

GEN-SNiP: an online tool to find polymorphisms in a genome

David B. Whyte¹, Gopakumar Gopalakrishnan Nair^{2,*},

Achuthsankar S. Nair² and Oommen V. Oommen³

¹ Argus Biosciences, 2033 Ralston Ave #84, Belmont, CA 94002, USA

² Centre For Bioinformatics, University Of Kerala, Trivandrum 695581, India

³ Department Of Zoology, University Of Kerala, Trivandrum 695581, India

* Corresponding author

Email: gopakumar.cbi@gmail.com

Edited by E. Wingender; received April 03, 2009; revised May 07, 2009; accepted May 21, 2009; published June 01, 2009

Abstract

An online tool named GEN-SNiP that identifies variations in a set of test DNA sequences with respect to a standard reference sequence is developed and deployed successfully. The tool generates a list of substitutions, insertions and deletions for each test sequences, determined by the reference sequence. In the key batch mode feature, the tool allows multiple sequences to be compared and contrasted even when small insertions and deletions are present, with results sent to the user via email. Other distinguishing features of the tool are grouping of continuous deletions or insertions in the test sequence into a single entity for better output handling,

displaying of the alignment of test and reference sequence and the input sequence. The tool has been reported as unique in recent literature.

Availability: GEN-SNiP is freely available at

www.argusbio.com/sooryakiran/gensnip/gensnip.php.

Keywords: single nucleotide polymorphism (SNP), mitochondrial DNA (mtDNA), sequence analysis, genome, revised Cambridge reference sequence

Introduction

The tool GEN-SNiP was developed to identify differences in DNA sequences between a reference mitochondrial genome and a set of test genomes. Human mitochondrial genomes/DNAs (mtDNAs) are approximately 16,569 base pairs in length and are over 99% invariant - the average number of base pair differences between two random genomes is on the order of a few dozen [1]. Extracting the differences from the invariant bases allows for a considerable condensation of the data and facilitates comparing and contrasting different genomes. The differences can be in the form of single nucleotide polymorphisms (SNPs) or a set of insertions and deletions. To standardize mtDNA SNP names, the positions of polymorphisms and mutations in mtDNA genomes are determined by the position of the equivalent base in a standard reference sequence, the revised Cambridge Reference Sequence (rCRS) [2]. Specific mutations in mtDNA have been described that have a variety of medical implications [3]. For example, the mutation A3243G is associated with an increased risk of diabetes mellitus [4]. Mutations that become established in the population at a frequency >1% are traditionally referred to as polymorphisms. Polymorphisms in mtDNA are valuable tools for studying mitochondrial haplogroups and early human migrations [5]. The polymorphism A2706G, for example, is associated with mitochondrial haplogroup H, the most common haplogroup in Europe. GEN-SNiP uses rCRS as the default reference to identify and report mtDNA differences in the format used in mtDNA research. It also accepts user defined reference sequences other than rCRS for identifying mutations in a set of test sequences. Compared to another online tool for extracting differences with CRS, named Mitomaster, GEN-SNiP has additional features like batch mode and handling of partial sequences.

Methods

The tool GEN-SNiP takes one standard reference sequence and one or more test sequences as its input. The sequences can be fed to the tool by pasting them into the appropriate textbox on the webpage, or uploaded from the user's computer. The tool carries out pair wise alignment of the reference sequence with each of the test sequences. The alignment obtained is then analyzed with

respect to the numbering in the reference sequence. SNPs, insertions and deletions present in each of the test sequences are then identified by considering the numbering in the reference sequence. Insertions are noted as "ins"; for example 309insC indicates an insertion of a cytosine at position 309. Deletions are noted as "del"; for example, 3109delT indicates a deletion of a thymine base at position 3109. Polymorphisms are reported in the standard format for mitochondrial DNA research, i.e., reference base/position/alternate base: for example A263G. All numbering is done relative to the reference sequence, allowing comparison of query sequences even if they are out of register due to insertions or deletions.

A variable, say *num*, is used for handling the numbering with respect to the reference sequence and another variable, say *poly*, is set for storing the insertions, deletions and SNPs. The pair of aligned characters (x_i, y_i), x_i from reference sequence and y_i from test sequence, is then compared by GEN-SNiP using the following criteria for identifying insertions, deletions and polymorphisms.

1. If x_i is "-" and y_i any of the nucleotides, there exists an insertion in the test sequence. Append the insertion format string, "*numinsyi*" to *poly* using current value of *num*. Increment *i* and *num* by 1.
2. If x_i is any of the nucleotides and y_i is "-", there exists a deletion in the test sequence. Append the deletion format string, "*numdelxi*" to *poly* using current value of *num*. Increment *i* and *num* by 1.
3. If x_i and y_i are different nucleotides, then append the SNP format string, "*xinumyi*" to *poly* with current value of *num*. Increment *i* and *num* by 1.

If there is a single test sequence alone, GEN-SNiP does a pair wise alignment of reference sequence and the test sequence and displays the results in table format in the web page itself. But if the number of test sequences is more than one, GEN-SNiP works in background, enabling the user either to quit GEN-SNiP or run another analysis. In this case results will be sent by email to the user on completion of the entire run.

