Onderstepoort Journal of Veterinary Research, 60:263–269 (1993)

The rabies virus genome: an overview

NOËL TORDO and ANNE KOUKNETZOFF

Laboratoire des Lyssavirus, Institut Pasteur, 25, rue du Dr Roux, 75724 Paris Cedex 15, France

ABSTRACT

TORDO, N. & KOUKNETZOFF, A. 1993. The rabies virus genome: an overview. *Onderstepoort Journal of Veterinary Research*, 60:263–269

The replication strategy, genome organization and extent of variation within the genome of the genus *Lyssavirus* is briefly reviewed. Strategies used in the approach to genome studies are discussed.

INTRODUCTION

Viruses causing rabies-like encephalitis belong to the *Lyssavirus* genus (Francki, Fauquet, Knudson & Brown 1991). The *Lyssavirus* and *Vesiculovirus* (notably vesicular stomatitis virus) genera together form the family Rhadboviridae. Families Rhabdo-, Paramyxo- and Filoviridae have been taxonomically grouped into the order Mononegavirales because they share an unsegmented negative-stranded RNA genome. "Unsegmented" means that the genes are covalently linked within a single RNA molecule, rather than distributed in distinct segments, as in "segmented" negative-stranded RNA genomes of the families Orthomyxo-, Bunya-, and Arenaviridae (Tordo, Bourhy & Sacramento 1992a). "Negative-stranded" means that the genome cannot be directly translated into viral proteins by the cell ribosomes—a preliminary transcription step is required to produce the complementary positive-stranded messenger RNAs (mRNAs).

STRUCTURAL FEATURES OF THE GENOME

The lyssavirus genome is about 12 kb in length (Tordo, Poch, Ermine, Keith & Rougeon 1988c; Conzelmann, Cox, Schneider & Thiel 1990). From the 3′ to the 5′ end, the genome encodes a short leader RNA (about 50 nucleotides), followed by the genes for the nucleoprotein (N), phosphoprotein (M1), matrix protein (M2), glycoprotein (G) and polymerase (L) (Tordo 1994). Each gene is composed of an internal protein coding region flanked by non-coding regions. The non-coding regions are bordered by start and stop transcription signals consisting of nine nucleotide consensus sequences (Tordo 1986). They indicate where initiation and termination (and polyadenylation) of the mRNAs should occur. The limits of each gene are generally marked by this pair of transcription signals. Only the M2 and G genes of several vaccine strains of rabies virus escape this rule by having two consecutive stop signals (Tordo & Poch 1988a; Morimoto, Ohkubo & Kawai 1989; Sacramento, Bourhy & Tordo 1992). These are used alternately to produce either short or long mRNAs.

The genes are separated from each other by non-transcribed intergenic regions (Tordo *et al.* 1986a). They are generally short (less than five nucleotides) except for the M2-G and G-L intergenic regions when the short M2 and G genes are transcribed. It is of note from an evolutionary point of view that the M2-G and G-L intergenic regions are the most variable in the *Mononegavirales* genome (Tordo, De Haan, Goldbach & Poch 1992b). The G-L intergenic

region, notably, encodes the HN protein in paramyxoviruses or the NV protein in fish rhabdoviruses (Kurath & Leong 1985). Despite its considerable size in lyssaviruses (450 nucleotides), it encodes no substantial polypeptide and has been proposed as a vestigial gene (pseudogene) (Tordo *et al.* 1986a).

The genome is also flanked by external signals: at the 3′ end there is the polymerization promoter, recognized by the polymerase complex to initiate transcription and replication; at the 5′ end there is the encapsidation promoter, recognized by the first molecules of N protein for genome encapsidation (Tordo *et al.* 1988c). The conservation of both promoters results in an inverted complementarity of about ten nucleotides at the genome ends, although to date no evidence of a "hairpin" structure has been obtained.

