

VECTOR COMPRISING MULTIPLE HOMOLOGOUS NUCLEOTIDE SEQUENCES

BACKGROUND OF THE INVENTION

The phenomenon of homologous recombination of nucleic acids involves physical breaking and crosswise rejoining of nucleic acid strands within homologous sequences. Recombination and gene conversion in mammalian cells have been studied by many groups who have monitored the reconstruction of selectable genes after infection with appropriately constructed viral or plasmid substrates. (Chakrabarti et al., Mol. Cell. Biol. 6:2520-2526, 1986). The results of these experiments indicate that cells efficiently support both intra- and intermolecular recombination and gene conversion. (*Id.*) Intermolecular recombination refers to recombination between homologous sequences present on two different nucleic acid molecules, while intramolecular recombination refers to recombination between homologous sequences present on a single nucleic acid molecule.

Intermolecular recombination can occur between genes in a plasmid or virus and homologous sequences within a cell. (Miller et al., Mol. Cell. Biol. 6:2895-2902, 1986.) This type of recombination can cause the generation of an infectious virus from an attenuated virus. Fuller et al. codon-optimized the separated sequences of the HIV-1 *gag* and the HIV-1 *pol* gene to increase its expression in mammalian cells. These optimizations also reduced identity of nucleotides in an overlapping region of about 200 base pairs present in the *gag-pol* gene of HIV, which also resulted in reduced levels of intermolecular recombination between the *gag* and *pol* open reading frames placed on two independent plasmids and the truncated *gag* gene contained in a recombinant retroviral vector. (Fuller et al., Hum. Gene Ther. 12:2081-2093, 2001.)

Intramolecular recombination can occur with vectors in which duplicated regions of a gene or a gene fragment are present as direct repeats separated by intervening sequences. This type of recombination generally results in the deletion of the intervening sequences and one copy of the repeated sequences. The frequency of intramolecular recombination is generally a great deal higher than for intermolecular recombination.

The level of intramolecular recombination within a plasmid vector has been quantitated in mammalian cells. (Rubnitz and Subrami, Mol. Cell. Biol. 4:2253-2258, 1984.) Depending upon the size of the homologous regions, the frequency of intramolecular recombination within a

transfected plasmid DNA varied between 0.306% and 0.002%. (*Id.*) Low recombination efficiencies were seen with as little as 14 bases of homology. (*Id.*)

Intramolecular recombination between homologous sequences has been also documented in a number of animal viruses including picornaviruses, influenza virus, adenovirus, and poxviruses. (Gritz et al., J. Virol. 64:5948-5957, 1990). In vaccinia viruses, it has been shown that tandemly duplicated sequences are genetically unstable. (*Id.*) In viruses, a level of intramolecular recombination has been seen that is much higher than that seen with plasmid vectors.

For example, in a retrovirus, the frequency of recombination between two identical sequences in the same RNA molecule was found to be about 62%. (Zhang et al., J. Virol. 75:6348-6358, 2001). 99% of these recombinations were intramolecular (between two sequences on one RNA molecule), as opposed to intermolecular (between two RNA molecules). (*Id.*) With adeno-associated virus, intramolecular recombination was also found to be far more efficient than intermolecular recombination. (Choi et al., J. Virol. 79:6801-6807, 2005). Herpes simplex virus type 1 has also been shown to exhibit high levels of recombination. (Dutch et al., J. Virol. 66:277-285.) In poxviruses, a high frequency of homologous recombination has been seen. An experimental system was used to measure recombination in a vaccinia virus by placing a thymidine kinase (*tk*) gene between two direct repeats of 1.5 kb of DNA. (Ball, J. Virol. 61:1788-1795, 1987.) During each of the first eight passages under non-selective conditions, 40% of *tk*+ vaccinia viruses lost their *tk*+phenotype. (*Id.*) Under non-selective conditions, the *tk*- virus increased to an abundance of 99.73% of the total virus population. (*Id.*) Even under selective conditions, recombination occurred with such high frequency that the majority of infectious virus particles that could be isolate from single plaques contained DNA that had already undergone recombination with subsequent loss of the *tk* gene. (*Id.*) Using a recombinant vaccinia virus designed to express three heterologous genes, all expressed from VV p7.5-promoters, Howley et al., Gene 172:233-237, 1996, demonstrated recombination between the repeated promoter sequences. A vaccinia virus recombinant designed to contain a C-repeat region (CRR) from the M protein of *Streptococcus pyogenes* contained a complex mixture of variants containing from 1 to more than 20 copies of the CRR. (Hruby et al., P.N.A.S. 88:3190-3194, 1991.)

Although it has been shown that multiple genes with homology of about 60-75% inserted into different insertion sites of MVA resulted in a stable multiple recombinant virus (WO 03/097846), there is, however, a need in the art for compositions and methods that reduce the level of intramolecular recombination in vectors, such as, e.g., viral vectors to allow the generation of stable vectors including multiple homologous nucleotide sequences containing longer stretches of identity.

The invention is defined in the independent claims. Some optional features of the invention are defined in the dependent claims.

THE INVENTION

The present invention relates to recombinant vectors and methods for making and using them.

Disclosed herein is a vector comprising two nucleotide sequences of 300 nucleotides in size each coding for 100 amino acids, wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity and wherein one of the two nucleotide sequences has at least 75 nucleotides different from the other nucleotide sequence, wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences.

Surprisingly, it was shown that the risk of intramolecular recombination can not only be significantly reduced, but even be avoided by systematically substituting synonymous codons in at least two similar or identical nucleotide sequences within one nucleic acid molecule, such as, for example a vector, thus leading to the generation of stable vectors containing at least two or more similar or identical nucleotide sequences. Unexpectedly, the strategy employed herein is also applicable to vectors containing three or more similar nucleotide sequences.

The results obtained show that it is possible to substitute a high number of nucleotides in nucleotide sequences to reduce intramolecular recombination within a vector, while, surprisingly, at the same time expression of the encoded protein is still retained: When introducing a high number of nucleotide variants into long stretches of a nucleotide sequence as disclosed herein, the skilled practitioner would have expected that expression of said sequence or gene would not work properly any more, i.e., it was not expected that the changed nucleotide sequence would remain suitable for efficient expression. The strategy employed herein is not only applicable to short nucleotide sequence stretches of 300 nucleotides, but also to much longer stretches as, e.g., full-length genes which, of course, include a stretch of 300 nucleotides as claimed. The results are applicable to many different genes, vectors and viruses and are highly advantageous for vaccine development, such as for example the development of multivalent vaccines, but may also be advantageous for other technologies as, for example, expression of proteins or for the generation of recombinant cell lines.

The techniques disclosed herein also encompass methods for the generation of viruses and vectors, and methods for reducing intramolecular recombination.

A disclosed method for generating a vector as described above comprises the steps of a) providing a first nucleotide sequence of 300 nucleotides in size coding for 100 amino acids and b) providing a second nucleotide sequence of 300 nucleotides in size coding for 100 amino acids, wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity and wherein one of the two nucleotide sequences has at least 75 nucleotides different from the other nucleotide sequence, wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences; and c) inserting the two divergent nucleotide sequences into a vector.

In a particularly preferred method for reducing intramolecular recombination within a vector containing two nucleotide sequences of 300 nucleotides in size, each coding for 100 amino acids, wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity, said method comprises substituting nucleotides in one or both nucleotide sequence(s) to generate two divergent sequences which show differences in at least 75 nucleotides, wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences.

When using viral vectors, the method reduces the level of intramolecular recombination during each generation of viral propagation. Preferably, the homologous nucleotide sequences recombine in less than 20%, 15%, 10%, 5%, 3%, 2%, 1%, 0.5%, 0.1%, 0.05% or 0.01%. of the progeny viruses per generation.

In another preferred method for generating a virus, preferably a poxvirus, comprising two homologous nucleotide sequences, said method comprises the steps of a) providing a virus comprising a nucleotide sequence of 300 nucleotides in size coding for 100 amino acids and b) inserting a second nucleotide sequence of 300 nucleotides in size coding for 100 amino acids into the virus; wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity, and wherein one of the two nucleotide sequences has at least 75 nucleotides different from the other nucleotide sequence, wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences.

