Antigenic diversity by the recombination of pseudogenes

Geneviève Thon,1,3 Théo Baltz,2 and Harvey Eisen1

1Fred Hutchinson Cancer Research Center, Seattle, Washington 98104 USA; 2Université de Bordeaux II, Immunologie et Biologie Parasitaire, 33076 Bordeaux Cedex, France

During the course of an infection, the hemoflagellate Trypanosoma equiperdum sequentially expresses an extensive repertoire of surface glycoproteins. There is evidence that combinations of silent genes are involved in the generation of this repertoire, but the combination rules are not known. To gain insight into these rules, we determined the fine structure of a composite gene. The gene coding for the variant surface glycoprotein 20 of T. equiperdum is a late gene generated by the partial duplication of three silent pseudogenes. Two closely related but not identical ‘5’ donors’ form a mosaic coding for the antigenic portion of the protein. A telomeric ‘3’ donor’ provides the last 200 nucleotides of the expressed gene. The sequences of the 5’ and 3’ donors are not related except for a short segment in which the hybrid junction is formed. These results demonstrate that recombinational processes generate diversity by re assorting sequences and also allow the expression of pseudogenes. Furthermore, the use of a short sequence similarity for the formation of the 5’-3’ donor hybrid suggests a mechanism that may act in ordering the expression of the variant surface glycoproteins.

[Key Words: Trypanosoma equiperdum; antigenic variation; surface antigen; gene conversion; pseudogene]

Received March 6, 1989; revised version accepted June 2, 1989.
Results

Three BCs are duplicated partially to generate the VSG-20 ELC

Trypanosomes of the antigenic type 20 typically appear around the fourth week of a rabbit infection (Capbern et al. 1977). Because previous work by Longacre and Eisen (1986) suggests that antigens expressed by *T. equiperdum* at that time are encoded by composite genes, we investigated the structure of the VSG-20-coding gene. We probed Southern blots of *T. equiperdum* genomic DNA with probes complementary to different portions of the VSG-20 messenger. These probes hybridize to three loci (*BC1, BC2, and BC3*) in the genomic DNA of BoTat-1, the variant that was used to initiate the infection. They hybridize to the same BCs and also to an ELC in the genomic DNA of the variant 20 (Fig. 1). Interestingly, the ELC is the only gene seen by both a probe complementary to the 5' end and a probe complementary to the 3' end of the VSG-20 mRNA. BC1 and BC2 show homology only to the 5' cDNA probe, and BC3 only to the 3' cDNA probe. We infer from this differential hybridization (1) that the ELC is the template for the VSG-20 mRNA, and (2) that at least two BCs have been duplicated partially to produce the ELC, BC3 being one of them.

To determine the contribution of BC1 and BC2 to the 5' end of the ELC, we prepared a restriction map of the VSG-20 gene family (Fig. 2). The ELC is the only gene to have the same restriction pattern as the cloned cDNA. The restriction maps suggest that the VSG-20 ELC was derived from segments of the three silent BCs. These maps are identical in BoTat-1 and BoTat-20 DNA, indicating that there was no exchange of information between the BC genes from the start of infection until the time VSG-20 was expressed. Furthermore, no genetic exchange between these genes could be detected in any of the variants examined (BoTat-78, BoTat-1*8, BoTat-28).

---

**Figure 1.** VSG-20 gene family. Genomic DNA of BoTat-1 (lane 1) or BoTat-20 (lane 20) digested by the restriction enzyme Dral (A) or PstI (B) was subjected to Southern blot analysis and hybridized to various probes. (A) cDNA representing the entire VSG-20 mRNA hybridizes to three loci, BC1, BC2, and BC3, in the genomic DNA of the two variants and also to an ELC in BoTat-20 DNA. The DNA fragment containing BC3 is of different sizes in different variants and slightly heterogeneous in size within one variant, which are characteristics of telomeric fragments (De Lange and Borst 1982; Longacre et al. 1983b). (B) Probes corresponding to different segments of the VSG-20 mRNA hybridize selectively to one or another of the BCs. The 5' cDNA probe contains 900 nucleotides at the 5' end of the VSG-20 cDNA. The 3' cDNA probe contains 200 nucleotides from the 3' end. Three oligonucleotides also were used as probes. Their sequences are indicated (bottom), as well as the position of these sequences in BC1, BC2, and cDNA.
We cloned and sequenced BC1, BC1, and BC3. The sequence analysis is presented in Figure 3. As suggested by the restriction maps, the sequence of the VSG-20 cDNA is not the same as any of the BCs but can be accounted for by a combination of the three BC sequences. The simplest combination of the BCs that would generate the ELC is schematized in Figure 4.

