HIV-1 encodes a sequence overlapping *env* gp41 with highly significant similarity to selenium-dependent glutathione peroxidases*

Ethan Will Taylor† Ajita Bha† RamGopal Nadimpalli†
Weiqing Zhang† John Kececioglu‡

11 August 1997

To the Editor A progressive decline in plasma or serum selenium (Se) levels, paralleling the loss of CD4+ T cells, has been widely documented in HIV-1 infections. As recently reviewed (1), such observations have been reported in over 20 papers published over the last decade (originating from independent laboratories in at least six different countries), and have been noted in asymptomatic as well as symptomatic HIV patients, and children as well as adults [2, 3, 4, 5, 6, 7].

The observations of a number of investigators suggest that this HIV-associated decline in Se levels cannot be explained simply by malnutrition and/or nutrient malabsorption. According to Dworkin [3], levels of plasma Se, cardiac Se (at autopsy) and the antioxidant selenoprotein glutathione peroxidase in ARC and AIDS patients are "significantly correlated with total lymphocyte counts", but Se depletion can occur "irrespective of the presence or absence of diarrhea or gastrointestinal malabsorption" [3]. This suggests that the decline in Se levels which parallels the progression of HIV disease cannot be attributed entirely to malabsorption. Similarly, other authors have pointed out "a surprisingly high prevalence of low levels of selenium in early stages of the disease" [2], well before wasting is usually observed; and that: "... a low selenium intake seems unlikely [as a cause of low plasma Se], because urinary excretion, which closely reflects the actual selenium intake, was similar in HIV-1 infected patients and controls" [5]. Allavena et al. rule out malabsorption as the underlying cause, correlate Se levels with survival prognosis, and conclude that "the measurement of trace elements, especially Se, may be a useful marker to predict the HIV infection progression" [6]. Similarly, in a recent 12 month study involving 95 patients, Constans et al. show that serum Se is a statistically significant predictor of outcome in HIV infection [7].

The dramatic results of the new study by Baum et al. [8], a 3.5 yr. longitudinal study of 125 HIV+ intravenous drug users, show that Se deficiency is an exceptionally high risk factor for HIV-associated mortality. The relative risk (RR) for mortality associated with Se deficiency (RR = 10.8, $p < 0.002$) is far higher than that for any other nutrient deficiency,

*This letter to the editor appeared in *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 15:5, 393–394, 1997.
†Department of Medicinal Chemistry, The University of Georgia, Athens, GA 30602.
‡Department of Computer Science, The University of Georgia, Athens, GA 30602.

and 15 times higher than that associated with low CD4 count (defined as CD4 < 200; RR $=$ 0.69, $p < 0.04$).

In the light of this body of clinical evidence, we would like to report a new observation of potential significance for the molecular mechanisms underlying any interaction between HIV and dietary Se. Taylor et al. [9] previously pointed out several regions of HIV-1 with the potential to encode selenoproteins, due to the existence of potential frameshift sequences and conserved UGA codons (potentially encoding selenocysteine, Sec) in several locations overlapping known HIV genes. One of these overlaps the HIV-1 env gp41 coding region in the -1 reading frame, and contains a single UGA codon near the middle of a highly conserved novel open reading frame (ORF) of about 120 amino acids, following a conserved -1 frameshift signal (A AAA), consistent with the established "P-site" slippage mechanism.

We have now identified the sequence encoded in this ORF, which we call env-fs, as a homologue of glutathione peroxidase (GPx), the prototypical eukaryotic selenoprotein. The database search method we used in 1994 was not sensitive enough to pick up this similarity, probably due in part to the presence of several deletions in the active site region, as well as the high mutation rate of retroviral sequences, leading to a low overall similarity to any single GPx sequence.

Nonetheless, the presence in this HIV-1 genomic region of a common variant of the GPx active site consensus sequence, spanning the catalytic Sec, is very apparent upon examination of a multiple alignment of GPx sequences (Figure 1; YUGAT in HIV-1 vs. YUGLT in GPx, where U=Sec). This similarity occurs precisely at the location of the single conserved UGA codon in the -1 reading frame of the gp41 coding region of HIV-1. This UGA codon is highly conserved in the predominant North American-European HIV-1 strains (subtype B), but other HIV-1 subtypes (and most HIV-2's) have a conserved arginine codon (AGA) at this position. This substitution is also highly consistent with peroxidase activity, because arginine is the active site residue of many peroxidases (e.g. horseradish peroxidase), and arginine variants of the GPx active site motif (e.g. SLRG rather than SLUG or SYUG) are common in non-selenium dependent peroxidases.

