PALME: PAtients Like My gEnome

Lichang Wang[1], Yong Fang, Ph.D[1], Dima Aref[2], Suyash Rathi[3], Li Shen[4], Xiaoqian Jiang, Ph.D[5], Shuang Wang, Ph.D[5]

[1]College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China;
[2]New Jersey Institute of Technology, Newark, New Jersey;
[3]Electrical Engineering & Computer Science, Syracuse University, Syracuse, New York;
[4]Institute of Biological Sciences and Biotechnology, Donghua University, Shanghai, China;
[5]Department of Biomedical Informatics, University of California San Diego, La Jolla, California

Abstract

PAtients Like My gEnome (PALME) is a webservice that matches patients based on their genome and healthcare profiles. We support two types of inputs: (1) dual query (a variant + phenotype), and (2) genome sequences. For the first type of queries, we will show the patient profile matching the inputs. For the second type of queries, we will calculate similarity (based on Hamming distance) and show the distribution of phenotypes of similar patients given the input sequences of a target patient. Using the publicly available Personal Genome Project (PGP) dataset, we retrieved 4,360 patients' profiles along with their genome data, medical conditions, and treatments. We used a subset of these profiles to build PALME to be an interactive system to support healthcare profile matching. PALME is designed not only for biomedical researchers to support their studies on human genome but also for individuals to explore their own genetics and health. The webservice is accessible at (http://pgp.ucsd-dbmi.org:3838/GenAnaly/PatientGen/#) and the demo videos are available at (https://youtu.be/ycP0rXQizlc)

Introduction

Progress in the understanding of human genome variation identifies rare single nucleotide variants as the most prevalent form of genetic diversity[1]. The accumulation of rare variation in the population results from the recent population expansion and insufficient time for purifying selection. Evolutionary theory indicates that a proportion of those rare variants may carry some fitness cost. However, to this date, there has been limited research on the phenotypic consequences (e.g., traits, diseases, etc.) of the bulk of the observed variants. Genome wide association studies[2] analyzed the phenotypic consequences of common variants (in general, allele frequency $f > 0.05$) by relying on large cohorts providing sufficient statistical power. It is highly unlikely that such statistical approaches will be applicable at very low allele frequencies (e.g., $f < 0.001$), including situations of $n = 1$ patient. Some approaches have explored the use of rare variant gene collapsing strategies[3], the use of biological gene/pathway information, the use of conservation parameters, or other metrics of gene and sequence biological importance. In many cases, only the ingenuity of the researchers can solve the identity of the causal rare variant of a given trait or disease. Increasingly, the biomedical researchers confront the challenge of systematically attributing phenotypic relevance to rare variants (Variants of Unknown Significance, VUS) outside of the setting of a particular trait or clinical disease. Here, each variant needs to be linked with full range of possible phenotypes: PheWAS[4]. The rarest, and thus possibly the variants most likely to carry functional consequences, may fail to be identified through such systematic approaches. The considerations above call for specific solutions to identify the few individuals that may share a rare variant (e.g., 5 individuals among 100,000 or 1 million population). These individuals are likely to be found across different studies, populations, and countries. The required tools are those that search for those rare variants, recover and condense available metadata, and trigger additional phenotyping.

There are a few existing efforts on connecting rare VUS. Geno2MP[5], which stands for Genotype to Mendelian Phenotype browser, allows one to query a gene, chromosomal position or HPO term separately in a centralized database with aggregated variants. GeneMatcher[6] uses a different mechanism, which connects investigators who post the same gene of interest (by gene symbol or base pair position). The Beacon project[7] is another attempt to solve the issue of matching based on one genetic variant but no subsequent phenotype matching is provided. The proposed PALME framework aims to solve rare phenotype/genetic association by allowing comparison and annotation of a large number of rare VUS. Aggregating and sharing data is not the only purpose that PALME intends to achieve. Most significant part of our interests lie in organizing data in a meaningful way and providing

*Address correspondence to Yong Fang at the College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, 712100, China. Tel: +86 02987092619; Fax: +86 02987092353;
E-mail: yfang79@gmail.com.

tools for users to perform genome comparison and relevant statistical analysis. In PALME, clinicians and researchers may deduce potential correlation features of health conditions and rare VUS by analyzing the comparison results. PALME contains not only genome sequences of patients but also related observational healthcare and medication data. These resources enable PALME to serve as a prototype for searching patients like my genome.