Although it was never detected on a Southern blot, we examined the possibility that a third 5' donor existed, which would have the same 5' sequence as the cDNA. To avoid detection, such a gene would have to be refractory to the restriction enzymes HaeIII and AvaII, both of which have recognition sites within the ELC. As we could not rule out this possibility, we made three oligonucleotides complementary to three portions of the VSG-20 mRNA. The sequences to two of these oligonucleotides are present in BC1 and not in BC1. The sequence of the third is in BC1 but not in BC1. These oligonucleotides hybridize on a Southern blot to different restriction fragments of the sizes expected for BC1, BC1, and ELC but fail to reveal a single BC that would include the three sequences (Fig. 1).

**The three VSG-20 BCs are VSG pseudogenes**

All of the VSGs described so far in *T. equiperdum* and *T. brucei* are proteins of ~55-65 kD and can be classified into two subsets depending on their carboxy-terminal amino acid sequence (Boothroyd et al. 1981; Majumder et al. 1981; Matthysens et al. 1981; Rice-Ficht et al. 1981; Liu et al. 1983). Both carboxy termini are very hydrophobic and undergo proteolytic cleavage during the VSG maturation. The class-1 (class-1) VSGs are defined by their amino acid sequence surrounding the Asp (Ser), which is at the end of the mature protein.

The 3'-untranslated regions of the VSG mRNAs also display several conserved features. These include a region rich in cytosine, an octamer sequence, and a tetradecamer sequence (Liu et al. 1983). The conservation of these homology blocks between VSG genes suggests that they have a function, possibly in the duplication-transposition event.

We searched the nucleotide sequences of the three VSG-10 BCs and found that none of them contains all the features of a complete VSG gene listed above. BC1 has one long open reading frame (RF1 in Fig 5). The first ATG in this reading frame is the 198th codon. This would encode a protein of 227 amino acids, which is half the size of any VSG ever described. Furthermore, there is no in-frame VSG carboxyl terminus consensus. A VSG class-1 sequence is found in a different reading frame (Fig. 5). We propose that BC1 was derived from a functional gene by small insertions or deletions of nucleotides that would account for the frameshifts in its coding sequence. In a similar fashion, all three reading frames of BC1 are interrupted by stop codons. In addition, none of the reading frames encodes a conserved carboxyl terminus nor are the 3' homology blocks found in the 500 nucleotides following the homology of BC1 with the VSG-10 ELC. Finally, although BC3 ends with the consensus for a class-1 VSG, the longest possible peptide encoded by BC3 is only 300 amino acids long and, presumably, could not generate a VSG by itself either.

**The 5' end of the VSG-20 ELC is a mosaic of the two 5' BCs**

The simplest combination of the BCs that would account for the cDNA sequence is presented in Figure 4. Both BC1 and BC1 contribute to the sequence of the first...
Figure 3. Comparison of the VSG-20 cDNA nucleotide sequence with the sequences of VSG-20 BC1, BC2, and BC3. The complete nucleotide sequence of VSG-20 cDNA is depicted. The VSG-20 mRNA has at its 5' end a 35-mer miniexon (indicated as 35mer) encoded in an vmlinked locus in the genome and is polyadenylated. Only the nucleotides that differ from its sequence have been indicated in the sequences of the BCs, and some gaps (indicated by spaces) have been introduced to maximize the identity between the genes. The start and stop codons of the VSG-20 gene, the homology blocks of the VSG mRNAs 3'-untranslated regions, and the 85-bp segment in which the four genes of the VSG-20 family are >70% identical are boxed.
The combination of the 3' and 5' BCs occurs within a short region of high sequence identity

BC3 encodes only the carboxyl terminus of the VSG. This has no significant identity with BC1 and BC2 except for a segment of 85 nucleotides in which the identity is >70%. This 85-nucleotide region is not part of the conserved carboxy-terminal sequence of the VSGs that is used as an integration target by some BCs of T. brucei VSGs [Bernards et al. 1981; Liu et al. 1983].

