

Simon D.W. Frost^{*1}, Jon H. Condra², Douglas D. Richman¹, and Andrew J. Leigh Brown^{1,3}

¹University of California, San Diego, CA, US; ²Merck Laboratories, PA, US; and ³University of Edinburgh, UK.

Introduction

It has been known for some time that HIV-1 protease exhibits 'natural polymorphisms', amino acid sites which are variable in the absence of drug selection pressure (Barrie *et al.* 1996). In combination with other mutations, the presence of natural polymorphisms can result in poorer responses to protease inhibitor (PI) therapy, by reducing susceptibility to PIs or compensating for resistance-associated decreases in viral replication (Condra *et al.* 1996).

It is unknown whether natural polymorphisms are generated and maintained simply by high mutation rates, or whether positive selection, for example by cytotoxic T lymphocytes (CTLs), may play a role. Little direct evidence exists for the role of CTL pressure in the evolution of protease; few epitopes have been experimentally defined. However, evidence of positive selection in protease in the absence of therapy may indirectly support the role of immune responses. The excess of synonymous substitutions relative to nonsynonymous substitutions in the protease gene suggests that overall protease is subject to negative selection i.e. the removal of deleterious mutations rather than positive selection (Seibert *et al.* 1995, Rouzine and Coffin 1999). However, this does not preclude positive selection by the immune system at a few sites. Yang *et al.* (2000) showed using a maximum likelihood method that sites 10, 37, and 63 in protease were under positive selection pressure. However, a recent study by Suzuki and Nei (2001) has shown that the maximum likelihood methods used by Yang *et al.* may generate false positive results.

We present an analysis of a large (>500 sequences) dataset of clonal protease sequences obtained from plasma viral RNA. The patients from whom the sequences were derived were protease inhibitor naïve, and were infected prior to the advent of PI-based therapy, such that the virus in these individuals has never 'seen' selection by PIs. Multiple sequences were obtained from each individual, allowing us to compare within-host diversity with between-host divergence, and ask whether positive selection plays a role in the evolution of natural polymorphisms in protease.

Aims

- How much natural polymorphism is present in protease at baseline prior to the advent of protease therapy?
- Is the pattern of natural polymorphism within individuals different from that between individuals?
- Are natural polymorphisms generated by high mutation rates or positive selection pressure?

There is extensive amino acid variation in protease driven by positive selection prior to the advent of protease-based therapy

• Variability at an amino acid position was quantified using entropy, which summarizes both the number of different amino acid variants at a site, and their relative frequency. Monomorphic sites have an entropy of 0, whilst polymorphic sites with many different amino acid variants at equal frequencies have high entropies.

• A polymorphism may be a result of high mutation rates or positive selection. We tested this hypothesis by testing whether the rate of nonsynonymous substitution (d_n) was higher than the rate of synonymous substitution (d_s) at each site. Sites under negative selection have $d_n < d_s$, while sites under positive selection have $d_n > d_s$.



Figure 1. Polymorphic sites (*), sites under negative selection (N, $p < 0.05$; $n, 0.1 > p > 0.05$), and sites under positive selection (X, $p < 0.05$; $x, 0.1 > p > 0.05$) labelled against the consensus amino sequence.

• Natural polymorphisms at positions 12, 15, 35, 37, 41, 62, 63, 69, 72, 77, 88 and 93 were inferred to be under positive selection.

• However, the polymorphism at position 10 was inferred to be under negative selection, and there were many polymorphic sites where d_n did not differ significantly from d_s .

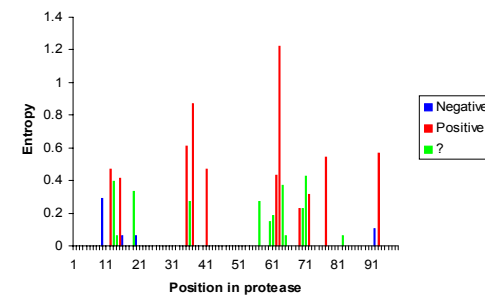


Figure 2. Amino acid variability (measured by entropy) of sites inferred to be under negative selection, positive selection or unclassified ('?').

Within host diversity at an amino acid residue is positively correlated with between host divergence

• For each amino acid site, variation within the host was calculated using entropy, and then averaged across hosts. The average within-host diversity was plotted against the between-host divergence on a site-by-site basis (Figure 3).

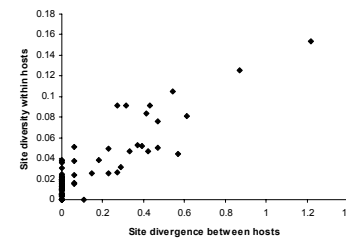


Figure 3.

• On average, within host variability mirrors between host variability, except for a small number of sites which are highly conserved between individuals but are polymorphic within individuals.

• This implies that on the timescale of a few years, positive selection is a continuous, ongoing process, and suggests that any differences in the strength of positive selection between individuals averages out.

Why are some natural polymorphisms not identified as being under positive selection?

Hypothesis 1: Lack of statistical power

• Sites where d_n is not significantly different from d_s may be evolving under positive selection, but low statistical power may prevent detection of these sites. However, at least for some sites, the errors in estimating d_n and d_s are small due to the large amount of data.

Hypothesis 2: Mutation rates vary along protease

• Some polymorphisms may be due to high mutation rates rather than positive selection pressure. This may be the case for position 10, which is under negative selection yet exhibits a number of different variants (L, V, or I in these data).

Hypothesis 3: Evolution may occur at unselected sites due to their proximity to selected sites

Polymorphisms can evolve not due to direct selection pressure, but because they are physically linked to sites which are under direct selection pressure.

Evolution of polymorphisms due to indirect positive selection

• Positive selection at a site affects evolution at linked sites.