As used herein, a "vector" may be any agent capable of delivering and expressing nucleic acid molecules in a host cell or subject. Thus, a vector may be a PCR product or any piece of nucleic acid introduced into a cell and/or integrated in the cellular genome; or a replicon, such as a plasmid, phage, or cosmid, into which another DNA segment may be inserted so as to bring about the replication of the inserted segment. Generally, a vector may replicate when associated with the proper control elements. Suitable vector backbones for use in the present techniques

include, for example, those routinely used in the art such as plasmids, viruses, artificial chromosomes, BACs, YACs, or PACs or even recombinant cells like bacteria and eukaryotic cells. The term “vector” includes cloning and expression vectors, as well as viral vectors and integrating vectors. An “expression vector” is a vector that includes a regulatory region. Suitable expression vectors for use in the present techniques include, without limitation, plasmids and viral vectors derived from, for example, plant viruses, bacteriophage, baculoviruses, tobacco mosaic virus, retroviruses, and poxviruses. Suitable non viral vectors include plasmids such as pREP4, pCEP4 (Invitrogen), pCI (Promega), pCDM8 (Seed, 1987, Nature 329, 840), pVAX and pgWiz (Gene Therapy System Inc; Himoudi et al, 2002, J. Virol. 76, 12735-12746). Numerous vectors and expression systems are commercially available from such corporations as Novagen (Madison, Wis.), Clontech (Palo Alto, Calif.), Stratagene (La Jolla, Calif.), and Invitrogen/Life Technologies (Carlsbad, Calif.).

In vaccine development, a recombinant virus can be used as the vehicle or vaccine vector for delivering genetic material to a cell. Once in the cell, genetic information is transcribed and translated into proteins, including the inserted antigen targeted against a specific disease. Treatment is successful if the antigen delivered by the vector into the cell induces the body's immune response against the antigen that protects against the disease.

In a preferred embodiment, the vector is a plasmid or a viral vector.

A viral vector can be based on an attenuated virus, which cannot replicate in the host, but is able to introduce and express a foreign gene in the infected cell. The virus or the recombinant virus is thereby able to make a protein and display it to the immune system of the host. Some key features of viral vectors are that they can elicit a humoral (B-cell) and/or cell-mediated (T-cell) immune response.

Viral vectors may be obtained from a variety of different viruses. In one embodiment, the virus is an animal virus. The vector may be obtained especially from a virus selected from the group consisting of retrovirus, picornavirus, influenza virus, adenovirus, adeno-associated virus (AAV), poxvirus, herpes virus (e.g., HSV-1), measles virus and foamy virus.

Viral vectors are commonly used by researchers to develop vaccines for the prevention and treatment of infectious diseases and cancer. Of these, poxviruses (including canary pox, vaccinia, and fowl pox) are belonging to the group of the most common vector vaccine candidates. Poxviruses are a preferred choice for transfer of genetic material into new hosts due to the relatively large capacity for insertion of sequences into the viral genome and because of their ability to replicate their genomes and perform transcription in the infected cell's cytoplasm

instead of the nucleus, thereby minimizing the risk of insertional mutagenesis by integrating genetic material into the genome of the host cell as seen with other vectors, e.g. retroviral vectors. The virions of poxviruses are large as compared to most other animal viruses (for more details see Fields et al., eds., Virology, 3rd Edition, Volume 2, Chapter 83, pages 2637 ff).

In a preferred embodiment, the viral vector is derived from a poxvirus (see for example Cox et al. in "Viruses in Human Gene Therapy" Ed J. M. Hos, Carolina Academic Press). It may be obtained from any member of the poxviridae and may be, in particular an avipoxvirus or an orthopoxvirus.

Examples for avipoxviruses suitable for use in the present techniques include any avipoxvirus such as Fowlpoxvirus, Canarypoxvirus, Uncopoxvirus, Mynahpoxvirus, Pigeonpoxvirus, Psittacinepoxvirus, Quailpoxvirus, Peacockpoxvirus, Penguinpoxvirus, Sparrowpoxvirus, Starlingpoxvirus and Turkeypoxvirus. Preferred avipoxviruses are Canarypoxvirus and Fowlpoxvirus.

Avipoxviruses are naturally host-restricted and productively replicate only in avian species and cells (Taylor et al., Biological and immunogenic properties of a canarypox-rabies recombinant, ALVAC-RG (vCP65) in non-avian species, Vaccine 13 :539-549, 1995). If human cells are infected with an avipoxvirus, heterologous genes are expressed from the viral genome. However, the avipoxvirus does not fully replicate in the human cells and there is, thus, no risk that the human being is harmed by productive virus replication. Various recombinant avipoxviruses have been constructed that express e. g. lentiviral gene products (US 5,766, 598), cytokines and/or tumor-associated antigens (US 5, 833, 975) or rabies G glycoprotein (Taylor et al., Biological and immunogenic properties of a canarypox-rabies recombinant, ALVAC-RG (vCP65) in non-avian species, Vaccine 13: 539-549, 1995). A recombinant canarypox virus expressing the four HIV genes gag, pol, env and nef has already been used in clinical trials (Peters, B. S., The basis for HIV immunotherapeutic vaccines, Vaccine 20: 688-705, 2001).

Since avipoxviruses productively replicate only in avian cells, these cells have to be used for the amplification of the virus and for the generation of recombinant viruses.

An example for a canarypox virus is strain Rentschler. A plaque purified Canarypox strain termed ALVAC (US 5,766, 598) was deposited under the terms of the Budapest treaty with the American Type Culture Collection (ATCC), accession number VR-2547. Another Canarypox strain is the commercial canarypox vaccine strain designated LF2 CEP 524 24 10 75, available from Institute Merieux, Inc.

Examples of a Fowlpox virus are strains FP-1, FP-5 and TROVAC (US 5,766, 598). FP-1 is a Duvette strain modified to be used as a vaccine in oneday old chickens. The strain is a commercial fowlpox virus vaccine strain designated 0 DCEP 25/CEP67/2309 October 1980 and is available from Institute Merieux, Inc. FP-5 is a commercial fowlpox virus vaccine strain of chicken embryo origin available from American Scientific Laboratories (Division of Schering Corp.) Madison, Wisconsin, United States Veterinary License No. 165, serial No. 30321.

Of the poxviruses, the vaccinia and variola species are the two best known. Variola virus is the cause of smallpox. In contrast to variola virus, vaccinia virus does not normally cause systemic disease in immune-competent individuals and it has therefore been used as a live vaccine to immunize against smallpox. Successful worldwide vaccination with vaccinia virus culminated in the eradication of smallpox as a natural disease in the 1980s (The global eradication of smallpox. Final report of the global commission for the certification of smallpox eradication; History of Public Health, No. 4, Geneva: World Health Organization, 1980). Since then, vaccination has been discontinued for many years, except for people at high risk of poxvirus infections (for example, laboratory workers). However, there is an increasing fear that, for example, variola causing smallpox may be used as a bio-terror weapon. Furthermore, there is a risk that other poxviruses such as cowpox, camelpox, and monkeypox may potentially mutate, through selection mechanisms, and obtain similar phenotypes as variola. Several governments are therefore building up stockpiles of vaccinia-based vaccines to be used either pre-exposure (before encounter with variola virus) or post-exposure (after encounter with variola virus) of a presumed or actual smallpox attack.

In a particular preferred embodiment, the vector is a vaccinia virus vector.

Vaccinia virus is highly immune-stimulating and provokes strong B- (humoral) and T-cell mediated (cellular) immunity to both, its own gene products and to many foreign gene products expressed from genes inserted in the vaccinia genome. Vaccinia virus is, therefore, seen as an ideal vector for vaccines against smallpox and other infectious diseases and cancer in the form of recombinant vaccines. Many of the recombinant vaccinia viruses described in the literature are based on the fully replication competent Western Reserve strain of Vaccinia virus. However, it is known that this strain has a high neurovirulence and is, thus, poorly suited for use in humans and animals (Morita et al. 1987, Vaccine 5, 65-70).

