(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2023 (26.01.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/004125 A2

(51) International Patent Classification:
C12N 15/86 (2006.01) *C12P 21/02* (2006.01)

(21) International Application Number:
PCT/US2022/038032

(22) International Filing Date:
22 July 2022 (22.07.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/225,212 23 July 2021 (23.07.2021) US
63/256,819 18 October 2021 (18.10.2021) US

(71) Applicant: **UNIVERSITY OF WASHINGTON** [US/US]; 4545 Roosevelt Way NE, Suite 400, Seattle, Washington 98105 (US).

(72) Inventors: **CHAMBERLAIN, Jeffrey S.**; c/o University of Washington, 4545 Roosevelt Way NE, Suite 400, Seattle, Washington 98105 (US). **TASFAOUT, Hichem**; c/o University of Washington, 4545 Roosevelt Way NE, Suite 400, Seattle, Washington 98105 (US).

(74) Agent: **FITZGERALD, Mark J.** et al.; Nixon Peabody LLP, 53 State Street, Exchange Place, Boston, Massachusetts 02109 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) **Title:** GENERATION OF LARGE PROTEINS BY CO-DELIVERY OF MULTIPLE VECTORS

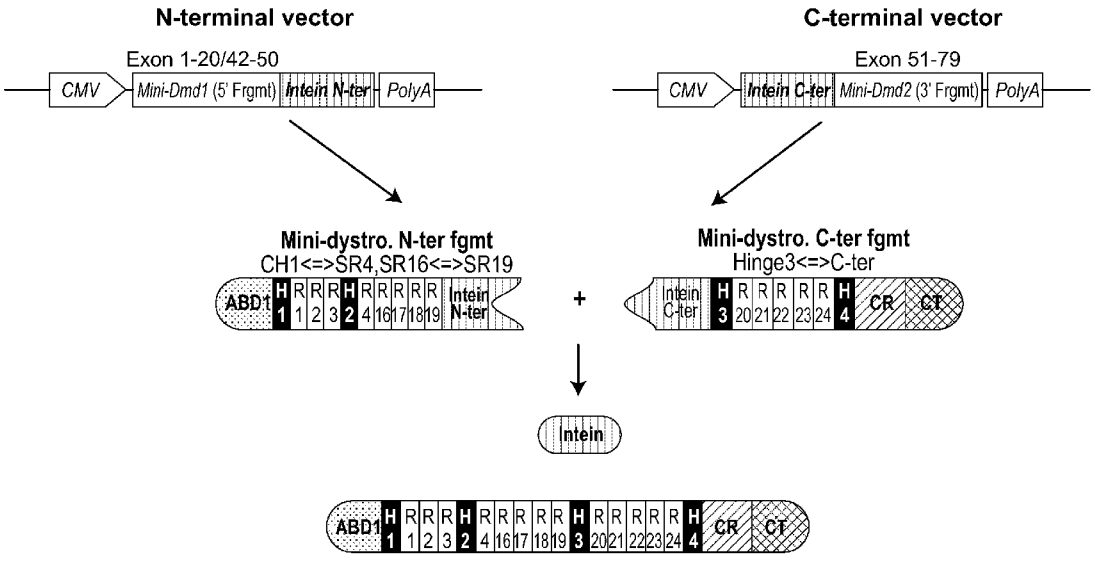


FIG. 5A

(57) **Abstract:** Provided herein are methods and compositions for delivering large proteins to a subject in need thereof for treatment of a disease or disorder. In certain embodiments, the methods and compositions described herein are useful in the delivery of large proteins to subjects using a protein expression system comprising at first and second AAV vector for the treatment of muscular or neuromuscular disease or disorders.

[Continued on next page]

WO 2023/004125 A2

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

GENERATION OF LARGE PROTEINS BY CO-DELIVERY OF MULTIPLE VECTORS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims benefit under 35 U.S.C. § 119(e) to U.S. Provisional Application No. 63/225,212 filed July 23, 2021 and U.S. Provisional Application No. 63/256,819 filed October 18, 2021, the contents of each which are incorporated herein by reference in their entirety.

TECHNICAL FIELD

[0002] The field of the invention relates to methods of delivering or inducing the production of large therapeutic proteins using multiple vectors.

BACKGROUND

[0003] Gene therapies using AAV vectors hold a promising future for treating different loss-of-function genetic disorders (Li 2020). However, this therapeutic modality has been challenged by the small packaging capacity of this viral vector (~5 kb).

SUMMARY

[0004] The methods and compositions described herein are based, in part, on the discovery that split inteins can permit the delivery of large polypeptides, including but not limited to dystrophin, using AAV vectors.

[0005] In one aspect, described herein is a method for delivering an exogenous polypeptide to a cell, the method comprising contacting the cell with: a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a split intein; and a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to a second portion of the split intein; wherein the first and second fusion polypeptides are produced in the cell from the first and second nucleic acids, and wherein the first and second portions of the split intein promote joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, thereby delivering the exogenous polypeptide to the cell; wherein the exogenous polypeptide delivered is larger than can be encoded by a single AAV vector particle.

[0006] In one embodiment of this and all other aspects described herein, the first and second nucleic acids comprise a muscle-specific expression cassette (MSEC).

[0007] In another embodiment of this and all other aspects described herein, the split intein is a naturally-occurring split intein.

[0008] In another embodiment of this and all other aspects described herein, the split intein is a genetically modified split intein. In another embodiment, the genetic modification of the split intein is selected from codon optimization for expression and/or stability in mammalian cells, shortening or lengthening of the

split intein, or changing encoded amino acids in the split intein to more closely match the sequence of the exogenous protein to be delivered.

[0009] In another embodiment of this and all other aspects described herein, the first and second portions of the exogenous polypeptide are substantially the same size.

[0010] In another embodiment of this and all other aspects described herein, the first and second portions of the exogenous polypeptide differ in size by no more than 50 amino acids.

[0011] In another embodiment of this and all other aspects described herein, the exogenous polypeptide comprises a footprint of less than four amino acids from the split intein.

[0012] In another embodiment of this and all other aspects described herein, the exogenous polypeptide comprises a footprint of 3 or fewer amino acids from the split intein.

[0013] In another embodiment of this and all other aspects described herein, the split site separating the first and second portions of the exogenous polypeptide is selected at a site having the same sequence as the split intein footprint, thereby producing the exogenous polypeptide without extra amino acids from the split intein.

[0014] In another embodiment of this and all other aspects described herein, the exogenous polypeptide is a therapeutic polypeptide.

[0015] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is selected from dystrophin, mini-dystrophin, utrophin and dysferlin, nebulin, titin, myosin, spectrin repeat containing nuclear envelope protein 1 (Syne-1), dystroglycan, ATP synthase, clotting factor IIX, lamin A/C, thyroglobulin, epidermal growth factor receptor (EGFR), alpha- and/or beta spectrin, muscle target of rapamycin (mTOR), and ryanodine receptor 1. In another embodiment, the mini-dystrophin is greater than 160kDa and smaller than full-length dystrophin.

[0016] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to the N-terminal portion of a split intein within or adjacent to a dystrophin hinge domain.

[0017] In another embodiment of this and all other aspects described herein, the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

[0018] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

[0019] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is dystrophin and the C-terminal portion of the dystrophin extein is joined to the C-terminal portion of the split intein within or adjacent to a dystrophin hinge domain or to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains. In another embodiment of this and all other aspects described herein, the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

[0020] In another embodiment of this and all other aspects described herein, the exogenous polypeptide is functional in the cell.

[0021] In another aspect, described herein is a method for delivering an exogenous polypeptide to a cell, the method comprising contacting the cell with: a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a first split intein, wherein the first portion of the split intein is fused to the carboxy terminus of the first portion of the exogenous polypeptide; a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to (i) a second portion of the first split intein at the amino terminus of the second portion of the exogenous polypeptide and (ii) a first portion of a second split intein at the carboxy terminus of the second portion of the exogenous polypeptide; and a third AAV vector particle comprising a third nucleic encoding a third fusion polypeptide comprising a third portion of the exogenous polypeptide fused to a second portion of the second split intein at the amino terminus of the third portion of the exogenous polypeptide, wherein the first, second, and third fusion polypeptides are produced in the cell from the first, second and third nucleic acids, and wherein the respective portions of the first and second split inteins promote joining of (a) the carboxy terminus of the first portion of the exogenous polypeptide to the amino terminus of the second portion of the exogenous polypeptide and (b) the carboxy terminus of the second portion of the exogenous polypeptide to the amino terminus of the third portion of the exogenous polypeptide, thereby delivering the exogenous polypeptide to the cell; wherein the exogenous polypeptide delivered is larger than can be encoded by a single AAV vector particle. In one embodiment of this and all other aspects described herein, the first and second split inteins do not cross-splice.

[0022] In another aspect, described herein is a protein expression system comprising a set of AAV vector particles comprising a first and second AAV particle, wherein the first AAV vector particle comprises a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a split intein; and wherein the second AAV vector particle comprises a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to a second portion of the split intein.

[0023] In one embodiment of this and all other aspects described herein, co-infection of a cell with the first and second AAV vector particles promotes joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the split intein.

[0024] In another embodiment of this and all other aspects described herein, joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the split intein generates an exogenous polypeptide larger than can be encoded in a single AAV particle.

[0025] In another aspect, described herein is a protein expression system comprising a set of AAV vector particles comprising a first, second, and third AAV particle, wherein the first AAV vector particle comprises a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a first split intein, wherein the first portion of the split intein is fused to the carboxy terminus of the first portion of the exogenous polypeptide; wherein the second AAV vector particle comprises a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to (i) a second portion of the first split intein at the amino terminus of the second portion of the exogenous polypeptide and (ii) a first portion of a second split intein at the carboxy terminus of the second portion of the exogenous polypeptide; and wherein the third AAV vector particle comprises a third nucleic encoding a third fusion polypeptide comprising a third portion of the exogenous polypeptide fused to a second portion of the second split intein at the amino terminus of the third portion of the exogenous polypeptide.

[0026] In one embodiment of this and all other aspects described herein, co-infection of a cell with the first, second and third AAV vector particles promotes joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the first split intein, and joining of the second portion of the exogenous polypeptide to the third portion of the exogenous polypeptide, with removal of the first and second portions of the second split intein.

[0027] In another embodiment of this and all other aspects described herein, joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the first split intein, and joining of the second portion of the exogenous polypeptide to the third portion of the exogenous polypeptide, with removal of the first and second portions of the second split intein generates an exogenous polypeptide larger than can be encoded in a single AAV particle.

[0028] In another embodiment of this and all other aspects described herein, expression of the first and second, or first, second and third fusion polypeptides is driven by a muscle-specific expression cassette.

[0029] In another aspect, described herein is a method of treating a disease or disorder in a subject in need thereof, the method comprising administering a protein expression system as described herein, thereby treating the subject.

[0030] In one embodiment of this and all other aspects described herein, the subject in need thereof has a muscular or neuromuscular disease or disorder.

[0031] In another embodiment, of this and all other aspects described herein, the exogenous polypeptide is dystrophin or mini-dystrophin and the subject in need thereof has Duchenne muscular dystrophy (DMD) or Becker muscular dystrophy (BMD).

[0032] In another embodiment, of this and all other aspects described herein, the dystrophin or mini-dystrophin increases the strength of dystrophic muscles by at least 10%.

[0033] In another embodiment, of this and all other aspects described herein, expression of the first and second, or first, second and third fusion polypeptides is driven by a muscle-specific expression cassette.

[0034] In another embodiment, of this and all other aspects described herein, the protein expression system is administered by infusion into the vasculature, or by direct injection into a tissue.

[0035] In another aspect, described herein is a method for inducing the production of an exogenous polypeptide in a cell, the method comprising contacting the cell with: a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a split intein; and a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to a second portion of the split intein; wherein the first and second fusion polypeptides are produced in the cell from the first and second nucleic acids, and wherein the first and second portions of the split intein promote joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, thereby inducing the production of the exogenous polypeptide in the cell; wherein the exogenous polypeptide produced is larger than can be encoded by a single AAV vector particle.

[0036] In one embodiment of this and all other aspects described herein, the first and second nucleic acids comprise a muscle-specific expression cassette (MSEC).

[0037] In another embodiment of this and all other aspects described herein, the split intein is a naturally-occurring split intein.

[0038] In another embodiment of this and all other aspects described herein, the split intein is a genetically modified split intein. In another embodiment, the genetic modification of the split intein is selected from codon optimization for expression and/or stability in mammalian cells, shortening or lengthening of the split intein, or changing encoded amino acids in the split intein to more closely match the sequence of the exogenous protein to be produced.

[0039] In another embodiment of this and all other aspects described herein, the first and second portions of the exogenous polypeptide are substantially the same size.

[0040] In another embodiment of this and all other aspects described herein, the first and second portions of the exogenous polypeptide differ in size by no more than 50 amino acids.

[0041] In another embodiment of this and all other aspects described herein, the exogenous polypeptide comprises a footprint of less than four amino acids from the split intein.

[0042] In another embodiment of this and all other aspects described herein, the exogenous polypeptide comprises a split intein footprint of 3 or fewer amino acids.

[0043] In another embodiment of this and all other aspects described herein, the split site separating the first and second portions of the exogenous polypeptide is selected at a site having the same sequence as the split intein footprint, thereby producing the exogenous polypeptide without extra amino acids from the split intein.

[0044] In another embodiment of this and all other aspects described herein, the exogenous polypeptide is a therapeutic polypeptide.

[0045] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is selected from dystrophin, mini-dystrophin, utrophin and dysferlin, nebulin, titin, myosin, spectrin repeat containing nuclear envelope protein 1 (Syne-1), dystroglycan, ATP synthase, clotting factor IIX, lamin A/C, thyroglobulin, epidermal growth factor receptor (EGFR), alpha- and/or beta spectrin, muscle target of rapamycin (mTOR), and ryanodine receptor 1.

[0046] In another embodiment of this and all other aspects described herein, the mini-dystrophin is greater than 160kDa and smaller than full-length dystrophin.

[0047] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to the N-terminal portion of a split intein within or adjacent to a dystrophin hinge domain.

[0048] In another embodiment of this and all other aspects described herein, the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

[0049] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

[0050] In another embodiment of this and all other aspects described herein, the therapeutic polypeptide is dystrophin and the C-terminal portion of the dystrophin extein is joined to the C-terminal portion of the split intein within or adjacent to a dystrophin hinge domain or to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

[0051] In another embodiment of this and all other aspects described herein, the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

[0052] In another embodiment of this and all other aspects described herein, the exogenous polypeptide is functional in the cell.

[0053] In another aspect, described herein is a method for inducing the production of an exogenous polypeptide in a cell, the method comprising contacting the cell with: a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a first split intein, wherein the first portion of the split intein is fused to the carboxy terminus of the first portion of the exogenous polypeptide; a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to (i) a second portion of the first split intein at the amino terminus of the second portion of the exogenous polypeptide and (ii) a first portion of a second split intein at the carboxy terminus of the second portion of the exogenous polypeptide; and a third AAV vector particle comprising a third nucleic encoding a third fusion polypeptide comprising a third portion of the exogenous polypeptide fused to a second portion of the second split intein at the amino terminus of the

third portion of the exogenous polypeptide, wherein the first, second, and third fusion polypeptides are produced in the cell from the first, second and third nucleic acids, and wherein the respective portions of the first and second split inteins promote joining of (a) the carboxy terminus of the first portion of the exogenous polypeptide to the amino terminus of the second portion of the exogenous polypeptide and (b) the carboxy terminus of the second portion of the exogenous polypeptide to the amino terminus of the third portion of the exogenous polypeptide, thereby producing the exogenous polypeptide in the cell; wherein the exogenous polypeptide produced is larger than can be encoded by a single AAV vector particle. In one embodiment of this and all other aspects described herein, the first and second split inteins do not cross-splice.

[0054] In another aspect, provided herein is a composition(s) as described herein for use in the treatment of a disease or disorder in a subject in need thereof (e.g., a subject having a muscular or neuromuscular disorder).

BRIEF DESCRIPTION OF DRAWINGS

[0055] FIGs. 1A-1B. (FIG. 1A) Schematic representation of the *DMD* coding sequences (top) encoding the full-length "muscle-specific" isoform of dystrophin (bottom), which consists of an amino-terminal globular domain that binds the actin cytoskeleton, followed by a flexible and elastic rod domain composed of 24 Spectrin-like repeats interspersed with four proline-rich "hinge" regions. A dystroglycan-binding domain (DgBD) is located after the rod domain, followed by the carboxy-terminal (CT) domain that contains binding sites for the syntrophin and dystrobrevin protein families. The DgBD and the CT domain nucleate the assembly of the dystrophin–glycoprotein protein complex (DGC). **(FIG. 1B)** Different μDys constructs being evaluated in clinical trials by Pfizer, Sarepta Therapeutics, and Solid Biosciences.

[0056] FIG. 2. Dual AAV vector homologous recombination strategy to reconstitute mini-Dys (ΔH2-SR19). Two AAV vectors encode either N- (top) or C-terminal (bottom) mini-Dys fragments. Both vectors carry a recombinant sequence (exon 51 to 53) which allows the formation of larger and functional mini-Dys (ΔH2-SR19).

[0057] FIG. 3. Schematic representation of protein trans-splicing mediated by contiguous (more common) or split inteins.

[0058] FIGs. 4A-4B Example of GFP reconstitution using split Npu intein in HEK293 cells. (FIG. 4A) Brightfield and fluorescent microscopy pictures of living HEK293 cells transfected with Wild-type (WT) GFP, N-terminal and/or C-terminal GFP/Npu intein plasmids. **(FIG. 4B)** The GFP fluorescence intensity was measured using a spectrophotometer. Values are represented as mean ± s.e.m (n=3).

[0059] FIGs. 5A-5C *in vitro* validation of mini-Dys reconstitution. (FIG. 5A) Schematic representation of intein-mediated mini-Dys reconstitution. Left: N-terminal vector encoding human *DMD* sequences from exons 1 to 50, but lacking exons 21 to 41. Right: C-terminal vector encoding human *DMD* sequences from exons 51 to 79. The mini-Dys sequences are fused to N- or C-terminal halves of the selected intein. **(FIG.**

5B) Western blot analysis of HEK293 cells lysates showing the 290 kDa mini-Dys. In control mini-dys, cells were transfected with plasmid expressing the entire mini-Dys ΔSR5-15. In split mini-Dys/intein, cells were co-transfected with both N- and C-terminal vectors. Each lane represents a selected split site between SR19 and Hinge3. **(FIG. 5C)** Densitometry quantification of mini-Dys normalized to loading control GAPDH (*n*= 4-5 independent experiments). Data are shown as mean ± s.e.m. M.W: molecular weight. kDa: kiloDalton.

[0060] FIGs. 6A-6B Schematic representation of AAV-based Dystrophin replacement using SIMPL-GT (Split Intein-Mediated Protein Ligation for Gene Therapy) approach. (FIG. 6A) Dual vector strategy which consists of simultaneous administration of two AAV vectors that express two halves of a mini-Dys (ΔSR5-15) fused to split intein. Following in-frame transcription and translation with the N- or C-terminal mini-Dys fragments, the intein polypeptides are self- excised and join the adjacent peptides, thus expressing a highly functional mini-Dys (ΔSR5-15). **(FIG. 6B)** Expression of full-length Dystrophin *via* triple AAV vectors administration. The 1st AAV vector encodes proteins from N-terminus to SR8 of Dystrophin fused to N-terminal fragment of split intein1. The 2nd AAV vector encodes a middle fragment of Dystrophin (SR9-19) flanked by both C-terminal half of intein1 and N-terminal half of intein2. While the 3rd AAV vector encodes for C-terminal fragment of Dystrophin which is fused to the C-terminal half of intein2. The double trans-splicing of intein1 and 2, respectively, will lead to the ligation of three Dystrophin fragments into full-length protein.

[0061] FIG. 7 Split intein screening using the split GFP system. N- or the C-terminal half of GFP were cloned in-frame with the N- or the C-terminal half of our codon optimized split inteins. The human embryonic kidney 293 (HEK293) cells were co-transfected with both N- and C-terminal GFP/intein plasmids. 24 hours later, The GFP fluorescence intensity was measured using a spectrophotometer. Values are represented as mean ± s.e.m of percentage *versus* the WT GFP (*n*=5-6). The protein ligation efficiency of each split intein (GFP fluorescence of a given intein/internal control) is labeled on the bar.

[0062] FIG. 8 Split intein specificity and cross-reactivity using split GFP system. To test the cross-reactivity of a given N-terminal split intein with a C-terminal of another split intein, N- and C-terminal split GFP-inteins were tested on HEK293 cells. As split inteins from the 1st group present amino-acid similarities, they showed poor specificity and cross-reacted with different inteins of the same group. However, split inteins from the 2nd group, i.e. gp41.1, IMPDH and Nrdj1, were more specific toward the other half of the same intein and did not cross-react with any other split-intein. Values are represented as mean ± s.e.m of percentage *versus* the WT GFP (*n*=3-4). These observations are very important specially for triple vector strategy, where 2 very specific split inteins are needed to ligate 3 dystrophin fragments into full-length protein.

[0063] FIGs. 9A-9C. Split intein footprint importance and optimization for Dystrophin reconstitution. (FIG. 9A) The intein-mediated protein trans-splicing is highly dependent on juxtaposed amino acids that are found in native bacterial extein proteins. When N- and C-terminal split intein

fragments fuse and splice out, these native extein amino-acids (AEY and CFN for both **Aha** and **Sel**; and SGY and SSS for **gp41.1**; GGG and SIC for **IMPDH**; NPC and SEI for **Nrdj1**) are left as a footprint in the reconstituted protein. Here, the inventors tested several combinations to reduce this footprint to a minimum. These data show that with Aha split intein, the same splicing efficiency is achievable with AEY as AEY/CFN when tested on split GFP system, while with gp41.1 GY/S are sufficient for efficient GFP ligation. This footprint was even shorter with IMPDH and Nrdj1 with only G/S and C/S, respectively. Values are represented as mean ± s.e.m of percentage *versus* the WT GFP (*n*=4). Schematic example of protein trans-splicing using Aha and/or gp41.1 with minimal footprint leftover for full-length (**FIG. 9B**) or mini-Dystrophin (**FIG. 9C**) reconstitution.

[0064] FIG. 10 Identification of several split sites in human Dystrophin protein where some native amino acids can be used as part of the intein footprint.

[0065] FIGs. 11A-11B. (FIG. 11A) Western blot analysis of HEK293 cells lysates showing the 290 kDa mini-Dys. In control mini-dys, cells were transfected with plasmid expressing the entire mini-Dys ΔSR5-15. In split mini-Dys/intein, cells were co-transfected with both N- and C-terminal vectors. Each lane represents a selected split site between SR19 and Hinge3. (**FIG. 11B**) Densitometry quantification of mini-Dys normalized to loading control GAPDH (*n*= 4-5 independent experiments). Data are shown as mean ± s.e.m. M.W: molecular weight. kDa: kiloDalton. These data show that inserting split gp41.1 in 4 different splitting sites efficiently reconstitutes the mini-Dystrophin ΔSR-5-15.

[0066] FIGs. 12A-12C. *in vivo* mini-Dystrophin ΔSR-5-15 expression after AAV intramuscular injections. The split mini-Dystrophin/intein clones were inserted into pAAV plasmid containing the muscle-specific creatine kinase 8 (CK8) regulatory cassette and small synthetic polyA flanked by two AAV serotype 2 inverted terminal repeats (ITRs). The final pAAV plasmids were co-transfected with the pDG6 packaging plasmid into HEK293 cells to generate recombinant AAV2/6 vectors and purified via heparin-affinity chromatography then concentrated using sucrose gradient centrifugation. A dose of $5x10^{10}$ viral genome (v.g) of AAV encoding the N- and/or C-terminal split mini-Dystrophin/intein was administrated into tibialis anterior muscles (T.A) of three- week-old C57BL/6-*mdx*^{4cv}. Four weeks post-injection, the injected muscles were harvested, and total proteins were extracted and separated on SDS gel for western blotting (**FIG. 12A**). A strong expression of mini-Dystrophin ΔSR5-15 was detected in 4 T.A muscle tested, highlighting the efficacy of SIMPLI-GT approach. Muscles were cryo-sectioned and immunostained for dystrophin (**FIG. 12B**) or stained with Hematoxylin and Eosin, and myofiber size and nuclei position were measured (**FIG. 12C**). The reconstituted mini-Dystrophin ΔSR5-15 was correctly localized at the myofiber sarcolemma of *mdx*^{4cv} injected with dual AAV N- and C-terminal vectors. These muscles exhibit a general muscle histology improvement with absence of inflammation.

[0067] FIG. 13. *in vitro* proof-of-concept of full-length Dystrophin expression *via* triple vector strategy. Western blot analysis of HEK293 cells lysates transfected with 3 plasmids expressing either N-,

C- or middle fragments of human Dystrophin. Split intein gp41.1 was used to ligate the middle with the C-terminal fragment, while 6 different split inteins were tested for N-terminal and middle fragment ligation.

[0068] FIGs. 14A-14B. *in vitro* proof-of-concept of full-length Dysferlin expression. (FIG. 14A) Western blot analysis of HEK293 cells lysates transfected with plasmid expressing either the full-length human Dysferlin or split Dysferlin/gp41.1 intein or Dysferlin/IMPDH. 3 splitting sites were tested. **(FIG. 14B)** Densitometry quantification of full-length Dysferlin expression normalized to loading control GAPDH (n=5 independent experiments). Data are shown as mean ± s.e.m. These data show that inserting split gp41.1 in 2 different splitting sites efficiently reconstitutes the full-length Dysferlin.

[0069] FIGs. 15A-15W. Split intein DNA and protein sequences. FIG. 15A, Aha (SEQ ID Nos: 1 & 2). **FIG. 15B,** Aov (SEQ ID Nos: 3 & 4). **FIG. 15C,** Asp (SEQ ID Nos: 5 & 6). **FIG. 15D,** Ava (SEQ ID Nos: 7 & 8). **FIG. 15E,** Cra (SEQ ID Nos: 9 & 10). **FIG. 15F,** Csp-CCY (SEQ ID Nos: 11 & 12). **FIG. 15G,** Csp-PCC7424 (SEQ ID Nos: 13 & 14). **FIG. 15H,** Csp-PCC8801 (SEQ ID Nos: 15 & 16). **FIG. 15I,** Cwa (SEQ ID Nos: 17 & 18). **FIG. 15J,** gp41.1 (SEQ ID Nos: 19 & 20). **FIG. 15K,** gp41.8 (SEQ ID Nos: 21 & 22). **FIG. 15L,** IMPDH (SEQ ID Nos: 23 & 24). **FIG. 15M,** Maer (SEQ ID Nos: 25 & 26). **FIG. 15N,** Mcht (SEQ ID Nos: 27 & 28). **FIG. 15O,** Npu (SEQ ID Nos: 29 & 30). **FIG. 15P,** Nrdj (SEQ ID Nos: 31 & 32). **FIG. 15Q,** Oli (SEQ ID Nos: 33 & 34). **FIG. 15R,** Sel (SEQ ID Nos: 35 & 36). **FIG. 15S,** Ssp-PCC6803 (SEQ ID Nos: 37 & 38). **FIG. 15T,** Ssp-PCC7002 (SEQ ID Nos: 39 & 40). **FIG. 15U,** Tel (SEQ ID Nos: 41 & 42). **FIG. 15V,** Ter (SEQ ID Nos: 43 & 44). **FIG. 15W,** Tvu (SEQ ID Nos: 45 & 46).

[0070] FIG. 16. Full-length dysferlin split sites.

[0071] FIG. 17. Full-length dystrophin split sites (IMPDH intein).

[0072] FIG. 18. Full-length dystrophin split sites (Nrdj intein).

[0073] FIG. 19. Full-length dystrophin split sites.

[0074] FIG. 20. Full-length dystrophin split sites (gp41.1 intein).

[0075] FIG. 21. Mini-dystrophin ΔSR5-15 split sites.

[0076] FIGs. 22A-22D *in vivo* expression of full-length dystrophin following intramuscular administration of 3 intein vectors. Split dystrophin/intein clones for each combination were packaged into an AAV6 vector using the CK8e promoter, and were administrated locally into TA muscles of 3-week-old *mdx*^{4cv} mice at 5x10^{10} v.g per construct. Four weeks after the injection, total proteins were analyzed by western blot using an antibody that recognizes the C-terminal end of dystrophin **(FIG. 22A)**. Similar to *in vitro* observations, a 427 kDa band was detected with both combinations with a higher expression detected with split Nrdj1 & split gp41.1 combination (2 to 4-fold more *versus* WT expression). More importantly, a dramatic reduction of centrally nucleated myofiber was observed at this early time point, from ~65% in untreated *mdx*^{4cv} TA muscles *versus* 30 to 40% in muscles treated with this triple vector strategy, with remarkable improvements of the general muscle histology in this short time point **(FIGs. 22B, 22C)**. The reconsituted full-length dystrophin was correctly localized at the sarcolema membrane when assessed using any of three diffrent antibodies that recognize the N-, C-terminal or middle fragments of the full-length

dystrophin protein. **FIG. 22A**, Western blot (above) showing the expression of full-length dystrophin following triple vector administration in *mdx*^{4cv} TA muscles. **FIG. 22B,** Visualization of centrally-nucleated myofibers in cross-sections of *mdx*^{4cv} TA muscles treated with different vector combinations (or saline) and stained with Hematoxylin and Eosin. Also shown are untreated wild-type (WT) or *mdx*^{4cv} TA muscles from age-matched mice. N-ter only: muscles injected with only a single vector, in this case the N-terminal vector; middle only: muscles injected with only a single vector, in this case the middle vector; C-ter only: muscles injected with only a single vector, in this case the C-terminal vector. Other panels show muscles injected with combinations of two vectors or all 3 (triple). **FIG. 22C,** Quantification of centrally-nucleated myofibers in cross-sections of mdx4cv TA muscles treated with the indicated triple vector combinations (or saline) and stained with Hematoxylin and Eosin. Also shown are values from untreated wild-type (WT) mouse TA muscles. Data is from counting ~400 myofibers from various muscles. **FIG. 22D**, Triple immunolabeling of TA muscle cryosections using antibodies against the N-, C-terminal or middle fragments of dystrophin. Wild-type muscles were uninjected; the *mdx*^{4cv} muscles were injected with saline or 2 triple vector combinations as in FIG. 22A.

[0077] FIGs. 23A-23C *in vivo* expression of mini-Dys and full-length dystrophin following intravenous infusion of dual or triple vectors. 8-week-old *mdx*^{4cv} were systemically treated with a total dose of $2x10^{14}$ vg/kg for three months of treatment. Both hindlimb and diaphragm muscle contractile properties were assessed using a muscle force transducer (**FIG. 23A, FIG. 23B**). Mice treated with Dual or triple vector exhibited significant improvements of muscle specific force development of the tibialis anterior and diaphragm muscles *versus* saline-treated *mdx*^{4cv} and wild-type mouse muscles. Using western blot, these muscles showed strong expression of mini-Dys and full-length dystrophin (**FIG. 23C**). **FIG. 23A,** *in vivo* specific force of *tibialis anterior* muscles. **FIG. 23B,** *in vitro* specific force of isolated diaphragm muscle strips. **FIG. 23C,** Western blot showing expression of mini-Dys and full-length dystrophin in *tibialis anterior* muscles following systemic administration of dual or triple vectors.

DETAILED DESCRIPTION

[0078] Provided herein are methods and compositions useful for the delivery of exogenous polypeptides that are too large to fit in a single adenoviral, adeno-associated, lentiviral or retroviral vector. The methods and compositions described herein employ the use of split inteins, which mediate the fusion of a first and second portion of a large exogenous polypeptide delivered using at least two viral vectors (e.g., AAV vectors), thereby permitting delivery of a large exogenous polypeptide to a cell (e.g., a muscle cell). The methods and compositions also relate to muscle-specific cell expression of such exogenous polypeptides (e.g., dystrophin, utrophin and dysferlin).

Definitions

[0079] For convenience, certain terms employed in the entire application (including the specification, examples, and appended claims) are collected here. Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs.

[0080] As used herein, the term "splice" or "splices" means to excise an internal portion of a polypeptide, with joinder of the portions flanking the internal portion to form two or more smaller polypeptide molecules (e.g., an excised polypeptide and a spliced polypeptide. In some cases, splicing also includes the step of fusing together two or more of the smaller polypeptides to form a new polypeptide. Splicing can also refer to the joining of two polypeptides encoded on two separate nucleic acid sequences or in two separate vectors through the action of a split intein.

[0081] As used herein, the term "cleave" or "cleaves" means to divide a single polypeptide to form two or more smaller polypeptide molecules. In some cases, cleavage is mediated by the addition of an extrinsic endopeptidase, which is often referred to as "proteolytic cleavage." In other cases, cleaving can be mediated by the intrinsic activity of one or both of the cleaved peptide sequences, which is often referred to as "self-cleavage." Cleavage can also refer to the self-cleavage of two polypeptides that is induced by the addition of a non-proteolytic third peptide, as in the action of a split intein system as described herein.

[0082] By the term "fused" is meant covalently bonded to. For example, a first peptide is fused to a second peptide when the two peptides are covalently bonded to each other (e.g., via a peptide bond).

[0083] As used herein, the term "intein" refers to a naturally occurring, self-splicing protein subdomain that is capable of excising out its own protein subdomain from a larger protein structure while simultaneously joining the two formerly flanking peptide regions ("exteins") together to form a mature host protein. In some inteins, the precursor protein comes from two genes, which is referred to as a 'split intein.'

[0084] As used herein, the term "split intein" refers to an intein that is comprised of two or more separate components not fused to one another. Split inteins can occur naturally, or can be engineered by splitting contiguous inteins. Typically, the term "split intein" refers to any intein in which one or more peptide bond breaks exists between the N-terminal intein segment and the C-terminal intein segment such that the N-terminal and C-terminal intein segments become separate molecules that can non-covalently reassociate, or reconstitute, into an intein that is functional for splicing or cleaving reactions. Any catalytically active intein, or fragment thereof, can be used to derive a split intein for use in the systems and methods disclosed herein. For example, in one aspect the split intein can be derived from a eukaryotic intein. In another aspect, the split intein can be derived from a bacterial intein. In another aspect, the split intein can be derived from an archaeal intein. Preferably, the split intein so-derived will possess only the amino acid sequences essential for catalyzing splicing reactions.

[0085] As used herein, the "N-terminal intein segment" refers to any intein sequence that comprises an N-terminal amino acid sequence that is functional for splicing and/or cleaving reactions when combined with

a corresponding C-terminal intein segment. An N-terminal intein segment thus also comprises a sequence that is spliced out when splicing occurs. An N-terminal intein segment can comprise a sequence that is a modification of the N-terminal portion of a naturally occurring (native) intein sequence. For example, an N-terminal intein segment can comprise additional amino acid residues and/or mutated residues so long as the inclusion of such additional and/or mutated residues does not render the intein non-functional for splicing or cleaving. Preferably, the inclusion of the additional and/or mutated residues improves or enhances the splicing activity and/or controllability of the intein. Non-intein residues can also be genetically fused to intein segments to provide additional functionality, such as the ability to be affinity purified or to be covalently immobilized.

[0086] As used herein, the "C-terminal intein segment" refers to any intein sequence that comprises a C-terminal amino acid sequence that is functional for splicing or cleaving reactions when combined with a corresponding N-terminal intein segment. In one aspect, the C-terminal intein segment comprises a sequence that is spliced out when splicing occurs. In another aspect, the C-terminal intein segment is cleaved from a peptide sequence fused to its C-terminus. The sequence which is cleaved from the C-terminal intein's C-terminus is a protein for the treatment of a muscular disorder, such as dystrophin, utrophin, dysferlin, mini-dystrophin, or the like. A C-terminal intein segment can comprise a sequence that is a modification of the C-terminal portion of a naturally occurring (native) intein sequence. For example, a C terminal intein segment can comprise additional amino acid residues and/or mutated residues so long as the inclusion of such additional and/or mutated residues does not render the C-terminal intein segment non-functional for splicing or cleaving. Preferably, the inclusion of the additional and/or mutated residues improves or enhances the splicing and/or cleaving activity of the intein.

[0087] As used herein, the term "larger than can be encoded by a single AAV vector particle" refers to a polypeptide for which nucleic acid encoding it exceeds the packaging limits of an AAV vector particle. While exact packaging limits can vary slightly with serotype or variant of AAV vector used, the maximum genome-packaging capacity of AAV vectors that efficiently infect and transduce target cells is about 5 kb (the wild-type AAV genome is about 4.7 kb; larger genomes up to 5.5 kb or more can be packaged under certain conditions, but they do not efficiently infect and transduce target cells). Excluding ITRs, this permits the incorporation of about 3.5 kb of DNA for the promoter, transgene coding region, polyadenylation sequence and other regulatory elements for a transgene construct to be carried by a single AAV vector particle. Thus, transgenes requiring more than about 3.5 kb to direct expression of a desired protein are larger than can be encoded by a single AAV vector particle as the term is used herein. In some embodiments, the protein that is larger than can be encoded by a single vector particle requires at least 4 kb, at least 4.5 kb, at least 5 kb, at least 5.5 kb, at least 6 kb, at least 6.5 kb, at least 7 kb, at least 7.5 kb, at least 8 kb, at least 8.5 kb, at least 9 kb, at least 9.5 kb, at least 10 kb, at least 10.5 kb, at least 11 kb, at least 11.5 kb, at least 12 kb, at least 12.5 kb, at least 13 kb, at least 13.5 kb, at least 14 kb or more to encode the transgene polypeptide. As also discussed elsewhere herein, when the target protein requires more nucleic

acid sequence than will fit into two separate viral vectors to generate a full length target polypeptide (or sub-full length polypeptide with functional improvements over a more truncated mini- or microgene construct), the polypeptide can be split over separate vectors including three or potentially more split intein constructs. In this instance co-infection with the set of vectors can generate the full length or improved sub-full length polypeptide.

[0088] As used herein, the terms "first portion of an exogenous polypeptide fused to a first portion of a split intein" and "second portion of the exogenous polypeptide fused to a second portion of the split intein" as used in regard to methods for delivering an exogenous polypeptide to a cell, producing an exogenous polypeptide in a cell or methods of treatment or prophylaxis based on such delivery or production or compositions therefor as described herein refer to fragments of a target polypeptide that is larger than can be encoded by a single AAV vector particle. The first portion and second portion fragments of the target polypeptide are fused respectively to amino and carboxy-terminal portions of a split intein in a manner that permits excision of the intein and covalent joining of the first and second portion (engineered extein) polypeptides to reconstitute the target protein when both fusion protein are expressed in a cell. The sizes of the first portion and second portion of the target protein can vary, e.g., with the amino-terminal fragment being shorter than, approximately the same size as or larger than the carboxy-terminal fragment (and the corresponding carboxy-terminal fragment varying such that it is longer than, approximately the same size as or shorter than the amino-terminal fragment, respectively), but it is preferred, where a target is divided into two fragments, that the target is split approximately near the middle of the target protein. Where a target protein is divided into three fragments as described herein, the sizes can vary, but it is preferred that the three fragments are also approximately the same length. It can be considered to split the target protein between or at the junction of structural domains, rather than within them, e.g., between alpha helices, beta sheets, or between any two such structural domains. As a non-limiting example, in the context of a dystrophin or utrophin polypeptide, it is contemplated that the protein be split between spectrin-like repeat domains, or between a spectrin-like repeat domain and a hinge domain. The various domains of exemplary large proteins dystrophin, utrophin and dysferlin are discussed further herein below. The boundaries of various domains for dystrophin and dysferlin polypeptides are also described herein below, and one of ordinary skill in the art can determine boundaries between domains in other proteins.

[0089] As used herein the term "comprising" or "comprises" is used in reference to compositions, methods, and respective component(s) thereof, that are essential to the invention, yet open to the inclusion of unspecified elements, whether essential or not.

[0090] As used herein the term "consisting essentially of" refers to those elements required for a given embodiment. The term permits the presence of additional elements that do not materially affect the basic and novel or functional characteristic(s) of that embodiment of the invention.

[0091] The term "consisting of" refers to compositions, methods, and respective components thereof as described herein, which are exclusive of any element not recited in that description of the embodiment.

[0092] As used in this specification and the appended claims, the singular forms “a," "an," and "the" include plural references unless the context clearly dictates otherwise. Thus for example, references to "the method" includes one or more methods, and/or steps of the type described herein and/or which will become apparent to those persons skilled in the art upon reading this disclosure and so forth. It is understood that the foregoing detailed description and the following examples are illustrative only and are not to be taken as limitations upon the scope of the invention. Various changes and modifications to the disclosed embodiments, which will be apparent to those of skill in the art, may be made without departing from the spirit and scope of the present invention. Further, all patents, patent applications, and publications identified are expressly incorporated herein by reference for the purpose of describing and disclosing, for example, the methodologies described in such publications that might be used in connection with the present invention. These publications are provided solely for their disclosure prior to the filing date of the present application. Nothing in this regard should be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention or for any other reason. All statements as to the date or representation as to the contents of these documents are based on the information available to the applicants and do not constitute any admission as to the correctness of the dates or contents of these documents.

[0093] The disclosure described herein, in a preferred embodiment, does not concern a process for cloning human beings, processes for modifying the germ line genetic identity of human beings, uses of human embryos for industrial or commercial purposes or processes for modifying the genetic identity of animals which are likely to cause them suffering without any substantial medical benefit to man or animal, and also animals resulting from such processes.

Muscular Dystrophies

[0094] Muscular dystrophy is a group of inherited disorders characterized by progressive muscle weakness and loss of muscle tissue.

[0095] Muscular dystrophies include many inherited disorders, including Becker muscular dystrophy and Duchenne muscular dystrophy, which are both caused by mutations in the dystrophin gene (i.e., DMD). Both of the disorders have similar symptoms, although Becker muscular dystrophy is a slower progressing form of the disease. Duchenne muscular dystrophy is a rapidly progressive form of muscular dystrophy.

[0096] Both disorders are characterized by progressive muscle weakness of the legs and pelvis which is associated with a loss of muscle mass (wasting). Muscle weakness also occurs in the arms, neck, and other areas, but not as severely as in the lower half of the body. Calf muscles initially enlarge (an attempt by the body to compensate for loss of muscle strength), the enlarged muscle tissue is eventually replaced by fat and connective tissue (pseudohypertrophy). Muscle contractions occur in the legs and heels, causing inability to use the muscles because of shortening of muscle fibers and fibrosis of connective tissue. Bones develop abnormally, causing skeletal deformities of the chest and other areas. Cardiomyopathy occurs in

almost all cases. A mouse model for DMD exists, and is proving useful for furthering understanding of both the normal function of dystrophin and the pathology of the disease. In particular, experiments that enhance the production of utrophin, a dystrophin relative, in order to compensate for the loss of dystrophin are promising, and may lead to the development of effective therapies for this devastating disease.

[0097] Dysferlinopathy is a muscular dystrophy that is caused by mutations in the dysferlin gene. The symptoms of dysferlinopathy vary significantly between individuals. Clinical presentations most commonly associated with dysferlinopathy include limb girdle muscular dystrophy (LGMD2B), Miyoshi myopathy, distal myopathy with anterior tibial onset (DMAT), proximodistal weakness, pseudometabolic myopathy, and hyperCKemia. Most commonly, patients report distal muscle weakness in the second decade of life with loss of distal motor function within the ensuing decade. Patients generally require a wheelchair for motility with varying degrees of overall body control. As dysferlinopathy is often misdiagnosed, its incidence has not been determined. To date, there is no effective treatment to slow the loss of muscle function or reverse/improve the dystrophic phenotype.

Exogenous Polypeptides

[0098] A distinct advantage of the methods and compositions described herein is the ability to encode and deliver large proteins to a cell, e.g., a muscle cell, among others. Vectors, such as adenoviral associated vectors (AAV), are limited in their capacity to package nucleic acids and, as such, large proteins cannot be encoded and delivered on a single AAV vector. The methods and compositions described herein utilize split inteins, where an N-terminal region of a split intein and a portion of a desired exogenous polypeptide are encoded on a first AAV vector, and a C-terminal region of the split intein and a second portion of the desired exogenous polypeptide is encoded on a second AAV vector. Expression of the products from each AAV together in a cell permit the first and second portions of the split intein to promote joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide. As one of skill in the art will appreciate, the methods and compositions described herein can be utilized for any large gene products, which need not be limited by function.

[0099] The methods and compositions described herein are exemplified in the working examples using muscle proteins including dystrophin, utrophin, and dysferlin.

[00100] *Dystrophin:* Dystrophin is a 427 kDa cytoskeletal protein and is a member of the spectrin/α-actinin superfamily (See e.g., Blake et al., Brain Pathology, 6:37 (1996); Winder, J. Muscle Res. Cell. Motil., 18:617 (1997); and Tinsley el al., PNAS, 91:8307 (1994)). The N-terminus of dystrophin binds to actin, having a higher affinity for non-muscle actin than for sarcomeric actin. Dystrophin is involved in the submembranous network of non-muscle actin underlying the plasma membrane. Dystrophin is associated with an oligomeric, membrane spanning complex of proteins and glycoproteins, the dystrophin-associated protein complex (DPC). The C-terminus of dystrophin binds to the cytoplasmic tail of β-dystroglycan, and in concert with actin, anchors dystrophin to the sarcolemma. Also bound to the C-terminus of dystrophin

are the cytoplasmic members of the DPC. Dystrophin thereby provides a link between the actin-based cytoskeleton of the muscle fiber and the extracellular matrix. It is this link that is disrupted in muscular dystrophy.

[00101] The central rod domain of dystrophin is composed of a series of 24 weakly repeating units of approximately 110 amino acids, similar to those found in spectrin (i.e., spectrin-like repeats). This domain constitutes the majority of dystrophin and gives dystrophin a flexible rod-like structure. The rod-domain is interrupted by four hinge regions that are rich in proline. It is contemplated that the rod-domain provides a structural link between members of the DPC.

Table 1: Domains of Full-length dystrophin cDNA

Nucleotides	*Domain*	*SEQ ID NO:*
1-208	5′ untranslated region	SEQ ID NO: 47
209-211	Start codon (ATC)	----
209-964	N terminus	SEQ ID NO: 48
965-1219	Hinge 1	SEQ ID NO: 49
1220-1546	Spectrin-like repeat No. 1	SEQ ID NO: 50
1547-1879	Spectrin-like repeat No. 2	SEQ ID NO: 51
1880-2212	Spectrin-like repeat No. 3	SEQ ID NO: 52
2213-2359	Hinge 2	SEQ ID NO: 53
2360-2692	Spectrin-like repeat No. 4	SEQ ID NO: 54
2693-3019	Spectrin-like repeat No. 5	SEQ ID NO: 55
3020-3346	Spectrin-like repeat No. 6	SEQ ID NO: 56
3347-3673	Spectrin-like repeat No. 7	SEQ ID NO: 57
3674-4000	Spectrin-like repeat No. 8	SEQ ID NO: 58
4001-4312	Spectrin-like repeat No. 9	SEQ ID NO: 59
4313-4588	Spectrin-like repeat No. 10	SEQ ID NO: 60
4589-4915	Spectrin-like repeat No. 11	SEQ ID NO: 61
4916-5239	Spectrin-like repeat No. 12	SEQ ID NO: 62
5340-5551	Spectrin-like repeat No. 13	SEQ ID NO: 63
5552-5833	Spectrin-like repeat No. 14	SEQ ID NO: 64
5834-6127	Spectrin-like repeat No. 15	SEQ ID NO: 65
6128-6187	20 amino acid insert (not hinge)	SEQ ID NO: 66
6188-6514	Spectrin-like repeat No. 16	SEQ ID NO: 67
6515-6835	Spectrin-like repeat No. 17	SEQ ID NO: 68
6836-7186	Spectrin-like repeat No. 18	SEQ ID NO: 69
7187-7489	Spectrin-like repeat No. 19	SEQ ID NO: 70
7490-7612	Hinge 3	SEQ ID NO: 71
7613-7942	Spectrin-like repeat No. 20	SEQ ID NO: 72
7943-8269	Spectrin-like repeat No. 21	SEQ ID NO: 73
8270-8617	Spectrin-like repeat No. 22	SEQ ID NO: 74

8618-9004	Spectrin-like repeat No. 23	SEQ ID NO: 75
9005-9328	Spectrin-like repeat No. 24	SEQ ID NO: 76
9329-9544	Hinge 4	SEQ ID NO: 77
9545-10431	Start of C terminus	SEQ ID NO: 78
10432-11254	Alternatively spliced exons 71-78	SEQ ID NO: 79
11255-11266	End of Coding Region	SEQ ID NO: 80
11267-13957	3' untranslated region	SEQ ID NO: 81

[00102] Homologs of dystrophin have been identified in a variety of organisms, including mouse (Genbank accession number M68859); dog (Genbank accession number AF070485); and chicken (Genbank accession number X13369). Similar comparisons can be generated with homologs from other species, including but not limited to those described above, by using any of a variety of available computer programs (e.g., BLAST, from NCBI). Candidate homologs can be screened for biological activity using any suitable assay, including, but not limited to those described herein.

[00103] *Utrophin:* Utrophin is an autosomally-encoded homolog of dystrophin and it has been postulated that the proteins play a similar physiological role (For a recent review, See e.g., Blake et al., Brain Pathology, 6:37 [1996]). Human utrophin shows substantial homology to dystrophin, with the major difference occurring in the rod domain, where utrophin lacks repeats 15 and 19 and two hinge regions (See e.g., Love et al., Nature 339:55 [1989]; Winder et al., FEBS Lett., 369:27 [1995]). Utrophin thus contains 22 spectrin-like repeats and two hinge regions.

[00104] *Dysferlin:* Dysferlin comprises the following domains: C2A, C2B, C2C, FerA, DysF, C2D, C2E, C2F, C2G, and TM. The exact boundaries of each domain may vary among orthologs and variants. The approximate amino acid range for each domain in human dysferlin is shown in Table 2. The listed domain boundaries may vary by up to about 20 residues, e.g., about 5, 10, 15, or 20 residues.

Table 2: Dysferlin domains

Domain	Amino Acid Range
C2A	1-124
C2B	219-352
C2C	366-515
FerA	670-782
DysF	864-1097
C2D	1137-1281
C2E	1314-1465
C2F	1579-1696
C2G	1789-1994
TM	2045-2067

[00105] *Protein Variants:* Moreover, as described above, variant forms (e.g., mutants) of an exogenous polypeptide, such as dystrophin, utrophin, a mini-dystrophin or dysferlin, are also contemplated for use with the methods and compositions described herein. For example, it is contemplated that an isolated replacement of a leucine with an isoleucine or valine, an aspartate with a glutamate, a threonine with a serine, or a similar replacement of an amino acid with a structurally related amino acid (i.e., conservative mutations) will not necessarily have a major effect on the biological activity of the resulting molecule. Accordingly, in some embodiments, the exogenous polypeptide can comprise one or more conservative amino acid replacements. Conservative replacements are those that take place within a family of amino acids that are related in their side chains. Genetically encoded amino acids can be divided into four families: (1) acidic (aspartate, glutamate); (2) basic (lysine, arginine, histidine); (3) nonpolar (alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan); and (4) uncharged polar (glycine, asparagine, glutamine, cysteine, serine, threonine, tyrosine). Phenylalanine, tryptophan, and tyrosine are sometimes classified jointly as aromatic amino acids. In similar fashion, the amino acid repertoire can be grouped as (1) acidic (aspartate, glutamate); (2) basic (lysine, arginine histidine), (3) aliphatic (glycine, alanine, valine, leucine, isoleucine, serine, threonine), with serine and threonine optionally be grouped separately as aliphatic-hydroxyl; (4) aromatic (phenylalanine, tyrosine, tryptophan); (5) amide (asparagine, glutamine); and (6) sulfur-containing (cysteine and methionine) (See e.g., Stryer (ed.), Biochemistry, 2nd ed, W H Freeman and Co. [1981]). Whether a change in the amino acid sequence of a peptide results in a functional homolog can be readily determined by assessing the ability of the variant peptide to function in a fashion similar to the wild-type protein. Peptides in which more than one replacement has taken place can readily be tested in the same manner.

[00106] In some embodiments, a variant of an exogenous polypeptide is engineered to comprise an enhanced biological activity. Such polypeptides, when expressed from recombinant DNA constructs, can be used in therapeutic embodiments as described herein.

[00107] In some embodiments, a variant of an exogenous polypeptide can comprise an increased intracellular half-life as compared to the corresponding wild-type protein. For example, such variant protein can be more stable or less stable to proteolytic degradation or other cellular process that result in destruction of, or otherwise inactivation of the variant. Such variants, and the genes that encode them, can be utilized to alter the pharmaceutical activity of constructs expressing variant exogenous polypeptides by modulating the half-life of the protein. For instance, a short half-life can give rise to more transient biological effects. As above, such proteins find use in pharmaceutical applications or for the treatment of a muscular disease or disorder.

[00108] A wide range of techniques are known in the art for screening gene products of combinatorial libraries made by point mutations, and for screening cDNA libraries for gene products having a certain property. Such techniques are generally adaptable for rapid screening of the gene libraries generated by the combinatorial mutagenesis of a given exogenous polypeptide. The most widely used techniques for

screening large gene libraries typically comprise cloning the gene library into replicable expression vectors, transforming appropriate cells with the resulting library of vectors, and expressing the combinatorial genes under conditions in which detection of a desired activity facilitates relatively easy isolation of the vector encoding the gene whose product was detected.

[00109] In some embodiments, the exogenous polypeptide comprises a mini-dystrophin or micro-dystrophin. As used herein, a "mini-dystrophin" comprises an amino terminal actin-binding domain, a β-dystroglycan binding domain and a plurality (e.g, at least 2) spectrin-like repeat domains.

Adenoviral Associated Vectors (AAVs)

[00110] AAV is a small virus that presents very low immunogenicity and is not associated with any known human disease, making it attractive as a vector for delivery of exogenous genetic material (e.g. for gene therapy). However, the size of the AAV capsid imposes a limit on the amount of DNA that can be packaged within it. The AAV genome is approximately 4.7 kilobases (kb) in size

[00111] The methods and compositions described herein permit the delivery of large proteins (e.g., greater than 4.7 kb) by administering two (or more) AAV vectors, each having a portion of an exogenous polypeptide to be expressed and a portion of a split intein. In one embodiment, the methods and compositions described herein use at least two different adeno-associated viral (AAV) vectors. The first AAV vector comprises an N-terminal portion of a split intein fused to a first portion of an exogenous polypeptide (e.g., dystrophin, dysferlin, utrophin or other desired therapeutic protein, e.g., for a muscular or other disease or disorder) and a second AAV vector comprises a C-terminal portion of a split intein fused to a second portion of the exogenous polypeptide. Upon expression of the first and second fusion polypeptides in the cell, the first and second portions of the split intein promote joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, thereby delivering the exogenous polypeptide to the cell. This system or arrangement permits delivery of an exogenous polypeptide that is larger than can be encoded by a single AAV vector particle.

[00112] Embodiments of the first and second AAV vectors are provided herein and include the following non-limiting embodiments. An AAV vector as used herein can be in the form of a mature AAV particle or virion, i.e. nucleic acid surrounded by an AAV protein capsid. The AAV vector can comprise an AAV genome or a portion or derivative thereof. An AAV genome is a polynucleotide which encodes functions needed for production of an AAV particle. These functions include those operating in the replication and packaging cycle of AAV in a host cell, including encapsidation of the AAV genome into an AAV particle. Naturally occurring AAVs are replication-deficient and rely on the provision of helper functions in trans for completion of a replication and packaging cycle. Accordingly, an AAV genome of a vector as used herein is typically replication-deficient.

[00113] The AAV genome can be in single-stranded form, either positive or negative-sense, or alternatively in double-stranded form. The use of a double-stranded form allows bypass of the DNA

replication step in the target cell and so can accelerate transgene expression. In one embodiment, the AAV genome is in single-stranded form. The AAV genome can be from any naturally derived serotype, isolate or clade of AAV. Thus, the AAV genome can be the full genome of a naturally occurring AAV or a recombinant, engineered AAV. As is known to the skilled person, AAVs occurring in nature may be classified according to various biological systems.

[00114] Commonly, AAVs are referred to in terms of their serotype. A serotype corresponds to a variant subspecies of AAV which, owing to its profile of expression of capsid surface antigens, has a distinctive reactivity which can be used to distinguish it from other variant subspecies. Typically, a virus having a particular AAV serotype does not efficiently cross-react with neutralizing antibodies specific for any other AAV serotype. AAV serotypes include AAV1, AAV2, AAV3, AAV4, AAV5, AAV6, AAV7, AAV8, AAV9, AAV10 and AAV11, and also recombinant serotypes, such as Rec2 and Rec3. Any of these AAV serotypes can be used with the methods and compositions described herein. Reviews of AAV serotypes can be found in Choi et al. (2005) Curr. Gene Ther. 5: 299-310 and Wu et al. (2006) Molecular Therapy 14: 316-27. The sequences of AAV genomes or of elements of AAV genomes including ITR sequences, rep or cap genes can be derived from the following accession numbers for AAV whole genome sequences: Adeno-associated virus 1 NC_002077, AF063497; Adeno-associated virus 2 NC_001401; Adeno-associated virus 3 NC_001729; Adeno-associated virus 3B NC_001863; Adeno-associated virus 4 NC_001829; Adeno-associated virus 5 Y18065, AF085716; Adeno-associated virus 6 NC_001862; Avian AAV ATCC VR-865 AY186198, AY629583, NC_004828; Avian AAV strain DA-1 NC_006263, AY629583; Bovine AAV NC_005889, AY388617.

[00115] AAV can also be referred to in terms of clades or clones. This refers to the phylogenetic relationship of naturally derived AAVs, and typically to a phylogenetic group of AAVs which can be traced back to a common ancestor, and includes all descendants thereof.

[00116] Additionally, AAVs can be referred to in terms of a specific isolate, i.e. a genetic isolate of a specific AAV found in nature. The term genetic isolate describes a population of AAVs which has undergone limited genetic mixing with other naturally occurring AAVs, thereby defining a recognizably distinct population at a genetic level.

[00117] The skilled person can select an appropriate serotype, clade, clone or isolate of AAV for use with the methods and compositions described herein on the basis of their common general knowledge of the particular AAV characteristics.

[00118] The AAV serotype determines the tissue specificity of infection (or tropism) of an AAV virus. Accordingly, preferred AAV serotypes for use in AAVs administered to patients in accordance with the methods and compositions described herein are those which, for example, have natural tropism for or a high efficiency of infection of target cells within a muscle.

[00119] Typically, the AAV genome of a naturally derived serotype, isolate or clade of AAV comprises at least one inverted terminal repeat sequence (ITR). An ITR sequence acts in *cis* to provide a functional origin of replication and allows for integration and excision of the vector from the genome of a cell.