On the 5' side of these 85 nucleotides, the VSG-20 cDNA, BC1, and BC2 sequences are very similar. On the 3' side, the VSG-20 cDNA is the exact copy of BC3, but the three BCs are nonhomologous.

We compared the sequence of this 85-nucleotide region with the hybrid junction of the T. equiperdum VSG-78 gene [Roth et al. 1986] and found an identical sequence of 11 bases in the two genes. This sequence [5'-GGGGTGGTTGCA-3'] is close to the junction point between 5' and 3' donors but in inverted orientations in the VSG-20 and VSG-78 genes.

The junction between 3' and 5' BCs could correspond to a junction between two domains of the protein

Proteolysis experiments have been conducted on VSG-1 and VSG-28 of T. equiperdum [Labastie et al. 1981]. Trypsin cleaves the native protein (57,000 daltons) at several positions after 30 min of treatment. One major product is 40,000 daltons and includes the amino terminus of the VSG. A similar cleavage occurs in the VSGs of T. brucei. The native proteins (65,000 daltons) are cleaved to a large amino-terminal fragment (48,000–55,000 daltons) and one or more small carboxy-terminal fragments [Johnson and Cross 1979].

According to the size of the products, the junction between the 3' and 5' BCs in the hybrid gene might correspond to the conserved site of trypsin cleavage in the VSGs.

This would be true for all of the other hybrid genes characterized in T. equiperdum by S1 protection experiments, in which the junction between 3' and 5' donors occurs at a similar position, at 200–300 nucleotides.
from the 3' end [Longacre and Eisen 1986]. This position also represents the end of the homology between the VSG-20 BC1 and BC2 [Fig. 4].

The carboxy-terminal domain contains the conserved amino acid sequences mentioned earlier. Interestingly, the 3' region of BC1 is predictive of a class-2 carboxyl terminus, whereas BC1 sequence is linked to a class-1 type sequence after the rearrangements that result in the ELC. The possibility of such interconversions might explain why no distinctive features can be assigned to the amino termini of class-1 versus class-2 VSGs [Metcalf et al. 1987].

Discussion

The VSG-20 ELC is a mosaic and a hybrid gene

We compared the sequences of the three BCs involved in the formation of the VSG-20 gene T. equiperdum and their contributions to the ELC. The comparison shows that the expressed gene results from the superimposition of at least two recombinational events. One combination, occurring between two homologous 5' donor genes, leads to the formation of a complex patchwork of sequences, which we will call a mosaic. A second combination occurs between nonhomologous genes: the 3' and 5' donors, which have only a very short region of sequence similarity. This results in a single point of recombination between the two largely unrelated genes. We call the resulting structure a hybrid.

The juxtaposition of sequences in the hybrid most likely corresponds to an assembly of protein domains [a membrane-bound carboxyl terminus encoded by the 3' donor and an external amino terminus provided by the 5' donors], whereas the mosaic formation more likely results in the replacement or creation of epitopes within the antigenic domain. Both features are very reminiscent of the structure of the chicken immunoglobulin light-chain genes [Reynaud et al. 1987]. The diversification of the light-chain preimmune repertoire is achieved in the chicken by a high number of segmental gene conversions that occur on the product of a single V11-I1 rearrangement and involve 25 pseudogenes as donors. The donor genes are not altered by the process; only the rearranged genes are converted.

In a similar fashion, pseudogenes are used as templates for the composite T. equiperdum VSG-20 gene, and these pseudogenes are not altered detectably by generating an ELC. The existence of entire families of pseudogenes like the VSG-20 family shows that there is no strong selection for the maintenance of complete BCs in T. equiperdum. Also, the use of these incomplete genes in the formation of an ELC suggests that the observed combinations represent only a subset of the combinations that occur. Altogether, this indicates the existence of an efficient mechanism of combination involved in the generation of antigenic diversity.

Whereas short elements of well-defined sequences direct the assembly of the variable and constant region of the immunoglobulin genes, the combination of the 3' and 5' VSG donors occurs within a small region of homology. We found a common sequence of 11 bases near the hybrid junctions of the VSG-20 and VSG-78 genes. It is possible that this element plays a role in the 5'–3' joining but that it must be surrounded by a more extensive homology for recognition.

