A gene from the VSG expression site of *Trypanosoma brucei* encodes a protein with both leucine-rich repeats and a putative zinc finger

Philippe Revelard, Stéphane Lips and Etienne Pays*
Department of Molecular Biology, University of Brussels, 67, rue des Chevaux, B1640 Rhode Saint Genèse, Belgium

Received September 24, 1990; Revised and Accepted November 19, 1990 EMBL accession no. X55978

ABSTRACT

The transcription unit of the gene for the variant specific glycoprotein (VSG) AnTat 1.3A of *Trypanosoma brucei* contains several associated genes (ESAGs, for Expression Site-Associated Genes), 7 of which have already been described (1). We report here the characterization of a further ESAG, which we term ESAG 8, present 1 kb downstream from the putative adenylate cyclase gene ESAG 4. ESAG 8 encodes a 70 kd protein whose sequence indicates that it is probably not exposed at the cell surface. With the exception of the N-terminal domain which contains a presumptive DNA-binding zinc finger, the ESAG 8 protein consists exclusively of leucine-rich repeats of 23 amino acids, typical of protein-interacting domains such as the RAS-interacting region of the yeast adenylate cyclase. ESAG 8 transcripts are only found in bloostream forms, and their level is particularly low, suggesting a high rate of degradation. The ESAG 8 protein may be involved in stage-specific regulatory processes, such as gene expression control or adenylate cyclase activation.

INTRODUCTION

African trypanosomes, such as *Trypanosoma brucei*, escape the immune response of their mammalian host by periodically changing their variant surface glycoprotein (VSG). Only one VSG gene is expressed at any one time, from a repertoire of several hundred different sequences, and antigenic variation is achieved either by the alternate use of different VSG gene expression sites, or by different DNA rearrangement events (reviewed in 2, 3).

The expression site of the VSG gene contains a battery of other genes (ESAGs), which belong to a 45–60 kb polycistronic transcription unit (1, 4). The function of ESAGs is not known, but it seems that most of them encode membrane proteins. This is the case for ESAG 4, which appears to synthesize a transmembrane protein homologous to all known eukaryotic adenylate/guanylate cyclases (1, 5). The function of ESAG 4 may correspond to a calcium-stimulable adenylate cyclase, found only in bloodstream forms of the parasite (6, 7). The relationship, if any, between VSG and adenylate cyclase is not clear, but stimulation of adenylate cyclase activity seems to be linked to the release of the VSG (8).

In a previous study, the 5.5 kb region between ESAG 4 and ESAG 3 did not appear to be transcribed into stable polyadenylated RNAs, and was therefore not characterized further (1). We report here the presence of an additional ESAG (ESAG 8) in this region. The poor abundance of ESAG 8 transcripts suggests a regulatory role for this gene. This is supported by the amino acid sequence of the corresponding protein, which shows motifs typical of protein-interacting domains, as well as a possible DNA-binding zinc finger.

MATERIALS AND METHODS

T. brucei bloodstream forms were from the AnTat 1.3A clone. Procyclic forms were derived from the AnTat 1.1B clone (7).

The procedures for DNA and RNA isolation, Southern and Northern blot hybridization as well as DNA cloning, were as described (32). cDNA libraries were constructed in lambda gt10 according to Gubler and Hoffman (33), using the Amersham cDNA synthesis and cloning kits. The sequences of DNA fragments, subcloned in bacteriophage M13 derivatives, were determined on both strands by the method of Sanger et al. (34), using a modified T7 DNA polymerase (Sequenase, United States Biochemical Corp.).

