Omicseq: a web-based search engine for exploring omics datasets

Xiaobo Sun, Emory University
William S Pittard, Emory University
Tianlei Xu, Emory University
Li Chen, Emory University
Michael Zwick, Emory University
Xiaoqian Jiang, University of California San Diego
Fusheng Wang, Emory University
Zhaohui Qin, Emory University

Journal Title: Nucleic Acids Research
Volume: Volume 45, Number W1
Publisher: Oxford University Press (OUP): Policy C - Option B | 2017-07-03,
Pages W445-W452
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1093/nar/gkx258
Permanent URL: https://pid.emory.edu/ark:/25593/s86h1

Final published version: http://dx.doi.org/10.1093/nar/gkx258

Copyright information:
© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/).

Accessed August 3, 2019 11:50 AM EDT
Omicseq: a web-based search engine for exploring omics datasets

Xiaobo Sun\(^1\), William S. Pittard\(^2\), Tianlei Xu\(^1\), Li Chen\(^1\), Michael E. Zwick\(^3\), Xiaoqian Jiang\(^4\), Fusheng Wang\(^5,6\) and Zhaohui S. Qin\(^2,7,*\)

\(^{1}\)Department of Mathematics and Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA, \(^{2}\)Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA, \(^{3}\)Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Atlanta, GA 30322, USA, \(^{4}\)Health Science Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, \(^{5}\)Department of Biomedical Informatics, Stony Brook University, HSC L3-043, Stony Brook, NY 11794, USA, \(^{6}\)Department of Computer Science, Stony Brook University, Computer Science Building, Stony Brook, NY 11794, USA and \(^{7}\)Department of Biomedical Informatics, Emory University School of Medicine, 36 Eagle Row, Atlanta, GA 30322, USA

Received February 20, 2017; Revised March 27, 2017; Editorial Decision March 31, 2017; Accepted April 04, 2017

ABSTRACT

The development and application of high-throughput genomics technologies has resulted in massive quantities of diverse omics data that continue to accumulate rapidly. These rich datasets offer unprecedented and exciting opportunities to address long standing questions in biomedical research. However, our ability to explore and query the content of diverse omics data is very limited. Existing dataset search tools rely almost exclusively on the metadata. A text-based query for gene name(s) does not work well on datasets wherein the vast majority of their content is numeric. To overcome this barrier, we have developed Omicseq, a novel web-based platform that facilitates the easy interrogation of omics datasets holistically to improve ‘findability’ of relevant data. The core component of Omicseq is trackRank, a novel algorithm for ranking omics datasets that fully uses the numerical content of the dataset to determine relevance to the query entity. The Omicseq system is supported by a scalable and elastic, NoSQL database that hosts a large collection of processed omics datasets. In the front end, a simple, web-based interface allows users to enter queries and instantly receive search results as a list of ranked datasets deemed to be the most relevant. Omicseq is freely available at http://www.omicseq.org.

INTRODUCTION

The rapid development of genomics technologies, such as microarrays and massively parallel DNA sequencing, have dramatically increased the size of experimental data, and the speed at which it arrives. These massive genomics data provide new information at an unprecedented scale and offer an attractive new source for biomedical knowledge. By design, these genome-wide profiling data such as ChIP-seq (1–3) and RNA-seq (4) offer unbiased, genome-wide information from transcription factor binding to histone modification. Measuring by size, these data contain more information, albeit less polished, than the scientific publications that report on them.

Data sharing policies promulgated by the National Institutes of Health of the United States and adopted elsewhere have caused genome-wide profiling experimental data to accumulate rapidly in public repositories such as Gene Expression Omnibus (GEO) (5), the Sequence Read Archive (SRA) (6) and ArrayExpress (7). As an example, GEO has archived 80,690 experimental studies that comprise more than 2,086,234 samples since 2001 (accessed January 10, 2017, http://www.ncbi.nlm.nih.gov/geo/). In addition to individual investigators, large consortia such as the 1000 Genomes (8), the Cancer Genome Atlas (TCGA) (9), International Cancer Genome Consortium (ICGC) (10), the ENCYclopedia Of DNA Elements (ENCODE) (11), mod-ENCODE (12,13) and Roadmap Epigenomics projects (14) are specifically tasked to provide high-quality genome-wide profile data as resources for the research community. Such a large volume of public omics data represents a tremendous resource for biomedical research because these genome-wide profiling data offer unbiased, genome-wide coverage. Therefore, a dataset generated in one study can be (and in