A suitable vaccinia virus can be selected from the group consisting of the Copenhagen strain (Goebel et al., 1990, Virol. 179, 247-266 and 517-563; Johnson et al., 1993, Virol. 196, 381-401), the Wyeth strain, NYVAC (see WO92/15672 and Tartaglia et al., 1992, Virology 188, 217-

232) and the highly attenuated modified Ankara (MVA) strain (Mayr et al., 1975, *Infection* 3, 6-16).

A preferred example of a suitable vaccinia virus is the highly attenuated vaccinia virus strain NYVAC, which was derived from a plaque-cloned isolate of the Copenhagen vaccine strain by deletion of 18 ORFs from the viral genome (Tartaglia et al., NYVAC: A highly attenuated strain of vaccinia virus, *Virology* 188, 217-232, 1992). NYVAC is characterized by a dramatically reduced ability to replicate on a variety of human tissue culture cells, but retains the ability to induce strong immune responses to extrinsic antigens.

All of the above-described viruses are equally suitable for use in the present techniques.

In a most preferred embodiment, the virus is a modified vaccinia virus Ankara (MVA) which is known to be exceptionally safe in vaccinations.

Modified Vaccinia virus Ankara (MVA) virus is related to Vaccinia virus, a member of the genus Orthopoxvirus in the family Poxviridae. MVA has been generated by 516 serial passages on chicken embryo fibroblasts of the dermal vaccinia strain Ankara (Chorioallantois vaccinia virus Ankara, CVA) (for review see Mayr, A., et al., Passage History: Abstammung, Eigenschaften und Verwendung des attenuierten Vaccinia-Stammes MVA, *Infection* 3, 6-14, 1975). As a consequence of these long-term passages the resulting MVA virus deleted about 31 kilobases of its genomic sequence and, therefore, was described as highly host cell restricted to avian cells (Meyer, H. et al., Mapping of deletions in the genome of the highly attenuated vaccinia virus MVA and their influence on virulence, *J. Gen. Virol.* 72, 1031-1038, 1991; (Meisinger-Henschel et al., Genomic sequence of chorioallantois vaccinia virus Ankara, the ancestor of modified vaccinia virus Ankara, *J. Gen. Virol.* 88, 3249-3259, 2007). It was shown, in a variety of animal models that the resulting MVA was significantly avirulent (Mayr, A. & Danner, K. Vaccination against pox diseases under immunosuppressive conditions, *Dev. Biol. Stand.* 41: 225-34, 1978). Additionally, this MVA strain has been tested in clinical trials as vaccine to immunize against the human smallpox disease (Mayr et al., *Zbl. Bakt. Hyg. I, Abt. Org. B* 167, 375-390 [1987], Stickl et al., MVA vaccination against smallpox: clinical tests with an attenuated live vaccinia virus strain (MVA) (author's transl), *Dtsch. med. Wschr.* 99, 2386-2392, 1974). These studies involved over 120,000 humans, including high risk patients, and proved that, compared to Vaccinia based vaccines, MVA had diminished virulence or infectiousness while it maintained good immunogenicity.

The techniques disclosed herein encompass recombinant MVA viruses generated with any and all MVA viruses. An example for an MVA strain is deposit VR-1508, deposited at the American

Type Culture collection (ATCC), Manassas, VA 20108, USA. In another embodiment the MVA-Vero strain or a derivative thereof can be used. The strain MVA-Vero has been deposited at the European Collection of Animal Cell Cultures under the deposition number ECACC V99101431 and ECACC 01021411. Further examples for MVA virus strains used are strains MVA 572 and 575 deposited at the European Collection of Animal Cell Cultures (ECACC), Salisbury (UK) with the deposition number ECACC V94012707 and ECACC V00120707, respectively. Particularly preferred MVA viruses are MVA variant strains MVA-BN® as, e.g., deposited at ECACC under number V00083008, and derivatives having the same properties as MVA-BN®.

MVA-BN® is a virus used in the manufacturing of a stand-alone third generation smallpox vaccine. MVA-BN® was developed by further passages from MVA strain 571/572. To date, more than 1500 subjects including subjects with atopic dermatitis (AD) and HIV infection have been vaccinated in clinical trials with MVA-BN® based vaccines.

Derivatives having the same properties as the deposited strain of MVA-BN® have the capability of reproductive replication in vitro in chicken embryo fibroblasts (CEF), but no capability of reproductive replication in human cells in which MVA 575 or MVA 572 can reproductively replicate. Most preferably, the MVA has no capability of reproductive replication in the human keratinocyte cell line HaCaT, the human embryo kidney cell line 293, the human bone osteosarcoma cell line 143B, and the human cervix adenocarcinoma cell line HeLa.

The term "not capable of reproductive replication" is used in the present application as defined in WO 02/42480 and U.S. Patent 6,761,893, respectively. Thus, said term applies to a virus that has a virus amplification ratio at 4 days after infection of less than 1 using the assays described in U.S. Patent 6,761,893, which assays are hereby incorporated by reference. The "amplification ratio" of a virus is the ratio of virus produced from an infected cell (Output) to the amount originally used to infect the cells in the first place (Input). A ratio of "1" between Output and Input defines an amplification status wherein the amount of virus produced from the infected cells is the same as the amount initially used to infect the cells.

In a most preferred embodiment, the MVA strain used is MVA-BN® or a derivative as described above. The features of MVA-BN®, the description of biological assays allowing evaluating whether an MVA strain is MVA-BN® or a derivative thereof and methods allowing to obtain MVA-BN® or an MVA having the properties of MVA-BN® are disclosed in WO 02/42480. The content of this application is included in the present application by reference. The highly attenuated MVA-BN® virus can be derived, e.g., by the further passage of a modified vaccinia virus Ankara (MVA), such as MVA-572 or MVA-575 and, optionally, by plaque or clone

purification. MVA-BN® lacks approximately 13% (26.5 kb from six major and multiple minor deletion sites) of the genome compared with ancestral CVA virus. The deletions affect a number of virulence and host range genes, as well as a large fragment of the gene coding for A-type inclusion protein (ATI) and a gene coding for a structural protein directing mature virus particles into A-type inclusion bodies.

In particular, reference is made to the definition of the properties of the MVA as described in WO 02/42480, such as the properties of MVA-BN® and the properties and definitions of the derivatives of MVA-BN®. Said reference also discloses how MVA and other vaccinia viruses can be propagated. Briefly, eukaryotic cells are infected with the virus. The eukaryotic cells are cells that are susceptible to infection with the respective poxvirus and allow replication and production of infectious virus. For MVA an example for this type of cells are chicken embryo fibroblasts (CEF) and BHK cells (Drexler et al., Highly attenuated modified vaccinia Ankara replicates in baby hamster kidney cells, a potential host for virus propagation, but not in various human transformed and primary cells, J. Gen. Virol. 79, 347-352, 1998). CEF cells can be cultivated under conditions known to the person skilled in the art. Preferably the CEF cells are cultivated in serum-free medium in stationary flasks or roller bottles. The incubation preferably takes place 48 to 96 hours at 37°C. For the infection MVA is preferably used at a multiplicity of infection (MOI) of 0,05 to 1 TCID₅₀ and the incubation preferably takes place 48 to 72 hours at 37°C.

The viruses as used can be propagated on various cell cultures, particularly animal cell cultures. The virus is allowed to infect susceptible cell cultures and reproductively replicate. Progeny viruses are collected by routine techniques in the art.

For example, with MVA viruses and other vaccinia viruses, chicken embryo fibroblasts (CEFs) in serum-containing or serum-free medium can be infected with the viruses. After the virus has been allowed to reproductively replicate, progeny viruses are collected.

The present disclosure also relates to a recombinant poxvirus, preferably vaccinia virus, in particular MVA, capable of expressing two or more homologous nucleotide sequences, in particular coding sequences. The virus can contain two, three, four or more homologous nucleotide coding sequences.

The vector of the present techniques may comprise two nucleotide sequences of 300 nucleotides in size. In a preferred embodiment, the vector comprises three, four, five, six or more nucleotide sequences, which, of course, encompass also two nucleotide sequences as claimed. 300 nucleotides may, of course, also be part of a longer nucleotide sequence.

