A PROPOSAL FOR NAMING HOST CELL DERIVED INSERTS IN RETROVIRUS GENOMES

John M. Coffin[1] and Harold E. Varmus[2]

[1]Department of Molecular Biology and Microbiology
Tufts University School of Medicine
Boston, Massachusetts 02111

[2]Department of Microbiology
University of California
San Francisco, California 94143

ABSTRACT

We propose a system for naming inserted sequences in transforming retroviruses (i.e. *onc* genes), based on using trivial names derived from a prototype strain of virus.

A number of retroviruses have been isolated from naturally occurring or laboratory-induced tumors. Some of these are able to induce rapid disease in laboratory animals and to induce transformation of morphological and/or growth properties of appropriate tissue culture cells (for review see .....). All such viruses whose genomes have been closely examined have been found to share a common feature: the presence of a nucleotide sequence which encodes a protein unnecessary for viral replication but required for the induction of the transformed phenotype (.....). Such sequences have been generally referred to as onc genes (.....). As shown in Table 1, there are at least twelve distinct onc genes which have been identified in at least twenty isolates of transforming retroviruses. Where tested, all such genes have been found to be closely related to a sequence present in the uninfected host cell, yet distinct from any endogenous viruses which might be present. It has been proposed that the transforming viruses have arisen by a mechanism involving recombination between virus and cellular information, with the consequence that an apparently normal cellular gene has come under the replicative and expression controls provided by the viral genome (.....), and by virtue of modification in structure and/or mode of expression has acquired the ability to cause cell transformation.

While there is general agreement among workers in the field concerning the nature of onc genes and their relationship to the host cell, there is substantial confusion surrounding the names of these sequences and their cellular relatives. For example, the name

src, originally used to designate the onc gene of Rous sarcoma virus (.....), has recently been applied generally to sequences which are completely unrelated in sequence, in nature of the gene product, and in location in the genome. The use of identical names for genes of unrelated sequence and function can lead to serious problems in communication. An additional problem has arisen in the description of the endogenous sequence related to an onc gene. The sequences related to the various src genes, for example, have been often called "sarc", with the result that the virus and cellular sequences have identical pronounciation. More cumbersome designation for such sequences have been proposed, but not widely accepted.

Retrovirus genes encoding replicative function (i.e. gag, pol, and env) been accorded three letter names derived from their function of some other feature (.....). We propose, for simplicity and readability, that this system be extended to include the non replicative inserts found in many strains of retrovirus. According to this proposed system, such inserts (or onc genes) will be given trivial three letter designations. These names are not meant to imply specific diseases, target cells, or functions, rather they are to be simply names of sequences which are not derived from viral replicative information, and which encode a protein (or a portion of a polyprotein) likely to be involved in transformation of the infected cell. We also propose a system for distinguishing the viral from the related cellular sequence and, where necessary, the sequences in related viral strains from one another.

The names for these sequences are to be assigned according to the following guidelines:

1. The names should be 3 letters, lower case italics.
2. The names should be trivial; that is no target cell specificity or functional significance is implied, and they are to be considered as names of coding sequences only.
3. They are to be derived in some mellifluous, yet mnemonic way from the name of the prototype virus or viruses or some other memorable feature of them.
4. Related sequences in different viruses from the same species are to be called by the same name, in a way that should (when completely resolved) point to the same or closely related cell sequence and the same or a closely related protein product, although it should not be necessary to have identified all of these to assign a name.
5. When necessary for clarity, the differences between inserts in related viruses can be indicated by prefixing the name with the abbreviation or name for the virus or virus strain.
6. The related sequence found in the cell of origin will be designaed with a lower case c- preceding the sequence name, e.g. c-src. The animal species of the cellular homologue should be indicated in paenthesis following the name of the sequence (e.g. c-src (chicken)). The unadorned name will always indicate the viral sequence only.
7. Protein products will be designated according to previous convention except that no superscripts will be used; thus, pp60src, P150c-abl, P110gag-abl stand for the product of

src, the product of the endogenous cell sequences related to abl, and the polyprotein containing both gag and abl specific information, respectively.

8. Should the same virus be found to have two independently expressed inserts (i.e. coding for different proteins through distinct mRNAs), then they can be distinguished by affixing -A, -B, etc. to the name.

9. Such names should be reserved for nonviral related sequences only. Such situations as spleen focus-forming virus, which seems to have only variants of viral replicative genes (.....) and the 30S region of Ha and Ki MSV which is apparently derived from an endogenous virus like element (.....) should not be so named. In this way, it can be assured that the names are unique.

10. Names along the same lines can also be given to nontransforming inserts if found in retroviruses or deliberately put there, but should be limited to genetically significant regions, i.e. those with protein (or functional RNA) product.

11. An exception to rule 4 can be made (although it need not) in the case where somewhat different yet related inserts are found in viruses of different species.

12. Strict genetic evidence is not required to assign a name, but it should be shown A) that the region is non-viral, and B) that it has either a protein (or functional RNA) product or a genetically identifiable funtion.

