Bioinformatics, 31(22), 2015, 3709–3711
doi: 10.1093/bioinformatics/btv449
Advance Access Publication Date: 29 July 2015
Applications Note

Genetic and population analysis

SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data

Chase W. Nelson[1,*], Louise H. Moncla[2] and Austin L. Hughes[1,*]

[1]Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA and [2]Department of Pathobiological Sciences, University of Wisconsin School of Veterinary Medicine, Madison, WI 53706, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 11, 2015; revised on July 23, 2015; accepted on July 24, 2015

Abstract

Summary: New applications of next-generation sequencing technologies use pools of DNA from multiple individuals to estimate population genetic parameters. However, no publicly available tools exist to analyse single-nucleotide polymorphism (SNP) calling results directly for evolutionary parameters important in detecting natural selection, including nucleotide diversity and gene diversity. We have developed SNPGenie to fill this gap. The user submits a FASTA reference sequence(s), a Gene Transfer Format (.GTF) file with CDS information and a SNP report(s) in an increasing selection of formats. The program estimates nucleotide diversity, distance from the reference and gene diversity. Sites are flagged for multiple overlapping reading frames, and are categorized by polymorphism type: nonsynonymous, synonymous, or ambiguous. The results allow single nucleotide, single codon, sliding window, whole gene and whole genome/population analyses that aid in the detection of positive and purifying natural selection in the source population.

Availability and implementation: SNPGenie version 1.2 is a Perl program with no additional dependencies. It is free, open-source, and available for download at https://github.com/hugheslab/snpgenie.

Contact: nelsoncw@email.sc.edu or austin@biol.sc.edu

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

1 Introduction

Next-generation sequencing (NGS) technologies allow the rapid sequencing of pooled DNA samples containing genetic material from multiple individuals. The resultant single-nucleotide polymorphism (SNP) data may be used to reliably estimate population genetic parameters with more accuracy and less expense than the separate sequencing of multiple individuals (Futschik and Schlötterer, 2010; Lynch *et al.*, 2014), especially when samples are large and coverage is high. Unfortunately, high coverage data also suffer from a substantial false-positive error rate. SNP calling techniques can address this issue, but the only software currently available for evolutionary analysis of pooled NGS data, PoPoolation (Kofler *et al.*, 2011), is inextricable from a problematic SNP caller that has an extremely high false-positive rate (Raineri *et al.*, 2012). Further, PoPoolation relies on large pileup files and problematic simplifications, including use of the reference sequence alone to determine the number of nonsynonymous and synonymous sites. Ideally, software for evolutionary analyses of these data would allow users to first call SNPs using any preferred method, and then process the results using standard methods for determining the numbers of nonsynonymous and synonymous differences and sites.

We have developed SNPGenie to meet this need. Using SNP calling results, SNPGenie estimates: (i) nucleotide diversity (π), and its nonsynonymous and synonymous partitions (π_N and π_S) for coding regions; (ii) mean nonsynonymous and synonymous divergence (d_N and d_S) from a reference sequence; (iii) gene diversity (H_O; Nei, 1987); (iv) site type classification (nonsynonymous, synonymous or ambiguous) for polymorphic coding loci (Knapp *et al.* 2011); and (v) the constraint imposed by overlapping open reading frames. These parameters do not depend on linkage (Nelson and Hughes, 2015), circumventing a major limitation of pooled data for other applications (Cutler and Jensen, 2010). Indefinitely large genomes with multi-nucleotide variants may be analysed at speeds exceeding those of PoPoolation's default settings. The results allow users to test evolutionary hypotheses on the roles of negative (purifying) selection, positive (Darwinian) selection and random genetic drift in the sampled population. In general, $\pi_N = \pi_S$ indicates neutrality, $\pi_N < \pi_S$ indicates purifying selection and $\pi_N > \pi_S$ may indicate positive selection favoring multiple amino acid changes (Hughes, 1999). Comparing H_O at distinct polymorphic site categories may also address these hypotheses (Hughes *et al.* 2011). Parameter estimates are available at the nucleotide, codon, sliding window, whole gene and whole genome/population levels.

