GEFERENCE: Reference Database for Construing Personal Genome Expression

Changwon Keum[3], Ju Han Kim[1], Young Soo Song[1], Kyoung Tai No[3,4] and Jung Hoon Woo[1, 2, ¶]
[1]Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul, Korea.
[2]Macrogen Inc., Seoul, Korea.
[3]Bioinformatics & Molecular Design Research Center, Seoul, Korea
[4]Dept. of Biotechnology, Yonsei University, Seoul, Korea
[¶]Corresponding author

Abstract - *With the lowering cost and the increasing discoveries, genomics research filed has materialized personal genome era. Especially, microarray gene expression platform has already made several success stories for prognosis/diagnosis in medical contexts. However, the genome expression profiles, although having extensive information, still not be used in usual clinical context due to its complexity. In this study, we developed a reference database system, GEFERENCE, a searchable space for personal genome expression profiles by reorganizing patient centric genome expression and clinical information extracted from public data repository, GEO. We assembled total 34,745 patient's gene expression data from the 28 different tissue types with corresponding clinical information. And we manually curate the relation of information among individual patient, gene expression data, and clinical information by tight collaboration with clinicians. With GEFERENCE, clinician or medical service provider could search genome expression profiles of patients and get insight for relevant decision making.*

Keywords: personal genome era, personal genome expression, reference database, clinical information, GEO, decision support

1 Introduction

Rapid development of next generation sequencing technologies, which allows reading genome efficiently, has been accelerating the advent of the personal genome era. Like genome sequence, there are different levels of personal genome information, such as, personal genome expression, personal metabolome, personal methylome, etc. Especially, recent advance (i.e. stabilization and lowering cost) of microarray technology, which allows for profiling whole genes' expression in a single experiment, has been realized personal genome era in different way. Being accumulated, mining knowledge from patient based gene expression data has produced plenty of prognostic or diagnostic models for several diseases [1-6]. For example, prognostic models based on patterns of genome expression (i.e. gene expression signature) have shown great performance in the context of predicting disease progression, survival of patient, recurrence, and drug responsibility. Furthermore, some of the prognostic model were approved by the United States Food And Drug Administration (FDA) and were deployed on a commercial scale (i.e. Mamma print (http://usa.agendia.com/en/mammaprint.html), Oncotype DX (http://www.genomichealth.com/OncotypeDX), Tissue of Origin test (http://www.pathworkdx.com/TissueOfOriginTest/) rather remained in research level. Personal genome era or what we call a 'personalized medicine' already has been materialized in clinical side.

In this manner, extracting relevant information from large-scale genome expression could provide clue for unsolved clinical problems or would support clinical decision for patients [6]. Gene expression of patients would even be routinely profiled and treated as conventional clinical data such as, history and physical examination, laboratory findings and radiologic featrures [7]. The conventional clinical information is intuitive but limited in providing enough information about treatment and prognosis because a set of clinical data is usually heterogeneous and the quality is inconsistent depending on the data collectors and institutions. On the other hand, personal genome expression would provide extensive information but hard to understand. Even if some results from previous studies might give meaning to some extent, they are limited by what type of tissue and type of sample collected, used for the studies. Simply, personal genome expression is numeric vector whose size is about thirty thousands. As it is hard to extract meaning from the personal genome expression directly, we hypothesized that researchers or clinicians might get relevant insights from clinical information of other patients anchored by similarity of genome expression with target individual (Fig 1).

2 Methods and Materials

2.1 Reference datasets

To establish reference datasets for personal genome expression, we downloaded gene expression datasets