RESULTS

Nucleotide sequence of ESAG 8

The nucleotide sequence of the beginning of the AnTat 1.3A VSG gene expression site was determined from the promoter region down to a 3′ HindIII site located about 450 bp downstream from ESAG 4 (1). A 2,517 bp HindIII-PvuII fragment starting at this HindIII site was sub-cloned from a plasmid encompassing the intergenic region between ESAG 4 and ESAG 3 (pES200.8), and its nucleotide sequence is presented in Fig. 1. This region contains an open reading frame of 630 codons starting from the ATG at position 538. As we show below that this region is transcribed

* To whom correspondence should be addressed

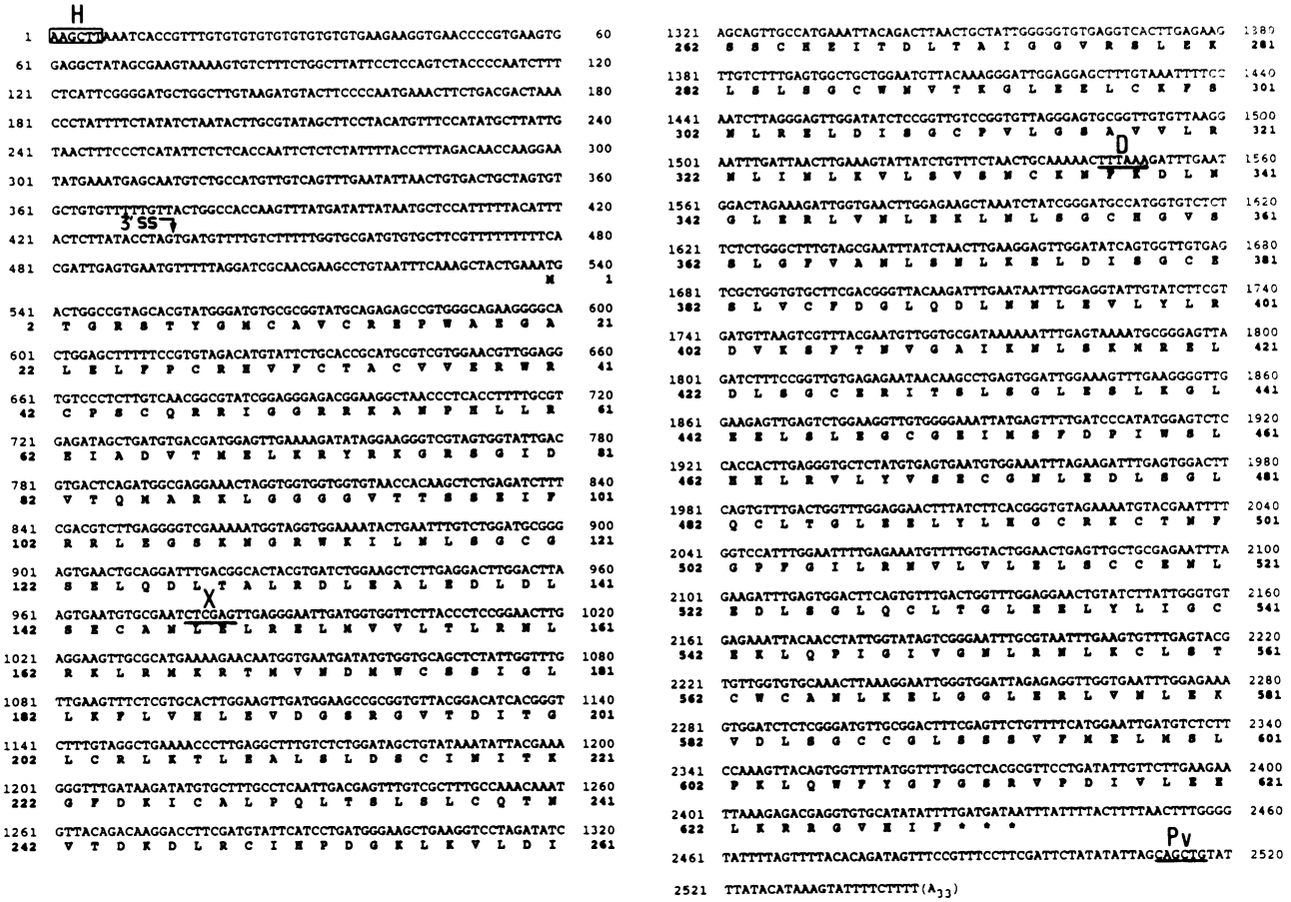

Fig. 1. Nucleotide sequence of ESAG 8 and its environment, and translation of the major open reading frame. The HindIII site at the beginning is that boxed at the end of the 18.6 kb sequence from the AnTat 1.3A VSG gene expression site, presented in ref. 1 (Fig. 2, GenBank number M20871). The 3′ splice site of the ESAG 8 mRNA could be determined from sequencing of a 2.18 kb cDNA, and is indicated by the arrow marked '3′ SS'. The terminal sequence between the PvuII site (underlined) and the poly(A) tail is only from the cDNA. The map of the AnTat 1.3A VSG gene expression site, with indication of the HindIII and PvuII sites underlined in the sequence, is shown below. The abbreviations for restriction endonuclease sites are: D=DraI; H=HindIII; Pv=PvuII; X=XhoI.