For single test sequence run, there are options in GEN-SNiP to view the input test sequence, alignment of the input test sequence and reference sequence and the feature "improved results". "Improved results" option involves post-processing steps to make the output conform to standard nomenclature for DNA variations. For example, where the output now lists "309insC 309insC" for a CC addition at position 309, the improved result will list it as 309insCC. Similarly continuous deletions, say from position 8276 to 8280 with nucleotides CTCTA, is displayed as 8276_8280delCTCGA. Other post-processing steps will be added in the future.

The tool accepts DNA sequence in Fasta format, and has a help page describing this format for new users. The tool also allows new users to run it with a default test sequence which will help them to start up with using the tool. The tool accepts sequences by pasting or uploading from a file and is implemented using PHP, Apache Web Server, Java Script and Fedora. GEN-SNiP is available free for use at the web site www.argusbio.com/sooryakiran/gensnip/gensnip.php and the test data and results are provided as [Supplementary material](#).

Results and discussion

The tool is tested with a number of reference as well as test sequences. For example, a single full length human mitochondrial genome, GenBank accession number AY713983, returned the following results, using the default rCRS reference sequence: A73G A93G A200G A263G 309insC 315insC 522delC 523delA A750G A1438G A2706G A2833G A4769G C7028T T8594C A8860G T8987C T9708C T10084C A10754G G11719A T14088C G14544A C14766T A15326G C16266T T16304C T16519C A16524G. Note that both single nucleotide polymorphisms (A73G, etc) and insertions and deletions are recorded. The time to process this single mitochondrial genome and to deliver the results online was 120 seconds. The SNP results reported by the program were verified by manual inspection of the alignments.

As an example of a test of the batch performance of the tool, 75 full length mitochondrial genomes [6] were uploaded and run against the default rCRS sequence. The test sequences vary in length from 16,567 to 16,577 due to small insertions or deletions. The output for the entire set of 75 mitochondrial genomes was completed and delivered by email in 2.5 hours (2 minutes per mtDNA genome). The test genomes and the output are available in the [Supplementary material](#).

To test GEN-SNiP with a user-defined reference sequence, and to demonstrate its use in other fields of research, the program was used to identify changing patterns of mutations in a set of Human immunodeficiency virus type 1 reverse transcriptase (RT) genes derived from sequential HIV-1 isolates in a patient with AIDS [7]. The results of the GEN-SNiP run, with six sequences, each 1299 base pairs long, using K03455 as the reference sequence, were delivered by email in approximately 2 minutes. The results are available in the [Supplementary material](#).

The tool has already been recognized as an important and unique tool in the area of biological sequence analysis, especially those involving mtDNA. The latest phylogenetic tree of global human mtDNA variation (mtDNA/Phylo Tree) has been created with the help of GEN-SNiP [8] where it was used to find haplotypes of all the available full length human mitochondrial genomes. In another work [9] GEN-SNiP was used for some proteomic research, too. Free encyclopedias like Wikipedia and genealogical communities like RootsWeb list GEN-SNiP as the sole tool to find single nucleotide polymorphisms of a DNA sequence due to its simplicity and unique features. The above mentioned citations of GEN-SNiP show its credibility as a computational research tool for biological sequence analysis and the spectrum of areas where it can be used.

Acknowledgements

One of the authors Gopakumar G would like to acknowledge the support received from University of Kerala's (www.keralauniversity.edu) Industry Incubation Facility and **SooryaKiran Bioinformatics (P) Ltd, India (www.sooryakiran.com)**.

References

1. [Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., Turnbull, D. M., Anderson, C., Ghosh, S. S., Olefsky, J. M., Beal, M. F., Davis, R. E. and Howell, N. \(2002\). Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am. J. Hum. Genet.* 70, 1152-1171.](#)
2. [Andrews, R. M., Kubacka I., Chinnery P. F., Lightowlers R. N., Turnbull D. M. and Howell N. \(1999\). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* 23, 147.](#)
3. [Wallace, D. C. \(2005\). A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* 39, 359-407.](#)
4. [Alcolado, J. C. and Thomas A. W. \(1995\). Maternally inherited diabetes mellitus: the role of mitochondrial DNA defects. *Diabet. Med.* 12, 102-108.](#)
5. [Torrioni, A., Achilli, A., Macaulay, V., Richards, M. and Bandelt, H.-J. \(2006\). Harvesting the fruit of the human mtDNA tree. *Trends Genet.* 22, 339-345.](#)
6. [Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-Y., Chaudhuri, T. K., Palla, V. and Zhang, Y.-P. \(2004\). Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am. J. Hum. Genet.* 75, 966-978.](#)
7. [Gurusinghe, A. D., Land, S. A., Birch, C., McGavin, C., Hooker, D. J., Tachedjian, G., Doherty, R. and Deacon, N. J. \(1995\). Reverse transcriptase mutations in sequential HIV-1 isolates in a patient with AIDS. *J. Med. Virol.* 46, 238-243.](#)

-
8. [Van Oven, M. and Kayser, M. \(2008\). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 29, E386-E394.](#)

-
9. [Wiwanitkit, V. \(2008\). Possible single nucleotide polymorphism \(SNP\) in the nucleic sequence of A-kinase-anchoring protein 9. J. Proteomics Bioinform. 1, 227-229.](#)