FUNCTIONAL FEATURES OF THE GENOME

The lyssavirus virion is composed of two distinct structural units: a lipoprotein envelope of cellular origin and an internal helically coiled ribonucleocaspid (RNP) structure embedding the genome so tightly that it is largely insensitive to ribonuclease activity (Tordo & Poch 1988b). The virion enters the cell by pinocytosis (Tsiang 1993): the transmembrane G protein is presumed to bind to an as yet uncharacterized receptor (Rustici, Bracci, Lozzi, Neri, Santucci, Soldani, Spreafico & Neri 1993) on the cell surface and, after fusion of the viral and lysosomal membranes, the RNP is released into the cell cytoplasm. The RNP possesses all the necessary viral elements to ensure transcription and replication (Kawai 1977; Flamand, Delagneau & Bussereau 1978): the N-protein-RNA genome template undergoes no uncoating and the polymerase complex comprises the L protein (actual polymerase) and the M1 protein (cofactor).

Chronologically, transcription preceeds replication. Both mechanisms start at the 3′ genome end promoter and progress towards the 5′ end (Flamand & Delagneau 1978). There is close synchrony between polymerization and encapsidation, the growing RNA being simultaneously coated with N protein from the encapsidation promoter at the 5′ genome end. Transcription produces monocistronic transcripts in response to start and stop signals (Tordo, Poch, Ermine & Keith 1986b; Bourhy, Tordo, Lafon & Sureau 1989). Replication leads to the synthesis of a complete positive-stranded genome that serves in turn to amplify negative-stranded genomes for the progeny virions. Put briefly, the transcriptative RNP is switched to replicative RNP when it ignores the start and stop transcription signals. The degree of replication is dependent on increasing amounts of N protein. Early in infection, only limited amounts of N protein are available. The transcriptase releases a 5′ leader RNA which keeps the encapsidation promoter uncoupled from the following mRNAs that are consecutively transcribed and translated. Once sufficient amounts of N protein are produced, the transcriptase is switched into a replicase. The leader region remains coupled with the rest of the positive-stranded genome whose encapsidation proceeds by virtue of the dual ability of the N protein to bind RNA and to self-assemble. This regulation process permits replication to begin only in the presence of sufficient quantities of N protein to encapsidate the growing template.

Transcription produces monocistronic transcripts in a cascade: first the non-capped, non-polyadenylated leader RNA and then the five capped and polyadenylated mRNAs. A progressive loss of transcriptional efficiency is observed from the 3′ to the 5′ encoded genes, suggesting that the control of gene expression is related to genomic location. This possibility is reinforced by the observation that all members of the order Mononegavirales share a similar genomic organization: the major "structural" proteins, such as the N protein which is required in sufficient quantities to encapsidate the genome, are encoded at the 3′ end while the L polymerase, required in catalytic amounts, is always encoded at the 5′ end (Tordo *et al.* 1992a). Polycistronic events, attributed to recognition failures of the transcription signals by the running transcription complex, are occasionally observed. They occur either

- accidentally and at a very low frequency; or
- sequence specifically, such as the unorthodox M1 stop signal of Mokola which results in a large amount of M1-M2 bicistronic mRNA (Bourhy *et al.* 1989); or
- by modulation of the signal recognition due to the local secondary structure and/or as a result of fixation of transcriptional factors (proteins, peptides) of viral or cellular origin.

The latter is notably the case during the alternative termination of the G and M2 genes where the proximal stop signal must be weakly recognized to allow the production of a long mRNA (Tordo & Poch 1988a; Tordo 1994). Modulation of this recognition during the course of infection is inferred by the changing ratio of short to long mRNAs. This modulation could involve tissue-specific factors because it is different during infection of fibroblastic and neuronal cell cultures. Within the context of sequential transcription, the most likely function of alternative termination is the regulation of the expression of the distal gene. This occurs by release of the transcription complex more or less upstream of the corresponding start signal, rendering reinitiation respectively less, or more, efficient.