\[^{*}\]To whom correspondence should be addressed. Tel: +1 404 712 9576; Fax: +1 404 727 1370; Email: zhaohui.qin@emory.edu
many cases is explicitly intended to be), used for completely different studies that may constitute secondary or even tertiary analyses of the original data.

Additionally, there are important statistics that are derived from sequencing data using computational and statistical methods. For example, PhastCons score (15), degree of centrality in the protein–protein interaction network, or the probability of being loss of function intolerant (pLI) of a gene, proposed by the Exome Aggregation Consortium (ExAC) (16). Like the above-mentioned genome-wide profiling data, such data also provides useful information from different perspectives. Currently, such data are scattered in isolated places.

Our inability to search through massive, disparate datasets poses a major barrier to biomedical research. Despite the obvious potential and promise of using omics data to address important research questions it is currently very difficult to explore and query omics data. First, genomics data is typically sequestered in disparate data repositories that impair the process of conveniently locating, retrieving, processing, and interrogating the data. Second, most data stored in public data repositories are unprocessed which means that users must process these data themselves prior to use. This often presents significant challenges for biomedical researchers who neither have, nor have access to, bioinformatics expertise. Third, the ability to identify datasets of interest is quite inadequate, given that existing search methods are usually limited to metadata only. Finally, it is a daunting task to rank and select among datasets in terms of their relevancy to a query term among a collection of diverse data types. Due to these issues, we believe most of the public omics data are under-utilized and, as a result, opportunities for novel discovery are being missed.

The search engine is perhaps the most useful tool to explore the internet. We believe a search engine (17) is also critically important for exploring the massive collection of omics datasets. For the internet search engine, the key lies in its ability to accurately rank websites in terms of their relevancy to the query term. For example, the success of Google can largely be attributed to the pageRank algorithm (18), and we believe the ranking idea is also the key to effectively explore massive genomics data.

To address this challenge, we have developed a novel informatics infrastructure named Omicseq that allows users to view, browse, and search processed, ready-to-use omics data. Our platform includes two parts: (i) a scalable and elastic database storing processed omics, or genome-wide profiling datasets produced from experiments such as ChIP-seq, RNA-seq and DNase-seq; (ii) a web interface to access, browse and search for omics data. Akin to an internet search engine, given a query entity (e.g. gene or gene set) researchers can easily identify important data by viewing a ranked listing of the most relevant genomic data sets related to the gene or gene set. What we want to achieve is to integrate diverse and disparate quantitative genomic information into a searchable format such that every aspect of the gene can be put in context and compared and contrasted in terms of its genome-wide significance in a comprehensive manner. The overarching goal of the Omicseq project is to develop enabling technologies to facilitate data-driven biomedical research that makes optimal maximum use of existing public omics resource.

**MATERIALS AND METHODS**

**Processing different data types**

For each dataset, we first retrieve sample metadata and source file information and import that into a metadata database that will complement the subsequently processed data. In cases wherein pre-processed data are directly available from the source, such as level 3 data in TCGA, we can download and import the data directly. Otherwise, we next download raw data (typically in FASTQ format) and subject them to a data processing workflow that makes use of third-party software such as SRATOOLKIT (19), bowtie2 (20) and TCGA assembler (21) to transform raw data into appropriate formats such as SAM, and BED. Subsequently, depending on the type of experiment, we call upon appropriate pipelines to calculate gene rank and percentile statistics which are maintained in the database.

**ChIP-seq.** Reads from ChIP and input samples will be mapped to the hg19 human reference genome separately using BWA (22,23). We calculate and save two types of scores, one for gene bodies and one for promoters (defined to be 5 kb up- and down-stream of the transcription start sites (TSS). This is because the promoter region is useful in the study of transcription factor (TF) binding, and the gene body is of interest for studying some histone marks such as H3K36me3. For each gene, the promoter score is calculated as the difference between ChIP read count and input read count. Next, the gene body score is calculated as the difference between ChIP read count and input read count normalized (divided) by the length of the coding region. Lastly, the two sets of scores are ranked and converted to percentiles separately.

