BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 16 2011, pages 2300–2301
doi:10.1093/bioinformatics/btr372

Sequence analysis
Advance access publication June 22, 2011

In-depth annotation of SNPs arising from resequencing projects using NGS-SNP

Jason R. Grant, Adriano S. Arantes, Xiaoping Liao and Paul Stothard*
Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB T6G2P5, Canada

Associate Editor: Martin Bishop

ABSTRACT

Summary: NGS-SNP is a collection of command-line scripts for providing rich annotations for SNPs identified by the sequencing of whole genomes from any organism with reference sequences in Ensembl. Included among the annotations, several of which are not available from any existing SNP annotation tools, are the results of detailed comparisons with orthologous sequences. These comparisons can, for example, identify SNPs that affect conserved residues, or alter residues or genes linked to phenotypes in another species.

Availability: NGS-SNP is available both as a set of scripts and as a virtual machine. The virtual machine consists of a Linux operating system with all the NGS-SNP dependencies pre-installed. The source code and virtual machine are freely available for download at http://stothard.afns.ualberta.ca/downloads/NGS-SNP/.

Contact: stothard@ualberta.ca

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

Received on April 27, 2011; revised on June 9, 2011; accepted on June 16, 2011

1 INTRODUCTION

The latest sequencing instruments in conjunction with SNP discovery tools can be used to identify huge numbers of putative SNPs. Whether the SNPs are discovered through genome or transcriptome sequencing the next problem after identification is often annotating and choosing functionally important SNPs. Here, we describe a collection of scripts called NGS-SNP (next-generation sequencing SNP), for performing in-depth annotation of SNPs identified by popular SNP discovery programs such as Maq (Li *et al.*, 2008) and SAMtools (Li *et al.*, 2009). NGS-SNP can be applied to data from any organism with reference sequences in Ensembl, and provides numerous annotation fields, several of which are not available from other tools.

2 IMPLEMENTATION

The main component of NGS-SNP is a Perl script called 'annotate_SNPs.pl' that accepts a SNP list as input and generates as output a SNP list with annotations added (Table 1). Information used for SNP annotation is retrieved from Ensembl (Hubbard *et al.*, 2009), NCBI (Maglott *et al.*, 2011) and UniProt (UniProt Consortium, 2011). Using a locally installed version of Ensembl the annotation script can process 4 million SNPs in about 2 days on a standard desktop system. Users analyzing many SNP lists, from different individuals of the same species for example, can take advantage of the script's ability to create a local database of annotation results. This database allows all the annotations and the flanking sequence for any previously processed SNPs to be obtained much more quickly. Additional components of NGS-SNP include a script for merging, filtering and sorting SNP lists as well as scripts for obtaining reference chromosome and transcript sequences from Ensembl that can be used with SNP discovery tools such as Maq.

When the annotation script identifies an amino acid-changing SNP it calculates an 'alignment score change' value a. This process involves comparing the reference amino acid and the non-reference amino acid to each orthologue. Briefly, the amino acid encoded by the variant (i.e. non-reference) allele v is compared to each available orthologous amino acid o using a log-odds scoring matrix (BLOSUM62 by default). This provides a score $s(v,o)$ for each of the n orthologues. Similarly, the amino acid encoded by the reference allele r is compared to the orthologues. Any set of species in Ensembl can be used as the source of orthologous sequences. The average score for the reference amino acid is subtracted from the average score for the variant amino acid (1), and the result is scaled to between –1 and 1, by dividing by the maximum possible value for the scoring matrix. A positive value indicates that the variant amino acid is more similar to the orthologues than the reference amino acid, whereas a negative value indicates that the reference amino acid is more similar to the orthologues. SNPs with large positive or negative values may be of more initial interest as candidates for further study.

$$a=\frac{\sum_{o} s(v,o)}{n}-\frac{\sum_{o} s(r,o)}{n} \tag{1}$$

The annotation script includes a 'model' option that can be used to specify a well-studied species to use as an additional annotation source. When a SNP is located near or within a gene, annotations describing the model species orthologue of the gene are obtained from Ensembl, Entrez Gene and UniProt. These annotations are used to generate values that appear in a 'Model_Annotations' field, in the form of key-value pairs. Examples of information provided in this field include KEGG pathway names (Kanehisa *et al.*, 2010), the number of interacting proteins, phenotypes associated with the orthologue, the names of protein features overlapping with the SNP site in the orthologue, and phenotypes associated with mutations affecting the SNP site in the orthologue. The sample output given in Supplementary File 1 begins with the results for a contrived SNP designed to change a residue in the bovine HBB protein, to resemble a mutation responsible for sickle-cell disease in humans.

*To whom correspondence should be addressed.

