov, M., Klacansky, I., and Corbeil, J. (2001). Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, Vol.17, No.4, pp. 319-326. ISSN 1367-4803

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet*, Vol.80, No.4, pp. 588-604. ISSN 1432-1203

Nakazato, T., Takinaka, T., Mizuguchi, H., Matsuda, H., Bono, H., and Asogawa, M. (2008). BioCompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of genes. *In Silico Biol*, Vol.8, No.1, pp. 53-61. ISSN 1386-6338

Nakazato, T., Bono, H., Matsuda, H., and Takagi, T. (2009). Gendoo: functional profiling of gene and disease features using MeSH vocabulary. *Nucleic Acids Res*, Vol.37, Web Server issue, pp. W166-169. ISSN 1362-4962

Nat. Biotechnol. Editorial Board, (2008). Prepare for the deluge. *Nat Biotechnol*, Vol.26, No.10, pp. 1099. ISSN 1546-1696

Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J. L., and Arluk, N. (2004). The MeSH translation maintenance system: structure, interface design, and implementation. *Stud Health Technol Inform*, Vol.107, No.Pt 1, pp. 67-69. ISSN 0926-9630

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N., and Brazma, A. (2011). ArrayExpress update--an archive

of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, Vol.39, Database issue, pp. D1002-1004. ISSN 1362-4962

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat Genet*, Vol.31, No.3, pp. 316-319. ISSN 1061-4036

Perez-Iratxeta, C., Wjst, M., Bork, P., and Andrade, M. A. (2005). G2D: a tool for mining genes associated with disease. *BMC Genet*, Vol.6, pp. 45. ISSN 1471-2156

Perez-Iratxeta, C., Bork, P., and Andrade-Navarro, M. A. (2007). Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res*, Vol.35, Web Server issue, pp. W212-216. ISSN 1362-4962

Rother, K. I. (2007). Diabetes treatment--bridging the divide. *N Engl J Med*, Vol.356, No.15, pp. 1499-1501. ISSN 1533-4406

Shatkay, H. (2005). Hairpins in bookstacks: information retrieval from biomedical text. *Brief Bioinform*, Vol.6, No.3, pp. 222-238. ISSN 1467-5463

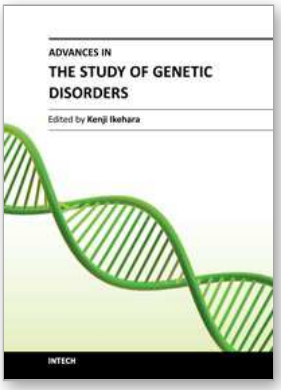
Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol*, Vol.26, No.10, pp. 1135-1145. ISSN 1546-1696

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, Vol.96, No.6, pp. 2907-2912. ISSN 0027-8424

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *Eur J Hum Genet*, Vol.14, No.5, pp. 535-542. ISSN 1476-5438

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, Vol.4, No.4, pp. R28. ISSN 1465-6914

Zeeberg, B. R., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D. W., Reimers, M., Stephens, R. M., Bryant, D., Burt, S. K., Elnekave, E., Hari, D. M., Wynn, T. A., Cunningham-Rundles, C., Stewart, D. M., Nelson, D., and Weinstein, J. N. (2005). High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*, Vol.6, pp. 168. ISSN 1471-2105



Advances in the Study of Genetic Disorders

Edited by Dr. Kenji Ikehara

ISBN 978-953-307-305-7
Hard cover, 472 pages
Publisher InTech
Published online 21, November, 2011
Published in print edition November, 2011

The studies on genetic disorders have been rapidly advancing in recent years as to be able to understand the reasons why genetic disorders are caused. The first Section of this volume provides readers with background and several methodologies for understanding genetic disorders. Genetic defects, diagnoses and treatments of the respective unifactorial and multifactorial genetic disorders are reviewed in the second and third Sections. Certainly, it is quite difficult or almost impossible to cure a genetic disorder fundamentally at the present time. However, our knowledge of genetic functions has rapidly accumulated since the double-stranded structure of DNA was discovered by Watson and Crick in 1956. Therefore, nowadays it is possible to understand the reasons why genetic disorders are caused. It is probable that the knowledge of genetic disorders described in this book will lead to the discovery of an epoch of new medical treatment and relieve human beings from the genetic disorders of the future.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821