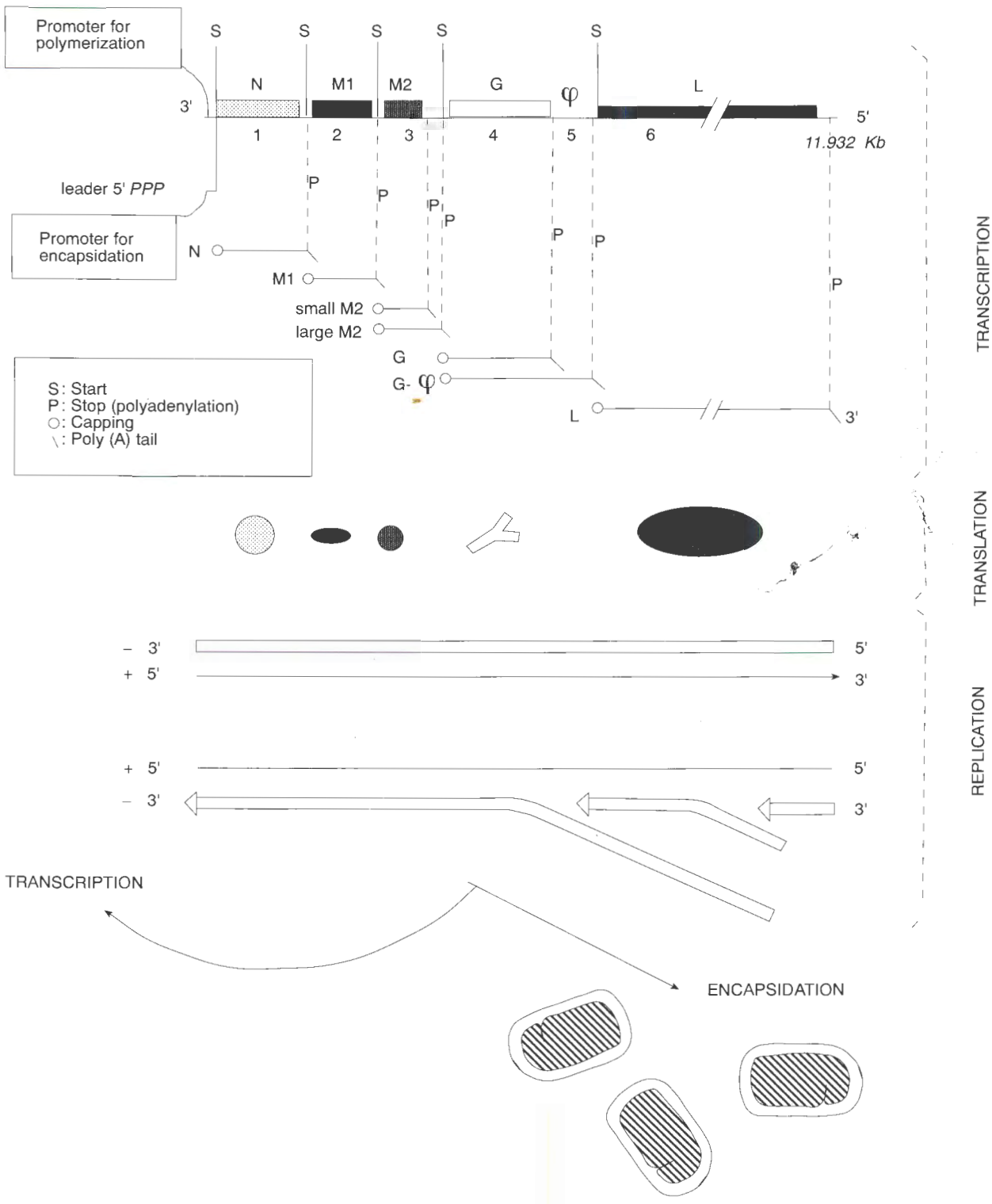

FIG. 1 Transcription and replication mechanism of the lyssavirus genome

TWELVE YEARS OF LYSSAVIRUS MOLECULAR GENETICS

Since 1981, when the G mRNA of the ERA strain became the first rabies sequence published (Anilionis, Wunner & Curtis 1981), knowledge of the molecular genetics of the *Lyssavirus* genus has increased dramatically (Tordo 1994). Studies were initially focused on the vaccinal strains, two of which, PV and SADBI9, were cloned and sequenced (Tordo *et al.* 1986a; Tordo *et al.* 1986b; Tordo *et al.* 1988; Conzelmann *et al.* 1990). In addition, the complete genome of the rabies-related Mokola virus was cloned (Bourhy *et al.* 1989) and sequenced (Bourhy *et al.* 1993; Tordo, Bourhy, Sather & Ollo 1993). Since 1991, the development of a combined reverse transcription/polymerase chain reaction/sequencing (RT/PCR/sequencing) method has allowed the isolation and identification of all lyssavirus genes present in minimal amounts in infected material (Sacramento, Bourhy & Tordo 1991). Besides its use in diagnosis and typing by restriction fragment length polymorphism (RFLP) (Sacramento *et al.* 1991; Smith, Fishbein, Rupprecht & Clark 1991; Bourhy, Kissi, Lafon, Sacramento & Tordo 1992; Nadin-Davies, Casey & Wandeler 1993), this method has enabled intensive molecular epidemiological studies of both rabies and rabies-related viruses to be carried out. Several papers describing a relationship between viral genome variability, isolate taxonomy and geographical location and host species specificity have been published (Benmansour, Brahimi, Tuffereau, Coulon, Lafay & Flamand 1992; Sacramento *et al.* 1992; Smith, Orciari, Yager, Seidel & Warner 1992; Bourhy *et al.* 1992; Bourhy *et al.* 1993; Nadin-Davies *et al.* 1993), as have several reviews which discuss these findings (Tordo 1991; Tordo *et al.* 1992a; Smith & Seidel 1993; Tordo 1994; Tordo, Bourhy & Sacramento 1994).

