BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 22 2011, pages 3216–3217
doi:10.1093/bioinformatics/btr540

Databases and ontologies

Advance Access publication September 28, 2011

Kaviar: an accessible system for testing SNV novelty

Gustavo Glusman*, Juan Caballero, Denise E. Mauldin, Leroy Hood and Jared C. Roach

Institute for Systems Biology, Seattle, WA 98109, USA

Associate Editor: Janet Kelso

ABSTRACT

Summary: With the rapidly expanding availability of data from personal genomes, exomes and transcriptomes, medical researchers will frequently need to test whether observed genomic variants are novel or known. This task requires downloading and handling large and diverse datasets from a variety of sources, and processing them with bioinformatics tools and pipelines. Alternatively, researchers can upload data to online tools, which may conflict with privacy requirements. We present here Kaviar, a tool that greatly simplifies the assessment of novel variants. Kaviar includes: (i) an integrated and growing database of genomic variation from diverse sources, including over 55 million variants from personal genomes, family genomes, transcriptomes, SNV databases and population surveys; and (ii) software for querying the database efficiently.

Availability: Kaviar is programmed in Perl and offered free of charge as Open Source Software. Kaviar may be used online as a programmatic web service or downloaded for local use from http://db.systemsbiology.net/kaviar. The database is also provided.

Contact: gustavo@systemsbiology.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on June 22, 2011; revised on September 21, 2011; accepted on September 22, 2011

The advent of personalized systems medicine (Auffray *et al.*, 2009) will be predicated on the availability of precise genomic information for each patient, which will be gleaned by genotyping of known variants, by exome and transcriptome sequencing and through whole-genome sequencing. A fraction of the personal variants observed are false positive artifacts of random sequencing error or cell line mutations, which may confound the search for disease-causing mutations. When family genomes are available, over half of the errors may be identified by inheritance state analysis (Roach *et al.*, 2010). Variants frequently observed in other personal genomes are less likely to be random artifacts. Given the list of personal variations, one of the first analytical tasks is thus to determine, for each variant, whether it has already been observed in humans. Depending on the type and extent of sequencing done, a physician or medical researcher may need to test from just a few to potentially a very large number of observed genome variants, most of which are single nucleotide variants (SNVs).

dbSNP is a freely available, periodically updated general catalog of genome variation (Sherry *et al.*, 2001). The recent explosion of genomic sequencing projects have identified vast numbers of variants that have not yet been incorporated to dbSNP, a trend expected to increase as genomic sequencing becomes more economical. Researchers therefore need to compare observed SNVs with data from several different sources to ascertain whether a variant is novel or known, and to determine the population frequencies of the alleles. Access to these datasets is offered via various web interfaces, or as voluminous downloadable files. These files come in a variety of formats and may specify coordinates relative to different genome versions, thus requiring significant processing.

Importantly, the personal nature of the data may impose significant restrictions on the methods used for studying them: Institutional Review Board (IRB) and other human subject protocols may have specified that the data should be kept within the institution's intranet, or even limited to a specific machine, with tight controls on data access. In this case, the use of web applications to study personal data may be restricted or entirely disallowed, leaving researchers with the single option of downloading huge files, developing bioinformatics pipelines to use them and maintaining the entire system updated as more data are produced. This effort is being replicated in laboratories worldwide.

We have created Kaviar (Known VARiants), a compilation of human SNVs collected from many and diverse sources (Table 1). We obtained genomic shotgun data (as BAM files) and mapped RNA-seq reads to hg19 using blat (Kent, 2002), then used VarScan 2.2.5 [Koboldt *et al.* (2009), with parameters: –min-coverage 8 –min-reads2 4 –min-var-freq 0.4 –p-value 0.05] to identify SNVs. We used liftOver to translate coordinates between genome references as needed. See Methods in Supplementary Material for a detailed description of processing pipelines, parameters and the querying tools provided.

Kaviar answers a very specific question: what variants are known for a given specific genomic location? For each SNV in a query, Kaviar reports all known variants and their sources (in which population or individual genomes they were observed), or the fact that no variants are known. Where available, dbSNP identifiers are displayed. Output formats include tabular html annotated with relevant links, tab-delimited text, JavaScript Object Notation (JSON) and Variant Call Format (VCF, Danecek *et al.*, 2011). The database encodes SNV positions, identities and sources. SNVs are identified by their genomic location using standard hg18 (NCBI Build 36) or hg19 (GRCh37) coordinates. The Kaviar database is compact, at 3.44% the size in Genome Variation Format (GVF) (Reese *et al.*, 2010) and 23.8% of the gzipped GVF. Table 1 summarizes the various data sources represented in Kaviar's database as of July 26, 2011. Over half of the current SNVs lack dbSNP identifiers. The largest contributor of novel SNVs is the 1000 Genomes Project (1000 Genomes Project Consortium, 2010). Individual genomes and variation databases

*To whom correspondence should be addressed.