2 Methods

SNPGenie is a data processing Perl script, with no additional dependencies. The program accepts a reference sequence(s) (.FASTA), a Gene Transfer Format (.GTF) file with CDS annotations and an arbitrary number of SNP reports, currently including Geneious (Variations/SNPs Annotations Table) and CLC Genomics Workbench (Annotated Variant File) formats. The download package includes the Perl code, a detailed manual (README), various example files and scripts to aid in data preparation.

Nucleotide diversity (π) is the mean number of pairwise differences per site in a population of sequences. SNPGenie estimates this for all sites, and then separately for nonsynonymous and synonymous coding sites (π_N and π_S), using a new method (Nelson and Hughes, 2015) based on that of Nei and Gojobori (1986). Differences are calculated using all comparisons within every polymorphic codon and including all mutational pathways. To calculate the numbers of nonsynonymous and synonymous sites, SNPGenie weights by the sample allele frequencies. This becomes especially important when populations diverge from the reference sequence(s). Gene diversity is calculated as $H_0 = 1 - \sum x_i^2$, where x_i is the population frequency of nucleotide *i*. Polymorphic coding sites are classified following the methods of Knapp *et al.* (2011), with gene diversities given for each category.

SNPGenie (version 1.2) and PoPoolation (version 1.2.2) were used to analyse pooled H5N1 data from ferret #3501-DPI5, obtained from Wilker *et al.* (2013) (Jorge Dinis, *personal correspondence*). SNPGenie used SNP calling results from Geneious Version 5.6.3, while PoPoolation necessarily performed its own SNP calling. For SNPGenie, all default values were used. For PoPoolation, the Syn-nonsyn-sliding.pl script was used with default settings, except max-coverage = 100 000, dissable-corrections = on, min-count = 1, window-size = 3, step-size = 3 (single codon analysis). Statistical analyses were performed using RStudio version 0.98.1049.

3 Results

To validate SNPGenie's execution of the Nei-Gojobori (1986) method, we constructed sequences with all 61 non-STOP codons and known numbers of differences. MEGA Version 6 (Tamura *et al.*, 2013) was used to calculate π_N and π_S. SNP reports and GTF files were then constructed to reflect the known variant frequencies and reference sequence, and SNPGenie was used to estimate the same parameters. All results agreed to the last decimal.

Table 1. Mean nonsynonymous (N) and synonymous (S) differences and sites in hemagglutinin (HA) and neuraminidase (NA) genes of an H5N1 influenza population, estimated by PoPoolation and SNPGenie

Gene	Param.	R^2	PoPoolation	SNPGenie	*P*
HA	N diffs	0.991	0.0039 ± 0.0015	0.0033 ± 0.0016	0.001
	S diffs	0.995	0.0011 ± 0.0006	0.0008 ± 0.0006	0.002
	N sites	0.830	2.3844 ± 0.0144	2.3483 ± 0.0144	<0.001
	S sites	0.830	0.6156 ± 0.0144	0.6517 ± 0.0144	<0.001
NA	N diffs	0.437	0.0015 ± 0.0001	0.0089 ± 0.0007	<0.001
	S diffs	0.231	0.0007 ± 0.0001	0.0029 ± 0.0002	<0.001
	N sites	0.882	2.3667 ± 0.0159	2.3347 ± 0.0158	<0.001
	S sites	0.884	0.6333 ± 0.0159	0.6647 ± 0.0158	<0.001

Values shown are means ± standard errors. *P*-values refer to a paired *t*-test comparing PoPoolation and SNPGenie, with the codon as the unit. For all R^2, $P < 0.001$ (*F* test).