Methods

PALME is built on the latest technology advances in R, Rshiny, HTML5, JS, and CSS, which enable webservice to deal with clients request promptly. To utilize the great advantages of statistical and graphical functionalities in R, the backend of our webservice is running on the R Shiny-server[8], which also allows us to adopt R-Shiny to provide an interactive user experience. We use MongoDB, an open source and Non-SQL database, as the backend database to store large-scale genome data and healthcare data in the consideration of irregularities of the datasets. It offers scheme-free feature to developers, by which various formatted data can be dumped into a BSON format (similar to JSON format) without having to be transformed into a table-based representation. All genome data and healthcare data used in our demo webservice are retrieved from the Personal Genome Project (PGP)[9], where over 4360 individuals were available in the PGP, but only 22% of them have complete medical records. We identified over 1000 medical conditions and 430 treatments in the prepressed database. We selected a subset of patients with both genome and medical profiles from 23andMe.

In PALME, four major operations, including search, statistical comparison, match, and statistical analysis, are available for users. As illustrated in Figure 1, datasets with genotype data, gene summary information, condition information, and medication records are stored in a MongoDB database. PALME provides multiple web interfaces to present these pre-organized datasets from the backend. Additionally, we also integrated the Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)[10] into our system. In PALME, we provide an interactive interface for browsing through SNOMET CT for any given record. As aforementioned in this paper, PALME, in particular, aims to develop effective scientific webservice for identifying patients with a similar genetic profile and for revealing the correlation between rare variants and certain phenotype conditions. In the comparison process, we leverage Hamming distance[11] as a metric to quantify the similarity among genome sequences. By uploading genome profiles in Variant Call Format (VCF)[12], PALME service can evaluate the Hamming distance between the querying VCF file and those of participants in our database. Based on Hamming distance, PALME returns sorted list of similar participants in ascending order. Besides, the corresponding conditions and medications will also be presented as outputs through interactive graphic interfaces with both tables and charts. Detailed descriptions of each module will be presented in the next sections.