Analyses of sequence data have led to a better understanding of structure-function relationships of viral elements. Firstly, the genetic bases of previously known structural and immunological features have been elucidated—e.g. characterization of genomic transcriptional and replicational signals, delineation of signal and transmembrane peptides on the G protein, location of sites for post translational modifications (glycosylation, palmitoylation, phosphorylation) (Tuffereau, Fischer & Flamand 1985; Gaudin, Tuffereau, Benmansour & Flamand 1991) and the mapping of regions important in humoral and cellular immunity along the G, N, M1 and M2 proteins (Benmansour, Leblois, Coulon, Tuffereau, Gaudin & Flamand 1991; Coulon, Lafay & Flamand 1993; Tordo 1994). In addition, sequence analysis has revealed new features related to presumptive functions—e.g. a central 19-amino-acid segment of the M2 protein has been shown to be sufficiently hydrophobic to mediate interaction with the membrane during virion morphogenesis (Tordo *et al.* 1986b); the L protein appears to be composed of functional domains linked by hinge regions, a structure consistent with its multifunctional nature (RNA syntheses, capping, polyadenylation, phosphorylation, etc.) (Poch, Blumberg, Bougueleret & Tordo 1990). Further, each functional domain appears to retain relative autonomy as shown by the intragenic complementation which exists between VSV mutants affecting the L gene (Flamand 1970). By sequence comparison of the L proteins of Mononegavirales and sequences available from data banks, several activities have been tentatively attributed to certain domains, providing the first available guideline for future study of the L protein by site-directed mutagenesis (Poch *et al.* 1990). In particular, the active site for polymerization, within domain 3, appears similar among polymerases showing RNA template specificity (RNA polymerases and reverse transcriptases) and among DNA-dependent polymerases (Poch, Sauvaget, Delarue & Tordo 1989; Delarue, Poch, Tordo, Moras & Argos 1990). Such ubiquity strongly suggests the existence of an ancestral "polymerase module" that would have been propagated through the viral kingdom by RNA or DNA recombination. This key finding also illustrates the more widespread relevance of lyssavirus research.