into a fully processed RNA, we propose to name it ESAG 8. This gene is located about 1 kb downstream from ESAG 4 (see map in Fig. 1).

The ESAG 8-encoded protein

Translation of the ESAG 8 open reading frame predicts a protein of approximately 70 kd, with no evidence for signal peptide or putative membrane-spanning domains, as determined by hydropathy analysis (Fig. 2). This protein is characterized by a high leucine and cysteine content: 18 and 6%, respectively. Many of these residues can be aligned, revealing an internal tandem repetition of 21 conserved and 2 to 3 degenerate repeats of 23 amino acids (Fig. 3). The repeats can be subdivided into two alternating types (A and B), whose consensus sequences are presented at the bottom of Fig. 3 (some residues characteristic of type A are encircled in Fig. 3). Overall, it appears that between the regularly spaced hydrophobic residues, the protein exhibits many interspersed charged amino acids, both basic and acidic.

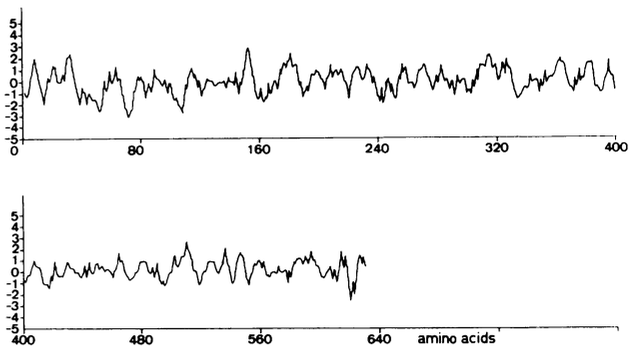

Fig. 2. Hydropathy analysis of the ESAG 8-encoded protein. The hydropathy profile has been determined according to Kyte and Doolittle (35), with a window of 6 amino acids. The relative level of hydrophobicity or hydrophilicity is represented by positive or negative values, respectively.