**DNase-seq.** Similar to ChIP-seq data, we record the difference between numbers of reads from DNase sample and the control sample in the promoter regions of all the genes. The read count differences are then sorted and converted to percentile.

**RNA-seq.** We use the transcript abundance measure represented by read per million per kilobases mapped (RPKM) value (4) as the gene-based score. We choose RPKM since it is a widely-used gene expression measure. We are aware of alternative measures such as TPM (24) and RSEM measures (25). We plan to provide such alternative measures in the future release of Omicseq.

**Copy number variation.** Copy number variation (CNV) data is processed the same way as in the CNV analysis pipeline used by the TCGA consortium (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/) in which the estimated copy number for the coding region of each gene was transformed into the segment mean value, which is defined as log2(copy-number/2). Normal copy number (diploid regions) will have a segment mean of zero, amplified regions will have positive values, and deletions will have negative values. The absolute values
of the segment means will be sorted and converted to percentiles.

**Methylation.** Currently, all genome-wide methylation profiling data stored in Omicsq are obtained from array-based technology (such as the Infinium HumanMethylation 450 BeadChip array from Illumina Inc, (San Diego, CA, USA)). The measure of methylation is taken as the beta values, calculated from array intensities as beta = M/(M + U) using the minfi package (26). For each gene, the average methylation measure of CG sites falling into its promoter region (5 kb upstream or downstream of the TSS) is used as the quantitative measure for each gene and subsequently converted into percentiles after ranking.

**GRO-seq.** The global run-on-sequencing (GRO-seq) assay (27) is a sequencing-based assay used mainly to evaluate promoter-proximal pausing on all genes. We treat the pausing index of each gene as the gene-based measure, and subsequently convert it into percentiles after ranking.

**Microarray gene expression.** For microarray data, typically processed gene-level gene expression data are available from the lab where the experiment is accomplished. We take these measures, sort and then convert them to percentiles.

**Somatic mutations.** Currently, the number of single nucleotide somatic mutations found within the coding region of a gene is tallied, sorted and converted to percentiles. Perhaps a better approach is to give known cancer-related genes higher weight due to its elevated importance. As an example, the cancer gene census project (28) of COSMIC (29) maintain a list of known cancer genes. By doing this, when querying these cancer genes, somatic mutation count data will be ranked higher. We are implementing this weighting scheme and will adopt it in the next release of Omicsq.

**DNA sequence motif.** The number of a specific type of detected motif (such as CTCF) that is located within the promoter region (within 5 kb upstream and downstream of the TSS) for a gene is taken as the gene-based measure. These values are then sorted and converted to percentiles.

**Summary statistics.** Each of the datasets (a track) we have described thus far represents a single sample. In some cases, the study-level summary statistics is of greater interest. The average measurement from all samples in a study (e.g. the average gene expression level among all prostate cancer tumor samples in TCGA) is taken as the measurement. These summary statistics are then sorted and converted to percentiles.

The trackRank algorithm

For omics datasets, a typical query entity is either a gene name or a set of gene names. Hence the key is to effectively rank omics datasets of diverse types given a query. The trackRank algorithm is designed to facilitate such ranking. There are three steps in trackRank. First, as described above, for each dataset (track), regardless of the data type, we can obtain a numerical and continuous score (such as promoter region read counts for ChiP-seq, RPKM values for RNA-seq) for each gene. Second, to make the scores comparable across tracks, within each track, we convert all the scores into percentiles, which is straightforward for a set of fixed features (genes). Third, datasets are ranked based on the percentiles of the query gene. Here we present a detailed description of the underlying trackRank component which is at the heart of the Omicsq system.

**trackRank.** Suppose there are $n$ datasets (tracks) in the omics data collection. We use $x_{ij}$ to denote the gene-based score of gene $j$ in the $i$th dataset where $i = 1, \ldots, n, j = 1, \ldots, n_i$, $n_i$ is the number of genes with scores in the $i$th dataset and this number may not be the same across different datasets.