[00120] The AAV genome typically also comprises packaging genes, such as rep and/or cap genes which encode packaging functions for an AAV particle. The rep gene encodes one or more of the proteins Rep78, Rep68, Rep52 and Rep40 or variants thereof. The cap gene encodes one or more capsid proteins such as VP1, VP2 and VP3 or variants thereof. These proteins make up the capsid of an AAV particle. Capsid variants are discussed below. A promoter can be operably linked to each of the packaging genes. Specific examples of such promoters include the p5, p19 and p40 promoters (Laughlin et al. (1979) Proc. Natl. Acad. Sci. USA 76: 5567-5571). For example, the p5 and p19 promoters are generally used to express the rep gene, while the p40 promoter is generally used to express the cap gene.

[00121] Typically, the AAV genome for use with the methods and compositions described herein will be derivatized for the purpose of administration to patients. Such derivatization is standard in the art (see *e.g.,* Coura and Nardi (2007) Virology Journal 4: 99). Derivatives of an AAV genome include any truncated or modified forms of an AAV genome which allow for expression of a transgene *in vivo*. Typically, it is possible to truncate the AAV genome significantly to include minimal viral sequence yet retain the above function. This is preferred for safety reasons to reduce the risk of recombination of the vector with wild-type virus, and also to avoid triggering a cellular immune response by the presence of viral gene proteins in the target cell.

[00122] Typically, a derivative of an AAV genome will include at least one inverted terminal repeat sequence (ITR), preferably more than one ITR, such as two ITRs or more. One or more of the ITRs may be derived from AAV genomes having different serotypes, or may be a chimeric or mutant ITR. A preferred mutant ITR is one having a deletion of a trs (terminal resolution site). This deletion allows for continued replication of the genome to generate a single-stranded genome, which contains both coding and complementary sequences, i.e. a self-complementary AAV genome. This allows for bypass of DNA replication in the target cell, and so enables accelerated transgene expression.

[00123] The inclusion of one or more ITRs is preferred to aid concatamer formation of the vector in the nucleus of a host cell, for example following the conversion of single-stranded vector DNA into double-stranded DNA by the action of host cell DNA polymerases. The formation of such episomal concatamers protects the vector construct during the life of the host cell, thereby allowing for prolonged expression of the transgene *in vivo*.

[00124] In some embodiments, ITR elements are the only sequences retained from the native AAV genome in the derivative. Thus, a derivative will preferably not include the rep and/or cap genes of the native genome and any other sequences of the native genome. This is preferred for the reasons described above, and also to reduce the possibility of integration of the vector into the host cell genome. The following portions could therefore be removed in a derivative: one inverted terminal repeat (ITR) sequence, the

replication (rep) and capsid (cap) genes. However, in some embodiments, derivatives may additionally include one or more rep and/or cap genes or other viral sequences of an AAV genome. Naturally occurring AAV integrates with a high frequency at a specific site on human chromosome 19, and shows a negligible frequency of random integration, such that retention of an integrative capacity in the vector may be tolerated in a therapeutic setting. Where a derivative comprises capsid proteins i.e. VP1, VP2 and/or VP3, the derivative can be a chimeric, shuffled or capsid-modified derivative of one or more naturally occurring AAVs. In particular, the methods and compositions described herein encompass the provision of capsid protein sequences from different serotypes, clades, clones, or isolates of AAV within the same vector (i.e. a pseudotyped vector).

[00125] Chimeric, shuffled or capsid-modified derivatives are typically selected to provide one or more desired functionalities for the viral vector. Thus, these derivatives may display increased efficiency of gene delivery, decreased immunogenicity (humoral or cellular), an altered tropism range and/or improved targeting of a particular cell type compared to an AAV vector comprising a naturally occurring AAV genome, such as that of AAV2. Increased efficiency of gene delivery can be effected by improved receptor or co-receptor binding at the cell surface, improved internalization, improved trafficking within the cell and into the nucleus, improved uncoating of the viral particle and/or improved conversion of a single-stranded genome to double-stranded form. Increased efficiency may also relate to an altered tropism range or targeting of a specific cell population, such that the vector dose is not diluted by administration to tissues where it is not needed.

[00126] Chimeric capsid proteins include those generated by recombination between two or more capsid coding sequences of naturally occurring AAV serotypes. This can be performed, for example, by a marker rescue approach in which non-infectious capsid sequences of one serotype are co-transfected with capsid sequences of a different serotype, and directed selection is used to select for capsid sequences having desired properties. The capsid sequences of the different serotypes can be altered by homologous recombination within the cell to produce novel chimeric capsid proteins.

[00127] Chimeric capsid proteins also include those generated by engineering of capsid protein sequences to transfer specific capsid protein domains, surface loops or specific amino acid residues between two or more capsid proteins, for example between two or more capsid proteins of different serotypes.

[00128] Shuffled or chimeric capsid proteins can also be generated by DNA shuffling or by error-prone PCR. Hybrid AAV capsid genes can be created by randomly fragmenting the sequences of related AAV genes e.g. those encoding capsid proteins of multiple different serotypes and then subsequently reassembling the fragments in a self-priming polymerase reaction, which may also cause crossovers in regions of sequence homology. A library of hybrid AAV genes created in this way by shuffling the capsid genes of several serotypes can be screened to identify viral clones having a desired functionality. Similarly, error prone PCR may be used to randomly mutate AAV capsid genes to create a diverse library of variants which may then be selected for a desired property.

[00129] The sequences of the capsid genes can also be genetically modified to introduce specific deletions, substitutions or insertions with respect to the native wild-type sequence. In particular, capsid genes may be modified by the insertion of a sequence of an unrelated protein or peptide within an open reading frame of a capsid coding sequence, or at the N- and/or C-terminus of a capsid coding sequence.

[00130] The vectors used herein can encompass the provision of sequences of an AAV genome in a different order and configuration to that of a native AAV genome. The vector(s) can also include the replacement of one or more AAV sequences or genes with sequences from another virus or with chimeric genes composed of sequences from more than one virus. Such chimeric genes can be composed of sequences from two or more related viral proteins of different viral species.

[00131] AAV vectors for use as described herein can include transcapsidated forms wherein an AAV genome or derivative having an ITR of one serotype is packaged in the capsid of a different serotype. Such AAV vectors can also include mosaic forms wherein a mixture of unmodified capsid proteins from two or more different serotypes makes up the viral capsid. An AAV vector can also include chemically modified forms bearing ligands adsorbed to the capsid surface. For example, such ligands may include antibodies for targeting a particular cell surface receptor.

[00132] The first and second AAV vectors of the AAV vector system as described herein together comprise all of the components necessary for a fully functional exogenous polypeptide to be re-assembled in a target cell following transduction by both vectors. A skilled person will be aware of additional genetic elements commonly used to ensure transgene expression in a viral vector-transduced cell. These may be referred to as expression control sequences. Thus, the AAV vectors of the AAV viral vector system described herein typically comprise expression control sequences (e.g. comprising a promoter sequence) operably linked to the nucleotide sequences encoding the desired exogenous polypeptide (e.g., dystrophin, utrophin, dysferlin and the like).

[00133] Any suitable promoter can be used. The promoter sequence can be constitutively active (i.e. operational in any host cell background), or alternatively may be active only in a specific host cell environment, thus allowing for targeted expression of the transgene in a particular cell type (e.g. a tissue-specific promoter). The promoter can show inducible expression in response to presence of another factor, for example a factor present in a host cell. In any event, where the vector is administered for therapy, it is preferred that the promoter should be functional in the target cell background.

[00134] In some embodiments, it is preferred that the promoter is highly efficacious in muscle cells in order to allow for the transgene to be preferentially or only expressed in muscle cell populations. Thus, expression from the promoter may be muscle-cell specific. In one embodiment, a muscle-specific promoter is comprised by a muscle-specific expression cassette, as that term is used herein.

[00135] At least one of the vectors described herein can comprise an untranslated region (UTR) located between the promoter and the upstream polypeptide-encoding nucleic acid sequence (i.e. a 5′ UTR). Any suitable UTR sequence can be used. The UTR can comprise one or more of the following elements:

a Gallus gallus β-actin (CBA) intron 1 fragment, an Oryctolagus cuniculus β-globin (RBG) intron 2 fragment, and an Oryctolagus cuniculus β-globin exon 3 fragment. The UTR can comprise a Kozak consensus sequence. Any suitable Kozak consensus sequence can be used.

[00136] At least one of the vectors described herein can further comprise a post-transcriptional response element (also known as post-transcriptional regulatory element) or PRE. Any suitable PRE can be used. The presence of a suitable PRE can enhance expression of the desired transgene. In one embodiment, the PRE is a Woodchuck Hepatitis Virus PRE (WPRE). The one or more vectors can also comprise a poly-adenylation sequence located 3′ to the protein-encoding nucleic acid sequence. Any suitable poly-adenylation sequence can be used. In one embodiment, the poly-adenylation sequence is a bovine Growth Hormone (bGH) poly-adenylation sequence.

[00137] Expression of a given exogenous protein requires that the target cell be transduced with both the first AAV vector and the second AAV vector; however, the order is not important. Thus, the target cell can be transduced with the first AAV vector and the second AAV vector in any order (first AAV vector followed by second AAV vector, or second AAV vector followed by first AAV vector) or simultaneously. Methods for transducing target cells with AAV vectors are known in the art and will be familiar to a skilled person. The target cell is preferably a muscular cell, preferably a skeletal muscle cell or cardiac muscle cell.

[00138] While the methods and compositions described herein relate to the use of at least two adeno-associated vectors, the methods and compositions can utilize alternative vectors including, *e.g.,* second generation adenoviral vectors, lentiviral vectors, or retroviral vectors.

[00139] Second generation adenoviral vectors delete the early regions of the Ad genome (E2A, E2B, and E4). Highly modified second generation adenoviral vectors are less likely to generate replication-competent virus during large-scale vector preparation. Host immune response against late viral proteins is thus reduced (See Amalfitano et al., "Production and Characterization of Improved Adenovirus Vectors With the E1, E2b, and E3 Genes Deleted," J. Virol. 72:926-933 (1998)). The elimination of E2A, E2B, and E4 genes from the adenoviral genome also provides increased cloning capacity. This, combined with the split intein approach described herein can further increase the size of the exogenously-encoded polypeptide introduced.

[00140] Lentivirus-based vectors infect non-dividing cells as part of their normal life cycles, and are produced by expression of a package-able vector construct in a cell line that expresses viral proteins. The small size of lentiviral particles constrains the amount of exogenous DNA they are able to carry to about 10 kb.

[00141] Vectors based on Moloney murine leukemia viruses (MMLV) and other retroviruses have emerged as useful for gene therapy applications. These vectors stably transduce actively dividing cells as part of their normal life cycles, and integrate into host cell chromosomes. Retroviruses can be employed as

described herein, for example, in the context of infection and transduction of muscle precursor cells such as myoblasts, satellite cells, or other muscle stem cells.

Split inteins

[00142] Inteins are naturally occurring, self-splicing protein subdomains that are capable of excising out their own protein subdomain from a larger protein structure while simultaneously joining the two formerly flanking peptide regions ("exteins") together to form a mature host protein.

[00143] The ability of inteins to rearrange flanking peptide bonds, and retain activity when in fusion to proteins other than their native exteins, has led to a number of intein-based biotechnologies. These include various types of protein ligation and activation applications, as well as protein labeling and tracing applications. An important application of inteins is in the production of purified recombinant proteins. In particular, inteins have the ability to impart self-cleaving activity to a number of conventional affinity and purification tags, and thus provide a major advance in the production of recombinant protein products for research, medical and other commercial applications.

[00144] The use of split inteins with the methods and compositions provided herein permits large protein-encoding sequences to be divided amongst two (or more) different vectors, such as AAV vectors, which, upon expression in a cell, are ligated together to form the full protein. Given that AAV vectors are limited by the size of protein-encoding sequence they can carry, the use of split inteins permits the delivery of large proteins to a cell, which could not be encoded on a single AAV vector alone.

[00145] Any catalytically active intein, or fragment thereof, can be used to derive a split intein for use in the methods of the invention. For example, in one aspect the split intein can be derived from a eukaryotic intein. In another aspect, the split intein can be derived from a bacterial intein. In another aspect, the split intein can be derived from an archaeal intein. Preferably, the split intein so-derived will possess only the amino acid sequences essential for catalyzing trans-splicing reactions.

[00146] The N-terminal split intein, as that term is used herein, can comprise a sequence that is a modification of the N-terminal portion of a naturally occurring intein sequence. For example, an N-terminal split intein sequence can comprise additional amino acid residues and/or mutated residues so long as the inclusion of such additional and/or mutated residues does not render the intein non-functional with respect to splicing of two portions of the exogenous polypeptide. Preferably, the inclusion of the additional and/or mutated residues improves or enhances the splicing activity of the intein.

[00147] A C-terminal split intein for use with the methods and compositions described herein can be any intein sequence that comprises a C-terminal amino acid sequence that is functional for trans-splicing reactions. In one aspect, the C-terminal split intein comprises 4 to 7 contiguous amino acid residues, at least 4 amino acids of which are from the last β-strand of the intein from which it was derived. A C-terminal split intein region thus also comprises a sequence that is spliced out when trans-splicing occurs. A C-terminal split intein region can comprise a sequence that is a modification of the C-terminal portion of a

naturally occurring intein sequence. For example, an C-terminal split intein region can comprise additional amino acid residues and/or mutated residues so long as the inclusion of such additional and/or mutated residues does not render the intein non-functional with respect to splicing. Preferably, the inclusion of the additional and/or mutated residues improves or enhances the splicing activity of the C-terminal split intein region.

[00148] In some embodiments, a peptide linked to a C-terminal or N-terminal split intein region can comprise an additional chemical moiety including, among others, fluorescence groups, biotin, polyethylene glycol (PEG), amino acid analogs, unnatural amino acids, phosphate groups, glycosyl groups, radioisotope labels, and pharmaceutical molecules. In other embodiments, a peptide linked to a C-terminal split intein region can comprise one or more chemically reactive groups including, among others, ketone, aldehyde, Cys residues and Lys residues. The N-intein and C-intein of a split intein can associate non-covalently to form an active intein and catalyze the splicing reaction when an "intein-splicing polypeptide (ISP)" is present. An "intein-splicing polypeptide (ISP)" is a portion of the amino acid sequence of a split intein that remains when the C-terminal or N-terminal split intein region or both, are removed from the split intein. In certain embodiments, the N-terminal split intein region comprises the ISP. In another embodiment, the C-terminal split intein region comprises the ISP. In yet another embodiment, the ISP is a separate peptide that is not covalently linked to either the C-terminal or N-terminal split intein region.

[00149] In protein trans-splicing, one precursor protein consists of an N-extein part followed by the N-intein, another precursor protein consists of the C-intein followed by a C-extein part, and a trans-splicing reaction (catalyzed by the N- and C-inteins together) excises the two intein sequences and links the two extein sequences with a peptide bond. Protein trans-splicing, being an enzymatic reaction, can work with very low (e.g. micromolar) concentrations of proteins and can be carried out under physiological conditions.

[00150] In some embodiments, the split intein sequences used herein are codon optimized for expression in particular cells, such as eukaryotic cells (e.g., eukaryotic muscle cells). The eukaryotic cells can be those of or derived from a particular organism, such as a mammal, including but not limited to human, mouse, rat, rabbit, dog, or non-human primate. In general, codon optimization refers to a process of modifying a nucleic acid sequence for enhanced expression in the host cells of interest by replacing at least one codon (e.g. about or more than about 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, or more codons) of the native sequence with codons that are more frequently or most frequently used in the genes of that host cell while maintaining the native amino acid sequence. Various species exhibit particular bias for certain codons of a particular amino acid. Codon bias (differences in codon usage between organisms) often correlates with the efficiency of translation of messenger RNA (mRNA), which is in turn believed to be dependent on, among other things, the properties of the codons being translated and the availability of particular transfer RNA (tRNA) molecules. The predominance of selected tRNAs in a cell is generally a reflection of the codons used most frequently in peptide synthesis. Accordingly, genes can be tailored for optimal gene expression in a given

organism based on codon optimization. Codon usage tables are readily available, for example, at the "Codon Usage Database", and these tables can be adapted in a number of ways. See Nakamura, Y., et al. "Codon usage tabulated from the international DNA sequence databases: status for the year 2000" Nucl. Acids Res. 28:292 (2000). Computer algorithms for codon optimizing a particular sequence for expression in a particular host cell are also available, such as Gene Forge (Aptagen; Jacobus, Pa.).

[00151] In some embodiments, the methods and compositions described herein utilize one or more split inteins present in the following Table. Exemplary split inteins for use herein are shown herein in FIGs. 15A-15U.

Table 3: Exemplary Split Inteins

Intein Name	*Organism Name*	*Organism Description*
Eucarya		
APMV Pol	*Acanthomoeba polyphaga* Mimivirus	isolate = "Rowbotham-Bradford", Virus, infects Amoebae, taxon: 212035
Abr PRP8	*Aspergillus brevipes* FRR2439	Fungi, ATCC 16899, taxon: 75551
Aca-G186AR PRP8	*Ajellomyces capsulatus* G186AR	Taxon: 447093, strain G186AR
Aca-H143 PRP8	*Ajellomyces capsulatus* H143	Taxon: 544712
Aca-JER2004 PRP8	*Ajellomyces capsulatus* (anamorph: *Histoplasma capsulatum*)	strain = JER2004, taxon: 5037, Fungi
Aca-NAm1 PRP8	*Ajellomyces capsulatus* NAm1	strain = "NAm1", taxon: 339724
Ade-ER3 PRP8	*Ajellomyces dermatitidis* ER-3	Human fungal pathogen. taxon: 559297
Ade-SLH14081 PRP8	*Ajellomyces dermatitidis* SLH14081,	Human fungal pathogen
Afu-Af293 PRP8	*Aspergillus fumigatus* var. *ellipticus*, strain Af293	Human pathogenic fungus, taxon: 330879
Afu-FRR0163 PRP8	*Aspergillus fumigatus* strain FRR0163	Human pathogenic fungus, taxon: 5085
Afu-NRRL5109 PRP8	*Aspergillus fumigatus* var. *ellipticus*, strain NRRL 5109	Human pathogenic fungus, taxon: 41121
Agi-NRRL6136 PRP8	*Aspergillus giganteus* Strain NRRL 6136	Fungus, taxon: 5060
	Aspergillus nidulans FGSC A	Filamentous fungus,

Ani-FGSCA4 PRP8		taxon: 227321
Avi PRP8	*Aspergillus viridinutans* strain FRR0577	Fungi, ATCC 16902, taxon: 75553
Bci PRP8	*Botrytis cinerea* (teleomorph of *Botryotinia fuckeliana* B05.10)	Plant fungal pathogen
Bde-JEL197 RPB2	*Batrachochytrium dendrobatidis* JEL197	Chytrid fungus, isolate = "AFTOL-ID 21", taxon: 109871
Bde-JEL423 PRP8-1	*Batrachochytrium dendrobatidis* JEL423	Chytrid fungus, isolate JEL423, taxon 403673
Bde-JEL423 PRP8-2	*Batrachochytrium dendrobatidis* JEL423	Chytrid fungus, isolate JEL423, taxon 403673
Bde-JEL423 RPC2	*Batrachochytrium dendrobatidis* JEL423	Chytrid fungus, isolate JEL423, taxon 403673
Bde-JEL423 eIF-5B	*Batrachochytrium dendrobatidis* JEL423	Chytrid fungus, isolate JEL423, taxon 403673
Bfu-B05 PRP8	*Botryotinia fuckeliana* B05.10	Taxon: 332648
CIV RIR1	Chilo iridescent virus	dsDNA eucaryotic virus, taxon: 10488
CV-NY2A ORF212392	*Chlorella* virus NY2A infects *Chlorella* NC64A, which infects *Paramecium bursaria*	dsDNA eucaryotic virus, taxon: 46021, Family Phycodnaviridae
CV-NY2A RIR1	*Chlorella* virus NY2A infects *Chlorella* NC64A, which infects *Paramecium bursaria*	dsDNA eucaryotic virus, taxon: 46021, Family Phycodnaviridae
CZIV RIR1	*Costelytra zealandica* iridescent virus	dsDNA eucaryotic virus, Taxon: 68348
Cba-WM02.98 PRP8	*Cryptococcus bacillisporus* strain WM02.98 (aka *Cryptococcus neoformans gattii*)	Yeast, human pathogen, taxon: 37769
Cba-WM728 PRP8	*Cryptococcus bacillisporus* strain WM728	Yeast, human pathogen, taxon: 37769
Ceu ClpP	*Chlamydomonas eugametos* (chloroplast)	Green alga, taxon: 3053
Cga PRP8	*Cryptococcus gattii* (aka *Cryptococcus bacillisporus*)	Yeast, human pathogen

Cgl VMA	*Candida glabrata*	Yeast, taxon: 5478
Cla PRP8	*Cryptococcus laurentii* strain CBS139	Fungi, Basidiomycete yeast, taxon: 5418
Cmo ClpP	*Chlamydomonas moewusii*, strain UTEX 97	Green alga, chloroplast gene, taxon: 3054
Cmo RPB2 (RpoBb)	*Chlamydomonas moewusii*, strain UTEX 97	Green alga, chloroplast gene, taxon: 3054
Cne-A PRP8 (Fne-A PRP8)	*Filobasidiella neoformans* (*Cryptococcus neoformans*) Serotype A, PHLS_8104	Yeast, human pathogen
Cne-AD PRP8 (Fne-AD PRP8)	*Cryptococcus neoformans* (*Filobasidiella neoformans*), Serotype AD, CBS132).	Yeast, human pathogen, ATCC32045, taxon: 5207
Cne-JEC21 PRP8	*Cryptococcus neoformans* var. *neoformans* JEC21	Yeast, human pathogen, serotype = "D" taxon: 214684
Cpa ThrRS	*Candida parapsilosis*, strain CLIB214	Yeast, Fungus, taxon: 5480
Cre RPB2	*Chlamydomonas reinhardtii* (nucleus)	Green algae, taxon: 3055
CroV Pol	*Cafeteria roenbergensis* virus BV-PW1	taxon: 693272, Giant virus infecting marine heterotrophic nanoflagellate
CroV RIR1	*Cafeteria roenbergensis* virus BV-PW1	taxon: 693272, Giant virus infecting marine heterotrophic nanoflagellate
CroV RPB2	*Cafeteria roenbergensis* virus BV-PW1	taxon: 693272, Giant virus infecting marine heterotrophic nanoflagellate
CroV Top2	*Cafeteria roenbergensis* virus BV-PW1	taxon: 693272, Giant virus infecting marine heterotrophic nanoflagellate
Cst RPB2	*Coelomomyces stegomyiae*	Chytrid fungus, isolate = "AFTOL-ID 18", taxon: 143960
Ctr ThrRS	*Candida tropicalis* ATCC750	Yeast

Ctr VMA	*Candida tropicalis* (nucleus)	Yeast
Ctr-MYA3404 VMA	*Candida tropicalis* MYA-3404	Taxon: 294747
Ddi RPC2	*Dictyostelium discoideum* strain AX4 (nucleus)	Mycetozoa (a social amoeba)
Dhan GLT1	*Debaryomyces hansenii* CBS767	Fungi, Anamorph: *Candida famata*, taxon: 4959
Dhan VMA	*Debaryomyces hansenii* CBS767	Fungi, taxon: 284592
Eni PRP8	*Emericella nidulans* R20 (anamorph: *Aspergillus nidulans*)	taxon: 162425
Eni-FGSCA4 PRP8	*Emericella nidulans* (anamorph: *Aspergillus nidulans*) FGSC A4	Filamentous fungus, taxon: 162425
Fte RPB2 (RpoB)	*Floydiella terrestris*, strain UTEX 1709	Green alga, chloroplast gene, taxon: 51328
Gth DnaB	*Guillardia theta* (plastid)	Cryptophyte Algae
HaV01 Pol	*Heterosigma akashiwo* virus 01	Algal virus, taxon: 97195, strain HaV01
Hca PRP8	*Histoplasma capsulatum* (anamorph: *Ajellomyces capsulatus*)	Fungi, human pathogen
IIV6 RIR1	Invertebrate iridescent virus 6	dsDNA eucaryotic virus, taxon: 176652
Kex-CBS379 VMA	*Kazachstania exigua*, formerly *Saccharomyces exiguus*, strain CBS379	Yeast, taxon: 34358
Kla-CBS683 VMA	*Kluyveromyces lactis*, strain CBS683	Yeast, taxon: 28985
Kla-IFO1267 VMA	*Kluyveromyces lactis* IFO1267	Fungi, taxon: 28985
Kla-NRRLY1140 VMA	*Kluyveromyces lactis* NRRL Y-1140	Fungi, taxon: 284590
Lel VMA	*Lodderomyces elongisporus*	Yeast
Mca-CBS113480 PRP8	*Microsporum canis* CBS 113480	Taxon: 554155
Nau PRP8	*Neosartorya aurata* NRRL 4378	Fungus, taxon: 41051

Nfe-NRRL5534 PRP8	*Neosartorya fennelliae* NRRL 5534	Fungus, taxon: 41048
Nfi PRP8	*Neosartorya fischeri*	Fungi
Ngl-FR2163 PRP8	*Neosartorya glabra* FRR2163	Fungi, ATCC 16909, taxon: 41049
Ngl-FRR1833 PRP8	*Neosartorya glabra* FRR1833	Fungi, taxon: 41049, (preliminary identification)
Nqu PRP8	*Neosartorya quadricincta*, strain NRRL 4175	taxon: 41053
Nspi PRP8	*Neosartorya spinosa* FRR4595	Fungi, taxon: 36631
Pabr-Pb01 PRP8	*Paracoccidioides brasiliensis* Pb01	Taxon: 502779
Pabr-Pb03 PRP8	*Paracoccidioides brasiliensis* Pb03	Taxon: 482561
Pan CHS2	*Podospora anserina*	Fungi, Taxon 5145
Pan GLT1	*Podospora anserina*	Fungi, Taxon 5145
Pbl PRP8-a	*Phycomyces blakesleeanus*	Zygomycete fungus, strain NRRL155
Pbl PRP8-b	*Phycomyces blakesleeanus*	Zygomycete fungus, strain NRRL155
Pbr-Pb18 PRP8	*Paracoccidioides brasiliensis* Pb18	Fungi, taxon: 121759
Pch PRP8	*Penicillium chrysogenum*	Fungus, taxon: 5076
Pex PRP8	*Penicillium expansum*	Fungus, taxon27334
Pgu GLT1	*Pichia* (*Candida*) *guilliermondii*	Fungi, Taxon 294746
Pgu-alt GLT1	*Pichia* (*Candida*) *guilliermondii*	Fungi
Pno GLT1	*Phaeosphaeria nodorum* SN15	Fungi, taxon: 321614
Pno RPA2	*Phaeosphaeria nodorum* SN15	Fungi, taxon: 321614
Ppu DnaB	*Porphyra purpurea* (chloroplast)	Red Alga
Pst VMA	*Pichia stipitis* CBS 6054, taxon: 322104	Yeast
Ptr PRP8	*Pyrenophora tritici-repentis* Pt-1C-BF	Ascomycete fungus, taxon: 426418
Pvu PRP8	*Penicillium vulpinum* (formerly *P. claviforme*)	Fungus
Pye DnaB	*Porphyra yezoensis* chloroplast, cultivar U-51	Red alga, organelle = "plastid: chloroplast", "taxon: 2788

Sas RPB2	*Spiromyces aspiralis* NRRL 22631	Zygomycete fungus, isolate = "AFTOL-ID 185", taxon: 68401
Sca-CBS4309 VMA	*Saccharomyces castellii*, strain CBS4309	Yeast, taxon: 27288
Sca-IFO1992 VMA	*Saccharomyces castellii*, strain IFO1992	Yeast, taxon: 27288
Scar VMA	*Saccharomyces cariocanus*, strain = "UFRJ 50791	Yeast, taxon: 114526
Sce VMA	*Saccharomyces cerevisiae* (nucleus)	Yeast, also in Sce strains OUT7163, OUT7045, OUT7163, IFO1992
Sce-DH1-1A VMA	*Saccharomyces cerevisiae* strain DH1-1A	Yeast, taxon: 173900, also in Sce strains OUT7900, OUT7903, OUT7112
Sce-JAY291 VMA	*Saccharomyces cerevisiae* JAY291	Taxon: 574961
Sce-OUT7091 VMA	*Saccharomyces cerevisiae* OUT7091	Yeast, taxon: 4932, also in Sce strains OUT7043, OUT7064
Sce-OUT7112 VMA	*Saccharomyces cerevisiae* OUT7112	Yeast, taxon: 4932, also in Sce strains OUT7900, OUT7903
Sce-YJM789 VMA	*Saccharomyces cerevisiae* strain YJM789	Yeast, taxon: 307796
Sda VMA	*Saccharomyces dairenensis*, strain CBS 421	Yeast, taxon: 27289, Also in Sda strain IFO0211
Sex-IFO1128 VMA	*Saccharomyces exiguus*, strain = "IFO1128"	Yeast, taxon: 34358
She RPB2 (RpoB)	*Stigeoclonium helveticum*, strain UTEX 441	Green alga, chloroplast gene, taxon: 55999
Sja VMA	*Schizosaccharomyces japonicus* yFS275	Ascomycete fungus, taxon: 402676

Spa VMA	*Saccharomyces pastorianus* IFO11023	Yeast, taxon: 27292
Spu PRP8	*Spizellomyces punctatus*	Chytrid fungus,
Sun VMA	*Saccharomyces unisporus*, strain CBS 398	Yeast, taxon: 27294
Tgl VMA	*Torulaspora globosa*, strain CBS 764	Yeast, taxon: 48254
Tpr VMA	*Torulaspora pretoriensis*, strain CBS 5080	Yeast, taxon: 35629
Ure-1704 PRP8	*Uncinocarpus reesii*	Filamentous fungus
Vpo VMA	*Vanderwaltozyma polyspora*, formerly *Kluyveromyces polysporus*, strain CBS 2163	Yeast, taxon: 36033
WIV RIR1	*Wiseana* iridescent virus	dsDNA eucaryotic virus, taxon: 68347
Zba VMA	*Zygosaccharomyces bailii*, strain CBS 685	Yeast, taxon: 4954
Zbi VMA	*Zygosaccharomyces bisporus*, strain CBS 702	Yeast, taxon: 4957
Zro VMA	*Zygosaccharomyces rouxii*, strain CBS 688	Yeast, taxon: 4956
Eubacteria		
AP-APSE1 dpol	*Acyrthosiphon pisum* secondary endosymbiot phage 1	Bacteriophage, taxon: 67571
AP-APSE2 dpol	Bacteriophage APSE-2, isolate = T5A	Bacteriophage of *Candidatus Hamiltonella defensa*, endosymbiot of *Acyrthosiphon pisum*, taxon: 340054
AP-APSE4 dpol	Bacteriophage of *Candidatus Hamiltonella defensa* strain 5ATac, endosymbiot of *Acyrthosiphon pisum*	Bacteriophage, taxon: 568990
AP-APSE5 dpol	Bacteriophage APSE-5	Bacteriophage of *Candidatus Hamiltonella defensa*, endosymbiot of *Uroleucon rudbeckiae*, taxon: 568991

AP-Aaphi23 MupF	Bacteriophage Aaphi23, *Haemophilus* phage Aaphi23	*Actinobacillus actinomycetemcomitans* Bacteriophage, taxon: 230158
Aae RIR2	*Aquifex aeolicus* strain VF5	Thermophilic chemolithoautotroph, taxon: 63363
Aave-AAC001 Aave1721	*Acidovorax avenae* subsp. *citrulli* AAC00-1	taxon: 397945
Aave-AAC001 RIR1	*Acidovorax avenae* subsp. *citrulli* AAC00-1	taxon: 397945
Aave-ATCC19860 RIR1	*Acidovorax avenae* subsp. *avenae* ATCC 19860	Taxon: 643561
Aba Hyp-02185	*Acinetobacter baumannii* ACICU	taxon: 405416
Ace RIR1	*Acidothermus cellulolyticus* 11B	taxon: 351607
Aeh DnaB-1	*Alkalilimnicola ehrlichei* MLHE-1	taxon: 187272
Aeh DnaB-2	*Alkalilimnicola ehrlichei* MLHE-1	taxon: 187272
Aeh RIR1	*Alkalilimnicola ehrlichei* MLHE-1	taxon: 187272
AgP-S1249 MupF	*Aggregatibacter* phage S1249	Taxon: 683735
Aha DnaE-c	*Aphanothece halophytica*	Cyanobacterium, taxon: 72020
Aha DnaE-n	*Aphanothece halophytica*	Cyanobacterium, taxon: 72020
Alvi-DSM180 GyrA	*Allochromatium vinosum* DSM 180	Taxon: 572477
Ama MADE823	phage uncharacterized protein [*Alteromonas macleodii* 'Deep ecotype']	Probably prophage gene, taxon: 314275
Amax-CS328 DnaX	*Arthrospira maxima* CS-328	Taxon: 513049
Aov DnaE-c	*Aphanizomenon ovalisporum*	Cyanobacterium, taxon: 75695
Aov DnaE-n	*Aphanizomenon ovalisporum*	Cyanobacterium, taxon: 75695
Apl-C1 DnaX	*Arthrospira platensis*	Taxon: 118562, strain C1
Arsp-FB24 DnaB	*Arthrobacter* species FB24	taxon: 290399

Asp DnaE-c	*Anabaena* species PCC7120, (*Nostoc* sp. PCC7120)	Cyanobacterium, Nitrogen-fixing, taxon: 103690
Asp DnaE-n	*Anabaena* species PCC7120, (*Nostoc* sp. PCC7120)	Cyanobacterium, Nitrogen-fixing, taxon: 103690
Ava DnaE-c	*Anabaena variabilis* ATCC29413	Cyanobacterium, taxon: 240292
Ava DnaE-n	*Anabaena variabilis* ATCC29413	Cyanobacterium, taxon: 240292
Avin RIR1 BIL	*Azotobacter vinelandii*	taxon: 354
Bce-MCO3 DnaB	*Burkholderia cenocepacia* MC0-3	taxon: 406425
Bce-PC184 DnaB	*Burkholderia cenocepacia* PC184	taxon: 350702
Bse-MLS10 TerA	*Bacillus selenitireducens* MLS10	Probably prophage gene, Taxon: 439292
BsuP-M1918 RIR1	*B. subtilis* M1918 (prophage)	Prophage in *B. subtilis* M1918. taxon: 157928
BsuP-SPBc2 RIR1	*B. subtilis* strain 168 Sp beta c2 prophage	*B. subtilis* taxon 1423. SPbeta c2 phage, taxon: 66797
Bvi IcmO	*Burkholderia vietnamiensis* G4	plasmid = "pBVIE03". taxon: 269482
CP-P1201 Thy1	*Corynebacterium* phage P1201	lytic bacteriophage P1201 from *Corynebacterium glutamicum* NCHU 87078. Viruses; dsDNA viruses, taxon: 384848
Cag RIR1	*Chlorochromatium aggregatum*	Motile, phototrophic consortia
Cau SpoVR	*Chloroflexus aurantiacus* J-10-fl	Anoxygenic phototroph, taxon: 324602
CbP-C-St RNR	*Clostridium botulinum* phage C-St	Phage, specific_host = "*Clostridium botulinum* type C strain C-Stockholm, taxon: 12336
CbP-D1873 RNR	*Clostridium botulinum* phage D	Ssp. phage from *Clostridium botulinum* type D strain, 1873, taxon: 29342
Cbu-Dugway DnaB	*Coxiella burnetii* Dugway 5J108-111	Proteobacteria; Legionellales; taxon: 434922

Cbu-Goat DnaB	*Coxiella burnetii* 'MSU Goat Q177'	Proteobacteria; Legionellales; taxon: 360116
Cbu-RSA334 DnaB	*Coxiella burnetii* RSA 334	Proteobacteria; Legionellales; taxon: 360117
Cbu-RSA493 DnaB	*Coxiella burnetii* RSA 493	Proteobacteria; Legionellales; taxon: 227377
Cce Hyp1-Csp-2	*Cyanothece* sp. ATCC 51142	Marine unicellular diazotrophic cyanobacterium, taxon: 43989
Cch RIR1	*Chlorobium chlorochromatii* CaD3	taxon: 340177
Ccy Hyp1-Csp-1	*Cyanothece* sp. CCY0110	Cyanobacterium, taxon: 391612
Ccy Hyp1-Csp-2	*Cyanothece* sp. CCY0110	Cyanobacterium, taxon: 391612
Cfl-DSM20109 DnaB	*Cellulomonas flavigena* DSM 20109	Taxon: 446466
Chy RIR1	*Carboxydothermus hydrogenoformans* Z-2901	Thermophile, taxon = 246194
Ckl PTerm	*Clostridium kluyveri* DSM 555	plasmid = "pCKL555A", taxon: 431943
Cra-CS505 DnaE-c	*Cylindrospermopsis raciborskii* CS-505	Taxon: 533240
Cra-CS505 DnaE-n	*Cylindrospermopsis raciborskii* CS-505	Taxon: 533240
Cra-CS505 GyrB	*Cylindrospermopsis raciborskii* CS-505	Taxon: 533240
Csp-CCY0110 DnaE-c	*Cyanothece* sp. CCY0110	Taxon: 391612
Csp-CCY0110 DnaE-n	*Cyanothece* sp. CCY0110	Taxon: 391612
Csp-PCC7424 DnaE-c	*Cyanothece* sp. PCC 7424	Cyanobacterium, taxon: 65393
Csp-PCC7424 DnaE-n	*Cyanothece* sp. PCC7424	Cyanobacterium, taxon: 65393
Csp-PCC7425 DnaB	*Cyanothece* sp. PCC 7425	Taxon: 395961

Csp-PCC7822 DnaE-n	*Cyanothece* sp. PCC 7822	Taxon: 497965
Csp-PCC8801 DnaE-c	*Cyanothece* sp. PCC 8801	Taxon: 41431
Csp-PCC8801 DnaE-n	*Cyanothece* sp. PCC 8801	Taxon: 41431
Cth ATPase BIL	*Clostridium thermocellum*	ATCC27405, taxon: 203119
Cth-ATCC27405 TerA	*Clostridium thermocellum* ATCC27405	Probable prophage, ATCC27405, taxon: 203119
Cth-DSM2360 TerA	*Clostridium thermocellum* DSM 2360	Probably prophage gene, Taxon: 572545
Cwa DnaB	*Crocosphaera watsonii* WH 8501 (*Synechocystis* sp. WH 8501)	taxon: 165597
Cwa DnaE-c	*Crocosphaera watsonii* WH 8501 (*Synechocystis* sp. WH 8501)	Cyanobacterium, taxon: 165597
Cwa DnaE-n	*Crocosphaera watsonii* WH 8501 (*Synechocystis* sp. WH 8501)	Cyanobacterium, taxon: 165597
Cwa PEP	*Crocosphaera watsonii* WH 8501 (*Synechocystis* sp. WH 8501)	taxon: 165597
Cwa RIR1	*Crocosphaera watsonii* WH 8501 (*Synechocystis* sp. WH 8501)	taxon: 165597
Daud RIR1	*Candidatus Desulforudis audaxviator* MP104C	taxon: 477974
Dge DnaB	*Deinococcus geothermalis* DSM11300	Thermophilic, radiation resistant
Dha-DCB2 RIR1	*Desulfitobacterium hafniense* DCB-2	Anaerobic dehalogenating bacteria, taxon: 49338
Dha-Y51 RIR1	*Desulfitobacterium hafniense* Y51	Anaerobic dehalogenating bacteria, taxon: 138119
Dpr-MLMSI RIR1	delta proteobacterium MLMS-1	Taxon: 262489
Dra RIR1	*Deinococcus radiodurans* R1, TIGR strain	Radiation resistant, taxon: 1299

Dra Snf2-c	*Deinococcus radiodurans* R1, TIGR strain	Radiation and DNA damage resistent, taxon: 1299
Dra Snf2-n	*Deinococcus radiodurans* R1, TIGR strain	Radiation and DNA damage resistent, taxon: 1299
Dra-ATCC13939 Snf2	*Deinococcus radiodurans* R1, ATCC13939/Brooks & Murray strain	Radiation and DNA damage resistent, taxon: 1299
Dth UDP GD	*Dictyoglomus thermophilum* H-6-12	strain = "H-6-12; ATCC 35947, taxon: 309799
Dvul ParB	*Desulfovibrio vulgaris* subsp. *vulgaris* DP4	taxon: 391774
EP-Min27 Primase	Enterobacteria phage Min27	bacteriophage of host = "*Escherichia coli* O157: H7 str. Min27"
Fal DnaB	*Frankia alni* ACN14a	Plant symbiot, taxon: 326424
Fsp-Ccl3 RIR1	*Frankia* species Ccl3	taxon: 106370
Gob DnaE	*Gemmata obscuriglobus* UQM2246	Taxon 114, TIGR genome strain, budding bacteria
Gob Hyp	*Gemmata obscuriglobus* UQM2246	Taxon 114, TIGR genome strain, budding bacteria
Gvi DnaB	*Gloeobacter violaceus*, PCC 7421	taxon: 33072
Gvi RIR1-1	*Gloeobacter violaceus*, PCC 7421	taxon: 33072
Gvi RIR1-2	*Gloeobacter violaceus*, PCC 7421	taxon: 33072
Hhal DnaB	*Halorhodospira halophila* SL1	taxon: 349124
Kfl-DSM17836 DnaB	*Kribbella flavida* DSM 17836	Taxon: 479435
Kra DnaB	*Kineococcus radiotolerans* SRS30216	Radiation resistant
LLP-KSY1 PolA	*Lactococcus* phage KSY1	Bacteriophage, taxon: 388452
LP-phiHSIC Helicase	*Listonella pelagia* phage phiHSIC	taxon: 310539, a pseudotemperate marine phage of *Listonella pelagia*
Lsp-PCC8106 GyrB	*Lyngbya* sp. PCC 8106	Taxon: 313612

MP-Be DnaB	Mycobacteriophage Bethlehem	Bacteriophage, taxon: 260121
MP-Be gp51	Mycobacteriophage Bethlehem	Bacteriophage, taxon: 260121
MP-Catera gp206	Mycobacteriophage Catera	Mycobacteriophage, taxon: 373404
MP-KBG gp53	*Mycobacterium* phage KBG	Taxon: 540066
MP-Mcjw1 DnaB	Mycobacteriophage CJW1	Bacteriophage, taxon: 205869
MP-Omega DnaB	Mycobacteriophage Omega	Bacteriophage, taxon: 205879
MP-U2 gp50	Mycobacteriophage U2	Bacteriophage, taxon: 260120
Maer-NIES843 DnaB	*Microcystis aeruginosa* NIES-843	Bloom-forming toxic cyanobacterium, taxon: 449447
Maer-NIES843 DnaE-c	*Microcystis aeruginosa* NIES-843	Bloom-forming toxic cyanobacterium, taxon: 449447
Maer-NIES843 DnaE-n	*Microcystis aeruginosa* NIES-843	Bloom-forming toxic cyanobacterium, taxon: 449447
Mau-ATCC27029 GyrA	*Micromonospora aurantiaca* ATCC 27029	Taxon: 644283
Mav-104 DnaB	*Mycobacterium avium* 104	taxon: 243243
Mav-ATCC25291 DnaB	*Mycobacterium avium* subsp. *avium* ATCC 25291	Taxon: 553481
Mav-ATCC35712 DnaB	*Mycobacterium avium*	ATCC35712, taxon 1764
Mav-PT DnaB	*Mycobacterium avium* subsp. *paratuberculosis* str. k10	taxon: 262316
Mbo Pps1	*Mycobacterium bovis* subsp. *bovis* AF2122/97	strain = "AF2122/97", taxon: 233413
Mbo RecA	*Mycobacterium bovis* subsp. *bovis* AF2122/97	taxon: 233413
Mbo SufB (Mbo Pps1)	*Mycobacterium bovis* subsp. *bovis* AF2122/97	taxon: 233413

Mbo-1173P DnaB	*Mycobacterium bovis* BCG Pasteur 1173P	strain = BCG Pasteur 1173P2,, taxon: 410289
Mbo-AF2122 DnaB	*Mycobacterium bovis* subsp. *bovis* AF2122/97	strain = "AF2122/97", taxon: 233413
Mca MupF	*Methylococcus capsulatus* Bath, prophage MuMc02	prophage MuMc02, taxon: 243233
Mca RIR1	*Methylococcus capsulatus* Bath	taxon: 243233
Mch RecA	*Mycobacterium chitae*	IP14116003, taxon: 1792
Mcht-PCC7420 DnaE-1	*Microcoleus chthonoplastes* PCC7420	Cyanobacterium, taxon: 118168
Mcht-PCC7420 DnaE-2c	*Microcoleus chthonoplastes* PCC7420	Cyanobacterium, taxon: 118168
Mcht-PCC7420 DnaE-2n	*Microcoleus chthonoplastes* PCC7420	Cyanobacterium, taxon: 118168
Mcht-PCC7420 GyrB	*Microcoleus chthonoplastes* PCC 7420	Taxon: 118168
Mcht-PCC7420 RIR1-1	*Microcoleus chthonoplastes* PCC 7420	Taxon: 118168
Mcht-PCC7420 RIR1-2	*Microcoleus chthonoplastes* PCC 7420	Taxon: 118168
Mex Helicase	*Methylobacterium extorquens* AM1	Alphaproteobacteria
Mex TrbC	*Methylobacterium extorquens* AM1	Alphaproteobacteria
Mfa RecA	*Mycobacterium fallax*	CITP8139, taxon: 1793
Mfl GyrA	*Mycobacterium flavescens* Fla0	taxon: 1776, reference #930991
Mfl RecA	*Mycobacterium flavescens* Fla0	strain = Fla0, taxon: 1776, ref. #930991
Mfl-ATCC14474 RecA	*Mycobacterium flavescens*, ATCC14474	strain = ATCC14474, taxon: 1776, ref #930991

Mfl-PYR-GCK DnaB	*Mycobacterium flavescens* PYR-GCK	taxon: 350054
Mga GyrA	*Mycobacterium gastri*	HP4389, taxon: 1777
Mga RecA	*Mycobacterium gastri*	HP4389, taxon: 1777
Mga SufB (Mga Pps1)	*Mycobacterium gastri*	HP4389, taxon: 1777
Mgi-PYR-GCK DnaB	*Mycobacterium gilvum* PYR-GCK	taxon: 350054
Mgi-PYR-GCK GyrA	*Mycobacterium gilvum* PYR-GCK	taxon: 350054
Mgo GyrA	*Mycobacterium gordonae*	taxon: 1778, reference number 930835
Min-1442 DnaB	*Mycobacterium intracellulare*	strain 1442, taxon: 1767
Min-ATCC13950 GyrA	*Mycobacterium intracellulare* ATCC 13950	Taxon: 487521
Mkas GyrA	*Mycobacterium kansasii*	taxon: 1768
Mkas-ATCC12478 GyrA	*Mycobacterium kansasii* ATCC 12478	Taxon: 557599
Mle-Br4923 GyrA	*Mycobacterium leprae* Br4923	Taxon: 561304
Mle-TN DnaB	*Mycobacterium leprae*, strain TN	Human pathogen, taxon: 1769
Mle-TN GyrA	*Mycobacterium leprae* TN	Human pathogen, STRAIN = TN, taxon: 1769
Mle-TN RecA	*Mycobacterium leprae*, strain TN	Human pathogen, taxon: 1769
Mle-TN SufB (Mle Pps1)	*Mycobacterium leprae*	Human pathogen, taxon: 1769
Mma GyrA	*Mycobacterium malmoense*	taxon: 1780
Mmag Magn8951 BIL	*Magnetospirillum magnetotacticum* MS-1	Gram negative, taxon: 272627
Msh RecA	*Mycobacterium shimodei*	ATCC27962, taxon: 29313
Msm DnaB-1	*Mycobacterium smegmatis* MC2 155	MC2 155, taxon: 246196
Msm DnaB-2	*Mycobacterium smegmatis* MC2 155	MC2 155, taxon: 246196
Msp-KMS DnaB	*Mycobacterium* species KMS	taxon: 189918
Msp-KMS GyrA	*Mycobacterium* species KMS	taxon: 189918
Msp-MCS DnaB	*Mycobacterium* species MCS	taxon: 164756

Msp-MCS GyrA	*Mycobacterium* species MCS	taxon: 164756
Mthe RecA	*Mycobacterium thermoresistibile*	ATCC19527, taxon: 1797
Mtu SufB (Mtu Pps1)	*Mycobacterium tuberculosis* strains H37Rv & CDC1551	Human pathogen, taxon: 83332
Mtu-C RecA	*Mycobacterium tuberculosis* C	Taxon: 348776
Mtu-CDC1551 DnaB	*Mycobacterium tuberculosis*, CDC1551	Human pathogen, taxon: 83332
Mtu-CPHL RecA	*Mycobacterium tuberculosis* CPHL_A	Taxon: 611303
Mtu-Canetti RecA	*Mycobacterium tuberculosis*/ strain = "*Canetti*"	Taxon: 1773
Mtu-EAS054 RecA	*Mycobacterium tuberculosis* EAS054	Taxon: 520140
Mtu-F11 DnaB	*Mycobacterium tuberculosis*, strain F11	taxon: 336982
Mtu-H37Ra DnaB	*Mycobacterium tuberculosis* H37Ra	ATCC 25177, taxon: 419947
Mtu-H37Rv DnaB	*Mycobacterium tuberculosis* H37Rv	Human pathogen, taxon: 83332
Mtu-H37Rv RecA	*Mycobacterium tuberculosis* H37Rv, Also CDC1551	Human pathogen, taxon: 83332
Mtu-Haarlem DnaB	*Mycobacterium tuberculosis* str. Haarlem	Taxon: 395095
Mtu-K85 RecA	*Mycobacterium tuberculosis* K85	Taxon: 611304
Mtu-R604 RecA-n	*Mycobacterium tuberculosis* '98-R604 INH-RIF-EM'	Taxon: 555461
Mtu-So93 RecA	*Mycobacterium tuberculosis* So93/sub_species = "*Canetti*"	Human pathogen, taxon: 1773
Mtu-T17 RecA-c	*Mycobacterium tuberculosis* T17	Taxon: 537210
Mtu-T17 RecA-n	*Mycobacterium tuberculosis* T17	Taxon: 537210
Mtu-T46 RecA	*Mycobacterium tuberculosis* T46	Taxon: 611302
Mtu-T85 RecA	*Mycobacterium tuberculosis* T85	Taxon: 520141
Mtu-T92 RecA	*Mycobacterium tuberculosis* T92	Taxon: 515617

Mvan DnaB	*Mycobacterium vanbaalenii* PYR-1	taxon: 350058
Mvan GyrA	*Mycobacterium vanbaalenii* PYR-1	taxon: 350058
Mxa RAD25	*Myxococcus xanthus* DK1622	Deltaproteobacteria
Mxe GyrA	*Mycobacterium xenopi* strain IMM5024	taxon: 1789
Naz-0708 RIR1-1	*Nostoc azollae* 0708	Taxon: 551115
Naz-0708 RIR1-2	*Nostoc azollae* 0708	Taxon: 551115
Nfa DnaB	*Nocardia farcinica* IFM 10152	taxon: 247156
Nfa Nfa15250	*Nocardia farcinica* IFM 10152	taxon: 247156
Nfa RIR1	*Nocardia farcinica* IFM 10152	taxon: 247156
Nosp-CCY9414 DnaE-n	*Nodularia spumigena* CCY9414	Taxon: 313624
Npu DnaB	*Nostoc punctiforme*	Cyanobacterium, taxon: 63737
Npu GyrB	*Nostoc punctiforme*	Cyanobacterium, taxon: 63737
Npu-PCC73102 DnaE-c	*Nostoc punctiforme* PCC73102	Cyanobacterium, taxon: 63737, ATCC29133
Npu-PCC73102 DnaE-n	*Nostoc punctiforme* PCC73102	Cyanobacterium, taxon: 63737, ATCC29133
Nsp-JS614 DnaB	*Nocardioides* species JS614	taxon: 196162
Nsp-JS614 TOPRIM	*Nocardioides* species JS614	taxon: 196162
Nsp-PCC7120 DnaB	*Nostoc* species PCC7120, (*Anabaena* sp. PCC7120)	Cyanobacterium, Nitrogen-fixing, taxon: 103690
Nsp-PCC7120 DnaE-c	*Nostoc* species PCC7120, (*Anabaena* sp. PCC7120)	Cyanobacterium, Nitrogen-fixing, taxon: 103690
Nsp-PCC7120 DnaE-n	*Nostoc* species PCC7120, (*Anabaena* sp. PCC7120)	Cyanobacterium, Nitrogen-fixing, taxon: 103690
Nsp-PCC7120 RIR1	*Nostoc* species PCC7120, (*Anabaena* sp. PCC7120)	Cyanobacterium, Nitrogen-fixing, taxon: 103690
Oli DnaE-c	*Oscillatoria limnetica* str. 'Solar Lake'	Cyanobacterium, taxon: 262926

Oli DnaE-n	*Oscillatoria limnetica* str. 'Solar Lake'	Cyanobacterium, taxon: 262926
PP-PhiEL Helicase	*Pseudomonas aeruginosa* phage phiEL	Phage infects *Pseudomonas aeruginosa*, taxon: 273133
PP-PhiEL ORF11	*Pseudomonas aeruginosa* phage phiEL	phage infects *Pseudomonas aeruginosa*, taxon: 273133
PP-PhiEL ORF39	*Pseudomonas aeruginosa* phage phiEL	Phage infects *Pseudomonas aeruginosa*, taxon: 273133
PP-PhiEL ORF40	*Pseudomonas aeruginosa* phage phiEL	phage infects *Pseudomonas aeruginosa*, taxon: 273133
Pfl Fha BIL	*Pseudomonas fluorescens* Pf-5	Plant commensal organism, taxon: 220664
Plut RIR1	*Pelodictyon luteolum* DSM 273	Green sulfur bacteria, Taxon 319225
Pma-EXH1 GyrA	*Persephonella marina* EX-H1	Taxon: 123214
Pma-ExH1 DnaE	*Persephonella marina* EX-H1	Taxon: 123214
Pna RIR1	*Polaromonas naphthalenivorans* CJ2	taxon: 365044
Pnuc DnaB	*Polynucleobacter* sp. QLW-P1DMWA-1	taxon: 312153
Posp-JS666 DnaB	*Polaromonas* species JS666	taxon: 296591
Posp-JS666 RIR1	*Polaromonas* species JS666	taxon: 296591
Pssp-A1-1 Fha	*Pseudomonas* species A1-1	
Psy Fha	*Pseudomonas syringae* pv. tomato str. DC3000	Plant (tomato) pathogen, taxon: 223283
Rbr-D9 GyrB	*Raphidiopsis brookii* D9	Taxon: 533247
Rce RIR1	*Rhodospirillum centenum* SW	taxon: 414684, ATCC 51521
Rer-SK121 DnaB	*Rhodococcus erythropolis* SK121	Taxon: 596309
Rma DnaB	*Rhodothermus marinus*	Thermophile, taxon: 29549
Rma-DSM4252 DnaB	*Rhodothermus marinus* DSM 4252	Taxon: 518766
Rma-DSM4252 DnaE	*Rhodothermus marinus* DSM 4252	Thermophile, taxon: 518766
Rsp RIR1	*Roseovarius* species 217	taxon: 314264

SaP-SETP12 dpol	*Salmonella* phage SETP12	Phage, taxon: 424946
SaP-SETP3 Helicase	*Salmonella* phage SETP3	Phage, taxon: 424944
SaP-SETP3 dpol	*Salmonella* phage SETP3	Phage, taxon: 424944
SaP-SETP5 dpol	*Salmonella* phage SETP5	Phage, taxon: 424945
Sare DnaB	*Salinispora arenicola* CNS-205	taxon: 391037
Sav RecG Helicase	*Streptomyces avermitilis* MA-4680	taxon: 227882, ATCC 31267
Sel-PC6301 RIR1	*Synechococcus elongatus* PCC 6301	taxon: 269084 Berkely strain 6301~equivalent name: Ssp PCC 6301~synonym: *Anacystis nudulans*
Sel-PC7942 DnaE-c	*Synechococcus elongatus* PC7942	taxon: 1140
Sel-PC7942 DnaE-n	*Synechococcus elongatus* PC7942	taxon: 1140
Sel-PC7942 RIR1	*Synechococcus elongatus* PC7942	taxon: 1140
Sel-PCC6301 DnaE-c	*Synechococcus elongatus* PCC 6301 and PCC7942	Cyanobacterium, taxon: 269084, "Berkely strain 6301~equivalent name: *Synechococcus* sp. PCC 6301~synonym: *Anacystis nudulans*"
Sel-PCC6301 DnaE-n	*Synechococcus elongatus* PCC 6301	Cyanobacterium, taxon: 269084 "Berkely strain 6301~equivalent name: *Synechococcus* sp. PCC 6301~synonym: *Anacystis nudulans*"
Sep RIR1	*Staphylococcus epidermidis* RP62A	taxon: 176279
ShP-Sfv-2a-2457T-n Primase	*Shigella flexneri* 2a str. 2457T	Putative bacteriophage
ShP-Sfv-2a-301-n Primase	*Shigella flexneri* 2a str. 301	Putative bacteriophage
ShP-Sfv-5 Primase	*Shigella flexneri* 5 str. 8401	Bacteriphage, isolation_source_epidemic,

SoP-SO1 dpol	*Sodalis* phage SO-1	taxon: 373384 Phage/isolation_source = "*Sodalis glossinidius* strain GA-SG, secondary symbiont of *Glossina austeni* (Newstead)"
Spl DnaX	*Spirulina platensis*, strain C1	Cyanobacterium, taxon: 1156
Sru DnaB	*Salinibacter ruber* DSM 13855	taxon: 309807, strain = "DSM 13855; M31"
Sru PolBc	*Salinibacter ruber* DSM 13855	taxon: 309807, strain = "DSM 13855; M31"
Sru RIR1	*Salinibacter ruber* DSM 13855	taxon: 309807, strain = "DSM 13855; M31"
Ssp DnaB	*Synechocystis* species, strain PCC6803	Cyanobacterium, taxon: 1148
Ssp DnaE-c	*Synechocystis* species, strain PCC6803	Cyanobacterium, taxon: 1148
Ssp DnaE-n	*Synechocystis* species, strain PCC6803	Cyanobacterium, taxon: 1148
Ssp DnaX	*Synechocystis* species, strain PCC6803	Cyanobacterium, taxon: 1148
Ssp GyrB	*Synechocystis* species, strain PCC6803	Cyanobacterium, taxon: 1148
Ssp-JA2 DnaB	*Synechococcus* species JA-2-3B'a(2-13)	Cyanobacterium, Taxon: 321332
Ssp-JA2 RIR1	*Synechococcus* species JA-2-3B'a(2-13)	Cyanobacterium, Taxon: 321332
Ssp-JA3 DnaB	*Synechococcus* species JA-3-3Ab	Cyanobacterium, Taxon: 321327
Ssp-JA3 RIR1	*Synechococcus* species JA-3-3Ab	Cyanobacterium, Taxon: 321327
Ssp-PCC7002 DnaE-c	*Synechocystis* species, strain PCC 7002	Cyanobacterium, taxon: 32049
Ssp-PCC7002 DnaE-n	*Synechocystis* species, strain PCC 7002	Cyanobacterium, taxon: 32049
Ssp-PCC7335 RIR1	*Synechococcus* sp. PCC 7335	Taxon: 91464