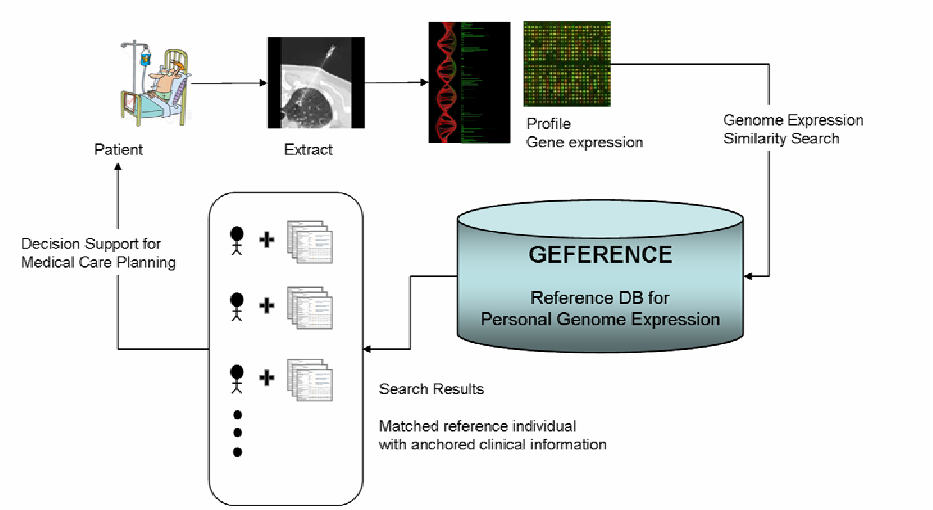

Fig 1. Representation of a general use case for GEFERENCE, reference database for personal genome expression

selectively from GEO (Version 9th Jan, 2009), a single largest gene expression data repository. There were total 10,445 gene expression datasets, which comprises of 233,775 gene expression samples, experimented on 2,326 different microarray platforms. Since GEO comprises gene expression datasets generated for various species, cell lines, and experimental design, we narrowed down our interest only on datasets which were experimented on human tissue. We considered each samples from the set of gene expression dataset, satisfying our filtering criteria as describe above, as an each person's genome expression data.

2.2 Extract reference gene expression data

Data for gene expression intensity, hybridization, PubMed id and the other data related to gene expression required for database construction were extracted from GEO series matrix data files (GSE files) and data for platform information were extracted from GEO platform data files (GPL files) with simple pattern matching techniques using Perl programming.

2.3 Extract reference clinical data

The clinical data could not be extracted fully automatically since there were more than hundreds of clinical data categories without standard ontology. Therefore, we first selected 'Sample_characteristics' and 'Sample_description' information, extracted from GEO series matrix data files, if they included at least one term matched to clinical term list made by pathologist. To exclude irrelevant information we screened ones, from the previously selected 'Sample_characteristics' and 'Sample_description', having at least one of the terms, strain, cell line, cell culture, stock, and the terms which represent other species such as, mouse, rat, monkey, swine, starfish, plant, c. elegans, macaque, pigs, etc. Reviewing the categories with clinicians considerably, we mapped the terms, different but have same meaning, to standardized LOINC [9] ontology. Finally, we assigned the extract clinical information to each of the corresponding individual gene expression data.

2.4 Gene function and pathway data

To annotate genes targeted by probes on each microarray platform, we utilized external data sources, Gene Ontology (GO) [10] for the gene function, KEGG [11] and C2 curated gene set of MsigDB[12] for the pathway information, respectively. We mapped probes to corresponding identifier in GO, KEGG and MsigDB using NCBI Entrez [13] across databases searching system by mapping probes ID to corresponding NCBI gene ID according to platform annotation of GPL files.

2.5 Database implementation

The GEFERENCE has been implemented as a web-based database. Patient centric reference gene expression datasets with corresponding clinical information resulting from the previous processing steps were inserted into relational database system, MySQL version 5.1

3 Results

3.1 Relational Database Schema

The GEFERENCE has been implemented as a web-based database system, powered by MySQL 5.1 (MySQL Inc, Cupertino, Calif) database management system. The relational schema is shown in Fig 2. The database was designed patient centric and composed of four major subsections that are origin of data, DNA level variome, clinical information, and gene expression data. As our system