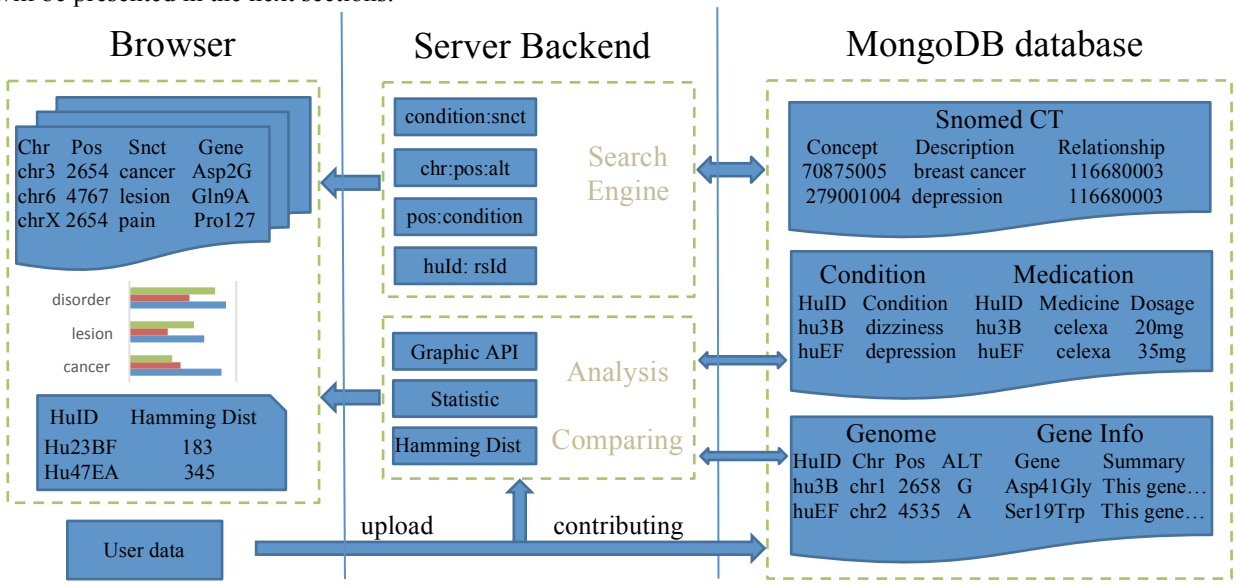

Figure 1. Workflow and key components of the proposed PALME webservice.

Results

Data Organization in PALME

Flexible and effective data query is essential for large-scale data. PALME stores datasets as collections using the MongoDB and creates some necessary collections to represent correlations between datasets to facilitate database access. To take advantages of the index mechanism in the MongoDB, we established several associated indices on attributes that will be frequently requested to remarkably improve the query throughput, where all the indexing processes are handled by mongoDB automatically. With above optimizations, most query results can be returned onto web pages within one second.

Multiple criteria-based searches were also supported by our service to meet different demands of different users. For example, users may have an interest in querying genome sequence by participant ID or by certain variants. To query the sequence of a specific participant in the PGP database, one needs to input participant ID (i.e., huID in PGP project) in the search field. In the same way, relevant information (e.g., summary, computational evidence, functional evidence, published research, etc.) of a gene can be returned from PALME by searching with Gene ID. In searching a gene, system will also match the rsID of SNPs that locate on the gene. In practice, researchers are more likely to deduce statistical conclusions based on several criteria. Searching across multiple criteria can be achieved by using 'colon' to connect the values of these fields. For instance, in Figure 2, the query of 1:75714554:C indicates that a user would like to search records with an allele C at the position 75714554 of chromosome 1 in the database. Furthermore, by selecting the Advanced Search option, users can search both genome data and conditions at the same time. The backend service will automatically generate the corresponding results (like JOIN statement in SQL databases) and display them on the web page as shown in Figure 2. In PALME, we adopted the standard VCF format to organize genome sequences in our database. Users can also check medication information of participants by selecting the related huID in the result table.

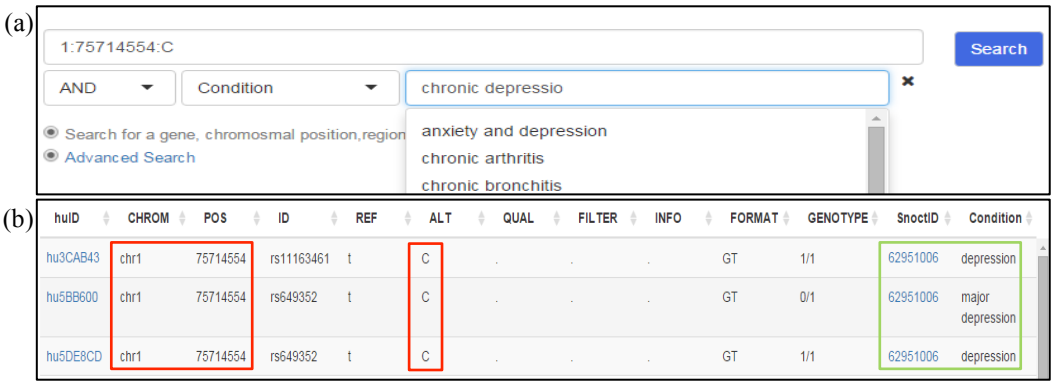

Figure 2. Example of PALME search engine and its results.