Additionally, in various embodiments, the two or more nucleotide sequences are 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, or 3000 or even more nucleotides in size which may all be part of longer nucleotide sequences and which, of course, all include 300 nucleotides as claimed.

As used herein, the terms “polynucleotide”, “nucleotide sequence”, “nucleic acid”, “nucleic acid molecule” “nucleic acid sequence” are used interchangeable and define a polymer of either polydeoxyribonucleotides (DNA) or polyribonucleotides (RNA) molecules or any combination thereof. The definition encompasses single or double-stranded, linear or circular, naturally occurring or synthetic polynucleotides.

The nucleotide sequences of the present techniques may be coding sequences and can contain complete genes, respectively. The term “coding sequence”, as used herein, refers to a nucleotide sequence that codes for a specific amino acid sequence. Non-coding sequences of genes include introns and control regions, such as promoters, operators, and terminators.

The nucleotide sequences can also contain gene fragments. The nucleotide sequences can contain synthetic sequences, such as nucleotide sequences encoding amino acid linker sequences or epitopes. The nucleotide sequences can be composed of a mixture of genes, gene fragments, and synthetic sequences. The nucleotide sequence may also contain analogs such as nucleotide analogs, phosphate ester analog and/or pentose sugar analog. Also included within the definition of nucleotide analogs are nucleotides in which the phosphate ester and/or sugar phosphate ester linkages are replaced with other types of linkages, such as N-(2-aminoethyl)-glycine amides and other amides (see, e.g., Nielsen et al., 1991, Science 254: 1497-1500; WO 92/20702; U.S. Pat. No. 5,719,262; U.S. Pat. No. 5,698,685;); morpholinos (see, e.g., U.S. Pat. No. 5,698,685; U.S. Pat. No. 5,378,841; U.S. Pat. No. 5,185,144); carbamates (see, e.g., Stirchak & Summerton, 1987, J. Org. Chem. 52: 4202); methylene(methylimino) (see, e.g., Vasseur et al., 1992, J. Am. Chem. Soc. 114: 4006); 3'thioformacetals (see, e.g., Jones et al., 1993, J. Org. Chem. 58: 2983); sulfamates (see, e.g., U.S. Pat. No. 5,470,967); 2-aminoethylglycine, commonly referred to as PNA (see, e.g., Buchardt, WO 92/20702; Nielsen (1991) Science 254:1497-1500); and others (see, e.g., U.S. Pat. No. 5,817,781; Frier & Altman, 1997, Nucl. Acids Res. 25:4429 and the references cited therein). Phosphate ester analogs include, but are not limited to, (i) C1-C4 alkylphosphonate, e.g. methylphosphonate; (ii) phosphoramidate; (iii) C1-C6 alkyl-phosphotriester; (iv) phosphorothioate; and (v) phosphorodithioate.

Further modifications include chemical modifications (e.g. see WO 92/03568; US 5,118,672) in order to increase the in vivo stability of the nucleic acid, enhance the delivery thereof, or reduce the clearance rate from the host subject.

Furthermore, in one embodiment, the nucleotide sequence can contain fusion genes, artificial genes and polypeptides.

A fusion gene, as denoted herein, is a hybrid gene formed from two previously separate genes, gene fragments or artificial DNA or epitopes. It can occur as the result of a translocation, interstitial deletion, or inversion.

A fusion gene can be constructed by linking at least two DNA fragments, wherein the DNA fragments encode identical or different amino acid sequences

Fusion proteins may facilitate the expression and/or purification of proteins. For example, a polypeptide may be generated as a glutathione-S-transferase (GST) fusion protein. Such GST fusion proteins may be used to simplify purification of a polypeptide, such as through the use of glutathione-derivatized matrices (see, for example, Current Protocols in Molecular Biology, eds. Ausubel et al., (N.Y.: John Wiley & Sons, 1991)). In another embodiment, a fusion gene coding for a purification leader sequence, such as a poly-(His)/enterokinase cleavage site sequence at the N-terminus of the desired portion of the recombinant protein, may allow purification of the expressed fusion protein by affinity chromatography using a Ni²⁺ metal resin. The purification leader sequence may then be subsequently removed by treatment with enterokinase to provide the purified protein (e.g., see Hochuli et al., (1987) J. Chromatography 411 : 177; and Janknecht et al., PNAS USA 88:8972). Further heterologous sequences encoding a polypeptide permitting the detection, isolation, solubilization and/or stabilization of the polypeptide to which it is fused, include poly His tag, myc, HA, protein A, protein G, calmodulin-binding peptide, thioredoxin, maltose-binding protein, poly arginine, poly His-Asp, FLAG, a portion of an immunoglobulin protein, and a transcytosis peptide.

Techniques for making fusion genes are well known. Essentially, the joining of various DNA fragments coding for different polypeptide sequences is performed in accordance with conventional techniques, employing blunt-ended or stagger-ended termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and enzymatic ligation. In another embodiment, the fusion gene may be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments may be carried out using anchor primers which give rise to complementary overhangs between two consecutive

gene fragments which may subsequently be annealed to generate a chimeric gene sequence (see, for example, Current Protocols in Molecular Biology, eds. Ausubel et al., John Wiley & Sons: 1992) and by fusion PCR wherein two or more polynucleotides are sharing a stretch of identity, which in a PCR reaction can result in fused polynucleotide sequences.

In another preferred embodiment, the nucleotide sequence of the present techniques encodes a polyepitope. A polyepitope is a chimeric protein containing isolated epitopes from at least one protein/antigen, preferably from more than one protein/antigen.

Said epitopes can be "isolated" or "biologically pure". The term "isolated" refers to material that is substantially free from components that normally accompany it as found in its naturally occurring environment. An "isolated" epitope refers to an epitope that does not include the neighbouring amino acids of the whole sequence of the antigen or protein from which the epitope was derived.

With regard to a particular amino acid sequence, an "epitope" is a set of amino acid residues which is involved in recognition by a particular immunoglobulin, or in the context of T cells, those residues necessary for recognition by T cell receptor proteins and/or Major Histocompatibility Complex (MHC) molecules. The term "peptide" designates a series of amino acids, connected one to the other, typically by peptide bonds between the amino and carboxyl groups of adjacent amino acids.

The epitopes are of a certain length and bind to a molecule functioning in the immune system, preferably a HLA class I and a T-cell receptor. The epitopes in a polyepitope construct can be HLA class I epitopes and optionally HLA class II epitopes. HLA class I epitopes are referred to as CTL epitopes and HLA class II epitopes are referred to as HTL epitopes. Some polyepitope constructs can have a subset of HLA class I epitopes and another subset of HLA class II epitopes. A CTL epitope usually consists of 13 or less amino acid residues in length, 12 or less amino acids in length, or 11 or less amino acids in length, preferably from 8 to 13 amino acids in length, most preferably from 8 to 11 amino acids in length (i.e. 8, 9, 10, or 11). A HTL epitope consists of 50 or less amino acid residues in length, and usually from 6 to 30 residues, more usually from 12 to 25, and preferably consists of 15 to 20 (i.e. 15, 16, 17, 18, 19, or 20) amino acids in length. The polyepitope construct preferably includes 2 or more, 5 or more, 10 or more, 13 or more, 15 or more, 20 or more, or 25 or more CTL epitopes. More specific, the polyepitope construct comprises at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 55, 60 or more CTL epitopes.