Next, both SNPGenie and PoPoolation were used to analyse a pooled H5N1 sample. The nonsynonymous and synonymous mean numbers of pairwise differences per site and numbers of sites (the numerator and denominator of π_N and π_S) were then estimated for the hemagglutinin (HA; Supplementary Tables S1A and S1B) and neuraminidase (NA; Supplementary Tables S2A and S2B) genes.

When PoPoolation estimates were regressed on those from SNPGenie, all R^2 values were significant ($P < 0.001$; F-test), but smaller for differences in NA. PoPoolation overestimated differences for HA and underestimated them for NA, while overestimating the number of nonsynonymous sites (Table 1). π was significantly lower in HA ($P < 0.01$ for π_N; $P < 0.001$ for π_S; two-sample *t*-tests), consistent with previous evidence for a population bottleneck upon viral transmission that is driven by selection for specific HA residues (Wilker *et al.*, 2013). Because PoPoolation overestimated differences in HA, this suggests that its false discovery rate may be exacerbated in low-diversity (e.g. bottlenecked) contexts.

Most differences between SNPGenie and PoPoolation can be attributed to: (i) differences in SNP calling; (ii) PoPoolation's treatment of STOP codon variants as nonsynonymous; and (iii) SNPGenie's use of allele frequency data in determining the number of sites, contrasted to PoPoolation's use of the reference sequence alone. PoPoolation also reports $\pi_S = 0$ for codons with no synonymous sites, where π_S should be undefined. This could highly inflate the π_N/π_S ratio, overestimating the prevalence of positive natural selection. If the false positive calls are random, ~75% will be nonsynonymous (Graur and Li, 2000), exacerbating this problem.

Planned future improvements in SNPGenie include additional SNP report formats (e.g. VCF) and weighted mutational pathways.

Acknowledgements

The authors thank Samuel Sibley, Adam Bailey, Shelby O'Connor, Jorge Dinis and Thomas Friedrich for invaluable help.

Funding

This research was supported by NIH grant AI077376 to David H. O'Connor (D.H.O.) and A.L.H.; by NIH grant AI096882 to Christopher Walker and A.L.H; by NIH R01 grant AI084787 to D.H.O. and Thomas C. Friedrich (L.H.M.); by UW-Madison Molecular Biosciences Training Grant T32 GM07215 to L.H.M.; and by an NSF GRF (DGE-0929297), a University of

South Carolina (USC) Presidential Fellowship, and a USC Dept. Biol. Sci. Kathryn Hinnant-Johnson, M.D. Memorial Fellowship to C.W.N.

Conflict of Interest: none declared.

References

Cutler,D.J. and Jensen,J.D. (2010) To pool, or not to pool? *Genetics*, **186**, 41–43.

Futschik,A. and Schlötterer,C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Graur,D. and Li,W.-H. (2000) *Fundamentals of Molecular Evolution*, 2nd Ed., Sinauer Associates, Inc., Sunderland, MA.

Hughes,A.L. (1999) *Adaptive Evolution of Genes and Genomes*, Oxford University Press, New York.

Hughes,A.L. *et al*. (2011) Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc. Natl. Acad. Sci. USA*, **100**, 15754–15757.

Knapp,E.W. *et al*. (2011) PolyAna: analyzing synonymous and nonsynonymous polymorphic sites. *Conserv. Genet. Resourc.*, **3**, 429–431.

Kofler,R. *et al*. (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*, **6**, e15925.

Lynch,M. *et al*. (2014) Population-genetic inference from pooled-sequencing data. *Genome Biol. Evol.*, **6**, 1210–1218.

Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.

Nelson,C.W. and Hughes,A.L. (2015) Within-host nucleotide diversity of virus populations: Insights from next-generation sequencing. *Infect. Genet. Evol.*, **30**, 1–7.

Raineri,E. *et al*. (2012) SNP calling by sequencing pooled samples. *BMC Bioinformatics*, **13**, 239.

Tamura,K. *et al*. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725–2729.

Wilker,P.R. *et al*. (2013) Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. *Nat. Commun.*, **4**, 2636.