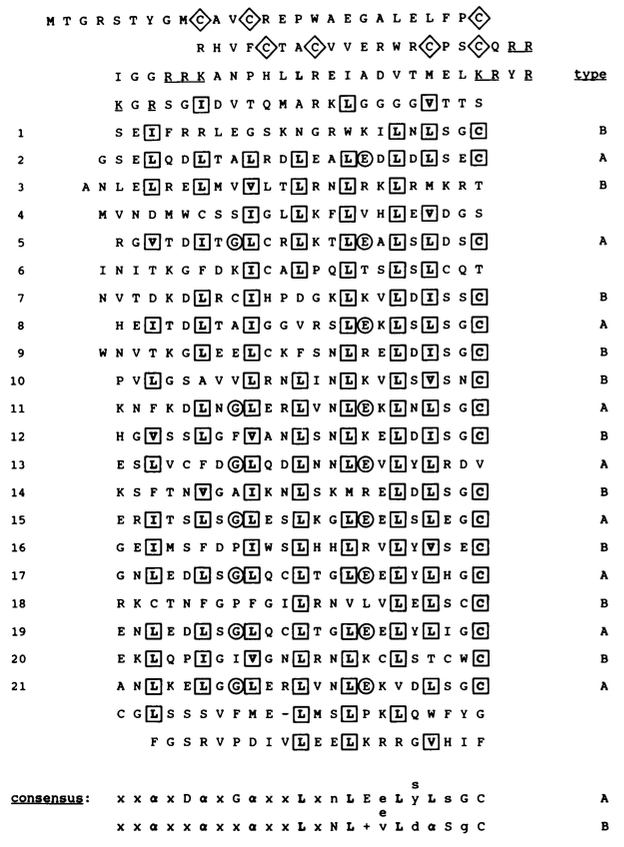

Fig. 3. Internal organization of the ESAG 8-encoded protein. Conserved amino acids allow the alignment of 21 to 24 repeats of 23 amino acids. The key residues defining the consensus sequences are boxed. Encircled amino acids are the most characteristic of the consensus subtype A; the consensus sequences for A and B subtypes are shown below. In these sequences, amino acids in upper case are conserved in at least 50% of the repeats, while those in lower case are present in at least three repeats; α and + stem for aliphatic and basic amino acids, respectively. In the N-terminal region, the cysteines are highlighted by lozenges; these residues are organized in pairs reminiscent of some DNA-binding zinc fingers. Clusters of basic residues are underlined.

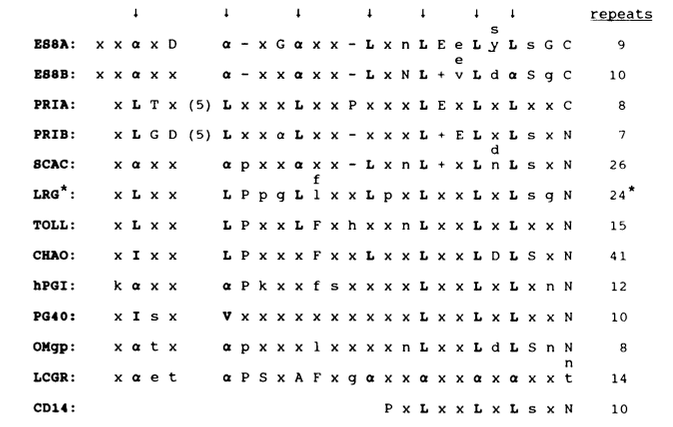

Fig. 4. The leucine-rich repeat family. The consensus sequences for the repeated leucine-rich domain of several proteins are aligned so as to maximize homologies (same abbreviations as for Fig. 3). The key hydrophobic residues are indicated by vertical arrows. The number of repeats for each consensus is shown at the right, the asterisk indicating that the LRG consensus is for a pool of five sequences. ES8 = ESAG 8 (Fig. 3); PRI = placental ribonuclease inhibitor (24, 25); SCAC = *Saccharomyces cerevisiae* adenylate cyclase (11); LRG = consensus of five serum glycoproteins: GP Iα2 (14), Ibα (15), Ibβ (16), V (17) and IX (18); TOLL = Drosophila toll protein (13); CHAO = *Drosophila* chaoptin (12); hPGI = human proteoglycan I (19); PG40 = proteoglycan PG40 (20); OMgp = oligodendrocyte-myelin glycoprotein (21); LCGR = lutropin-choriogonadotropin receptor (23); CD14 = monocyte/granulocyte glycoprotein CD14 (22).

