Published in final edited form as:

Conserv Genet Resour. 2011 July 1; 3(3): 429–431. doi:10.1007/s12686-010-9372-5.

PolyAna: analyzing synonymous and nonsynonymous polymorphic sites

Ethan W. Knapp,

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

Program in Integrated Bioscience, Department of Biology, University of Akron, Akron, OH 44325, USA

Stephanie Jiménez Irausquin,

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

Robert Friedman, and

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

Austin L. Hughes

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

Austin L. Hughes: austin@biol.sc.edu

Abstract

PolyAna (available at <http://www.biol.sc.edu/~austin/>) aids in the analysis of natural selection on protein-coding sequences by estimating sample gene diversity at individual polymorphic nucleotide sites, which are classified as synonymous, non-synonymous, or ambiguous.

Keywords

Effective population size; Purifying selection; Single nucleotide polymorphism

Introduction

Technological advances of the genomic era have led to improved methods of genetic analysis for use in population-based studies (Gray et al. 2000). In the human species, common single nucleotide polymorphisms (SNPs) have attracted a great deal of attention because of their potential association with common complex diseases and other phenotypes of medical interest (Chanock 2001; Hirschhorn et al. 2002; Landegren et al. 1998; Matsuzaki et al. 2004). In species of conservation concern, understanding patterns of polymorphism can offer valuable insights into population histories and ongoing population processes, including the roles of recombination, genetic drift, and natural selection (Hughes and Hughes 2007; Hughes et al. 2008b; Morin et al. 2004; Rafalski 2002; Seddon et al. 2005).

The most common form of natural selection in nature is purifying selection (acting to eliminate deleterious mutations). Slightly deleterious variants present in a population may be subject to more or less effective ongoing purifying selection acting to decrease their frequency. In humans, such slightly deleterious variants, present at much higher frequencies

than those of classic Mendelian disease genes, may be important contributors to complex disease (Hughes et al. 2003), while in non-human species, they may likewise contribute to phenotypic traits of relatively reduced fitness. In coding regions, most slightly deleterious variants are likely to be nonsynonymous; thus, comparing the level of polymorphism at synonymous and nonsynonymous SNP sites provides evidence of ongoing purifying selection (Dufour et al. 2000; Hughes 2005, 2007, 2009a, b; Hughes and Hughes 2007; Hughes et al. 2003, 2005a, b, 2007, 2008a, b; Hughes and Piontkivska 2008; Irausquin and Hughes 2008, 2010; Kearney et al. 2008; Putaporntip et al. 2010; Rozen et al. 2009).

In the human population, where most SNPs are bi-allelic, the minor allele frequency (MAF), or the frequency of the less common allele, is often used as a measure of diversity at SNP sites. However, a more general measure, usable for any polymorphic site, is the single locus “heterozygosity” or gene diversity (Nei 1987). This may be particularly useful in the case of many non-human populations, such as those of invertebrates or microbes, which may have much greater effective population sizes than the human population and consequently much greater genetic diversity (Li and Sadler 1991; Hughes et al. 2008a). Here we present a simple software program, PolyAna (Polymorphism Analysis), which computes and outputs sample gene diversity at individual nucleotide sites and classifies each site within a population of aligned protein-coding DNA sequences as synonymous, nonsynonymous, or ambiguous. Although programs such as MEGA 4 (Tamura et al. 2007) and DnaSP (Librado and Rozas 2009) provide numerous sophisticated methods for analyzing nucleotide sequence data, this type of calculation is not available in those programs.

Methods

Both the Graphical User Interface (GUI) and Command-Line Interface (CLI) applications of PolyAna v0.1 are written in C++ and are designed to run under 32-bit Microsoft Windows®. The GUI was developed using Borland C++ Builder™ Professional and incorporates Borland’s Visual Class Library (VCL). Both programs have been optimized to process large sequence files and can accommodate up to $2^{32}-1$ sequences of $2^{32}-1$ nucleotides in length. Both versions of PolyAna accept aligned nucleotide sequences in MEGA format (Tamura et al. 2007), with each MEGA file containing 2 or more aligned sequences.

