BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 5 2012, pages 724–725
doi:10.1093/bioinformatics/bts032

Genome analysis

Advance Access publication January 18, 2012

AnnTools: a comprehensive and versatile annotation toolkit for genomic variants

Vladimir Makarov[1,2,*], Tina O'Grady[3], Guiqing Cai[1,2], Jayon Lihm[1,2,4], Joseph D. Buxbaum[1,2,5,6,7] and Seungtai Yoon[1,2,*]

[1]The Seaver Autism Center for Research and Treatment, [2]Department of Psychiatry, [3]Levy Library, Mount Sinai School of Medicine, New York, NY 10029, USA, [4]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, USA, [5]Department of Neuroscience, [6]Department of Genetics and Genomic Sciences and [7]The Friedman Brain Institute, Mount Sinai School of Medicine, New York, NY 10029, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: AnnTools is a versatile bioinformatics application designed for comprehensive annotation of a full spectrum of human genome variation: novel and known single-nucleotide substitutions (SNP/SNV), short insertions/deletions (INDEL) and structural variants/copy number variation (SV/CNV). The variants are interpreted by interrogating data compiled from 15 constantly updated sources. In addition to detailed functional characterization of the coding variants, AnnTools searches for overlaps with regulatory elements, disease/trait associated loci, known segmental duplications and artifact prone regions, thereby offering an integrated and comprehensive analysis of genomic data. The tool conveniently accepts user-provided tracks for custom annotation and offers flexibility in input data formats. The output is generated in the universal Variant Call Format. High annotation speed makes AnnTools suitable for high-throughput sequencing facilities, while a low-memory footprint and modest CPU requirements allow it to operate on a personal computer. The application is freely available for public use; the package includes installation scripts and a set of helper tools.

Availability: http://anntools.sourceforge.net/

Contact: vladimir.makarov@mssm.edu; chris.yoon@mssm.edu

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

Received on November 21, 2011; revised on January 10, 2012; accepted on January 11, 2012

1 INTRODUCTION

Increasing volumes of next-generation sequencing and microarray data demand efficient bioinformatics analysis tools capable of high-throughput processing and accurate interpretation of various types of detected changes. Detailed annotation of the genomic variants with assessment of their possible functional or regulatory roles allows researchers to highlight potentially significant findings and prioritize candidates for further analyses. While there are a number of bioinformatics tools developed specifically for this purpose (Chen *et al*., 2010; Ge *et al*., 2011; Karchin, 2009; Wang, 2010) to the best of our knowledge, the majority of the published methods are limited to interpretation of single nucleotide substitutions (SNP/SNV) and small insertions/deletions (INDEL) and do not currently offer an integrated analysis of the full spectrum of variants occurring in the coding and non-coding regions of the human genome.

Here we present AnnTools, a fast, versatile and user-friendly application designed for complete annotation of novel and publicly known SNP/SNV, INDEL and SV/CNV, called from next-generation sequencing and microarray data. The main advantages of our method are its integrative approach not limited by the specific type or genomic location of the variant, the ability to evaluate both functional and regulatory regions and incorporation of variant frequency data to highlight potentially significant variants. Our application allows individual and batch query processing and offers flexibility in input data formats. The output is generated in the universal Variant Call Format (VCF), as the current *de facto* community standard (Danecek *et al*., 2011). Complete documentation is available on the AnnTools website (http://anntools.sourceforge.net/). We provide installation scripts, demos and a set of helper tools to perform custom annotation and data transformation for future visualization and analysis.

2 METHODS

2.1 Components and Implementation

To offer the most comprehensive and accurate annotation for the detected variants, AnnTools mines data compiled from 15 constantly updated sources (Supplementary Table S1 and Fig. S1, Supplementary Materials), including the latest build of dbSNP, various UCSC Genome Browser tables, GATK refGene, Genetic Association Database (GAD), published lists of common structural genomic variation (Conrad *et al*, 2010; McCarroll *et al*, 2008; Database of Genomic Variants (http://projects.tcag.ca/variation/)), lists of conserved transcription factor-binding sites (TFBs), microRNA (miRNA) regulatory sites and promoter predictions. We run updates on a monthly basis to include the latest versions of these tables. In addition, we offer the 'db_update' helper tool as part of the AnnTools distribution package. The tool retrieves data from the source tables/databases to allow for easy aggregate update of MySQL database at the user's convenience. The complied genomic data tables are stored in MySQL database and queried with a set of tools developed in Python. The application requires Python version 2.6 (or later), has no dependencies other than MySQLdb (Python-MySQL driver) and can operate on multiple platforms. AnnTools can be used as a stand-alone application or in a distributed computing environment, or can be integrated into the user's own application code via Python API.

*To whom correspondence should be addressed.

Acceptable input data formats include VCF, variant pileup format generated by SAMTools and user-specified tabular formats. The latter must have at least four columns (chromosome, position, reference and alternative) for SNP and INDEL, and at least three columns for SV/CNV (chromosome, position start and position end).