The N-terminal region, which does not contain leucine repeats, is rich in cysteines (lozenges in Fig. 3). In particular, this region contains three C-X2-C motives, with a spacing similar to that of DNA- or RNA-binding zinc fingers of both C4 and C6 types (C-X2-C-X13-C-X2-C, and C-X2-C-X6-C-X6-C-X2-C-X6-C, respectively) (9). The full pattern (C-X2-C-X13-C-X4-C-X2-C-X6-C-X2-C) does not exactly match any of the zinc finger consensus sequences described so far, but falls within the range of observed deviations (10). Clusters of basic residues are present in the direct vicinity (underlined in Fig. 3), and may correspond to nuclear localization signals.

The ESAG 8 protein belongs to the family of 'leucine-rich repeats'

The 23 amino acid-repeats of the ESAG 8 protein can be aligned with the 'leucine-rich' repeats found in a growing number of proteins, particularly surface glycoproteins (Fig. 4). A notable difference between ESAG 8 and the majority of these proteins is the replacement of a highly conserved asparagine by a cysteine at the end of the repeat. Such a substitution has also been observed in one other case, in half of the alternating repeats of the ribonuclease inhibitor, which is so far the only cytosolic protein known to exhibit leucine-rich repeats (Fig. 4). In most cases, there are more than 10 copies of the tandem repeats, making up a substantial portion of the protein. In ESAG 8, the repeats constitute the majority of the protein. The function of this domain would appear to be interaction with other proteins, as discussed below.

Transcription of ESAG 8

The level of transcription in the region of the AnTat 1.3A VSG gene expression site encompassing ESAG 8 has been analysed previously (1). Run-on transcription assays indicated that this region was transcribed at a level comparable to that of other regions from the same transcription unit, but Northern blot hybridization data showed that the level of steady-state transcripts was very low. Fig. 5 shows that, upon long exposure of the autoradiograms, polyadenylated transcripts of about 1.7 and 2.2 kb can be revealed by the ESAG 8 probe, in bloodstream forms only. These transcripts are at least ten fold-less abundant than those of ESAG 4 (1), and about 6-fold less abundant than the actin mRNA (Fig. 5). A full-size ESAG 8 cDNA was cloned into lambda gt10. The size of this cDNA (2.18 kb) is in accordance with that of one of the two major ESAG 8 transcripts (Fig. 5, arrowhead), and its nucleotide sequence was found to be identical to that of the gene. The sites of 3′ splicing and polyadenylation of this ESAG 8 mRNA are indicated in Fig. 1. The origin of the 1.7 kb transcript is unknown.

The ESAG 8 family

Like the other ESAGs, ESAG 8 belongs to a family of several related sequences in different sub-species of the *T. brucei* group. Under relaxed hybridization conditions, ESAG 8-related DNA fragments can also be detected in *T. congolense*. DNA from other kinetoplastids, such as *T. vivax*, *T. cruzi*, *Crithidia fasciculata* and *Leishmania donovani*, does not seem to hybridize with the ESAG 8 probe (data not shown).

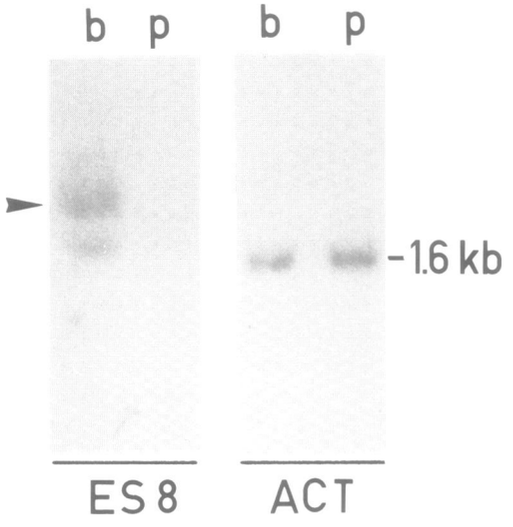

Fig. 5. Steady-state transcripts of ESAG 8. Northern blots of poly(A)+ RNA from bloodstream (b) and procyclic (p) trypanosomes were hybridized with either a 0.57 kb XhoI-DraI fragment from ESAG 8 (ES8) (restriction sites underlined in Fig. 1), or a 1.6 kb SalI fragment encompassing a *T. brucei* actin gene (A-CT), as a control for equal loading of the two RNA samples (36). The arrowhead points to ESAG 8 transcripts with a size (2.2 kb) similar to a cloned cDNA extending from the spliced leader to the poly(A) tail. The exposure of the autoradiograms was 6-fold longer for ESAG 8 than for actin.