The homologous nucleotide sequences can be derived from any organism, microorganism, such as any virus, any bacterium, any fungus or parasite. The homologous nucleotide sequences can be either heterologous to the sequence of the vector, but also homologous thereto: When, for example, a virus is used as a vector, also viral own nucleotide sequences can be multiplied, for example, in order to overexpress a protein of the virus for getting enhanced immune reactivity or safety. Preferably, the homologous nucleotide sequences are derived from an infectious or pathogenic microorganism and most preferably from different strains or clades, variants, subtypes or serotypes of said microorganism. The terms "strain" or "clade" are technical terms, well known to the practitioner, referring to the taxonomy of microorganisms. The taxonomic system classifies all so far characterised microorganisms into the hierarchic order of Families, Genera, Species, Strains (Fields Virology, ed. by Fields B. N., Lippincott-Raven Publishers, 4th edition 2001). While the criteria for the members of Family is their phylogenetic relationship, a Genera comprises all members which share common characteristics, and a Species is defined as a polythetic class that constitutes a replicating lineage and occupies a particular ecological niche. The term "strain" or "clade" describes a microorganism, i. e. virus, which shares the common characteristics, like basic morphology or genome structure and organisation, but varies in biological properties, like host range, tissue tropism, geographic distribution, attenuation or pathogenicity. The term "variants" or "serotypes" further distinguishes between members of the same strain, also called subtypes, which show individual infection spectra or antigenic properties due to minor genomic variations.

According to a further embodiment the homologous nucleotide sequences are preferably selected from viruses. Representative examples of viruses include without limitation HIV (HIV-I or HIV-2), herpes viruses (e.g. HSVI or HSV2), cytomegalovirus (CMV), Epstein Barr virus (EBV), hepatitis viruses (e.g. hepatitis A virus (HAV), HBV, HCV and hepatitis E virus), flaviviruses (e.g. Yellow Fever Virus), varicella-zoster virus (VZV), paramyxoviruses, respiratory syncytial viruses (RSV), parainfluenza viruses, measles virus, influenza viruses, and papillomaviruses.

According to another embodiment, the homologous nucleotide sequences are selected from Dengue virus genes. Most preferred are homologous genes derived from different serotypes of the virus, wherein said genes may be derived from one, two, three or from all of the 4 Dengue virus serotypes.

In a preferred embodiment, the two homologous nucleotide sequences encode respiratory syncytial virus (RSV) genes. In a preferred embodiment, the homologous nucleotide sequences

encode RSV-F and/or RSV-G proteins. Preferably, one of the RSV genes is full-length and the other is truncated.

In another preferred embodiment, the two, preferably three homologous nucleotide sequences encode Ebola virus (EBOV) proteins. Three homologous nucleotide sequences encoding Ebola virus (EBOV) proteins do, of course, also cover two homologous nucleotide sequences. In a preferred embodiment, the homologous nucleotide sequences encode EBOV glycoproteins (GP). In a particular preferred embodiment, the nucleotide sequences encode glycoprotein precursor proteins from the EBOV strains EBOV-B (Bundibugyo), EBOV-S (Sudan ebolavirus strain Gulu) and EBOV-Z (Zaire ebola virus strain Mayinga).

In another embodiment, the homologous nucleotide sequences are selected from bacteria. Representative examples of suitable bacteria include without limitation *Neisseria* (e.g. *N. gonorrhea* and *N. meningitidis*); *Bordetella* (e.g. *B. pertussis*, *B. parapertussis* and *B. bronchiseptica*); *Mycobacteria* (e.g. *M. tuberculosis*, *M. bovis*, *M. leprae*, *M. avium*, *M. paratuberculosis*, *M. smegmatis*); *Legionella* (e.g. *L. pneumophila*); *Escherichia* (e.g. enterotoxigenic *E. coli*, enterohemorrhagic *E. coli*, enteropathogenic *E. coli*); *Shigella* (e.g. *S. sonnei*, *S. dysenteriae*, *S. flexnerii*); *Salmonella* (e.g. *S. typhi*, *S. paratyphi*, *S. choleraesuis*, *S. enteritidis*); *Listeria* (e.g. *L. monocytogenes*); *Helicobacter* (e.g. *H. pylori*); *Pseudomonas* (e.g. *P. aeruginosa*); *Staphylococcus* (e.g. *S. aureus*, *S. epidermidis*); *Enterococcus* (e.g. *E. faecalis*, *E. faecium*); *Bacillus* (e.g. *B. anthracis*); *Corynebacterium* (e.g. *C. diphtheriae*), and *Chlamydia* (e.g. *C. trachomatis*, *C. pneumoniae*, *C. psittaci*). Representative examples of parasites include without limitation *Plasmodium* (e.g. *P. falciparum*); *Toxoplasma* (e.g. *T. gondii*); *Leshmania* (e.g. *L. major*); *Pneumocystis* (e.g. *P. carinii*); and *Schistosoma* (e.g. *S. mansoni*). Representative examples of fungi include without limitation *Candida* (e.g. *C. albicans*) and *Aspergillus*.

The at least two nucleotide sequences can be of the same size or of different sizes. In a preferred embodiment, one of the two nucleotide sequences is truncated relative to the other. The truncation can be at the 5' or 3' end.

In various embodiments, the 300 nucleotides of the two nucleotide sequences encode 100 amino acids, which have at least 75%, 80%, 85%, 90%, 95%, 98%, 99%, or 100% amino acid identity. In a preferred embodiment, said amino acid identity is within a stretch of 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900 or 1000 or more contiguous amino acids.

In a particular preferred embodiment, the amino acids have at least 75%, 80%, 85%, 90%, 95%, 98%, 99%, or 100% amino acid identity within a stretch of at least 150 or 200 contiguous amino acids.

In other preferred embodiments, the proteins encoded by the two nucleotide sequences have at least 75%, 80%, 85%, 90%, 95%, 98%, 99%, or 100% amino acid identity within a stretch of 300 or 500 contiguous amino acids. In other preferred embodiments, the proteins encoded by the at least two nucleotide sequences have 85%-100%, in particular 100% amino acid identity within a stretch of 100, 200, 400, 600, or 800 contiguous amino acids in pairwise comparison.

As used herein, any term referring to "percent sequence identity", such as "amino acid identity" refers to the degree of identity between any given query sequence and a subject sequence.

Specifically, the following terms are used to describe the sequence relationships between two or more nucleic acids, polynucleotides or amino acid sequences: "reference sequence", "comparison window", "sequence identity", "percentage of sequence identity", and "substantial identity". A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence.

The terms "identical" or percent "identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same (e.g., 75% identity, 80% identity, 85% identity, 90% identity, 99%, or 100% identity in pairwise comparison), when compared and aligned for maximum correspondence over a comparison window, or designated region as measured using a sequence comparison algorithm or by manual alignment and visual inspection. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the results by 100 to yield the percentage of sequence identity.

The phrase "substantially identical ", in the context of two nucleic acids or polypeptides, refers to two or more sequences or subsequences that have at least about 85% identity, at least about 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% nucleotide or amino acid residue identity, when compared and aligned pairwise for maximum correspondence, as measured using a sequence comparison algorithm or by visual inspection. In an exemplary embodiment, the substantial identity exists over a region of the sequences that is at least about 50 residues in length. In another exemplary embodiment, the substantial identity

exists over a region of the sequences that is at least about 100 residues in length. In still another exemplary embodiment, the substantial identity exists over a region of the sequences that is at least about 150 residues or more, in length. In one exemplary embodiment, the sequences are substantially identical over the entire length of nucleic acid or protein sequence.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters.

A "comparison window", as used herein, includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 20 to 600, usually 20 to 50, about 50 to about 100, about 100 to about 200, more usually about 100 to about 150, or of about 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, or 3000 or even more in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned.

Percent identity can be determined using the alignment method of Needleman and Wunsch (J. Mol. Biol. 48; 443-453 (1970)) that has been shown to be equivalent to Sellers (SIAM J. of Applied Math 26; 787-793 (1974)). The percent identity may be determined, for example, by comparing sequence information using the GAP computer program, version 6.0 described by Devereux et al. (Nucl. Acids Res. 12:387, 1984) and available from the University of Wisconsin Genetics Computer Group (UWGCG), which utilizes this alignment method. The preferred default parameters for the GAP program include: (1) a unary comparison matrix (containing a value of 1 for identities and 0 for non-identities) for nucleotides, and the weighted comparison matrix of Gribskov and Burgess, Nucl Acids Res. 14:6745, 1986, as described by Schwartz and Dayhoff, eds., Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, pp. 353-358, 1979; (2) a penalty of 3.0 for each gap and an additional 0.10 penalty for each symbol in each gap; and (3) no penalty for end gaps. An other suitable tool is to use the ContigExpress from the VectorNTI Advance program (INVITROGEN), e.g. version 10.3.1 from 2007.

