Genetic and population analysis

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

Chase W. Nelson¹*, Louise H. Moncla² and Austin L. Hughes¹,*

¹Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA and ²Department 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/snp genie.

**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 (πN and πS) for coding
analyses were performed using RStudio version 0.98.1049. A Syn-nonsyn-sliding.pl script was used with default settings, except for SNPGenie, all default values were used. For PoPoolation, the 5.6.3, while PoPoolation necessarily performed its own SNP calling.

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 πN and πS. 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.

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 πS and πN) 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 R² 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 in hemagglutinin (Table 1). π was significantly lower in HA (P < 0.01 for πS; P < 0.001 for πG 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 synonymous; 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 πS = 0 for codons with no synonymous variants, where πS should be undefined. This could highly inflate the πN/πS 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

| 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. | R²  | PoPoolation | SNPGenie | P    |
| HA     | N diffs | 0.991 | 0.0039 ± 0.0015 | 0.0033 ± 0.0016 | 0.001 |
|        | S diffs | 0.995 | 0.0011 ± 0.0008 | 0.0088 ± 0.0007 | 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 R², P < 0.001 (F-test).
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.

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