At the beginning, we sort the gene-based scores within each dataset, i.e. sort elements in each of the vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$, from the most significant to the least significant (depending on the data type, for RNA-seq data, from high to low, for p-value data, from low to high, for CNV data, take absolute value, then from high to low), next, convert every $x_i$ to $p_i = (p_{i1}, p_{i2}, \ldots, p_{in_i})$ where $p_{ij}$ is the corresponding percentile of $x_{ij}$ in $x_i$ which is defined as

$$p_{ij} = \frac{\sum_{k=1}^{n_i} I(x_{ik} \leq x_{ij})}{n_i}$$

Suppose we are querying gene $g$. First, among all the percentiles that correspond to gene $g$ in these $n$ datasets, $(p_{1g}, p_{2g}, \ldots, p_{ng})$, select a subset such that they satisfy the significance criteria, say <1%. Suppose there are $m$ such percentiles (or $m$ datasets). Then sort these $m$ percentiles into ascending order $(p_{s(1)g}, p_{s(2)g}, \ldots, p_{s(m)g})$ such that $p_{s(1)g} \leq p_{s(2)g} \leq \cdots \leq p_{s(m)g}$ and $p_{s(k)g} \leq 0.01$ where $s(k)$ is that original order of the dataset, $k = 1, \ldots, m$. As a result, these $m$ datasets are considered significant given the query gene and they will be displayed in the order $s(1), s(2), \ldots, s(m)$. This procedure is illustrated in Figure 1 using a simple example. To speed up the search, we pre-calculate the percentile for each gene and pre-sort them within each dataset. Therefore only one round of sorting is needed to return the query result.

**trackRank set.** In many studies, it is often of interest to find the biological properties for a set of genes. For example, a group of genes located in the same region of the genome or belonging to the same biological pathway.

When querying multiple genes, the key is to estimate the ranking percentile for an arbitrary set of genes. To achieve this, we derive a statistically-sound approach to derive combined ranks for a group of genes, and thus assess the significance of the query gene set in each track. Suppose the query is a group of $k$ genes. Using the same notation, it is reasonable to assume that percentiles $p_{ij}$ are independent random variables following a Uniform distribution between 0 and 1. Let $q_j = \sum_{i=1}^{k} p_{ij}$, when $k$ is large (say, $k > 10$), we can use Normal approximation to calculate the percentile of $q_j$ courtesy of the central limit theorem (CLT). We have $\sqrt{k}(\frac{q_j}{k} - \frac{1}{2}) \sim N(0, \frac{1}{k})$. All we need to do next is to rank all the tracks based on $q_j$'s
Figure 1. Illustration of ranking genomic datasets of different types. In this illustrative example, assume there are exactly eight genes (A–H) in the genome. We have five datasets (001–005) to rank. (A) The values in the table are the gene-based scores. These scores are different for different data types which are represented by different row colors. (B) Gene-based scores are sorted and converted to percentiles (1/8–8/8). The shade of the color reflects the order. (C) Suppose a user query gene E, then we sort the five percentiles (bolded) for gene E in these five datasets. And then reorder the five datasets based on the five percentiles. For example, gene E is the most significant gene hence has the lowest percentile in dataset 003–1/8, so dataset 003 is ranked at the top. Dataset 005 has the second lowest percentile for gene E, so it is ranked the second. Suppose the significance threshold is 30%, then only datasets 003 and 005 will be considered significant for gene E and displayed as the search result.

RESULTS

The current release of the Omicseq search engine

We provide a Web based portal (http://www.omicseq.org) as the user interface to provide query capability on individual genes or a pathway with the intent to accommodate more general query types in the future. This website is free and open to all users with no login requirement.