StP-Twort ORF6	*Staphylococcus* phage Twort	Phage, taxon 55510
Susp-NBC371 DnaB intein	*Sulfurovum* sp. NBC37-1	taxon: 387093
Taq-Y51MC23 DnaE	*Thermus aquaticus* Y51MC23	Taxon: 498848
Taq-Y51MC23 RIR1	*Thermus aquaticus* Y51MC23	Taxon: 498848
Tcu-DSM43183 RecA	*Thermomonospora curvata* DSM 43183	Taxon: 471852
Tel DnaE-c	*Thermosynechococcus elongatus* BP-1	Cyanobacterium, taxon: 197221
Tel DnaE-n	*Thermosynechococcus elongatus* BP-1	Cyanobacterium,
Ter DnaB-1	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter DnaB-2	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter DnaE-1	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter DnaE-2	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter DnaE-3c	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter DnaE-3n	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter GyrB	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter Ndse-1	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter Ndse-2	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter RIR1-1	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter RIR1-2	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter RIR1-3	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter RIR1-4	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter Snf2	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Ter ThyX	*Trichodesmium erythraeum* IMS101	Cyanobacterium, taxon: 203124
Tfus RecA-1	*Thermobifida fusca* YX	Thermophile, taxon: 269800
Tfus RecA-2	*Thermobifida fusca* YX	Thermophile, taxon: 269800

Tfus Tfu2914	*Thermobifida fusca* YX	Thermophile, taxon: 269800
Thsp-K90 RIR1	*Thioalkalivibrio* sp. K90mix	Taxon: 396595
Tth-DSM571 RIR1	*Thermoanaerobacterium thermosaccharolyticum* DSM 571	Taxon: 580327
Tth-HB27 DnaE-1	*Thermus thermophilus* HB27	thermophile, taxon: 262724
Tth-HB27 DnaE-2	*Thermus thermophilus* HB27	thermophile, taxon: 262724
Tth-HB27 RIR1-1	*Thermus thermophilus* HB27	thermophile, taxon: 262724
Tth-HB27 RIR1-2	*Thermus thermophilus* HB27	thermophile, taxon: 262724
Tth-HB8 DnaE-1	*Thermus thermophilus* HB8	thermophile, taxon: 300852
Tth-HB8 DnaE-2	*Thermus thermophilus* HB8	thermophile, taxon: 300852
Tth-HB8 RIR1-1	*Thermus thermophilus* HB8	thermophile, taxon: 300852
Tth-HB8 RIR1-2	*Thermus thermophilus* HB8	thermophile, taxon: 300852
Tvu DnaE-c	*Thermosynechococcus vulcanus*	Cyanobacterium, taxon: 32053
Tvu DnaE-n	*Thermosynechococcus vulcanus*	Cyanobacterium, taxon: 32053
Tye RNR-1	*Thermodesulfovibrio yellowstonii* DSM 11347	taxon: 289376
Tye RNR-2	*Thermodesulfovibrio yellowstonii* DSM 11347	taxon: 289376
Archaea		
Ape APE0745	*Aeropyrum pernix* K1	Thermophile, taxon: 56636
Cme-boo Pol-II	*Candidatus Methanoregula boonei* 6A8	taxon: 456442
Fac-Fer1 RIR1	*Ferroplasma acidarmanus*, taxon: 97393 and taxon 261390	strain Fer1, eats iron
Fac-Fer1 SufB (Fac Pps1)	*Ferroplasma acidarmanus*	strain fer1, eats iron, taxon: 97393
Fac-TypeI RIR1	*Ferroplasma acidarmanus* type I,	Eats iron, taxon 261390
Fac-typeI SufB (Fac Pps1)	*Ferroplasma acidarmanus*	Eats iron, taxon: 261390

Hma CDC21	*Haloarcula marismortui* ATCC 43049	taxon: 272569,
Hma Pol-II	*Haloarcula marismortui* ATCC 43049	taxon: 272569,
Hma PolB	*Haloarcula marismortui* ATCC 43049	taxon: 272569,
Hma TopA	*Haloarcula marismortui* ATCC 43049	taxon: 272569
Hmu-DSM12286 MCM	*Halomicrobium mukohataei* DSM 12286	taxon: 485914 (Halobacteria)
Hmu-DSM12286 PolB	*Halomicrobium mukohataei* DSM 12286	Taxon: 485914
Hsa-R1 MCM	*Halobacterium salinarum* R-1	Halophile, taxon: 478009, strain = "R1; DSM 671"
Hsp-NRC1 CDC21	*Halobacterium* species NRC-1	Halophile, taxon: 64091
Hsp-NRC1 Pol-II	*Halobacterium salinarum* NRC-1	Halophile, taxon: 64091
Hut MCM-2	*Halorhabdus utahensis* DSM 12940	taxon: 519442
Hut-DSM12940 MCM-1	*Halorhabdus utahensis* DSM 12940	taxon: 519442
Hvo PolB	*Haloferax volcanii* DS70	taxon: 2246
Hwa GyrB	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa MCM-1	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa MCM-2	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa MCM-3	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa MCM-4	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 =

		HBSQ001
Hwa Pol-II-1	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa Pol-II-2	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa PolB-1	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa PolB-2	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa PolB-3	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa RCF	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa RIR1-1	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa RIR1-2	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa Top6B	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Hwa rPol A"	*Haloquadratum walsbyi* DSM 16790	Halophile, taxon: 362976, strain: DSM 16790 = HBSQ001
Maeo Pol-II	*Methanococcus aeolicus* Nankai-3	taxon: 419665
Maeo RFC	*Methanococcus aeolicus* Nankai-3	taxon: 419665
Maeo RNR	*Methanococcus aeolicus* Nankai-3	taxon: 419665
Maeo-N3 Helicase	*Methanococcus aeolicus* Nankai-3	taxon: 419665
Maeo-N3 RtcB	*Methanococcus aeolicus* Nankai-3	taxon: 419665

Maeo-N3 UDP GD	*Methanococcus aeolicus* Nankai-3	taxon: 419665
Mein-ME PEP	*Methanocaldococcus infernus* ME	thermophile, Taxon: 573063
Mein-ME RFC	*Methanocaldococcus infernus* ME	Taxon: 573063
Memar MCM2	*Methanoculleus marisnigri* JR1	taxon: 368407
Memar Pol-II	*Methanoculleus marisnigri* JR1	taxon: 368407
Mesp-FS406 PolB-1	*Methanocaldococcus* sp. FS406-22	Taxon: 644281
Mesp-FS406 PolB-2	*Methanocaldococcus* sp. FS406-22	Taxon: 644281
Mesp-FS406 PolB-3	*Methanocaldococcus* sp. FS406-22	Taxon: 644281
Mesp-FS406-22 LHR	*Methanocaldococcus* sp. FS406-22	Taxon: 644281
Mfe-AG86 Pol-1	*Methanocaldococcus fervens* AG86	Taxon: 573064
Mfe-AG86 Pol-2	*Methanocaldococcus fervens* AG86	Taxon: 573064
Mhu Pol-II	*Methanospirillum hungateii* JF-1	taxon 323259
Mja GF-6P	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja Helicase	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja Hyp-1	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja IF2	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja KlbA	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja PEP	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii*	Thermophile, DSM 2661, taxon: 2190

	DSM 2661)	
Mja Pol-1	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja Pol-2	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja RFC-1	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja RFC-2	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja RFC-3	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja RNR-1	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja RNR-2	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja RtcB (Mja Hyp-2)	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja TFIIB	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja UDP GD	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja r-Gyr	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mja rPol A'	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii*	Thermophile, DSM 2661, taxon: 2190

	DSM 2661)	
Mja rPol A"	*Methanococcus jannaschii* (*Methanocaldococcus jannaschii* DSM 2661)	Thermophile, DSM 2661, taxon: 2190
Mka CDC48	*Methanopyrus kandleri* AV19	Thermophile, taxon: 190192
Mka EF2	*Methanopyrus kandleri* AV19	Thermophile, taxon: 190192
Mka RFC	*Methanopyrus kandleri* AV19	Thermophile, taxon: 190192
Mka RtcB	*Methanopyrus kandleri* AV19	Thermophile, taxon: 190192
Mka VatB	*Methanopyrus kandleri* AV19	Thermophile, taxon: 190192
Mth RIR1	*Methanothermobacter thermautotrophicus* (*Methanobacterium thermoautotrophicum*)	Thermophile, delta H strain
Mvu-M7 Helicase	*Methanocaldococcus vulcanius* M7	Taxon: 579137
Mvu-M7 Pol-1	*Methanocaldococcus vulcanius* M7	Taxon: 579137
Mvu-M7 Pol-2	*Methanocaldococcus vulcanius* M7	Taxon: 579137
Mvu-M7 Pol-3	*Methanocaldococcus vulcanius* M7	Taxon: 579137
Mvu-M7 UDP GD	*Methanocaldococcus vulcanius* M7	Taxon: 579137
Neq Pol-c	*Nanoarchaeum equitans* Kin4-M	Thermophile, taxon: 228908
Neq Pol-n	*Nanoarchaeum equitans* Kin4-M	Thermophile, taxon: 228908
Nma-ATCC43099 MCM	*Natrialba magadii* ATCC 43099	Taxon: 547559
Nma-ATCC43099 PolB-1	*Natrialba magadii* ATCC 43099	Taxon: 547559
Nma-ATCC43099 PolB-2	*Natrialba magadii* ATCC 43099	Taxon: 547559
Nph CDC21	*Natronomonas pharaonis* DSM 2160	taxon: 348780
Nph PolB-1	*Natronomonas pharaonis* DSM 2160	taxon: 348780
Nph PolB-2	*Natronomonas pharaonis* DSM 2160	taxon: 348780
Nph rPol A"	*Natronomonas pharaonis* DSM 2160	taxon: 348780
Pab CDC21-1	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292

Pab CDC21-2	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab IF2	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab KlbA	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab Lon	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab Moaa	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab Pol-II	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab RFC-1	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab RFC-2	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab RIR1-1	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab RIR1-2	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab RIR1-3	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab RtcB (Pab Hyp-2)	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Pab VMA	*Pyrococcus abyssi*	Thermophile, strain Orsay, taxon: 29292
Par RIR1	*Pyrobaculum arsenaticum* DSM 13514	taxon: 340102
Pfu CDC21	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu IF2	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu KlbA	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu Lon	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu RFC	*Pyrococcus furiosus*	Thermophile, DSM3638, taxon: 186497
Pfu RIR1-1	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638

Pfu RIR1-2	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu RtcB (Pfu Hyp-2)	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu TopA	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pfu VMA	*Pyrococcus furiosus*	Thermophile, taxon: 186497, DSM3638
Pho CDC21-1	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho CDC21-2	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho IF2	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho KlbA	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho LHR	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho Lon	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho Pol I	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho Pol-II	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho RFC	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho RIR1	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho RadA	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho RtcB (Pho Hyp-2)	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho VMA	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Pho r-Gyr	*Pyrococcus horikoshii* OT3	Thermophile, taxon: 53953
Psp-GBD Pol	*Pyrococcus* species GB-D	Thermophile
Pto VMA	*Picrophilus torridus*, DSM 9790	DSM 9790, taxon: 263820, Thermoacidophile
Smar 1471	*Staphylothermus marinus* F1	taxon: 399550
Smar MCM2	*Staphylothermus marinus* F1	taxon: 399550
Tac-ATCC25905 VMA	*Thermoplasma acidophilum*, ATCC 25905	Thermophile, taxon: 2303
Tac-DSM1728 VMA	*Thermoplasma acidophilum*, DSM1728	Thermophile, taxon: 2303
Tag Pol-1 (Tsp-TY Pol-1)	*Thermococcus aggregans*	Thermophile, taxon: 110163
Tag Pol-2 (Tsp-TY Pol-2)	*Thermococcus aggregans*	Thermophile, taxon: 110163

Tag Pol-3 (Tsp-TY Pol-3)	*Thermococcus aggregans*	Thermophile, taxon: 110163
Tba Pol-II	*Thermococcus barophilus* MP	taxon: 391623
Tfu Pol-1	*Thermococcus fumicolans*	Thermophilem, taxon: 46540
Tfu Pol-2	*Thermococcus fumicolans*	Thermophile, taxon: 46540
Thy Pol-1	*Thermococcus hydrothermalis*	Thermophile, taxon: 46539
Thy Pol-2	*Thermococcus hydrothermalis*	Thermophile, taxon: 46539
Tko CDC21-1	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko CDC21-2	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko Helicase	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko IF2	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko KlbA	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko LHR	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko Pol-1 (Pko Pol-1)	*Pyrococcus/Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko Pol-2 (Pko Pol-2)	*Pyrococcus/Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko Pol-II	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko RFC	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko RIR1-1	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko RIR1-2	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko RadA	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko TopA	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tko r-Gyr	*Thermococcus kodakaraensis* KOD1	Thermophile, taxon: 69014
Tli Pol-1	*Thermococcus litoralis*	Thermophile, taxon: 2265
Tli Pol-2	*Thermococcus litoralis*	Thermophile, taxon: 2265
Tma Pol	*Thermococcus marinus*	taxon: 187879
Ton-NA1 LHR	*Thermococcus onnurineus* NA1	Taxon: 523850
Ton-NA1 Pol	*Thermococcus onnurineus* NA1	taxon: 342948
Tpe Pol	*Thermococcus peptonophilus* strain SM2	taxon: 32644
Tsi-MM739 Lon	*Thermococcus sibiricus* MM 739	Thermophile, Taxon: 604354

Tsi-MM739 Pol-1	*Thermococcus sibiricus* MM 739	Taxon: 604354
Tsi-MM739 Pol-2	*Thermococcus sibiricus* MM 739	Taxon: 604354
Tsi-MM739 RFC	*Thermococcus sibiricus* MM 739	Taxon: 604354
Tsp AM4 RtcB	*Thermococcus* sp. AM4	Taxon: 246969
Tsp-AM4 LHR	*Thermococcus* sp. AM4	Taxon: 246969
Tsp-AM4 Lon	*Thermococcus* sp. AM4	Taxon: 246969
Tsp-AM4 RIR1	*Thermococcus* sp. AM4	Taxon: 246969
Tsp-GE8 Pol-1	*Thermococcus* species GE8	Thermophile, taxon: 105583
Tsp-GE8 Pol-2	*Thermococcus* species GE8	Thermophile, taxon: 105583
Tsp-GT Pol-1	*Thermococcus* species GT	taxon: 370106
Tsp-GT Pol-2	*Thermococcus* species GT	taxon: 370106
Tsp-OGL-20P Pol	*Thermococcus* sp. OGL-20P	taxon: 277988
Tthi Pol	*Thermococcus thioreducens*	Hyperthermophile
Tvo VMA	*Thermoplasma volcanium* GSS1	Thermophile, taxon: 50339
Tzi Pol	*Thermococcus zilligii*	taxon: 54076
Unc-ERS PFL	uncultured archaeon Gzfos13E1	isolation_source = "Eel River sediment", clone = "GZfos13E1", taxon: 285397
Unc-ERS RIR1	uncultured archaeon GZfos9C4	isolation_source = "Eel River sediment", taxon: 285366, clone = "GZfos9C4"
Unc-ERS RNR	uncultured archaeon GZfos10C7	isolation_source = "Eel River sediment", clone = "GZfos10C7", taxon: 285400
Unc-MetRFS MCM2	uncultured archaeon (Rice Cluster I)	Enriched methanogenic consortium from rice field soil, taxon: 198240

[00152] In some embodiments, the coding sequence of the exogenous polypeptide can be split into three (or more) portions.

[00153] Split inteins can mediate the efficient post-translational splicing of two or more heterologous extein polypeptides. However, the resulting spliced product generally includes three to five amino acids

of intein sequence introduced at the junction of the spliced amino- and carboxy-terminal extein polypeptides. In some instances, these three to five "intein footprint" (or simply "footprint") amino acids do not appreciably affect the function of the final spliced polypeptide, but in others, the presence of such inserted amino acids can negatively impact the structure and function of the final product. As such, there can be a benefit to minimizing or even altogether avoiding an intein footprint in the trans-spliced protein product.

[00154] In one aspect, an intein footprint insert can be minimized or even completely avoided in the methods and compositions as described herein. To achieve this, one can, for example, analyze the sequence of a target protein relative to known split intein footprints to identify sequences within the target that match or closely approximate a split intein's footprint.

Table 4: Exemplary sequences to minimize split intein footprints

Split Intein ID	6A.A conventional footprint		Footprint tested	Minimal footprint with high protein trans-splicing rate
	N-ter	C-ter		
Split Aha, Split Sel	AEY	CFN	AEYCFN, AEYCF, AEYC, AEY EYCFN, EYCF, EYC, EY, YCFN, YCF, YC, Y CFN, CF, C	AEY (only for split Aha using GFP)
Split gp41.1	SGY	SSS	SGYSSS, SGYSS, SGYS, SGY GYSSS, GYSS, GYS, GY YSSS, YSS, YS, Y SSS, SS, S	GYS
Split IMPDH	GGG	SIC	GGGSIC, GGGSI, GGGS, GGG GGSIC, GGSI, GGS, GG GSIC, GSI, GS, G SIC, SI, S	GS
Nrdj-1	NPC	SEI	NPCSEI, NPCSE, NPCS, NPC PCSEI, PCSE, PCS, PC CSEI, CSE, CS, C SEI, SE, S	CS

Amino-acids abbreviation:

A: Alanine
C: Cysteine
E: Glutamic Acid
F: Phenylalanine
G: Glycine
I: Isoleucine
N: Asparagine
P: Proline
S: Serine
Y: Tyrosine

[00155] One can then use such a split intein that has a footprint naturally occurring in a given target protein to design the heterologous extein-intein fusions to be separately expressed, and thereby minimize or even avoid the insertion of non-naturally occurring amino acids in the spliced polypeptide product. For example,

after screening the target protein sequence for sequences that match split intein footprint sequence, one can prepare sequences encoding amino- and carboxy-terminal fusions of the target polypeptide fragments to the respective amino and carboxy-terminal spit intein fragments in which the footprint amino acids are omitted from the extein fusion polypeptide sequences. In this situation, upon cleavage and joinder of the extein sequences, the intein footprint insert reconstitutes the native target polypeptide sequence, resulting in a spliced target polypeptide that does not differ in amino acid sequence from the natural target polypeptide. That is, while there is technically still a footprint insert characteristic of that split intein, its sequence matches sequence occurring in the target protein, such that there is no non-native footprint in the resulting spliced polypeptide product.

[00156] In some instances, a given target polypeptide may lack an exact match to a split intein footprint, or an exact match may be located so close to the amino or carboxy terminus of the target protein that splitting the sequence encoding the target protein at that point does not divide the target protein coding sequence into fragments that will each fit into a delivery vector. In such instances, it can still be beneficial to identify sequences within the target protein that are similar, but not identical to a split intein footprint sequence. Such similarity can be, for example, matching four out of five footprint amino acids, three out of five footprint amino acids, or even two out of five footprint amino acids. Similarity in this context can also include, for example, the inclusion of amino acids with similar properties to those in the footprint, e.g., amino acids that are conservative substitutions for the naturally-occurring amino acids, or a combination of matches and conservative substitutions. Used analogously to the situation in which an exact match to an intein footprint can be identified in a beneficial location in a target protein, such an approach based on footprint similarity can minimize the intein footprint and/or its impact on function of the spliced target protein. Thus, a spliced product with an intein footprint of four or fewer differences, three or fewer, two or fewer, one or fewer, or no differences relative to the naturally occurring or desired target protein sequence can be generated as described herein.

[00157] In some embodiments, an "engineered" split intein differs from a naturally occurring polypeptide or nucleic acid by one or more amino acid or nucleic acid deletions, additions, substitutions or side-chain modifications, yet retains one or more specific functions or biological activities of the naturally occurring split intein sequence. Amino acid substitutions include alterations in which an amino acid is replaced with a different naturally-occurring or a non-conventional amino acid residue. Some substitutions can be classified as "conservative," in which case an amino acid residue contained in a polypeptide is replaced with another naturally occurring amino acid of similar character either in relation to polarity, side chain functionality or size. Substitutions encompassed by variants as described herein can also be "non-conservative," in which an amino acid residue which is present in a peptide is substituted with an amino acid having different properties (e.g., substituting a charged or hydrophobic amino acid with an uncharged or hydrophilic amino acid), or alternatively, in which a naturally-occurring amino acid is substituted with a non-conventional amino acid.

[00158] In one embodiment, the split intein comprises at least two of SEQ ID Nos: 1-46. In another embodiment, the split intein comprises SEQ ID NO: 1 and SEQ ID NO: 2. In another embodiment, the split intein comprises SEQ ID NO: 3 and SEQ ID NO: 4. In another embodiment, the split intein comprises SEQ ID NO: 5 and SEQ ID NO: 6. In another embodiment, the split intein comprises SEQ ID NO: 7 and SEQ ID NO: 8. In another embodiment, the split intein comprises SEQ ID NO: 9 and SEQ ID NO: 10. In another embodiment, the split intein comprises SEQ ID NO: 11 and SEQ ID NO: 12. In another embodiment, the split intein comprises SEQ ID NO: 13 and SEQ ID NO: 14. In another embodiment, the split intein comprises SEQ ID NO: 15 and SEQ ID NO: 16. In another embodiment, the split intein comprises SEQ ID NO: 17 and SEQ ID NO: 18. In another embodiment, the split intein comprises SEQ ID NO: 19 and SEQ ID NO: 20. In another embodiment, the split intein comprises SEQ ID NO: 21 and SEQ ID NO: 22. In another embodiment, the split intein comprises SEQ ID NO: 23 and SEQ ID NO: 24. In another embodiment, the split intein comprises SEQ ID NO: 25 and SEQ ID NO: 26. In another embodiment, the split intein comprises SEQ ID NO: 27 and SEQ ID NO: 28. In another embodiment, the split intein comprises SEQ ID NO: 29 and SEQ ID NO: 30. In another embodiment, the split intein comprises SEQ ID NO: 31 and SEQ ID NO: 32. In another embodiment, the split intein comprises SEQ ID NO: 33 and SEQ ID NO: 34. In another embodiment, the split intein comprises SEQ ID NO: 35 and SEQ ID NO: 36. In another embodiment, the split intein comprises SEQ ID NO: 37 and SEQ ID NO: 38. In another embodiment, the split intein comprises SEQ ID NO: 39 and SEQ ID NO: 40. In another embodiment, the split intein comprises SEQ ID NO: 41 and SEQ ID NO: 42. In another embodiment, the split intein comprises SEQ ID NO: 43 and SEQ ID NO: 44. In another embodiment, the split intein comprises SEQ ID NO: 45 and SEQ ID NO: 46.

[00159] Exemplary sequences for specific dystrophin, dysferlin, utrophin, and mini-dystrophin fragments with corresponding inteins are provided herein in Example 2.

MSECs

[00160] In certain embodiments, the split intein constructs described herein can benefit from cell-type-specific expression. Such a design can ensure expression, including high level, moderate level or low level or regulated expression of the target protein not only where it is most needed, but also avoid or limit potential negative impact of ectopic expression in non-target cells or tissues. Inclusion of a tissue-specific expression cassette can thus maximize therapeutic benefit of transgene introduction. Such a design can also, for example, facilitate or permit systemic administration of vectors, in that while infection may occur in non-target cells or tissues, expression of the transgene polypeptide(s) will substantially only occur in the desired cell or tissue type. When used in combination with, for example, a vector that has a tropism or enhanced tropism for transduction of a given tissue or cell type, the use of a tissue specific expression cassette to drive expression of each target protein-split intein construct as described herein can be highly beneficial. When used in the context of delivery of two or more vectors, multiple tissue specific expression

cassettes can be used to generate balanced ratios of, for example, mRNA production or accumulation, or protein translation, production or accumulation.

[00161] A "tissue-specific expression cassette," as the term is used herein, provides expression of a target protein in a manner restricted to a particular tissue or cell type. By "restricted to" or "in a restricted manner" in this context is meant that expression from the construct is at least 5-fold higher in the target tissue or cell type than in other tissues or cell types, e.g., at least 5-fold higher, 10-fold higher, 15-fold higher, 20-fold higher or more. Expression can be measured at the level of, for example, mRNA production or accumulation, or at the level of protein translation, production or accumulation. In one embodiment, a tissue-specific expression cassette is a "muscle-specific expression cassette," or "MSEC" as described herein. An MSEC will drive expression of a linked construct in a muscle cell- or muscle tissue-restricted manner as that term is defined herein above.

[00162] MSECs generally include elements of muscle-specific promoters and enhancers. See, for example, Salva et al., Molecular Therapy 15: 320-329 (2007), which is incorporated herein by reference, for examples and discussion of muscle-specific expression cassettes designed for use in rAAV vectors to drive heterologous protein expression in skeletal and cardiac muscle. Muscle-specific expression cassettes include, for example, promoter and enhancer sequence elements derived from muscle-specific genes including muscle creatine kinase (MCK), skeletal α-actin and α-myosin heavy-chain genes, among others. The murine MCK gene includes a 206 bp enhancer located approximately 1.2 kb upstream of the transcription start site, and a 358 bp proximal promoter. However, for use in gene therapy vectors such as AAV, the viral packaging limits as discussed herein require that regulatory elements designed to drive muscle-specific expression be kept to a minimum (about 800 bp or less) in order to maximize the amount of payload protein coding sequence for a given vector. Thus, muscle-specific expression cassettes useful in the methods and compositions described herein are comprised of truncated/modified muscle-specific regulatory elements that provide binding sites for myogenic regulatory factors, as well as Inr (initiator element) and/or TATA box sequences, and can include, for example, additional sequences from the 5' untranslated region of muscle-specific genes. The MHCK7 cassette described by Salva et al. is but one example of an MSEC useful in the methods and compositions described herein. That cassette drives expression to a higher degree than the constitutively active CMV promoter in MM14 myocytes, but is essentially inactive in non-muscle cells (e.g., HEK 293 fibroblasts, murine L cell fibroblasts, and JAWSII dendritic cells). See also the expression cassettes described in U.S. 10,479,821, which is incorporated herein by reference. As but one example, SEQ ID NO: 19 described therein and referred to as CK8, is highly active in cardiac and skeletal muscle. It is contemplated that variants of such MSEC sequences can also provide highly active, muscle-specific expression of therapeutic transgenes. For example, a sequence having at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or greater identity to such MSECs can also be of use in the methods and compositions described herein. One of skill in the art can determine the

activity of a given MSEC in muscle cells or tissue, e.g., using assays as described in the Salva et al. publication.

Pharmaceutical Compositions

[00163] Provided herein are vector compositions that are useful for treating or preventing a variety of different diseases and/or disorders in a subject. An important subset of disease and disorders is muscle diseases and disorders. In one embodiment, the composition is a pharmaceutical composition. The composition can comprise a therapeutically or prophylactically effective amount of at least two vectors encoding an exogenous polynucleotide or therapeutic agent. The at least two vectors utilize split inteins to aid in delivery of large protein-encoding nucleic acids to a given cell.

[00164] The composition can optionally include a carrier, such as a pharmaceutically acceptable carrier. Pharmaceutically acceptable carriers are determined in part by the particular composition being administered, as well as by the particular method used to administer the composition. Accordingly, there is a wide variety of suitable formulations of pharmaceutical compositions. Formulations suitable for parenteral administration can be formulated, for example, for intravenous, intramuscular, intradermal, intraperitoneal, and subcutaneous routes. Carriers can include aqueous isotonic sterile injection solutions, which can contain antioxidants, buffers, bacteriostats, and solutes that render the formulation isotonic with the blood of the intended recipient, and aqueous and non-aqueous sterile suspensions that can include suspending agents, solubilizers, thickening agents, stabilizers, preservatives, liposomes, microspheres and emulsions.

[00165] In one embodiment, the composition is formulated for intramuscular delivery.

[00166] Therapeutic compositions contain a physiologically tolerable carrier together with the vectors described herein, dissolved or dispersed therein as an active ingredient. As used herein, the terms "pharmaceutically acceptable", "physiologically tolerable" and grammatical variations thereof, as they refer to compositions, carriers, diluents and reagents, are used interchangeably and represent that the materials are capable of administration to or upon a mammal without the production of undesirable physiological effects such as nausea, dizziness, gastric upset and the like. A pharmaceutically acceptable carrier will not promote the raising of an immune response to an agent with which it is admixed, unless so desired. The preparation of a pharmaceutical composition that contains active ingredients dissolved or dispersed therein is understood in the art and need not be limited based on formulation. Typically, such compositions are prepared as injectable either as liquid solutions or suspensions; however, solid forms suitable for solution, or suspension in liquid prior to use can also be prepared. The preparation can also be emulsified or presented as a liposome composition. The active ingredient can be mixed with excipients which are pharmaceutically acceptable and compatible with the active ingredient and in amounts suitable for use in the therapeutic methods described herein. Suitable excipients include, for example, water, saline,

dextrose, glycerol, ethanol or the like and combinations thereof. In addition, if desired, the composition can contain minor amounts of auxiliary substances such as wetting or emulsifying agents, pH buffering agents and the like which enhance the effectiveness of the active ingredient. The therapeutic composition for use with the methods described herein can include pharmaceutically acceptable salts of the components therein. Pharmaceutically acceptable salts include the acid addition salts (formed with the free amino groups of the polypeptide) that are formed with inorganic acids such as, for example, hydrochloric or phosphoric acids, or such organic acids as acetic, tartaric, mandelic and the like. Salts formed with the free carboxyl groups can also be derived from inorganic bases such as, for example, sodium, potassium, ammonium, calcium or ferric hydroxides, and such organic bases as isopropylamine, trimethylamine, 2-ethylamino ethanol, histidine, procaine and the like. Physiologically tolerable carriers are well known in the art. Exemplary liquid carriers are sterile aqueous solutions that contain no materials in addition to the active ingredients and water, or contain a buffer such as sodium phosphate at physiological pH value, physiological saline or both, such as phosphate-buffered saline. Still further, aqueous carriers can contain more than one buffer salt, as well as salts such as sodium and potassium chlorides, dextrose, polyethylene glycol and other solutes. Liquid compositions can also contain liquid phases in addition to and to the exclusion of water. Examples of such additional liquid phases are glycerin, vegetable oils such as cottonseed oil, and water-oil emulsions. The amount of a vector to be administered herein that will be effective in the treatment of a particular disorder or condition will depend on the nature of the disorder or condition, the expression of the therapeutic agent, and can be determined by standard clinical techniques.

[00167] While any suitable carrier known to those of ordinary skill in the art can be employed in the pharmaceutical composition, the type of carrier will vary depending on the mode of administration. Compositions for use as described herein can be formulated for any appropriate manner of administration, including for example, topical, oral, nasal, intravenous, intracranial, intraperitoneal, subcutaneous or intramuscular administration. For parenteral administration, such as intramuscular or subcutaneous injection, the carrier preferably comprises water, saline, alcohol, a fat, a wax or a buffer. Alternatively, compositions as described herein can be formulated as a lyophilizate. Compounds can also be encapsulated within liposomes.

Dosage and Administration

[00168] Treatment using the methods and compositions described herein includes both prophylaxis/prevention of disease onset and therapy of an active disease. Prophylaxis or treatment can be accomplished by a single direct injection at a single time point or multiple time points. Administration can also be nearly simultaneous to multiple sites. Patients or subjects include mammals, such as human, bovine, equine, canine, feline, porcine, and ovine animals as well as other veterinary subjects. Preferably, the patients or subjects are human.

[00169] In one aspect, the methods described herein provide a method for treating a disease or disorder in a subject (e.g., a muscle disease or disorder). In one embodiment, the subject can be a mammal. In another embodiment, the mammal can be a human, although the approach is effective with respect to all mammals. The method comprises administering to the subject an effective amount of a pharmaceutical composition comprising vector as described herein in a pharmaceutically acceptable carrier.

[00170] The dosage range for the agent depends upon the potency, the expression level of the therapeutic protein and includes amounts large enough to produce the desired effect, *e.g.,* reduction in at least one symptom of the disease to be treated. The dosage should not be so large as to cause unacceptable adverse side effects. Generally, the dosage will vary with the type of exogenous protein expressed from the vector (*e.g.,* recombinant polypeptide, peptide, peptidomimetic, small molecule, etc.), the therapeutic protein characteristics (*e.g.,* dystrophin, utrophin, dysferlin, etc) and with the age, condition, and sex of the patient. The dosage can be determined by one of skill in the art and can also be adjusted by the individual physician in the event of any complication.

[00171] In some embodiments, the vectors are administered at a multiplicity of infection (MOI) of at least 5, at least 10, at least 20, at least 30, at least 40, at least 50, at least 100, at least 200, at least 500 or more.

[00172] In certain embodiments, the vectors are administered at a titer of at least 1x 10^5, 1 x 10^6, 1 x 10^7, 1 x 10^8, 1 x 10^9, 1 x 10^{10}, 1 x 10^{11}, 1 x 10^{12} viral particles or more.

[00173] Repeated administration can be performed as necessary to maintain therapeutic efficacy. As used herein, the term "therapeutically effective amount" refers to an amount of a vector or expressed therapeutic agent that is sufficient to produce a statistically significant, measurable change in at least one symptom of a disease (see "Efficacy Measurement" below). Alternatively, a therapeutically effective amount is an amount of a vector or expressed therapeutic protein that is sufficient to produce a statistically significant, measurable change in the expression level of a biomarker associated with the disease in the subject. Such effective amounts can be gauged in clinical trials as well as animal studies for a given agent.

[00174] The vector compositions can be administered directly to a particular site (*e.g.,* intramuscular injection, intravenous, into a specific organ) or can be administered orally. It is also contemplated herein that the agents can also be delivered intravenously (by bolus or continuous infusion), by inhalation, intranasally, intraperitoneally, intramuscularly, subcutaneously, intracavity, and can be delivered by peristaltic means, if desired, or by other means known by those skilled in the art. The agent can be administered systemically, if so desired.

[00175] Therapeutic compositions containing at least one agent can be conventionally administered in a unit dose. The term "unit dose" when used in reference to a therapeutic composition refers to physically discrete units suitable as unitary dosage for the subject, each unit containing a predetermined quantity of active material calculated to produce the desired therapeutic effect in association with the required physiologically acceptable diluent, i.e., carrier, or vehicle.

[00176] Precise amounts of active ingredient required to be administered depend on the judgment of the practitioner and are particular to each individual. However, suitable dosage ranges for systemic application are disclosed herein and depend on the route of administration. Suitable regimes for administration are also variable, but are typified by an initial administration followed by repeated doses at one or more intervals by a subsequent injection or other administration. Alternatively, continuous intravenous infusion sufficient to maintain concentrations in the blood in the ranges specified for *in vivo* therapies are contemplated.

Efficacy measurement

[00177] The efficacy of a given treatment for a disease can be determined by the skilled clinician. However, a treatment is considered "effective treatment," as the term is used herein, if any one or all of the signs or symptoms of the disease to be treated is/are altered in a beneficial manner, other clinically accepted symptoms or markers of disease are improved, or even ameliorated, *e.g.,* by at least 10% following treatment with a vector as described herein. Efficacy can also be measured by failure of an individual to worsen as assessed by stabilization of the disease, hospitalization or need for medical interventions (*i.e.,* progression of the disease is halted or at least slowed). Methods of measuring these indicators are known to those of skill in the art and/or described herein. Treatment includes any treatment of a disease in an individual or an animal (some non-limiting examples include a human, or a mammal) and includes: (1) inhibiting the disease, *e.g.,* arresting, or slowing progression of the disease; or (2) relieving the disease, *e.g.,* causing regression of symptoms; and (3) preventing or reducing the likelihood of the development of the disease or preventing secondary issues associated with the disease.

[00178] In some embodiments, efficacy of treatment of a muscle disease or disorder can be determined by assessing one or more parameters of muscle function including, but not limited to, specific force generation, mobility, spasticity, tension, stability etc. In some embodiments, clinical tests for determining an improvement in muscle function, such as electromyography, magnetic resonance imaging (MRI) or muscle biopsies, can be used to assess efficacy of a method of treatment as described herein.

[00179] It is understood that the foregoing description and the following examples are illustrative only and are not to be taken as limitations upon the scope of the invention. Various changes and modifications to the disclosed embodiments, which will be apparent to those of skill in the art, may be made without departing from the spirit and scope of the present invention. Further, all patents, patent applications, and publications identified are expressly incorporated herein by reference for the purpose of describing and disclosing, for example, the methodologies described in such publications that might be used in connection with the present invention. These publications are provided solely for their disclosure prior to the filing date of the present application. Nothing in this regard should be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention or for any other reason. All statements as to the date or representation as to the contents of these documents are

based on the information available to the applicants and do not constitute any admission as to the correctness of the dates or contents of these documents.

EXAMPLES

[00180] The following provides non-limiting Examples demonstrating and supporting the technology as described herein.

EXAMPLE 1

Summary

[00181] Duchenne muscular dystrophy (DMD) is among the most common human genetic disorders, affecting approximately 1 in 5,000 newborn males (Emery 2003). It results from genetic mutations in the *DMD* gene that prevent expression of functional dystrophin (Monaco 1985, Kunkel 1986), which is one of the largest proteins made by the human cells. Adeno-associated viral (AAV) vector-based gene delivery has been actively used to treat DMD (Crudele 2019). However, the main limitation associated with the delivery of the *DMD* gene is its large coding sequences (11 kb) (Koenig 1989, Chamberlain 1989), while the maximum AAV cargo capacity is less than 5 kb (Srivastava 1983). Previous work by the inventors included the development of miniaturized 'micro-dystrophin' (μDys) that can be packaged within a single AAV vector (Harper 2002, Gregorevic 2006). These truncated, albeit functional, dystrophins have been shown to improve the skeletal muscle force and histology when administrated to dystrophin-deficient animal models despite missing critical domains of dystrophin required for its mechanical and signaling functions and the impossibility of transporting them all in one AAV vector.

[00182] Gene therapies using AAV vectors hold a promising future for treating different loss-of-function genetic disorders (Li 2020). However, this therapeutic modality has been challenged by the small packaging capacity of this viral vector (~5 kb). Due to the large size of the *DMD* coding sequences (~11.2 kb), the current therapeutic approaches in development for DMD using AAV vectors aim to express either a miniaturized μDys or genetic elements that restore the open-reading frame (CRISPR/Cas9-mediated gene editing or exon-skipping induced by U7 small nuclear RNA) (Harper 2002, Goyenvalle 2004, Long 2016, Nelson 2016, Bengtsson 2017). In both strategies, smaller than normal dystrophins are produced. In addition, very low expression and unstable secondary structures are observed when one or multiple exons are skipped using CRISPR/Cas9 or U7 tools (most likely because of improperly 'phased' repeats; Harper 2002). Previous work has shown the feasibility of reconstituting larger mini- or full-length dystrophins *via* homologous recombination and following the administration of two or three vectors (**FIG. 2**) (Odom 2011, Koo 2014, Lostal 2014). However, these events were rare and, in some cases, led to unwanted and potentially toxic products that would jeopardize their clinical use.

[00183] Provided herein is a novel therapeutic approach that can deliver larger, possibly up to full-length, dystrophins with high expression and in a precise manner. This method is not limited by unwanted recombination products, and it can be adapted to clinical use for any patient with Duchenne or Becker muscular dystrophies (BMD). This improved strategy allows for the expression of large and stable proteins with high specificity and efficiency (SIMPLI-GT (**S**plit **I**ntein-**M**ediated **P**rotein **Li**gation for **G**ene **T**herapy). This approach takes advantage of the intrinsic ability of split inteins to mediate protein trans-splicing, and therefore to reconstitute larger therapeutic constructs, which extends the usage of AAV-based gene replacement approach to any gene exceeding the maximum cargo capacity of AAV vectors.

Study

[00184] Gene replacement therapies using AAV vectors hold a great promise for treating genetic disorders caused by loss-of-function mutations. Currently, hundreds of primate serotypes have been isolated and more serotypes are in development (Li 2020). The inventors have previously shown that AAV can be used for systemic gene delivery to both cardiac and skeletal muscles at high efficiency (Gregorevic 2004). Numerous studies have also demonstrated very stable expression lasting up to 8 years in different animal models (Rivera 2005, Niemeyer 2009). In the clinic, several patients affected by neuromuscular disorders such as myotubular myopathy or spinal muscular atrophy have been recently treated with a single dose of AAV vectors to replace the defective gene, *MTM* or *SMN* respectively (NCT03199469, NCT02122952) and the clinical data have shown very promising physiological outcomes with the lack of cellular immunogenicity to both transgene and vector.

[00185] However, one of the main drawbacks of AAV vectors is their limited packaging capacity (~5 kb), which excludes many genetic disorders from using these vectors as a gene transporter. Due to the large coding sequences of the defective gene in muscular dystrophies like Duchenne or limb-girdle type 2B, a single AAV vector cannot be used to deliver the *DMD,* or *DYSF* genes respectively, to the affected muscles. For DMD, a series of miniaturized μDys were previously developed that can be delivered by single AAV vector (**FIG. 1**) (Harper 2002, Gregorevic 2006, Banks 2010, Ramos 2019). These truncated, albeit functional, dystrophins have been shown to improve the skeletal muscle force and histology when administrated to dystrophin-deficient animal models, and three constructs are currently being tested in human clinical trials, by Sarepta and Solid Biosciences (NCT03375164, NCT03368742).

[00186] It is noteworthy that most of the μDys developed to date contain 4 to 6 Spectrin-repeats (SR) with 2 or 3 Hinges, plus the N-terminal actin-binding domain (ABD) and the dystroglycan-binding domain (DgBD)(Ramos, 2019). Thousands of combinations are possible to generate μDys from the 24 SR and 4 Hinges domains that originally constitute the full-length dystrophin. This high number of possible combinations together with the limited AAV carrying capacities make it challenging to predict which μDys would have the best physiological outcomes. Indeed, previous studies revealed that many μDys are not functional and the best constructs cloned so far exhibit incomplete cardiac and skeletal rescue (Wasla 2018,

Ramos 2019). This suggests that the therapeutic candidate should include more domains to stabilize the dystrophin structure and provide additional protein functions.

[00187] On the other hand, two patients carrying a deletion of 46% (Δexon17-48) of the *DMD* gene were reported to express a truncated but highly functional dystrophin lacking most of the region between hinge 2 and the middle of SR19, and present a very mild phenotype with normal life span (England 1990), which inspired the development of larger 'mini-dystrophin' (mini-Dys) constructs (Phelps 1995, Harper, 2002, Odom 2011). Simultaneous administration of two AAV vectors that express two halves of a mini-Dys (ΔH2-SR19) flanked by a short region of homology allows the recombination of AAV genomes, which leads to the reconstitution the mini-Dys (**FIG. 2**). This highly functional protein showed a great therapeutic potential when tested in *mdx*^{4cv} mouse model with improvements in muscle histology and physiological performances. Nonetheless, unwanted products, obtained most likely from unspecific annealing of single-stranded genomes and the formation of concatamers, were detected by Southern blot analysis and by polymerase chain reaction. These unspecific products could potentially be toxic and trigger an immunoreaction which prevents its clinical use.

[00188] In the present study, the inventors describe a novel method named SIMPLI-GT (abbreviation of Split Intein-Mediated Protein Ligation for Gene Therapy), which aims to reconstitute larger therapeutic constructs using the protein trans-splicing mechanism mediated by inteins. This approach can overcome the main hurdle related to AAV-based gene replacement and extends its applications to numerous genetic disorders caused by loss-of-function in large genes.

[00189] Inteins are genetic elements that are found in unicellular organisms. They are embedded within essential genes that are involved in DNA transcription, replication and maintenance (e.g. DNA or RNA polymerase subunits, helicases, gyrases, and ribonucleotide reductase) or in other housekeeping genes including essential proteases and metabolic enzymes (Shah 2014). Following their in-frame transcription and translation with the host gene, the intein polypeptides (size varies between 138 to 844 amino acids) are self-excised from the precursor protein (also called extein) and join the adjacent peptides. This post-translational modification, known as protein splicing, does not require energy supply, cofactors or exogenous protease intervention. Over 600 inteins have been identified to date and around 30 have the particularity to be encoded by two separate genes. Unlike the more common contiguous inteins, these split inteins are transcribed and translated separately in N- and C-intein fragments. Then, they associate and form one reconstituted complex (N-extein/N-intein/C-intein/C-extein) before spontaneous splicing of the intein, resulting in reconstituted and fully functional extein (host protein) (**FIG. 3**).

[00190] This protein trans-splicing mechanism is used in biotechnological applications including protein purification and labeling steps (Li 2015). The inventors propose to utilize split inteins to reconstitute larger proteins that cannot be delivered by a single AAV vector due to its packaging limitation. Therefore, they have generated a library of 23 split inteins in order to screen for their ability to reconstitute two polypeptide fragments into one functional protein. This pre-screening will be performed using the green fluorescent

protein (GFP) as screening platform, which will permit testing of several inteins under the same conditions and in an unbiased and reliable manner. GFP is a widely used protein that has revolutionized different biology fields due to its small size (238 amino acids), easiness, specificity and lack of cell toxicity. It was previously adapted as a scaffold to screen aptamers and small anti-bacterial peptides (Abedi 1988, Soundrarajan 2016).

[00191] To begin, the inventors identified a splitting site in the GFP protein sequence where N- and C-terminal inteins can be inserted. In a preliminary test, two plasmids were cloned that encode either the N- or the C-terminal half of GFP fused to the N- or the C-terminal half of Npu intein (one of the most studied intein, which is found in nostoc punctiforme cyanobacterium). Then, human embryonic kidney 293 (HEK293) cells were co-transfected with both N- and/or C-terminal GFP/intein plasmids. 24 hours later, GFP fluorescence was detected only in cells transfected with either WT GFP (full-length GFP expressed from one plasmid) or dual split GFP/intein plasmids but not with the single N- or C-terminal plasmid (**FIG. 4A**). These data indicate that GFP was efficiently reassembled through protein trans-splicing mediated by Npu intein. The GFP fluorescence intensity was measured in living cells using a spectrophotometer. It was found that the GFP signal from the reconstituted protein was lower than that from WT GFP (**FIG. 4B**). This could be due to the short transfection period and the time required by the different steps prior to the formation of functional GFP (i.e., transcription, translation, fusion of N- and C-terminal fragments, intein splicing and ligation of the two N- and C-terminal GFP). In addition, the inventors observed other inteins that are currently being screened have a higher activity. The initial test showed the feasibility of using split GFP to test the ability of inteins to reconstitute a full-length GFP that is able to emit a measurable fluorescence. The inventors continued to screen the intein library using this accurate system, and results from a variety of additional split inteins is shown in **FIG. 7**.

[00192] Next, the ability of inteins to join two halves of mini-Dys was tested. With the small size of the split inteins, a dual AAV vector approach allows the expression of the largest mini-Dys construct tested so far. *In silico* modeling showed that two AAV vectors can transport the coding sequences of two dystrophin fragments: an N-terminal clone encoding proteins from N-terminus to the end of the SR19, but lacking SRs 5-15, while the C-terminal clone encodes sequences from Hinge3 through the C-terminus domain. Thus, the reconstituted mini-Dys mediated by the split intein trans-splicing contains 4 hinges, 13 SRs, the ABD, CR and CT domains. Unlike the improperly-phased-repeats in dystrophins that are frequently generated with exon skipping or gene editing (Harper, 2002), this novel mini-Dys (ΔSR5-15) carries only full spectrin-like repeats that will stabilize its secondary structure and molecular folding. More importantly, this mini-Dys (ΔSR5-15) is larger than the highly functional ΔH2-SR19 dystrophin found in very mild Becker patients (discussed in the background section). The new mini-Dys harbors several functional domains including actin, dystrotroglycan, dystrobrevin, syntrophin, and the neuronal nitric oxide synthase (nNOS) binding sites, which are important for its mechanical and signaling roles.

[00193] In a test, the inventors cloned two plasmids expressing either the N- or C-terminal mini-Dys/intein fragments (**FIG. 5A**). The inventors have tested four different splitting sites located in the linker between SR19 and Hinge3. (See, for example, **FIG. 10**). HEK293 cells were transfected with control plasmid encoding the entire mini-Dys (ΔSR5-15) or N and C-terminal vectors which encode for split mini-Dys (ΔSR5-15) fused to split intein. 48 hours later, total proteins were harvested for western blot analysis. Surprisingly, it was found that mini-Dys protein level was higher (5 to 11 fold) with the split vectors compared to the control plasmid (**FIGs. 5B & 5C**). This can possibly be explained by the short time required to simultaneously process two halves encoded by two vectors *versus* a long construct expressed by a single vector, or by transfection efficiencies. Interestingly, the inventors have noticed that all four selected splitting sites result in efficient mini-Dys formation, and the highest expression was obtained with site #2. In addition, no other bands were detected on the western blot membrane even after overexposure, which highlights the specificity of this method.

[00194] These *in vitro* data showed that this new split mini-Dys system can be used to validate the preselected inteins *in vitro*. Once the most efficient inteins are identified, the split mini-Dys/intein sets are cloned in AAV vectors for *in vivo* validation.

Pre-screening of intein library using split GFP system.

[00195] Numerous split intein pairs have been described, and the inventors compared the splicing of more of these (compared with the data in **FIGs. 4 and 5**) using a split intein GFP system. Plasmids expressing CMV-eGFP (control), dual plasmids carrying the N- or C-terminal half of eGFP with the N- or C-terminal half of codon optimized split inteins, & single plasmids carrying CMV-eGFP with the footprint that would be left by split intein joining of the dual vectors were cloned and transfected into HEK293 cells & the relative fluorescence was monitored (**FIG. 7**). Data show that this system worked well for comparing split intein splicing. While no pairs generated fluorescence as strong as that of WT GFP, several split intein vectors generated levels similar to that from a single vector carrying the corresponding footprint left behind after extein joining (e.g. the split Aha Intein; **FIG. 7**).

[00196] To evaluate whether 2 sets of split inteins could be used in 3 vectors to make full-length Dys, combinations of split intein halves were tested to see whether some might cross-splice with different split inteins, which could prevent joining of the exteins from 3 vectors by skipping the middle extein. It was observed that many split inteins do cross-splice, but that newer classes, 'group 2 inteins' generally do not (**FIG. 8**).

[00197] It was next tested whether the footprint could be reduced in size from its normal 6 amino acids (AA), and found that some split inteins splice efficiently when modified to leave behind only a 3 AA footprint (**FIG. 9A**). These results show the one can efficiently screen 'wild-type' & synthetic split inteins using a rapid GFP assay, and that some split intein pairs display efficient & specific splicing, while leaving minimal footprints after extein joining.

***in vitro* validation of the pre-selected inteins using split mini-Dys system**

[00198] This intein system was adapted for mini- & full-length dystrophin (Dys). Two or three vectors were prepared with one or two sets of split inteins & tested in HEK293 cells. Controls used single plasmids expressing the corresponding mini- (ΔSR5-15) or full-Dys. All split intein vectors made the correct protein, at levels higher than with the single vector (perhaps reflecting reduced transfection efficiency by the larger plasmid; **FIGs. 9B, 9C; FIGs. 5A-5C; FIG. 13)**. Some smaller products accumulated in the cells (dual vector) and some 'mini-Dys' in the triple vector studies, the latter suggesting splicing of the N-terminal extein directly to the C-terminal extein.

***in vivo* validation of the best mini-Dys/intein sets by intramuscular injection**

[00199] An example of one set of dual vectors that has been tested in *mdx*^{4cv} muscles reveals efficient expression of the ΔSR5-15 mini-dystrophin (**FIG. 12**). In this example, the split mini-Dystrophin/intein clones were inserted into AAV plasmid containing the muscle-specific creatine kinase 8 (CK8) regulatory cassette and small synthetic polyA flanked by two AAV serotype 2 inverted terminal repeats (ITRs) and used to make AAV vectors. A dose of 5x10^{10} viral genome (v.g) of AAV encoding the N- and/or C-terminal split mini-Dystrophin/intein was injected into tibialis anterior muscles (T.A) of three- week-old C57BL/6-*mdx*^{4cv}. Four weeks post-injection, the injected muscles were harvested and analyzed. Strong expression of mini-Dystrophin ΔSR5-15 was detected in 4 T.A muscle tested, highlighting the efficacy of SIMPLI-GT approach (**FIG. 12A**). Muscles were also cryo-sectioned and immunostained for dystrophin or stained with Hematoxylin and Eosin. The reconstituted mini-Dystrophin ΔSR5-15 was correctly localized at the myofiber sarcolemma of *mdx*^{4cv} injected with dual AAV N- and C-terminal vectors (**FIG. 12B**). These muscles exhibit a general muscle histology improvement with absence of inflammation (**FIG. 12C**).

[00200] For any given split-intein, there is flexibility on precisely where to split the mini-Dys exteins, which can affect splicing efficiency & minimize the footprint remaining after extein joining (FIGs B-E). Some of our efforts have focused on the ΔSR5-15 mini-Dys (**FIGs. 10 & 9C)** in which spectrin-like repeat (SR) 4 is joined to SR16. This mini-Dys is modified from the Dys made in a mildly-affected patient with an exon 17-48 deletion, who remained ambulatory until his late 70s (England, 1989). However, the design described herein removes a partial SR encoded on exon 49 (Harper, 2002) and adds the nNOS-localization domain in SRs 16-17. (Adams et al, 2018; Lai et al., 2009) This ΔSR5-15 mini-Dys is expected to assemble the entire DGC. Unlike exon-skipping or Cas9 cutting methods, it is not needed to split the Dys sequence at a precise exon (which often leaves behind partial domains). Instead, one can put portions of the SR16-19 region in either vector followed by the 5' split intein, while the 3' vector could carry the 3' split intein joined to the remaining SRs. The only difference is the location of the footprint. Multiple splitting sites have and can be tested to minimize footprints while monitoring splicing efficiency in 293 cells (**FIGs. 10,**

9B, C & 13). Engineered split inteins have also been developed to generate smaller inteins with a closer footprint match to candidate splitting sites (*e.g.* near the middle of mini-Dys clones; **FIG. 10**. The point here is that there are a wide choice of splitting targets such that the intein footprint can be matched to the Dys sequence while maximizing efficiency of splicing (& protein stability/function). Ideally, a Dys region that contains a perfect 6 AA match with the intein footprint would be used. While the inventors have not found such 6 AA stretches in a useful location, they have found regions where the footprint differs from Dys at only 1 or 2 amino acids. Given the size of the ideal mini-Dys proteins, there at least 600 bps in the middle of Dys to target the split inteins, as each extein vector must fit within the AAV cloning capacity (<5 kb) which includes the ITRs (360 bp), a polyA site (~60 bp) and an MSEC. An additional way to minimize footprints is to target Dys sequences that display conservative amino acid differences from a given footprint.

[00201] To adapt this system to genes other than dystrophin, the inventors split the dysferlin cDNA into two pieces, using 3 different split sites, and cloned 3 sets of plasmids each carrying one of the sets of split inteins, similar to what was done with the dual dystrophin vector studies. In this study, the three sets of split intein dysferlin plasmids were separately co-transfected into HEK293 cells followed by harvesting of the cells and analysis by western blot against dysferlin protein (**FIG. 14**). As shown in **FIG. 14A**, the full-length dysferlin protein was produced in the HEK293 cells with both sets of split-intein dysferlin clones. Both sets produced similar levels of dysferlin as did a control plasmid carrying the full-length dysferlin cDNA. **FIG. 14B** shows quantitation of the protein levels, illustrating the similar efficiencies that were obtained.

[00202] The new SIMPLI-GT approach presents several advantages and can be applied to any genetic disorder with a defective gene larger than the packaging capacities of AAV vectors. It relies on the usage of AAV vectors, which are widely used in gene therapy field due to their efficiency, serotype diversity, and tissue tropism. Unlike CRISPR-Cas9 gene editing and U7 exon skipping methods, this method will promote high expression of larger dystrophin with properly phased domains, which will stabilize the dystrophin structure. This strategy can be applied to any DMD or BMD patient regardless of their genetic mutations, and ultimately, will lead to the manufacturing of one therapeutic candidate with less variability and regulatory hurdles.

