Los Alamos National Laboratory
Los Alamos New Mexico 87545

Dr. John Nutter and
Dr. John LeMontagne
NIH-NIAID AIDS Program
Bldg 31, Rm 7A46
Bethesda MD 20892

6 April 1987

Dear John and John,

I regret catching you up in these problems at such a busy time. You know that I am genuinely concerned, enough so that I am asking a small number of colleagues as you can see, to also study this letter and send their comments to you. Feel free to have others critique it.

For various reasons I have pursued nucleotide alignments across the HIV genomes in the database, even though our notions of conservation and variability have become somewhat set in terms of protein sequences. We say, for instance, that GAG is conserved, ENV is variable, except for the conserved regions which offer hope for...etc.,etc. The true variability of HIV has not struck us, I believe, because the early sequences from LAV, BH10, BH5, HXB2, and their relatives weren't that different. More recent sequences from the U.S. -- ARV2 (SF2) and CDC451 -- begin to show the diversity within the country, a diversity I shall argue which began back in the early 80's, and which was conspicuous by 1985. Of course the African HIV1 sequences -- MAL, ELI, etc. -- are somewhat different.

A mutation rate for HIV has been estimated though a "tree" has not been proposed; that mutation rate was stated to be a minimum: between 0.002 and 0.003 nucleotide substitutions per site per year for GAG, and between 0.010 and 0.001 per site per year for ENV (Mann, et.al., Science 232, 1548-1553, 1986). I think that the higher values are frightening and that they are underestimates: we estimate the point mutation rate for GAG and POL to be at least 0.003 and other genes, with exception of the ENV gene, will be of that order, about 0.01 substitutions per nucleotide per year (LTR is less). This rate holds for 3rd base positions so the issue of selection and viability need not be raised now (though there are things to be said along those lines at the appropriate time). We have worked out some preliminary "trees" as you shall see; while they show exceptional agreement with one another, we are still developing them for publication and ask that they be treated confidentially for now. Temple Smith has just sent a fresh batch, of which I include a tentative ENV tree.

A good way to begin is by comparing the gross point mutational data across the genome with the well-characterized Influenza A NS gene. Fitch and his colleagues recently published their analysis of that system, choosing it for the abundance of sequence data for isolates going back 50 years and for the likelihood that the NS gene is under indirect selection. A copy of their paper is enclosed.

for now, we are working on the 2nd base position rate and upon amino acid alignments to determine the amino acid substitution rate per site (codon) per year. This figure, placed against estimates of vaccine development time, should predict how much variation will occur from the start of the vaccine development to its application. Again, the fact that there are some conserved regions in some of the proteins cannot be idly accepted as a strong ground for optimism if, with a high mutation and recombination rate and a weak field or selection pressure, the variation is galloping. We must also consider the possible detrimental aspects of a vaccine, not from a typical toxicity standpoint but rather from an ecological standpoint.

My immediate concern is to reconstruct our perception of the virus as it has evolved. Consider the tree for SAB with respect to the cluster of isolates known as BH10, BH5, BH8, PV22 and BRU (LAV): because some of these were stated to have arisen from blood pooled from many patients, we fixed our attention initially over the variation represented therein (tree length of 10 or so). Last Spring, new sequences came forward and we became understandably focussed upon envelope variation, constructing the notion of SAB conservation - ENV variation; there is much truth to this from a comparative point of view, but the construction left as is would woefully mislead us about the full variational potential or process which is unfolding before us.

With these trees, our perspective must change to some realization that there is no "monovirus" so to speak, and that even the HIV1 - HIV2 dichotomy is merely a temporary construct. If we err in this regard, it seems we would nevertheless err in the right direction. We have tree lengths of 200 - 300 for HIV1 in 1985. This theoretical proposition has as its clinical correlate the isolation of "libraries" of distinct genotypes from individual patients (Hahn and undoubtedly others). We have not seen in the sequences to date any overt sign of recombination but we certainly should expect such. We might ask if the discrepancies reported about neurological phenomena in the recent Science article stem from viral variation. Surely we should not place much emphasis upon epidemiological models which leave out the emergence of new forms.

The question is bound to come up, could the variation be an artifact of culturing and cloning? The trees show slight evidence of that for the H9 cell-grown viruses but otherwise there is no patent sign of unnaturalness. Beatrice Hahn has addressed this question in her variation paper, and more recently with experiments designed specifically for that issue; she should report those results which argue that the variation is in vivo...is natural for the most part.

Please let me hear your thoughts and criticisms. In the meantime, Temple Smith is generating intriguing analyses which we can report soon.

Sincerely,

Gerald Myers

Gerald Myers
Theoretical Division
T-10, MS K710
505-665-0400

Encl: data
Cy: G. Bell, W. Fitch, W. Goad,
B. Hahn, M. Martin, A. Rabson,
H. Temin

Addendum for LaMontagne, Nutter, Robson, Martin, Bell, Smith only.

Literally a "double fraud" took place when the H9 cell-derived isolates -- HXB2, BH10, BH8, BH5, HXB3, PV22 (Muesing) -- were declared to be i) independent from LAV (BRU) and ii) derived from blood pooled from several patients. The probability of either account being true is very very small by this analysis, and I predict that it will become smaller with each U.S. isolate sequenced in the future. We did not set out to clarify this dispute: the conclusion is inextricably connected to the analysis of variation which we have pursued. Undoubtedly others will take the same path, see the same picture in weeks ahead.