GENETIC VARIABILITY OF THE LYSSAVIRUS GENOME

The genomes of lyssaviruses have been extensively compared both within the genus and with other members of the *Mononegavirales* (Tordo *et al.* 1992b; Tordo 1994): From this it is clear that selection pressure differs along the length of the genome. As an example, Fig. 2 shows a pairwise comparison between rabies and Mokola genomes which represent the two most divergent lyssaviruses of medical interest. It is obvious that the non-coding regions (5′ and 3′ non-translated regions of each gene and the intergenes) are more divergent, particularly the M2-G and G-L intergenes. The requirement of stringent conservation for expression has resulted in preservation of the control and effector elements involved in transcription and replication. At the signal level, the 9 nucleotide start and stop transcription signals are consensus sequences; the genomic ends are strictly conserved up to position 11, reflecting the co-conservation of promoters for RNA synthesis (3′ ends) and encapsidation (5′ ends). At the protein level also, conservation favours the polypeptides involved in transcription and replication (decreasing order: L > N > M2 > G > M1). The L polymerase is the only polypeptide retaining substantial sequence similarity between rhabdoviruses and paramyxoviruses (Poch *et al.* 1990). It includes long stretches of identical amino acids, notably within domain 3 that encompass the "polymerase module" (Poch *et al.* 1989; Delarue *et*