Literature cited:

[00203] Banks G. B., Judge L. M., Allen J. M., Chamberlain J. S. (2010). The polyproline site in hinge 2 influences the functional capacity of truncated dystrophins, PLoS Genet., 6(5):e1000958

[00204] Bengtsson N. E., Hall J. K., Odom G. L., Phelps M. P., Andrus C. R., Hawkins R. D., Hauschka S. D., Chamberlain J. R., Chamberlain J. S. (2017). Muscle-specific CRISPR/Cas9 dystrophin gene editing

ameliorates pathophysiology in a mouse model for Duchenne muscular dystrophy, Nat Commun., 14;8:14454

[00205] Chamberlain J. S., Gibbs R. A., Ranier J. E., Nguyen P. N., Caskey C .T. (1988). Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification, Nucleic Acids Res., 16(23): 11141–11156

[00206] Crudele J. M., Chamberlain J.S. (2019). AAV-based gene therapies for the muscular dystrophies, Human Molecular Genetics, Volume 28, Issue R1, Pages R102–R107

[00207] Emery E., Muntoni F. (2003). Duchenne muscular dystrophy, Oxford University Press, New York, 3rd edition, ISBN No. 0198515316

[00208] England S. B., Nicholson L. V., Johnson M. A., Forrest S. M., Love D. R., Zubrzycka-Gaarn E. E., Bulman D. E., Harris J. B., Davies K. E. (1990). Very mild muscular dystrophy associated with the deletion of 46% of dystrophin, Nature, 343(6254):180-2

[00209] Goyenvalle A., Vulin A., Fougerousse F., Leturcq F., Kaplan J., Garcia L., Danos O. (2004). Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping, Science, 306 (5702):1796-9

[00210] Gregorevic P., Allen J. M., Minami E., Blankinship M. J., Haraguchi M., Meuse L., Finn E., Adams M. E., Froehner S. C., Murry C. E., Chamberlain J. S. (2006). rAAV6-microdystrophin preserves muscle function and extends lifespan in severely dystrophic mice, Nat Med. 12(7):787-9

[00211] Gregorevic P., Blankinship M. J., Allen J. M., Crawford R. W., Meuse L., Miller D. G., Russell D. W., Chamberlain J. S. (2004). Systemic delivery of genes to striated muscles using adeno-associated viral vectors, Nat Med., 10(8):828-34

[00212] Harper S. Q., Hauser M. A., DelloRusso C., Duan D., Crawford R. W., Phelps S. F., Harper H. A., Robinson A. S., Engelhardt J. F., Brooks S. V., Chamberlain J. S. (2002). Modular flexibility of dystrophin: implications for gene therapy of Duchenne muscular dystrophy, Nat Med., 8(3):253-61

[00213] Koenig M., Beggs A. H., Moyer M., Scherpf S., Heindrich K., Bettecken T., Meng G., Müller C. R., Lindlöf M., Kaariainen H., de la Chapellet A., Kiuru A., Savontaus M. L., Gilgenkrantz H., Récan D., Chelly J., Kaplan J. C., Covone A. E., Archidiacono N., Romeo G., Liechti-Gailati S., Schneider V., Braga S., Moser H., Darras B. T., Murphy P., Francke U., Chen J. D., Morgan G., Denton M., Greenberg C. R., Wrogemann, Blonden L. A., van Paassen M. B., van Ommen G. J., Kunkel L. M. (1989). The molecular basis for Duchenne versus Becker muscular dystrophy: correlation of severity with type of deletion, Am J Hum Genet., 45(4):498-506

[00214] Koo T., Popplewell L., Athanasopoulos T., Dickson G. (2014). Triple trans-splicing adeno-associated virus vectors capable of transferring the coding sequence for full-length dystrophin protein into dystrophic mice, Hum Gene Ther, 25(2):98-108

[00215] Kunkel L. M. & co-authors (1986). Analysis of deletions in DNA from patients with Becker and Duchenne muscular dystrophy, Nature, 322, pages73–77

[00216] Li C & Samulski R. J. (2020). Engineering adeno-associated virus vectors for gene therapy, Nature Reviews Genetics, 21, p255–272

[00217] Li Y. (2015). Split-inteins and their bioapplications, Biotechnol Lett., 37(11):2121-37

[00218] Abedi M. R., Caponigro G., Kamb A. (1998). Green fluorescent protein as a scaffold for intracellular presentation of peptides, Nucleic Acids Res., 26(2): 623–630

[00219] Long C., Amoasii L., Mireault A. A., McAnally J. R., Li H., Sanchez-Ortiz E., Bhattacharyya S., Shelton J. M., Bassel-Duby R., Olson E. N (2016). Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy, Science, Vol. 351, Issue 6271, pp. 400-403

[00220] Lostal W., Kodippili K., Yue Y., Duan D. (2014). Full-length dystrophin reconstitution with adeno-associated viral vectors, Hum Gene Ther., 25(6):552-62

[00221] Monaco A. P., Bertelson C. J., Middlesworth W., Colletti C., Aldridge J., Fischbeck K.H., Bartlett R., Pericak-Vance M. A., Roses A.D. & Kunkel L. M. (1985), Detection of deletions spanning the Duchenne muscular dystrophy locus using a tightly linked DNA segment, Nature, 316, pages842–845

[00222] Nelson C. E., Hakim C. H., Ousterout D. G., Thakore P. I., Moreb E. A., Castellanos Rivera R. M., Madhavan S., Pan X., Ann Ran F., Yan W. X., Asokan A., Zhang F., Duan D., Gersbach C. A. (2016). In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy, Science, Vol. 351, Issue 6271, pp. 403-407

[00223] Niemeyer G. P., Herzog R. W., Mount J., Arruda V. R., Tillson D. M., Hathcock J., van Ginkel F. W., High K. A., Lothrop Jr C. D. (2009). Long-term correction of inhibitor-prone hemophilia B dogs treated with liver-directed AAV2-mediated factor IX gene therapy, Blood, 113(4):797-806

[00224] Ramos J. N., Hollinger K., Bengtsson N. E., Allen J. M., Hauschka S. D., Chamberlain J. S. (2019). Development of Novel Micro-dystrophins with Enhanced Functionality, Mol Ther., 27(3): 623-635

[00225] Rivera V. M., Gao G., Grant R. L., Schnell M. A., Zoltick P. W., Rozamus L. W., Clackson T., Wilson J. M. (2005). Long-term pharmacologically regulated expression of erythropoietin in primates following AAV-mediated gene transfer, Blood, 105(4):1424-30

[00226] Odom G.L., Gregorevic P., Allen J. M., Chamberlain J. S. (2011). Gene therapy of mdx mice with large truncated dystrophins generated by recombination using rAAV6, Mol Ther., 19(1):36-45

[00227] Shah N. H., Muir T. W. (2015). Inteins: nature's gift to protein chemists, 2014, Chemical Science, 5 (2), 446-461

[00228] Phelps S. F., Hauser M. A., Cole N. M., Rafael J. A., Hinkle R. T., Faulkner J. A., Chamberlain J. S. (1995). Expression of full-length and truncated dystrophin mini-genes in transgenic mdx mice, Hum. mol. gen., Volume 4, Issue 8, Pages 1251–1258

[00229] Srivastava A., Lusby E. W., Berns K. I. (1983). Nucleotide sequence and organization of the adeno-associated virus 2 genome, J Virol, 45(2):555-64

[00230] Soundrarajan N., Cho H., Ahn B., Choi M., Thong L. M., Choi H., Cha S., Kim J., Park C., Seo K., Park C. (2016). Green fluorescent protein as a scaffold for high efficiency production of functional bacteriotoxic proteins in Escherichia coli, Sci Rep., 6: 20661

[00231] Wasala N.B., Shin J.H., Lai Y., Yue Y., Montanaro F., Duan D. (2018). Cardiac specific expression of ΔH2-R15 mini-dystrophin normalized all ECG abnormalities and the end-diastolic volume in a 23-m-old mouse model of Duchenne dilated cardiomyopathy, Hum. Gene Ther., Vol. 29, No. 7

[00232] Adams ME, Odom GL, Kim MJ, Chamberlain JS and Froehner SC: Syntrophin binds directly to multiple spectrin-like repeats in dystrophin and mediates binding of nNOS to repeats 16-17. Hum Mol Genet 2018; 27:2978-2985.

[00233] Lai Y, Thomas GD, Yue Y, Yang HT, Li D, Long C, Judge L, Bostick B, Chamberlain JS, Terjung RL and Duan D: Dystrophins carrying spectrin-like repeats 16 and 17 anchor nNOS to the sarcolemma and enhance exercise performance in a mouse model of muscular dystrophy. J Clin Invest 2009; 119:624- 635.

EXAMPLE 2: Exemplary sequences of split inteins with mini-dystrophin, dystrophin dysferlin or utrophin

MiniDys ΔSR5-15:

1) **Split site #1:**

- **MiniDys-gp41.1 N1 (SEQ ID NO. 82):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
SYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDIIHSKKT
AALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFNQWLTE
AEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASILQEKLGS
LNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQLKEKLEQ
VKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRALPEKQGEIE
AQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQAKQPDVEEI
LSKGQHLYKEKPGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKIT
LEDGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **MiniDys-gp41.1 C1 (SEQ ID NO. 83):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSATQPVKRKLEDLSSEWKAVNRLLQELR
AKQPDLAPGLTTIGASPTQTVTLVTQPVVTKETAISKLEMPSSLMLEVPALADFNRAWTELTDW
LSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLEELITAAQNLKNKTSNQEARTII
TDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEAEQVLGQARAKLESWKEGPYT

VDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSADDTRKVHMITENINASWRSIHK
RVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDATRKERLLEDSKGVKELMKQW
QDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDNMNFKWSELRKKSLNIRSHLEA
SSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQNDVHRAFKRELKTKEPVIMST
LETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAEEVNTEWEKLNLHSADWQRKI
DETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQDHLEKVKALRGEIAPLKENVS
HVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQLHEAHRDFGPASQHFLSTSV
QGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRLQKALC
LDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLN
VYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGE
VASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAETAKHQAKC
NICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVEYCTPTTSGEDVRDFAKVLKN
KFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSAPASSPQLSHDDTHSRIEHYAS
RLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQPRSPAQILISLESEERGELERIL
ADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPRDAELIAEAKLLRQHKGRLEA
RMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSLQRSDSSQPMLLRVVGSQTSD
SMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMREDTM*

2) **Split site #2:**

- **MiniDys-gp41.1 N2 (SEQ ID NO. 84):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
SYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDIIHSKKT
AALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFNQWLTE
AEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASILQEKLGS
LNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQLKEKLEQ
VKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRALPEKQGEIE
AQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQAKQPDVEEI
LSKGQHLYKEGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLE
DGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **MiniDys-gp41.1 C2 (SEQ ID NO. 85):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSTQPVKRKLEDLSSEWKAVNRLLQELRA
KQPDLAPGLTTIGASPTQTVTLVTQPVVTKETAISKLEMPSSLMLEVPALADFNRAWTELTDWLS
LLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLEELITAAQNLKNKTSNQEARTIITD
RIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEAEQVLGQARAKLESWKEGPYTVD
AIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSADDTRKVHMITENINASWRSIHKRV
SEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDATRKERLLEDSKGVKELMKQWQD
LQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDNMNFKWSELRKKSLNIRSHLEASS
DQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQNDVHRAFKRELKTKEPVIMSTLE
TVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAEEVNTEWEKLNLHSADWQRKIDE
TLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQDHLEKVKALRGEIAPLKENVSHV
NDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQLHEAHRDFGPASQHFLSTSVQG

PWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRLQKALCLD
LLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLNV
YDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGEV
ASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAETAKHQAKCNI
CKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVEYCTPTTSGEDVRDFAKVLKNK
FRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSAPASSPQLSHDDTHSRIEHYASR
LAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQPRSPAQILISLESEERGELERILA
DLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPRDAELIAEAKLLRQHKGRLEAR
MQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSLQRSDSSQPMLLRVVGSQTSDS
MGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMREDTM*

3) **Split site #3:**

- **MiniDys-gp41.1 N3 (SEQ ID NO. 86):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
SYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDIIHSKKT
AALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFNQWLTE
AEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASILQEKLGS
LNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQLKEKLEQ
VKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRALPEKQGEIE
AQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQAKQPDVEEI
LSKGQHLYKEKPATQPVKRKLEDLSSEWKAVNRLLQELRAKQPDLAPGYCLDLKTQVQTPQG
MKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEEHLFPTQTGEMNISGGLK
EGMCLYVKE*

- **MiniDys-gp41.1 C3 (SEQ ID NO. 87):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSLTTIGASPTQTVTLVTQPVVTKETAISKL
EMPSSLMLEVPALADFNRAWTELTDWLSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQ
RRPQLEELITAAQNLKNKTSNQEARTIITDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWL
EAKEEAEQVLGQARAKLESWKEGPYTVDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLL
RDYSADDTRKVHMITENINASWRSIHKRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTA
NVLQDATRKERLLEDSKGVKELMKQWQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLL
QRRLDNMNFKWSELRKKSLNIRSHLEASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFP
AVQKQNDVHRAFKRELKTKEPVIMSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLL
RKQAEEVNTEWEKLNLHSADWQRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLL
IDSLQDHLEKVKALRGEIAPLKENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAV
EDRVRQLHEAHRDFGPASQHFLSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSL
ADLNNVRFSAYRTAMKLRRLQKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYD
RLEQEHNNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQV
ASSTGFCDQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLE
PQSMVWLPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKM
HYPMVEYCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINF
WPVDSAPASSPQLSHDDTHSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQ

DSPLSQPRSPAQILISLESEERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPT
SPQSPRDAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVS
SPSTSLQRSDSSQPMLLRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPG
KPMREDTM*

4) **Split site #4:**

- **MiniDys-gp41.1 N4 (SEQ ID NO. 88):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
SYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDIIHSKKT
AALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFNQWLTE
AEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASILQEKLGS
LNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQLKEKLEQ
VKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRALPEKQGEIE
AQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQAKQPDVEEI
LSKGQHLYKEKPATQPVKRKLEDLSSEWKAVNRLLQELRAKQPDLAPGLTTIGYCLDLKTQVQ
TPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEEHLFPTQTGEMNIS
GGLKEGMCLYVKE*

- **MiniDys-gp41.1 C4 (SEQ ID NO. 89):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSPTQTVTLVTQPVVTKETAISKLEMPSSL
MLEVPALADFNRAWTELTDWLSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLE
ELITAAQNLKNKTSNQEARTIITDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEA
EQVLGQARAKLESWKEGPYTVDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSAD
DTRKVHMITENINASWRSIHKRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDA
TRKERLLEDSKGVKELMKQWQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDN
MNFKWSELRKKSLNIRSHLEASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQ
NDVHRAFKRELKTKEPVIMSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAE
EVNTEWEKLNLHSADWQRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQD
HLEKVKALRGEIAPLKENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQ
LHEAHRDFGPASQHFLSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNV
RFSAYRTAMKLRRLQKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEH
NNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFC
DQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVW
LPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVE
YCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSA
PASSPQLSHDDTHSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQ
PRSPAQILISLESEERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPR
DAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSL
QRSDSSQPMLLRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMRE
DTM*

Full-length Dystrophin:

a. **Using IMPDH and gp41.1:**

1) **Split site ImS1:**

- **N-ter fragment (SEQ ID NO. 90):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVM
KNIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVV
CQQSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVE
MLPRPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSP
FPSQHLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFH
THEGYMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQ
SNLHRVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQ
VRVNSLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLT
EEQCLFSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDL
LSTLKNKSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTT
REQILVKHAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFS
DLKEKVNAIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLS
ERLNWLEYQNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQP
QIERLKIQSIALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMS
AIRTWVQQSETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEI
SRKYQSEFEEIEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWP
ALGGCFVPGTLVNTENGLKKIEEIKVGDKVFSHTGKLQEVVDTLIFDRDEEIISINGIDCTKNHEF
YVIDKENANRVNEDNIHLFARWVHAEELDMKKHLLIELE*

- **Middle fragment (SEQ ID NO. 91):**

MKFKLKEITSIETKHYKGKVHDLTVNQDHSYNVRGTVVHNSEILKKQLKQCRLLVSDIQTIQPS
LNSVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEALKGGLEKTVSLQKDL
SEMHEWMTQAEEEYLERDFEYKTPDELQKAVEEMKRAKEEAQQKEAKVKLLTESVNSVIAQA
PPVAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELLSYLEKANKWLNEVEFKL
KTTENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMDELINEELETFNSRWREL
HEEAVRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDAAQMPQEAQKIQSDLTS
HEISLEEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPANFELRLQESKMILDEVK
MHLPALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIVQKKQTENPKELDERV
TALKLHYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATDMELTKRSAVEGMPSN
LDSEVAWGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLSLLNSNWIAVTSRAEE
WLNLLLEYQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVLKRLKAELNDIRPKVD
STRDQAANLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKELEQFNSDIQKLLEPL
EAEIQQGVNLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEIKIKQQLLQTKHNA
LKDLRSQRRKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDERKIKEIDRELQKKKE
ELNAVRRQAEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIHTVREETMMVMTE
DMPLEISYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRI
DIIHSKKTAALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKI
FNQWLTEAEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDA
SILQEKLGSLNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKE
QQLKEKLEQVKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSR
ALPEKQGEIEAQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAV
QAKQPDVEEILSKGQHLYKEGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPK
SKKKSYKITLEDGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 92):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSTQPVKRKLEDLSSEWKAVNRLLQELR
AKQPDLAPGLTTIGASPTQTVTLVTQPVVTKETAISKLEMPSSLMLEVPALADFNRAWTELTDW
LSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLEELITAAQNLKNKTSNQEARTII
TDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEAEQVLGQARAKLESWKEGPYT

VDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSADDTRKVHMITENINASWRSIH
KRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDATRKERLLEDSKGVKELMKQ
WQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDNMNFKWSELRKKSLNIRSHL
EASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQNDVHRAFKRELKTKEPVI
MSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAEEVNTEWEKLNLHSADW
QRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQDHLEKVKALRGEIAPLK
ENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQLHEAHRDFGPASQHF
LSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRL
QKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMC
LNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSI
QIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAET
AKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVEYCTPTTSGEDVR
DFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSAPASSPQLSHDDT
HSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQPRSPAQILISLES
EERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPRDAELIAEAKLL
RQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSLQRSDSSQPML
LRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMREDTM*

2) **Split site ImS2:**

- **N-ter fragment (SEQ ID NO. 93):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
QNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQPQIERLKIQSI
ALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMSAIRTWVQQS
ETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEISRKYQSEFEE
IEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWPALGCFVPGTLV
NTENGLKKIEEIKVGDKVFSHTGKLQEVVDTLIFDRDEEIISINGIDCTKNHEFYVIDKENANRVN
EDNIHLFARWVHAEELDMKKHLLIELE*

- **Middle fragment (SEQ ID NO. 94):**

MKFKLKEITSIETKHYKGKVHDLTVNQDHSYNVRGTVVHNSEILKKQLKQCRLLVSDIQTIQPSL
NSVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEALKGGLEKTVSLQKDLSE
MHEWMTQAEEEYLERDFEYKTPDELQKAVEEMKRAKEEAQQKEAKVKLLTESVNSVIAQAPP
VAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELLSYLEKANKWLNEVEFKLKTT
ENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMDELINEELETFNSRWRELHEEA
VRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDAAQMPQEAQKIQSDLTSHEISL
EEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPANFELRLQESKMILDEVKMHLP
ALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIVQKKQTENPKELDERVTALKL
HYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATDMELTKRSAVEGMPSNLDSEVA
WGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLSLLNSNWIAVTSRAEEWLNLLLE
YQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVLKRLKAELNDIRPKVDSTRDQAA
NLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKELEQFNSDIQKLLEPLEAEIQQGV
NLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEIKIKQQLLQTKHNALKDLRSQR
RKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDERKIKEIDRELQKKKEELNAVRRQ

AEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIHTVREETMMVMTEDMPLEISYV
PSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDIIHSKKTAAL
QSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFNQWLTEAEQ
FLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASILQEKLGSLNL
RWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQLKEKLEQVK
LLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRALPEKQGEIEAQI
KDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQAKQPDVEEILSK
GQHLYKEGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGK
EIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 95):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSTQPVKRKLEDLSSEWKAVNRLLQELRA
KQPDLAPGLTTIGASPTQTVTLVTQPVVTKETAISKLEMPSSLMLEVPALADFNRAWTELTDWLS
LLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLEELITAAQNLKNKTSNQEARTIITD
RIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEAEQVLGQARAKLESWKEGPYTVD
AIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSADDTRKVHMITENINASWRSIHKRV
SEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDATRKERLLEDSKGVKELMKQWQD
LQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDNMNFKWSELRKKSLNIRSHLEASS
DQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQNDVHRAFKRELKTKEPVIMSTLE
TVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAEEVNTEWEKLNLHSADWQRKIDE
TLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQDHLEKVKALRGEIAPLKENVSHV
NDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQLHEAHRDFGPASQHFLSTSVQG
PWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRLQKALCLD
LLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLNV
YDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGEV
ASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAETAKHQAKCNI
CKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVEYCTPTTSGEDVRDFAKVLKNK
FRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSAPASSPQLSHDDTHSRIEHYASR
LAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQPRSPAQILISLESEERGELERILA
DLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPRDAELIAEAKLLRQHKGRLEAR
MQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSLQRSDSSQPMLLRVVGSQTSDS
MGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMREDTM*

3) **Split site ImS3:**

- **N-ter fragment (SEQ ID NO. 96):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
QNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQPQIERLKIQSI
ALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMSAIRTWVQQS
ETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEISRKYQSEFEE
IEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWPALGDSEILKKQ
LKQCRLLVSDIQTIQPSLNSVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEA
LKGGLEKTVSLQKDLSEMHEWMTQAEEEYLERDFEYGCFVPGTLVNTENGLKKIEEIKVGDKV

FSHTGKLQEVVDTLIFDRDEEIISINGIDCTKNHEFYVIDKENANRVNEDNIHLFARWVHAEELD
MKKHLLIELE*

- **Middle fragment (SEQ ID NO. 97):**

MKFKLKEITSIETKHYKGKVHDLTVNQDHSYNVRGTVVHNSKTPDELQKAVEEMKRAKEEAQ
QKEAKVKLLTESVNSVIAQAPPVAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHEL
LSYLEKANKWLNEVEFKLKTTENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVM
DELINEELETFNSRWRELHEEAVRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVD
AAQMPQEAQKIQSDLTSHEISLEEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPA
NFELRLQESKMILDEVKMHLPALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQI
VQKKQTENPKELDERVTALKLHYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATD
MELTKRSAVEGMPSNLDSEVAWGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLS
LLNSNWIAVTSRAEEWLNLLLEYQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVL
KRLKAELNDIRPKVDSTRDQAANLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKE
LEQFNSDIQKLLEPLEAEIQQGVNLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEI
KIKQQLLQTKHNALKDLRSQRRKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDER
KIKEIDRELQKKKEELNAVRRQAEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIH
TVREETMMVMTEDMPLEISYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLK
NIKDSLQQSSGRIDIIHSKKTAALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSV
EKWRRFHYDIKIFNQWLTEAEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATG
EEIIQQSSKTDASILQEKLGSLNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEAD
NIASIPLEPGKEQQLKEKLEQVKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQ
TNLQWIKVSRALPEKQGEIEAQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGP
FDVQETEIAVQAKQPDVEEILSKGQHLYKEKPATQPVKRKLEDLSSEWKAVNRLLQELRAKQPD
LAPGLTTIGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGK
EIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 98):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSPTQTVTLVTQPVVTKETAISKLEMPSSL
MLEVPALADFNRAWTELTDWLSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLE
ELITAAQNLKNKTSNQEARTIITDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEA
EQVLGQARAKLESWKEGPYTVDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSAD
DTRKVHMITENINASWRSIHKRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDA
TRKERLLEDSKGVKELMKQWQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDN
MNFKWSELRKKSLNIRSHLEASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQ
NDVHRAFKRELKTKEPVIMSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAE
EVNTEWEKLNLHSADWQRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQD
HLEKVKALRGEIAPLKENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQ
LHEAHRDFGPASQHFLSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNV
RFSAYRTAMKLRRLQKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEH
NNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFC
DQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVW
LPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVE
YCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSA
PASSPQLSHDDTHSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQ
PRSPAQILISLESEERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPR
DAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSL
QRSDSSQPMLLRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMRE
DTM*

4) **Split site ImS4:**

- **N-ter fragment (SEQ ID NO. 99):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
QNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQPQIERLKIQSI
ALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMSAIRTWVQQS
ETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEISRKYQSEFEE
IEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWPALGDSEILKKQ
LKQCRLLVSDIQTIQPSLNSVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEA
LKGGLEKTVSLQKDLSEMHEWMTQAEEEYLERDFEGCFVPGTLVNTENGLKKIEEIKVGDKVFS
HTGKLQEVVDTLIFDRDEEIISINGIDCTKNHEFYVIDKENANRVNEDNIHLFARWVHAEELDMK
KHLLIELE*

- **Middle fragment (SEQ ID NO. 100):**

MKFKLKEITSIETKHYKGKVHDLTVNQDHSYNVRGTVVHNSTPDELQKAVEEMKRAKEEAQQ
KEAKVKLLTESVNSVIAQAPPVAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELL
SYLEKANKWLNEVEFKLKTTENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMD
ELINEELETFNSRWRELHEEAVRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDA
AQMPQEAQKIQSDLTSHEISLEEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPAN
FELRLQESKMILDEVKMHLPALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIV
QKKQTENPKELDERVTALKLHYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATD
MELTKRSAVEGMPSNLDSEVAWGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLS
LLNSNWIAVTSRAEEWLNLLLEYQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVL
KRLKAELNDIRPKVDSTRDQAANLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKE
LEQFNSDIQKLLEPLEAEIQQGVNLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEI
KIKQQLLQTKHNALKDLRSQRRKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDER
KIKEIDRELQKKKEELNAVRRQAEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIH
TVREETMMVMTEDMPLEISYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLK
NIKDSLQQSSGRIDIIHSKKTAALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSV
EKWRRFHYDIKIFNQWLTEAEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATG
EEIIQQSSKTDASILQEKLGSLNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEAD
NIASIPLEPGKEQQLKEKLEQVKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQ
TNLQWIKVSRALPEKQGEIEAQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGP
FDVQETEIAVQAKQPDVEEILSKGQHLYKEKPATQPVKRKLEDLSSEWKAVNRLLQELRAKQPD
LAPGLTTIGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGK
EIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 101):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSPTQTVTLVTQPVVTKETAISKLEMPSSL
MLEVPALADFNRAWTELTDWLSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLE
ELITAAQNLKNKTSNQEARTIITDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEA
EQVLGQARAKLESWKEGPYTVDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSAD
DTRKVHMITENINASWRSIHKRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDA
TRKERLLEDSKGVKELMKQWQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDN

MNFKWSELRKKSLNIRSHLEASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQ
NDVHRAFKRELKTKEPVIMSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAE
EVNTEWEKLNLHSADWQRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQD
HLEKVKALRGEIAPLKENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQ
LHEAHRDFGPASQHFLSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNV
RFSAYRTAMKLRRLQKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEH
NNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFC
DQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVW
LPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVE
YCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSA
PASSPQLSHDDTHSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQ
PRSPAQILISLESEERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPR
DAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSL
QRSDSSQPMLLRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMRE
DTM*

b. Using Nrdj-1:

1) Split site NrS1:

- **N-ter fragment (SEQ ID NO. 102):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVM
KNIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVV
CQQSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVE
MLPRPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSP
FPSQHLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFH
THEGYMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQ
SNLHRVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQ
VRVNSLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLT
EEQCLFSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDL
LSTLKNKSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTT
REQILVKHAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFS
DLKEKVNAIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLS
ERLNWLEYQNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQP
QIERLKIQSIALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMS
AIRTWVQQSETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEI
SRKYQSEFEEIEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWP
ALGCCLVGSSEIITRNYGKTTIKEVVEIFDNDKNIQVLAFNTHTDNIEWAPIKAAQLTRPNAELV
ELEIDTLHGVKTIRCTPDHPVYTKNRGYVRADELTDDDELVVAI*

- **Middle fragment (SEQ ID NO. 103):**

MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNSEILKKQLKQCRLLVSDIQTIQPSLN
SVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEALKGGLEKTVSLQKDLSE
MHEWMTQAEEEYLERDFEYKTPDELQKAVEEMKRAKEEAQQKEAKVKLLTESVNSVIAQAPP
VAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELLSYLEKANKWLNEVEFKLKT
TENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMDELINEELETFNSRWRELHE
EAVRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDAAQMPQEAQKIQSDLTSHE
ISLEEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPANFELRLQESKMILDEVKM
HLPALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIVQKKQTENPKELDERVT
ALKLHYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATDMELTKRSAVEGMPSNL
DSEVAWGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLSLLNSNWIAVTSRAEEW
LNLLLEYQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVLKRLKAELNDIRPKVDS
TRDQAANLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKELEQFNSDIQKLLEPLE
AEIQQGVNLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEIKIKQQLLQTKHNAL
KDLRSQRRKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDERKIKEIDRELQKKKEE

LNAVRRQAEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIHTVREETMMVMTED
MPLEISYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDII
HSKKTAALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFN
QWLTEAEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASIL
QEKLGSLNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQ
LKEKLEQVKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRAL
PEKQGEIEAQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQA
KQPDVEEILSKGQHLYKEGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSK
KKSYKITLEDGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 104):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSTQPVKRKLEDLSSEWKAVNRLLQELRA
KQPDLAPGLTTIGASPTQTVTLVTQPVVTKETAISKLEMPSSLMLEVPALADFNRAWTELTDWLS
LLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLEELITAAQNLKNKTSNQEARTIITD
RIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEAEQVLGQARAKLESWKEGPYTVD
AIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSADDTRKVHMITENINASWRSIHKRV
SEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDATRKERLLEDSKGVKELMKQWQD
LQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDNMNFKWSELRKKSLNIRSHLEASS
DQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQNDVHRAFKRELKTKEPVIMSTLE
TVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAEEVNTEWEKLNLHSADWQRKIDE
TLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQDHLEKVKALRGEIAPLKENVSHV
NDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQLHEAHRDFGPASQHFLSTSVQG
PWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRLQKALCLD
LLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLNV
YDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGEV
ASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAETAKHQAKCNI
CKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVEYCTPTTSGEDVRDFAKVLKNK
FRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSAPASSPQLSHDDTHSRIEHYASR
LAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQPRSPAQILISLESEERGELERILA
DLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPRDAELIAEAKLLRQHKGRLEAR
MQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSLQRSDSSQPMLLRVVGSQTSDS
MGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMREDTM*

2) **Split site NrS2:**

- **N-ter fragment (SEQ ID NO. 105):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
QNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQPQIERLKIQSI
ALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMSAIRTWVQQS
ETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEISRKYQSEFEE
IEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWPALGDCCLVGSS
EIITRNYGKTTIKEVVEIFDNDKNIQVLAFNTHTDNIEWAPIKAAQLTRPNAELVELEIDTLHGVK
TIRCTPDHPVYTKNRGYVRADELTDDDELVVAI*

- **Middle fragment (SEQ ID NO. 106):**

MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNSEILKKQLKQCRLLVSDIQTIQPSLNS
VNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEALKGGLEKTVSLQKDLSEM
HEWMTQAEEEYLERDFEYKTPDELQKAVEEMKRAKEEAQQKEAKVKLLTESVNSVIAQAPPVA
QEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELLSYLEKANKWLNEVEFKLKTTENI
PGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMDELINEELETFNSRWRELHEEAVR
RQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDAAQMPQEAQKIQSDLTSHEISLEE
MKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPANFELRLQESKMILDEVKMHLPAL
ETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIVQKKQTENPKELDERVTALKLHY
NELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATDMELTKRSAVEGMPSNLDSEVAW
GKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLSLLNSNWIAVTSRAEEWLNLLLEY
QKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVLKRLKAELNDIRPKVDSTRDQAAN
LMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKELEQFNSDIQKLLEPLEAEIQQGVN
LKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEIKIKQQLLQTKHNALKDLRSQRR
KKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDERKIKEIDRELQKKKEELNAVRRQA
EGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIHTVREETMMVMTEDMPLEISYVP
STYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDSLQQSSGRIDIIHSKKTAAL
QSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWRRFHYDIKIFNQWLTEAEQ
FLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQQSSKTDASILQEKLGSLNL
RWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASIPLEPGKEQQLKEKLEQVK
LLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQWIKVSRALPEKQGEIEAQI
KDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQETEIAVQAKQPDVEEILSK
GQHLYKEGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGK
EIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 107):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSTQPVKRKLEDLSSEWKAVNRLLQELRA
KQPDLAPGLTTIGASPTQTVTLVTQPVVTKETAISKLEMPSSLMLEVPALADFNRAWTELTDWLS
LLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLEELITAAQNLKNKTSNQEARTIITD
RIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEAEQVLGQARAKLESWKEGPYTVD
AIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSADDTRKVHMITENINASWRSIHKRV
SEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDATRKERLLEDSKGVKELMKQWQD
LQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDNMNFKWSELRKKSLNIRSHLEASS
DQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQNDVHRAFKRELKTKEPVIMSTLE
TVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAEEVNTEWEKLNLHSADWQRKIDE
TLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQDHLEKVKALRGEIAPLKENVSHV
NDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQLHEAHRDFGPASQHFLSTSVQG
PWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNVRFSAYRTAMKLRRLQKALCLD
LLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEHNNLVNVPLCVDMCLNWLLNV
YDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFCDQRRLGLLLHDSIQIPRQLGEV
ASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVWLPVLHRVAAAETAKHQAKCNI
CKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVEYCTPTTSGEDVRDFAKVLKNK
FRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSAPASSPQLSHDDTHSRIEHYASR
LAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQPRSPAQILISLESEERGELERILA
DLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPRDAELIAEAKLLRQHKGRLEAR
MQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSLQRSDSSQPMLLRVVGSQTSDS
MGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMREDTM*

3) **Split site NrS3:**

- **N-ter fragment (SEQ ID NO. 108):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ

QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
QNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQPQIERLKIQSI
ALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMSAIRTWVQQS
ETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEISRKYQSEFEE
IEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWPALGDSEILKKQ
LKQCRLLVSDIQTIQPSLNSVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEA
LKGGLEKTVSLQKDLSEMHEWMTQAEEEYLERDFEYCCLVGSSEIITRNYGKTTIKEVVEIFDND
KNIQVLAFNTHTDNIEWAPIKAAQLTRPNAELVELEIDTLHGVKTIRCTPDHPVYTKNRGYVRAD
ELTDDDELVVAI*

- **Middle fragment (SEQ ID NO. 109):**

MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNSKTPDELQKAVEEMKRAKEEAQQK
EAKVKLLTESVNSVIAQAPPVAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELLS
YLEKANKWLNEVEFKLKTTENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMD
ELINEELETFNSRWRELHEEAVRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDA
AQMPQEAQKIQSDLTSHEISLEEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPAN
FELRLQESKMILDEVKMHLPALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIV
QKKQTENPKELDERVTALKLHYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATD
MELTKRSAVEGMPSNLDSEVAWGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLS
LLNSNWIAVTSRAEEWLNLLLEYQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVL
KRLKAELNDIRPKVDSTRDQAANLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKE
LEQFNSDIQKLLEPLEAEIQQGVNLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEI
KIKQQLLQTKHNALKDLRSQRRKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDER
KIKEIDRELQKKKEELNAVRRQAEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIH
TVREETMMVMTEDMPLEISYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLK
NIKDSLQQSSGRIDIIHSKKTAALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSV
EKWRRFHYDIKIFNQWLTEAEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATG
EEIIQQSSKTDASILQEKLGSLNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEAD
NIASIPLEPGKEQQLKEKLEQVKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQ
TNLQWIKVSRALPEKQGEIEAQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGP
FDVQETEIAVQAKQPDVEEILSKGQHLYKEKPATQPVKRKLEDLSSEWKAVNRLLQELRAKQPD
LAPGLTTIGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGK
EIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 110):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSPTQTVTLVTQPVVTKETAISKLEMPSSL
MLEVPALADFNRAWTELTDWLSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLE
ELITAAQNLKNKTSNQEARTIITDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEA
EQVLGQARAKLESWKEGPYTVDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSAD
DTRKVHMITENINASWRSIHKRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDA
TRKERLLEDSKGVKELMKQWQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDN
MNFKWSELRKKSLNIRSHLEASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQ
NDVHRAFKRELKTKEPVIMSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAE
EVNTEWEKLNLHSADWQRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQD
HLEKVKALRGEIAPLKENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQ

LHEAHRDFGPASQHFLSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNV
RFSAYRTAMKLRRLQKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEH
NNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFC
DQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVW
LPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVE
YCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSA
PASSPQLSHDDTHSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQ
PRSPAQILISLESEERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPR
DAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSL
QRSDSSQPMLLRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMRE
DTM*

4) **Split site NrS4:**

- **N-ter fragment (SEQ ID NO. 111):**

MLWWEEVEDCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRLLDLLEGLTGQKL
PKEKGSTRVHALNNVNKALRVLQNNNVDLVNIGSTDIVDGNHKLTLGLIWNIILHWQVKNVMK
NIMAGLQQTNSEKILLSWVRQSTRNYPQVNVINFTTSWSDGLALNALIHSHRPDLFDWNSVVCQ
QSATQRLEHAFNIARYQLGIEKLLDPEDVDTTYPDKKSILMYITSLFQVLPQQVSIEAIQEVEMLP
RPPKVTKEEHFQLHHQMHYSQQITVSLAQGYERTSSPKPRFKSYAYTQAAYVTTSDPTRSPFPSQ
HLEAPEDKSFGSSLMESEVNLDRYQTALEEVLSWLLSAEDTLQAQGEISNDVEVVKDQFHTHEG
YMMDLTAHQGRVGNILQLGSKLIGTGKLSEDEETEVQEQMNLLNSRWECLRVASMEKQSNLH
RVLMDLQNQKLKELNDWLTKTEERTRKMEEEPLGPDLEDLKRQVQQHKVLQEDLEQEQVRVN
SLTHMVVVVDESSGDHATAALEEQLKVLGDRWANICRWTEDRWVLLQDILLKWQRLTEEQCL
FSAWLSEKEDAVNKIHTTGFKDQNEMLSSLQKLAVLKADLEKKKQSMGKLYSLKQDLLSTLKN
KSVTQKTEAWLDNFARCWDNLVQKLEKSTAQISQAVTTTQPSLTQTTVMETVTTVTTREQILVK
HAQEELPPPPPQKKRQITVDSEIRKRLDVDITELHSWITRSEAVLQSPEFAIFRKEGNFSDLKEKVN
AIEREKAEKFRKLQDASRSAQALVEQMVNEGVNADSIKQASEQLNSRWIEFCQLLSERLNWLEY
QNNIIAFYNQLQQLEQMTTTAENWLKIQPTTPSEPTAIKSQLKICKDEVNRLSGLQPQIERLKIQSI
ALKEKGQGPMFLDADFVAFTNHFKQVFSDVQAREKELQTIFDTLPPMRYQETMSAIRTWVQQS
ETKLSIPQLSVTDYEIMEQRLGELQALQSSLQEQQSGLYYLSTTVKEMSKKAPSEISRKYQSEFEE
IEGRWKKLSSQLVEHCQKLEEQMNKLRKIQNHIQTLKKWMAEVDVFLKEEWPALGDSEILKKQ
LKQCRLLVSDIQTIQPSLNSVNEGGQKIKNEAEPEFASRLETELKELNTQWDHMCQQVYARKEA
LKGGLEKTVSLQKDLSEMHEWMTQAEEEYLERDFECCLVGSSEIITRNYGKTTIKEVVEIFDNDK
NIQVLAFNTHTDNIEWAPIKAAQLTRPNAELVELEIDTLHGVKTIRCTPDHPVYTKNRGYVRADE
LTDDDELVVAI*

- **Middle fragment (SEQ ID NO. 112) :**

MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNSTPDELQKAVEEMKRAKEEAQQKE
AKVKLLTESVNSVIAQAPPVAQEALKKELETLTTNYQWLCTRLNGKCKTLEEVWACWHELLSY
LEKANKWLNEVEFKLKTTENIPGGAEEISEVLDSLENLMRHSEDNPNQIRILAQTLTDGGVMDEL
INEELETFNSRWRELHEEAVRRQKLLEQSIQSAQETEKSLHLIQESLTFIDKQLAAYIADKVDAAQ
MPQEAQKIQSDLTSHEISLEEMKKHNQGKEAAQRVLSQIDVAQKKLQDVSMKFRLFQKPANFEL
RLQESKMILDEVKMHLPALETKSVEQEVVQSQLNHCVNLYKSLSEVKSEVEMVIKTGRQIVQKK
QTENPKELDERVTALKLHYNELGAKVTERKQQLEKCLKLSRKMRKEMNVLTEWLAATDMELT
KRSAVEGMPSNLDSEVAWGKATQKEIEKQKVHLKSITEVGEALKTVLGKKETLVEDKLSLLNS
NWIAVTSRAEEWLNLLLEYQKHMETFDQNVDHITKWIIQADTLLDESEKKKPQQKEDVLKRLK
AELNDIRPKVDSTRDQAANLMANRGDHCRKLVEPQISELNHRFAAISHRIKTGKASIPLKELEQF
NSDIQKLLEPLEAEIQQGVNLKEEDFNKDMNEDNEGTVKELLQRGDNLQQRITDERKREEIKIKQ
QLLQTKHNALKDLRSQRRKKALEISHQWYQYKRQADDLLKCLDDIEKKLASLPEPRDERKIKEI
DRELQKKKEELNAVRRQAEGLSEDGAAMAVEPTQIQLSKRWREIESKFAQFRRLNFAQIHTVRE
ETMMVMTEDMPLEISYVPSTYLTEITHVSQALLEVEQLLNAPDLCAKDFEDLFKQEESLKNIKDS
LQQSSGRIDIIHSKKTAALQSATPVERVKLQEALSQLDFQWEKVNKMYKDRQGRFDRSVEKWR
RFHYDIKIFNQWLTEAEQFLRKTQIPENWEHAKYKWYLKELQDGIGQRQTVVRTLNATGEEIIQ
QSSKTDASILQEKLGSLNLRWQEVCKQLSDRKKRLEEQKNILSEFQRDLNEFVLWLEEADNIASI

PLEPGKEQQLKEKLEQVKLLVEELPLRQGILKQLNETGGPVLVSAPISPEEQDKLENKLKQTNLQ
WIKVSRALPEKQGEIEAQIKDLGQLEKKLEDLEEQLNHLLLWLSPIRNQLEIYNQPNQEGPFDVQ
ETEIAVQAKQPDVEEILSKGQHLYKEKPATQPVKRKLEDLSSEWKAVNRLLQELRAKQPDLAPG
LTTIGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGKEIICS
EEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 113):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSPTQTVTLVTQPVVTKETAISKLEMPSSL
MLEVPALADFNRAWTELTDWLSLLDQVIKSQRVMVGDLEDINEMIIKQKATMQDLEQRRPQLE
ELITAAQNLKNKTSNQEARTIITDRIERIQNQWDEVQEHLQNRRQQLNEMLKDSTQWLEAKEEA
EQVLGQARAKLESWKEGPYTVDAIQKKITETKQLAKDLRQWQTNVDVANDLALKLLRDYSAD
DTRKVHMITENINASWRSIHKRVSEREAALEETHRLLQQFPLDLEKFLAWLTEAETTANVLQDA
TRKERLLEDSKGVKELMKQWQDLQGEIEAHTDVYHNLDENSQKILRSLEGSDDAVLLQRRLDN
MNFKWSELRKKSLNIRSHLEASSDQWKRLHLSLQELLVWLQLKDDELSRQAPIGGDFPAVQKQ
NDVHRAFKRELKTKEPVIMSTLETVRIFLTEQPLEGLEKLYQEPRELPPEERAQNVTRLLRKQAE
EVNTEWEKLNLHSADWQRKIDETLERLQELQEATDELDLKLRQAEVIKGSWQPVGDLLIDSLQD
HLEKVKALRGEIAPLKENVSHVNDLARQLTTLGIQLSPYNLSTLEDLNTRWKLLQVAVEDRVRQ
LHEAHRDFGPASQHFLSTSVQGPWERAISPNKVPYYINHETQTTCWDHPKMTELYQSLADLNNV
RFSAYRTAMKLRRLQKALCLDLLSLSAACDALDQHNLKQNDQPMDILQIINCLTTIYDRLEQEH
NNLVNVPLCVDMCLNWLLNVYDTGRTGRIRVLSFKTGIISLCKAHLEDKYRYLFKQVASSTGFC
DQRRLGLLLHDSIQIPRQLGEVASFGGSNIEPSVRSCFQFANNKPEIEAALFLDWMRLEPQSMVW
LPVLHRVAAAETAKHQAKCNICKECPIIGFRYRSLKHFNYDICQSCFFSGRVAKGHKMHYPMVE
YCTPTTSGEDVRDFAKVLKNKFRTKRYFAKHPRMGYLPVQTVLEGDNMETPVTLINFWPVDSA
PASSPQLSHDDTHSRIEHYASRLAEMENSNGSYLNDSISPNESIDDEHLLIQHYCQSLNQDSPLSQ
PRSPAQILISLESEERGELERILADLEEENRNLQAEYDRLKQQHEHKGLSPLPSPPEMMPTSPQSPR
DAELIAEAKLLRQHKGRLEARMQILEDHNKQLESQLHRLRQLLEQPQAEAKVNGTTVSSPSTSL
QRSDSSQPMLLRVVGSQTSDSMGEEDLLSPPQDTSTGLEEVMEQLNNSFPSSRGRNTPGKPMRE
DTM*

Full-length Dysferlin:

1) **Split site #1 (using split gp41.1):**
   - **N-ter fragment (SEQ ID NO. 114):**

MLRVFILYAENVHTPDTDISDAYCSAVFAGVKKRTKVIKNSVNPVWNEGFEWDLKGIPLDQGSE
LHVVVKDHETMGRNRFLGEAKVPLREVLATPSLSASFNAPLLDTKKQPTGASLVLQVSYTPLPG
AVPLFPPPTPLEPSPTLPDLDVVADTGGEEDTEDQGLTGDEAEPFLDQSGGPGAPTTPRKLPSRPP
PHYPGIKRKRSAPTSRKLLSDKPQDFQIRVQVIEGRQLPGVNIKPVVKVTAAGQTKRTRIHKGNS
PLFNETLFFNLFDSPGELFDEPIFITVVDSRSLRTDALLGEFRMDVGTIYREPRHAYLRKWLLLSD
PDDFSAGARGYLKTSLCVLGPGDEAPLERKDPSEDKEDIESNLLRPTGVALRGAHFCLKVFRAE
DLPQMDDAVMDNVKQIFGFESNKKNLVDPFVEVSFAGKMLCSKILEKTANPQWNQNITLPAMF
PSMCEKMRIRIIDWDRLTHNDIVATTYLSMSKISAPGGEIEEEPAGAVKPSKASDLDDYLGFLPTF
GPCYINLYGSPREFTGFPDPYTELNTGKGEGVAYRGRLLLSLETKLVEHSEQKVEDLPADDILRV
EKYLRRRKYSLFAAFYSATMLQDVDDAIQFEVSIGNYGNKFDMTCLPLASTTQYSRAVFDGCH
YYYLPWGNVKPVVVLSSYWEDISHRIETQNQLLGIADRLEAGLEQVHLALKAQCSTEDVDSLV
AQLTDELIAGCSQPLGDIHETPSATHLDQYLYQLRTHHLSQITEAALALKLGYCLDLKTQVQTPQ
GMKEISNIQVGDLVLSNTGYNEVLNVFPKSKKKSYKITLEDGKEIICSEEHLFPTQTGEMNISGGL
KEGMCLYVKE*

   - **C-ter fragment (SEQ ID NO. 115):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSELPAALEQAEDWLLRLRALAEEPQNSLP
DIVIWMLQGDKRVAYQRVPAHQVLFSRRGANYCGKNCGKLQTIFLKYPMEKVPGARMPVQIR
VKLWFGLSVDEKEFNQFAEGKLSVFAETYENETKLALVGNWGTTGLTYPKFSDVTGKIKLPKDS
FRPSAGWTWAGDWFVCPEKTLLHDMDAGHLSFVEEVFENQTRLPGGQWIYMSDNYTDVNGEK
VLPKDDIECPLGWKWEDEEWSTDLNRAVDEQGWEYSITIPPERKPKHWVPAEKMYYTHRRRR
WVRLRRRDLSQMEALKRHRQAEAEGEGWEYASLFGWKFHLEYRKTDAFRRRRWRRRMEPLE

KTGPAAVFALEGALGGVMDDKSEDSMSVSTLSFGVNRPTISCIFDYGNRYHLRCYMYQARDLA
AMDKDSFSDPYAIVSFLHQSQKTVVVKNTLNPTWDQTLIFYEIEIFGEPATVAEQPPSIVVELYDH
DTYGADEFMGRCICQPSLERMPRLAWFPLTRGSQPSGELLASFELIQREKPAIHHIPGFEVQETSRI
LDESEDTDLPYPPPQREANIYMVPQNIKPALQRTAIEILAWGLRNMKSYQLANISSPSLVVECGG
QTVQSCVIRNLRKNPNFDICTLFMEVMLPREELYCPPITVKVIDNRQFGRRPVVGQCTIRSLESFL
CDPYSAESPSPQGGPDDVSLLSPGEDVLIDIDDKEPLIPIQEEEFIDWWSKFFASIGEREKCGSYLE
KDFDTLKVYDTQLENVEAFEGLSDFCNTFKLYRGKTQEETEDPSVIGEFKGLFKIYPLPEDPAIPM
PPRQFHQLAAQGPQECLVRIYIVRAFGLQPKDPNGKCDPYIKISIGKKSVSDQDNYIPCTLEPVFG
KMFELTCTLPLEKDLKITLYDYDLLSKDEKIGETVVDLENRLLSKFGARCGLPQTYCVSGPNQW
RDQLRPSQLLHLFCQQHRVKAPVYRTDRVMFQDKEYSIEEIEAGRIPNPHLGPVEERLALHVLQQ
QGLVPEHVESRPLYSPLQPDIEQGKLQMWVDLFPKALGRPGPPFNITPRRARRFFLRCIIWNTRD
VILDDLSLTGEKMSDIYVKGWMIGFEEHKQKTDVHYRSLGGEGNFNWRFIFPFDYLPAEQVCTI
AKKDAFWRLDKTESKIPARVVFQIWDNDKFSFDDFLGSLQLDLNRMPKPAKTAKKCSLDQLDD
AFHPEWFVSLFEQKTVKGWWPCVAEEGEKKILAGKLEMTLEIVAESEHEERPAGQGRDEPNMN
PKLEDPRRPDTSFLWFTSPYKTMKFILWRRFRWAIILFIILFILLLFLAIFIYAFPNYAAMKLVKPFS
*

2) **Split site #2 (using split gp41.1):**

- **N-ter fragment (SEQ ID NO. 116):**

MLRVFILYAENVHTPDTDISDAYCSAVFAGVKKRTKVIKNSVNPVWNEGFEWDLKGIPLDQGSE
LHVVVKDHETMGRNRFLGEAKVPLREVLATPSLSASFNAPLLDTKKQPTGASLVLQVSYTPLPG
AVPLFPPPTPLEPSPTLPDLDVVADTGGEEDTEDQGLTGDEAEPFLDQSGGPGAPTTPRKLPSRPP
PHYPGIKRKRSAPTSRKLLSDKPQDFQIRVQVIEGRQLPGVNIKPVVKVTAAGQTKRTRIHKGNS
PLFNETLFFNLFDSPGELFDEPIFITVVDSRSLRTDALLGEFRMDVGTIYREPRHAYLRKWLLLSD
PDDFSAGARGYLKTSLCVLGPGDEAPLERKDPSEDKEDIESNLLRPTGVALRGAHFCLKVFRAE
DLPQMDDAVMDNVKQIFGFESNKKNLVDPFVEVSFAGKMLCSKILEKTANPQWNQNITLPAMF
PSMCEKMRIRIIDWDRLTHNDIVATTYLSMSKISAPGGEIEEEPAGAVKPSKASDLDDYLGFLPTF
GPCYINLYGSPREFTGFPDPYTELNTGKGEGVAYRGRLLLSLETKLVEHSEQKVEDLPADDILRV
EKYLRRRKYSLFAAFYSATMLQDVDDAIQFEVSIGNYGNKFDMTCLPLASTTQYSRAVFDGCH
YYYLPWGNVKPVVVLSSYWEDISHRIETQNQLLGIADRLEAGLEQVHLALKAQCSTEDVDSLV
AQLTDELIAGCSQPLGDIHETPSATHLDQYLYQLRTHHLSQITEAALALKLGHSELPAALEQAED
WLLRLRALAEEPQNSLPDIVIWMLQGDKRVAYQRVPAHQVLFSRRGANYCGKNCGKLQTIFLK
YPMEKVPGARMPVQIRVKLWFGYCLDLKTQVQTPQGMKEISNIQVGDLVLSNTGYNEVLNVFP
KSKKKSYKITLEDGKEIICSEEHLFPTQTGEMNISGGLKEGMCLYVKE*

- **C-ter fragment (SEQ ID NO. 117):**

MMLKKILKIEELDERELIDIEVSGNHLFYANDILTHNSVDEKEFNQFAEGKLSVFAETYENETKL
ALVGNWGTTGLTYPKFSDVTGKIKLPKDSFRPSAGWTWAGDWFVCPEKTLLHDMDAGHLSFV
EEVFENQTRLPGGQWIYMSDNYTDVNGEKVLPKDDIECPLGWKWEDEEWSTDLNRAVDEQGW
EYSITIPPERKPKHWVPAEKMYYTHRRRRWVRLRRRDLSQMEALKRHRQAEAEGEGWEYASLF
GWKFHLEYRKTDAFRRRRWRRRMEPLEKTGPAAVFALEGALGGVMDDKSEDSMSVSTLSFGV
NRPTISCIFDYGNRYHLRCYMYQARDLAAMDKDSFSDPYAIVSFLHQSQKTVVVKNTLNPTWD
QTLIFYEIEIFGEPATVAEQPPSIVVELYDHDTYGADEFMGRCICQPSLERMPRLAWFPLTRGSQP
SGELLASFELIQREKPAIHHIPGFEVQETSRILDESEDTDLPYPPPQREANIYMVPQNIKPALQRTAI
EILAWGLRNMKSYQLANISSPSLVVECGGQTVQSCVIRNLRKNPNFDICTLFMEVMLPREELYCP
PITVKVIDNRQFGRRPVVGQCTIRSLESFLCDPYSAESPSPQGGPDDVSLLSPGEDVLIDIDDKEPLI
PIQEEEFIDWWSKFFASIGEREKCGSYLEKDFDTLKVYDTQLENVEAFEGLSDFCNTFKLYRGKT
QEETEDPSVIGEFKGLFKIYPLPEDPAIPMPPRQFHQLAAQGPQECLVRIYIVRAFGLQPKDPNGK
CDPYIKISIGKKSVSDQDNYIPCTLEPVFGKMFELTCTLPLEKDLKITLYDYDLLSKDEKIGETVV
DLENRLLSKFGARCGLPQTYCVSGPNQWRDQLRPSQLLHLFCQQHRVKAPVYRTDRVMFQDKE
YSIEEIEAGRIPNPHLGPVEERLALHVLQQQGLVPEHVESRPLYSPLQPDIEQGKLQMWVDLFPK
ALGRPGPPFNITPRRARRFFLRCIIWNTRDVILDDLSLTGEKMSDIYVKGWMIGFEEHKQKTDVH
YRSLGGEGNFNWRFIFPFDYLPAEQVCTIAKKDAFWRLDKTESKIPARVVFQIWDNDKFSFDDFL
GSLQLDLNRMPKPAKTAKKCSLDQLDDAFHPEWFVSLFEQKTVKGWWPCVAEEGEKKILAGK

LEMTLEIVAESEHEERPAGQGRDEPNMNPKLEDPRRPDTSFLWFTSPYKTMKFILWRRFRWAIIL
FIILFILLLFLAIFIYAFPNYAAMKLVKPFS*

3) **Split site #3 (using split IMPDH):**

- **N-ter fragment (SEQ ID NO. 118):**

MLRVFILYAENVHTPDTDISDAYCSAVFAGVKKRTKVIKNSVNPVWNEGFEWDLKGIPLDQGSE
LHVVVKDHETMGRNRFLGEAKVPLREVLATPSLSASFNAPLLDTKKQPTGASLVLQVSYTPLPG
AVPLFPPPTPLEPSPTLPDLDVVADTGGEEDTEDQGLTGDEAEPFLDQSGGPGAPTTPRKLPSRPP
PHYPGIKRKRSAPTSRKLLSDKPQDFQIRVQVIEGRQLPGVNIKPVVKVTAAGQTKRTRIHKGNS
PLFNETLFFNLFDSPGELFDEPIFITVVDSRSLRTDALLGEFRMDVGTIYREPRHAYLRKWLLLSD
PDDFSAGARGYLKTSLCVLGPGDEAPLERKDPSEDKEDIESNLLRPTGVALRGAHFCLKVFRAE
DLPQMDDAVMDNVKQIFGFESNKKNLVDPFVEVSFAGKMLCSKILEKTANPQWNQNITLPAMF
PSMCEKMRIRIIDWDRLTHNDIVATTYLSMSKISAPGGEIEEEPAGAVKPSKASDLDDYLGFLPTF
GPCYINLYGSPREFTGFPDPYTELNTGKGEGVAYRGRLLLSLETKLVEHSEQKVEDLPADDILRV
EKYLRRRKYSLFAAFYSATMLQDVDDAIQFEVSIGNYGNKFDMTCLPLASTTQYSRAVFDGCH
YYYLPWGNVKPVVVLSSYWEDISHRIETQNQLLGIADRLEAGLEQVHLALKAQCSTEDVDSLV
AQLTDELIAGCSQPLGDIHETPSATHLDQYLYQLRTHHLSQITEAALALKLGHSELPAALEQAED
WLLRLRALAEEPQNSLPDIVIWMLQGDKRVAYQRVPAHQVLFSRRGANYCGKNCGKLQTIFLK
YPMEKVPGARMPVQIRVKLWFGLSVDEKEFNQFAEGKLSVFAETYENETKLALVGNWGTTGLT
YPKFSDVTGKIKLPKDSFRPSAGWTWAGDWFVCPEKTLLHDMDAGHLSFVEEVFENQTRLPGG
QWIYMSDNYTDVNGEKVLPKDDIECPLGWKWEDEEWSTDLNRAVDEQGWEYSITIPPERKPKH
WVPAEKMYYTHRRRRWVRLRRRDLSQMEALKRHRQAEAEGEGWEYASLFGWKFHLEYRKTD
AFRRRRWRRRMEPLEKTGPAAVFALEGALGGVMDDKSEDSMSVSTLSFGVNRPTISCIFDYGNR
YHLRCYMYQARDLAAMDKDSFSDPYAIVSFLHQSQKTVVVKNTLNPTWDQTLIFYEIEIFGEPA
TVAEQPPSIVVELYDHDTYGADEFMGRCICQPSLERMPRLAWFPLTRGCFVPGTLVNTENGLKKI
EEIKVGDKVFSHTGKLQEVVDTLIFDRDEEIISINGIDCTKNHEFYVIDKENANRVNEDNIHLFAR
WVHAEELDMKKHLLIELE*

- **C-ter fragment (SEQ ID NO. 119):**

MKFKLKEITSIETKHYKGKVHDLTVNQDHSYNVRGTVVHNSQPSGELLASFELIQREKPAIHHIP
GFEVQETSRILDESEDTDLPYPPPQREANIYMVPQNIKPALQRTAIEILAWGLRNMKSYQLANISS
PSLVVECGGQTVQSCVIRNLRKNPNFDICTLFMEVMLPREELYCPPITVKVIDNRQFGRRPVVGQ
CTIRSLESFLCDPYSAESPSPQGGPDDVSLLSPGEDVLIDIDDKEPLIPIQEEEFIDWWSKFFASIGER
EKCGSYLEKDFDTLKVYDTQLENVEAFEGLSDFCNTFKLYRGKTQEETEDPSVIGEFKGLFKIYP
LPEDPAIPMPPRQFHQLAAQGPQECLVRIYIVRAFGLQPKDPNGKCDPYIKISIGKKSVSDQDNYI
PCTLEPVFGKMFELTCTLPLEKDLKITLYDYDLLSKDEKIGETVVDLENRLLSKFGARCGLPQTY
CVSGPNQWRDQLRPSQLLHLFCQQHRVKAPVYRTDRVMFQDKEYSIEEIEAGRIPNPHLGPVEE
RLALHVLQQQGLVPEHVESRPLYSPLQPDIEQGKLQMWVDLFPKALGRPGPPFNITPRRARRFFL
RCIIWNTRDVILDDLSLTGEKMSDIYVKGWMIGFEEHKQKTDVHYRSLGGEGNFNWRFIFPFDY
LPAEQVCTIAKKDAFWRLDKTESKIPARVVFQIWDNDKFSFDDFLGSLQLDLNRMPKPAKTAKK
CSLDQLDDAFHPEWFVSLFEQKTVKGWWPCVAEEGEKKILAGKLEMTLEIVAESEHEERPAGQ
GRDEPNMNPKLEDPRRPDTSFLWFTSPYKTMKFILWRRFRWAIILFIILFILLLFLAIFIYAFPNYAA
MKLVKPFS*