Table 1. Summary of Kaviar sources as of July 26, 2011 (hg19 version)

Dataset	SNVs	Unique[a] (%)	Novel[b] (%)	References
dbSNP v. 132	25 775 925	22.0	0	Sherry (2001)
1000 Genomes	43 170 073	46.9	56.9	1000 genomes.org
Personal genomes	9 578 233	5.2	9.1	various[c]
Bushman project	6 276 178	0.3	1.2	Schuster (2010)
GMI[d]	8 845 383	10.8	19.3	tiara.gmi.ac.kr
PGP[e]	5 643 933	1.4	4.5	snp.med.harvard.edu
69 Genome Set[f]	19 061 252	11.1	29.9	completegenomics .com
ISB 44 genomes[g]	7 158 208	0.1	3.5	systemsbiology.org
200 exomes	121 857	19.7	28.9	Li (2010)
RNA-seq[h]	734 413	64.8	67.7	Wang (2008), Blekhman (2010)
Saqqaq genome	2 206 894	6.5	10.2	Rasmussen (2010)
5 humans (NP)[i]	2 599 325	12.9	16.9	Green (2010)
7 humans (DP)[j]	96 990	9.3	14.8	Reich (2010)
Kaviar total	55 006 392	55.4	53.1	

[a]Unique SNVs are those observed solely in that source, but possibly in more than one individual represented by that source.
[b]Novel SNVs are those lacking dbSNP ids.
[c]Eleven individual genomes including J. C. Venter, J. Watson, S. Quake, S. J. Kim, G. Lucier, J. West, D. E. Duncan and anonymous Chinese, Yoruban and Irish individuals.
[d]Korean Genome Project.
[e]Personal Genome Project.
[f]Panel of 69 genomes released by Complete Genomics.
[g]Panel of 44 unrelated genomes sequenced in Institute for Systems Biology projects. Only SNVs observed in three or more individuals are reported.
[h]See Methods in Supplementary Material.
[i]Five modern humans from the Neanderthal Genome Project.
[j]Seven modern humans from the Denisovan Genome Project.
The full list of sources is given in the Supplementary Material.

may include vast numbers of sequencing errors (Day, 2010). By tracking of the provenance of each variant to its source individual genome(s), Kaviar facilitates the selection of SNVs confirmed by independent observation. Caution should be used when interpreting results, as SNV novelty need not imply functional or clinical relevance. Conversely, known SNVs may have unknown functional implications.

A variety of tools and services exist for collecting and annotating genome variations, including Varietas (Paananen *et al.*, 2010), SeqAnt (Shetty *et al.*, 2010), SVA (Pelak *et al.*, 2010), PanSNPdb (Ngamphiw *et al.*, 2011) and ENGINES (Amigo *et al.*, 2011). Most of these tools focus on functional annotation of variants and only offer web-based interfaces, or downloadables with difficult technical requirements. Some report only dbSNP data. Kaviar offers the widest range of different SNV sources integrated into one package. While most of the data sources in Kaviar are public, local installations can be configured to include data particular to an institution's IRB protections and not made visible to outside researchers. Alternatively, aggregate anonymized data can also be reported. Indeed, the current distribution includes such aggregate data from 44 unrelated individuals of diverse origins, which were sequenced by the Institute for Systems Biology (see Methods in Supplementary Material).

Kaviar is easily incorporated into automated workflows for genome analysis, and results can be easily used by downstream tools such as MAGMA (Hubley *et al.*, 2003). Kaviar output can be used to study genomic admixture and to judge the population frequency of alleles, in turn used by many algorithms, including those that integrate data across highly linked SNV (Roach *et al.*, 2006).

We expect Kaviar to be of use to the casual researcher interested in determining the novelty of sets of observed SNVs, to the growing number of people with their own personal genome information, and to researchers studying genome-wide personal variation requiring strict confidentiality.

ACKNOWLEDGEMENTS

We appreciate technical support, comments and discussion by Robert Hubley, Lee Rowen and Chris Witwer. All human subject protocols at Institute for Systems Biology were reviewed by the Western IRB.

Funding: National Institutes of Health (RO1GM081083 to G.G.); University of Luxembourg – Institute for Systems Biology Program.

Conflict of Interest: none declared.

REFERENCES

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
Amigo,J. *et al.* (2011) ENGINES: exploring single nucleotide variation in entire human genomes. *BMC Bioinformatics*, **12**, 105.
Auffray,C. *et al.* (2009) Systems medicine: the future of medical genomics and healthcare. *Genome Med.*, **1**, 2.
Blekhman,R. *et al.* (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, **20**, 180–189.
Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
Day,I.N.M. (2010) dbSNP in the detail and copy number complexities. *Hum. Mutat.*, **31**, 2–4.
Green,R.E. *et al.* (2010) A draft sequence of the Neanderthal genome. *Science*, **328**, 710–722.
Kent,W.J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
Koboldt,D.C. *et al.* (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.
Hubley,R.M. *et al.* (2003) Evolutionary algorithms for the selection of single nucleotide polymorphisms. *BMC Bioinformatics*, **4**, 30.
Li,Y. *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.*, **42**, 969–972.
Ngamphiw,C. *et al.* (2011) PanSNPdb: the Pan-Asian SNP genotyping database. *PLoS One*, **6**, e21451.
Paananen,J. *et al.* (2010) Varietas: a functional variation database portal. *Database*, **2010**, baq016.
Pelak,K. *et al.* (2010) The characterization of twenty sequenced human genomes. *PLoS Genet.*, **6**, e1001111.
Rasmussen,M. *et al.* (2010) Ancient human genome sequence of an extinct palaeo-eskimo. *Nature*, **463**, 757–762.
Reese,M.G. *et al.* (2010) A standard variation file format for human genome sequences. *Genome Biol.*, **11**, R88.
Roach,J.C. *et al.* (2006) Genetic mapping at 3-kilobase resolution reveals inositol 1,4,5-triphosphate receptor 3 as a risk factor for type 1 diabetes in Sweden. *Am. J. Hum. Genet.*, **79**, 614–627.
Roach,J.C. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636–639.
Schuster,S.C. *et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature*, **463**, 943–947.
Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
Shetty,A.C. *et al.* (2010) SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics*, **11**, 471.
Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.