The use of a short sequence similarity to make a junction between 5' and 3' donors may be of some advantage to the trypanosome, and we propose that it is involved in defining the loose order of expression of the VSGs.

Composite ELCs and the order of expression of the VSGs

Antigenic variation is an effective way of escaping the immune response of the host only as long as the whole repertoire of antigens is not expressed simultaneously. A low frequency of switching and/or the use of a very wide repertoire could avoid that problem. T. equiperdum has evolved an apparently more complex mechanism, which causes the different antigenic types to appear sequentially in a loosely predictable order [Capbem et al. 1977].

A model for how this ordered expression of VSGs is accomplished was proposed [Longacre and Eisen 1986]. This model is based on the result of SI protection experiments, which indicate that early ELCs of T. equiperdum are made by the duplication of a single BC, and late ELCs, including the VSG-78 ELC, are made by the duplication of several BCs. The investigators hypothesized that the early BCs were complete VSG genes and that the late ones were somehow defective and could be expressed only in a hybrid structure. The transposition of the complete BCs would be more likely than the formation of the hybrid ELCs, and therefore the complete genes would be expressed first. Later on in the infection, as the immune response is presumably not as efficient, the hybrid genes would have a better chance to be made and the corresponding VSGs to be expressed.

In support of this model, Roth et al. [1986] showed that two of the VSG-78 BCs are indeed pseudogenes. We extend their observation to another family of VSG genes by showing that the three BCs or the late expressed VSG-20 are pseudogenes also. None of these three BCs could possibly have gene expressed alone in a first step before being converted by the other two. This provides some information on how the VSG-20 ELC was generated. A first possibility is that several recombinational events occurred within one cell so as to replace a previous unrelated ELC by a functional VSG-20 ELC. An alternative possibility is that a different expression site was active for several generations, while the VSG-20 ELC was built step by step, until a functional gene was made and activated. A third possibility is that one of the VSG-20 BCs can be combined with a fourth donor gene to encode a VSG of a different antigenicity. Such a composite gene could be the parent of the VSG-20 ELC. The 5' end of the VSG-20 BC3, which has no homology with the VSG-20 ELC, actually hybridizes to at least one other gene on Southern blots of T. equiperdum genomic
DNA, and we have preliminary evidence that this gene can be expressed in combination with BC3 (data not shown).

The use of short regions of sequence similarity in VSG genes otherwise nonrelated could be the way by which an order of expression finer than "early" and "late" is set. Such regions allow an interaction between genes or pseudogenes of different families. Each family could be related through them to a set of other families, which would be expressed at about the same time during the infection. The replacement of part of the ELC by the successive conversion-like events would remove and replace these short sequences and determine which gene family is more likely to be expressed next.

Methods

Trypanosomes

BoTat-20 was derived from BoTat-1, as described by Capbern et al. [1977]. Trypanosomes were grown in Wistar Furth rats, and their DNA and RNA was isolated according to Longacre et al. [1983a].

Molecular cloning
cDNA clones A BoTat-20 cDNA library was made in the plasmid pBR322 [Bolivar et al. 1977], following the method described by Longacre et al. [1983a]. BoTat-20-specific clones were selected by differential hybridization to 32P-labeled BoTat-20 and BoTat-1 cDNA and identified to VSG-20 cDNA clones by Northern blot and sequence analysis.

Genomic clones Genomic libraries were made from BoTat-20 DNA. BC1 and BC2 were cloned in the bacteriophage λ EMBL3 [Frischuh et al. 1983] as two genomic SacI restriction fragments of 12.5 and 19 kb, respectively. The restriction maps of these fragments are as expected for BC1 and BC2. Because BC3 is close to a telomere, few restriction enzymes could be used to clone it. We obtained its 3'end on a 650-bp Haelll restriction fragment and its 5' end on an overlapping 850-bp Hael1 restriction fragment. The Haell genomic library was made in the Smal restriction site of the plasmid vector Bluescribe [Vector Cloning Systems], and the Haell library in the same vector modified in the following way: The oligonucleotide 5'- CCAACGGCCTTC-3' was cloned in the Bluescribe Smal site, creating a BstXI restriction site. Cohesive ends compatible with Haell restriction ends are cut away on the engineered vector by BstXI.