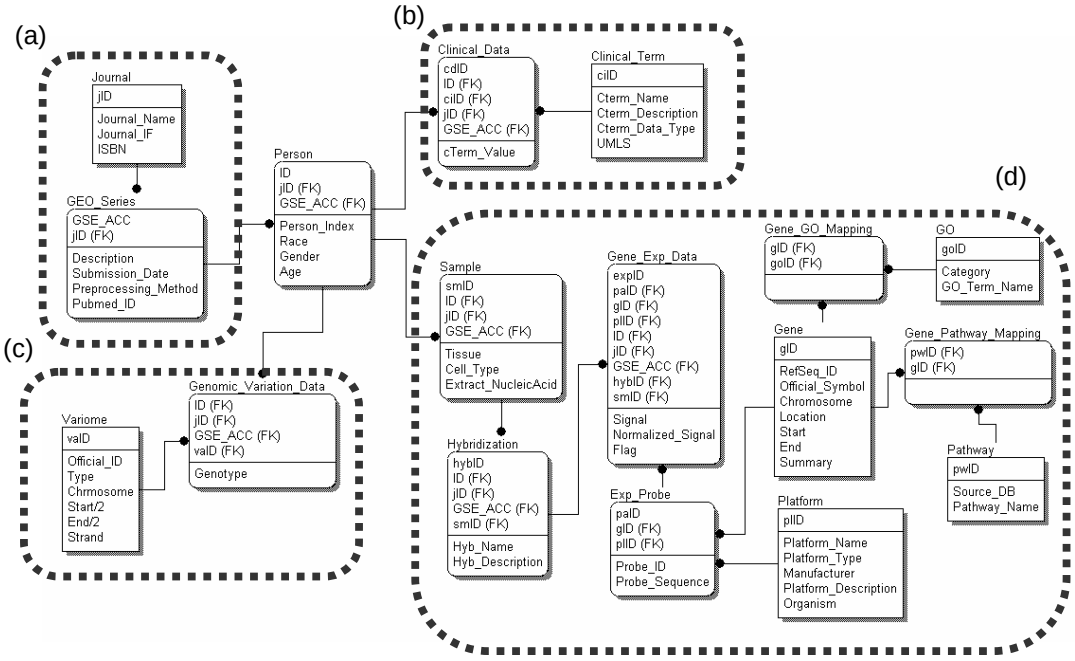

Fig 2. GEFERENCE database schema. Four major subsection with total 17 independent tables comprise GEFERENCE

stored data localized from public repositories, we designed GEO_series and Journal tables as external description of individual data point (Fig 2(a)). Correspond to recent study design, genetical genomics [14], conducting on not only gene expression data but also genotype data (i.e. single nucleotide polymorphism (SNP) or copy number variation (CNV)) extracted from single individual, we designed Variome and Genomic_Variation_Data tables to store that types of information (Fig 2(b)). We designed Clincal_Data and Clinical Term table to store clinical information, such as age, sex, and disease status, manually curated and assigned to each patients (Fig 2(c)). In Fig 2(d) represents tables contain information and annotation on the event of microarray experiment. Sample table was designed for storing information of biomaterials used in array experiment, tissue, cell type and the hybridized extract such as mRNA, DNA or microRNA. Because there used to be technical repeat for the extract, we designed Hybridization table to represent physical unit of microarray experiment. Gene_Exp_Data table stored actual expression signal of probes basically target corresponding genes and were annotated by information stored at Gene, GO, and Pathway tables.

3.2 General features of database

The database consists of 17 independent data tables. There are total 34,745 personal genome expression samples on 204 different platforms. Of these personal genome expression samples, 5,678 samples were annotated with 28 different tissue types from which the samples were derived. Breast (1,511 samples), blood (1,473 samples), lung (811 samples), brain (769 samples), prostate (492 samples) and liver (476 samples) were most frequently observed tissues. All 28 tissue types and corresponding number of personal genome expression data samples are shown in Table 1.

4 Conclusions

In this study, we designed patient centric database and localized plenty of datasets for specific purpose. Having genome expression of 34,745 individuals, GEFERENCE could provide extensive reference space for searching personal genome expression. Not giving direct results or meaning, for the input personal genome expression, GEFERENCE would provide corresponding clinical information of matched reference individuals. With our database system, clinician or medical service provider could search their patients with their genome information and could get insights for stratifying medical treatments. In brief, searching GEFERENCE might be the very first step of personal genome expression touching personalized medicine.

