BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 8 2011, pages 1155–1156
doi:10.1093/bioinformatics/btr100

Genome analysis

Advance Access publication March 2, 2011

Human variation database: an open-source database template for genomic discovery

Anthony P. Fejes*, Alireza Hadj Khodabakhshi, Inanc Birol and Steven J. M. Jones
Genome Sciences Centre, BC Cancer Agency, Suite 100 570 West 7th Avenue, Vancouver, British Columbia, Canada V5Z 4S6

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Current public variation databases are based upon collaboratively pooling data into a single database with a single interface available to the public. This gives little control to the collaborator to mine the database and requires that they freely share their data with the owners of the repository. We aim to provide an alternative mechanism: providing the source code and application programming interface (API) of a database, enabling researchers to set up local versions without investing heavily in the development of the resource and allowing for confidential information to remain secure.

Results: We describe an open-source database that can be installed easily at any research facility for the storage and analysis of thousands of next-generation sequencing variations. This database is built using PostgreSQL 8.4 (The PostgreSQL Global Development Group. postgres 8.4: *http://www.postgresql.org*) and provides a novel method for collating and searching across the reported results from thousands of next-generation sequence samples, as well as rapidly accessing vital information on the origin of the samples. The schema of the database makes rapid and insightful queries simple and enables easy annotation of novel or known genetic variations. A modular and cross-platform Java API is provided to perform common functions, such as generation of standard experimental reports and graphical summaries of modifications to genes. Included libraries allow adopters of the database to quickly develop their own queries.

Availability: The software is available for download through the Vancouver Short Read Analysis Package on Sourceforge, http://vancouvershortr.sourceforge.net. Instructions for use and deployment are provided on the accompanying wiki pages.

Contact: afejes@bcgsc.ca

Received on November 5, 2010; revised on February 8, 2011; accepted on February 17, 2011

1 INTRODUCTION

With the introduction of next-generation sequencing technologies, there has been a rapid increase in the amount of genomic and transcriptomic sequence data generated. Consequently, an ecosystem of bioinformatics applications and resources has grown up around the ever-increasing volume of data being produced, including aligners, assemblers, repositories and protocol specific applications—mainly designed to investigate one genotype at a time. Other databases and tools are available for performing analyses of small groups or conducting genome wide association study (Fong *et al*., 2010); however, a significant keystone of this ecosystem has yet to be introduced: a simple way to collate and search across large numbers of sequenced libraries, particularly those for which the analysis can not be supplied by tools external to the organizations collecting the data.

Databases have long been a staple of bioinformatics research, with the most common organizational model based upon collaboratively pooling data into a single publicly available database. This allows the developers to maintain control of the hosting and formatting of the data, and also requires that collaborators should freely share their data with the owners of the repository. In contrast, our project provides an alternative mechanism: providing the source code and application programming interface (API) for the implementation of a local database, enabling researchers to set up private repositories without investing heavily in the development of the resource and allowing for confidential information to remain secure. This is an especially important model, when the genomic data is of medical nature and cannot be publicly shared. However, distributing the burden of development can significantly provide a collective bootstrap to researchers involved in the collection and analysis of private genome or genome-derived datasets.

There are several advantages to collecting genomic information in the proposed database versus analysis of a set of independent DNA and RNA sequencing experiments. In contrast to publicly available databases, such as dbSNP (Sherry *et al*., 2001), a local repository can be used to track more data about each observation and more information can be stored about the library of origin. Thus, common variations will quickly become easy to identify, as will those that cluster into broad categories. This makes it easy to conduct metagenomics experiments on larger datasets and enables data mining for each new set of sequencing data, building on previously performed experiments. Finally, the database itself does not require any genomic annotation system, and annotations can be imposed at the time of data export. We provide an API that utilizes an Ensembl database (http://agd.vital-it.ch/info/software/java/index.html) to identify and analyze genes/exons, applicable for most analysis needs.

2 METHODS

Novel functions: the database enables four novel functions that would be otherwise difficult to accomplish: (i) the ability to rapidly access genetic variation information across multiple datasets (e.g. to determine the frequency of a change in the population); (ii) storing information in an annotation-free manner, allowing the user to select the appropriate annotation

*To whom correspondence should be addressed.

set to use (e.g. selecting which version of ensemble annotations to use); (iii) rapid comparison of data from any sequencing platform, regardless of the origin; and (iv) perform aggregate analyses, such as the assignment of validation probabilities to observed variants based on sequencing characteristics across a population. These functions are accomplished by storing each variation observed in a manner indexed both by the library to which they belong, as well as the location of origin in the genome to which the reads were aligned.