Currently, Omicseq contains 50,484 unique, high quality genome-wide profiling datasets. The vast majority are from Human and the rest are from Mouse. The majority of these data is collected from large international consortia including: 36,694 datasets from TCGA, 3,935 from ENCODE and 2,331 from Roadmap Epigenome, 2,079 from Cancer Cell Line Encyclopedia (CCLE) (30), 661 from ICGC, 660 from GEUVADIS (31). We included diverse data types including ChIP-seq, RNA-seq, DNase-seq, CNV, methylation, GRO-seq, microarray gene expression, DNA motifs and summary-level data. More data types will be added later. Currently, for query terms we allow a single gene name or a known pathway name. There are 32,745 Human gene names (from RefSeq version hg19) in our database. Gene queries are flexible as we accept gene names as well as known aliases. For example OCT4, an alias for POU5F1 will be recognized. There are 10,023 pathways (from mSigDB) stored in Omicseq that may be queried. Users can select them from a list arranged in alphabetical order or enter text corresponding to the pathway of interest. Partially typed gene or pathway names will be auto-completed for the convenience of the user. We are also beginning to provide gene search capability on mouse data. There are 30,493 Mouse gene names (from RefSeq version mm9) in our database. More mouse data and the pathway search capability will be added in the next release.

After a user submits a search, a result page will be displayed. Directly underneath the search box, Omicseq reports the total number of datasets containing the scores of the query gene (a datasets may not contain scores for all genes), and the number of datasets deemed relevant (that the query gene ranked within the top 1% among all genes in that dataset). A link is also provided for users to view a pie chart that show the breakdown of data types among the top ranked datasets. For the convenience of users, we also supply a number of external links about the query gene such as PubMed, Wikipedia and WikiGenes (32).

The main part of the result page is a list of the most relevant datasets determined by the trackRank algorithm. For each dataset, key metadata is displayed, including cell type, factor/experimental condition and source of the data. In addition, we provide buttons to allow curious users to
know more about the dataset: a ‘Meta Data’ button link to a popup window to display more metadata about the dataset; a ‘PubMed’ button link to the publication that this dataset is originated; a ‘GEO’ button (the name is used metaphorically to indicate a public data repository) link to the public repository where the dataset is stored and an optional ‘download’ button, if the raw dataset has a direct download link.

Like PubMed, Omicseq also provides an ‘advanced search’ function that allows users to narrow their searches. For example, the search can be constrained to a specific experiment type, such as ChIP-seq, RNA-seq or DNase-seq. It will also be possible to constrain searches to a specific data sources such as the ENCODE (11), TCGA (9) or Roadmap Epigenomics (33), or combination of multiple ones. Querying selected cell lines, such as MCF7, LNCaP or selected histone marks such as H3K4me3 is also supported.

We believe the Omicseq system is intuitive to use. Nevertheless, we have created a tutorial document with detailed instructions. This tutorial page (http://www.omicseq.org/tutorial.htm) can be easily accessed from any page within the Omicseq web portal. In addition, we have created a video tutorial, which is embedded in the tutorial page and accessible at https://youtu.be/tfmjh6ADVu0.

Our recent testing indicates that a typical gene-based search takes only 0.1 seconds with a nine node MongoDB setup. In general, loading/ingesting data into our database takes significant time though by implementing a parallel loading strategy, we are able to achieve a loading metric of 7.3 s per dataset. Our performance study also demonstrates that MongoDB based indexing is highly scalable on query and loading performance. Given that data loading is an incremental one-time cost, we expect data ingestion to be prompt.

System architecture

The Omicseq system is based upon the SpringMVC architecture using Spring 4.0 and Servlet 3.0 techniques. It runs as a cluster of databases and web servers on Amazon Web Services Elastic Computing instances, which can be easily scaled in response to demand. The database server cluster is built on a sharded TokuMX database with the supporting web servers using Apache Tomcat for the user interface application. As shown in Figure 2, the overall system consists of seven major subsystem components including: (i) the webcrawler system (under construction); (ii) the source data download system; (iii) the data processing system; (iv) the cache system; (v) the database system; (vi) The task-scheduling system and (vii) the web application service system. They work collaboratively to collect and process data from multiple data sources, and provide query services to users.