Being designed for the specific motivation, the resulting GEFERENCE database could be used for various purposes. In the research perspective, our patient centric database and the clinical information based gene expression searching availability shed new lights on disease research, which was thought to be impossible for the reason, requiring extensive manual curate process of data collection. Even though the gene expression data stored in GEFERENCE have already been existed, it was thought to be difficult to select subset of

Table 1. Number of samples according to tissue type

Tissue	# of Samples	Tissue	# of Samples
Breast	1,511	Heart	43
Blood	1,473	Bladder	36
Lung	811	Spleen	32
Brain	769	Testis	24
Prostate	492	Placenta	20
Liver	476	Stomach	17
Ovary	365	Spinal cord	13
Skin	233	Thalamus	4
Colon	230	Corpus	3
Skeletal muscle	219	Thymus	2
Thyroid	191	Pancreas	2
Caudate nucleus	83	Intestine	2
Cerebellum	79	Adrenal gland	2
Uterus	56	Bowel	1

gene expression profiles for one's interest especially when one need gene expression profiles on specific clinical interests. Researchers, for example, using GEFERENCE, could compile meta-datasets with gene expression profiles of lung cancer patients with whose detailed clinical information such as age, sex, tobacco usage, alcohol usage, list of medications taken and so on. Multi-dimensional analysis of detailed clinical data provided gene expression data would identify the characteristic gene expression patterns associated with clinical conditions, which were neglected before.

5 Limitations

Currently, the category of clinical term was defined both by automatic category identification and in part by manual curation. To improve the quality, continuous manual curattion would be needed. At the same time, we would aim to improve the process more advanced automatic clinical category extraction algorithms. Composing GEFERENCE with the GEO datasets only, we would extend our system to incorporate another public gene expression database, ArrayExpress [15].

6 Acknowledgment

This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare (A040002) and a grant (09172KFDA637) from Korea Food & Drug Administration.

7 References

[1] Butte A. "The use and analysis of microarray data". Nat Rev Drug Discov., 1, 951–960, 2002

[2] Perez EA, Pusztai L, Van de Vijver M. "Improving patient care through molecular diagnostics". Semin Oncol., 31, Suppl10, 14–20, 2004

[3] Pusztai L, Symmans FW, Hortobagyi GN. "Development of pharmacogenomic markers to select preoperative chemotherapy for breast cancer". Breast Cancer, 12, 73–85, 2005

[4] Simon R. "Roadmap for developing and validating therapeutically relevant genomic classifiers". J Clin Oncol., 23, 7332–7341, 2005

[5] Dupuy A, Simon RM. "Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting". J Natl Cancer Inst., 99, 2,147-157, 2007

[6] Liu ET."Mechanism-derived gene expression signatures and predictive biomarkers in clinical oncology". Proc. Natl. Acad. Sci. USA ,102, 3531-3532, 2005

[7] Hoffman MA. "The genome enabled electronic medical record". J Biomed Inform., 40, 1, 44-46, 2007

[8] Barrett T. "NCBI GEO: mining tens of millions of expression profiles—database and tools update", Nucleic Acids Res., 35, D760-D765, 2007

[9] McDonald CJ. "LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update", Clinical chemistry, 49, 624-633, 2003

[10] Ashburner M. "Gene Ontology: tool for the unification of biology", Nature Genet., 25, 25-29, 2000

[11] Kanehisa M. "KEGG: kyoto encyclopedia of genes and genomes", Nucleic Acids Res., 28, 1, 27-30, 2000

[12] Subramanian A. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", Proc. Natl. Acad. Sci., 102, 15545–15550, 2005

[13] Maglott D. "Entrez Gene: gene-centered information at NCBI", Nucleic Acids Res., 35, Database issue, D26-31, 2007

[14] Jansen RC, Nap JP, "Genetical genomics: the added value from segregation". Trends Genet., 17, 388-391, 2001

[15] Parkinson H. "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression", Nucleic Acids Res., 37, Database issue, D868-872, 2009