<u>**Mini-Utrophin ΔSR5-15 (Site #1) or ΔSR6-14 (site #2):**</u>

1) **Split site #1 (using split IMPDH):**

- **N-ter fragment (SEQ ID NO. 120):**

MAKYGEHEASPDNGQNEFSDIIKSRSDEHNDVQKKTFTKWINARFSKSGKPPINDMFTDLKDGR
KLLDLLEGLTGTSLPKERGSTRVHALNNVNRVLQVLHQNNVELVNIGGTDIVDGNHKLTLGLL
WSIILHWQVKDVMKDVMSDLQQTNSEKILLSWVRQTTRPYSQVNVLNFTTSWTDGLAFNAVL
HRHKPDLFSWDKVVKMSPIERLEHAFSKAQTYLGIEKLLDPEDVAVQLPDKKSIIMYLTSLFEVL
PQQVTIDAIREVETLPRKYKKECEEEAINIQSTAPEEEHESPRAETPSTVTEVDMDLDSYQIALEEV
LTWLLSAEDTFQEQDDISDDVEEVKDQFATHEAFMMELTAHQSSVGSVLQAGNQLITQGTLSDE
EEFEIQEQMTLLNARWEALRVESMDRQSRLHDVLMELQKKQLQQLSAWLTLTEERIQKMETCP

LDDDVKSLQKLLEEHKSLQSDLEAEQVKVNSLTHMVVIVDENSGESATAILEDQLQKLGERWT
AVCRWTEERWNRLQEINILWQELLEEQCLLKAWLTEKEEALNKVQTSNFKDQKELSVSVRRLAI
LKEDMEMKRQTLDQLSEIGQDVGQLLDNSKASKKINSDSEELTQRWDSLVQRLEDSSNQVTQA
VAKLGMSQIPQKDLLETVRVREQAITKKSKQELPPPPPPKKRQIHVDIEAKKKFDAISAELLNWIL
KWKTAIQTTEIKEYMKMQDTSEMKKKLKALEKEQRERIPRADELNQTGQILVEQMGKEGLPTE
EIKNVLEKVSSEWKNVSQHLEDLERKIQLEEWRQFHCDLNDLTQWITEAEELLVDTCAPGGSLD
LEKARIHQQELEVGISSHQPSFAALNRTGDGIVQKLSQADGSFLKEKLAGLNQRWDAIVAEVKD
RQPRLKGESKQVMKYRHQLDEIICWLTKAEHAMQKRSTTELGENLQELRDLTQEMEVHAEKLK
WLNRTELEMLSDKSLSLPERDKISESLRTVNMTWNKICREVPTTLKECIQEPSSVSQTRIAAHPNV
QKVVLVSSASDIPVQSHRTSEISIPADLDKTITELADWLVLIDQMLKSNIVTVGDVEEINKTVSRM
KITKADLEQRHPQLDYVFTLAQNLKNKASSSDMRTAITEKLERVKNQWDGTQHGVELRQQQLE
DMIIDSLQWDDHREETEELMRKYEARLYILQQARRDPLTKQISDNQILLQELGPGDGIVMAFDN
VLQKLLEEYGCFVPGTLVNTENGLKKIEEIKVGDKVFSHTGKLQEVVDTLIFDRDEEIISINGIDCT
KNHEFYVIDKENANRVNEDNIHLFARWVHAEELDMKKHLLIELE*

- **C-ter fragment (SEQ ID NO. 121):**

MKFKLKEITSIETKHYKGKVHDLTVNQDHSYNVRGTVVHNSDDTRNVKETTEYLKTSWINLKQ
SIADRQNALEAEWRTVQASRRDLENFLKWIQEAETTVNVLVDASHRENALQDSILARELKQQM
QDIQAEIDAHNDIFKSIDGNRQKMVKALGNSEEATMLQHRLDDMNQRWNDLKAKSASIRAHLE
ASAEKWNRLLMSLEELIKWLNMKDEELKKQMPIGGDVPALQLQYDHCKALRRELKEKEYSVL
NAVDQARVFLADQPIEAPEEPRRNLQSKTELTPEERAQKIAKAMRKQSSEVKEKWESLNAVTSN
WQKQVDKALEKLRDLQGAMDDLDADMKEAESVRNGWKPVGDLLIDSLQDHIEKIMAFREEIA
PINFKVKTVNDLSSQLSPLDLHPSLKMSRQLDDLNMRWKLLQVSVDDRLKQLQEAHRDFGPSS
QHFLSTSVQLPWQRSISHNKVPYYINHQTQTTCWDHPKMTELFQSLADLNNVRFSAYRTAIKIRR
LQKALCLDLLELSTTNEIFKQHKLNQNDQLLSVPDVINCLTTTYDGLEQMHKDLVNVPLCVDM
CLNWLLNVYDTGRTGKIRVQSLKIGLMSLSKGLLEEKYRYLFKEVAGPTEMCDQRQLGLLLHD
AIQIPRQLGEVAAFGGSNIEPSVRSCFQQNNNKPEISVKEFIDWMHLEPQSMVWLPVLHRVAAAE
TAKHQAKCNICKECPIVGFRYRSLKHFNYDVCQSCFFSGRTAKGHKLHYPMVEYCIPTTSGEDV
RDFTKVLKNKFRSKKYFAKHPRLGYLPVQTVLEGDNLETPITLISMWPEHYDPSQSPQLFHDDT
HSRIEQYATRLAQMERTNGSFLTDSSSTTGSVEDEHALIQQYCQTLGGESPVSQPQSPAQILKSVE
REERGELERIIADLEEEQRNLQVEYEQLKDQHLRRGLPVGSPPESIISPHHTSEDSELIAEAKLLRQ
HKGRLEARMQILEDHNKQLESQLHRLRQLLEQPESDSRINGVSPWASPQHSALSYSLDPDASGP
QFHQAAGEDLLAPPHDTSTDLTEVMEQIHSTFPSCCPNVPSRPQAM*

2) **Split site #2 (using split Nrdj-1):**
   - **N-ter fragment (SEQ ID NO. 122):**

MAKYGEHEASPDNGQNEFSDIIKSRSDEHNDVQKKTFTKWINARFSKSGKPPINDMFTDLKDGR
KLLDLLEGLTGTSLPKERGSTRVHALNNVNRVLQVLHQNNVELVNIGGTDIVDGNHKLTLGLL
WSIILHWQVKDVMKDVMSDLQQTNSEKILLSWVRQTTRPYSQVNVLNFTTSWTDGLAFNAVL
HRHKPDLFSWDKVVKMSPIERLEHAFSKAQTYLGIEKLLDPEDVAVQLPDKKSIIMYLTSLFEVL
PQQVTIDAIREVETLPRKYKKECEEEAINIQSTAPEEEHESPRAETPSTVTEVDMDLDSYQIALEEV
LTWLLSAEDTFQEQDDISDDVEEVKDQFATHEAFMMELTAHQSSVGSVLQAGNQLITQGTLSDE
EEFEIQEQMTLLNARWEALRVESMDRQSRLHDVLMELQKKQLQQLSAWLTLTEERIQKMETCP
LDDDVKSLQKLLEEHKSLQSDLEAEQVKVNSLTHMVVIVDENSGESATAILEDQLQKLGERWT
AVCRWTEERWNRLQEINILWQELLEEQCLLKAWLTEKEEALNKVQTSNFKDQKELSVSVRRLAI
LKEDMEMKRQTLDQLSEIGQDVGQLLDNSKASKKINSDSEELTQRWDSLVQRLEDSSNQVTQA
VAKLGMSQIPQKDLLETVRVREQAITKKSKQELPPPPPPKKRQIHVDIEAKKKFDAISAELLNWIL
KWKTAIQTTEIKEYMKMQDTSEMKKKLKALEKEQRERIPRADELNQTGQILVEQMGKEGLPTE
EIKNVLEKVSSEWKNVSQHLEDLERKIQLQEDINAYFKQLDELEKVIKTKEEWVKHTSISESSRQ
SLPSLKDSCQRELTNLLGLHPKIEMARASCSALMSQPSAPDFVQRGFDSFLGRYQAVQEAVEDR
QQHLENELKLLPTDYLVEINKILLCMDDVELSLNVPELNTAIYEDFSFQEDSLKNIKDQLDKLGE
QIAVIHEKQPDVILEASGPEAIQIRDTLTQLNAKWDRINRMYSDRKGCFDRAMEEWRQFHCDLN
DLTQWITEAEELLVDTCAPGGSLDLEKARIHQQELEVGISSHQPSFAALNRTGDGIVQKLSQADG
SFLKEKLAGLNQRWDAIVAEVKDRQPRLKGESKQVMKYRHQLDEIICWLTKAEHAMQKRSTTE

LGENLQELRDLTQEMEVHAEKLKWLNRTELEMLSDKSLSLPERDKISESLRTVNMTWNKICREV
PTTLKECIQEPSSVSQTRIAAHPNVQKVVLVSSASDIPVQSHRTCCLVGSSEIITRNYGKTTIKEVV
EIFDNDKNIQVLAFNTHTDNIEWAPIKAAQLTRPNAELVELEIDTLHGVKTIRCTPDHPVYTKNR
GYVRADELTDDDELVVAI*

- **C-ter fragment (SEQ ID NO. 123):**

MEAKTYIGKLKSRKIVSNEDTYDIQTSTHNFFANDILVHNSEISIPADLDKTITELADWLVLIDQM
LKSNIVTVGDVEEINKTVSRMKITKADLEQRHPQLDYVFTLAQNLKNKASSSDMRTAITEKLER
VKNQWDGTQHGVELRQQQLEDMIIDSLQWDDHREETEELMRKYEARLYILQQARRDPLTKQIS
DNQILLQELGPGDGIVMAFDNVLQKLLEEYGSDDTRNVKETTEYLKTSWINLKQSIADRQNALE
AEWRTVQASRRDLENFLKWIQEAETTVNVLVDASHRENALQDSILARELKQQMQDIQAEIDAH
NDIFKSIDGNRQKMVKALGNSEEATMLQHRLDDMNQRWNDLKAKSASIRAHLEASAEKWNRL
LMSLEELIKWLNMKDEELKKQMPIGGDVPALQLQYDHCKALRRELKEKEYSVLNAVDQARVF
LADQPIEAPEEPRRNLQSKTELTPEERAQKIAKAMRKQSSEVKEKWESLNAVTSNWQKQVDKA
LEKLRDLQGAMDDLDADMKEAESVRNGWKPVGDLLIDSLQDHIEKIMAFREEIAPINFKVKTVN
DLSSQLSPLDLHPSLKMSRQLDDLNMRWKLLQVSVDDRLKQLQEAHRDFGPSSQHFLSTSVQLP
WQRSISHNKVPYYINHQTQTTCWDHPKMTELFQSLADLNNVRFSAYRTAIKIRRLQKALCLDLL
ELSTTNEIFKQHKLNQNDQLLSVPDVINCLTTTYDGLEQMHKDLVNVPLCVDMCLNWLLNVYD
TGRTGKIRVQSLKIGLMSLSKGLLEEKYRYLFKEVAGPTEMCDQRQLGLLLHDAIQIPRQLGEVA
AFGGSNIEPSVRSCFQQNNNKPEISVKEFIDWMHLEPQSMVWLPVLHRVAAAETAKHQAKCNIC
KECPIVGFRYRSLKHFNYDVCQSCFFSGRTAKGHKLHYPMVEYCIPTTSGEDVRDFTKVLKNKF
RSKKYFAKHPRLGYLPVQTVLEGDNLETPITLISMWPEHYDPSQSPQLFHDDTHSRIEQYATRLA
QMERTNGSFLTDSSSTTGSVEDEHALIQQYCQTLGGESPVSQPQSPAQILKSVEREERGELERIIA
DLEEEQRNLQVEYEQLKDQHLRRGLPVGSPPESIISPHHTSEDSELIAEAKLLRQHKGRLEARMQI
LEDHNKQLESQLHRLRQLLEQPESDSRINGVSPWASPQHSALSYSLDPDASGPQFHQAAGEDLL
APPHDTSTDLTEVMEQIHSTFPSCCPNVPSRPQAM*

CLAIMS

1. A method for delivering an exogenous polypeptide to a cell, the method comprising contacting the cell with:

a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a split intein; and

a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to a second portion of the split intein;

wherein the first and second fusion polypeptides are produced in the cell from the first and second nucleic acids, and wherein the first and second portions of the split intein promote joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, thereby delivering the exogenous polypeptide to the cell;

wherein the exogenous polypeptide delivered is larger than can be encoded by a single AAV vector particle.

2. The method of claim 1, wherein the first and second nucleic acids comprise a muscle-specific expression cassette (MSEC).

3. The method of claim 1 or claim 2, wherein the split intein is a naturally-occurring split intein.

4. The method of any one of claims 1-3, wherein the split intein is a genetically modified split intein.

5. The method of claim 4, wherein the genetic modification of the split intein is selected from codon optimization for expression and/or stability in mammalian cells, shortening or lengthening of the split intein, or changing encoded amino acids in the split intein to more closely match the sequence of the exogenous protein to be delivered.

6. The method of claim any one of claims 1-5, wherein the first and second portions of the exogenous polypeptide are substantially the same size.

7. The method of any one of claims 1-5, wherein the first and second portions of the exogenous polypeptide differ in size by no more than 50 amino acids.

8. The method of any one of claims 1-7, wherein the exogenous polypeptide comprises a footprint of less than four amino acids from the split intein.

9. The method of claim 8, wherein the exogenous polypeptide comprises a footprint of 3 or fewer amino acids from the split intein.

10. The method of claim 9, wherein the split site separating the first and second portions of the exogenous polypeptide is selected at a site having the same sequence as the split intein footprint, thereby producing the exogenous polypeptide without extra amino acids from the split intein.

11. The method of any one of claims 1-10, wherein the exogenous polypeptide is a therapeutic polypeptide.

12. The method of claim 11, wherein the therapeutic polypeptide is selected from dystrophin, mini-dystrophin, utrophin and dysferlin, nebulin, titin, myosin, spectrin repeat containing nuclear envelope protein 1 (Syne-1), dystroglycan, ATP synthase, clotting factor IIX, lamin A/C, thyroglobulin, epidermal growth factor receptor (EGFR), alpha- and/or beta spectrin, muscle target of rapamycin (mTOR), and ryanodine receptor 1.

13. The method of claim 12, wherein the mini-dystrophin is greater than 160kDa and smaller than full-length dystrophin.

14. The method of claim 12, wherein the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to the N-terminal portion of a split intein within or adjacent to a dystrophin hinge domain.

15. The method of claim 14, wherein the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

16. The method of claim 12, wherein the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

17. The method of claim 12, wherein the therapeutic polypeptide is dystrophin and the C-terminal portion of the dystrophin extein is joined to the C-terminal portion of the split intein within or adjacent to

a dystrophin hinge domain or to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

18. The method of claim 17, wherein the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

19. The method of any one of claims 1-18, wherein the exogenous polypeptide is functional in the cell.

20. A method for delivering an exogenous polypeptide to a cell, the method comprising contacting the cell with:

a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a first split intein, wherein the first portion of the split intein is fused to the carboxy terminus of the first portion of the exogenous polypeptide;

a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to (i) a second portion of the first split intein at the amino terminus of the second portion of the exogenous polypeptide and (ii) a first portion of a second split intein at the carboxy terminus of the second portion of the exogenous polypeptide; and

a third AAV vector particle comprising a third nucleic encoding a third fusion polypeptide comprising a third portion of the exogenous polypeptide fused to a second portion of the second split intein at the amino terminus of the third portion of the exogenous polypeptide,

wherein the first, second, and third fusion polypeptides are produced in the cell from the first, second and third nucleic acids, and wherein the respective portions of the first and second split inteins promote joining of (a) the carboxy terminus of the first portion of the exogenous polypeptide to the amino terminus of the second portion of the exogenous polypeptide and (b) the carboxy terminus of the second portion of the exogenous polypeptide to the amino terminus of the third portion of the exogenous polypeptide, thereby delivering the exogenous polypeptide to the cell;

wherein the exogenous polypeptide delivered is larger than can be encoded by a single AAV vector particle.

21. The method of claim 20, wherein the first and second split inteins do not cross-splice.

22. A protein expression system comprising a set of AAV vector particles comprising a first and second AAV particle,

wherein the first AAV vector particle comprises a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a split intein; and

wherein the second AAV vector particle comprises a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to a second portion of the split intein.

23. The protein expression system of claim 22, wherein co-infection of a cell with the first and second AAV vector particles promotes joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the split intein.

24. The protein expression system of claim 22 or 23, wherein joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the split intein generates an exogenous polypeptide larger than can be encoded in a single AAV particle.

25. A protein expression system comprising a set of AAV vector particles comprising a first, second, and third AAV particle,

wherein the first AAV vector particle comprises a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a first split intein, wherein the first portion of the split intein is fused to the carboxy terminus of the first portion of the exogenous polypeptide;

wherein the second AAV vector particle comprises a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to (i) a second portion of the first split intein at the amino terminus of the second portion of the exogenous polypeptide and (ii) a first portion of a second split intein at the carboxy terminus of the second portion of the exogenous polypeptide; and

wherein the third AAV vector particle comprises a third nucleic encoding a third fusion polypeptide comprising a third portion of the exogenous polypeptide fused to a second portion of the second split intein at the amino terminus of the third portion of the exogenous polypeptide.

26. The protein expression system of claim 25, wherein co-infection of a cell with the first, second and third AAV vector particles promotes joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the first

split intein, and joining of the second portion of the exogenous polypeptide to the third portion of the exogenous polypeptide, with removal of the first and second portions of the second split intein.

27. The protein expression system of claim 25 or 26, wherein joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, with removal of the first and second portions of the first split intein, and joining of the second portion of the exogenous polypeptide to the third portion of the exogenous polypeptide, with removal of the first and second portions of the second split intein generates an exogenous polypeptide larger than can be encoded in a single AAV particle.

28. The protein expression system of any one of claims 22-27, wherein expression of the first and second, or first, second and third fusion polypeptides is driven by a muscle-specific expression cassette.

29. A method of treating a disease or disorder in a subject in need thereof, the method comprising administering a protein expression system of any one of claims 22 - 28, thereby treating the subject.

30. The method of claim 29, wherein the subject in need thereof has a muscular or neuromuscular disease or disorder.

31. The method of claim 29 or 30, wherein the exogenous polypeptide is dystrophin or mini-dystrophin and the subject in need thereof has Duchenne muscular dystrophy (DMD) or Becker muscular dystrophy (BMD).

32. The method of claim 31, wherein the dystrophin or mini-dystrophin increases the strength of dystrophic muscles by at least 10%.

33. The method of any one of claims 29-32, wherein expression of the first and second, or first, second and third fusion polypeptides is driven by a muscle-specific expression cassette.

34. The method of any one of claims 29 - 33, wherein the protein expression system is administered by infusion into the vasculature, or by direct injection into a tissue.

35. A method for inducing the production of an exogenous polypeptide in a cell, the method comprising contacting the cell with:

a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a split intein; and

a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to a second portion of the split intein;

wherein the first and second fusion polypeptides are produced in the cell from the first and second nucleic acids, and wherein the first and second portions of the split intein promote joining of the first portion of the exogenous polypeptide to the second portion of the exogenous polypeptide, thereby inducing the production of the exogenous polypeptide in the cell;

wherein the exogenous polypeptide produced is larger than can be encoded by a single AAV vector particle.

36. The method of claim 35, wherein the first and second nucleic acids comprise a muscle-specific expression cassette (MSEC).

37. The method of claim 35 or 36, wherein the split intein is a naturally-occurring split intein.

38. The method of any one of claims 35-37, wherein the split intein is a genetically modified split intein.

39. The method of claim 38, wherein the genetic modification of the split intein is selected from codon optimization for expression and/or stability in mammalian cells, shortening or lengthening of the split intein, or changing encoded amino acids in the split intein to more closely match the sequence of the exogenous protein to be produced.

40. The method of any one of claims 35-39, wherein the first and second portions of the exogenous polypeptide are substantially the same size.

41. The method of any one of clams 35-40, wherein the first and second portions of the exogenous polypeptide differ in size by no more than 50 amino acids.

42. The method of claim any one of claims 35-41, wherein the exogenous polypeptide comprises a footprint of less than four amino acids from the split intein.

43. The method of claim 42, wherein the exogenous polypeptide comprises a split intein footprint of 3 or fewer amino acids.

44. The method of claim 43, wherein the split site separating the first and second portions of the exogenous polypeptide is selected at a site having the same sequence as the split intein footprint, thereby producing the exogenous polypeptide without extra amino acids from the split intein.

45. The method of any one of claims 35-44, wherein the exogenous polypeptide is a therapeutic polypeptide.

46. The method of claim 45, wherein the therapeutic polypeptide is selected from dystrophin, mini-dystrophin, utrophin and dysferlin, nebulin, titin, myosin, spectrin repeat containing nuclear envelope protein 1 (Syne-1), dystroglycan, ATP synthase, clotting factor IIX, lamin A/C, thyroglobulin, epidermal growth factor receptor (EGFR), alpha- and/or beta spectrin, muscle target of rapamycin (mTOR), and ryanodine receptor 1.

47. The method of claim 46, wherein the mini-dystrophin is greater than 160kDa and smaller than full-length dystrophin.

48. The method of claim 45 or 46, wherein the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to the N-terminal portion of a split intein within or adjacent to a dystrophin hinge domain.

49. The method of claim 48, wherein the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

50. The method of claim 45 or 46, wherein the therapeutic polypeptide is dystrophin and the N-terminal portion of the dystrophin extein is joined to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

51. The method of claim 45 or 46, wherein the therapeutic polypeptide is dystrophin and the C-terminal portion of the dystrophin extein is joined to the C-terminal portion of the split intein within or adjacent to a dystrophin hinge domain or to a loop domain joining helix b to helix c, or helix c to helix a' within one of the 24 dystrophin spectrin-like repeat domains.

52. The method of claim 51, wherein the hinge domain comprises hinge 1, 2, 3, or 4 of dystrophin.

53. The method of any one of claims 35-52, wherein the exogenous polypeptide is functional in the cell.

54. A method for inducing the production of an exogenous polypeptide in a cell, the method comprising contacting the cell with:

a first adeno-associated virus (AAV) vector particle comprising a first nucleic acid encoding a first fusion polypeptide comprising a first portion of an exogenous polypeptide fused to a first portion of a first split intein, wherein the first portion of the split intein is fused to the carboxy terminus of the first portion of the exogenous polypeptide;

a second AAV vector particle comprising a second nucleic acid encoding a second fusion polypeptide comprising a second portion of the exogenous polypeptide fused to (i) a second portion of the first split intein at the amino terminus of the second portion of the exogenous polypeptide and (ii) a first portion of a second split intein at the carboxy terminus of the second portion of the exogenous polypeptide; and

a third AAV vector particle comprising a third nucleic encoding a third fusion polypeptide comprising a third portion of the exogenous polypeptide fused to a second portion of the second split intein at the amino terminus of the third portion of the exogenous polypeptide,

wherein the first, second, and third fusion polypeptides are produced in the cell from the first, second and third nucleic acids, and wherein the respective portions of the first and second split inteins promote joining of (a) the carboxy terminus of the first portion of the exogenous polypeptide to the amino terminus of the second portion of the exogenous polypeptide and (b) the carboxy terminus of the second portion of the exogenous polypeptide to the amino terminus of the third portion of the exogenous polypeptide, thereby producing the exogenous polypeptide in the cell;

wherein the exogenous polypeptide produced is larger than can be encoded by a single AAV vector particle.

55. The method of claim 54, wherein the first and second split inteins do not cross-splice.

56. A composition of any one of claims 11-28 for use in the treatment of a disease or disorder in a subject in need thereof.

57. The composition of claim 56, wherein the subject in need thereof comprises a subject having a muscular or neuromuscular disorder.

FIG. 1A

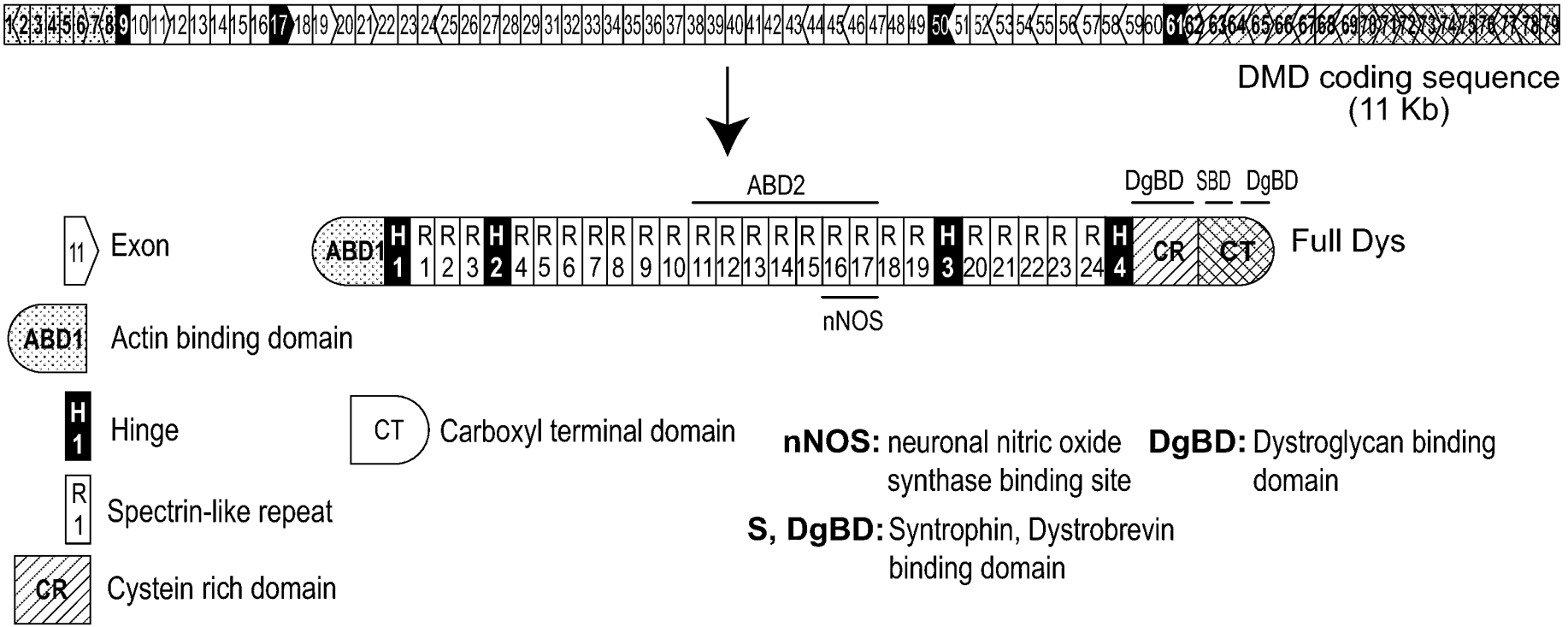
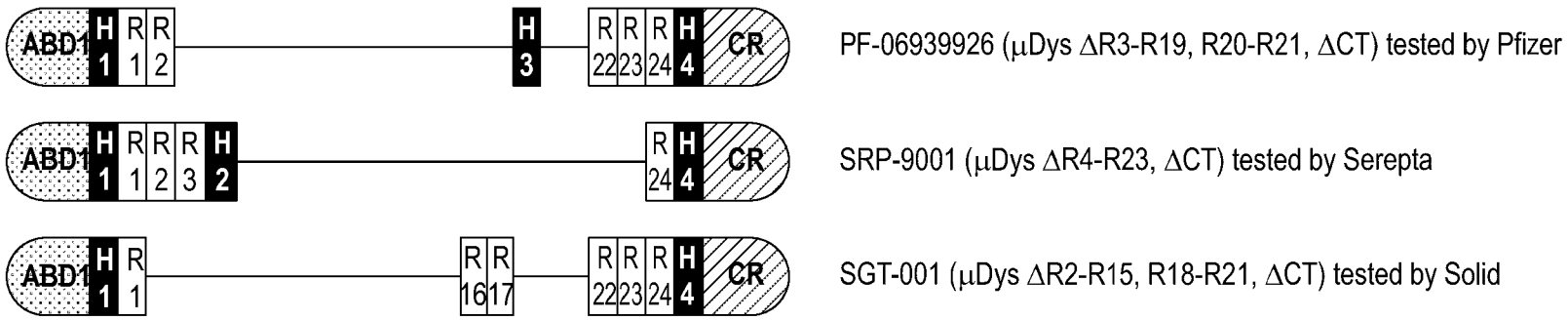

DMD coding sequence
(11 Kb)
ABD2
DgBD SBD DgBD
Full Dys
ABD1 H1 R1 R2 R3 H2 R4 R5 R6 R7 R8 R9 R10 R11 R12 R13 R14 R15 R16 R17 R18 R19 H3 R20 R21 R22 R23 R24 H4 CR CT
nNOS
11 Exon
ABD1 Actin binding domain
H1 Hinge
CT Carboxyl terminal domain
R1 Spectrin-like repeat
CR Cystein rich domain
nNOS: neuronal nitric oxide synthase binding site
DgBD: Dystroglycan binding domain
S, DgBD: Syntrophin, Dystrobrevin binding domain

FIG. 1B


ABD1 H1 R1 R2 H3 R22 R23 R24 H4 CR
PF-06939926 (μDys ΔR3-R19, R20-R21, ΔCT) tested by Pfizer
ABD1 H1 R1 R2 R3 H2 R24 H4 CR
SRP-9001 (μDys ΔR4-R23, ΔCT) tested by Serepta
ABD1 H1 R1 R16 R17 R22 R23 R24 H4 CR
SGT-001 (μDys ΔR2-R15, R18-R21, ΔCT) tested by Solid

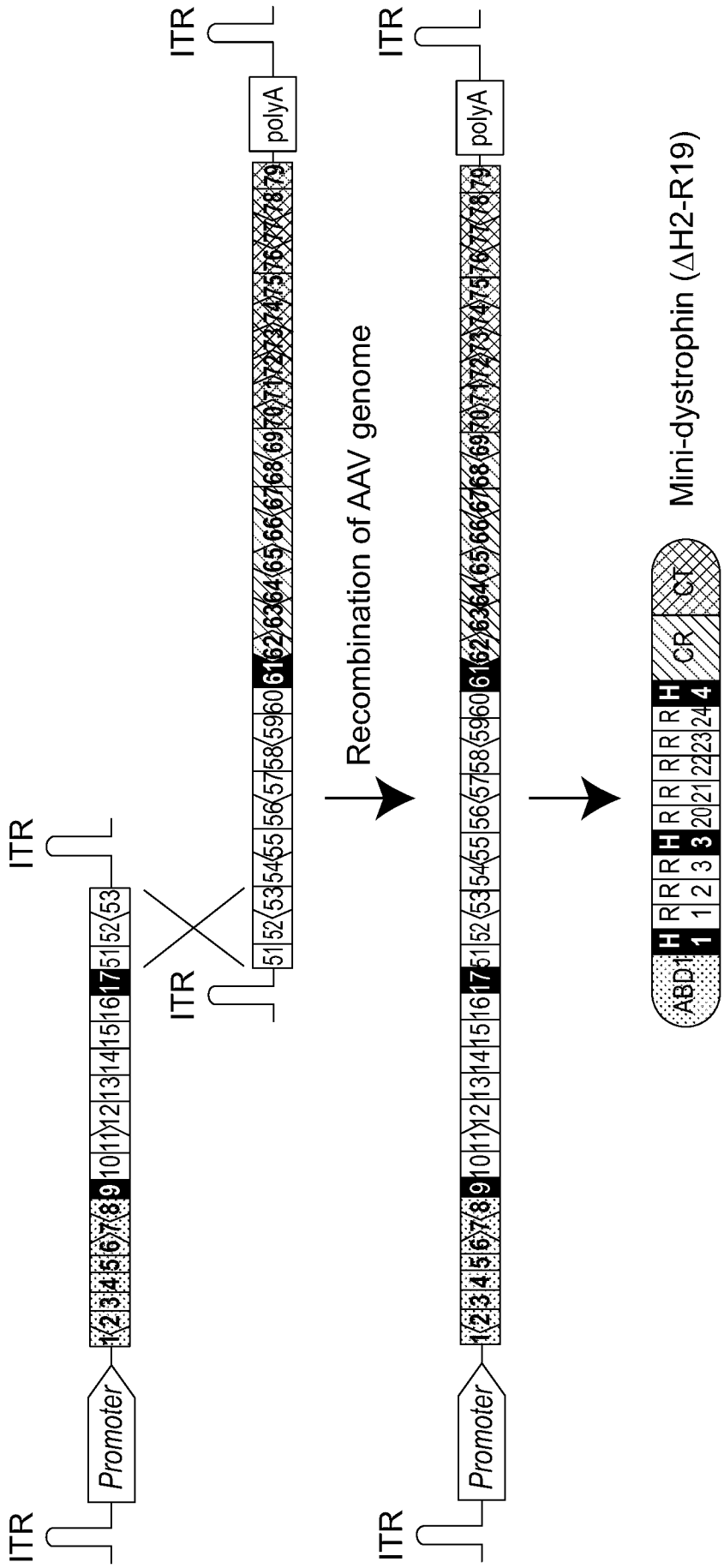

ITR
Promoter
ITR
ITR
polyA
ITR
Recombination of AAV genome
ITR
Promoter
ITR
polyA
ABD1
CR
CT
Mini-dystrophin (ΔH2-R19)

FIG. 2

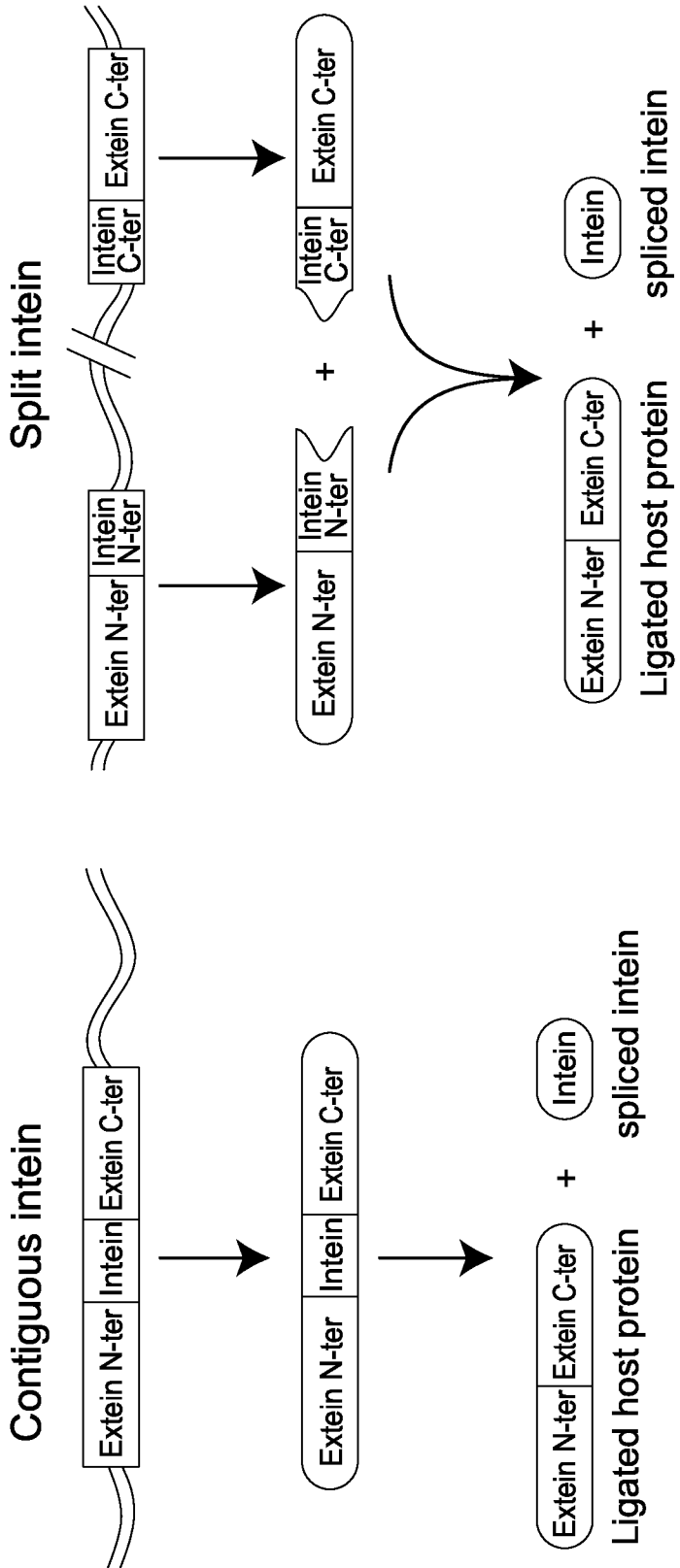

Split intein
Extein N-ter
Intein N-ter
Intein C-ter
Extein C-ter
Extein N-ter
Intein N-ter
+
Intein C-ter
Extein C-ter
Extein N-ter
Extein C-ter
+
Intein
Ligated host protein
spliced intein
Contiguous intein
Extein N-ter
Intein
Extein C-ter
Extein N-ter
Intein
Extein C-ter
Extein N-ter
Extein C-ter
+
Intein
Ligated host protein
spliced intein

FIG. 3

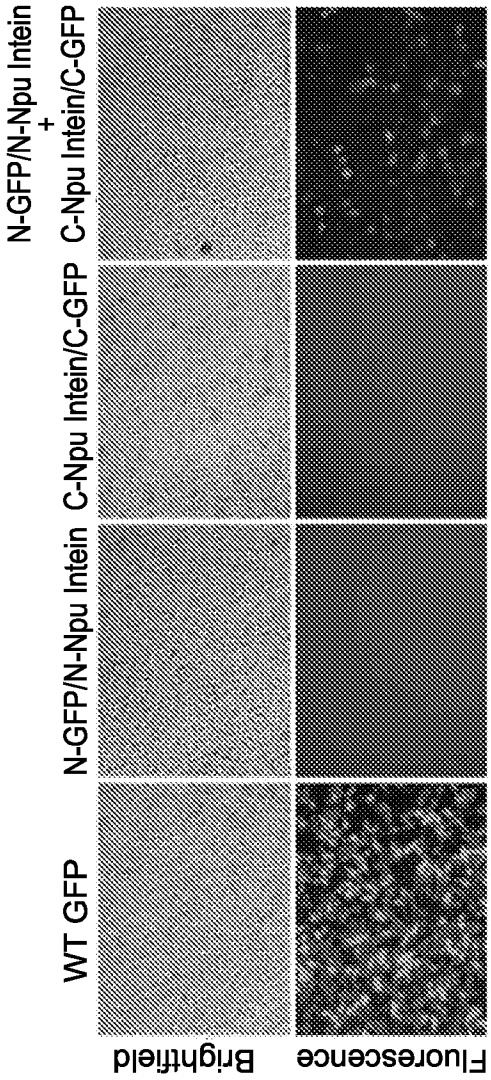

WT GFP
N-GFP/N-Npu Intein
C-Npu Intein/C-GFP
N-GFP/N-Npu Intein + C-Npu Intein/C-GFP
Brightfield
Fluorescence

FIG. 4A

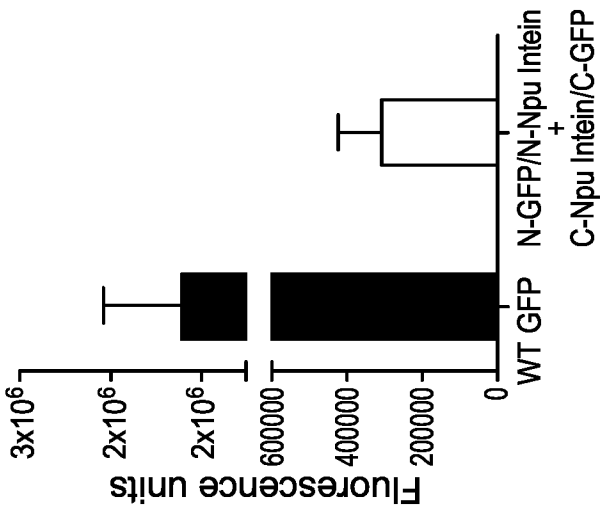

Fluorescence units
3x10^6
2x10^6
2x10^6
600000
400000
200000
0
WT GFP
N-GFP/N-Npu Intein + C-Npu Intein/C-GFP

FIG. 4B

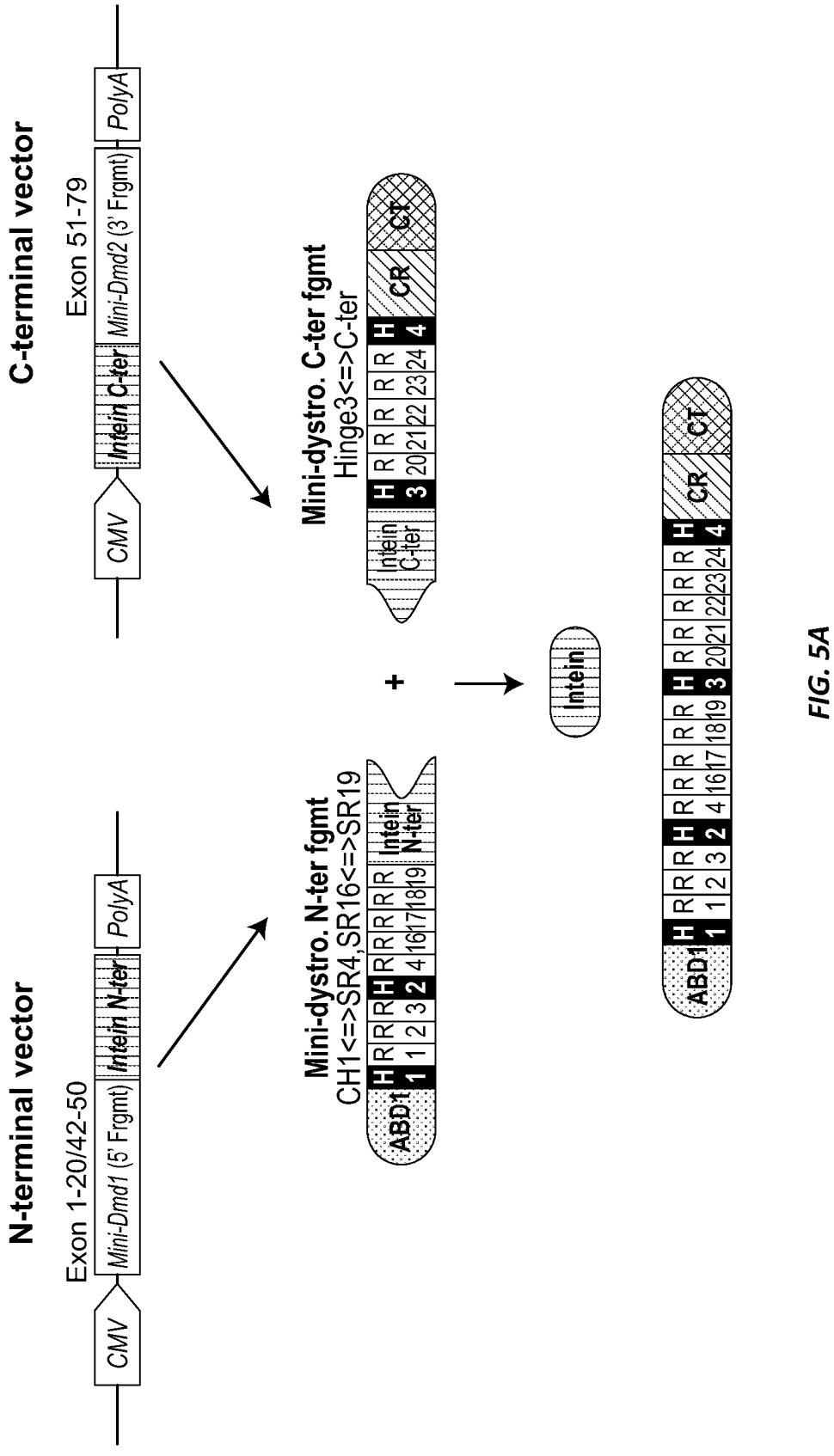

N-terminal vector
Exon 1-20/42-50
CMV
Mini-Dmd1 (5' Frgmt)
Intein N-ter
PolyA
C-terminal vector
Exon 51-79
CMV
Intein C-ter
Mini-Dmd2 (3' Frgmt)
PolyA
Mini-dystro. N-ter fgmt
CH1<=>SR4,SR16<=>SR19
ABD1
Intein N-ter
Mini-dystro. C-ter fgmt
Hinge3<=>C-ter
Intein C-ter
CR
CT
Intein

FIG. 5A

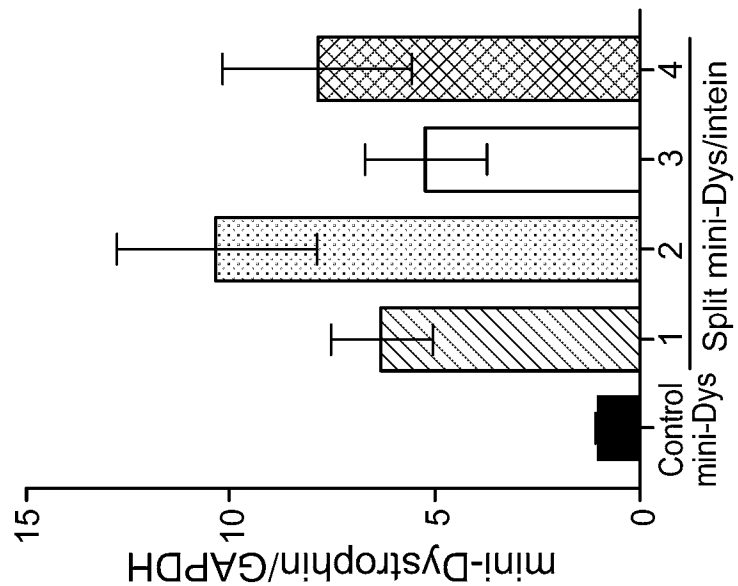

mini-Dystrophin/GAPDH
15
10
5
0
Control mini-Dys
1
2
3
4
Split mini-Dys/intein

FIG. 5C

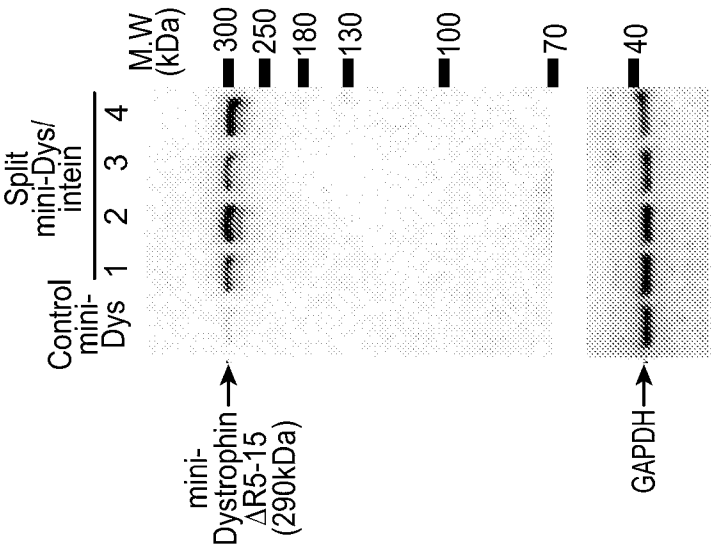

Control mini-Dys
Split mini-Dys/ intein
1 2 3 4
M.W (kDa)
300
250
180
130
100
70
40
mini-Dystrophin ΔR5-15 (290kDa)
GAPDH

FIG. 5B

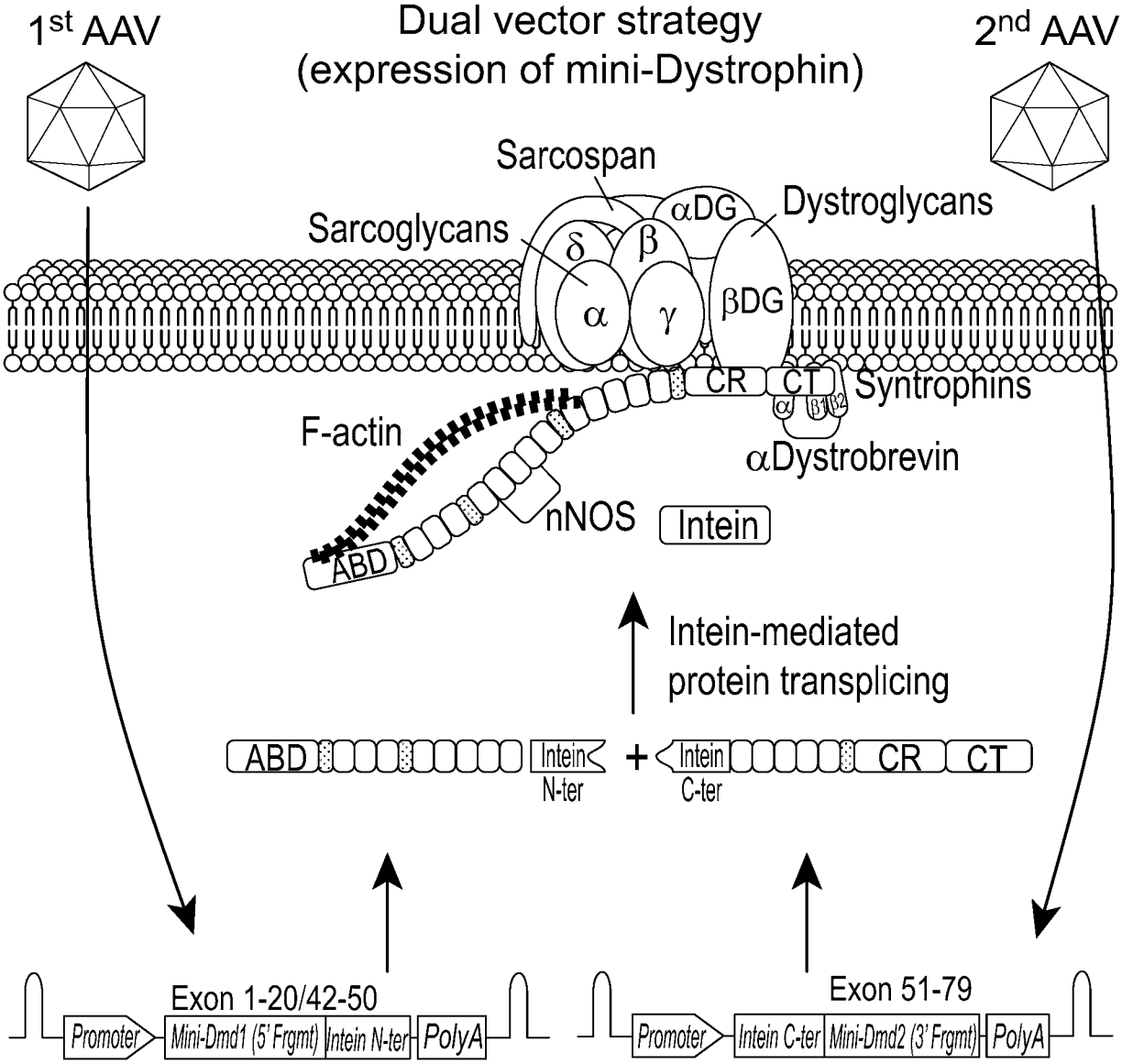

1st AAV
Dual vector strategy
(expression of mini-Dystrophin)
2nd AAV
Sarcospan
Sarcoglycans
αDG
Dystroglycans
δ
β
α
γ
βDG
CR
CT
Syntrophins
α
β1
β2
F-actin
αDystrobrevin
nNOS
Intein
ABD
Intein-mediated
protein transplicing
ABD
Intein
N-ter
+
Intein
C-ter
CR
CT
Exon 1-20/42-50
Promoter
Mini-Dmd1 (5' Frgmt)
Intein N-ter
PolyA
Exon 51-79
Promoter
Intein C-ter
Mini-Dmd2 (3' Frgmt)
PolyA

FIG. 6A

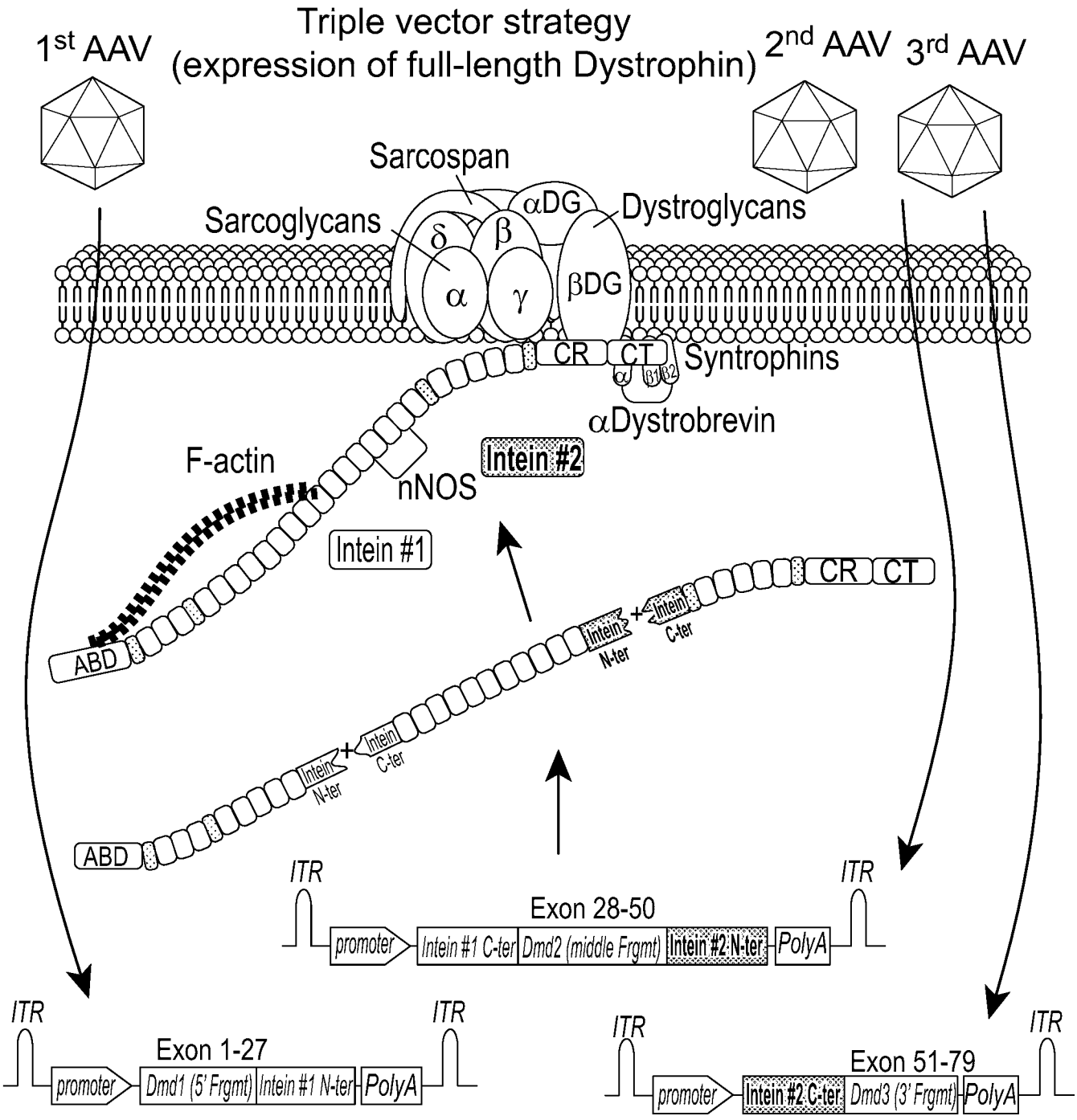

Triple vector strategy
(expression of full-length Dystrophin)
1st AAV
2nd AAV
3rd AAV
Sarcospan
Sarcoglycans
αDG
Dystroglycans
δ
β
α
γ
βDG
CR
CT
Syntrophins
αDystrobrevin
F-actin
nNOS
Intein #2
Intein #1
ABD
Intein N-ter
Intein C-ter
ITR
promoter
Exon 1-27
Dmd1 (5' Frgmt)
Intein #1 N-ter
PolyA
Exon 28-50
Intein #1 C-ter
Dmd2 (middle Frgmt)
Intein #2 N-ter
Exon 51-79
Intein #2 C-ter
Dmd3 (3' Frgmt)

FIG. 6B

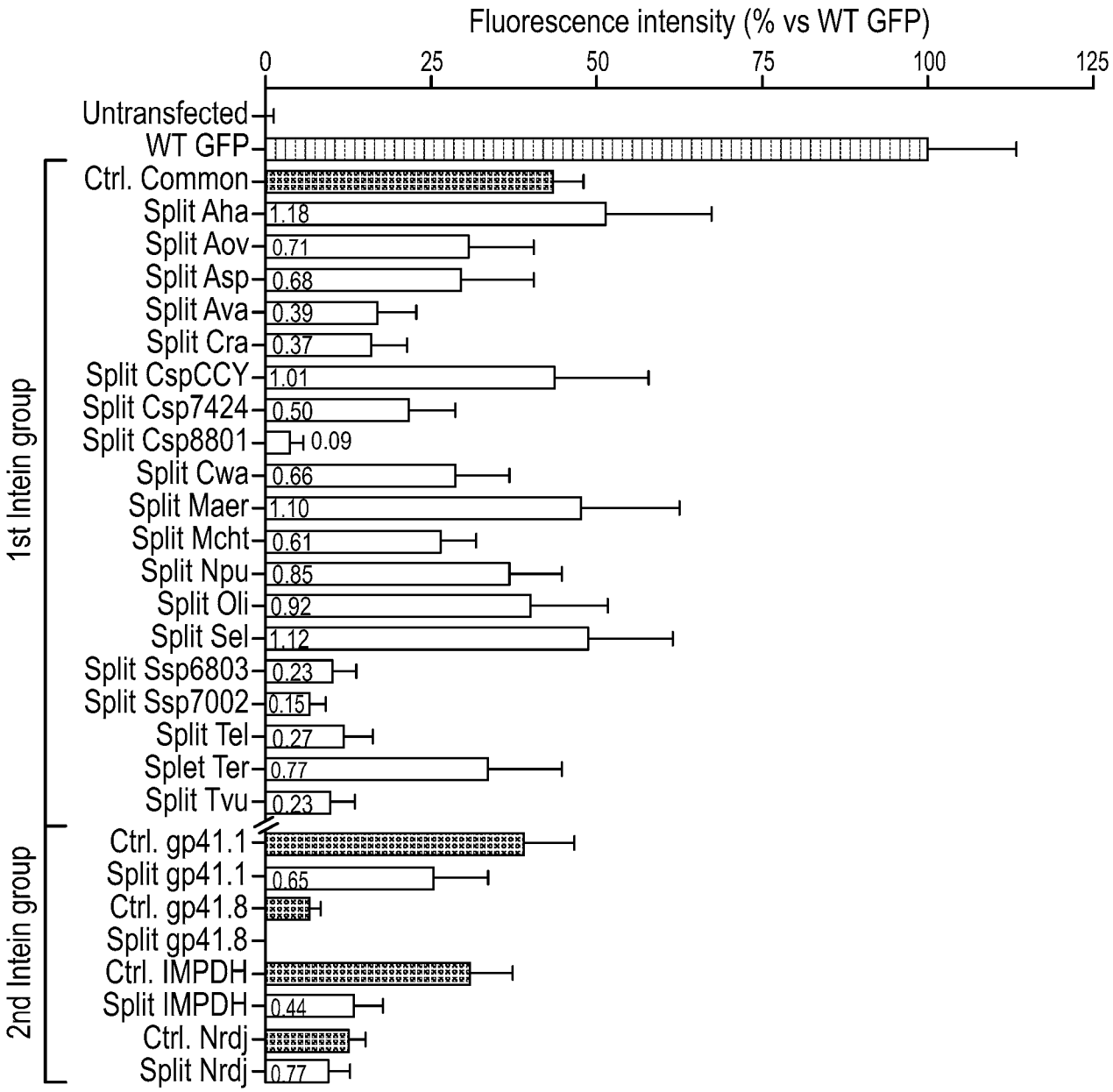

Fluorescence intensity (% vs WT GFP)
0
25
50
75
100
125
Untransfected
WT GFP
Ctrl. Common
Split Aha
1.18
Split Aov
0.71
Split Asp
0.68
Split Ava
0.39
Split Cra
0.37
Split CspCCY
1.01
Split Csp7424
0.50
Split Csp8801
0.09
Split Cwa
0.66
Split Maer
1.10
Split Mcht
0.61
Split Npu
0.85
Split Oli
0.92
Split Sel
1.12
Split Ssp6803
0.23
Split Ssp7002
0.15
Split Tel
0.27
Splet Ter
0.77
Split Tvu
0.23
Ctrl. gp41.1
Split gp41.1
0.65
Ctrl. gp41.8
Split gp41.8
Ctrl. IMPDH
Split IMPDH
0.44
Ctrl. Nrdj
Split Nrdj
0.77
1st Intein group
2nd Intein group