Considering the large size of the data, two mechanisms were adopted to optimize the overall system performance. First, we implemented a task scheduling system initiated at server startup which manages the assignment and reclamation of computing resources such as webpage crawling, data downloading and processing tasks. Second, a memcached system has been implemented on a separate server to alleviate database workload and boost system performance and improve query speed. With memcached, frequently queried data are preloaded into memory at startup. In addition, intermediate data generated between data processing steps are also cached for later use. So, when a request for such data is issued, either by users or internal system components, the memcached system is first consulted to avoid redundant database lookups. If the data is not found then the query would be routed to the database. This approach significantly reduces the load on the database system and thereby improves query speed.
Omicseq web server

The Web portal is based on the well-performing Apache Server with tight integration of the Tomcat Application Server. For purposes of scalability, high availability, and security the infrastructure has been deployed on the Amazon Web Services Elastic Computing environment. The Web portal application has two layers: application logic layer and presentation layer. The application logic layer provides the mapping of queries into various backend database operations or computation operations, caching management of repeated queries, and search job management of multiple query requests. Many common operations in the application layer are implemented as RESTful WebAPIs to facilitate easy integration with various applications. REST is a software architecture style for distributed hypermedia systems such as the Web. REST is lightweight in representation, and can be used to build applications easily. In RESTful Web Services, data are viewed as resources, which can be identified by their URIs. Normally implemented on the HTTP, RESTful Web Services are very efficient for transporting data over the Web. We define a set of common queries as RESTful Web Services to allow flexible integration of the search engine into users’ applications.

The presentation layer provides the actual interfaces for interactive searching and visualization of query results. The front page provides a search box where a user can type in a keyword and select the corresponding query type to start a search. Auto-completion is implemented using an Ajax-based request based on the JQuery framework. The Ajax request is supported by a RESTful Web API which generates possible completed keywords based on a collection of keywords related to common names or IDs such as gene names stored in the indexing database.

The online search portal is built on top of a NoSQL databases (MongoDB is used to manage index data for gene search). The web portal provides modeling, managing and searching of results from the databases on top of an Apache Web Server integrated with Tomcat Application Server.

Database

The database system consists of a cluster of five nodes and can be functionally divided into two components. The first component is a single, unsharded database built on MongoDB 2.4.9 that stores the extracted metadata of the crawled raw data. For example, metadata information including sample ID, source type, data-processing state are stored in a collection indexed on sample ID. The second component is a distributed sharded database built on TokuMX 2.0.1 that extends the MongoDB protocol but with additional performance-enhancing features such as improved multi-threaded I/O, high throughput transaction support and efficient creation and management of clustering indexes. This second component includes: (i) three Shard servers where the bulk of data are maintained, (ii) a Configuration server which holds metadata about which shards hold what data and (iii) a Service server which acts as a proxy to route client requests to the appropriate shard (Figure 2). The shard servers maintains about one billion records of gene rank and other types of information. In such a large data scenario, the choice of shard key is a critical step in optimizing the system performance. Since almost all client queries are based on gene ID or pathway ID, these are used as a primary shard key to horizontally split the data of the corresponding collection into separate shards. The internal balancer of TokuMX automatically balances the data load among shard servers. In addition, since some queries may be based on sample ID or need sorting according to percentile value, clustering indexes are also built on these fields which significantly improve range queries as well as data migration among shards.

User cases

As an example of Omicseq’s utility, it has long been established that prostate-specific antigen (PSA), encoded by KLK3 gene, is an important biomarker for prostate cancer (34). When we query the KLK3 gene, it turns out that RNA-seq data on prostate tumor samples from TCGA and ICGC show up on top. Hence, users who query KLK3 will immediately recognize the importance of this gene for prostate cancer. Similarly, a query of the HER2 gene (ERBB2) identified hundreds of RNA-seq data from tumor samples of breast cancer patients in TCGA. As another example, when query PTEN, the top ranked datasets are dominated by CNV data show strong deletion, which clearly indicates that PTEN is a tumor suppressor. Search results for these three genes can be found in Figure 3.