The search is performed based on genomic coordinates, chromosome and allele information or dbSNP identifiers. Reference SNP and INDEL IDs are retrieved from the latest dbSNP build (ftp://ftp.ncbi.nih.gov/snp). For known SNP and INDEL, allele frequencies are reported if publicly available. If no rsID matches to the given parameters, a period ('.') is inserted as a field value to signify a novel variant. For each variant, we report gene and transcript names, strand, position (CDS, intronic, 5′UTR, 3′UTR, intergenic) and any overlaps with putative promoter sites, miRNA regulatory sites, conserved TFBs, disease/trait associated loci and artifact prone regions. For coding variants, we specify nucleotide, codon and amino acid changes, functional class (silent, missense, nonsense/stop, read-through), exon number and a total number of exons for each isoform. Annotation of SV/CNV includes percent overlap with reported common CNV and segmental duplications. AnnTools takes advantage of the pre-computed GATK (McKenna *et al*., 2010) refGene-big-table-b37.txt data table to identify functional classes of not only known but also novel SNPs. The use of a pre-calculated table significantly improves annotation speed and accuracy. AnnTools appends annotation to the INFO field of the VCF without re-writing or modifying any of the existing information from the variant caller.

Putative promoter regions are predicted by AnnTools based on their common location reaching 500 bp upstream from the transcription start and overlap with the CpG islands (Antequera and Bird, 1999). An example of a non-coding SNP mapped to a predicted promoter site is illustrated in Supplementary Figure S2 (Supplementary Materials).

2.2 Custom annotation

While we provide annotation of general interest, we also allow users to create custom tables of special interest. Users requiring custom annotation may either download the UCSC tables in BED format or generate their own BED file and import it to the database with the provided helper tool.

3 RESULTS

AnnTools was applied to three sets of next-generation sequencing data: SNP calls generated at Mount Sinai School of Medicine from whole-exome sequencing of the Caucasian trios (HapMap CEU), and sets of INDEL and SV calls downloaded from the pilot project of the 1000 genomes consortium. Since the available INDEL and SV calls were in the hg18 assembly of the human genome, we converted the coordinates to hg19 using the UCSC lift-over tool.

Out of a total of 31 782 SNPs, 30 932 (97%) were previously reported in dbSNP and 850 (3%) were novel. The results are summarized in Supplementary Table S2 (Supplementary Materials). Examples of the AnnTools output are presented in Supplementary Table S3 (Supplementary Materials). Comparison of the AnnTools promoter prediction with the RegionMiner and Gene2Promoter tools (both part of the popular commercial Genomatix software suite, http://www.genomatix.de) demonstrated 98% concordance. Genomatix localized 113 variants to putative promoter regions while AnnTools identified 115, including all 113 identified by the Genomatix tools.

Annotation speed for each set was ∼500 lines/min on the x86_64 GNU/Linux Ubuntu 10.04.3 LTS server. Direct comparison with two popular annotation programs (ANNOVAR, http://www.openbioinformatics.org/annovar/ and variant effect predictor tool 'snpEff', http://snpeff.sourceforge.net) confirmed that our method is sufficiently fast while having the advantage of providing a more comprehensive annotation for the genomic variants (see Performance Testing in Supplementary Materials). To evaluate the impact of the expanding size of the database on annotation speed, we tested AnnTools performance under different conditions and determined that a 10-fold expansion of the database caused only a 2-fold increase in annotation time as a result of records being sorted and indexed (see Performance Testing in Supplementary Material). Complete annotated VCF files for each dataset are available from the AnnTools website.

4 CONCLUSION

In summary, we developed a fast and comprehensive tool capable of annotating multiple types of human genomic variants in a high-throughput setting. Easily parallelized Python source code allows for the simultaneous annotation of multiple call sets in high-performance computer cluster environments. On the other hand, a low-memory footprint and modest CPU requirements enable AnnTools to run on a personal computer. The tool is designed with clear separation of application logic from the user interface and data, permitting independent development, testing and maintenance.

ACKNOWLEDGEMENTS

We thank Dr Violetta Barbashina for critical reading of the manuscript and Henry Escobar and Tim Conrad of MSSM UNIX cluster support group for server administration.

Funding: Seaver Foundation (to S.Y. and J.D.B.); Simons Foundation (to J.D.B.); NIMH (grants MH089025 and MH093725, to J.D.B.). S.Y. and G.C. are Seaver Foundation Fellows.

Conflict of Interest: none declared.

REFERENCES

Antequera,F. and Bird,A. (1999) CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.*, **9**, R661–R667.

Chen,Y. *et al.* (2010) Ensembl variation resources. *BMC Genomics*, **11**, 293–309.

Conrad,D.F. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Ge,D. *et al.* (2011) SVA: software for annotating and visualizing sequenced human genomes. *Bioinformatics*, **27**, 1998–2000.

Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief. Bioinform.*, **10**, 35–52.

McCarroll,S.A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.