SNOMED CT browser

PALME provides the SNOMED CT browsing module with the aims of mapping condition information in our database onto standard clinical terms and further augmenting practicability for healthcare professionals. SNOMET CT has an acyclic taxonomic hierarchy structure, which categorizes all clinical terms and relevant processes in concept. There are four core components in SNOMED CT including concepts, descriptions, relationships, and references. Each component is represented as a text file or a relational table in SQL schema (shown in Figure 3(a)). The concept file defines numerical codes of clinical concepts while description file includes the term of each concept in Fully Specified Name (FSN) or Synonym, where negated concepts are also considered in the SNOMED CT database. The directed hierarchy structure is recorded in the relationship, to identify that one concept '|is a|' type of another concept. One concept may have more than one parent concepts. Similarly, one concept can possess one or more children concepts. As illustrated in Figure 3(b), concept with ID 53430007 is the common child of concepts with IDs 29857009, 279001004, and 290078008. On the other hand, it is a higher level abstract of concepts with IDs 225573003, 237453001, 247415009, and 5522007.

In our database, each personal condition was mapped onto a unique SNOMED CT concept ID (i.e., the SnoctID column in Figure 2(a)). When users search for a condition, the system will match it to the concept automatically. SNOMED CT browser will be invoked, when a concept ID is selected by a user. First, PALME searches the Relationship database of SNOMED CT to find records, whose source ID and destination ID match the concept ID. Source IDs and destination IDs are set to concept IDs in the concept file, identifying the concepts that start and end certain types of relationships. Currently, the platform only supports the '|is a|' relationship in the terminology

hierarchy structure. Second, PALME returns descriptions of source concepts and destination concepts from the Descriptions database. Finally, the results will be rearranged and presented to users on a web page. Figure 3(c) is an example of the SNOMED CT hierarchy in PALME. The green node (i.e., Breast pain) is the target disease while the red nodes represent the synonyms that can be classified as Breast Pain, and the blue nodes are parent concepts of Breast Pain in broader clinical scopes. By clicking the tree nodes, users can traverse the whole SNOMED CT terminology database along with the hierarchy.

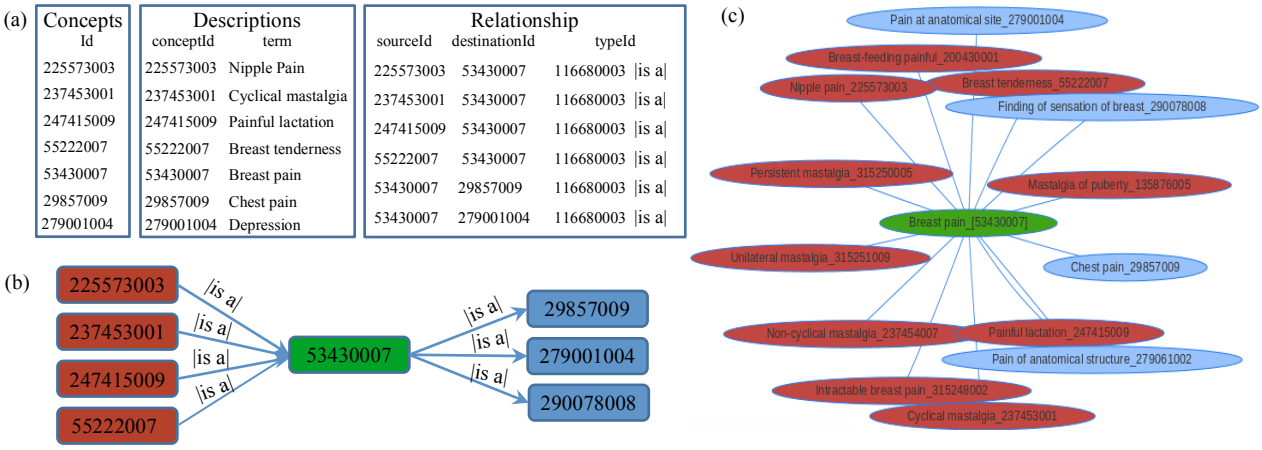

Figure 3. SNOMED CT Browser**:** (a) the actual representation of the hierarchy in SNOMED CT files; (b) an example of hierarchy structure of concept IDs; (c) the user interface in PALME.