According to the present techniques, the degeneracy of the genetic code is used to make homologous or identical nucleotide sequences less homologous in order to prevent intramolecular recombination. Said differences may already be included in the nucleotide sequences by nature and/or are included artificially by substitutions. In various embodiments, the number of different nucleotides originating from nature plus from artificial substitution is at least 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, or 500. Preferably, the number of different bases is at least 75, 200 or 450. The number of differences does, of course, vary and increase, respectively, with the number of nucleotides of the nucleotide sequences.

In a preferred embodiment, at least 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, or 500 nucleotides are substituted. Said substitutions are artificially introduced independently of already present numbers of different nucleotides included, for example, by silent mutations.

In various embodiments, two nucleotide sequences with stretches of identity of no more than 13, 12, 11, 10, 9, 8, 7, 6, 5, or 4 contiguous nucleotides after substitution are preferred. In case of more than two nucleotide sequences, stretches of identity of no more than 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5 or 4 contiguous nucleotides after substitution are preferred.

In another embodiment, the nucleotide sequences can have at least 75, 100, 150, 200, 250, 300, 350, 400, or 450 nucleotides substituted out of 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, or 1600 or more nucleotides.

In the context of this disclosure, substitution of nucleotides with different nucleotides means the technical or artificial replacement of nucleotides by other nucleotides. Preferably, the substituted nucleotides do not alter the encoded amino acid sequence. Substitution can be performed by identifying codons in the two homologous nucleotide sequences encoding the same amino acids and altering codons in one of the two homologous nucleotide sequences such that the codons still encodes the same amino acids. The alterations can be made in one, both or all of the homologous nucleotide sequences.

For example the amino acid proline is encoded by the codons CCA, CCC, CCG and CCU (on the DNA level the U is replaced by a T). A simple nucleotide sequence, CCCCCC, initially encoding two prolines in two homologous nucleotide sequences could be changed to CCACCG, also encoding two prolines, in one of the two homologous nucleotide sequences. Alternatively, one of the sequences encoding proline-proline could be changed to CCCCCG, and the other to CCACCC.

A more complicated example is the amino acid serine, which is encoded by UCA, UCC, UCG, UCU, AGC and AGU. Similarly, UCAUCA, initially encoding two different serines could be changed in multiple homologous sequences, to AGCAGC (sharing no common nucleotide with UCAUCA) and UCGAGU (sharing only one position with UCAUCA or two position with AGCAGC) and so on. This allows a higher flexibility in introducing different nucleotide variants into two or more nucleotide sequences encoding a serine-serine.

Preferably codon optimization as described herein avoids the use of rare codons for a desired host since rare codons may block or reduce expression of the encoded protein. Also, substitutions that may introduce nucleic acid signals for the desired host are preferably avoided. Such signals include, but are not limited to, splice signals, termination signals, and initiation signals. Preferably, the following sequence motifs may be avoided depending on the type of vector used, e.g., the vaccinia virus early transcription termination signal needs not to be avoided in many other vectors, being no poxvirus vectors:

- internal TATA-boxes, chi-sites, and ribosomal entry sites;
- AT-rich and GC-rich sequence stretches;
- ARE, INS, and CRS sequence elements;
- repeat sequences and RNA secondary structures;
- (cryptic) splice donor and acceptor sites, and branch points; and
- vaccinia early transcription termination signals: (TTTTTNT).

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is more fully understood with reference to the drawings, in which:

Figure 1 depicts an alignment of the nucleotide sequence encoding the full-length RSV-F (F) protein with the nucleotide sequence encoding the substituted, truncated RSV-F_{trunc} (F_{trunc}) protein. The identical sequences are highlighted in black, and the substituted nucleotides remain unhighlighted. The locations of primers A1 and B2 are indicated.

Figure 2 depicts an alignment of the full-length RSV-F (F) protein with the truncated RSV-F_{trunc} (F_{trunc}) protein. The full length sequence of RSV-F is truncated by 50 aa to result in the truncated RSV-F_{trunc} protein. The RSV-F_{trunc} protein covers approximately 91% of the full length protein.

Figure 3 depicts expression of RSV-F and RSV-F_{trunc} from recombinant MVA-BN® viruses in a human cell line. Western blot with extracts from infected human cells upon infection with different MVA-BN® based viruses with an MOI of 10 and lysis at 24h post infection. MVA-BN® (empty vector control; lane 1), MVA-mBN172B (recombinant MVA-BN® with full length RSV-F; lane 2), MVA-mBN173B (recombinant MVA-BN® with truncated RSV-F_{trunc}; lane 3) and lane 4: MVA-mBN175B (recombinant MVA-BN® with RSV-F and RSV-F_{trunc}). The calculated molecular weight of the proteins is: RSV-F (61.6 kDa) and RSV-F_{trunc} (56.1 kDa).

Figures 4A-C depict PCR analysis of MVA-mBN175B. RSV-F (F) and RSV-F_{trunc} (F_{trunc}) are shown. A. PCR results with various primer pairs. M=markers (1 kb-ladder, New England Biolabs). Lane 1 is MVA-mBN175B. Lane 2 is a positive control plasmid (pBN345). Lane 3 is MVA-mBN®. Lane 4 is a water control. Lane 5 is a positive control plasmid (pBN343). B. Schematic of MVA-mBN175B showing locations of primers used for the PCRs shown in Figure 4A. C. Schematic of wild type MVA-mBN® showing locations of primers.

Figures 5A-C depict the hypothetical recombination F/F_{trunc} between the full length RSV-F gene (F) and the truncated F gene (F_{trunc}) in the double recombinant MVA and the locations of the PCR primers in the recombinant and non-recombinant viruses and control plasmids. A. MVA-mBN175B. B. pMISC173. C. pMISC172.

Figure 6 depicts PCR analysis of DNA isolated from cells infected with MVA-mBN175B. Lanes 1 and 7 are marker lanes. Lane 2 is MVA-mBN175B. Lane 3 is a plasmid control for the F gene (pBN343). Lane 4 is a plasmid control for the truncated F gene (pBN345). Lane 5 is MVA-BN®. Lane 6 is a water control. The expected PCR product from a hypothetical recombination

between the RSV-F gene and truncated F gene RSV-F_trunc in MVA-mBN175B is 613 base pairs.

Figure 7 depicts an alignment of three EBOV (ebolavirus) GP (glycoprotein) protein sequences. The amino acid sequences of three GP proteins of the ebola virus strains EBOV-B, EBOV-S and EBOV-Z are aligned. No gaps were allowed in the alignment. The overall identity in all three protein sequences is 48.5%. Gray background: identical in all three protein sequences. Black background: identical in two proteins.

Figures 8A and 8B depict an alignment of three EBOV GP coding sequences used in the recombinant MVA-BN® based construct. The coding sequences for the GP genes originating from three EBOV strains EBOV-B, -S and -Z were aligned before (non-opt; see Figure 8A) and after (opt; see Figure 8B) optimization. No gaps were allowed in the alignment. Gray background: identical nucleotide positions in three coding sequences. Black background: identical nucleotide positions in two coding sequences. The identity in nucleotide positions of three genes prior optimization (non-opt) is 45.3%, while after optimization (opt) it is 44.6%.

Figure 9 depicts pairwise alignments of three EBOV GP coding sequences used in the recombinant MVA-BN® based construct. The coding sequences for the GP genes originating from three EBOV strains EBOV-B, -S and -Z were aligned pairwise before (non-opt; see Figure 9A) and after (opt; see Figure 9B) optimization. No Gaps were allowed in the alignments. Gray background: identical nucleotide positions in the coding sequence. The identity in nucleotide positions of three genes prior (non-opt) and after (opt) optimization is tabulated in Table C.

Figure 10 depicts a restriction enzyme digest and plasmid map of plasmid pMISC210 comprising the full-length (RSV-F) and truncated (RSV-F_trunc) protein. Lane 1: plasmid pMISC210 comprising RSV-F and RSV-F_trunc; Lane 2: control plasmid pMISC209 comprising RSV-F_trunc only; Lane 3: Molecular weight marker. The size of the marker-bands in base pairs (bp) is shown.