FIG. 7

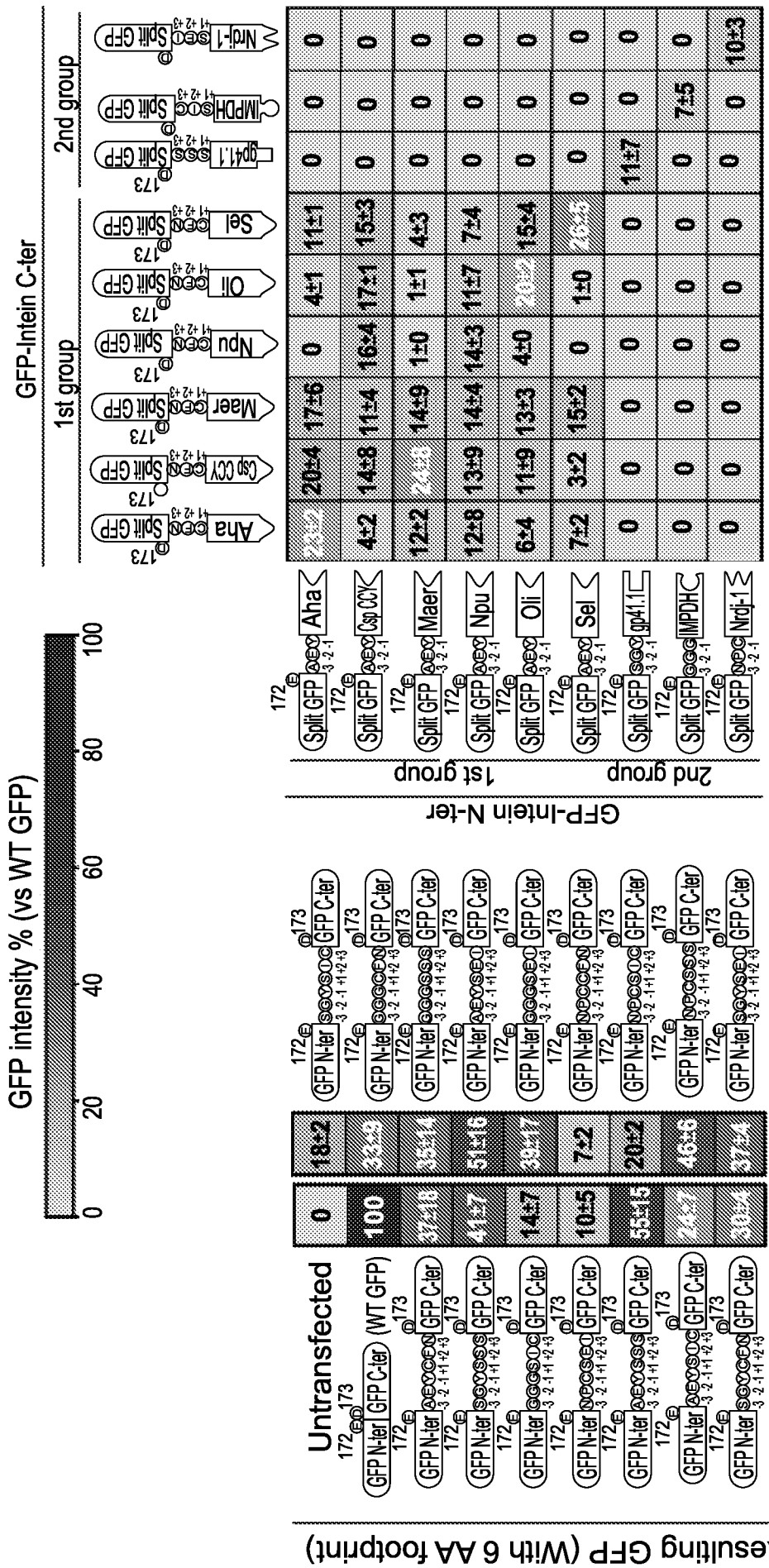

Resulting GFP (With 6 AA footprint)
Untransfected
100
18±2
0
GFP intensity % (vs WT GFP)
0
20
40
60
80
100
GFP-Intein N-ter
1st group
2nd group
GFP-Intein C-ter
1st group
2nd group
Aha
Csp CCK
Maer
Npu
Oli
Sel
gp41.1
IMPDH
Nrdj-1
Split GFP
GFP N-ter
GFP C-ter
20±4
17±6
4±1
11±1
14±8
11±4
16±4
17±1
15±3
14±9
1±0
1±1
4±3
13±9
14±4
14±3
11±7
7±4
11±9
13±3
4±0
15±4
3±2
15±2
0
1±0
11±7
7±5
10±3

FIG. 8

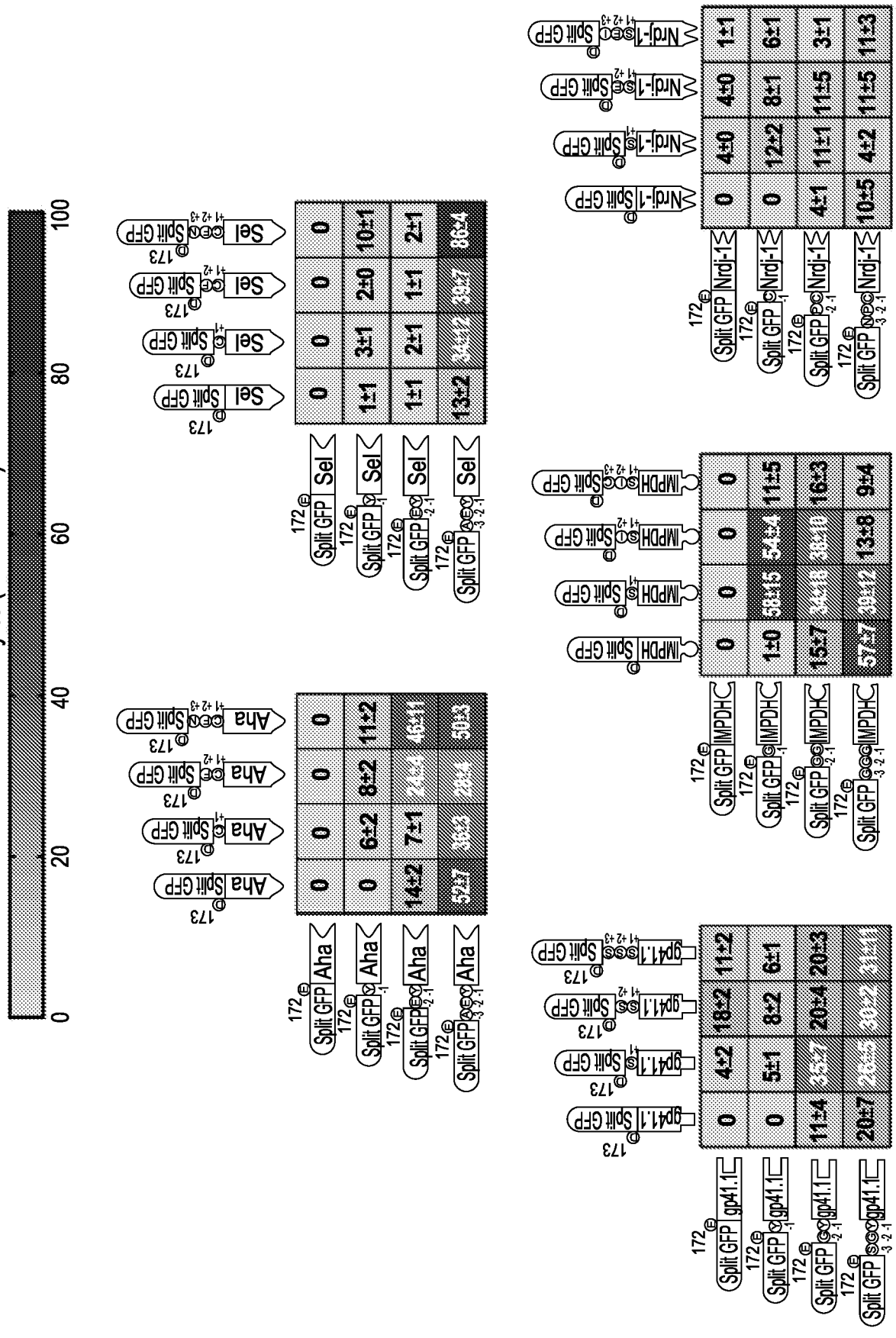

GFP intensity % (vs WT GFP)
0
20
40
60
80
100
Split GFP Aha
Split GFP Sel
Split GFP gp41.1
Split GFP IMPDHC
Split GFP Nrdj-1
Aha Split GFP
Sel Split GFP
gp41.1 Split GFP
IMPDH Split GFP
Nrdj-1 Split GFP
0 0 0
0 0 6±2 8±2 11±2
0 0 7±1
14±2
0 0 0
1±1 3±1 2±0 10±1
1±1 2±1 1±1 2±1
13±2 86±4
0 0 0
0 4±2 18±2 11±2
0 5±1 8±2 6±1
11±4 20±3
20±7
0 0 0
1±0 11±5
15±7 16±3
37±7 13±8 9±4
0 0 0
0 4±0 4±0 1±1
0 12±2 8±1 6±1
4±1 11±1 11±5 3±1
10±5 4±2 11±5 11±3

FIG. 9A

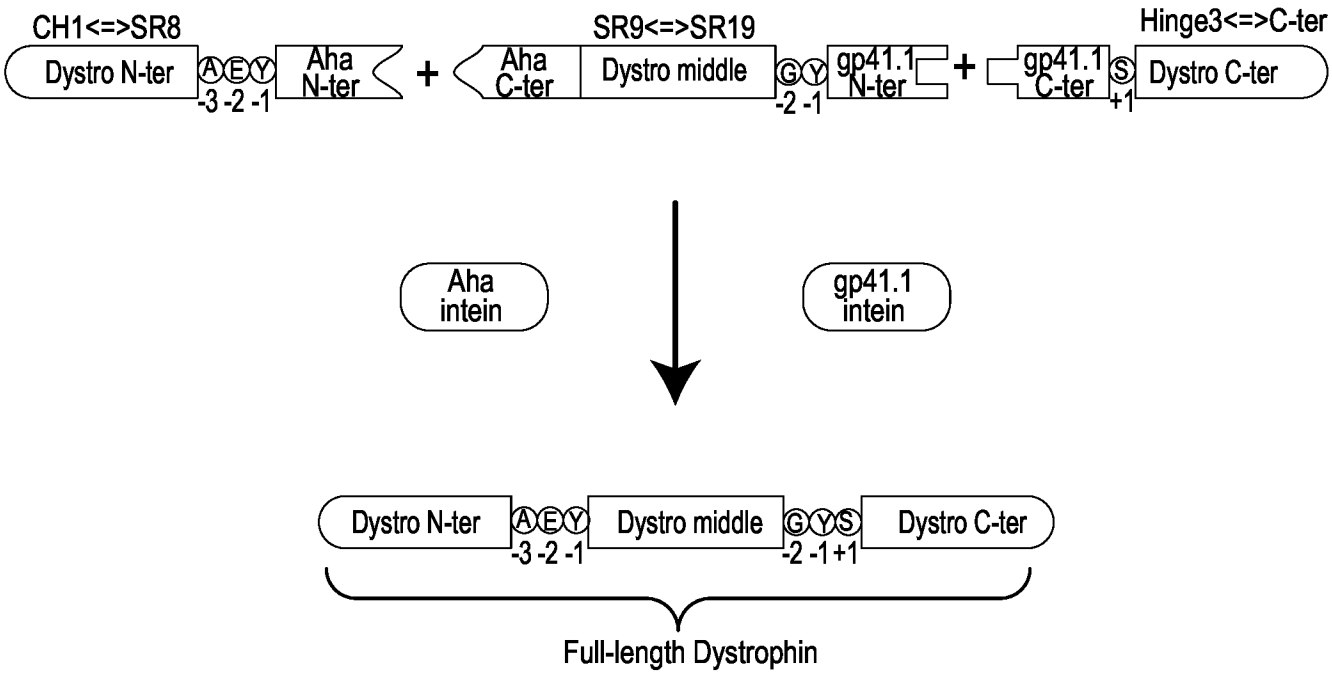

CH1<=>SR8
Dystro N-ter
A E Y
-3 -2 -1
Aha N-ter
+
Aha C-ter
SR9<=>SR19
Dystro middle
G Y
-2 -1
gp41.1 N-ter
+
gp41.1 C-ter
S
+1
Hinge3<=>C-ter
Dystro C-ter
Aha intein
gp41.1 intein
Dystro N-ter
A E Y
-3 -2 -1
Dystro middle
G Y S
-2 -1 +1
Dystro C-ter
Full-length Dystrophin

FIG. 9B

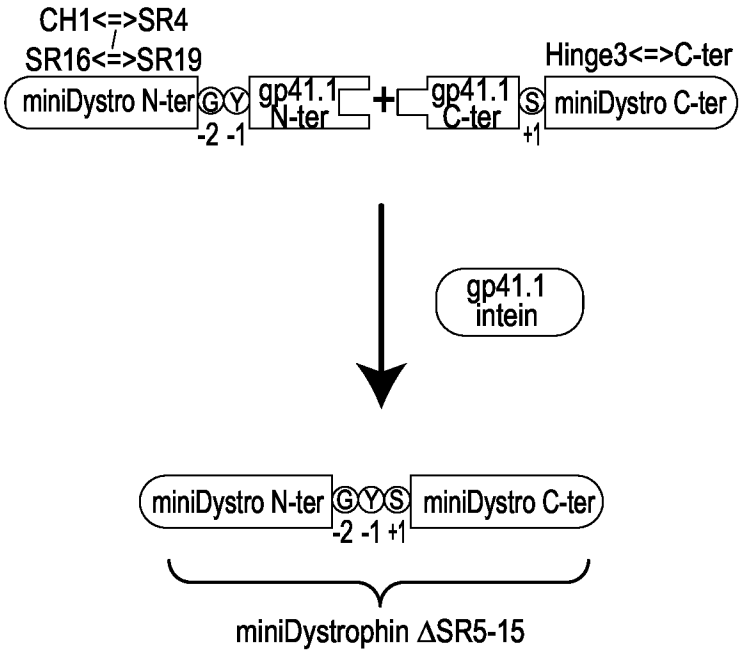

CH1<=>SR4
SR16<=>SR19
miniDystro N-ter
G Y
-2 -1
gp41.1 N-ter
+
gp41.1 C-ter
S
+1
Hinge3<=>C-ter
miniDystro C-ter
gp41.1 intein
miniDystro N-ter
G Y S
-2 -1 +1
miniDystro C-ter
miniDystrophin ΔSR5-15

FIG. 9C

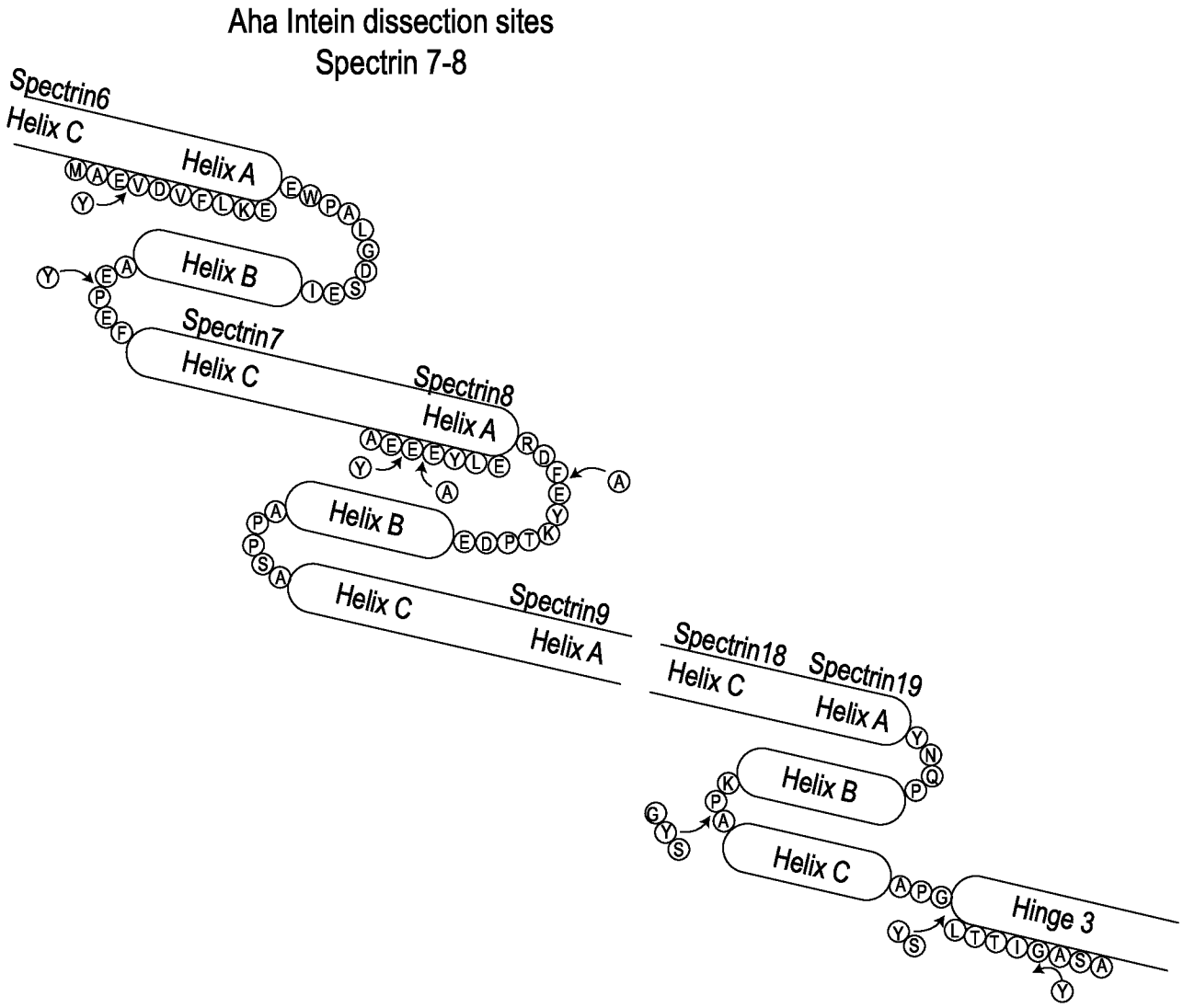

Aha Intein dissection sites
Spectrin 7-8
Spectrin6
Helix C
Helix A
Helix B
Spectrin7
Helix C
Spectrin8
Helix A
Helix B
Helix C
Spectrin9
Helix A
Spectrin18
Helix C
Spectrin19
Helix A
Helix B
Helix C
Hinge 3
gp41.1 Intein dissection sites
Spectrin 19-Hinge3

FIG. 10

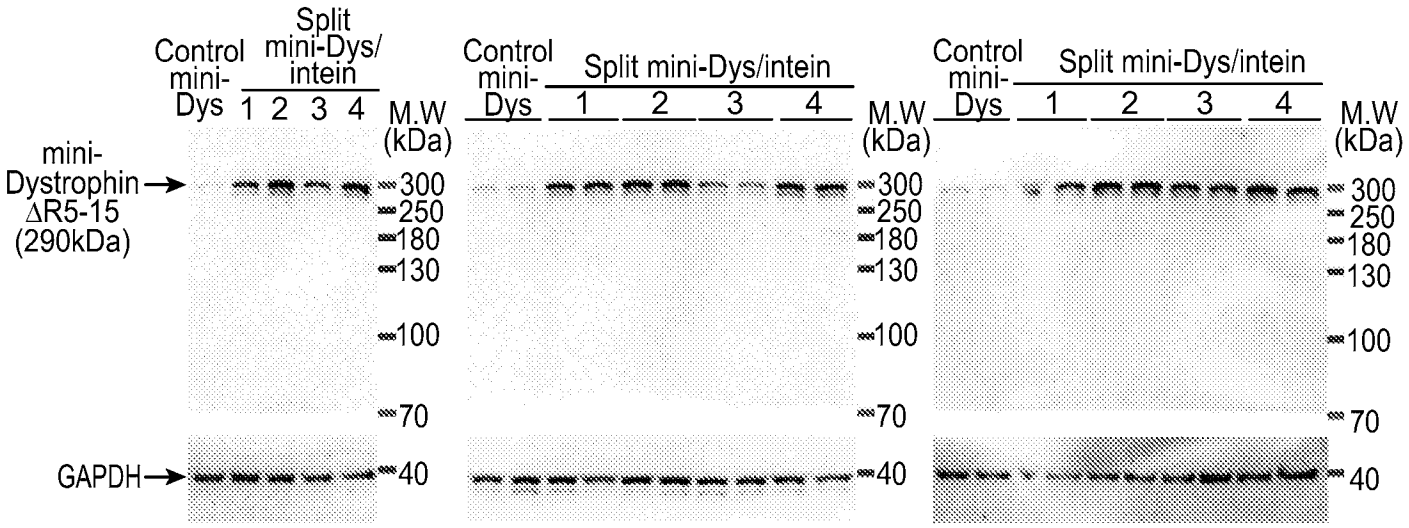


FIG. 11A

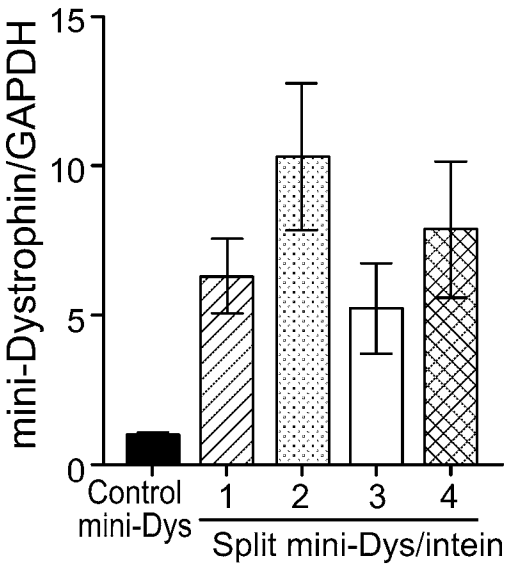


FIG. 11B

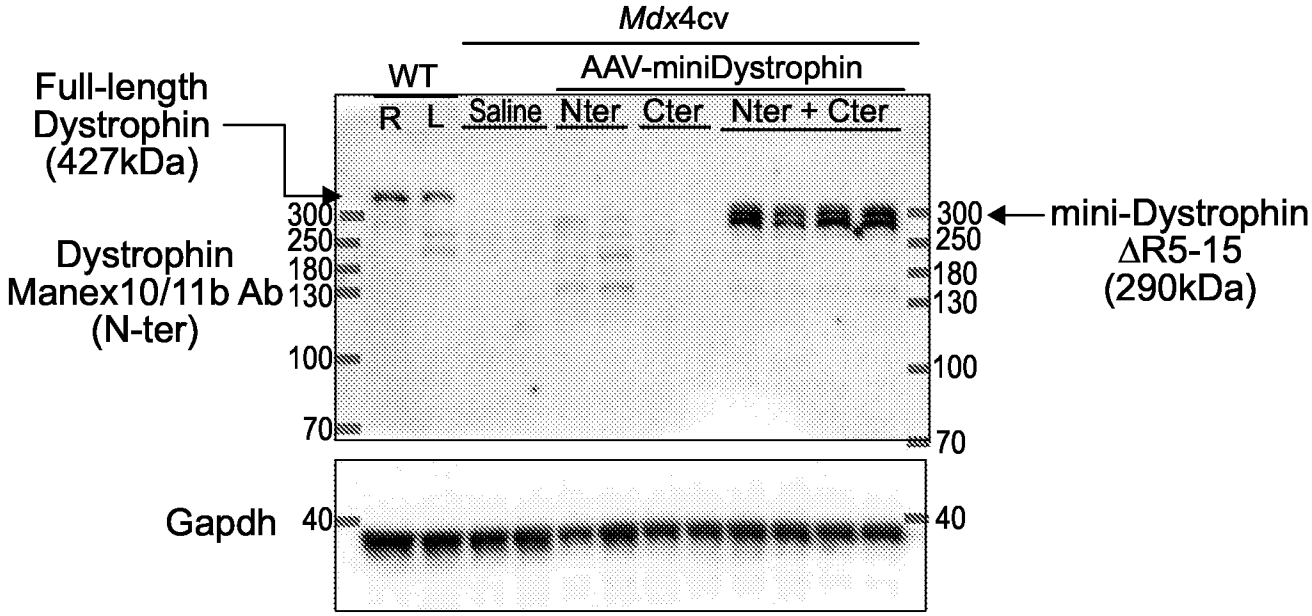


FIG. 12A

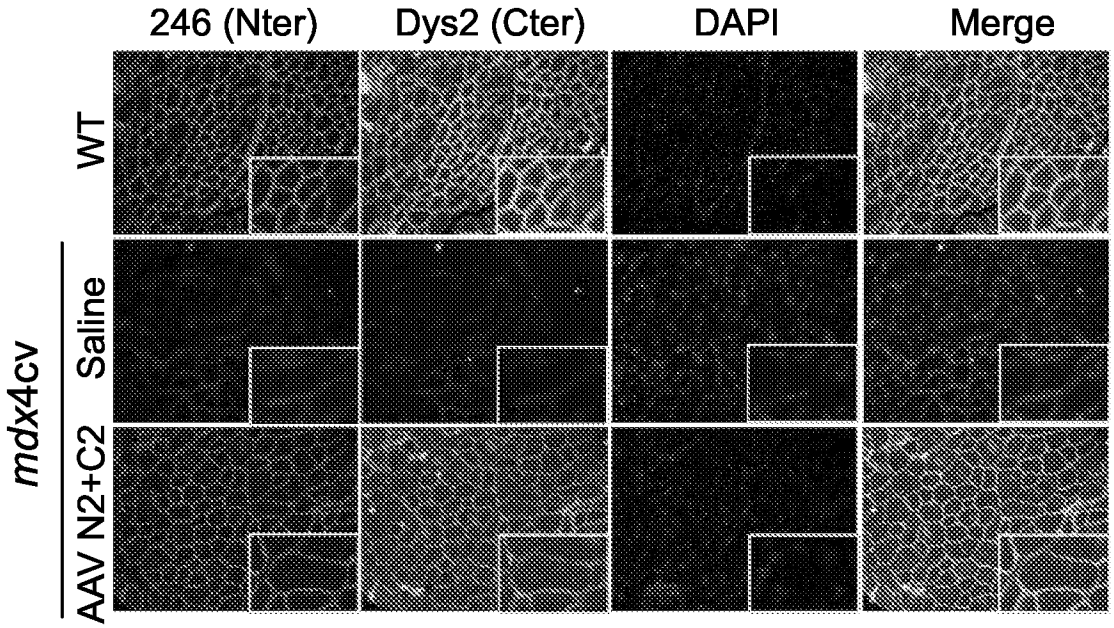


FIG. 12B

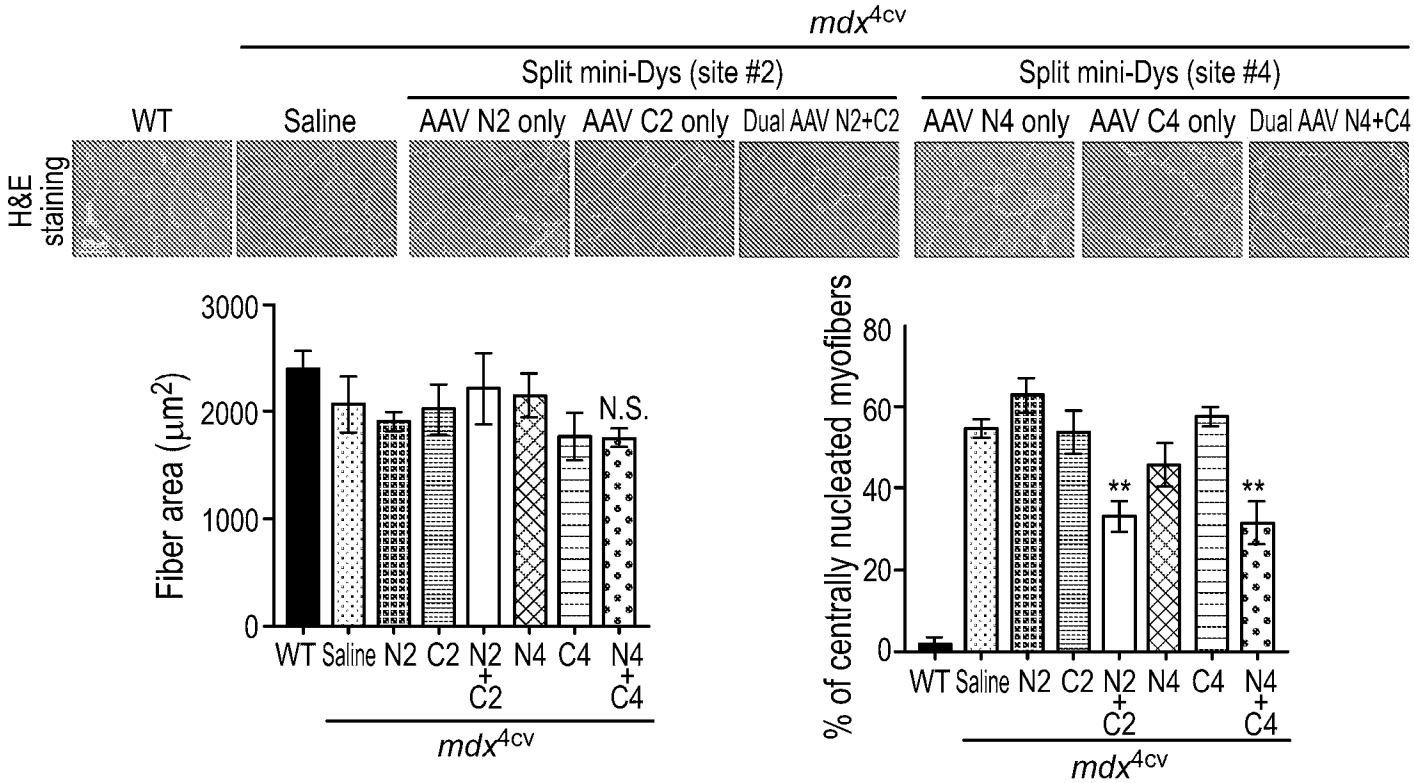


FIG. 12C

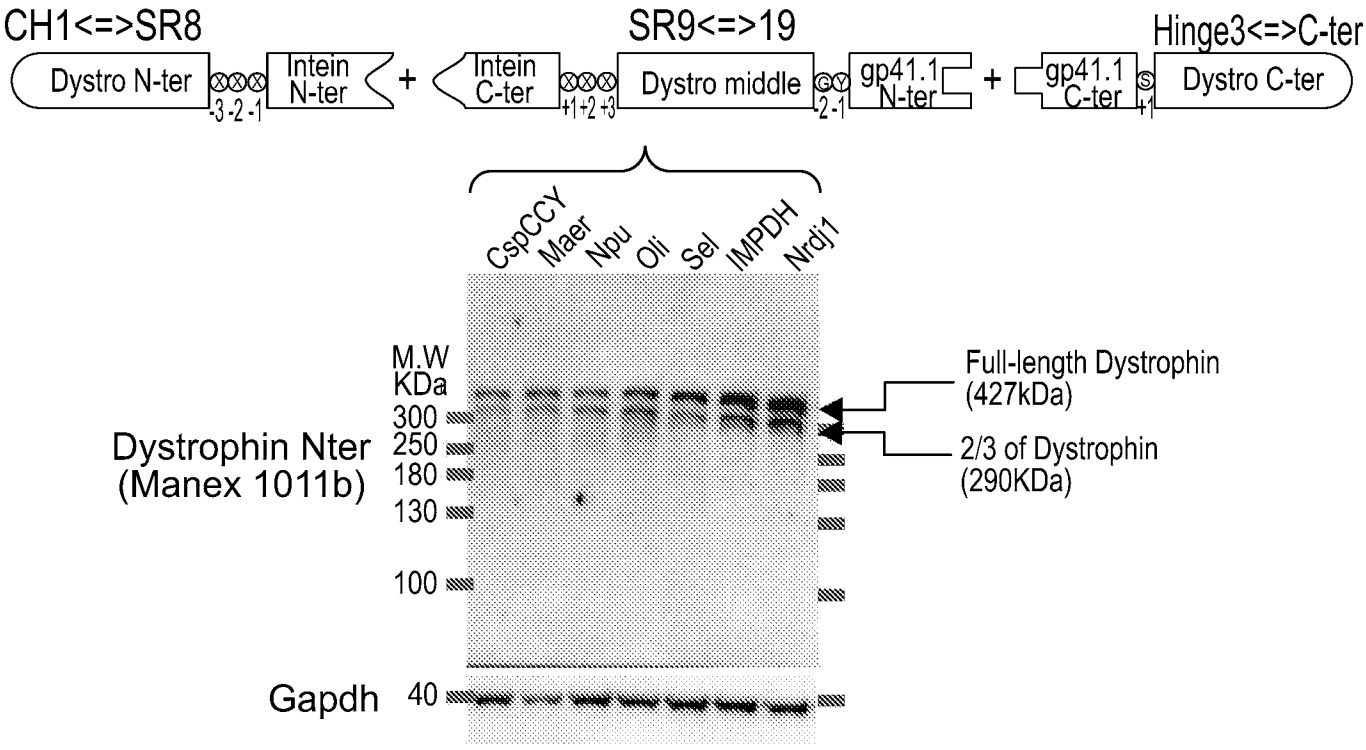


FIG. 13

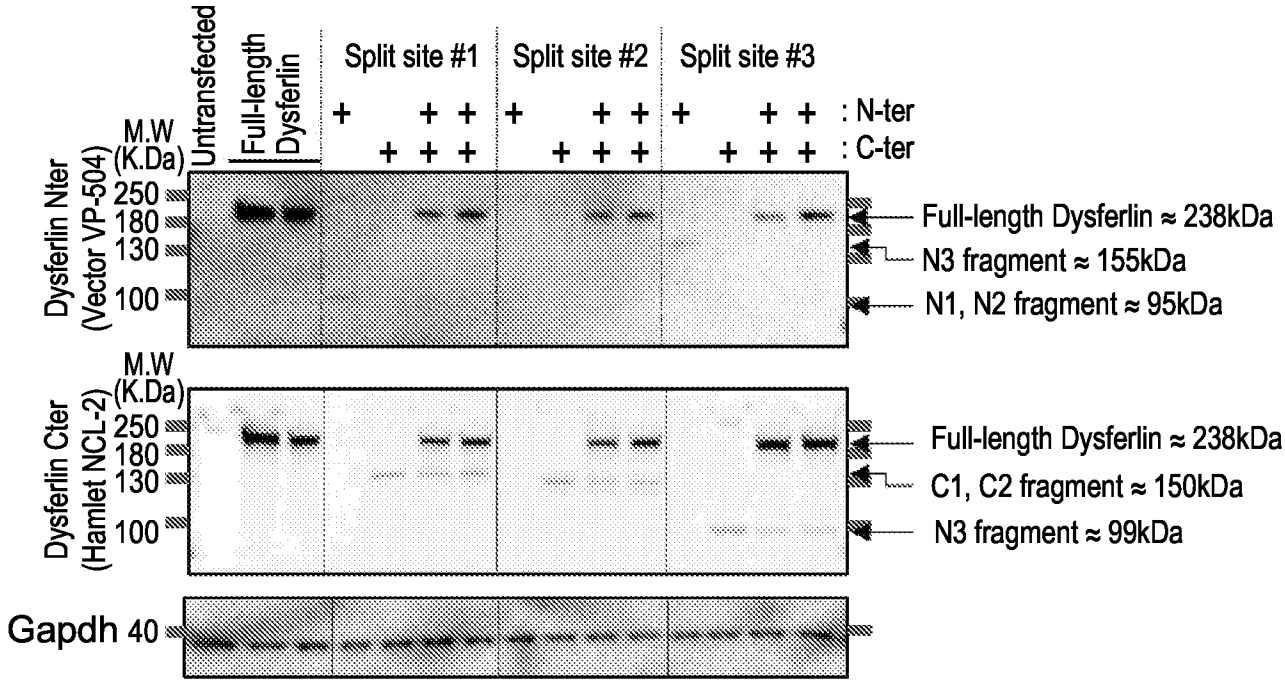


FIG. 14A

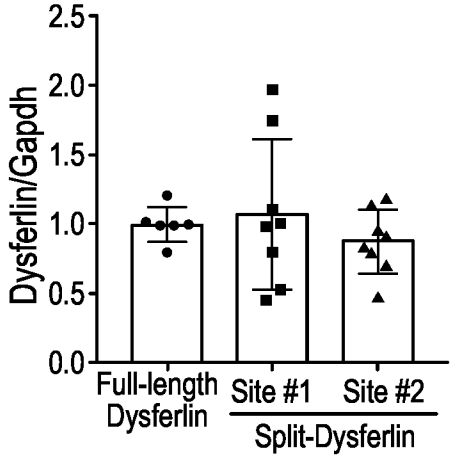


FIG. 14B

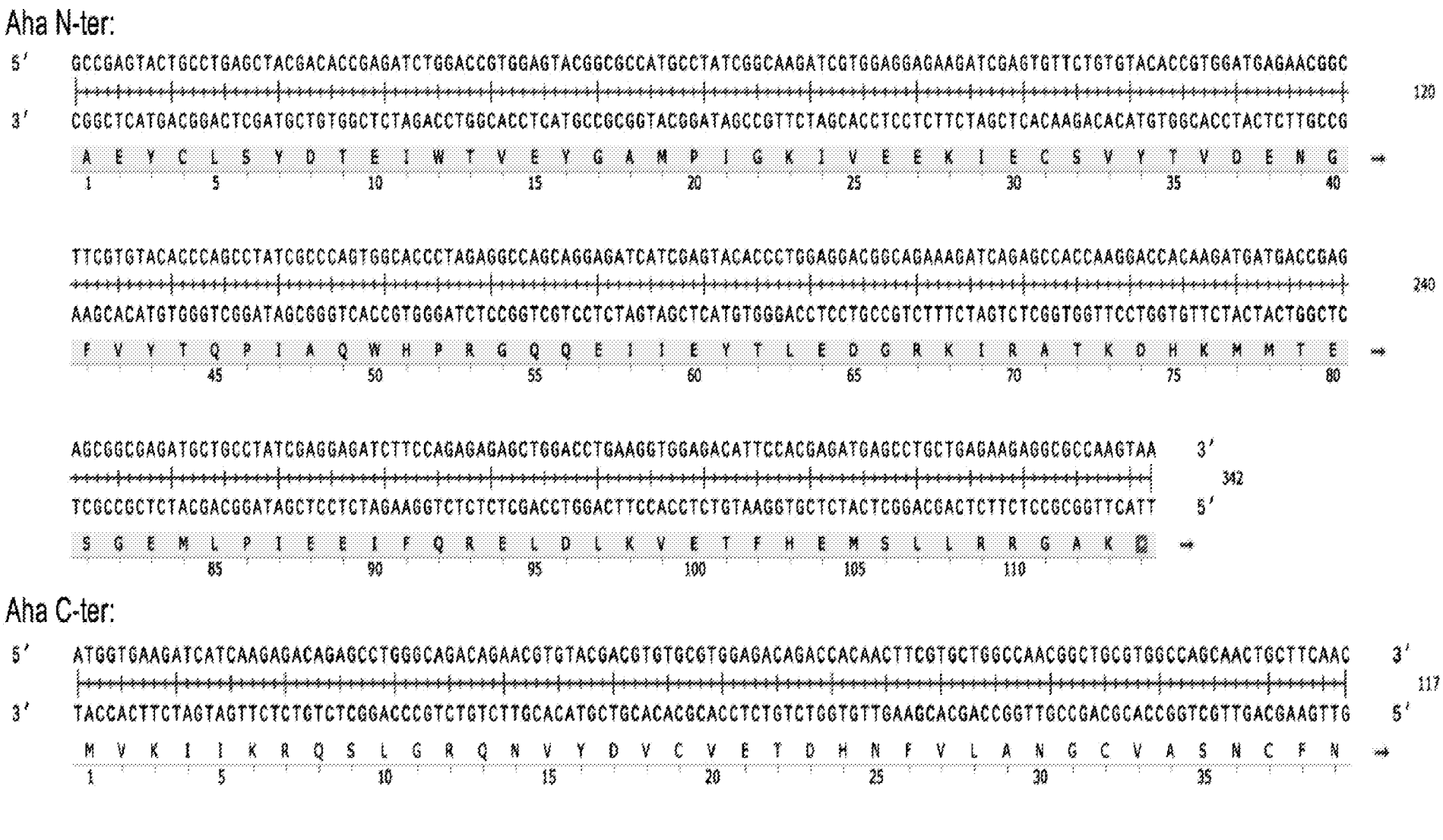

Aha N-ter:
5′ GCCGAGTACTGCCTGAGCTACGACACCGAGATCTGGACCGTGGAGTACGGCGCCATGCCTATCGGCAAGATCGTGGAGGAGAAGATCGAGTGTTCTGTGTACACCGTGGATGAGAACGGC
120
3′ CGGCTCATGACGGACTCGATGCTGTGGCTCTAGACCTGGCACCTCATGCCGCGGTACGGATAGCCGTTCTAGCACCTCCTCTTCTAGCTCACAAGACACATGTGGCACCTACTCTTGCCG
A E Y C L S Y D T E I W T V E Y G A M P I G K I V E E K I E C S V Y T V D E N G
1 5 10 15 20 25 30 35 40
TTCGTGTACACCCAGCCTATCGCCCAGTGGCACCCTAGAGGCCAGCAGGAGATCATCGAGTACACCCTGGAGGACGGCAGAAAGATCAGAGCCACCAAGGACCACAAGATGATGACCGAG
240
AAGCACATGTGGGTCGGATAGCGGGTCACCGTGGGATCTCCGGTCGTCCTCTAGTAGCTCATGTGGGACCTCCTGCCGTCTTTCTAGTCTCGGTGGTTCCTGGTGTTCTACTACTGGCTC
F V Y T Q P I A Q W H P R G Q Q E I I E Y T L E D G R K I R A T K D H K M M T E
45 50 55 60 65 70 75 80
AGCGGCGAGATGCTGCCTATCGAGGAGATCTTCCAGAGAGAGCTGGACCTGAAGGTGGAGACATTCCACGAGATGAGCCTGCTGAGAAGAGGCGCCAAGTAA 3′
342
TCGCCGCTCTACGACGGATAGCTCCTCTAGAAGGTCTCTCTCGACCTGGACTTCCACCTCTGTAAGGTGCTCTACTCGGACGACTCTTCTCCGCGGTTCATT 5′
S G E M L P I E E I F Q R E L D L K V E T F H E M S L L R R G A K
85 90 95 100 105 110
Aha C-ter:
5′ ATGGTGAAGATCATCAAGAGACAGAGCCTGGGCAGACAGAACGTGTACGACGTGTGCGTGGAGACAGACCACAACTTCGTGCTGGCCAACGGCTGCGTGGCCAGCAACTGCTTCAAC 3′
117
3′ TACCACTTCTAGTAGTTCTCTGTCTCGGACCCGTCTGTCTTGCACATGCTGCACACGCACCTCTGTCTGGTGTTGAAGCACGACCGGTTGCCGACGCACCGGTCGTTGACGAAGTTG 5′
M V K I I K R Q S L G R Q N V Y D V C V E T D H N F V L A N G C V A S N C F N
1 5 10 15 20 25 30 35

FIG. 15A

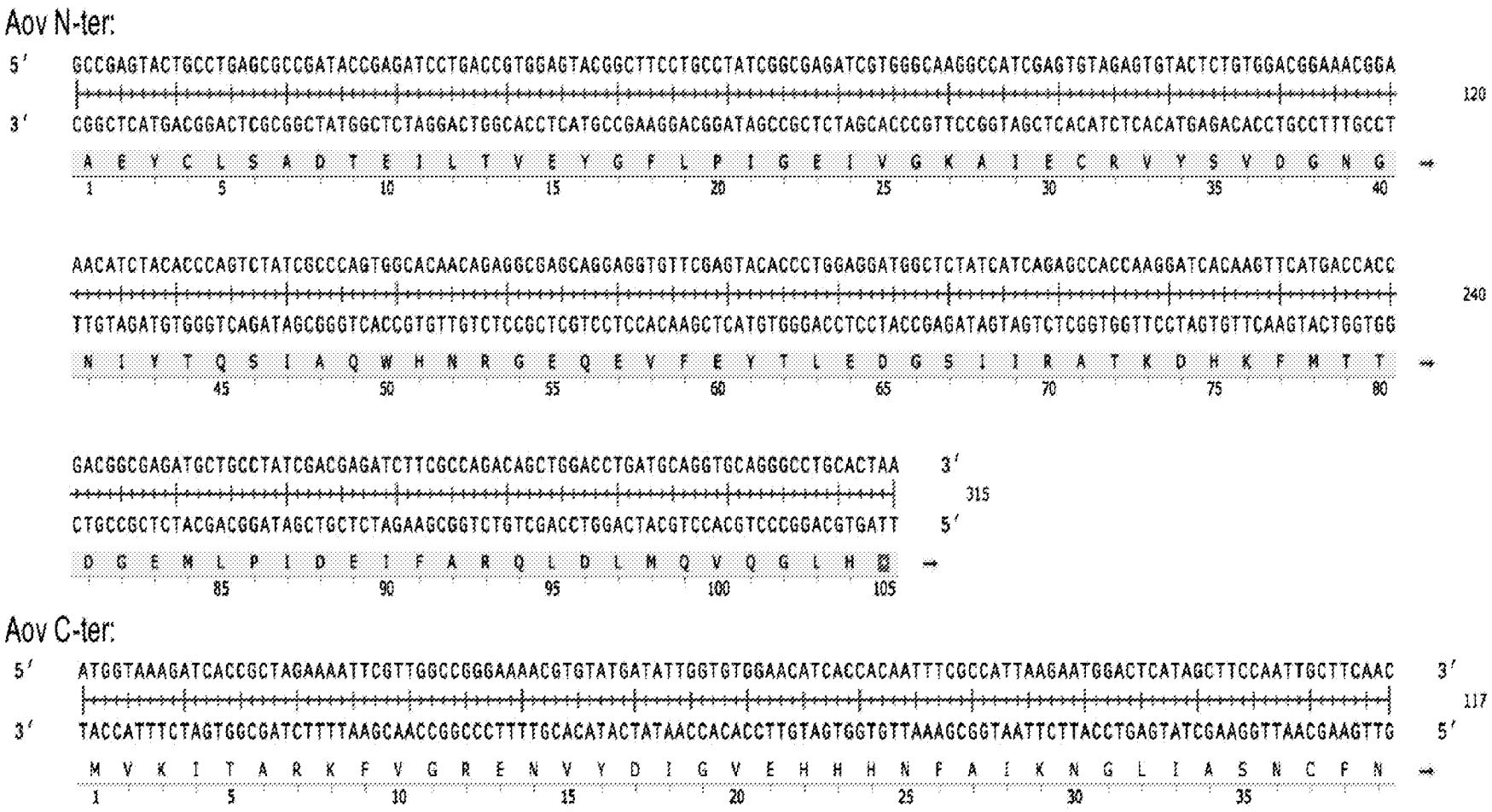

Aov N-ter:
5′ GCCGAGTACTGCCTGAGCGCCGATACCGAGATCCTGACCGTGGAGTACGGCTTCCTGCCTATCGGCGAGATCGTGGGCAAGGCCATCGAGTGTAGAGTGTACTCTGTGGACGGAAACGGA
120
3′ CGGCTCATGACGGACTCGCGGCTATGGCTCTAGGACTGGCACCTCATGCCGAAGGACGGATAGCCGCTCTAGCACCCGTTCCGGTAGCTCACATCTCACATGAGACACCTGCCTTTGCCT
A E Y C L S A D T E I L T V E Y G F L P I G E I V G K A I E C R V Y S V D G N G
1 5 10 15 20 25 30 35 40
AACATCTACACCCAGTCTATCGCCCAGTGGCACAACAGAGGCGAGCAGGAGGTGTTCGAGTACACCCTGGAGGATGGCTCTATCATCAGAGCCACCAAGGATCACAAGTTCATGACCACC
240
TTGTAGATGTGGGTCAGATAGCGGGTCACCGTGTTGTCTCCGCTCGTCCTCCACAAGCTCATGTGGGACCTCCTACCGAGATAGTAGTCTCGGTGGTTCCTAGTGTTCAAGTACTGGTGG
N I Y T Q S I A Q W H N R G E Q E V F E Y T L E D G S I I R A T K D H K F M T T
45 50 55 60 65 70 75 80
GACGGCGAGATGCTGCCTATCGACGAGATCTTCGCCAGACAGCTGGACCTGATGCAGGTGCAGGGCCTGCACTAA 3′
315
CTGCCGCTCTACGACGGATAGCTGCTCTAGAAGCGGTCTGTCGACCTGGACTACGTCCACGTCCCGGACGTGATT 5′
D G E M L P I D E I F A R Q L D L M Q V Q G L H
85 90 95 100 105
Aov C-ter:
5′ ATGGTAAAGATCACCGCTAGAAAATTCGTTGGCCGGGAAAACGTGTATGATATTGGTGTGGAACATCACCACAATTTCGCCATTAAGAATGGACTCATAGCTTCCAATTGCTTCAAC 3′
117
3′ TACCATTTCTAGTGGCGATCTTTTAAGCAACCGGCCCTTTTGCACATACTATAACCACACCTTGTAGTGGTGTTAAAGCGGTAATTCTTACCTGAGTATCGAAGGTTAACGAAGTTG 5′
M V K I T A R K F V G R E N V Y D I G V E H H H N F A I K N G L I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15B

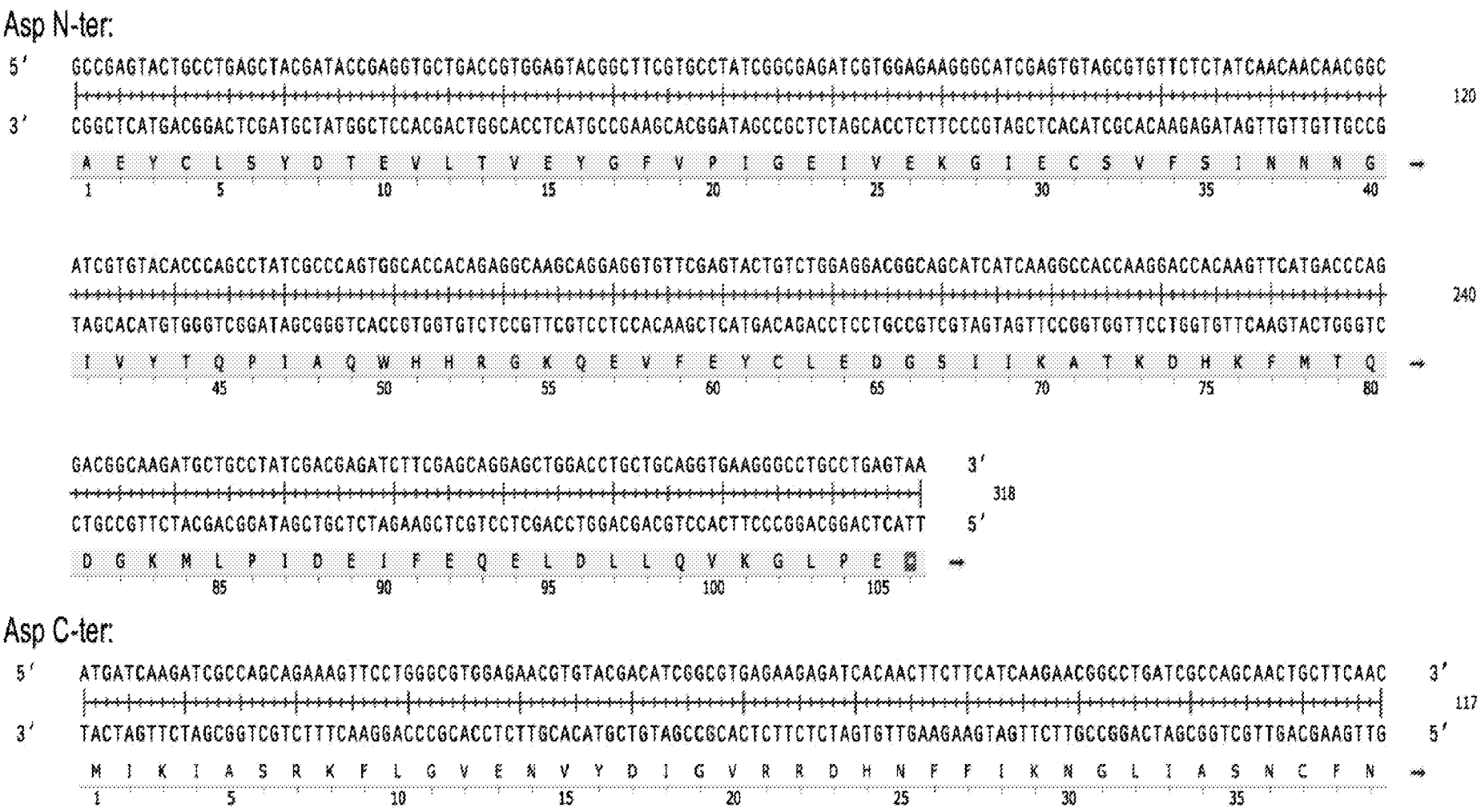

Asp N-ter:
5′ GCCGAGTACTGCCTGAGCTACGATACCGAGGTGCTGACCGTGGAGTACGGCTTCGTGCCTATCGGCGAGATCGTGGAGAAGGGCATCGAGTGTAGCGTGTTCTCTATCAACAACAACGGC
120
3′ CGGCTCATGACGGACTCGATGCTATGGCTCCACGACTGGCACCTCATGCCGAAGCACGGATAGCCGCTCTAGCACCTCTTCCCGTAGCTCACATCGCACAAGAGATAGTTGTTGTTGCCG
A E Y C L S Y D T E V L T V E Y G F V P I G E I V E K G I E C S V F S I N N N G
1 5 10 15 20 25 30 35 40
ATCGTGTACACCCAGCCTATCGCCCAGTGGCACCACAGAGGCAAGCAGGAGGTGTTCGAGTACTGTCTGGAGGACGGCAGCATCATCAAGGCCACCAAGGACCACAAGTTCATGACCCAG
240
TAGCACATGTGGGTCGGATAGCGGGTCACCGTGGTGTCTCCGTTCGTCCTCCACAAGCTCATGACAGACCTCCTGCCGTCGTAGTAGTTCCGGTGGTTCCTGGTGTTCAAGTACTGGGTC
I V Y T Q P I A Q W H H R G K Q E V F E Y C L E D G S I I K A T K D H K F M T Q
45 50 55 60 65 70 75 80
GACGGCAAGATGCTGCCTATCGACGAGATCTTCGAGCAGGAGCTGGACCTGCTGCAGGTGAAGGGCCTGCCTGAGTAA 3′
318
CTGCCGTTCTACGACGGATAGCTGCTCTAGAAGCTCGTCCTCGACCTGGACGACGTCCACTTCCCGGACGGACTCATT 5′
D G K M L P I D E I F E Q E L D L L Q V K G L P E
85 90 95 100 105
Asp C-ter:
5′ ATGATCAAGATCGCCAGCAGAAAGTTCCTGGGCGTGGAGAACGTGTACGACATCGGCGTGAGAAGAGATCACAACTTCTTCATCAAGAACGGCCTGATCGCCAGCAACTGCTTCAAC 3′
117
3′ TACTAGTTCTAGCGGTCGTCTTTCAAGGACCCGCACCTCTTGCACATGCTGTAGCCGCACTCTTCTCTAGTGTTGAAGAAGTAGTTCTTGCCGGACTAGCGGTCGTTGACGAAGTTG 5′
M I K I A S R K F L G V E N V Y D I G V R R D H N F F I K N G L I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15C