The program estimates sample gene diversity at individual nucleotide sites utilizing the formula:

$$1 - \sum_{i=1}^n x_i^2$$

where n is the number of alleles and x_i is the frequency in the sample of the i th allele (Nei 1987, p. 177). Polymorphic nucleotide sites are classified as synonymous, nonsynonymous, or ambiguous according to the genetic code specified by the user. All naturally occurring genetic codes are available.

A site is classified as ambiguous if either of the following conditions holds: (1) both synonymous and nonsynonymous variants occur at that site in the dataset; or (2) because of polymorphisms at other sites in the same codon, the polymorphism at the site could be considered synonymous or non-synonymous depending on the pathway taken by evolution. The latter can occur only at codons with two or more changes. For example, consider a codon site at which the codons CTA and TTC occur in the population of sequences to be analyzed. A mutation $C \rightarrow T$ at the first position of this codon will be synonymous or nonsynonymous depending on which nucleotide is present in the third position. Likewise A

→ C at the third position will be synonymous or nonsynonymous depending on which nucleotide is present in the first position. Unless the level of sequence polymorphism is very high, the fraction of polymorphic sites counted as ambiguous is usually small.

Applications

The computation of sample gene diversity separately for synonymous and nonsynonymous sites has a number of applications in studying the evolution of protein-coding sequences and a number of statistical analyses can be conducted by copying the output obtained by PolyAna into a spreadsheet. First, comparison of sample gene diversity at synonymous and nonsynonymous sites provides a test of the hypothesis that a population harbors multiple nonsynonymous variants subject to ongoing purifying selection, since the latter will have reduced gene diversity in comparison to synonymous sites in the same genes (Hughes et al. 2003). The sample median gene diversity at nonsynonymous SNP sites can be compared with that at synonymous SNP tests by a Mann–Whitney test; because gene diversities are ordinarily not normally distributed, it is preferable to use nonparametric tests on medians rather than parametric tests on means (Hughes and Hughes 2007).

This type of analysis can also be used to provide information regarding population history and structure, particularly because purifying selection becomes less effective when the effective population size (N_e) is small (Hughes and Hughes 2007). If samples of approximately the same number of sequences are obtained from two species differing substantially in N_e , gene diversity at nonsynonymous SNPs is predicted to be higher relative to that at synonymous SNPs in the population of small N_e than in the population of large N_e . This pattern is clearly seen in highly bottlenecked populations; for example, in live vaccine strains of viruses or other passaged viral strains (Hughes 2009a, b, c). It also is to be expected that a similar effect will be seen in populations of conservation concern that have been subject to severe and prolonged bottlenecks.

Availability

Both the Graphical User Interface (GUI) and Command-Line Interface (CLI) applications of PolyAna v0.1, as well as a user's manual provided as a Portable Document Format (PDF), are available at <http://www.biol.sc.edu/~austin/>.

Acknowledgments

This work was supported by the National Institutes of Health [GM43940 to A.L.H.].

References

- Chanock S. Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis Markers*. 2001; 17(2):89–98. Review. [PubMed: 11673655]
- Dufour C, Casane D, Denton D, Wickings J, Corvol P, Jeunemaitre X. Human-chimpanzee DNA sequence variation in the four major genes of the renin angiotensin system. *Genomics*. 2000; 69(1): 14–26. [PubMed: 11013071]
- Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet*. 2000; 9(16):2403–2408. Review. [PubMed: 11005795]
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med*. 2002; 4(2):45–61. Review. [PubMed: 11882781]
- Hughes AL. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics*. 2005; 169(2):533–538. [PubMed: 15545641]
- Hughes AL. Micro-scale signature of purifying selection in Marburg virus genomes. *Gene*. 2007; 392(1–2):266–272. [PubMed: 17306473]