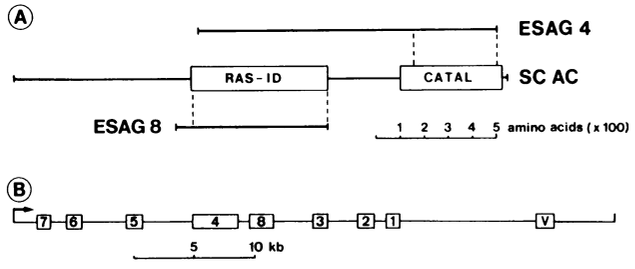

Fig. 6. Relationship between ESAG 8 and genes for adenylate cyclase. Panel A shows the extent of sequence homology between yeast adenylate cyclase (SC AC) and the proteins encoded by ESAG 4 and ESAG 8. The regions homologous to ESAG 4 and ESAG 8 correspond respectively to domains for the adenylate cyclase catalytic activity (CATAL)(28) and responsiveness to RAS (RAS-ID, for RAS-interacting domain)(29, 30). Panel B shows the arrangement of ESAG 4 and ESAG 8 in the AnTat 1.3A VSG gene expression site (1), highlighting the proximity of ESAG 8 to a putative gene for adenylate cyclase (ESAG 4: see ref. 5). Numbered boxes correspond to the ESAG open reading frames, and V indicates the VSG gene. The vertical bar at the end of the map corresponds to the chromosome end, while the horizontal arrow at the beginning shows the start of transcription.

DISCUSSION

We report here the characterization of an eighth non-VSG gene (ESAG) from the 45 kb AnTat 1.3A expression site of *T. brucei*. The discovery of ESAG 8 came as a surprise, as this region from the VSG expression site did not appear to be transcribed into stable polyadenylated RNA (1). Following the present work, a sequence of about 3 kb downstream from ESAG 8 still remains to be examined in detail, so we cannot exclude the possibility that additional ESAGs are also present between ESAG 8 and ESAG 3.

The nucleotide sequence of ESAG 8 predicts a protein of 70 kd. With the exception of the N-terminal region, which may contain a C7 zinc finger, this protein is almost entirely made up of leucine-rich repeats of 23 amino acids. These repeats can be aligned with those found in a variety of proteins, such as adenylate cyclase from yeast (11), chaoptin and the *toll* protein from *Drosophila* (12, 13), many serum glycoproteins and proteoglycans (14–20), the oligodendrocyte-myelin glycoprotein (21), the monocyte/granulocyte cell surface glycoprotein CD14 (22), the lutropin-choriogonadotropin receptor (23) and the placental ribonuclease inhibitor (24, 25). With the exception of the latter, all these proteins are either extracellular, or bound to the plasma membrane. They appear to be involved in either adhesion to, or interaction with, other proteins.

Interestingly, the repeats of the ESAG 8 protein contain a conserved cysteine at a position where asparagine is frequently found. This particularity is shared with only one other protein, the placental ribonuclease inhibitor, where the cysteine is found in half of the repeats, alternating with asparagine. In the ribonuclease inhibitor, this residue seems to be important, since the integrity of the free sulfhydryl groups is essential for protein activity (24, 26). Free sulfhydryl groups could not exist if the protein were exposed outside the cell (27). As the majority of proteins with leucine-rich repeats are extracellular or bound to the plasma membrane, this would explain the absence of cysteine from these repeats. In the ESAG 8 protein, most of the repeats contain a conserved cysteine instead of asparagine. Based on the above considerations, it can be suggested that this protein is not exposed at the cell surface. This hypothesis is in accordance with the hydropathy analysis of the protein, which does not reveal any hydrophobic region characteristic of membrane-interacting domains.