The original cDNA and genomic clones were subcloned subsequently in Bluescribe (+) and Bluescribe (−) for the sequence analysis. The Bluescribe plasmids were grown in Escherichia coli TG1 in the presence of the helper phage M13K07 for the single-strand preparations.

Southern blot analysis

Trypanosome genomic DNA digested with various restriction enzymes was electrophoresed in agarose gels and transferred to nitrocellulose or GeneScreen Plus nylon membranes [NEN Research products] according to the method of Southern [Southern 1975]. Hybridization to 32P-labeled cDNA probes was performed under conditions used by Longacre et al. [1983a]. Hybridization to 32P-labeled oligonucleotide probes was performed as suggested by Schleicher and Schuell for S & S Nytran membranes (Sequences Application Update 390).

DNA sequencing

Nucleotide sequences were determined by the dideoxynucleotide chain termination method of Sanger et al. [1980], using single-stranded templates of the appropriate cDNA and genomic clones and either commercially available M13 primers or specific 18-mer oligonucleotides synthesized on an Applied Biosystems DNA synthesizer. Both strands were sequenced and the sequences were analyzed and compared using GenePro sequence analysis software version 4.1 (Riverside Scientific Enterprises).

Acknowledgments

We thank Anna Raibaud, Ursula Hibner, Gregory Buck, and Charles Roth for their help and comments at the beginning of this work and Stanley Brown, Andrew Strand, Wesley Van Voorhis for a critical reading of this manuscript and for the improvements they suggested. This research was supported by grants from the Centre National de la Recherche Scientifique and the Institut de la Santé et de la Recherche Médicale [France] and National Science Foundation grant DDC-8602729 and an American Cancer Society grant NP-585 (USA).

References

Bernards, A., L.H.T. Van der Ploeg, A.C.C. Frasch, and P. Borst. 1981. Activation of trypanosome surface glycoprotein genes involves a duplication-transposition leading to an altered 3'end. Cell 27: 497–505.

Bernards, A., T. De Lange, P. Michels, A. Liu, M.J. Huisman, and P. Borst. 1984. Two modes of activation of a single surface antigen gene of Trypanosoma brucei. Cell 36: 163–170.

Bolivar, F., R.L. Rodriguez, P.J. Greene, M.C. Betlach, H.L. Heynecker, H.W. Boyer, J.H. Croso, and S. Falkow. 1977. construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene 2: 95.

Boothroyd, J.C., G.A.M. Cross, J.H.J. Hoeijmakers, and P. Borst. 1980. A variant surface glycoprotein of Trypanosoma brucei synthesized with a C-terminal hydrophobic ‘tail’ absent from purified glycoprotein. Nature 288: 624–626.

Boothroyd, J.C., C.A. Paynter, G.A.M. Cross, A. Bernard, and P. Borst. 1981. Variant surface glycoproteins of Trypanosoma brucei are synthesized with cleavable hydrophilic sequences at the carboxy and amino termini. Nucleic Acids Res. 9: 4735–4743.

Buck, G., C. Jacquemot, T. Baltz, and H. Eisen. 1984a. Re-expression of an inactivated surface glycoprotein gene in Trypanosoma equiperdum. Gene 32: 329–336.

Buck, G., S. Longacre, A. Raibaud, U. Hibner, C. Giroud, T. Baltz, and H. Eisen. 1984b. Stability of expression-linked surface antigen genes in Trypanosoma equiperdum. Nature 307: 563–566.

Capbern, A., C. Giroud, T. Baltz, and P. Mattern. 1977. Etude des variations antigeniques au cours de la trypanosomie experimentale du Lapin. Exp. Parasitol. 42: 6–13.

De Lange, T. and P. Borst. 1982. Genomic environment of the expression-linked extra copies of genes for surface antigens of Trypanosoma brucei resembles the end of a chromosome. Nature 299: 451–453.

Frischuh, A.-M., H. Lehrach, A. Pousta, and N. Murray. 1983. Lambda replacement vectors carrying poly linker sequences. J. Mol. Biol. 170: 827–842.

Hoeijmakers, J.H.J., A.C.C. Frasch, A. Bernard, P. Borst, and G.A.M. Cross. 1980. Novel expression-linked copies of the genes for variant surface antigens in trypanosomes. Nature 284: 78–80.
Thon et al.