Similarity measurement based on Hamming distance

PALME utilizes Hamming distance to measure the similarity between different genome sequences. To reduce database access and redundant comparison, as shown in Figure 4, we align all SNPs' locations in genome datasets to the same reference human genome V3 from 23andMe[13]. Then, we encode the REF and ALT pairs with numerical values according to the Coding table in Figure 4. We also map all columns of these numerical values to one collection in the database to speed up data query and analyses. The coding table in Figure 4 covers all cases that would occur in REF and ALT pairs, such as insertion in which ALT is a dot while REF is not, or deletion with a dot in both ALT and REF fields. The system will return ranked results based on Hamming distances.

Coding table

REF	ALT	CODE
A|G|T|C	A	1
A|G|T|C	G	2
A|G|T|C	T	3
A|G|T|C	C	4
A|G|T|C	.	5
.	A|G|T|C	6
.	.	7

V3 reference table

REF	ALT
225573003	G
256736706	T
3445566635	A
45005655655	.
56760067607	.
96768786654	C
96768786654	A

REF and ALT pair coding table

POS	hu004512	hu204B25	hu8E4A51	huFE653A
225573003	2	4	1	3
256736706	3	4	2	4
3445566635	4	1	5	2
45005655655	6	6	7	6
56760067607	6	7	7	7
96768786654	4	2	5	3
96768786654	5	3	1	5

Figure 4. Coding REF and ALT pair according to Coding table and V3 reference table.

Getting Hamming distance is the first step to enable similarity comparison in our framework. PALME provides an analytical process by combining genome similarity results with patient-level observational healthcare data to reveal the potential relation between genome and health conditions. When Hamming distance between querying sequences and those of participants in our database are calculated, a sorted list of similar patients based on Hamming distance in ascending order will be presented to users. We implemented an interactive graphic interface for users to evaluate the distribution of conditions. For example, as shown in Figure 5, depression has the highest frequency (about 77 patients) among all conditions based on a certain user's query, which may imply that the genome data in the query may more likely be shared by patients with depression. In other words, the person with the querying genome sequence may be more susceptible to depression. Furthermore, we also integrate medication datasets in our database. Therefore, through the medication information on those people with depression symptoms, the potential chronic side effects of some medications frequently used by these patients may be identified by integrating more data.

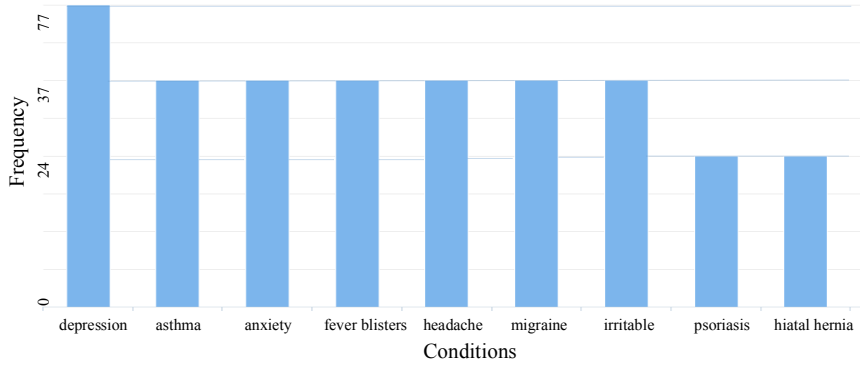

Figure 5. Histogram of health conditions given a certain query.