The particularly low level of ESAG 8 transcripts suggests a regulatory function for this gene. The presence of a putative DNA-binding zinc finger, close to a cluster of basic residues perhaps involved in nuclear targeting, suggests a possible interaction of the ESAG 8 protein with DNA. As ESAG 8 is specific to the bloodstream form, an obvious function in this respect would be an involvement in stage-specific control of gene expression. However, as one of many other possibilities, the ESAG 8 protein could also be involved in the control of adenylate cyclase activity of bloodstream forms. This hypothesis stems from the comparison of the adenylate cyclase from *Saccharomyces cerevisiae* with the ESAG 8 protein. Besides the C-terminal catalytic domain, which is homologous to the third half of the ESAG 4-encoded protein (1, 5, 28), the yeast adenylate cyclase contains a domain with tandem leucine-rich repeats homologous to the ESAG 8 protein (11). Thus the yeast adenylate cyclase contains domains corresponding to both ESAG 4 and ESAG 8 polypeptides (Fig. 6A). Whether this observation can be related to the close proximity of ESAG 4 and ESAG 8 in the VSG gene expression site (Fig. 6B) is open to conjecture. In the yeast adenylate cyclase, the leucine-rich domain has been shown to be crucial for interaction with RAS (29, 30). Therefore, it may be hypothesized that ESAG 8 encodes a G protein-interacting domain, perhaps required for adenylate cyclase activity of the ESAG 4 protein. As transient activity assays of plasmid constructs transfected into trypanosomes is now feasible (31), the two main hypotheses concerning the function of the ESAG 8 protein, either DNA binding or control of adenylate cyclase, are amenable to direct experimental verification.

ACKNOWLEDGEMENTS

We thank Dr. D. Jefferies for comments on the manuscript. This work was supported by the Belgian Fonds de la Recherche Scientifique Médicale and research contracts between the University of Brussels and Solvay & Cie (Brussels), and between the University and the Commission of the European Communities (TDS-M-023B). P. R. acknowledges an IRSIA fellowship (Belgium).

