Data and text mining

database.bio: a web application for interpreting human variations

Min Ou1,†, Ricky Ma2,†, Jeanno Cheung2, Katie Lo2, Patrick Yee2, Tewei Luo1, T.L. Chan3, Chun Hang Au3, Ava Kwong4, Ruibang Luo1,2,5,* and Tak-Wah Lam1,2,*

1HKU-BGI Bioinformatics Algorithms Research Laboratory and Department of Computer Science, University of Hong Kong, 2L3 Bioinformatics Limited, 3Department of Pathology, Hong Kong Sanatorium and Hospital, 4Department of Surgery, University of Hong Kong and Hong Kong Hereditary Breast Cancer Family Registry, Hong Kong and 5United Electronics Co., Ltd., Beijing, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on March 29, 2015; revised on August 18, 2015; accepted on August 18, 2015

Abstract

Summary: Rapid advances of next-generation sequencing technology have led to the integration of genetic information with clinical care. Genetic basis of diseases and response to drugs provide new ways of disease diagnosis and safer drug usage. This integration reveals the urgent need for effective and accurate tools to analyze genetic variants. Due to the number and diversity of sources for annotation, automating variant analysis is a challenging task. Here, we present database.bio, a web application that combines variant annotation, prioritization and visualization so as to support insight into the individual genetic characteristics. It enhances annotation speed by preprocessing data on a supercomputer, and reduces database space via a unified database representation with compressed fields.

Availability and implementation: Freely available at https://database.bio
Contact: rb@l3-bioinfo.com or twlam@cs.hku.hk
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The decreasing cost of sequencing technologies and growing knowledge in the clinical relevance of genomic data have led to rapid adoption of genetic tests in healthcare (Feero et al., 2011). While huge amount of data is generated through these tests, the management and analysis of these data present many challenges as discussed hereafter. Variant annotations come from diverse sources, which may not use the same coordinate system and naming convention. Once annotations have been integrated, these database files consume large storage space and queries to the databases can be very slow without proper indexing and preprocessing. Existing tools have solved some of the stated issues but are lacking in other features. For example, some software can visualize records from different databases across the genome, such as the UCSC genome browser (Kent et al., 2002) and Alamut Visual 2.4 (Interactive Biosoftware; last accessed August 5, 2014). However, it is hard to display all relevant annotations, such as relationships between genes and pathways, by only using a genome browser. Moreover, the scale of genome browsers is usually small as they are designed to show information of small genomic regions. It is also not possible for users to sort or filter variants in the browser. In contrast, tables can better summarize variant information and help users prioritize their findings. QIAGEN’s Ingenuity Variant Analysis software (IVA) (www.qiagen.com/ingenuity) is a popular table-based tool and allows users to configure a filter cascade for data prioritization. However, without an embedded genome browser, it is inconvenient for users to picture genomic characteristics of variants and study-related data from databases integrated in IVA. A similar tool, Tute Genomics (www.tutegenomics.com); last accessed January 30,
2 Materials and Methods

In database.bio, variant information is retrieved from different databases based on join conditions such as genomic positions, gene-drug, gene-disease and gene-pathway relationships. To unify the coordinate scheme of databases (see Supplementary Methods for details), sequence changes at RNA level are converted to DNA level using the Mutalyzer Position Converter (http://www.I.OV.D.nl/mutalyzer/). In addition, the left-aligned HGVS G-dot notation is adopted while diseases, drugs, and pathways are linked to related genes using the HGNC symbol (Supplementary Fig. S1). In order to allow users to access the original data from which annotation is derived, hyperlinks to the original data source are often provided.

Annotation is a very time-consuming part of data processing. In order to reduce annotation time, preprocessing is performed to generate annotation of all possible human single nucleotide polymorphism (SNP) variants at each genomic position. Afterwards, annotation information for any individual case can be retrieved quickly. There are ~8.6 billion possible human SNP variants to be preprocessed; using state-of-the-art annotators such as VEP would require over a 100,000 CPU hours. To this end, we used a supercomputer with 1000 nodes. The annotation information used by database.bio were fetched from more than 25 curated databases (see Supplementary Methods for details): (i) databases for variant annotation, (ii) databases for gene annotation and (iii) databases for real samples, such as ICGC (Zhang et al., 2011) and TCGA (The TCGA Research Network, http://cancergenome.nih.gov). As the data size in database.bio is very large, it is time-consuming to check the three key fields (chromosome, genomic position and nucleotide change) for every query. Moreover, the size of index files can be large when multiple fields are indexed, lowering query performance. For some large tables in database.bio, these three key fields are merged into an integer by bitwise operations, which result in smaller index files and faster query speed (Supplementary Materials).

3 Visualization

We used eight tools (see Supplementary Methods for details) to generate visualizations of various statistical properties of the input variant list. These tools include Circos (Krzywinski et al., 2009) and BCFTools (Li, 2011). We have also built our own genome viewer, named Genome Browser, which consolidates and displays information from most of the relevant databases in one single viewer. For a specified genomic region, the Genome Browser displays annotation details such as variant, gene and orthology information of human and 14 other mammals from related databases. In addition to data derived from VCF files, content from SNAPSHOT (Luo et al., 2014) files and alignment details from BAM files can also be plotted (Fig. 1).

In order to visualize the effect of variants on protein structures, protein-domain graphs and 3D structures of proteins with affected amino acid highlighted are provided. Similarly, pathway graphs with mutated gene highlighted are generated for each affected molecular pathway.

4 Usage

Database.bio accepts input in VCF, 23andMe, BAM, and SNAPSHOT formats. After authentication, users can browse, delete existing analyses or upload new files. On the variant upload page, users can augment the default categorization rules to prioritize the variants (Fig. 2) with a user-friendly filtering cascade to give higher priority in view of user-interested genes, population statistics and biological consequences. The customizable filtering cascade allows users to focus on a small number of variants based on different criteria. The categorization rules are similar to the Emory Genetic Laboratory’s classification definitions (Bean et al., 2013), which rely on terms such as consequence-of-variants and clinical-significance-type from dbSNP and EVS. In addition, we considered relationships among genes, genetic disorders, clinical significance (as defined in ClinVar), predicted categories from CADD and Align-GVGD prediction in formulating the categorization rules. The variants are then filtered based on user-defined filters and prioritized through classification based on associated disease information.

Fig. 1. Screenshot of Genome Browser
After annotated data are uploaded and processed, each variant is categorized into one of five severity levels (pathogenic, likely-pathogenic, Variant of Unknown Significance (VOUS), likely benign and benign). Per-column sorting and searching are also available. The variant count in each severity level is given in table entries on the Gene, Disease, Drug and Pathway Summary pages. The Genome Browser is embedded in the Variant Detail and Gene Detail pages. Details of samples from TCGA or ICGC with related variants are shown on the Variant Details and Disease Details pages. Furthermore, clinical trial records of diseases and drugs are provided on the Disease Detail and Drug Detail pages. These features can help researchers to efficiently integrate variant related information and form a whole picture of their data more quickly.

5 Conclusion

Database.bio, empowered by a space-efficient integration of diverse sources of annotations, provides an efficient way for interpreting human genomic variations and supports interactive visualization. As database.bio is a web-based application, the way users retrieve and manage data have been specially designed to ensure a security level up to the standard of HIPAA.

Acknowledgements

Thanks for the insightful comments of Dr. E.S.K. Ma, and the indispensable supports from Dr. Ruiqiang Li.

Funding

This research is benefited from GRF grant HKU-713512E and ITF grant GHP901112.

Conflict of Interest: none declared.

References