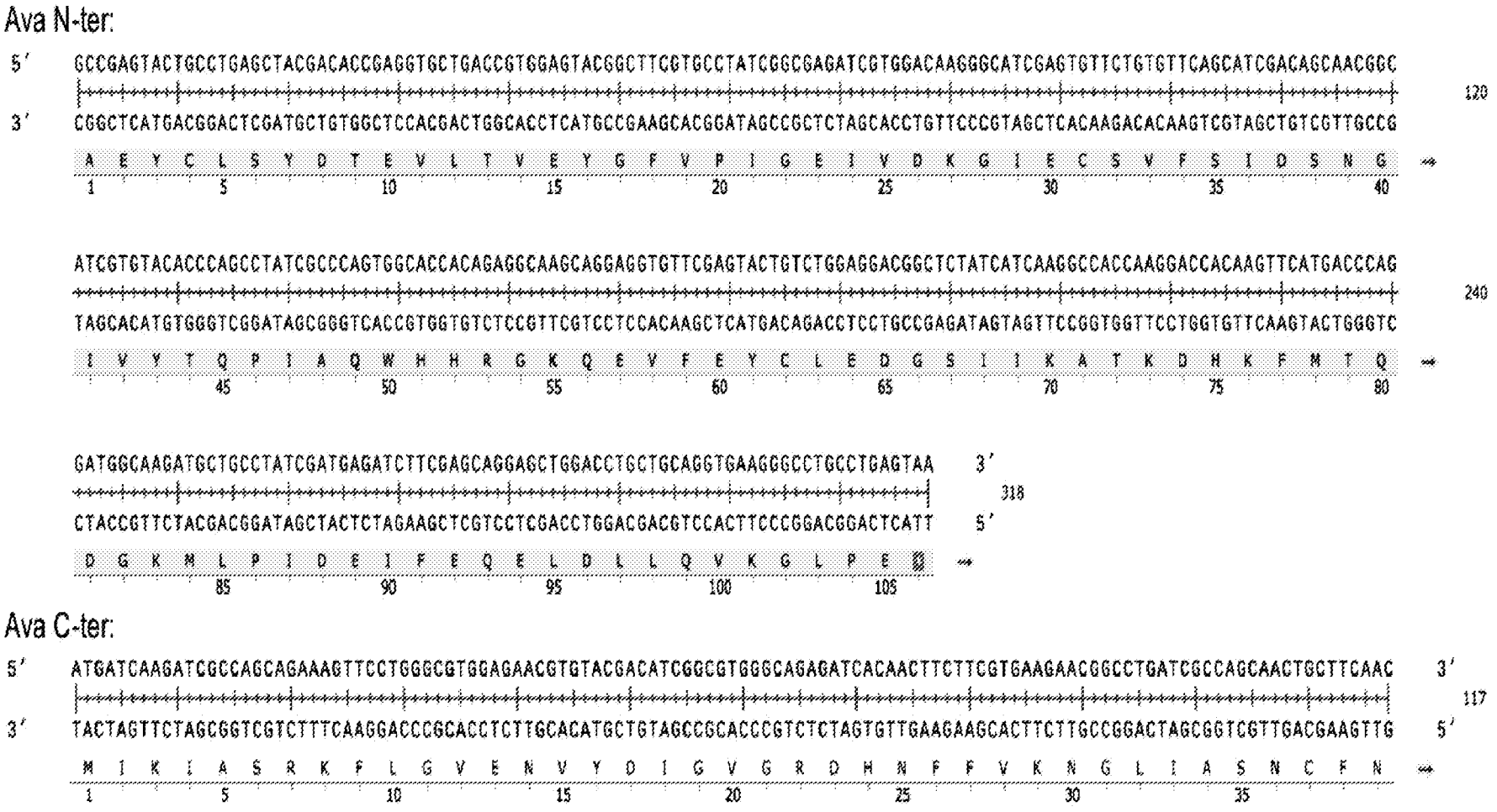

Ava N-ter:
5′ GCCGAGTACTGCCTGAGCTACGACACCGAGGTGCTGACCGTGGAGTACGGCTTCGTGCCTATCGGCGAGATCGTGGACAAGGGCATCGAGTGTTCTGTGTTCAGCATCGACAGCAACGGC
120
3′ CGGCTCATGACGGACTCGATGCTGTGGCTCCACGACTGGCACCTCATGCCGAAGCACGGATAGCCGCTCTAGCACCTGTTCCCGTAGCTCACAAGACACAAGTCGTAGCTGTCGTTGCCG
A E Y C L S Y D T E V L T V E Y G F V P I G E I V D K G I E C S V F S I D S N G
1 5 10 15 20 25 30 35 40
ATCGTGTACACCCAGCCTATCGCCCAGTGGCACCACAGAGGCAAGCAGGAGGTGTTCGAGTACTGTCTGGAGGACGGCTCTATCATCAAGGCCACCAAGGACCACAAGTTCATGACCCAG
240
TAGCACATGTGGGTCGGATAGCGGGTCACCGTGGTGTCTCCGTTCGTCCTCCACAAGCTCATGACAGACCTCCTGCCGAGATAGTAGTTCCGGTGGTTCCTGGTGTTCAAGTACTGGGTC
I V Y T Q P I A Q W H H R G K Q E V F E Y C L E D G S I I K A T K D H K F M T Q
45 50 55 60 65 70 75 80
GATGGCAAGATGCTGCCTATCGATGAGATCTTCGAGCAGGAGCTGGACCTGCTGCAGGTGAAGGGCCTGCCTGAGTAA 3′
318
CTACCGTTCTACGACGGATAGCTACTCTAGAAGCTCGTCCTCGACCTGGACGACGTCCACTTCCCGGACGGACTCATT 5′
D G K M L P I D E I F E Q E L D L L Q V K G L P E
85 90 95 100 105
Ava C-ter:
5′ ATGATCAAGATCGCCAGCAGAAAGTTCCTGGGCGTGGAGAACGTGTACGACATCGGCGTGGGCAGAGATCACAACTTCTTCGTGAAGAACGGCCTGATCGCCAGCAACTGCTTCAAC 3′
117
3′ TACTAGTTCTAGCGGTCGTCTTTCAAGGACCCGCACCTCTTGCACATGCTGTAGCCGCACCCGTCTCTAGTGTTGAAGAAGCACTTCTTGCCGGACTAGCGGTCGTTGACGAAGTTG 5′
M I K I A S R K F L G V E N V Y D I G V G R D H N F F V K N G L I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15D

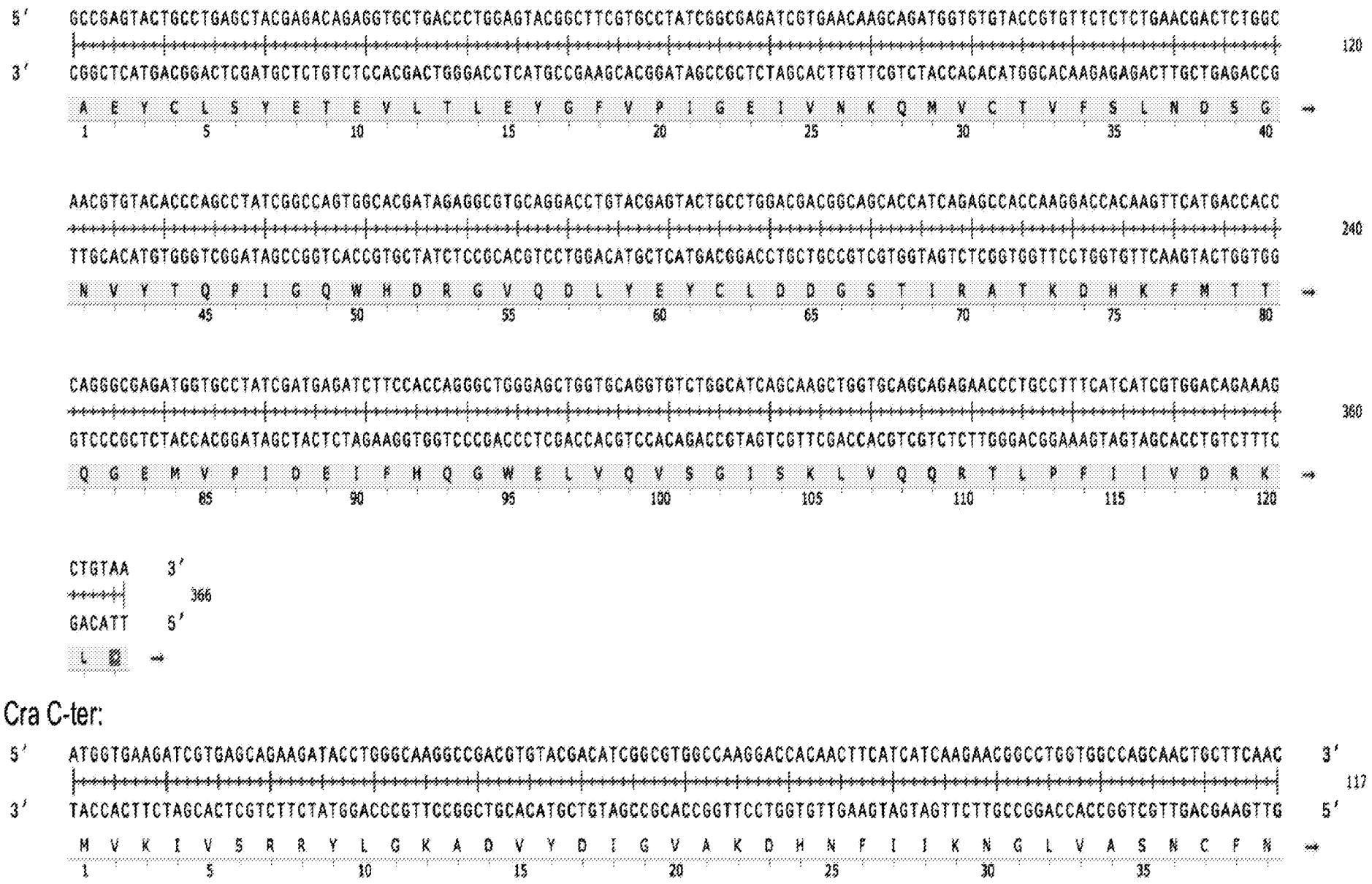
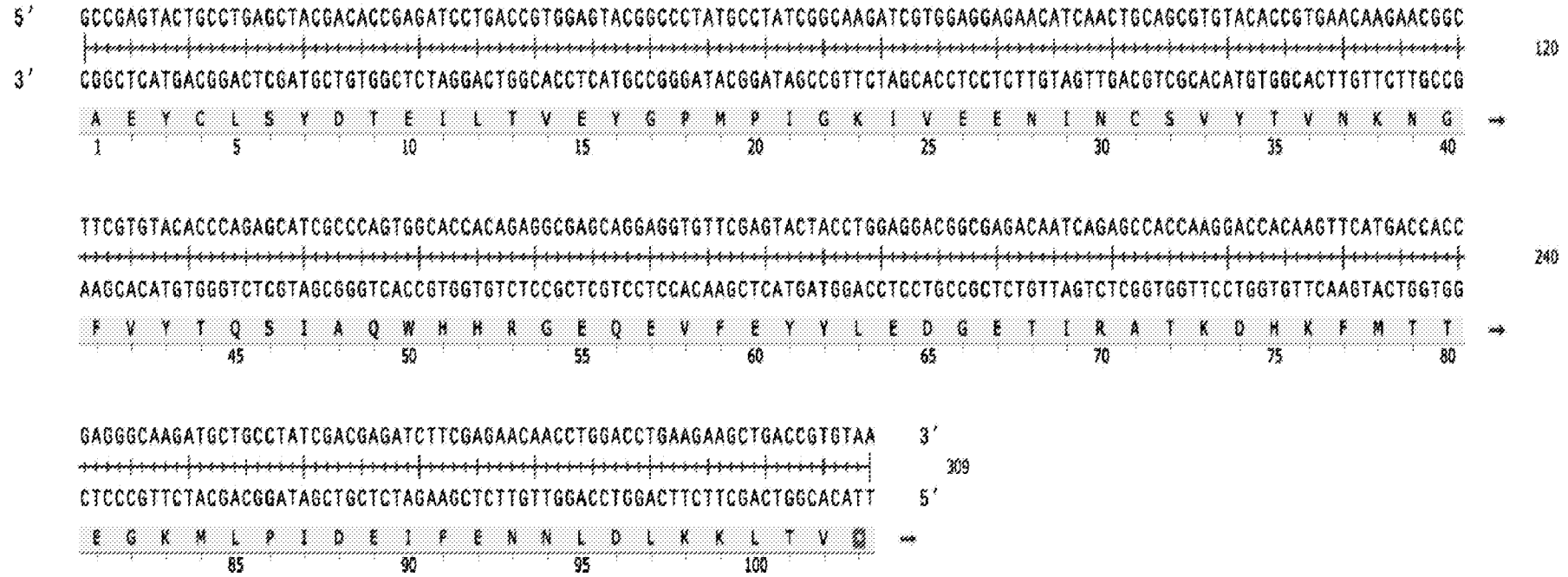

Cra N-ter:
5′ GCCGAGTACTGCCTGAGCTACGAGACAGAGGTGCTGACCCTGGAGTACGGCTTCGTGCCTATCGGCGAGATCGTGAACAAGCAGATGGTGTGTACCGTGTTCTCTCTGAACGACTCTGGC
3′ CGGCTCATGACGGACTCGATGCTCTGTCTCCACGACTGGGACCTCATGCCGAAGCACGGATAGCCGCTCTAGCACTTGTTCGTCTACCACACATGGCACAAGAGAGACTTGCTGAGACCG
A E Y C L S Y E T E V L T L E Y G F V P I G E I V N K Q M V C T V F S L N D S G
1 5 10 15 20 25 30 35 40
AACGTGTACACCCAGCCTATCGGCCAGTGGCACGATAGAGGCGTGCAGGACCTGTACGAGTACTGCCTGGACGACGGCAGCACCATCAGAGCCACCAAGGACCACAAGTTCATGACCACC
TTGCACATGTGGGTCGGATAGCCGGTCACCGTGCTATCTCCGCACGTCCTGGACATGCTCATGACGGACCTGCTGCCGTCGTGGTAGTCTCGGTGGTTCCTGGTGTTCAAGTACTGGTGG
N V Y T Q P I G Q W H D R G V Q D L Y E Y C L D D G S T I R A T K D H K F M T T
45 50 55 60 65 70 75 80
CAGGGCGAGATGGTGCCTATCGATGAGATCTTCCACCAGGGCTGGGAGCTGGTGCAGGTGTCTGGCATCAGCAAGCTGGTGCAGCAGAGAACCCTGCCTTTCATCATCGTGGACAGAAAG
GTCCCGCTCTACCACGGATAGCTACTCTAGAAGGTGGTCCCGACCCTCGACCACGTCCACAGACCGTAGTCGTTCGACCACGTCGTCTCTTGGGACGGAAAGTAGTAGCACCTGTCTTTC
Q G E M V P I D E I F H Q G W E L V Q V S G I S K L V Q Q R T L P F I I V D R K
85 90 95 100 105 110 115 120
CTGTAA 3′
366
GACATT 5′
L
Cra C-ter:
5′ ATGGTGAAGATCGTGAGCAGAAGATACCTGGGCAAGGCCGACGTGTACGACATCGGCGTGGCCAAGGACCACAACTTCATCATCAAGAACGGCCTGGTGGCCAGCAACTGCTTCAAC 3′
3′ TACCACTTCTAGCACTCGTCTTCTATGGACCCGTTCCGGCTGCACATGCTGTAGCCGCACCGGTTCCTGGTGTTGAAGTAGTAGTTCTTGCCGGACCACCGGTCGTTGACGAAGTTG 5′
M V K I V S R R Y L G K A D V Y D I G V A K D H N F I I K N G L V A S N C F N
1 5 10 15 20 25 30 35

FIG. 15E

Csp-CCY N-ter:



Csp-CCY C-ter:

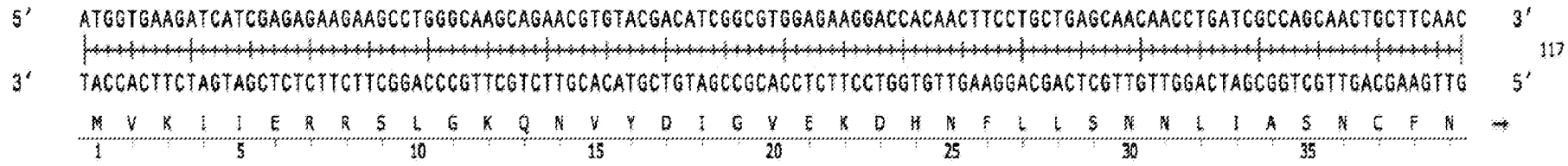


FIG. 15F

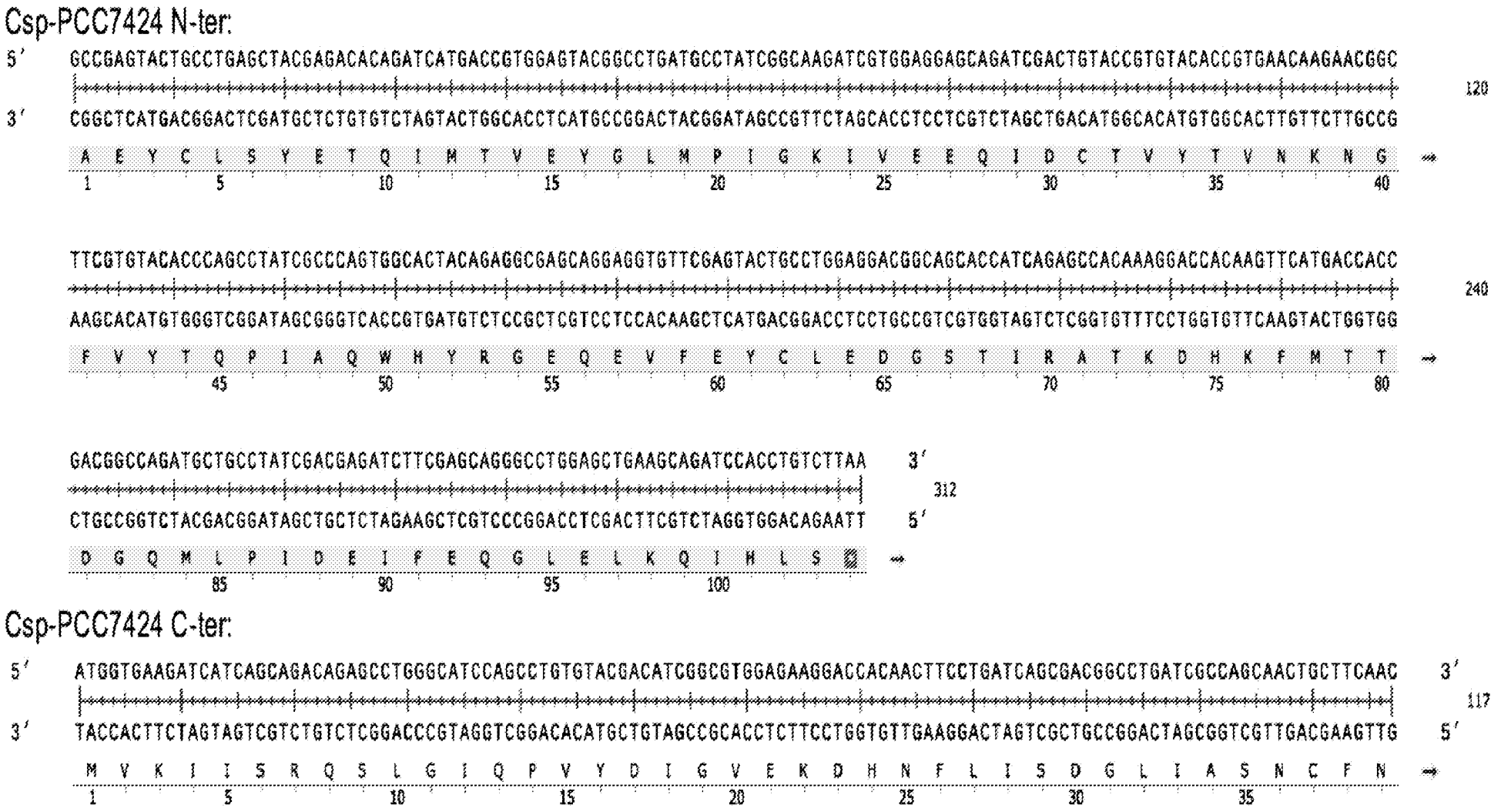

Csp-PCC7424 N-ter:
5′ GCCGAGTACTGCCTGAGCTACGAGACACAGATCATGACCGTGGAGTACGGCCTGATGCCTATCGGCAAGATCGTGGAGGAGCAGATCGACTGTACCGTGTACACCGTGAACAAGAACGGC
120
3′ CGGCTCATGACGGACTCGATGCTCTGTGTCTAGTACTGGCACCTCATGCCGGACTACGGATAGCCGTTCTAGCACCTCCTCGTCTAGCTGACATGGCACATGTGGCACTTGTTCTTGCCG
A E Y C L S Y E T Q I M T V E Y G L M P I G K I V E E Q I D C T V Y T V N K N G
1 5 10 15 20 25 30 35 40
TTCGTGTACACCCAGCCTATCGCCCAGTGGCACTACAGAGGCGAGCAGGAGGTGTTCGAGTACTGCCTGGAGGACGGCAGCACCATCAGAGCCACAAAGGACCACAAGTTCATGACCACC
240
AAGCACATGTGGGTCGGATAGCGGGTCACCGTGATGTCTCCGCTCGTCCTCCACAAGCTCATGACGGACCTCCTGCCGTCGTGGTAGTCTCGGTGTTTCCTGGTGTTCAAGTACTGGTGG
F V Y T Q P I A Q W H Y R G E Q E V F E Y C L E D G S T I R A T K D H K F M T T
45 50 55 60 65 70 75 80
GACGGCCAGATGCTGCCTATCGACGAGATCTTCGAGCAGGGCCTGGAGCTGAAGCAGATCCACCTGTCTTAA 3′
312
CTGCCGGTCTACGACGGATAGCTGCTCTAGAAGCTCGTCCCGGACCTCGACTTCGTCTAGGTGGACAGAATT 5′
D G Q M L P I D E I F E Q G L E L K Q I H L S
85 90 95 100
Csp-PCC7424 C-ter:
5′ ATGGTGAAGATCATCAGCAGACAGAGCCTGGGCATCCAGCCTGTGTACGACATCGGCGTGGAGAAGGACCACAACTTCCTGATCAGCGACGGCCTGATCGCCAGCAACTGCTTCAAC 3′
117
3′ TACCACTTCTAGTAGTCGTCTGTCTCGGACCCGTAGGTCGGACACATGCTGTAGCCGCACCTCTTCCTGGTGTTGAAGGACTAGTCGCTGCCGGACTAGCGGTCGTTGACGAAGTTG 5′
M V K I I S R Q S L G I Q P V Y D I G V E K D H N F L I S D G L I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15G

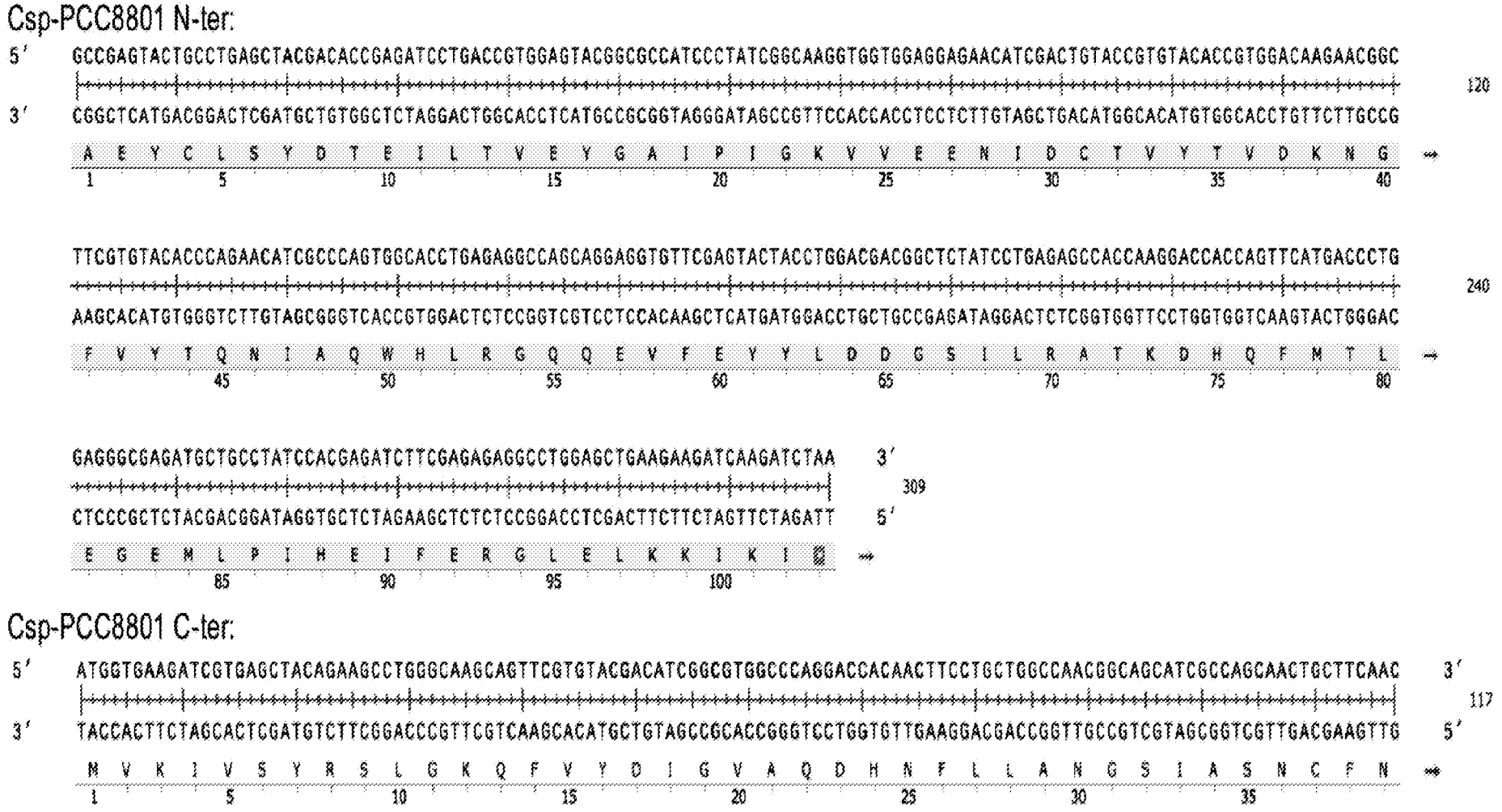
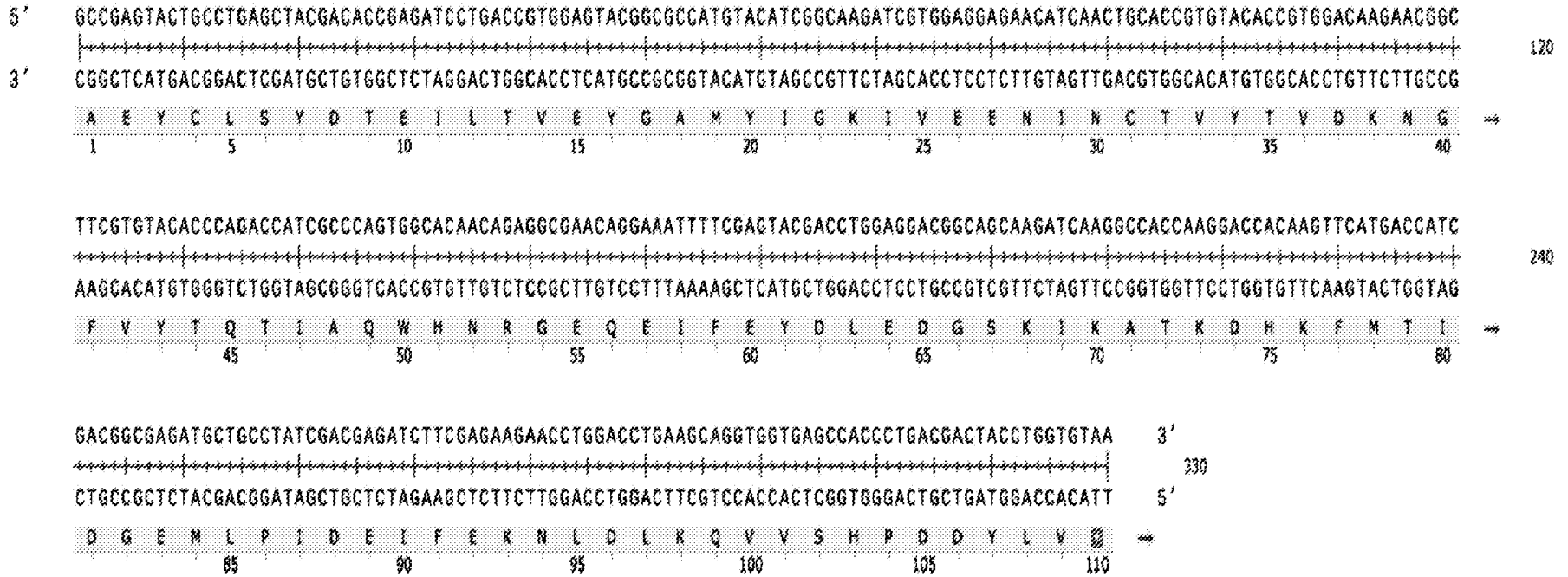

Csp-PCC8801 N-ter:
5′ GCCGAGTACTGCCTGAGCTACGACACCGAGATCCTGACCGTGGAGTACGGCGCCATCCCTATCGGCAAGGTGGTGGAGGAGAACATCGACTGTACCGTGTACACCGTGGACAAGAACGGC
120
3′ CGGCTCATGACGGACTCGATGCTGTGGCTCTAGGACTGGCACCTCATGCCGCGGTAGGGATAGCCGTTCCACCACCTCCTCTTGTAGCTGACATGGCACATGTGGCACCTGTTCTTGCCG
A E Y C L S Y D T E I L T V E Y G A I P I G K V V E E N I D C T V Y T V D K N G
1 5 10 15 20 25 30 35 40
TTCGTGTACACCCAGAACATCGCCCAGTGGCACCTGAGAGGCCAGCAGGAGGTGTTCGAGTACTACCTGGACGACGGCTCTATCCTGAGAGCCACCAAGGACCACCAGTTCATGACCCTG
240
AAGCACATGTGGGTCTTGTAGCGGGTCACCGTGGACTCTCCGGTCGTCCTCCACAAGCTCATGATGGACCTGCTGCCGAGATAGGACTCTCGGTGGTTCCTGGTGGTCAAGTACTGGGAC
F V Y Y T Q N I A Q W H L R G Q Q E V F E Y Y L D D G S I L R A T K D H Q F M T L
45 50 55 60 65 70 75 80
GAGGGCGAGATGCTGCCTATCCACGAGATCTTCGAGAGAGGCCTGGAGCTGAAGAAGATCAAGATCTAA 3′
309
CTCCCGCTCTACGACGGATAGGTGCTCTAGAAGCTCTCTCCGGACCTCGACTTCTTCTAGTTCTAGATT 5′
E G E M L P I H E I F E R G L E L K K I K I ▨
85 90 95 100
Csp-PCC8801 C-ter:
5′ ATGGTGAAGATCGTGAGCTACAGAAGCCTGGGCAAGCAGTTCGTGTACGACATCGGCGTGGCCCAGGACCACAACTTCCTGCTGGCCAACGGCAGCATCGCCAGCAACTGCTTCAAC 3′
117
3′ TACCACTTCTAGCACTCGATGTCTTCGGACCCGTTCGTCAAGCACATGCTGTAGCCGCACCGGGTCCTGGTGTTGAAGGACGACCGGTTGCCGTCGTAGCGGTCGTTGACGAAGTTG 5′
M V K I V S Y R S L G K Q F V Y D I G V A Q D H N F L L A N G S I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15H

Cwa N-ter:



Cwa C-ter:

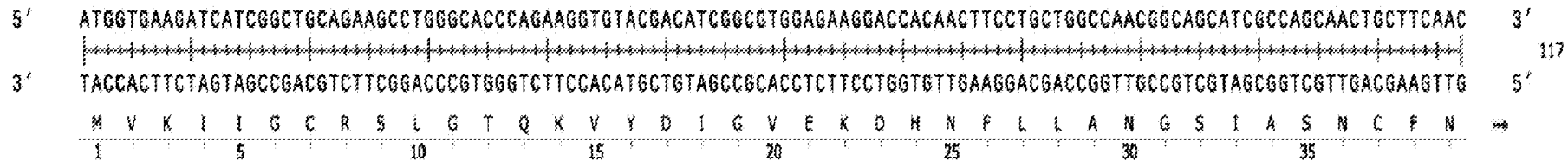


FIG. 15I

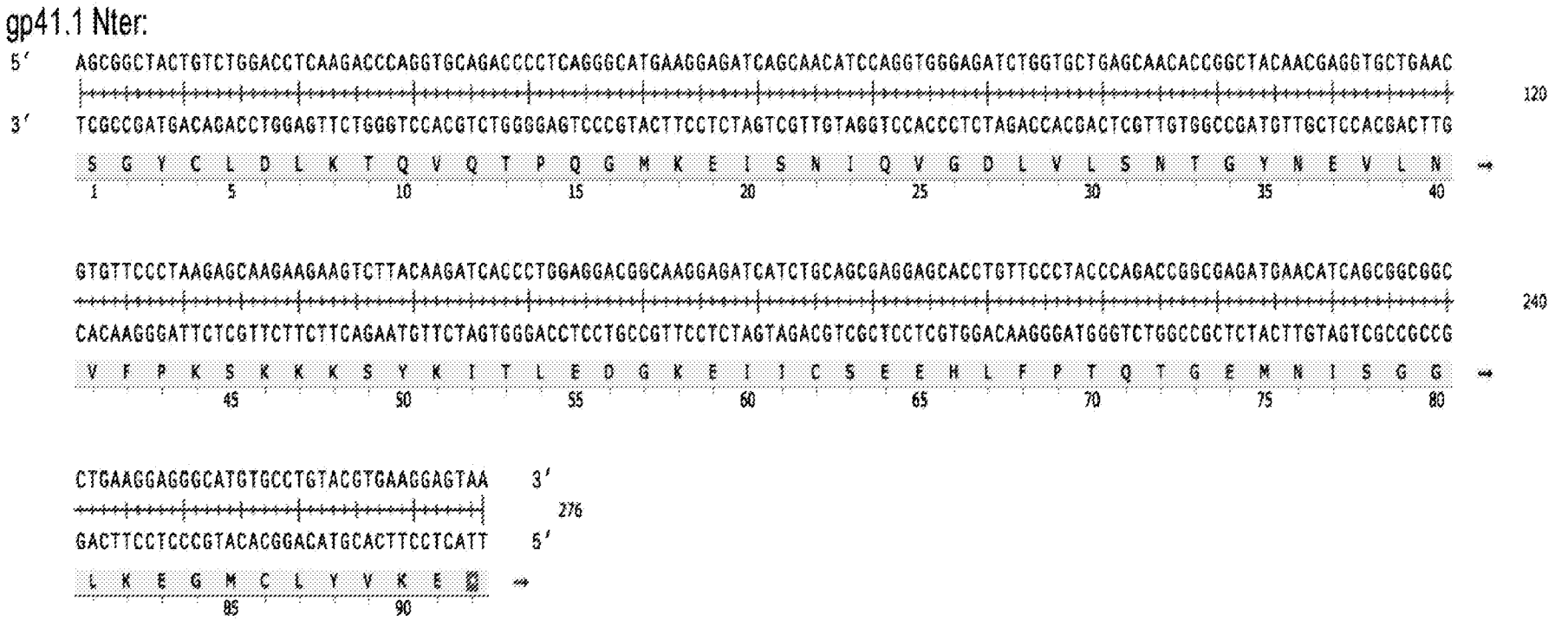

gp41.1 Nter:
5′ AGCGGCTACTGTCTGGACCTCAAGACCCAGGTGCAGACCCCTCAGGGCATGAAGGAGATCAGCAACATCCAGGTGGGAGATCTGGTGCTGAGCAACACCGGCTACAACGAGGTGCTGAAC
120
3′ TCGCCGATGACAGACCTGGAGTTCTGGGTCCACGTCTGGGGAGTCCCGTACTTCGTCTAGTCGTTGTAGGTCCACCGTCTAGACCACGACTCGTTGTGGCCGATGTTGCTCCACGACTTG
S G Y C L D L K T Q V Q T P Q G M K E I S N I Q V G D L V L S N T G Y N E V L N
1 5 10 15 20 25 30 35 40
GTGTTCCCTAAGAGCAAGAAGAAGTCTTACAAGATCACCCTGGAGGACGGCAAGGAGATCATCTGCAGCGAGGAGCACCTGTTCCCTACCCAGACCGGCGAGATGAACATCAGCGGCGGC
240
CACAAGGGATTCTCGTTCTTCTTCAGAATGTTCTAGTGGGACCTCCTGCCGTTCCTCTAGTAGACGTCGCTCCTCGTGGACAAGGGATGGGTCTGGCCGCTCTACTTGTAGTCGCCGCCG
V F P K S K K K S Y K I T L E D G K E I I C S E E H L F P T Q T G E M N I S G G
45 50 55 60 65 70 75 80
CTGAAGGAGGGCATGTGCCTGTACGTGAAGGAGTAA 3′
276
GACTTCCTCCCGTACACGGACATGCACTTCCTCATT 5′
L K E G M C L Y V K E
85 90
gp41.1 Cter:
5′ ATGATGCTGAAGAAGATCCTGAAGATCGAGGAGCTGGACGAGAGAGAGCTGATCGACATCGAGGTGAGCGGCAACCACCTGTTCTACGCCAACGACATCCTGACCCACAACAGCAGCAGC 3′
120
3′ TACTACGACTTCTTCTAGGACTTCTAGCTCCTCGACCTGCTCTCTCTCGACTAGCTGTAGCTCCACTCGCCGTTGGTGGACAAGATGCGGTTGCTGTAGGACTGGGTGTTGTCGTCGTCG 5′
M M L K K I L K I E E L D E R E L I D I E V S G N H L F Y A N D I L T H N S S S
1 5 10 15 20 25 30 35 40

FIG. 15J

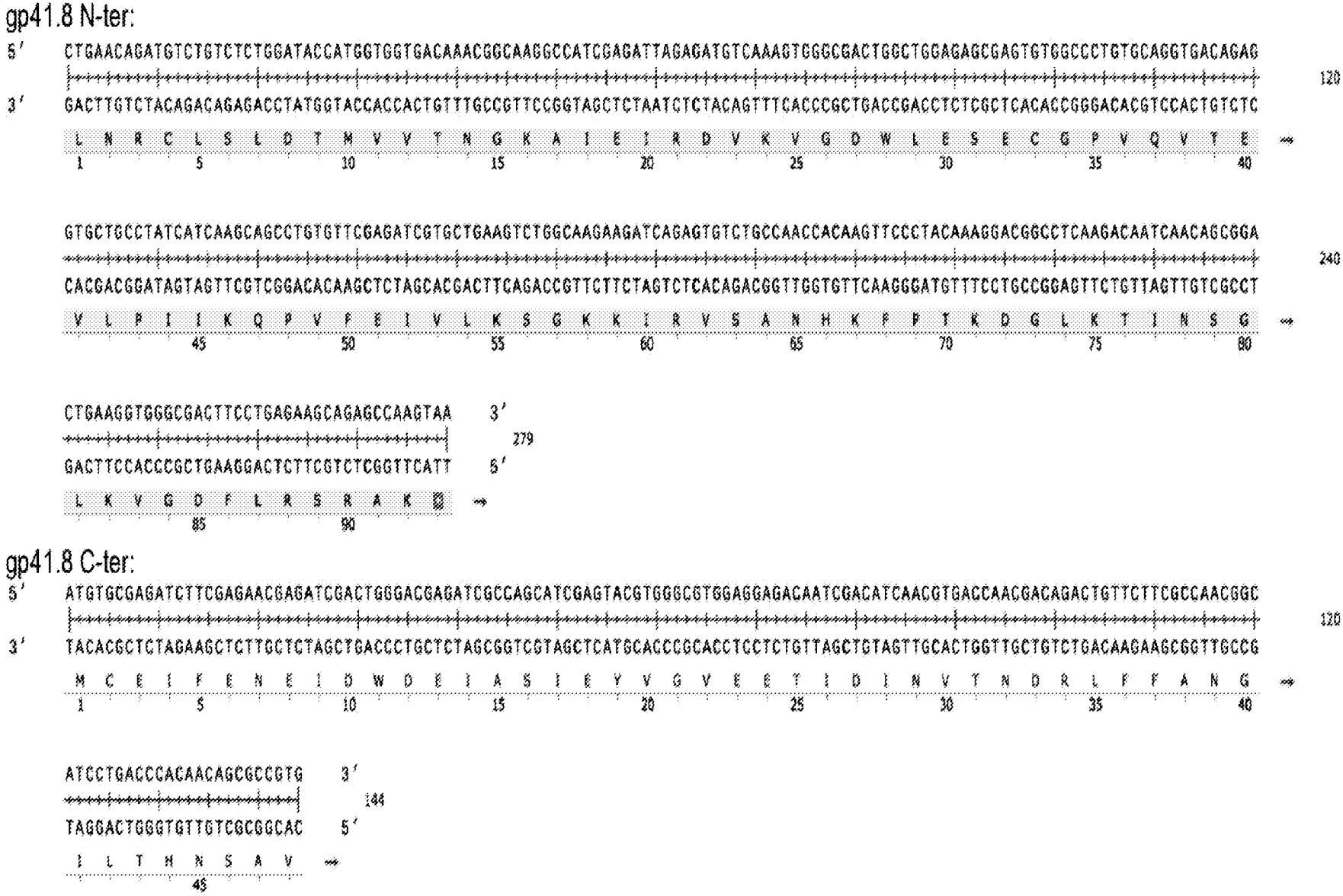

gp41.8 N-ter:
5′ CTGAACAGATGTCTGTCTCTGGATACCATGGTGGTGACAAACGGCAAGGCCATCGAGATTAGAGATGTCAAAGTGGGCGACTGGCTGGAGAGCGAGTGTGGCCCTGTGCAGGTGACAGAG
120
3′ GACTTGTCTACAGACAGAGACCTATGGTACCACCACTGTTTGCCGTTCCGGTAGCTCTAATCTCTACAGTTTCACCCGCTGACCGACCTCTCGCTCACACCGGGACACGTCCACTGTCTC
L N R C L S L D T M V V T N G K A I E I R D V K V G D W L E S E C G P V Q V T E
1 5 10 15 20 25 30 35 40
GTGCTGCCTATCATCAAGCAGCCTGTGTTCGAGATCGTGCTGAAGTCTGGCAAGAAGATCAGAGTGTCTGCCAACCACAAGTTCCCTACAAAGGACGGCCTCAAGACAATCAACAGCGGA
240
CACGACGGATAGTAGTTCGTCGGACACAAGCTCTAGCACGACTTCAGACCGTTCTTCTAGTCTCACAGACGGTTGGTGTTCAAGGGATGTTTCCTGCCGGAGTTCTGTTAGTTGTCGCCT
V L P I I K Q P V F E I V L K S G K K I R V S A N H K F P T K D G L K T I N S G
45 50 55 60 65 70 75 80
CTGAAGGTGGGCGACTTCCTGAGAAGCAGAGCCAAGTAA 3′
279
GACTTCCACCCGCTGAAGGACTCTTCGTCTCGGTTCATT 5′
L K V G D F L R S R A K
85 90
gp41.8 C-ter:
5′ ATGTGCGAGATCTTCGAGAACGAGATCGACTGGGACGAGATCGCCAGCATCGAGTACGTGGGCGTGGAGGAGACAATCGACATCAACGTGACCAACGACAGACTGTTCTTCGCCAACGGC
120
3′ TACACGCTCTAGAAGCTCTTGCTCTAGCTGACCCTGCTCTAGCGGTCGTAGCTCATGCACCCGCACCTCCTCTGTTAGCTGTAGTTGCACTGGTTGCTGTCTGACAAGAAGCGGTTGCCG
M C E I F E N E I D W D E I A S I E Y V G V E E T I D I N V T N D R L F F A N G
1 5 10 15 20 25 30 35 40
ATCCTGACCCACAACAGCGCCGTG 3′
144
TAGGACTGGGTGTTGTCGCGGCAC 5′
I L T H N S A V
45

FIG. 15K

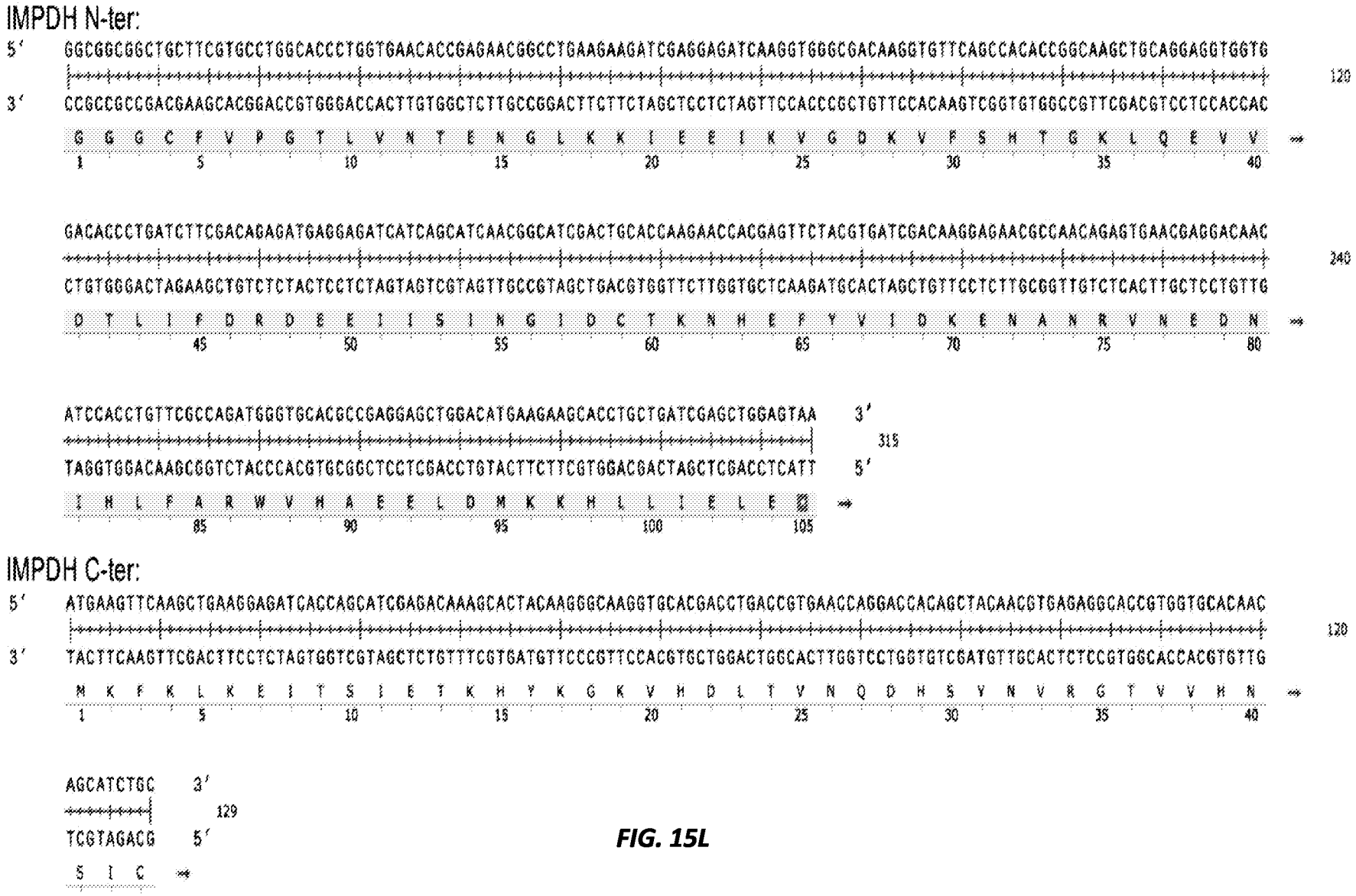

IMPDH N-ter:
5′ GGCGGCGGCTGCTTCGTGCCTGGCACCCTGGTGAACACCGAGAACGGCCTGAAGAAGATCGAGGAGATCAAGGTGGGCGACAAGGTGTTCAGCCACACCGGCAAGCTGCAGGAGGTGGTG
3′ CCGCCGCCGACGAAGCACGGACCGTGGGACCACTTGTGGCTCTTGCCGGACTTGTTCTAGCTCCTCTAGTTCCACCCGCTGTTCCACAAGTCGGTGTGGCCGTTCGACGTCCTCCACCAC
G G G C F V P G T L V N T E N G L K K I E E I K V G D K V F S H T G K L Q E V V
1 5 10 15 20 25 30 35 40
GACACCCTGATCTTCGACAGAGATGAGGAGATCATCAGCATCAACGGCATCGACTGCACCAAGAACCACGAGTTCTACGTGATCGACAAGGAGAACGCCAACAGAGTGAACGAGGACAAC
CTGTGGGACTAGAAGCTGTCTCTACTCCTCTAGTAGTCGTAGTTGCCGTAGCTGACGTGGTTCTTGGTGCTCAAGATGCACTAGCTGTTCCTCTTGCGGTTGTCTCACTTGCTCCTGTTG
D T L I F D R D E E I I S I N G I D C T K N H E F Y V I D K E N A N R V N E D N
45 50 55 60 65 70 75 80
ATCCACCTGTTCGCCAGATGGGTGCACGCCGAGGAGCTGGACATGAAGAAGCACCTGCTGATCGAGCTGGAGTAA 3′
315
TAGGTGGACAAGCGGTCTACCCACGTGCGGCTCCTCGACCTGTACTTCTTCGTGGACGACTAGCTCGACCTCATT 5′
I H L F A R W V H A E E L D M K K H L L I E L E
85 90 95 100 105
IMPDH C-ter:
5′ ATGAAGTTCAAGCTGAAGGAGATCACCAGCATCGAGACAAAGCACTACAAGGGCAAGGTGCACGACGTGACCGTGAACCAGGACCACAGCTACAACGTGAGAGGCACCGTGGTGCACAAC
3′ TACTTCAAGTTCGACTTCCTCTAGTGGTCGTAGCTCTGTTTCGTGATGTTCCCGTTCCACGTGCTGGACTGGCACTTGGTCCTGGTGTCGATGTTGCACTCTCCGTGGCACCACGTGTTG
M K F K L K E I T S I E T K H Y K G K V H D L T V N Q D H S Y N V R G T V V H N
1 5 10 15 20 25 30 35 40
AGCATCTGC 3′
129
TCGTAGACG 5′
S I C

FIG. 15L

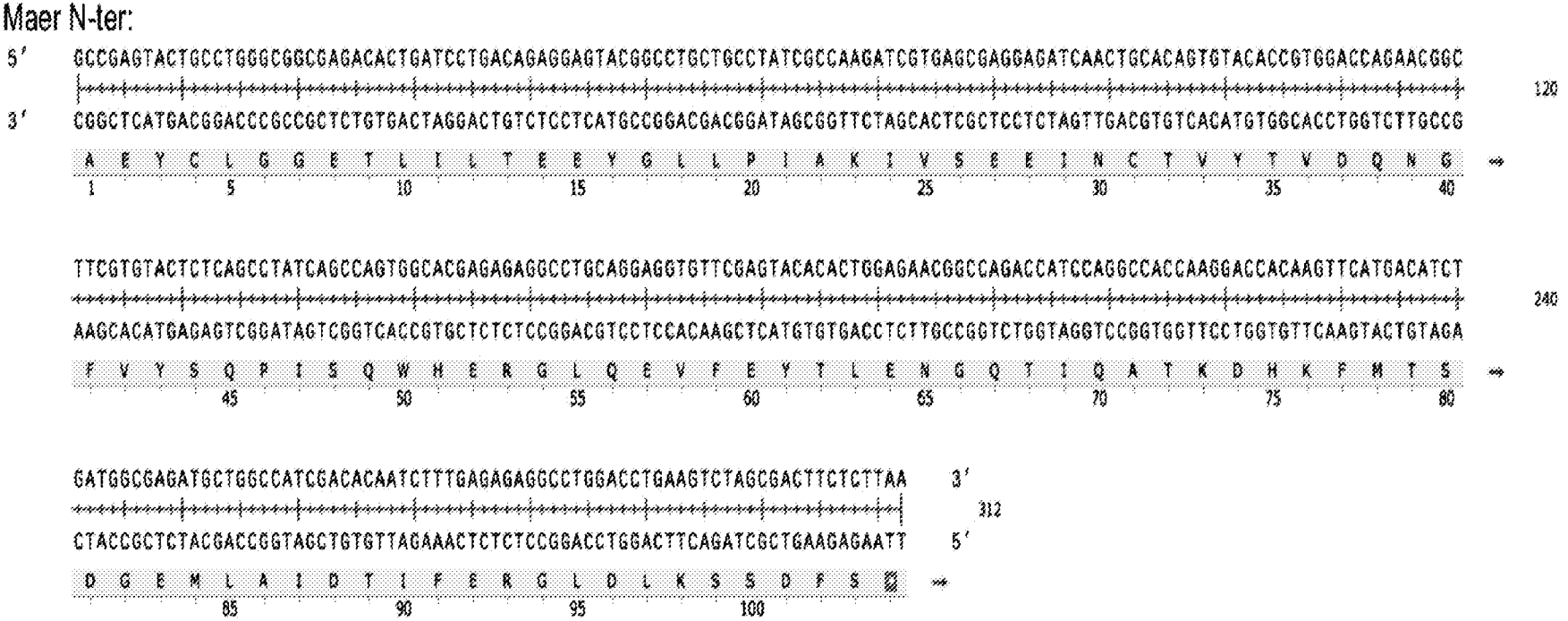

Maer N-ter:
5′ GCCGAGTACTGCCTGGGCGGCGAGACACTGATCCTGACAGAGGAGTACGGCCTGCTGCCTATCGCCAAGATCGTGAGCGAGGAGATCAACTGCACAGTGTACACCGTGGACCAGAACGGC 120
3′ CGGCTCATGACGGACCCGCCGCTCTGTGACTAGGACTGTCTCCTCATGCCGGACGACGGATAGCGGTTCTAGCACTCGCTCCTCTAGTTGACGTGTCACATGTGGCACCTGGTCTTGCCG
A E Y C L G G E T L I L T E E Y G L L P I A K I V S E E I N C T V Y T V D Q N G
1 5 10 15 20 25 30 35 40
TTCGTGTACTCTCAGCCTATCAGCCAGTGGCACGAGAGAGGCCTGCAGGAGGTGTTCGAGTACACACTGGAGAACGGCCAGACCATCCAGGCCACCAAGGACCACAAGTTCATGACATCT 240
AAGCACATGAGAGTCGGATAGTCGGTCACCGTGCTCTCTCCGGACGTCCTCCACAAGCTCATGTGTGACCTCTTGCCGGTCTGGTAGGTCCGGTGGTTCCTGGTGTTCAAGTACTGTAGA
F V Y S Q P I S Q W H E R G L Q E V F E Y T L E N G Q T I Q A T K D H K F M T S
45 50 55 60 65 70 75 80
GATGGCGAGATGCTGGCCATCGACACAATCTTTGAGAGAGGCCTGGACCTGAAGTCTAGCGACTTCTCTTAA 3′ 312
CTACCGCTCTACGACCGGTAGCTGTGTTAGAAACTCTCTCCGGACCTGGACTTCAGATCGCTGAAGAGAATT 5′
D G E M L A I D T I F E R G L D L K S S D F S
85 90 95 100
Maer C-ter:
5′ ATGGTGAAGATCATCGGCAGACAGAGCCTGGGCAGAAAGCCTGTGTACGACATCGGCGTGGAGAAGGACCACAACTTCCTGCTGGGCAACGGCCTGATCGCCAGCAACTGCTTCAAC 3′ 117
3′ TACCACTTCTAGTAGCCGTCTGTCTCGGACCCGTCTTTCGGACACATGCTGTAGCCGCACCTCTTCCTGGTGTTGAAGGACGACCCGTTGCCGGACTAGCGGTCGTTGACGAAGTTG 5′
M V K I I G R Q S L G R K P V Y D I G V E K D H N F L L G N G L I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15M

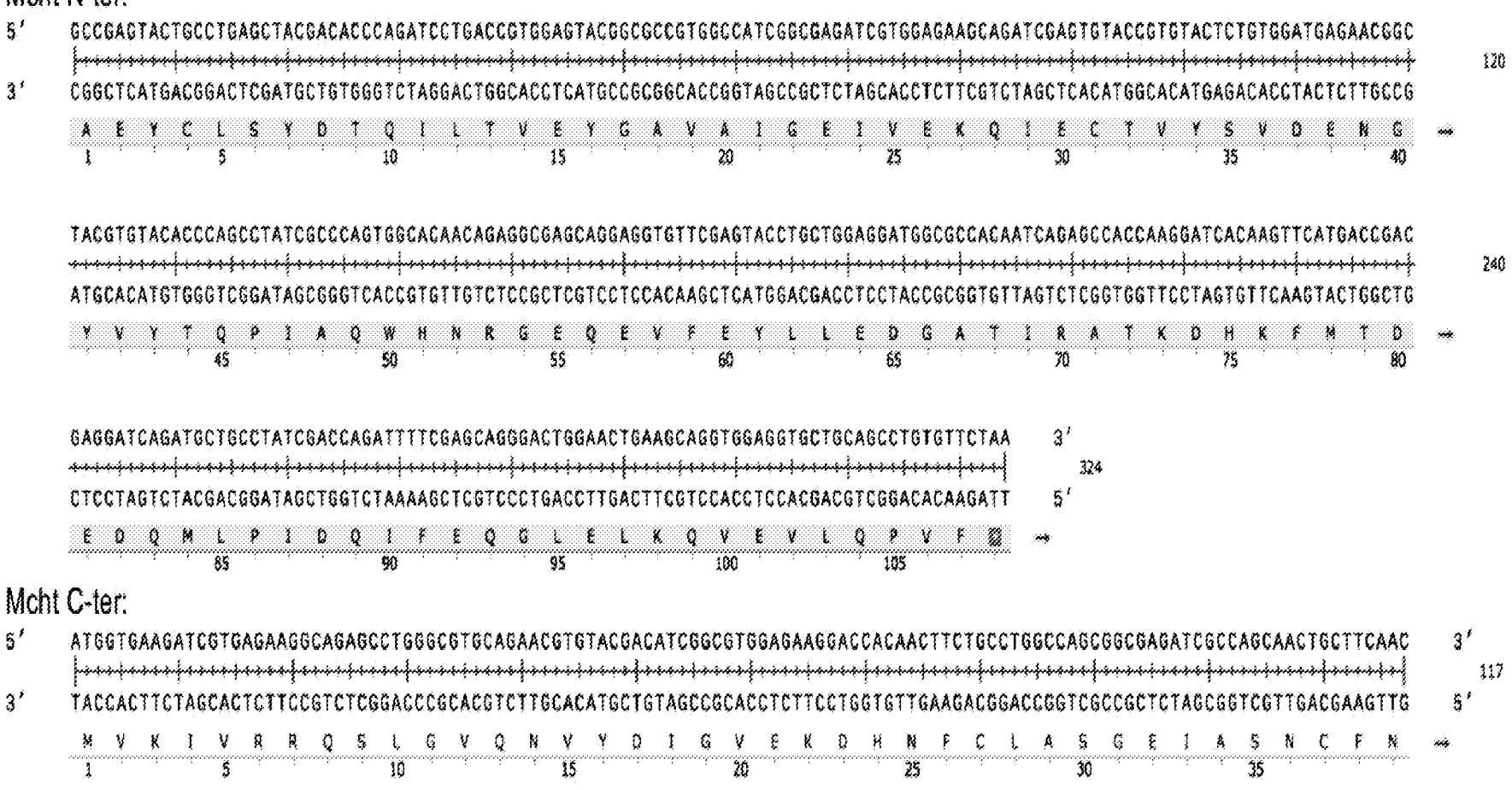

Mcht N-ter:
5′ GCCGAGTACTGCCTGAGCTACGACACCCAGATCCTGACCGTGGAGTACGGCGCCGTGGCCATCGGCGAGATCGTGGAGAAGCAGATCGAGTGTACCGTGTACTCTGTGGATGAGAACGGC
120
3′ CGGCTCATGACGGACTCGATGCTGTGGGTCTAGGACTGGCACCTCATGCCGCGGCACCGGTAGCCGCTCTAGCACCTCTTCGTCTAGCTCACATGGCACATGAGACACCTACTCTTGCCG
A E Y C L S Y D T Q I L T V E Y G A V A I G E I V E K Q I E C T V Y S V D E N G
1 5 10 15 20 25 30 35 40
TACGTGTACACCCAGCCTATCGCCCAGTGGCACAACAGAGGCGAGCAGGAGGTGTTCGAGTACCTGCTGGAGGATGGCGCCACAATCAGAGCCACCAAGGATCACAAGTTCATGACCGAC
240
ATGCACATGTGGGTCGGATAGCGGGTCACCGTGTTGTCTCCGCTCGTCCTCCACAAGCTCATGGACGACCTCCTACCGCGGTGTTAGTCTCGGTGGTTCCTAGTGTTCAAGTACTGGCTG
Y V Y T Q P I A Q W H N R G E Q E V F E Y L L E D G A T I R A T K D H K F M T D
45 50 55 60 65 70 75 80
GAGGATCAGATGCTGCCTATCGACCAGATTTTCGAGCAGGGACTGGAACTGAAGCAGGTGGAGGTGCTGCAGCCTGTGTTCTAA 3′
324
CTCCTAGTCTACGACGGATAGCTGGTCTAAAAGCTCGTCCCTGACCTTGACTTCGTCCACCTCCACGACGTCGGACACAAGATT 5′
E D Q M L P I D Q I F E Q G L E L K Q V E V L Q P V F
85 90 95 100 105
Mcht C-ter:
5′ ATGGTGAAGATCGTGAGAAGGCAGAGCCTGGGCGTGCAGAACGTGTACGACATCGGCGTGGAGAAGGACCACAACTTCTGCCTGGCCAGCGGCGAGATCGCCAGCAACTGCTTCAAC 3′
117
3′ TACCACTTCTAGCACTCTTCCGTCTCGGACCCGCACGTCTTGCACATGCTGTAGCCGCACCTCTTCCTGGTGTTGAAGACGGACCGGTCGCCGCTCTAGCGGTCGTTGACGAAGTTG 5′
M V K I V R R Q S L G V Q N V Y D I G V E K D H N F C L A S G E I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15N