A list of suggested names is shown in Table 1. We note that many of the assignments are tentative and that more names will likely be added in the future. Three of the names on this list (src, myb, erb) are already in use. Erb and myb were originally proposed with a different rationale; i.e. that they were indicative of transformed cell type (.....). We do not consider transformed cell type to be useful criterion for such assignments, since many of the viruses cause a variety of diseases, since at least seven of the onc sequences are in viruses that cause sarcoma as their most common disease, and since even in these viruses that do cause a relatively unique definable disease (such as Abelson MuLV), there is no general agreement concerning the nature of the transformed cell. The three names mentioned, however, should in this context be considered as trivial names derived from the name of the prototype virus, and we suggest they be so used. We do suggest changing the name proposed for the transforming insert of avian myelocytomatosis virus MC29and related viruses (mac;.....) to myc to match more closely the name of the protoype virus.

If the name of an onc "gene" is considered to desribe a name of inserted sequence, all or at least part of which enduces a functional product, then (at least in principle) it can be precisely defined as that sequence which is unrelated to the genome of any replication-competent nontransforming virus (i.e. not belonging to a gag, pol, or env gene or to some noncoding internal or terminal region of such a virus). With many of these sequences, it is quite difficult to obtain a definition by purely genetic techniques, since they are usually found in replication-defective viruses. In all cases, however,

it is possible to use physical, biochemical, and recombinant DNA techniques to define the limits of onc sequences with precision, for example by comparing nucleotide sequences of a transforming virus, its nontransforming but replication competent helper, and the related cellular sequence or sequences with each other and with the amino acid sequence of the suspected gene product. A region of a genome defined in this way is not, in the strictest sense, a "gene". However, to refer to a defined sequence as an onc gene, while imprecise, should not create serious confusion, so long as it is understood that not all of the sequence may be directly involved in encoding a product and that additional viral sequnces may encode part of the final gene product.

Some of the names proposed may not at first seem as mellifluous as might be desirable. However, with practice they seem to be fairly easy to pronounce; for example, abl can be pronounced like "able" and fps like "fips". We also suggest that mas be pronounced "mass" to avoid confusion with mos ("mos").

The following investigators have agreed to these guidelines: S. Aaronson, P. Balduzzi, J. Ball, D. Baltimore, H. Bauer, J. M. Bishop, D. Dina, R. Eisenman, R. Friis, D. Fujita, A. Goldberg, H. Hanafusa, S. Hughes, W. Joklik, G.S. Martin, S. Rasheed, F. Reynolds, N. Rosenberg, C. Sherr, J. Stephenson, H. Temin, G. Theilen, K. Toyoshima, G. Vande Woude, I. Verma, P. Vogt, M. Weber, R. Weinberg and M. Yoshida.

TABLE 1. PROPOSED NAMES FOR onc GENES

Viral Insert	Virus Strain	Probable Animal Origin	Protein Product
rel	avian reticuloendotheliosis virus-T	turkey	?
RSV-src	Rous sarcoma virus	chicken	pp60src
B77-src	B77 avian sarcoma virus	chicken	pp60src
rASV-src	recovered avian sarcoma virus	chicken, Japanese quail	pp60src
PR-RSV-src	Prague strain Rous sarcoma virus	chicken	pp60src
AMV-myb	avian myeloblastosis virus strain BA1-1	chicken	?
E26-myb	avian leukemia virus strain E26	chicken	?
MC29-myc	avian myelocytoma virus MC29	chicken	P110gag-mac
CMII-myc	avian myelocytoma virus CMII	chicken	P90gag-mac
MH2-myc	avian myelocytoma and carcinoma virus MH2	chicken	P100gag-mac
OK10-myc	avian myelocytoma virusOK10	chicken	?
AEV-erb-A	avian erythroblastosis virus	chicken	P75gag-erb-A
AEV-erb-B	avian erythroblastosis virus	chicken	P45gag-erb-B
FSV-fps	Fujinami sarcoma virus	chicken	P140gag-fps
PRCII-fps	PRCII sarcoma virus	chicken	P105gag-fps
Moloney-mos	Moloney murine sarcoma virus	mouse	?
Gazdar-mos	Gazdar murine sarcoma virus	mouse	?
Rasheed-ras	Rasheed rat sarcoma virus	rat	P29gag-ras
Kirsten-ras	Kirsten murine sarcoma virus	rat	P21ras
Harvey-ras	Harvey murine sarcoma virus	rat	P21ras
abl	Abelson murine leukemia virus	mouse	P120gag-abl
ST-fes	Snyder-Theilen feline sarcoma virus	cat	P85gag-fes
GA-fes	Gardner-Arnstein feline sarcoma virus	cat	P110gag-fes
MS-~~Mos~~ fms	McDonough feline sarcoma virus	cat	P170gag-~~mas~~ fms
wos	Woolly monkey sarcoma virus	woolly monkey	?
Y73-yes	Y73 avian sarcoma virus	chicken	P90gag-yes
ESV-yes	Esh sarcoma virus	chicken	P80gag-yes