EXAMPLES

Example 1

Preparation of Substituted, Truncated F Gene

Creation of a recombinant MVA expressing both a full-length RSV-F protein and a truncated Version RSV-F_trunc was desired. However, based on results with MVA and other vaccinia viruses containing repeat sequences, it was expected that intramolecular recombination would lead to recombination between the two copies of the F gene, resulting in deletion of one of the copies of the F gene.

To minimize the presence of long stretches of identical nucleotides between the two F genes, the codons in the nucleotide sequence encoding the RSV-F_trunc gene were substituted, while maintaining the amino acid sequence of the F genes. The use of rare codons for mammals and chickens was avoided. Also, substitutions that might introduce nucleic acid signals were avoided. Such signals included internal TATA-boxes, chi-sites, and ribosomal entry sites; AT-rich and GC-rich sequence stretches; ARE, INS, and CRS sequence elements; repeat sequences and RNA secondary structures; (cryptic) splice donor and acceptor sites, and branch points; and vaccinia termination signals (TTTTTNT). The substituted nucleotide sequence is shown in Fig. 1, compared to a coding sequence for a full-length RSV-F protein. Although significant identity remains throughout the two coding sequences, there are no remaining large stretches of identity greater than nine contiguous nucleotides within the two coding sequences. The proteins encoded by the two coding sequences are aligned in Fig. 2. The two proteins have 100% identity over the first 524 amino acids (the substituted F protein is truncated at the carboxy terminus). Thus, although these two coding nucleotide sequences encode a stretch of identical amino acids, one of the sequences has been substituted relative to the other.

Example 2

Preparation of Recombinant Viruses comprising RSV-F Genes

The DNA encoding the full-length RSV-F gene was inserted into MVA at two different integration sites to generate MVA-mBN170B and MVA-mBN172B (in the IGR88/89 site). The substituted, RSV-F_trunc gene was inserted into MVA at the IGR148/149 site to generate MVA-mBN173B.

A double recombinant MVA was then created containing the full-length RSV-F gene inserted into MVA at the IGR88/89 site and the substituted, RSV-F_trunc gene inserted into the same MVA at

the IGR148/149 site. The double recombinant virus was called MVA-mBN175B. A schematic of this virus is shown in Fig. 4B.

Example 3

Expression of F Proteins from Recombinant Viruses To determine whether protein was expressed from the substituted nucleotide sequence, western blot analysis was performed on protein extracts from a human cell line infected with a recombinant MVA-BN®-based virus encoding the full-length RSV-F gene (MVA-mBN172B), the virus encoding the substituted, RSV-F_trunc gene (MVA-mBN173B) and a double recombinant virus encoding both, the full length and the RSV-F_trunc gene (MVA-mBN175B). All three viruses showed the production of the appropriately sized RSV-F proteins by Western blot analysis (Fig. 3), while the MVA-BN® control (empty vector) did not show any bands, as expected. Thus, the full length and the truncated F protein expressed from the substituted coding nucleotide sequence were expressed individually from single recombinant MVA-BN® but both were also co-expressed from one double recombinant MVA-BN® virus (MVA-mBN175B) in a human cell line.

Example 4

Growth of Recombinant Viruses

Chicken embryo fibroblast cells were infected with MVA-mBN175B, a construct containing both the full-length F gene and the substituted, RSV-F_trunc gene, or a construct containing only the full-length F gene to receive a first virus crude stock. Similar titers of the double recombinant virus containing both full length F and truncated F genes (1.34×10^7 TCID₅₀) were seen in comparison with titers of the virus containing only the full length F gene (1.46×10^7 TCID₅₀). These results indicated that a stable double recombinant MVA was being produced, and that recombination between the two copies of the F gene had been limited by substituting nucleotide bases in the sequences.

Example 5

PCR Analysis of Recombinant Viruses

PCR analysis was performed on DNA from cells infected with MVA-mBN175B or MVA-BN® using the insert-specific and flank-specific primer pairs depicted in Fig. 4B and C. PCR A with primers A1/A2, which are specific for the full-length F gene, detected a band with the size of 663 base pairs (bp) in cells infected with MVA-mBN175B and in a specific plasmid positive control as expected. This band as expected is absent in cells infected with MVA-BN® or in the water

control (Fig. 4A). PCR B with primers B1/B2, which are specific for the substituted, truncated F gene, detected a band with the size of 625 bp in cells infected with MVA-mBN175B and in a specific plasmid positive control as expected. This band, as expected, is absent in cells infected with MVA-BN® or in the water control (Fig. 4A). PCR C with primers C1/C2, which detect insertions into the IGR88/89 site, detected a band with the size of 2047 bp in cells infected with MVA-mBN175B and in a specific plasmid positive control as expected. This band, as expected, is absent in cells infected with the empty vector control MVA-BN®, instead a band of 161 bp indicates the wildtype situation at IGR88/89 in MVA-BN® (Fig. 4A). PCR D with primers D1/D2, which detect insertions into the IGR148/149 site, detected a band with the size of 2062 bp in cells infected with MVA-mBN175B and in a specific plasmid positive control as expected. This band as expected is absent in cells infected with the empty vector control MVA-BN®, instead a band of 360 bp indicates the wildtype situation at IGR88/89 in MVA-BN®. (Fig. 4A).

Recombination between the F genes would yield a hybrid F gene having parts of the wild-type F gene and parts of the truncated F gene. (Fig. 5A.) To detect the presence of any such recombinants, PCR analysis was performed on DNA from cells infected with MVA-mBN175B or MVA-BN® using the primer pairs A1/B2 (Fig. 5B.), which should generate a 613 base pair product, specific for the recombinant F gene. The results of this PCR showed no detectable recombinants. (Fig. 6.) These results indicated that a stable double recombinant MVA was being produced, and that recombination between the two copies of the F gene had been limited.

Example 6

Preparation of recombinant glycoprotein (GP) genes of three different ebolavirus (EBOV) strains

Generation of a recombinant MVA expressing three ebolavirus (EBOV) glycoproteins (GP) was desired. The EBOV strains used herein are EBOV-B (Bundibugyo), EBOV-S (Sudan) and EBOV-Z (Zaire), all belonging to virus strains with high lethality in infected humans. Said three GP share an overall identity of 48.5%, indicating that nearly every second amino acid in the GP proteins is identical in all three strains, while the percent identities over the full-length protein sequences in comparison of combinations of two strains are between 57.0 % and 64.2% (Figure 7).

To minimize the presence of long stretches of identical nucleotides within the three EBOV GP genes, the codons in the three nucleotide sequences were substituted, while maintaining the encoded amino acid sequences of the three GP genes. The use of rare codons for mammals and chickens, as well as substitutions that might introduce nucleic acid signals were avoided.

Such signals included internal TATA-boxes, chi-sites, and ribosomal entry sites; AT-rich and GC-rich sequence stretches; ARE, INS, and CRS sequence elements; repeat sequences and RNA secondary structures; (cryptic) splice donor and acceptor sites, and branch points; and vaccinia termination signals (TTTTTNT). The G after the ATG start codon allows for high expression and is present in the original coding sequence of all three EBOV GP genes and was maintained.

Although 23.3 to 24.9% of the nucleotides in each of the 3 optimized EBOV GP coding sequences were exchanged (see Table A), the overall identities did not dramatically change between the three GP coding sequences (Table B). In two cases, the pair wise comparisons even showed marginally higher identities after optimization of the coding sequences, as shown below in Table B.