Table 1. Annotation fields provided by the NGS-SNP annotation script

Field	Description
Functional_Class	Type of SNP (e.g. nonsynonymous)
Chromosome	Chromosome containing the SNP
Chromosome_Position	Position of the SNP on the chromosome
Chromosome_Strand	Strand corresponding to the reported alleles
Chromosome_Reference	Base found in the reference genome
Chromosome_Reads	Base in genome supported by the reads
Gene_Description	Short description of the relevant gene
Ensembl_Gene_ID	Ensembl Gene ID of the relevant gene
Entrez_Gene_Name	Entrez Gene name of the relevant gene
Entrez_Gene_ID	Entrez Gene ID of the relevant gene
Ensembl_Transcript_ID	Ensembl Transcript ID of the transcript
Transcript_SNP_Position	Position of the SNP on the transcript
Transcript_SNP_Reference	Base found in the reference transcript
Transcript_SNP_Reads	Base in transcript according to the reads
Transcript_To_Chr_Strand	Chromosome strand matching transcript
Ensembl_Protein_ID	Ensembl Protein ID of the affected protein
UniProt_ID	UniProt ID of the relevant protein
Amino_Acid_Position	Position of the affected amino acid
Overlapping_Protein_Features	Protein features, obtained from UniProt, that overlap with the affected amino acid
Amino_Acid_Reference	Amino acid encoded by the reference
Amino_Acid_Reads	Amino acid encoded by the reads
Amino_Acids_In_Orthologues	Amino acids from orthologous sequences that align with the reference amino acid
Alignment_Score_Change	Effect of SNP on protein conservation
C_blosum	Conservation score when reference amino acid compared to orthologues using an amino acid scoring matrix
Context_Conservation	Average percent identity of the SNP region
Orthologue_Species	Source species of the orthologues used for previous four columns
Gene_Ontology	GO slim IDs and terms for the transcript
Model_Annotations	Functional information obtained from a model species, in the form of key-value pairs
Comments	Various annotations in the form of key-value pairs, such as protein sequence lost because of stop codon
Ref_SNPs	rs IDs of known SNPs sharing alleles with this SNP
Is_Fully_Known	Whether existing SNP records completely describe this SNP

Fields present in the input SNP list are also included in the output, preceding the fields described above.

The annotation script can optionally provide the genomic flanking sequence for each SNP, for use in the design of validation assays. Known SNP sites in the flanking sequence and at the SNP position can be included in the output, as lowercase IUPAC characters in the flanking, and as potentially additional alleles at the SNP site. Supplementary File 2 contains the flanking sequences provided by the annotation script (with known SNPs indicated in lowercase) for the 10 SNPs described in Supplementary File 1.

3 DISCUSSION

Many existing SNP annotation tools work only for human SNPs or SNPs already present in dbSNP, or can only be used to process a few thousand SNPs at a time (Chelala *et al*., 2009; Johnson *et al*., 2008; Schmitt *et al*., 2010). Apart from NGS-SNP we are aware of two tools designed to annotate the very large SNP lists generated by whole-genome resequencing of humans and non-human species. ANNOVAR (Wang *et al*., 2010) is a command-line tool that uses information from the UCSC Genome Browser to provide annotations. SeqAnt (Shetty *et al*., 2010) is web-based and can be downloaded, and also relies on resources from the UCSC Genome Browser. Both can place SNPs into functional classes, describe nearby genes, and indicate which SNPs are already described in dbSNP. Neither compares affected residues to orthologous sequences, reports overlapping protein features or domains, provides gene ontology information, or provides flanking sequence. The ability to map SNP-altered residues to a protein in another species to retrieve additional information is also not supported. However, ANNOVAR and SeqAnt provide a measure of DNA conservation at the SNP site, can handle indels, and return annotations much more quickly than NGS-SNP. These features and others give each tool some unique advantages. The option to submit SNPs to SeqAnt online may be particularly appealing to some users.

In summary, NGS-SNP can be used to annotate the SNP lists returned from programs such as Maq and SAMtools. SNPs are classified as synonymous, non-synonymous, 3′-UTR, etc., regardless of whether or not they match existing SNP records. Numerous additional fields of information are provided, several of which are not available from other tools.

Funding: Alberta Livestock and Meat Agency; the Natural Sciences and Engineering Research Council of Canada.

Conflict of Interest: none declared.

REFERENCES

Chelala,C. *et al*. (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, **25**, 655–661.

Hubbard,T.J.P. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*., **37**, D690–D697.

Johnson,A.D. *et al*. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24**, 2938–2939.

Kanehisa,M. *et al*. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*., **38**, D355–D360.

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*., **18**, 1851–1858.

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Maglott,D. *et al*. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **39**, D52–D57.

Schmitt,A.O. *et al*. (2010) CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. *Bioinformatics*, **26**, 969–970.

Shetty,A.C. *et al*. (2010) SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics*, **11**, 471.

UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*., **39**, D214–D219.

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