- Hughes AL, Hughes MA. Coding sequence polymorphism in avian mitochondrial genomes reflects population histories. *Mol Ecol*. 2007; 16(7):1369–1376. [PubMed: 17391262]
- Hughes AL. Relaxation of purifying selection on live attenuated vaccine strains of the family Paramyxoviridae. *Vaccine*. 2009a; 27(11):1685–1690. [PubMed: 19195493]
- Hughes AL. Relaxation of purifying selection on the SAD lineage of live attenuated oral vaccines for rabies virus. *Infect Genet Evol*. 2009b; 9(5):827–831. [PubMed: 19409512]
- Hughes AL. Small effective population sizes and rare nonsynonymous variants in potyviruses. *Virology*. 2009c; 393:127–134. [PubMed: 19695658]
- Hughes AL, Piontkivska H. Nucleotide sequence polymorphism in circoviruses. *Infect Genet Evol*. 2008; 8(2):130–138. [PubMed: 18093882]
- Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc Natl Acad Sci USA*. 2003; 100(26):15754–15757. [PubMed: 14660790]
- Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding loci. *Genetics*. 2005a; 170(3):1181–1187. [PubMed: 15911586]
- Hughes AL, Packer B, Welch R, Chanock SJ, Yeager M. High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. *Immunogenetics*. 2005b; 57(11):821–827. [PubMed: 16261383]
- Hughes AL, Hughes MA, Friedman R. Variable intensity of purifying selection on cytotoxic T-lymphocyte epitopes in hepatitis C virus. *Virus Res*. 2007; 123(2):147–153. [PubMed: 17005284]
- Hughes AL, Friedman R, Rivailler P, French JO. Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol Biol Evol*. 2008a; 25(10):2199–2209. [PubMed: 18667439]
- Hughes AL, Welch R, Puri V, Matthews C, Haque K, Chanock SJ, Yeager M. Genome-wide SNP typing reveals signatures of population history. *Genomics*. 2008b; 92(1):1–8. [PubMed: 18485661]
- Irausquin SJ, Hughes AL. Distinctive pattern of sequence polymorphism in the NS3 protein of hepatitis C virus type 1b reflects conflicting evolutionary pressures. *J Gen Virol*. 2008; 89(Pt 8):1921–1929. [PubMed: 18632963]
- Irausquin SJ, Hughes AL. Conflicting selection pressures target the NS3 protein in hepatitis C virus genotypes 1a and 1b. *Virus Res*. 2010; 147(2):202–207. [PubMed: 19896990]
- Kearney M, Palmer S, Maldarelli F, Shao W, Polis MA, Mican J, Rock-Kress D, Margolick JB, Coffin JM, Mellors JW. Frequent polymorphism at drug resistance sites in HIV-1 protease and reverse transcriptase. *AIDS*. 2008; 22(4):497–501. [PubMed: 18301062]
- Landegren U, Nilsson M, Kwok PY. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res*. 1998; 8(8):769–776. Review. [PubMed: 9724323]
- Li WH, Sadler LA. Low genetic diversity in man. *Genetics*. 1991; 129(2):513–523. [PubMed: 1743489]
- Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009; 25(11):1451–1452. [PubMed: 19346325]
- Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW, Mei R. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res*. 2004; 14(3):414–425. [PubMed: 14993208]
- Morin PA, Luikart G, Wayne RK, et al. SNPs in ecology, evolution and conservation. *Trends Ecol Evol*. 2004; 19(4):208–216. Review.
- Nei, M. *Molecular Evolutionary Genetics*. New York: Columbia University Press; 1987.
- Putaporntip C, Udomsangpet R, Pattanawong U, Cui L, Jongwutiwes S. Genetic diversity of the *Plasmodium vivax* merozoite surface protein-5 locus from diverse geographic origins. *Gene*. 2010; 456(1–2):24–35. [PubMed: 20178839]
- Rafalski JA. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci*. 2002; 162:329–333. Review.

- Rozen S, Marszalek JD, Alagappan RK, Skaletsky H, Page DC. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. *Am J Hum Genet.* 2009; 85(6):923–928. [PubMed: 20004767]
- Seddon JM, Parker HG, Ostrander EA, Ellegren H. SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Mol Ecol.* 2005; 14(2):503–511. [PubMed: 15660941]
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol.* 2007; 24(8):1596–1599. [PubMed: 17488738]