Despite the vast volume of the biomedical literature, many of the genes in the human genome receive little coverage. When querying all genes in the human genome against PubMed, we found ~30% of the genes are not mentioned in any paper, and the median number of publications is only six for all human genes. For example, for non-coding RNA (ncRNA) SLC04A1-AS1, a PubMed query returns zero record. But an Omicseq search returns 278 datasets in which this ncRNA gene has a score that ranked within the top 1% among all genes in the genome. Among the top ranked datasets, there are multiple ChIP-seq datasets of histone marks H3K4me1 and H3K36me3, both of which show significant enrichment in the promoter region of this gene in cancer cell lines HepG2 and HeLa-S3. There are also 205 CNV datasets, from various tumor samples collected by TCGA included among the top ranked datasets. All these datasets point to potential functional connection of this ncRNA gene with cancer.

DISCUSSION

Literature is the dominating source of biomedical knowledge today. The explosion of massive genomics data offers an attractive, alternative source for biomedical knowledge since these assays interrogate a large number of genes which provides a somewhat unbiased (no vetting from investigators), comprehensive view of the genome. As a result, genome-wide profiling experiments are supplementing the traditional hypothesis-driven research paradigm with a data-driven paradigm. However, key informatics infrastructure needs to be developed in order to make these omics data a truly useful resource. One of the major aims of the recent Big Data to knowledge (BD2K, https://datascience.nih.gov/bd2k) efforts (e.g. bioCADDIE) is focused on mak-
Figure 3. Search interface and result pages for KLK3, ERBB2 and PTEN. Most relevant datasets are displayed with links to metadata and paper, etc.

...ing biomedical data searchable and reusable to speed up discovery (35). Here, we describe our attempt to develop such an informatics infrastructure to fulfill this aim. A unique strength of our approach that distinguishes us from pure genomic browsers is our ability to automatically rank and prioritize among thousands of diverse datasets and display only the ones that are considered ‘interesting’. The ultimate overarching goal of Omicseq is to enhance findability of omics datasets, to facilitate re-utilization, or re-purposing of existing data for secondary, tertiary analyses.

Since we are pooling diverse datasets from multiple sources, maintaining high quality standards among the datasets is critically important and we are taking this seriously. We plan to add quality metrics for datasets collected from places other than large international consortia (where data quality is typically high). In addition, we have deployed a ‘crowd sourcing’ type of approach which allows users to leave comments on the data quality. These comments can alert future users about the potential quality issues so cautions can be exercised to avoid making conclusions based on faulty data.

Despite the significant effort put into Omicseq, it is still (and will continue to be) a continual work in progress. So far, datasets are collected manually. This is manageable for collecting datasets en masse from those large consortia. For public data repositories such as GEO, the format, file type and annotations vary substantially. We aim to collect data more efficiently and will use a web crawler system to discover and collect newly emerging omics data across multiple sources.

Genome-wide profiling experiments offer great opportunities for ‘data-centric’ knowledge discovery, which we believe represents an innovation that significantly improves upon, and effectively augments, the traditional knowledge discovery approach relied upon in text mining. In this sense, Omicseq is a novel tool that can facilitate discoveries using existing and emerging genomic datasets.

ACKNOWLEDGEMENTS

We want to thank the executive editor, Dr Gary Benson and two anonymous reviewers for their constructive comments and suggestions which we found extremely helpful. We thank Mr Yanjia Wang, Zhengyu Zhang, Qiushen Zhong, Min Wang and Ms Daojia Wu for coding and logistic help. We thank Dr Jindan Yu for helpful discussion at the early stage of the project. We thank the bioCADDIE team for support and valuable input.

FUNDING

Emory Integrated Computational Core (EICC), one of the Emory Integrated Core Facilities, which is subsidized by the Emory University School of Medicine and by the National Institutes of Health [UL1TR000454 to M.E.Z.]; Patient-Centered Outcomes Research Institute [ME-1310-07058 to X.J.]; National Institute of Health [R01HG008802, R01GM114612, R01GM118574, R01GM118609, R21LM012060, U01EB023685 to X.J.]; National Science Foundation [ACI 1443054, IIS 1350885 to F.W.]; National Institute of Health [P01GM085354 to Z.S.Q. and W.S.P.]. Funding for open access charge: Department fund.

Conflict of interest statement. None declared.

REFERENCES