Data: the database currently stores single nucleotide variations and polymorphisms (SNV/SNPs), as well as deletions and insertions (indels). We also import information from external databases such as dbSNP in the form of annotations. These data can be used for concordance analysis and to asses the quality of each imported dataset.

Interface: we include a Java API to handle communication with the database, both for managing insertion and deletion of records, as well as for extracting and analyzing recorded information. This API layer uses function from the Ensembl (Hubbard *et al*., 2007) Java API (http://agd.vital-it.ch/info/software/java) to gain genome annotation information used for queries such as the ExperimentalRecord summary.

ExperimentalRecord: a common task is to search for variants present within a single sample, and to compare them to variants in other samples in the database. To simplify this process, we have produced a standard, ExperimentalRecord query that provides a summary of the non-synonymous variants (indels and single nucleotide substitutions) and variations likely to interfere with splicing junctions. This report generating script makes use of the Ensemble API for Java to obtain annotations for genes, exons and transcripts, and to determine which variations in the database are likely to affect coding regions.

Concordance: scripts are provided to compute concordance with SNPs and indels from dbSNP. In the absence of validated sequencing results, this provides an approximate method for assessing the quality of calls with an expected background for the overall population. For indels, the concordance API calls accept a window parameter in order to compensate for the difficulty of rigidly defining indel boundaries.

Graphic output: scripts are provided along with the Java API for obtaining gene/exon coverage information from a variety of file formats. The application combines information obtained from the variation database and an Ensembl database with coverage information to provide an image in Scalable Vector Graphics (SVG) format that can be used for visualizing up to four groupings of data (e.g. RNA, exon capture, whole genome and controls) at once, each of which can contain multiple distinct samples. The end result is ideal for rapid comparison of groups of sequencing experiments.

Input Formats: the database currently accepts a wide variety of formats including: GFF3 (http://www.sanger.ac.uk/resources/software/gff/), SAM (Li *et al*., 2009), SNVmix (Shah *et al*., 2009), VCF and several custom formats for SNVs. The application is also designed for quick and simple addition of new formats as new variant callers become available.

Library information: in order to perform qualitative or quantitative analyses across the many source datasets stored in the database, each dataset is annotated with a minimal set of required information detailing the origin of the sample (e.g. cancer versus normal, cell line versus tissue). This enables users of the database to quickly identify the propensity of any given variation to appear in a subset of the data, and provides the ability to perform metaanalysis across the whole database rapidly to identify cancer-associated variants.

Variation annotations: the database also holds annotations from both external databases (e.g. dbSNP) and manual annotations. Java API utilities are provided to facilitate this process.

Common use-cases: the database is commonly used for several major tasks, including filtering, validation, analysis and discovery. Filtering can be done by utilizing annotations (e.g. dbSNP), matched pair datasets or datasets marked as non-cancer for separating polymorphisms from putative variants. Similar methods can also be used for identifying somatic mutations, driver mutations or other mutations of interest in non-cancer related samples (e.g. genetic diseases). API utilities exist for performing filtering on sample sizes ranging from single bases of interest to entire genome-wide studies. Limited validation of variants can also be performed by searching for support in other datasets, identifying the frequency of variants across cancer and normal samples. Further, the frequency of variants in both general and specific populations can be tested, lending credibility to commonly detected variants, while identifying those which are more likely to be sequencing errors (i.e. samples with low variant read depth in regions with high canonical read depth or variants only sparsely detected across many independent datasets). Several API utilities also exist for performing complex analysis, such as identifying variants common to one or more datasets, but not found in a second set of sequencing experiments. (e.g. this would be used for identifying variants found in data obtained from cancers, but not found in any of the matched normals.) Finally, the database is an excellent resource for performing discovery of variants commonly found in cancers but not in non-cancer samples or to identify common variations in genes of interest.

ACKNOWLEDGEMENTS

We would like to thank An He, Richard Varhol, Johnson Pang and Nina Theissen for their contributions in populating the database, Martin Hirst, Richard Moore and the GSC sequencing group for generating sequence data, and Lance Baily for his assistance with the hardware configuration.

Funding: S.J.M.J. is a scholar of the Michael Smith Foundation for Health Research. A.P.F. is supported by the Michael Smith Foundation for Health Research. Genome Canada; Genome British Columbia (in part).

Conflict of Interest: none declared.

REFERENCES

Fong,C. *et al*. (2010) GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics*. **26**, 560–564.

Hubbard,T.J.P. *et al*. (2007) Ensembl. *Nucleic Acids Res.*, **35**, D610–D617.

Li,H. *et al*. (1000) Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Shah,S.P. *et al*. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*. **461**, 809–813.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.