The total branch length over the "cluster" which holds for every tree across the genome, every base position, is about 28 to be compared to a nearest distance of 48 from BH5 to SF2 and CDC401, the other U.S. isolates. The time over the cluster is in months, as one would expect. What are the chances that the French and the Americans would isolate two relatives -- they must have had a proximate ancestor -- by chance? Call the ancestor "node 13" (POL tree) or "Pre-BRU". We are analyzing over 8000 nucleotides manifesting over 1000 sites of substitution (say a 3% error by random walk?) and we have all the distinctly different branches before us which focus attention upon the "cluster", the recent derivatives of a single taxonomic entity. The only question to be asked, as I see it, is whether the French might have acquired BRU from the BH stocks.

When Robson and Martin drew attention to the two shared restriction sites peculiar to LAV and the H9 -derived group, the fact that those were enzyme cleavage sites was incidental; they were seeing only a fraction of the shared "uniquenesses" which the parsimony program integrates over so to speak. Ultimately, though, it is the astonishing and unforeseen variation of the virus which exposes the fraud.(In passing, we should be cautious, I think, about calling isolates "Haitian" which were taken from individuals of Haitian origin who were residing in the U.S., as was the case for WMJ and RF. Ahead we should be able to precisely place sequences in space and in time.)

You are all in a better position to judge this matter since I have been out of science for awhile, but I suggest that we have paid for this deception in more than the usual ways. Scientific fraudulence always costs humanity -- Mendel's sweetpea data or Millikan's oil drop data being possible exceptions -- but here we have been additionally misdirected with regard to the extent of variation of the virus, which we can ill afford during the dog days of an epidemic let alone during halcyon times.

Obviously when we present the trees, there will be realizations -- it's hard to avoid, impossible to disguise even if we chose to do so. My interest is in analyzing the variation, but I have no trouble with calling the cluster what it is. I'm not close to NIH as most of you are, thus I am most concerned about your awkwardnesses given the situation. If I can help in any way short of suppressing the variation data, which I see no other way of presenting please let me know. Temple, of course, must have his own say. It would not have been fair for me to have involved him without identifying the sequences and alerting him to the implications, though he would have quickly caught on without prompting, as others will.

Corrected %, the sixth column, takes into account the multiple substitutions at a given site, summarized in the fourth and fifth columns. We see by inspection that HIV incurs as much or more point mutation in a few years as Influenza incurs in 50; the rate of multiple substitution per substitution site is especially telling. The mutation rate for NS was determined to be 0.00 substitutions per site per year, thus we can anticipate that HIV is mutating much faster or, more correctly, varying much faster. Also, the Influenza tree for the NS gene is very slender, and the authors reasonably surmise that it is under indirect or positive selection (the oldtimers called it the Ryan effect for Francis Ryan). Drift, then, typically offsets variation unless the direct selection is weak, or balancing selection, which I think may be happening here, is going on. Thus the HIV tree is not going to be slender but instead very bushy. This, as much as the mutation rate, is a matter of utmost concern.

The question of whether these diverse sequences are all viable is important in one sense, but in another -- a Darwinian sense -- the variation itself deserves some respect. The work of Luria and Delbruck, as amplified by Lea and Coulson, argued that variation would go to infinity under nonselective conditions. Indirect selection undoubtedly holds down variation and it is clear that not much of that is going on. The molecular manifestation of this state of tepid selection, if I interpret correctly, is the frequency with which multiple substitutions are found at sites of variation. When Hahn saw the WMJ isolates evolve independently rather than sequentially, I believe she observed the microcosmic version of what the trees macroscopically represent, and, sadly, what the epidemic may ultimately be all about.

Enough of the melancholy. I aligned sequences which had few gaps or insertions and Temple Smith of the MBCRR (Dana Farber Cancer Institute at Harvard) using Swofford's code (U. of Illinois Museum of Natural History) cranked out trees from which mutation rates could be deduced. I'll bring many of these with me to show you on April 23, but for now consider the trees for GAG and POL 3rd base positions (enclosed). The nucleotide substitution rate per site per year is about 0.01, depending upon the estimate of the time of divergence from nodes; this figure won't be that sensitive to errors of 2 - 3 years ...it's going to be higher than Influenza by a factor of 10- 20 (point changes only) and higher than Hahn's figures by 8 to 25-fold.

Each tree presumes maximum parsimony -- to yield the tree of minimum length which would be fewest changes. Branch lengths then are base substitutions which connect taxonomic entities by the least path. Even so, the distances grow quickly. To illustrate, SF2 (ARV2), a San Francisco isolate, was already 3 - 4 mutational years away from the HTLV-III/LAV group in 1983 and 7 - 8 years away from the African ELI and much further from the African MAL. The U.S. viruses are SF2, CDC451 and the cluster that was HTLV-III/LAV originally. In GAG we see that CDC451 is 4 - 5 years away from SF2 in 1983. WMJ in the GAG tree looks like a U.S. virus and indeed it was taken from a Miami child who was Haitian by descent. The so-called Haitian isolates may represent "American" viruses. Ahead you'll see trees with NY5, Z6, Z3 and RF which will show more of the African - American connections. This approach will undoubtedly place viral sequences in space and time with great precision, understanding that a sequence is like a fingerprint, a genetic fingerprint.

Talking with Temin, I realize the inadequacies of this approach for explaining the molecular mechanism of HIV mutation (though I think it offers some clues) however this is the very best approach for estimating the variation of the virus as it will dictate the evolution of the epidemic. Some experiments would help. When we meet, we can discuss those and I'll tell you what we're doing to refine and test the analysis.