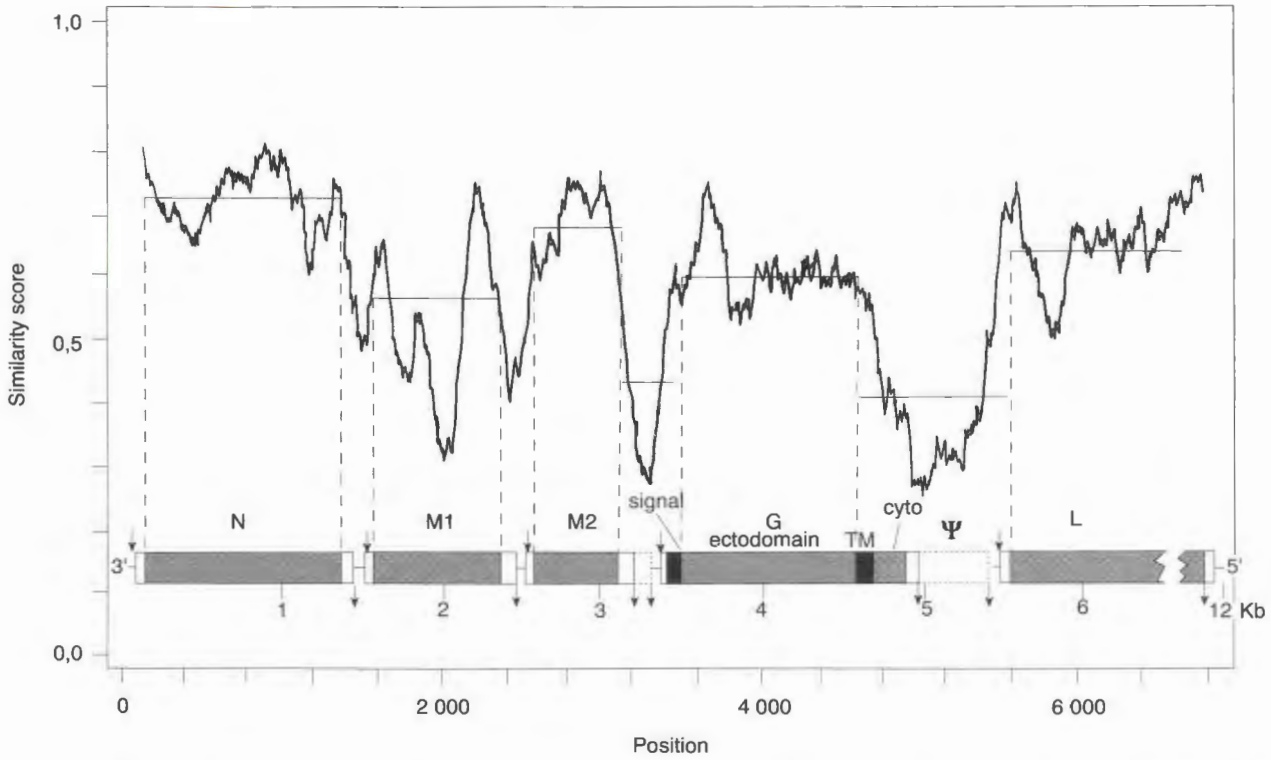

FIG. 2 Pairwise comparison of the first 6500 3′ nucleotides of the rabies (PV strain) and the Mokola (MOK5 isolate) genomes. The similarity profile was calculated by comparing successive windows of 150 nucleotides. The mean value of each region is outlined by horizontal lines. The different domains of the G glycoprotein (signal peptide, ectodomain, transmembrane domain, cytoplasmic domain) are shown

al. 1990; Poch *et al.* 1990). The N protein presents a more conserved central-COOH region where similarities with other Mononegavirales can be observed (Sanchez, Kiley, Klenk & Feldmann 1992). This region is possibly responsible for direct interaction with the genome back-bone. It encompasses candidate T cell epitopes which may extend the protective spectrum of future lyssavirus vaccines (Bourhy *et al.* 1993). Indeed, the M1 protein is the only poorly conserved replicative element, being very variable in the central part. Its role is more likely to be mediated by a highly electronegative charge (numerous acidic amino-acids and phosphate residues) than by defined sequence motifs.

The degree of conservation of the M2 protein comparative to the other proteins is interesting: it varies between good and poor, depending on whether distant [e.g. rabies and Mokola (Bourhy *et al.* 1993)] or closely related [the vaccinal PV and CVS strains (Poch *et al.* 1988)] lyssaviruses are compared. In closely related viruses such as the vaccinal PV and CVS strains, the M2 protein is one of the least conserved (Poch, Tordo & Keith 1988) whereas with the more distantly related rabies and Mokola viruses it is one of the more conserved proteins (Bourhy *et al.* 1993). This suggests that the mutation rate of each protein is not constant, but increases and decreases relative to the changes in evolutionary distance.

The glycoprotein is composed of four distinct domains:

- The signal peptide *(S)* that allows the translocation of the polypeptide through the membrane of the endoplasmic reticulum and that is cleaved from the final polypeptide.
- The ectodomain that is exposed on the outer surface of the virion and includes the glycosylation, palmytoylation and antigenic sites (Tuffereau *et al.* 1985; Benmansour *et al.* 1991; Gaudin *et al.* 1991; Coulon *et al.* 1993).
- The transmembrane peptide *(TM)* that anchors the protein within the viral envelope.
- The cytoplasmic domain *(cyto)* located in the inner part of the virion. *S* and *TM* are hydrophobic, while *cyto* is hydrophilic. The maintenance of their hydropathic nature constitutes the major constraint on these three peptides which accept numerous conservative amino-acid changes. In contrast, the sequence of the ectodomain is more conserved, notably in the NH2 regions which are implicated in the T-cell response and around the main glycosylation site (Tordo *et al.* 1993).

REFERENCES

ANILIONIS, A., WUNNER, W.H. & CURTIS, P. 1981. Structure of the glycoprotein gene in rabies virus. *Nature*, 294:275–278.

BENMANSOUR, A., LEBLOIS, H., COULON, P., TUFFEREAU, C., GAUDIN, Y., FLAMAND, A. & LAFAY, F. 1991. Antigenicity of rabies virus glycoprotein. *Journal of Virology*, 65: 4198–4203.

BENMANSOUR, A., BRAHIMI, M., TUFFEREAU, C., COULON, P., LAFAY, F. & FLAMAND, A. 1992. Rapid sequence evolution of street rabies glycoprotein is related to the heterogeneous nature of the viral population. *Virology*, 187:33–45.

BOURHY, H., TORDO, N., LAFON, M. & SUREAU, P. 1989. Complete cloning and molecular organization of a rabies-related virus: Mokola virus. *Journal of General Virology*, 70: 2063–2074.