Johnson, J.G. and G.A.M. Cross. 1979. Selective cleavage of variant surface glycoproteins from Trypanosoma brucei. Biochem. J. 178: 689–697.

Labastie, M.C., T. Baltz, C. Richet, C. Giroud, G. Duvillier, R. Pautrizel, and P. Degand. 1981. Variant specific glycoproteins of Trypanosoma equiperdum: Cross reacting determinants and chemical studies. Biochem. Biophys. Res. Comm. 99: 729–736.

Liu, A.Y.C., L.H.T. Van der Ploeg, F.A.M. Rijsewijk, and P. Borst. 1983. The transposition unit of variant surface glycoprotein gene 118 of Trypanosoma brucei. J. Mol. Biol. 167: 57–75.

Longacre, S. and H. Eisen. 1986. Expression of whole and hybrid genes in Trypanosoma equiperdum antigenic variation. EMBO J. 5: 1057–1063.

Longacre, S., U. Hibner, A. Raibaud, H. Eisen, T. Baltz, C. Giroud, and D. Baltz. 1983a. DNA rearrangements and antigenic variation in Trypanosoma equiperdum: Multiple expression-linked sites in independent isolates expressing the same antigen. Mol. Cell. Biol. 3: 399–409.

Longacre, S., A. Raibaud, U. Hibner, C. Buck, H. Eisen, T. Baltz, C. Giroud, and D. Baltz. 1983b. DNA rearrangements and antigenic variation in Trypanosoma equiperdum: Expression independent DNA rearrangements in the basic copy of a variant surface glycoprotein gene. Mol. Cell. Biol. 3: 410–414.

Majumder, H.K., J.C. Boothroyd, and H. Weber. 1981. Homologous 3’-terminal regions of mRNAs for surface antigens of different antigenic variants of Trypanosoma brucei. Nucleic Acids Res. 9: 4745–4753.

Matthyssens, G., F. Michiels, R. Hamers, E. Pays, and M. Steinert. 1981. Two variant surface glycoproteins of Trypanosoma brucei have a conserved C-terminus. Nature 293: 230–233.

Metcalfe, P., M. Blum, D. Freyman, M. Turner, and D.C. Wiley. 1987. Two variant surface glycoproteins of Trypanosoma brucei of different sequence classes have similar 6 resolution X-ray structure. Nature 325: 84–86.

Pays, E. and M. Steinert. 1988. Control of antigen gene expression in African trypanosomes. Annu. Rev. Genet. 22: 107–126.

Pays, E., N. Van Meirvenne, D. Le Ray, and M. Steinert. 1981. Gene duplication and transposition linked to antigenic variation in Trypanosoma brucei. Proc. Natl. Acad. Sci. 78: 2673–2677.

Pays, E., S. Van Assel, M. Laurent, M. Darville, T. Vervoort, N. Van Meirvenne, and M. Steinert. 1983a. Gene conversion as a mechanism for antigenic variation. Cell 34: 371–381.

Pays, E., M.F. Delauw, S. Van Assel, M. Laurent, T. Vervoort, N. Van Meirvenne, and M. Steinert. 1983b. Modification of a Trypanosoma brucei antigen repertoire by different DNA recombinational mechanisms. Cell 35: 721–731.

Pays, E., S. Houard, S. Pays, S. Van Assel, F. Dupont, D. Aerts, and M. Steinert. 1985. Trypanosoma brucei: The extent of conversion in antigen genes may be related to the DNA coding specificity. Cell 42: 821–829.

Reynaud, C.A., V. Anquez, H. Grimal, and J.C. Weill. 1987. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell 48: 379–388.

Rice-Ficht, A.C., K.K. Chen, and J.E. Donelson. 1981. Sequence homologies near the C-termini of the variable surface glycoproteins of Trypanosoma brucei. Nature 294: 53–57.

Roth, C., S. Longacre, A. Raibaud, T. Baltz, and H. Eisen. 1986. The use of incomplete genes for the construction of a Trypanosoma equiperdum variant surface glycoprotein gene. EMBO J. 5: 1065–1070.
Antigenic diversity by the recombination of pseudogenes.

G Thon, T Baltz and H Eisen

Genes Dev. 1989, 3:  
Access the most recent version at doi:10.1101/gad.3.8.1247