Discussion and limitations

In this paper, we proposed a PALME webservice to enable users to query similar patients based on their genetic profiles and healthcare data. The proposed PALME framework was evaluated with public data from the Personal Genome Project. In PALME, users can query genome data, patient-level health data, medication data, and SNOMED CT terminologies. The similarity of genetic profiles in PALME is measured by Hamming distance. Interactive graphic interfaces enable researchers to reveal potential correlations among similar genomes, observational healthcare data, and medication data. The proposed PALME service has several limitations. Currently, the PALME prototype only supports small datasets based on the PGP for the initial testing phase. To include more data into PALME, we plan to incorporate publicly available datasets such as the 1000 genome project and HapMap. We highly encourage users to contribute their own data to our platform anonymously. However, it takes time and effort to accumulate additional data. As healthcare data contain sensitive personal information, they can lead to patient re-identification and cause negative impact. This is a major challenge that very limited publicly datasets contain comprehensive patient information, including genomic data, medication records and treatments of the same patients. Another challenge is about the integration of heterogeneous data from different sources. We need to develop specific interface and methods to preprocess heterogeneous datasets using a common data model. The performance of PALME in handling certain queries still needs to be improved. We only considered the Hamming distance for quantifying the similarity. In our future study, we will support other distance measurements. To make it more user-friendly and useful, we will simplify the operations and develop more visualization modules in future work. In this study, we evaluated the PALME webservice with public dataset from PGP without considering the potential privacy risk[14,15], when involving private data. It is important to enhance the privacy protection[16,17], which warrants the further investigation along this line.

Conclusion

In this study, we present a PAtients Like My gEnome (PALME) webservice, which can match patients based on their genotype and phenotype data. Both dual query (a variant + phenotype), and genome sequence based query were supported in PALME. Given a dual query, PALME can output exactly matched patient profiles in the PGP database. For genome sequence based query in VCF file, PAMLE returns matched patient profiles based on Hamming distance similarity in an ascending order, as well as, the distribution of phenotypes of these similar patients. PALME is designed not only for biomedical researchers to support their studies on human genomes, but also for individuals who want to explore their own genetic profiles.

Authors' contribution

LW contributed the majority of the writing and designed the whole PALME webservice. DA collected the data from PGP. SR contributed to the methodology. LS and LW contributed to the SNOMED CT mapping. XJ, SW and YF provided the motivation for this work, detailed edits and critical suggestions.

Acknowledgement

This work was funded in part by the NHGRI (K99HG008175, R00HG008175, and R01HG008802), NLM (R00LM011392, and R21LM012060) and NHLBI (U54HL108460).

References

1. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet. Nature Publishing Group; 2013;14(10):681–91.
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012 Jan 13;90(1):7–24.
3. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. Elsevier; 2011;89(1):82–93.
4. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010 May;26(9):1205–10.
5. Geno2MP [Internet]. [cited 2015 Sep 9]. Available from: http://geno2mp.gs.washington.edu/Geno2MP/#/about
6. GeneMatcher [Internet]. [cited 2015 Sep 18]. Available from: https://genematcher.org/
7. Beacon Project [Internet]. [cited 2015 May 20]. Available from: http://ga4gh.org/#/beacon
8. Shiny-Server1.4 [Internet]. [cited 2015 Sep 9]. Available from: https://www.rstudio.com/products/shiny/download-server/
9. Personal Genome Project [Internet]. [cited 2015 Sep 9]. Available from: http://personalgenomes.org/
10. U.S. National Library of Medicine [Internet]. 2009 [cited 2015 Sep 10]. Available from: http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html
11. Hosangadi S. Distance measures for sequences. arXiv Prepr arXiv12085713. 2012;
12. 1000 Genomes [Internet]. [cited 2015 Sep 18]. Available from: http://www.1000genomes.org/wiki/analysis/variant call format/vcf-variant-call-format-version-41
13. 23andMe [Internet]. [cited 2015 Sep 9]. Available from: https://my.pgp-hms.org/public_genetic_data
14. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. Public Library of Science; 2008;4(8):e1000167.
15. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science (80-). 2013 Jan 18;339(6117):321–4.
16. Wang S, Mohammed N, Chen R. Differentially private genome data dissemination through top-down specialization. BMC Med Inform Decis Mak. 2014 Dec 8;14(Suppl 1):S2.
17. Jiang X, Zhao Y, Wang X, Malin B, Wang S, Ohno-Machado L, et al. A community assessment of privacy preserving techniques for human genomes. BMC Med Inform Decis Mak. 2014 Dec 8;14 Suppl 1(Suppl 1):S1.