Table A: Nucleotide exchanges in three optimized EBOV GP genes. The table shows the number of changed nucleotides at the corresponding positions in the optimized GP coding sequences (opt) compared to the non-optimized (non-opt) sequence of different EBOV strains based on the total number of nucleotides in [%]. The total number of nt is 1147.

	exchanged nt positions in optimized GP coding sequences compared to non-optimized sequences [%]
EBOV-B non-opt : EBOV-B opt	23.3
EBOV-S non-opt : EBOV-S opt	24.9
EBOV-Z non-opt : EBOV-Z opt	23.9

Table B: Identical nucleotide positions of three EBOV GP coding sequences. The table shows the number of identical nucleotides at the corresponding positions in two GP coding sequences of different EBOV strains based on the total number of nucleotides in [%].

pairwise comparison of GP genes	identity of nucleotides in non-optimized genes [%]	identity of nucleotides in optimized genes [%]
EBOV-B : EBOV-S	57.0	57.3
EBOV-B : EBOV-Z	64.2	61.1

EBOV-S : EBOV-Z	57.6	60.4
-----------------	------	------

Pairwise alignments of the GP coding sequences of three EBOV strains EBOV-B, -S and -Z showed the identities in nucleotide positions and the distribution of identities (Figure 9). Consequently, the disclosed method led to shorter stretches of nucleotide identity in the EBOV GP-sequences. When considering long stretches of identical consecutive nucleotides, it is evident that the interruption or shortening of such stretches of identities is an important part of the strategy to avoid recombination between sequences sharing a certain degree of nucleotide identities. In Table C (see below) the number of stretches of consecutive identical nucleotides from pair wise comparison of the GP coding sequences are shown. Prior to optimization, there are stretches of up to 23 bp length and in summary there are 41 stretches of 10 or more identical nucleotides. In the optimized version of the GP genes, only one 13 bp stretch is found and 7 stretches of 10 or more identical nucleotides can be found.

Table C: Long stretches of consecutive identical nucleotides. The table shows the number of stretches of consecutive identical nucleotides of a certain length in pair wise comparison of EBOV GP coding sequences before (non-opt) and after (opt) optimization. The numbers of the pairwise comparisons are summarized in the column 'combined numbers'. The longest stretch in the non-optimized comparisons are 23 consecutive identical nucleotides, while in the optimized genes, it is reduced to a maximum of 13 nucleotides. Only stretches of 10 or more nucleotides are listed.

length	EBOV-B : EBOV-S		EBOV-B : EBOV-Z		EBOV-S : EBOV-Z		combined numbers	
	non-opt	opt	non-opt	opt	non-opt	Opt	non-opt	opt
23 nt			1				1	
20 nt			2				2	
17 nt					1		1	
16 nt			2				2	
14 nt			2		2		4	
13 nt			1	1	1		2	1
12 nt	1		2				3	
11 nt	10	2	4	1	8		22	3
10 nt	1		2	1	1	2	4	3

Example 7

Preparation of recombinant MVA-BN® viruses with GP genes of EBOV strains.

The three EBOV GP genes were synthesized by GeneArt (Regensburg, Germany) and cloned into recombination vectors to allow for integration into MVA-BN®. A recombinant virus comprising the three optimized homologous GP gene sequences from three different EBOV strains was generated. The transcription of the three inserted GP coding sequences is controlled by different individual early-late promoters.

Specific PCR reactions for the three optimized EBOV-GP sequences showed the presence of the three individual genes in the recombinant MVA-BN®.

Example 8

Preparation of Plasmid comprising RSV-F Genes

The two versions of the RSV-F gene used in examples 1-5 and shown in Figure 1 were cloned into one plasmid and maintained in E.coli TZ101 (Trenzyme GmbH, Konstanz, Germany) using standard cloning techniques. The plasmid (see plasmid map in Figure 10) was isolated and digested with the restriction enzymes *Ale* I, *Dra* III and *Spe* I and separated on a 1% TAE agarose gel (see Figure 10). The band patterns for pMISC210 encoding the full-length RSV-F protein and RSV-F_trunc protein (lane 1) as well as the control plasmid pMISC209 encoding the RSV-F_trunc protein only (lane 2) were compared with the patterns expected from the results of analysis of the electronic sequence of the plasmids. The expected size of bands for pMISC210 was 404, 573, 809, 1923 and 4874 bp, while for pMISC209 a pattern of bands with sizes of 573, 661, 809 and 4874 bp was expected. All expected bands and no additional bands were found experimentally. In case recombination between the RSV-F variants in pMISC210 occurred, one or more of the smaller fragments would be lost, depending on the sites of recombination. This was clearly not found in the current example. Thus, the results show the stability of the plasmid pMISC210 with the two RSV-F genes (RSV-F and RSV-F_trunc) in E.coli.

Claims:

1. A modified vaccinia Ankara (MVA) virus vector comprising two nucleotide sequences of 300 nucleotides in size, each coding for 100 amino acids, wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity; and wherein one of the two nucleotide sequences has at least 75 nucleotides different from the other nucleotide sequence, wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences, wherein the two divergent sequences share stretches of identity of no more than 10 contiguous nucleotides.

2. A method for generating a modified vaccinia Ankara (MVA) virus vector according to claim 1, said method comprising the steps of: a) providing a first nucleotide sequence of 300 nucleotides in size coding for 100 amino acids; b) providing a second nucleotide sequence of 300 nucleotides in size coding for 100 amino acids, wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity and wherein one of the two nucleotide sequences has at least 75 nucleotides different from the other nucleotide sequence, wherein the at least 75 different nucleotides were substituted and wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences; and c) inserting the two divergent nucleotide sequences into a vector, wherein the two divergent sequences share stretches of identity of no more than 10 contiguous nucleotides.

3. A method for reducing intramolecular recombination within a modified vaccinia Ankara (MVA) virus vector containing two nucleotide sequences of 300 nucleotides in size, each coding for 100 amino acids, wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity, said method comprising substituting nucleotides in one or both nucleotide sequence(s) to generate two divergent sequences which show differences in at least 75 nucleotides, wherein the substituted nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences, wherein the two divergent sequences share stretches of identity of no more than 10 contiguous nucleotides.

4. A method for generating a modified vaccinia Ankara (MVA) virus, comprising two homologous nucleotide sequences, said method comprising the steps of: a) providing a virus comprising a nucleotide sequence of 300 nucleotides in size coding for 100 amino acids; and b) inserting a second nucleotide sequence of 300 nucleotides in size coding for 100 amino acids into the virus; wherein the 100 amino acids encoded by each of the two nucleotide sequences have at least 75% amino acid identity; and wherein one of the two nucleotide sequences has at least 75

nucleotides different from the other nucleotide sequence, wherein the at least 75 different nucleotides were substituted, wherein the different nucleotides do not alter the identical amino acids encoded by said two nucleotide sequences, wherein the two divergent sequences share stretches of identity of no more than 10 contiguous nucleotides.

5. The vector according to claims 1, wherein the two nucleotide sequences are two respiratory syncytial virus (RSV) genes or two Filovirus genes.

6. The vector according to claim 5, wherein the two respiratory syncytial virus (RSV) genes are RSV-F; RSV-G; or RSV-F and RSV-G genes.

7. The vector according to claim 6, wherein the two nucleotide sequences encode a full length and a truncated RSV-F protein, wherein the two divergent sequences share stretches of identity of no more than 9 contiguous nucleotides.

8. The vector according to claim 5, wherein the two Filovirus genes encode Ebola virus Sudan and Ebola virus Zaire glycoproteins.

9. The vector of claim 1, wherein the two divergent sequences comprise the nucleotide sequences of SEQ ID NO:1 and SEQ ID NO:2 or SEQ ID NO:12 and SEQ ID NO:13.

10. The method according to claim 2, wherein the two nucleotide sequences are two respiratory syncytial virus (RSV) genes or two Filovirus genes.

11. The method according to claim 2, wherein the two nucleotide sequences encode a full length and a truncated RSV-F protein, wherein the two divergent sequences share stretches of identity of no more than 9 contiguous nucleotides.

12. The method according to claim 10, wherein the two Filovirus genes encode Ebola virus Sudan and Ebola virus Zaire glycoproteins.

13. The method according to claim 3, wherein the two nucleotide sequences are two respiratory syncytial virus (RSV) genes or two Filovirus genes.

14. The method according to claim 3, wherein the two nucleotide sequences encode a full length and a truncated RSV-F protein, wherein the two divergent sequences share stretches of identity of no more than 9 contiguous nucleotides.

15. The method according to claim 13, wherein the two Filovirus genes encode Ebola virus Sudan and Ebola virus Zaire glycoproteins.