BOURHY, H., KISSI, B., LAFON, M., SACRAMENTO, D. & TORDO, N. 1992. Antigenic and molecular characterization of bat rabies virus in Europe. *Journal of Clinical Microbiology*, 30: 2419–2426.

BOURHY, H., KISSI, B. & TORDO, N. 1993. Molecular diversity of the Lyssavirus genus. *Virology*, 194:70–81.

CONZELMANN, K.K., COX, J.H., SCHNEIDER, L.G. & THIEL, H.J. 1990. Molecular cloning and complete nucleotide sequence of the attenuated rabies virus SAD B19. *Virology*, 175: 485–489.

COULON, P., LAFAY, F. & FLAMAND, A. 1993. Rabies virus antigenicity. an overview. *Onderstepoort Journal of Veterinary Research*, 60:271–275.

DELARUE, M., POCH, O., TORDO, N., MORAS, D. & ARGOS, P. 1990. An attempt to unify the structure of polymerases. *Protein Engineering*, 3:461–467.

FLAMAND, A. 1970. Etude genetique du virus de la stomatite vesiculaire. Classement des mutants ts spontanes en groupes de complementation. *Journal of General Virology*, 8:187–195.

FLAMAND, A. & DELAGNEAU, J.F. 1978. Transcriptional mapping of rabies virus in vivo. *Journal of Virology*, 28:518–523.

FLAMAND, A., DELAGNEAU, J.F. & BUSSEREAU, F. 1978. An RNA polymerase activity in purified rabies virions. *Journal of General Virology*, 40:233–238.

FRANCKI, R.I.B., FAUQUET, C.M., KNUDSON, D.L. & BROWN, F. (Eds) 1991. Classification and nomenclature of viruses. *Fifth report of the International Committee on Taxonomy of Viruses*. Wien: Springer-Verlag.

GAUDIN, Y., TUFFEREAU, C., BENMANSOUR, A. & FLAMAND, A. 1991. Fatty acylation of rabies virus proteins. *Virology*, 184: 441–444.

KAWAI, A. 1977. Transcriptase activity associated with rabies virion. *Journal of Virology*, 24:826–835.

KURATH, G. & LEONG, J.C. 1985. Characterization of infectious hematopoietic necrosis virus mRNA species reveals a nonvirion rhabdovirus protein. *Journal of Virology*, 53:462–468.

MORIMOTO, K., OHKUBO, A. & KAWAI, A. 1989. Structure and transcription of the glycoprotein gene of attenuated HEP- Flury strain of rabies virus. *Virology*, 173:465–477.

NADIN-DAVIES, S.A., CASEY, G.A. & WANDELER, A. 1993. Identification of regional variants of the rabies virus within the Canadian province of Ontario. *Journal of General Virology*, 74: 829–837.

POCH, O., TORDO, N. & KEITH, G. 1988. Sequence of the 3386 3′ nucleotides of the genome of the AvO1 strain rabies virus: structural similarities of the protein regions involved in transcription. *Biochemie*, 70:1019–1029.

POCH, O., SAUVAGET, I., DELARUE, M. & TORDO, N. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *The EMBO Journal*, 8:3867–3874.

POCH, O., BLUMBERG, B.M., BOUGUELERET, L. & TORDO, N. 1990. Sequence comparison of five polymerases (L. proteins) of unsegmented negative-strand RNA viruses: theoretical assignments of functional domains. *Journal of General Virology*, 71:1153–1162.

RUSTICI, M., BRACCI, L., LOZZI, L., NERI, P., SANTUCCI, A., SOLDANI, P., SPREAFICO, A. & NERI, N. 1993. A model of the rabies glycoprotein active site. *Biopolymers*, 33:961–969.

SACRAMENTO, D., BOURHY, H. & TORDO, N. 1991. PCR technique as an alternative method for diagnosis and molecular epidemiology of rabies virus. *Molecular and Cellular Probes*, 6:229–240.