Using more sophisticated bioinformatics approaches, a highly significant similarity between this HIV-encoded sequence and GPx can be demonstrated. For example, we have assessed the similarity of env-fs to the aligned set of GPx sequences by a variation on the standard method for comparing two sequences, which we have implemented in a computer program developed in our laboratories. In brief, the program produces an optimal alignment of the probe sequence (env-fs) vs. a fixed multiple alignment (GPx), by optimizing a scoring function that correctly takes into account linear gap penalties in the multiple alignment. The score of the optimal alignment of the probe sequence is the best possible value of the scores of the pairwise alignments induced between the probe sequence and the rows of the input multiple alignment. Then the probe sequence is randomly shuffled (here 50 times), and realigned by the same algorithm. The average alignment score and the standard deviation (SD) for random sequences of identical composition are then calculated, and the significance of the actual score expressed as a Z score, i.e. the number of SD from the "expected" random score.

Using a truncated GPx alignment as shown in Figure 1, the blosum62 similarity matrix, gap creation penalty of 5 and gap extension penalty of 2, the env-fs sequence is aligned essentially as shown, with a significance score of 5.0 SD. Using the full GPx alignment (including the region shown as the C-terminal), with slightly different gap parameters a Z-score of 4.9

```
Match    G SR  Y  TV DADG *   *  Y * A  F**  *YUG*TAS *A  S  *GHQ* PGKN       KGS  PG  G*L K  L** C LEC
env-fs   GSSRKHYGRTVNDADGTGQTIIVWYSAAAEQFAE..GYUGATAS.VATHS..LGHQAAPGKNPGCGKIPKGSTAPGDLGLLWKTHLHHCALEC
S05317   AAQSTVYAFSARPLTGGEPVSLGSLRGKVLLIENVASLUGTTIR%LGFPCNQFGHQV.YGARW....VALGSRLPRA.GHRLLSPLV..LQLEC
A53395   RCARSMHEFSAKDIDG.HMVNLDKYRGYVCIVTNVASQUGKTEV%LAFPCNQFGRQE.PGSDAEI....KE...FAA.GYNVKFDMFSKICVNG
S20501   SKPQSIYDFTVKDAKG.NDVDLSIYKGKVLIIVNVASQCGLTNS%LAFPCNQFGGQE.PGSIEEIQN.MVCTR......FKAEYPIFDKVDVNG
Jq0476   GISGTIYEYGALTIDGEEYIPFKQYAGKYVLFVNVASYUGLTG.%LGFPCNQFGKQE.PGENSEILPTLKYVR.PGG.GFVPNFQLFEKGDVNG
A55086   GMSGTIYEYGALTIDGEEYIPFKQYAGKYILFVNVASYUGLTD.%LGFPSNQFGKQE.PGENSEILPSLKYVR.PGG.GFVPNFQLFEKGDVNG
Jx0176   GMSGTIYEYGALTIDGEEYIPFKQYAGKYILFVNVASYUGLTD.%LGFPCNQFGKQE.PGENSEILPSLKYVR.PGG.GFVPNFQLFEKGDVNG
Jx0280   GVGGTIYEYGALTIDGEEYIPFKQYAGKYILFVNVASYUGLTG.%LGFPCNQFGKQE.PGENSEILATLKYVR.PGG.GFTPNFQLFEKGDVNG
S18912   DEKGTIYDYEAIALNKNEYVPFKQYVGKHILFVNVATYCGLTA.%LGFPCNQFGKQE.PGDNKEILPGLKYVR.PGG.GFVPNFQLFEKGDVNG
A47367   DVKGTIYDYEALSLNGKEDIPFKQYRGKHVLFVNVATYCGLTI.%LGFPCNEFGKQE.PGDNLEILPGLKYVR.PGK.GFLPNFQLFAKGDVNG
B40464   GVAGTVYEYGANTLDGGEYVQFQQYAGKHILFVNVASFCGLTA.%LGFPCNQFGKQE.PGKNSEILLGLKYVR.PGG.GFVPNFQLFEKGDVNG
A45207   FIAKSFYDLSAISLDG.EKVDFNTFRGRAVLIENVASLUGTTTR%LGFPCNQFGHQE.NCQNEEILNSLKYVR.PGG.GYQPTFTLVQKCEVNG
S21119   VAQSTVYAFSARPLAGGEPVSLGSLRGKVLLIENVASLUGTTTR%LGFPCNQFGHQE.NGKNEEILNSLKYVR.PGG.GFEPNFTLFEKCEVNG
S06304   AAAQSVYAFSARPLAGGEPVSLGSLRGKVLLIENVASLUGTTVR%LGFPCNQFGHQE.NAKNEEILNSLKYVR.PGG.GFEPNFMLFEKCEVNG
S03723   AAAQSVYSFSAHPLAGGEPVNLGSLRGKVLLIENVASLUGTTVR%LGFPCNQFGHQE.NAKNEEILNSLKYVR.PGG.GFEPNFMLFQKCEVNG
                                           ++++++++ +++++++++++++++
```