REFERENCES

1. Pays, E., Tebabi, P., Pays, A., Coquelet, H., Revelard, P., Salmon, D. and Steinert, M. (1989) Cell **57**, 835–845.
2. Pays, E. and Steinert, M. (1988) Annu. Rev. Genet. **22**, 107–126.
3. Cross, G.A.M. (1990) Annu. Rev. Immunol. **8**, 83–110.
4. Kooter, J.M., van der Spek, H.J., Wagter, R., d'Oliveira, C.E., van der Hoeven, F., Johnson, P.J. and Borst, P. (1987) Cell **51**, 261–272.
5. Alexandre, S., Paindavoine, P., Tebabi, P., Pays, A., Halleux, S., Steinert, M. and Pays, E. (1990) Mol. Biochem. Parasitol., in press.
6. Voorheis, H.P. and Martin, B.R. (1980) Eur. J. Biochem. **113**, 223–227.
7. Rolin, S., Halleux, S., Van Sande, J., Dumont, J.E., Pays, E. and Steinert, M. (1990) Exp. Parasitol. **71**, 350–352.
8. Bowles, D.J. and Voorheis, H.P. (1982) FEBS letters **139**, 17–21.
9. Evans, R.M. and Hollenberg, S.M. (1988) Cell **52**, 1–3.
10. Liebhaber, S.A., Emery, J.G., Urbanek, M., Wang, X. and Cooke, N.E. (1990) Nucl. Acids Res. **18**, 3871–3879.
11. Kataoka, T., Broek, D. and Wigler, M. (1985) Cell **43**, 493–505.
12. Krantz, D.E. and Zipursky, S.L. (1990) EMBO J. **9**, 1969–1977.
13. Hashimoto, C., Hudson, K.L. and Anderson, K.V. (1988) Cell **52**, 269–279.
14. Takahashi, N., Takahashi, Y. and Putnam, F.W. (1985) Proc. Nat. Acad. Sci. USA **82**, 1906–1910.
15. Lopez, J.A., Chung, D.W., Fujikawa, K., Hagen, F.S., Papayannopoulou, T. and Roth, G.J. (1987) Proc. Nat. Acad. Sci. USA **84**, 5615–5619.
16. Lopez, J.A., Chung, D.W., Fujikawa, K., Hagen, F.S., Davie, E.W. and Roth, G.J. (1988) Proc. Nat. Acad. Sci. USA **85**, 2135–2139.
17. Hickey, M.J., Williams, S.A. and Roth, G.J. (1989) Proc. Nat. Acad. Sci. USA **86**, 6773–6777.
18. Roth, G.J., Church, T.A., McMullen, B.A. and Williams, S.A. (1990) Biochem. Biophys. Res. Commun. **170**, 153–161.
19. Fisher, L., Termine, J.D. and Young, M.F. (1989) J. Biol. Chem. **264**, 4571–4576.
20. Krusius, T. and Ruoslahti, E. (1986) Proc. Nat. Acad. Sci. USA **83**, 7683–7687.
21. Mikol, D.D., Gulcher, J.R. and Stefansson, K. (1990) J. Cell Biol. **110**, 471–479.
22. Ferrero, E., Hsieh, C.L., Francke, U. and Goyert, S.M. (1990) J. Immunol. **145**, 331–336.
23. McFarland, K.C., Sprengel, R., Phillips, H.S., Köhler, M., Rosemblit, N., Nikolics, K., Segaloff, D.L. and Seeburg, P.H. (1989) Science **245**, 494–499.
24. Lee, F.S., Fox, E.A., Zhou, H.M., Strydom, D.J. and Vallee, B.L. (1988) Biochemistry **27**, 8545–8553.
25. Hofsteenge, J., Kieffer, B., Matthies, R., Hemmings, B.A. and Stone, S.R. (1988) Biochemistry **27**, 8537–8544.
26. Lee, F.S. and Vallee, B.L. (1990) Proc. Nat. Acad. Sci. USA **87**, 1879–1883.
27. Torchinskii, Y.M. (1974) In Dixon, H.B.F. (ed), Sulfhydryl and Disulfide Groups of Proteins, Plenum, New York, pp 212–214.
28. Yamawaki-Kataoka, Y., Tamaoki, T., Choe, H., Tanaka, H. and Kataoka, T. (1989) Proc. Nat. Acad. Sci. USA **86**, 5693–5697.
29. Colicelli, J., Field, J., Ballester, R., Chester, N., Young, D. and Wigler, M. (1990) Mol. Cell. Biol. **10**, 2539–2543.
30. Field, J., Xu, H.P., Michaeli, T, Ballester, R., Sass, P., Wigler, M. and Colicelli, J. (1990) Science **247**, 464–467.
31. Bellofatto, V. and Cross, G.A.M. (1989) Science **244**, 1167–1169.
32. Pays, E., Delronche, M., Lheureux, M., Vervoort, T., Bloch, J., Gannon, F. and Steinert, M. (1980) Nucl. Acids Res. **8**, 5965–5981.
33. Gubler, U. and Hoffman, B.J. (1983) Gene **25**, 263–269.
34. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B. (1980) J. Mol. Biol. **143**, 161–178.
35. Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. **157**, 105–132.
36. Ben Amar, M.F., Pays, A., Tebabi, P., Dero, B., Seebeck, T., Steinert, M. and Pays, E. (1988) Mol. Cell. Biol. **8**, 2166–2176.