SACRAMENTO, D., BADRANE, H., BOURHY, H. & TORDO, N. 1992. Molecular epidemiology of rabies in France: comparison with vaccinal strains. *Journal of General Virology*, 73:1149–1158.

SANCHEZ, A., KILEY, M.P., KLENK, H.D. & FELDMANN, H. 1992. Sequence analysis of the Marburg virus nucleoprotein gene: comparison to Ebola virus and other nonsegmented negative-strand RNA viruses. *Journal of General Virology*, 73: 347–357.

SMITH, J.S., FISHBEIN, D.B., RUPPRECHT, C.E. & CLARK, K. 1991. Unexplained rabies in three immigrants in the United States. A virologic investigation. *The New England Journal of Medicine*, 324:205–211.

SMITH, J.S., ORCIARI, L.A., YAGER, P.A., SEIDEL, H.D. & WARNER, C.K. 1992. Epidemiologic and historical relationships among 97 rabies virus isolates as determined by limited sequence analysis. *Journal of Infectious Diseases*, 166:296–307.

SMITH, J.S. & SEIDEL, H.D. 1993. Rabies: a new look at an old disease, in *Progress in Medical Virology*, edited by J.L. Melnick. Basel: S. Karger: 82–106.

TORDO, N., POCH, O., ERMINE, A., KEITH, G. & ROUGEON, F. 1986a. Walking along the rabies genome: is the large G-L intergenic region a remnant gene? *Proceedings of the National Academy of Sciences*, USA, 83:3914–3918.

TORDO, N., POCH, O., ERMINE, A. & KEITH, G. 1986b. Primary structure of leader RNA and nucleoprotein genes of the rabies genome: segmented homology with VSV. *Nucleic Acids Research*, 14:2671–2683.

TORDO, N. & POCH, O. 1988a. Strong and weak transcription signals within the rabies genome. *Virus Research*, sup. 2, 30.

TORDO, N. & POCH, O. 1988b. Structure of rabies virus, in: *Rabies*, edited by J.B. Campbell & K.M. Charlton. Boston: Kluwer Academic Publishers: 25–45.

TORDO, N., POCH, O., ERMINE, A., KEITH, G. & ROUGEON, F. 1988. Completion of the rabies virus genome sequence determination: highly conserved domains along the L (polymerase) proteins of unsegmented negative-strand RNA viruses. *Virology*, 165:565–576.

TORDO, N. 1991. Contribution of molecular biology to vaccine development and molecular epidemiology of rabies disease. *Memorias do Instituto Butantan*, 53:31–51.

TORDO, N., BOURHY, H. & SACRAMENTO, D. 1992a. Polymerase chain reaction technology for rabies virus, in *Frontiers in Virology*, cited by Y. Becker & G. Darai. Berlin: Springer-Verlag: 389–405.

TORDO, N., DE HAAN, P., GOLDBACH, R. & POCH, O. 1992b. Evolution of negative-stranded RNA genomes. *Seminars in Virology*, 3:341–357.

TORDO, N., BOURHY, H., SATHER, S. & OLLO, R. 1993. Structure and expression in the baculovirus of the Mokola virus glycoprotein: an efficient recombinant vaccine. *Virology*, 194:59–69.

TORDO, N. 1994. Characteristic and molecular biology of the rabies virus, in *Laboratory techniques in rabies*, edited by F.X. Meslin & K. Bogel. Geneva: World Health Organization. In press.

TORDO, N., BOURHY, H. & SACRAMENTO, D. 1994. PCR technology for Lyssavirus diagnosis, in *The polymerase chain reaction for human diagnosis*, edited by J.P. Clewley. London: CRC Press: In press.

TSIANG, H. 1993. Pathophysiology of rabies virus infection of the nervous system. *Advances in Virus Research*, 42:375–412.

TUFFEREAU, C., FISCHER, S. & FLAMAND, A. 1985. Phosphorylation of the N and M1 proteins of rabies virus. *Journal of General Virology*, 66:2285–2289.