Figure 1 Sequence alignment of hypothetical HIV-1 env-fs sequence vs. various selenium-dependent glutathione peroxidases. The HIV-1 sequence shown as env-fs is as given in Table 1 of Taylor et al. [9], beginning right after a predicted putative protease cleavage site. It is aligned with a set of eukaryotic glutathione peroxidases (GPx) from various sources (single letter amino acid code, with U as the symbol for selenocysteine, Sec, encoded by the UGA codon in RNA). A match to one or more of the aligned GPx sequences is shown as a letter in the "Match" line above the alignment; similar residues are indicated by an asterisk. Active site regions of GPx are indicated by "++++++" below the alignment. Note the matches in the region of the conserved catalytic selenocysteine, with env-fs residues YUGATA, compared to YUGLTA in GPx sequence S18912 (note that here UGA means Sec-Gly-Ala, not the UGA codon). The most common variant in the predominant North American-European HIV-1 subtype B is SYUGAT instead of GYUGAT, an even better match to many of the GPx (active site motif usually SYUGLT or SLUGTT). After a 19 residue deletion in the HIV sequence (shown as a conserved cysteine (also note that one GPx has S rather than C), followed by GHQ..PGKN, which are exact matches to several GPx sequences. The deletion (at part of the GPx active site, and thus structurally not a problem. The N-terminal of mature GPx protein begins at the fourth residue in the alignment, and one human GPx variant (S05317) ends only a few residues past the end of the alignment; thus, a known GPx variant is truncated almost identically as this putative HIV GPx homologue. The total similarity score of the env-fs sequence to the aligned GPx sequences (see text) is 5 standard deviations above the average similarity score of randomly shuffled sequences of identical composition, each optimally aligned by the same algorithm. Assuming a normal distribution of random scores, this significance level (5 SD) means that the probability of obtaining this degree of similarity purely by chance is less than one in a million.

SD is obtained, producing an alignment that is somewhat looser in the C-terminal region, but otherwise identical to Figure 1 up to the YUGLT active site region.

The possibility of a virally-encoded selenium-dependent GPx gene is not unprecedented: a GPx homologue was recently reported in Molluscum contagiosum virus [9]. We have also identified GPx-related sequences in coxsackie B virus, the cofactor in Keshan disease, a classical Se-deficiency disease [1].

Since a GPx with a UGA codon at the active site is certainly a selenoprotein, the existence of this sequence in HIV-1 strongly suggests that a conserved UGA codon in HIV may encode selenocysteine, rather than a conventional amino acid, which can also occur. This possibility merits serious consideration in the light of the high statistical significance of the env-fs vs. GPx sequence similarity, and the new clinical data strongly linking HIV disease progression to host Se status.

References

[1] Taylor E.W., Nadimpalli R.G., Ramanathan C.S. "Genomic structures of viral agents in relation to the biosynthesis of selenoproteins." *Biol. Trace Element Res.* 1997; Vol. 56, in press.

[2] Mantero-Atienza E., Beach R.S., Gavancho M.C., Morgan R., Shor-Posner G., Fordyce-Baum M.K. "Selenium status of HIV-1 infected individuals." *J. Parenteral Enteral Nutr.* 1991; 15:693–694.

[3] Dworkin B.M. "Selenium deficiency in HIV infection and the acquired immunodeficiency syndrome (AIDS)." *Chem. Biol. Interact.* 1994; 91:181–186.

[4] Bologna R., Indacochea F., Shor-Posner G., et al. "Selenium and immunity in HIV-1 infected pediatric patients." *J. Nutr. Immunol.* 1994; 3:41–49.

[5] Schumacher M., Peraire J., Domingo J.L., Vidai F., Richart C., Corbella J. "Trace elements in patients with HIV-1 infection." *Trace Elem. Electrolytes* 1994; 11:130–134.

[6] Allavena C., Dousset B., May T., Dubois F., Canton P., Belleville F. "Relationship of trace element, immunological markers, and HIV infection progression." *Biol. Trace Element Res.* 1995; 47:133–138.

[7] Constans J., Pellegrin J.L., Sergeant C., et al. "Serum selenium predicts outcome in HIV infection." *J. AIDS Human Retrovirol.* 1995; 10:392.

[8] Baum M.K., Shor-Posner G., Lai S., et al. "High risk of mortality in HIV infection is associated with selenium deficiency." *J. AIDS Human Retrovirol.* 1997; in press.

[9] Taylor E.W., Ramanathan C.S., Jalluri R.K., Nadimpalli R.G. "A basis for new approaches to the chemotherapy of AIDS: novel genes in HIV-1 potentially encode selenoproteins expressed by ribosomal frameshifting and termination suppression." *J. Med. Chem.* 1994; 37:2637–2654.

[10] Senkevich T.G., Bugert J.J., Sisler J.R., Koonin E.V., Darai G., Moss B. "Genome sequence of a human tumorigenic poxvirus: prediction of specific host response-evasion genes." *Science* 1996; 273:813–816.