• Hence polymorphisms may emerge if they 'hitch-hike' with sites which are under positive selection.

• Under this process, the total amount of genetic variation within the host is limited by the rate at which selected sites are fixed in the population (Gillespie 2000). In contrast, in the absence of selection, genetic variation is limited by the population size. Hence we predict no correlation between genetic variation and population size, which was indeed the case.

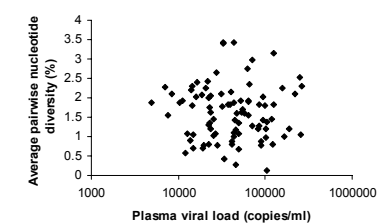


Figure 4. No correlation between within-host diversity and plasma viral load.

• This process also generates a large number of polymorphisms within the host which are present at a low frequency (Gillespie 2000). Consistent with this prediction, the number of low-frequency polymorphisms within each host (estimated by counting the number of mutations which occur in only one sequence) was on average high, and greater than that expected in the absence of selection.

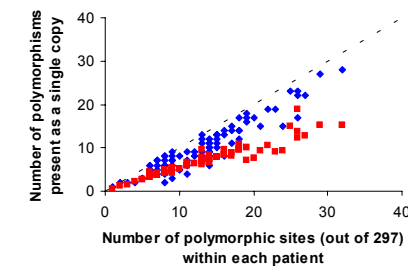


Figure 5. The number of low-frequency polymorphisms vs. the total number of polymorphisms within each patient. Blue = observed, red = expected under no selection pressure.

Conclusions

• Several natural polymorphisms appear to be maintained by positive selection, however, variability at position 10 may be driven by a high mutation rate.

• Within patient diversity mirrors between patient divergence in protease, implying a continuous, ongoing selection pressure rather than selection restricted to the time of transmission.

• Although selection pressure may vary between individuals, the effect of selection on viral genetic variation 'averages out' at the level of the population of infected hosts.

• The pattern of genetic variation in protease within the host supports the concept that some sites may evolve due to a "bystander" effect of positive selection at adjacent sites.

• Fluctuating positive selection pressure such as that exerted by the immune system is a good candidate for the force which (both directly and by bystander effects) generates and maintains natural polymorphisms in protease.

Methods

Sequence data

Primary isolates were obtained as part of the Merck 035 trial, a prospective, randomized, double-blind, placebo-controlled trial comparing AZT+3TC, IDV, and AZT+3TC+IDV (Galick *et al.* 1997). All patients were protease-inhibitor naïve. Multiple clonal sequences of the 297 base pairs encoding protease (median = 6 per patient) and viral load measurements were obtained from the plasma of 90 patients at baseline as described previously (Condra *et al.* 1995). Each protease sequence was derived from a different PCR reaction, so independence of clones was assured.

A consensus nucleotide sequence was generated for each patient using a simple majority rule. In the case of no consensus, bases were chosen according to the following rules based on the relative nucleotide frequencies of HIV genes (Cheyner *et al.* 2001): A=T>G>C, A=G>T>C, A=C>G>T, A=T>C>G, A=T>G>C, A=C>G>T or A=T>C>G, A was chosen. T=G>A>C, G=C>A>T or T=C>G>A, G was chosen. T=C>A>G, T was chosen.

Statistical analysis

Within-patient amino acid diversity was calculated using entropy performed in BioEdit v. 5.0.9 (Hall 1999). The number of single mutations and the total number of mutations within each patient was calculated using DNAsp v. 3.51 (Rozas and Rozas 1999). Viral phylogenies were reconstructed using the neighbor-joining method based on a Kimura two parameter model with a transition/transversion rate ratio parameter of 5, performed using the PHYLIP package (Felsenstein 2001). The number of nonsynonymous, k_a , and synonymous, k_s , changes were calculated using a parsimony based method (Suzuki and Gojobori 1999) implemented in the ADAPTSITE program (Suzuki *et al.* 2001). In contrast to Suzuki and Gojobori (1999), we tested whether the rate of nonsynonymous substitution, d_n , was greater than the rate of synonymous substitution, d_s , at each site using a log-linear regression rather than a chi-squared test.

References

- Barrie, K. A. *et al.* (1996) *Virology* 219:407-416.
Cheyner, R. *et al.* (2001) *J. Gen. Virol.* 82:1613-1619.
Condra, J. H. *et al.* (1995) *Nature* 374:569-571.
Condra, J. H. *et al.* (1996) *J. Virol.* 70:8270-8276.
Felsenstein, J. (2001) PHYLIP. Dept. of Genetics, University of Washington.
Gillespie, J. H. (2000) *Genetics* 155:909-919.
Galick, R. M. *et al.* (1997) *New Engl. J. Med.* 337:734-739.
Hall, T. (2001) BioEdit. Dept. of Microbiology, North Carolina State University.
Rouzine, I. & Coffin, J. M. (1999) *J. Virol.* 73:8167-8178.
Rozas, J. & Rozas, R. (1999) *Bioinformatics* 15:174-175.
Seibert, S. A. *et al.* (1995) *Mol. Biol. Evol.* 12:803-813.
Suzuki, Y. & Gojobori, T. (1999) *Mol. Biol. Evol.* 16:1315-1328.
Suzuki, Y. *et al.* (2001) *Bioinformatics* 17:660-661.
Suzuki, Y. & Nei, M. (2001) *Mol. Biol. Evol.* 18:2179-2185.
Yang, Z. *et al.* (2000) *Genetics* 155:431-449.

Acknowledgements

This work was supported by the National Institutes of Health (grant no. AI47745) and by a UCSD Center for AIDS Research Developmental Grant (NIAID 2 P30 AI 36214) to SDWF.