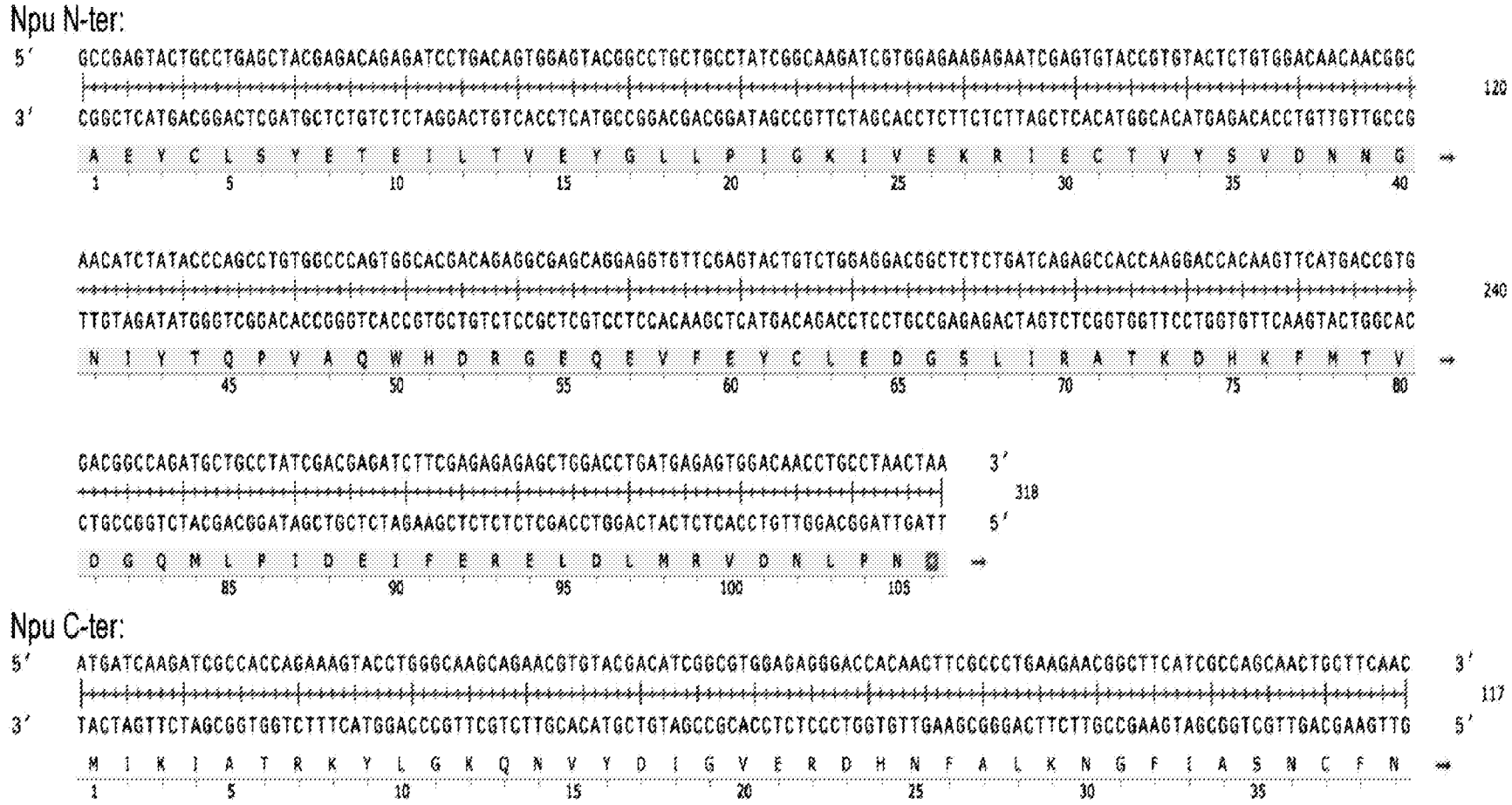

Npu N-ter:
5′ GCCGAGTACTGCCTGAGCTACGAGACAGAGATCCTGACAGTGGAGTACGGCCTGCTGCCTATCGGCAAGATCGTGGAGAAGAGAATCGAGTGTACCGTGTACTCTGTGGACAACAACGGC
3′ CGGCTCATGACGGACTCGATGCTCTGTCTCTAGGACTGTCACCTCATGCCGGACGACGGATAGCCGTTCTAGCACCTCTTCTCTTAGCTCACATGGCACATGAGACACCTGTTGTTGCCG
A E Y C L S Y E T E I L T V E Y G L L P I G K I V E K R I E C T V Y S V D N N G
1 5 10 15 20 25 30 35 40
AACATCTATACCCAGCCTGTGGCCCAGTGGCACGACAGAGGCGAGCAGGAGGTGTTCGAGTACTGTCTGGAGGACGGCTCTCTGATCAGAGCCACCAAGGACCACAAGTTCATGACCGTG
TTGTAGATATGGGTCGGACACCGGGTCACCGTGCTGTCTCCGCTCGTCCTCCACAAGCTCATGACAGACCTCCTGCCGAGAGACTAGTCTCGGTGGTTCCTGGTGTTCAAGTACTGGCAC
N I Y T Q P V A Q W H D R G E Q E V F E Y C L E D G S L I R A T K D H K F M T V
45 50 55 60 65 70 75 80
GACGGCCAGATGCTGCCTATCGACGAGATCTTCGAGAGAGAGCTGGACCTGATGAGAGTGGACAACCTGCCTAACTAA 3′
CTGCCGGTCTACGACGGATAGCTGCTCTAGAAGCTCTCTCTCGACCTGGACTACTCTCACCTGTTGGACGGATTGATT 5′
D G Q M L P I D E I F E R E L D L M R V D N L P N
85 90 95 100 105
Npu C-ter:
5′ ATGATCAAGATCGCCACCAGAAAGTACCTGGGCAAGCAGAACGTGTACGACATCGGCGTGGAGAGGGACCACAACTTCGCCCTGAAGAACGGCTTCATCGCCAGCAACTGCTTCAAC 3′
3′ TACTAGTTCTAGCGGTGGTCTTTCATGGACCGGTTCGTCTTGCACATGCTGTAGCCGCACCTCTCCCTGGTGTTGAAGCGGGACTTCTTGCCGAAGTAGCGGTCGTTGACGAAGTTG 5′
M I K I A T R K Y L G K Q N V Y D I G V E R D H N F A L K N G F I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15O

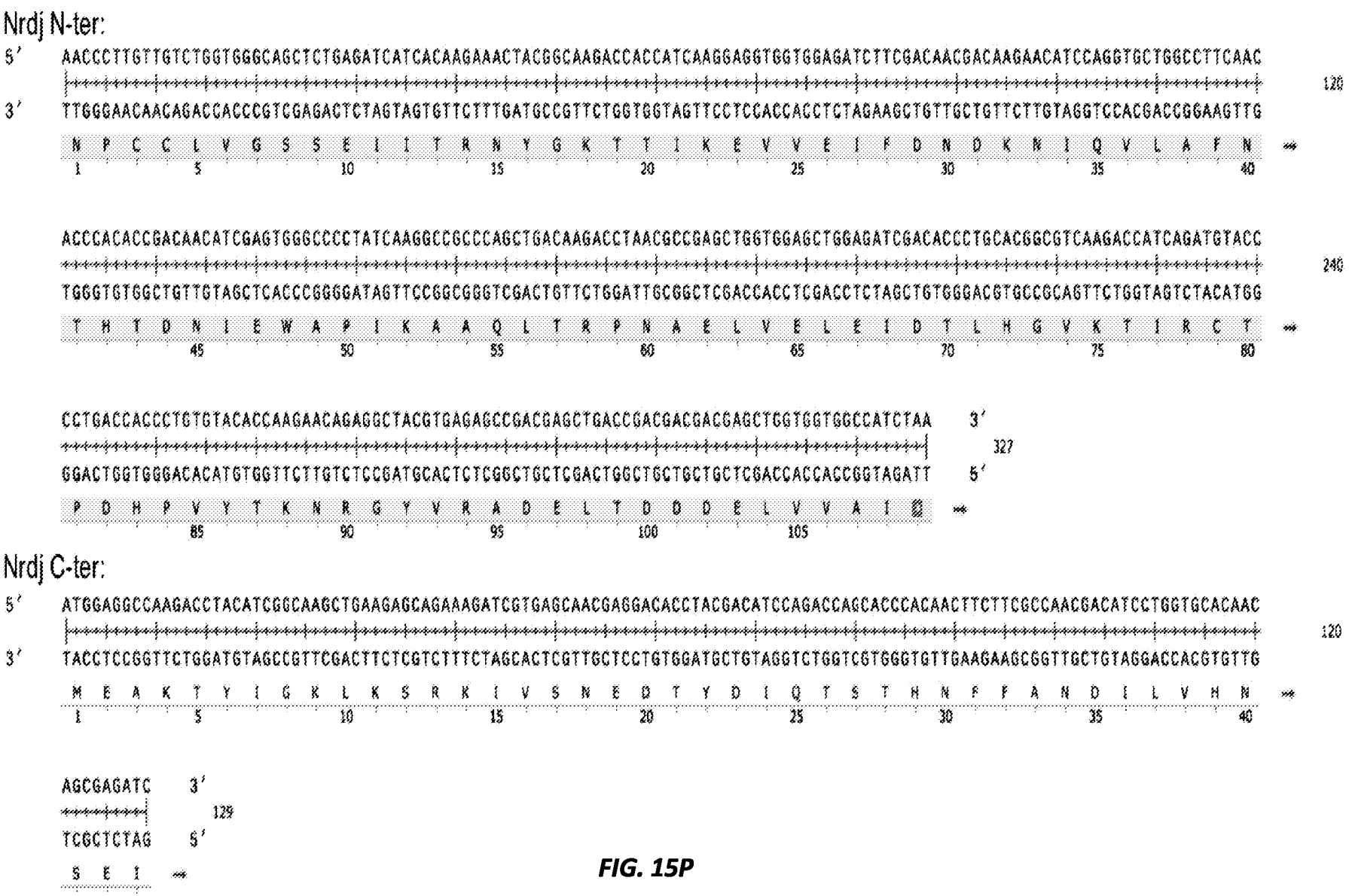

Nrdj N-ter:
5′ AACCCTTGTTGTCTGGTGGGCAGCTCTGAGATCATCACAAGAAACTACGGCAAGACCACCATCAAGGAGGTGGTGGAGATCTTCGACAACGACAAGAACATCCAGGTGCTGGCCTTCAAC
120
3′ TTGGGAACAACAGACCACCCGTCGAGACTCTAGTAGTGTTCTTTGATGCCGTTCTGGTGGTAGTTCCTCCACCACCTCTAGAAGCTGTTGCTGTTCTTGTAGGTCCACGACCGGAAGTTG
N P C C L V G S S E I I T R N Y G K T T I K E V V E I F D N D K N I Q V L A F N
1 5 10 15 20 25 30 35 40
ACCCACACCGACAACATCGAGTGGGCCCCTATCAAGGCCGCCCAGCTGACAAGACCTAACGCCGAGCTGGTGGAGCTGGAGATCGACACCCTGCACGGCGTCAAGACCATCAGATGTACC
240
TGGGTGTGGCTGTTGTAGCTCACCCGGGGATAGTTCCGGCGGGTCGACTGTTCTGGATTGCGGCTCGACCACCTCGACCTCTAGCTGTGGGACGTGCCGCAGTTCTGGTAGTCTACATGG
T H T D N I E W A P I K A A Q L T R P N A E L V E L E I D T L H G V K T I R C T
45 50 55 60 65 70 75 80
CCTGACCACCCTGTGTACACCAAGAACAGAGGCTACGTGAGAGCCGACGAGCTGACCGACGACGACGAGCTGGTGGTGGCCATCTAA 3′
327
GGACTGGTGGGACACATGTGGTTCTTGTCTCCGATGCACTCTCGGCTGCTCGACTGGCTGCTGCTGCTCGACCACCACCGGTAGATT 5′
P D H P V Y T K N R G Y V R A D E L T D D D E L V V A I
85 90 95 100 105
Nrdj C-ter:
5′ ATGGAGGCCAAGACCTACATCGGCAAGCTGAAGAGCAGAAAGATCGTGAGCAACGAGGACACCTACGACATCCAGACCAGCACCCACAACTTCTTCGCCAACGACATCCTGGTGCACAAC
120
3′ TACCTCCGGTTCTGGATGTAGCCGTTCGACTTCTCGTCTTTCTAGCACTCGTTGCTCCTGTGGATGCTGTAGGTCTGGTCGTGGGTGTTGAAGAAGCGGTTGCTGTAGGACCACGTGTTG
M E A K T Y I G K L K S R K I V S N E D T Y D I Q T S T H N F F A N D I L V H N
1 5 10 15 20 25 30 35 40
AGCGAGATC 3′
129
TCGCTCTAG 5′
S E I

FIG. 15P

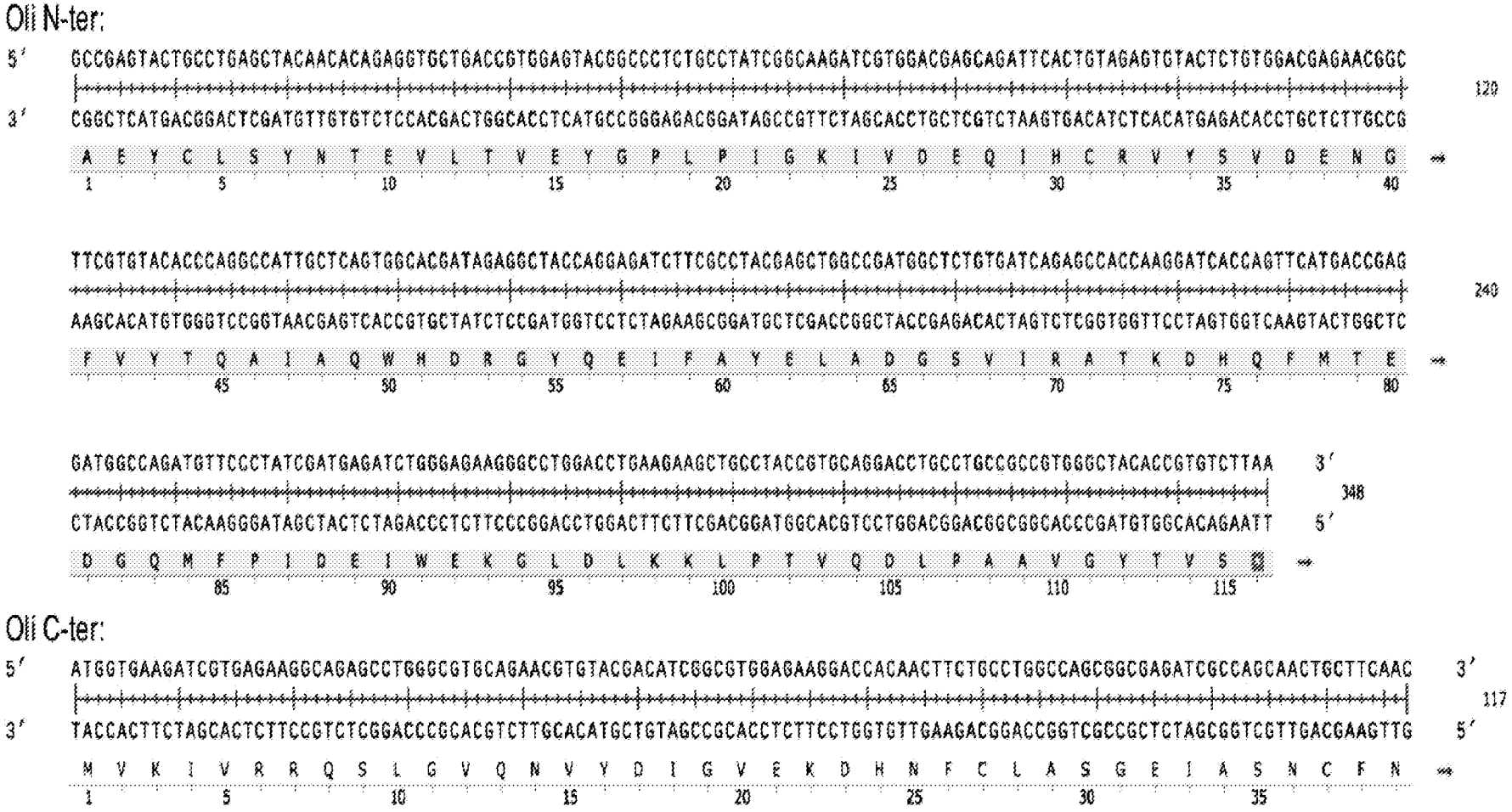

Oli N-ter:
5′ GCCGAGTACTGCCTGAGCTACAACACAGAGGTGCTGACCGTGGAGTACGGCCCTCTGCCTATCGGCAAGATCGTGGACGAGCAGATTCACTGTAGAGTGTACTCTGTGGACGAGAACGGC 120
3′ CGGCTCATGACGGACTCGATGTTGTGTCTCCACGACTGGCACCTCATGCCGGGAGACGGATAGCCGTTCTAGCACCTGCTCGTCTAAGTGACATCTCACATGAGACACCTGCTCTTGCCG
A E Y C L S Y N T E V L T V E Y G P L P I G K I V D E Q I H C R V Y S V D E N G
1 5 10 15 20 25 30 35 40
TTCGTGTACACCCAGGCCATTGCTCAGTGGCACGATAGAGGCTACCAGGAGATCTTCGCCTACGAGCTGGCCGATGGCTCTGTGATCAGAGCCACCAAGGATCACCAGTTCATGACCGAG 240
AAGCACATGTGGGTCCGGTAACGAGTCACCGTGCTATCTCCGATGGTCCTCTAGAAGCGGATGCTCGACCGGCTACCGAGACACTAGTCTCGGTGGTTCCTAGTGGTCAAGTACTGGCTC
F V Y T Q A I A Q W H D R G Y Q E I F A Y E L A D G S V I R A T K D H Q F M T E
45 50 55 60 65 70 75 80
GATGGCCAGATGTTCCCTATCGATGAGATCTGGGAGAAGGGCCTGGACCTGAAGAAGCTGCCTACCGTGCAGGACCTGCCTGCCGCCGTGGGCTACACCGTGTCTTAA 3′ 348
CTACCGGTCTACAAGGGATAGCTACTCTAGACCCTCTTCCCGGACCTGGACTTCTTCGACGGATGGCACGTCCTGGACGGACGGCGGCACCCGATGTGGCACAGAATT 5′
D G Q M F P I D E I W E K G L D L K K L P T V Q D L P A A V G Y T V S
85 90 95 100 105 110 115
Oli C-ter:
5′ ATGGTGAAGATCGTGAGAAGGCAGAGCCTGGGCGTGCAGAACGTGTACGACATCGGCGTGGAGAAGGACCACAACTTCTGCCTGGCCAGCGGCGAGATCGCCAGCAACTGCTTCAAC 3′ 117
3′ TACCACTTCTAGCACTCTTCCGTCTCGGACCCGCACGTCTTGCACATGCTGTAGCCGCACCTCTTCCTGGTGTTGAAGACGGACCGGTCGCCGCTCTAGCGGTCGTTGACGAAGTTG 5′
M V K I V R R Q S L G V Q N V Y D I G V E K D H N F C L A S G E I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15Q

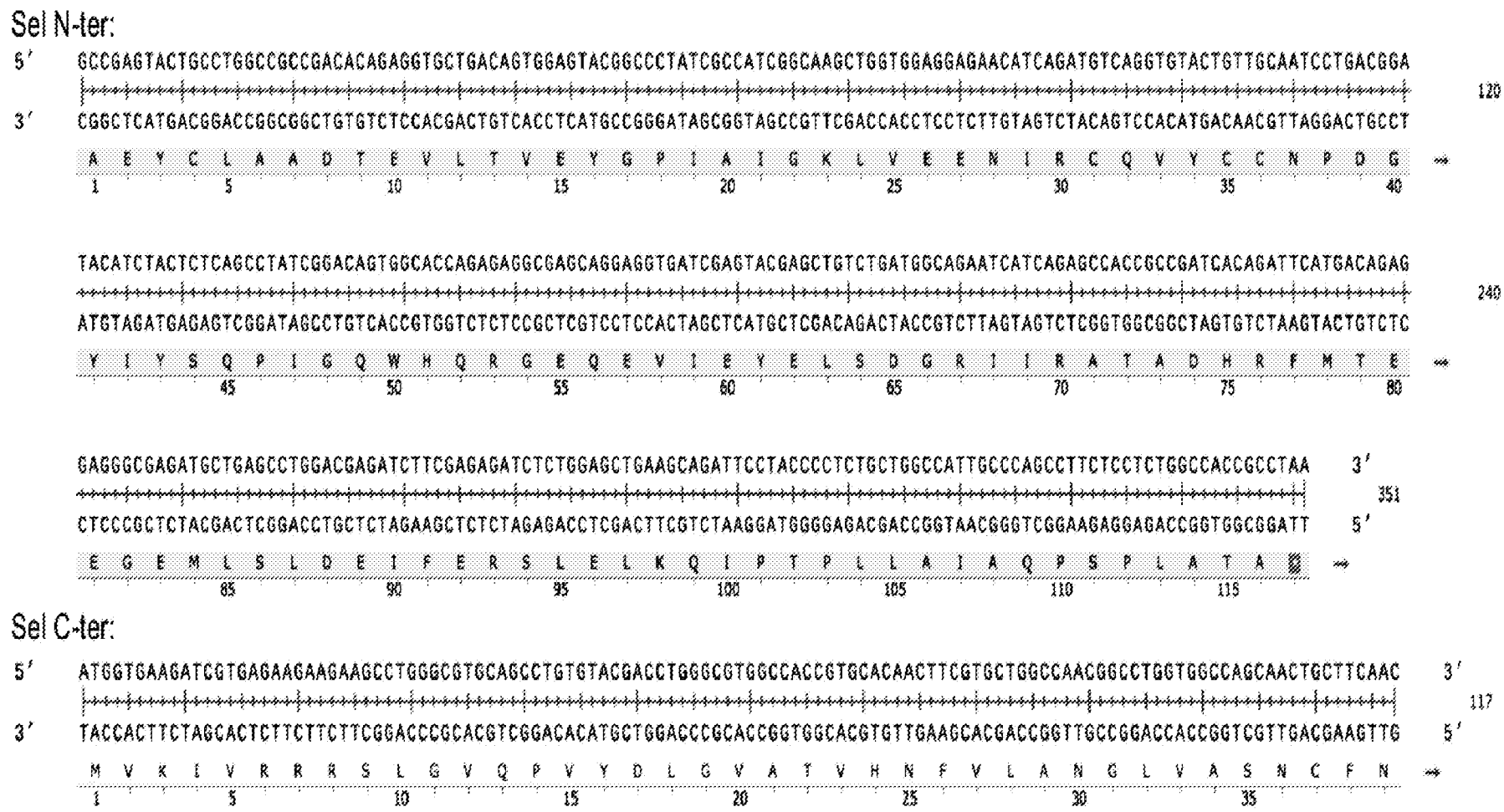

Sel N-ter:
5′ GCCGAGTACTGCCTGGCCGCCGACACAGAGGTGCTGACAGTGGAGTACGGCCCTATCGCCATCGGCAAGCTGGTGGAGGAGAACATCAGATGTCAGGTGTACTGTTGCAATCCTGACGGA 120
3′ CGGCTCATGACGGACCGGCGGCTGTGTCTCCACGACTGTCACCTCATGCCGGGATAGCGGTAGCCGTTCGACCACCTCCTCTTGTAGTCTACAGTCCACATGACAACGTTAGGACTGCCT
A E Y C L A A D T E V L T V E Y G P I A I G K L V E E N I R C Q V Y C C N P D G
1 5 10 15 20 25 30 35 40
TACATCTACTCTCAGCCTATCGGACAGTGGCACCAGAGAGGCGAGCAGGAGGTGATCGAGTACGAGCTGTCTGATGGCAGAATCATCAGAGCCACCGCCGATCACAGATTCATGACAGAG 240
ATGTAGATGAGAGTCGGATAGCCTGTCACCGTGGTCTCTCCGCTCGTCCTCCACTAGCTCATGCTCGACAGACTACCGTCTTAGTAGTCTCGGTGGCGGCTAGTGTCTAAGTACTGTCTC
Y I Y S Q P I G Q W H Q R G E Q E V I E Y E L S D G R I I R A T A D H R F M T E
45 50 55 60 65 70 75 80
GAGGGCGAGATGCTGAGCCTGGACGAGATCTTCGAGAGATCTCTGGAGCTGAAGCAGATTCCTACCCCTCTGCTGGCCATTGCCCAGCCTTCTCCTCTGGCCACCGCCTAA 3′ 351
CTCCCGCTCTACGACTCGGACCTGCTCTAGAAGCTCTCTAGAGACCTCGACTTCGTCTAAGGATGGGGAGACGACCGGTAACGGGTCGGAAGAGGAGACCGGTGGCGGATT 5′
E G E M L S L D E I F E R S L E L K Q I P T P L L A I A Q P S P L A T A
85 90 95 100 105 110 115
Sel C-ter:
5′ ATGGTGAAGATCGTGAGAAGAAGAAGCCTGGGCGTGCAGCCTGTGTACGACCTGGGCGTGGCCACCGTGCACAACTTCGTGCTGGCCAACGGCCTGGTGGCCAGCAACTGCTTCAAC 3′ 117
3′ TACCACTTCTAGCACTCTTCTTCTTCGGACCCGCACGTCGGACACATGCTGGACCCGCACCGGTGGCACGTGTTGAAGCACGACCGGTTGCCGGACCACCGGTCGTTGACGAAGTTG 5′
M V K I V R R R S L G V Q P V Y D L G V A T V H N F V L A N G L V A S N C F N
1 5 10 15 20 25 30 35

FIG. 15R

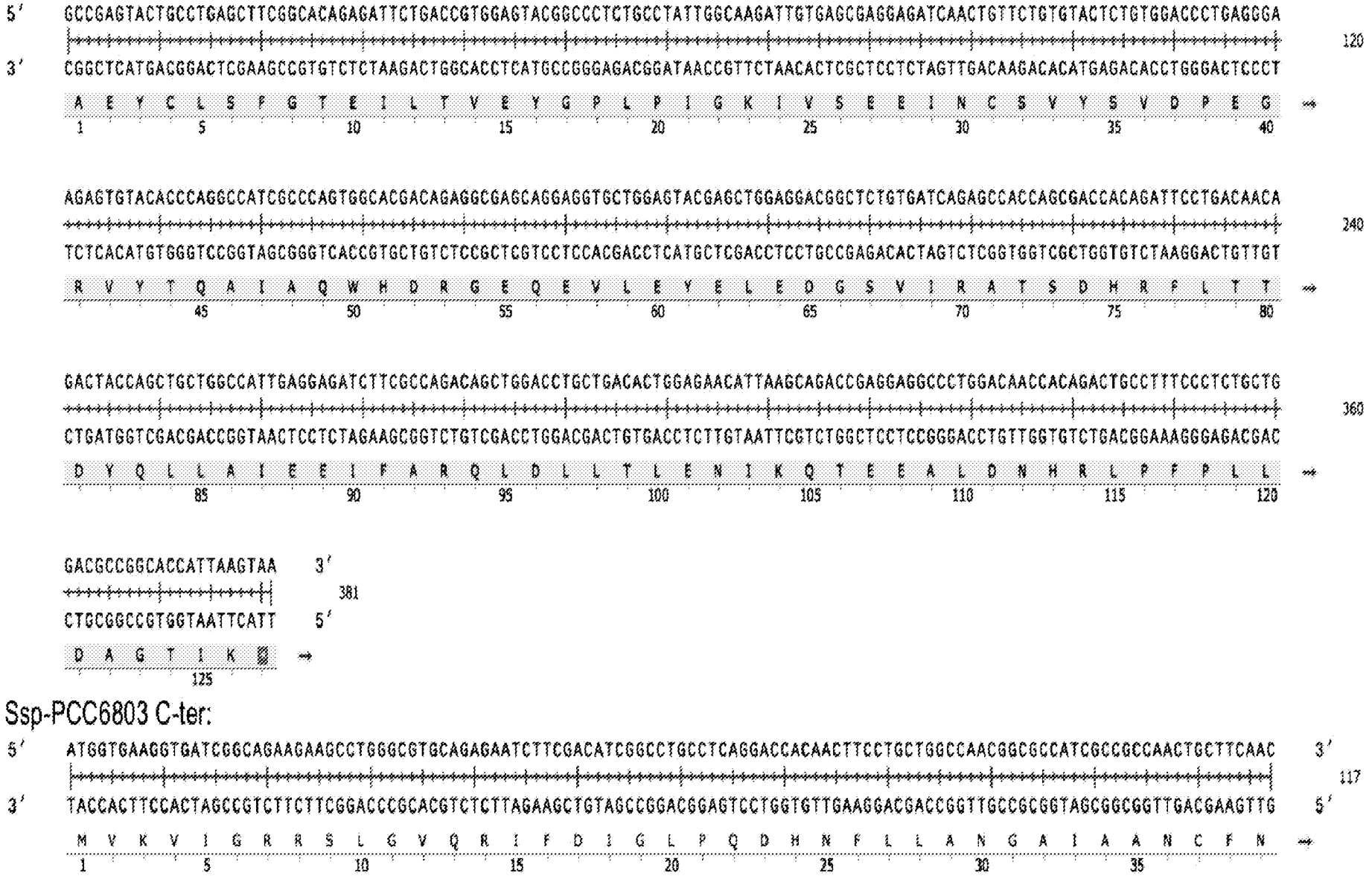

Ssp-PCC6803 N-ter:
5′ GCCGAGTACTGCCTGAGCTTCGGCACAGAGATTCTGACCGTGGAGTACGGCCCTCTGCCTATTGGCAAGATTGTGAGCGAGGAGATCAACTGTTCTGTGTACTCTGTGGACCCTGAGGGA
3′ CGGCTCATGACGGACTCGAAGCCGTGTCTCTAAGACTGGCACCTCATGCCGGGAGACGGATAACCGTTCTAACACTCGCTCCTCTAGTTGACAAGACACATGAGACACCTGGGACTCCCT
120
A E Y C L S F G T E I L T V E Y G P L P I G K I V S E E I N C S V Y S V D P E G
1 5 10 15 20 25 30 35 40
AGAGTGTACACCCAGGCCATCGCCCAGTGGCACGACAGAGGCGAGCAGGAGGTGCTGGAGTACGAGCTGGAGGACGGCTCTGTGATCAGAGCCACCAGCGACCACAGATTCCTGACAACA
TCTCACATGTGGGTCCGGTAGCGGGTCACCGTGCTGTCTCCGCTCGTCCTCCACGACCTCATGCTCGACCTCCTGCCGAGACACTAGTCTCGGTGGTCGCTGGTGTCTAAGGACTGTTGT
240
R V Y T Q A I A Q W H D R G E Q E V L E Y E L E D G S V I R A T S D H R F L T T
45 50 55 60 65 70 75 80
GACTACCAGCTGCTGGCCATTGAGGAGATCTTCGCCAGACAGCTGGACCTGCTGACACTGGAGAACATTAAGCAGACCGAGGAGGCCCTGGACAACCACAGACTGCCTTTCCCTCTGCTG
CTGATGGTCGACGACCGGTAACTCCTCTAGAAGCGGTCTGTCGACCTGGACGACTGTGACCTCTTGTAATTCGTCTGGCTCCTCCGGGACCTGTTGGTGTCTGACGGAAAGGGAGACGAC
360
D Y Q L L A I E E I F A R Q L D L L T L E N I K Q T E E A L D N H R L P F P L L
85 90 95 100 105 110 115 120
GACGCCGGCACCATTAAGTAA 3′
CTGCGGCCGTGGTAATTCATT 5′
381
D A G T I K
125
Ssp-PCC6803 C-ter:
5′ ATGGTGAAGGTGATCGGCAGAAGAAGCCTGGGCGTGCAGAGAATCTTCGACATCGGCCTGCCTCAGGACCACAACTTCCTGCTGGCCAACGGCGCCATCGCCGCCAACTGCTTCAAC 3′
3′ TACCACTTCCACTAGCCGTCTTCTTCGGACCCGCACGTCTCTTAGAAGCTGTAGCCGGACGGAGTCCTGGTGTTGAAGGACGACCGGTTGCCGCGGTAGCGGCGGTTGACGAAGTTG 5′
117
M V K V I G R R S L G V Q R I F D I G L P Q D H N F L L A N G A I A A N C F N
1 5 10 15 20 25 30 35

FIG. 15S

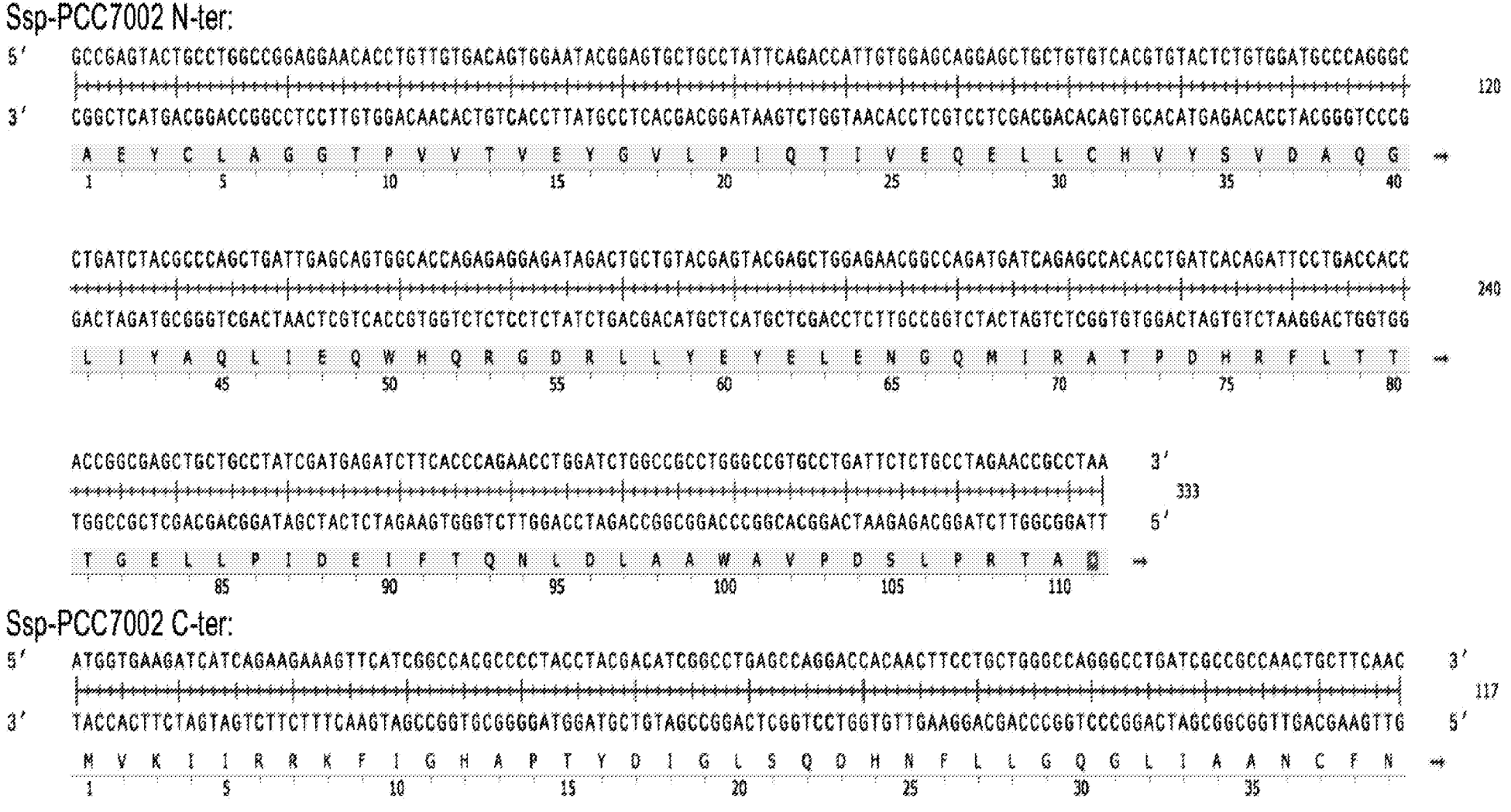

Ssp-PCC7002 N-ter:
5′ GCCGAGTACTGCCTGGCCGGAGGAACACCTGTTGTGACAGTGGAATACGGAGTGCTGCCTATTCAGACCATTGTGGAGCAGGAGCTGCTGTGTCACGTGTACTCTGTGGATGCCCAGGGC
120
3′ CGGCTCATGACGGACCGGCCTCCTTGTGGACAACACTGTCACCTTATGCCTCACGACGGATAAGTCTGGTAACACCTCGTCCTCGACGACACAGTGCACATGAGACACCTACGGGTCCCG
A E Y C L A G G T P V V T V E Y G V L P I Q T I V E Q E L L C H V Y S V D A Q G
1 5 10 15 20 25 30 35 40
CTGATCTACGCCCAGCTGATTGAGCAGTGGCACCAGAGAGGAGATAGACTGCTGTACGAGTACGAGCTGGAGAACGGCCAGATGATCAGAGCCACACCTGATCACAGATTCCTGACCACC
240
GACTAGATGCGGGTCGACTAACTCGTCACCGTGGTCTCTCCTCTATCTGACGACATGCTCATGCTCGACCTCTTGCCGGTCTACTAGTCTCGGTGTGGACTAGTGTCTAAGGACTGGTGG
L I Y A Q L I E Q W H Q R G D R L L Y E Y E L E N G Q M I R A T P D H R F L T T
45 50 55 60 65 70 75 80
ACCGGCGAGCTGCTGCCTATCGATGAGATCTTCACCCAGAACCTGGATCTGGCCGCCTGGGCCGTGCCTGATTCTCTGCCTAGAACCGCCTAA 3′
333
TGGCCGCTCGACGACGGATAGCTACTCTAGAAGTGGGTCTTGGACCTAGACCGGCGGACCCGGCACGGACTAAGAGACGGATCTTGGCGGATT 5′
T G E L L P I D E I F T Q N L D L A A W A V P D S L P R T A
85 90 95 100 105 110
Ssp-PCC7002 C-ter:
5′ ATGGTGAAGATCATCAGAAGAAAGTTCATCGGCCACGCCCCTACCTACGACATCGGCCTGAGCCAGGACCACAACTTCCTGCTGGGCCAGGGCCTGATCGCCGCCAACTGCTTCAAC 3′
117
3′ TACCACTTCTAGTAGTCTTCTTTCAAGTAGCCGGTGCGGGGATGGATGCTGTAGCCGGACTCGGTCCTGGTGTTGAAGGACGACCCGGTCCCGGACTAGCGGCGGTTGACGAAGTTG 5′
M V K I I R R K F I G H A P T Y D I G L S Q D H N F L L G Q G L I A A N C F N
1 5 10 15 20 25 30 35

FIG. 15T

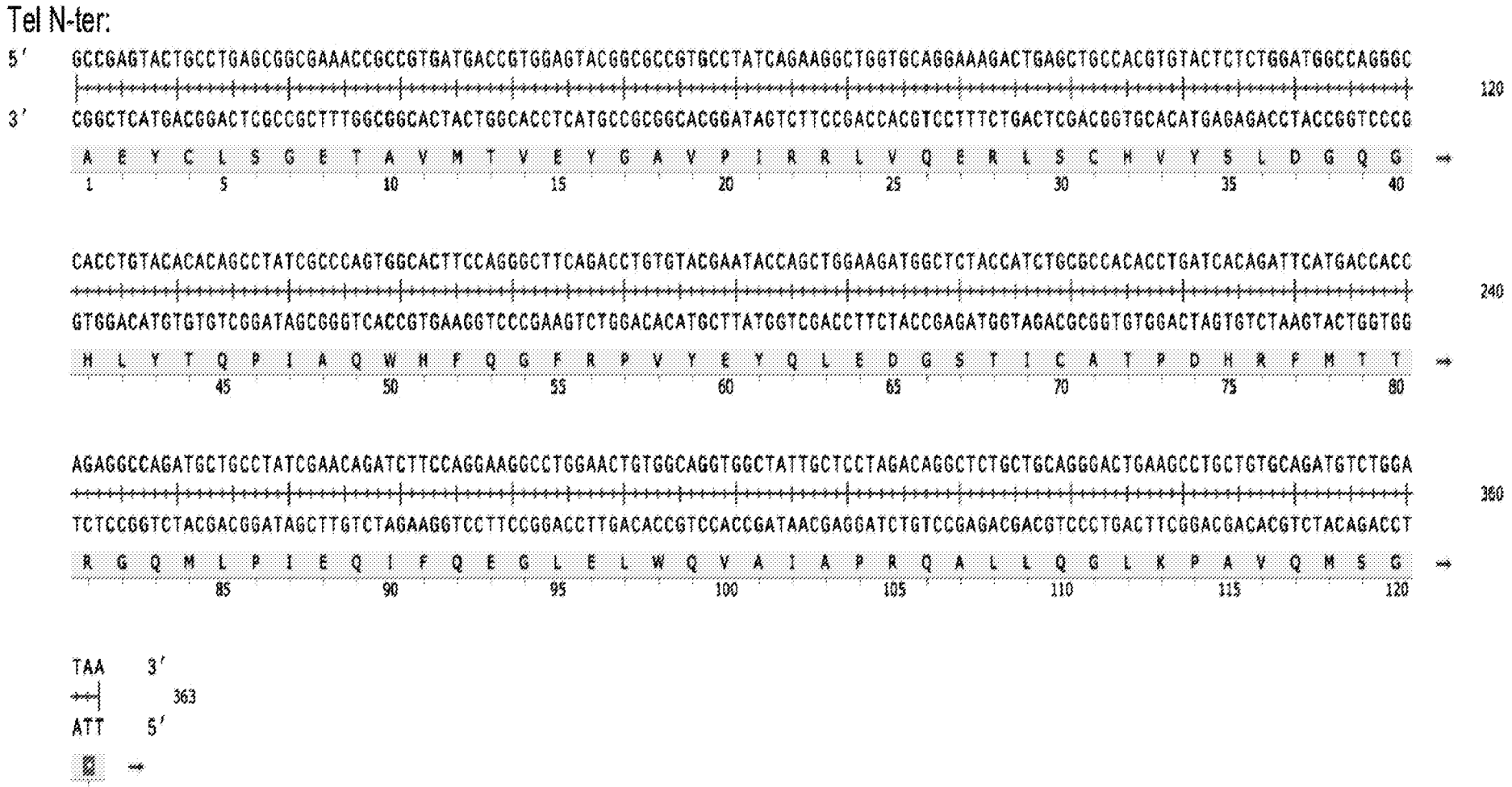

Tel N-ter:
5′ GCCGAGTACTGCCTGAGCGGCGAAACCGCCGTGATGACCGTGGAGTACGGCGCCGTGCCTATCAGAAGGCTGGTGCAGGAAAGACTGAGCTGCCACGTGTACTCTCTGGATGGCCAGGGC
120
3′ CGGCTCATGACGGACTCGCCGCTTTGGCGGCACTACTGGCACCTCATGCCGCGGCACGGATAGTCTTCCGACCACGTCCTTTCTGACTCGACGGTGCACATGAGAGACCTACCGGTCCCG
A E Y C L S G E T A V M T V E Y G A V P I R R L V Q E R L S C H V Y S L D G Q G
1 5 10 15 20 25 30 35 40
CACCTGTACACACAGCCTATCGCCCAGTGGCACTTCCAGGGCTTCAGACCTGTGTACGAATACCAGCTGGAAGATGGCTCTACCATCTGCGCCACACCTGATCACAGATTCATGACCACC
240
GTGGACATGTGTGTCGGATAGCGGGTCACCGTGAAGGTCCCGAAGTCTGGACACATGCTTATGGTCGACCTTCTACCGAGATGGTAGACGCGGTGTGGACTAGTGTCTAAGTACTGGTGG
H L Y T Q P I A Q W H F Q G F R P V Y E Y Q L E D G S T I C A T P D H R F M T T
45 50 55 60 65 70 75 80
AGAGGCCAGATGCTGCCTATCGAACAGATCTTCCAGGAAGGCCTGGAACTGTGGCAGGTGGCTATTGCTCCTAGACAGGCTCTGCTGCAGGGACTGAAGCCTGCTGTGCAGATGTCTGGA
360
TCTCCGGTCTACGACGGATAGCTTGTCTAGAAGGTCCTTCCGGACCTTGACACCGTCCACCGATAACGAGGATCTGTCCGAGACGACGTCCCTGACTTCGGACGACACGTCTACAGACCT
R G Q M L P I E Q I F Q E G L E L W Q V A I A P R Q A L L Q G L K P A V Q M S G
85 90 95 100 105 110 115 120
TAA 3′
363
ATT 5′
Tel C-ter:
5′ ATGAAGATCGTGGGCAGGAGACTGATGGGCTGGCAGGCCGTGTACGACATCGGCCTGGCCGCCGACCACAACTTCGTGCTGGCCAACGGCGCCATCGCCGCCAACTGCTTCAAC 3′
114
3′ TACTTCTAGCACCCGTCCTCTGACTACCCGACCGTCCGGCACATGCTGTAGCCGGACCGGCGGCTGGTGTTGAAGCACGACCGGTTGCCGCGGTAGCGGCGGTTGACGAAGTTG 5′
M K I V G R R L M G W Q A V Y D I G L A A D H N F V L A N G A I A A N C F N
1 5 10 15 20 25 30 35

FIG. 15U

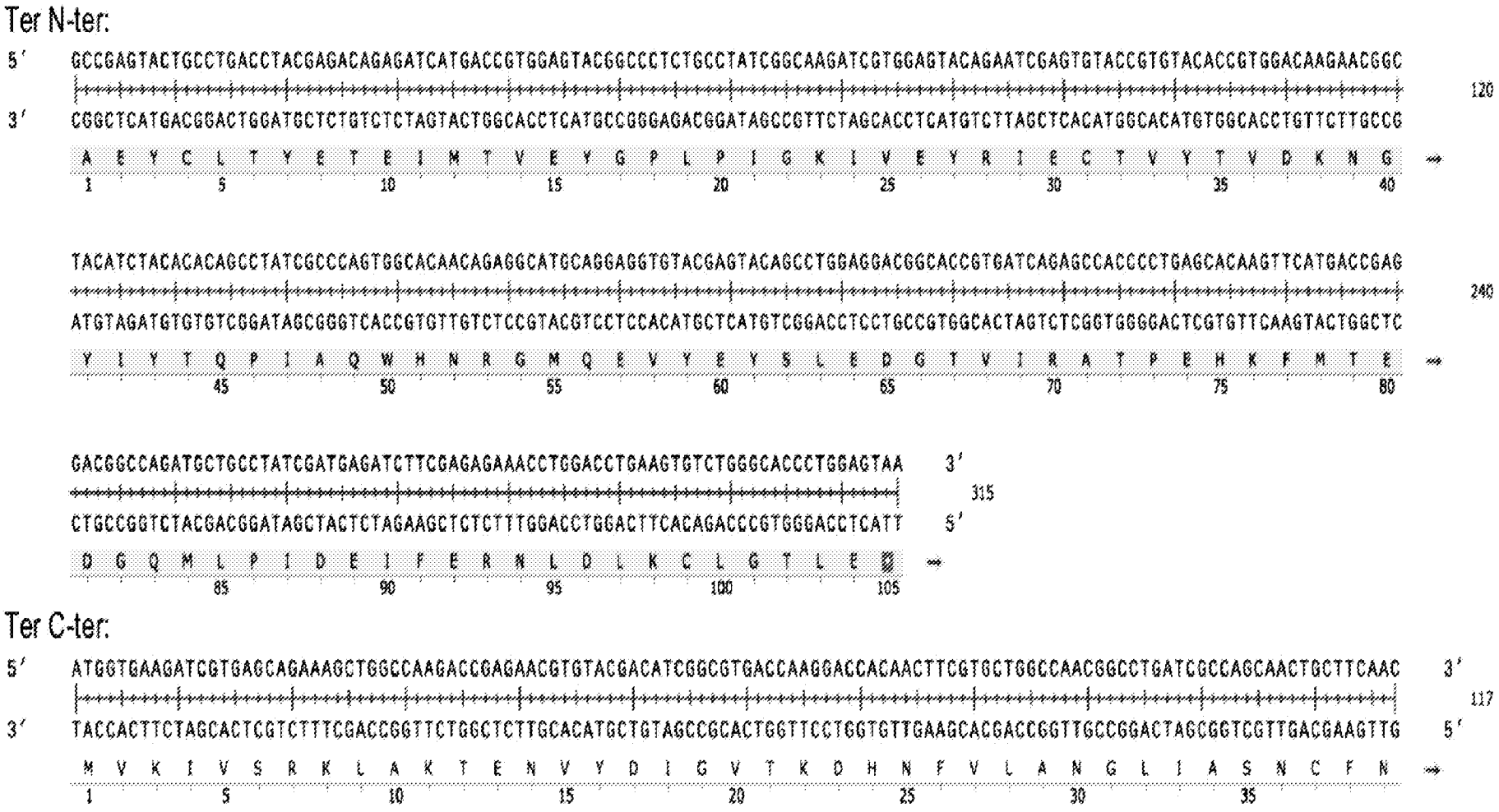

Ter N-ter:
5′ GCCGAGTACTGCCTGACCTACGAGACAGAGATCATGACCGTGGAGTACGGCCCTCTGCCTATCGGCAAGATCGTGGAGTACAGAATCGAGTGTACCGTGTACACCGTGGACAAGAACGGC 120
3′ CGGCTCATGACGGACTGGATGCTCTGTCTCTAGTACTGGCACCTCATGCCGGGAGACGGATAGCCGTTCTAGCACCTCATGTCTTAGCTCACATGGCACATGTGGCACCTGTTCTTGCCG
A E Y C L T Y E T E I M T V E Y G P L P I G K I V E Y R I E C T V Y T V D K N G
1 5 10 15 20 25 30 35 40
TACATCTACACACAGCCTATCGCCCAGTGGCACAACAGAGGCATGCAGGAGGTGTACGAGTACAGCCTGGAGGACGGCACCGTGATCAGAGCCACCCCTGAGCACAAGTTCATGACCGAG 240
ATGTAGATGTGTGTCGGATAGCGGGTCACCGTGTTGTCTCCGTACGTCCTCCACATGCTCATGTCGGACCTCCTGCCGTGGCACTAGTCTCGGTGGGGACTCGTGTTCAAGTACTGGCTC
Y I Y T Q P I A Q W H N R G M Q E V Y E Y S L E D G T V I R A T P E H K F M T E
45 50 55 60 65 70 75 80
GACGGCCAGATGCTGCCTATCGATGAGATCTTCGAGAGAAACCTGGACCTGAAGTGTCTGGGCACCCTGGAGTAA 3′ 315
CTGCCGGTCTACGACGGATAGCTACTCTAGAAGCTCTCTTTGGACCTGGACTTCACAGACCCGTGGGACCTCATT 5′
D G Q M L P I D E I F E R N L D L K C L G T L E
85 90 95 100 105
Ter C-ter:
5′ ATGGTGAAGATCGTGAGCAGAAAGCTGGCCAAGACCGAGAACGTGTACGACATCGGCGTGACCAAGGACCACAACTTCGTGCTGGCCAACGGCCTGATCGCCAGCAACTGCTTCAAC 3′ 117
3′ TACCACTTCTAGCACTCGTCTTTCGACCGGTTCTGGCTCTTGCACATGCTGTAGCCGCACTGGTTCCTGGTGTTGAAGCACGACCGGTTGCCGGACTAGCGGTCGTTGACGAAGTTG 5′
M V K I V S R K L A K T E N V Y D I G V T K D H N F V L A N G L I A S N C F N
1 5 10 15 20 25 30 35

FIG. 15V

FIG. 15W

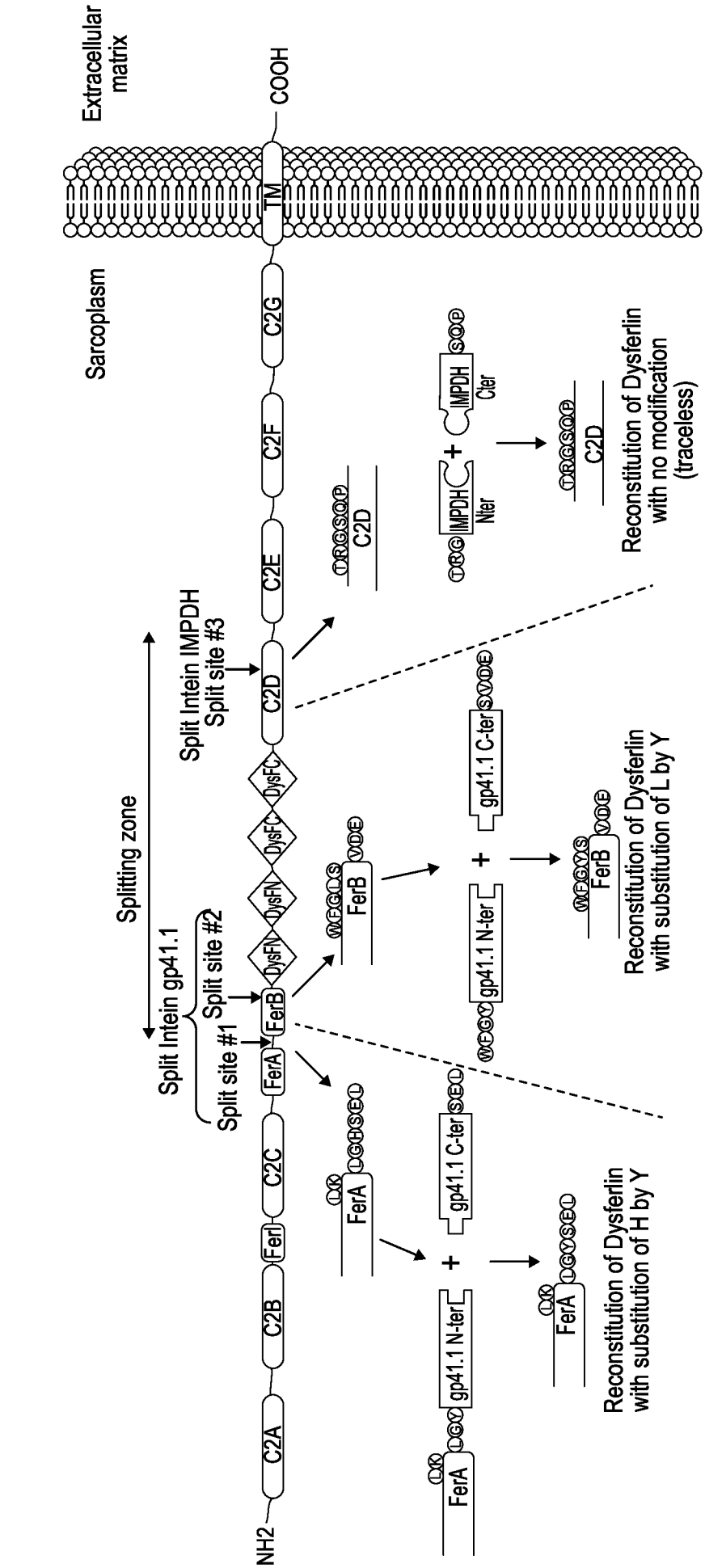

Extracellular matrix
COOH
TM
Sarcoplasm
C2G
C2F
C2E
C2D
DysFC
DysFC
DysFN
DysFN
FerB
FerA
C2C
FerI
C2B
C2A
NH2
Splitting zone
Split Intein gp41.1
Split site #1
Split site #2
Split Intein IMPDH
Split site #3
FerB
gp41.1 N-ter
gp41.1 C-ter
FerA
C2D
IMPDH Nter
IMPDH Cter
Reconstitution of Dysferlin with no modification (traceless)
Reconstitution of Dysferlin with substitution of L by Y
Reconstitution of Dysferlin with substitution of H by Y

FIG. 16

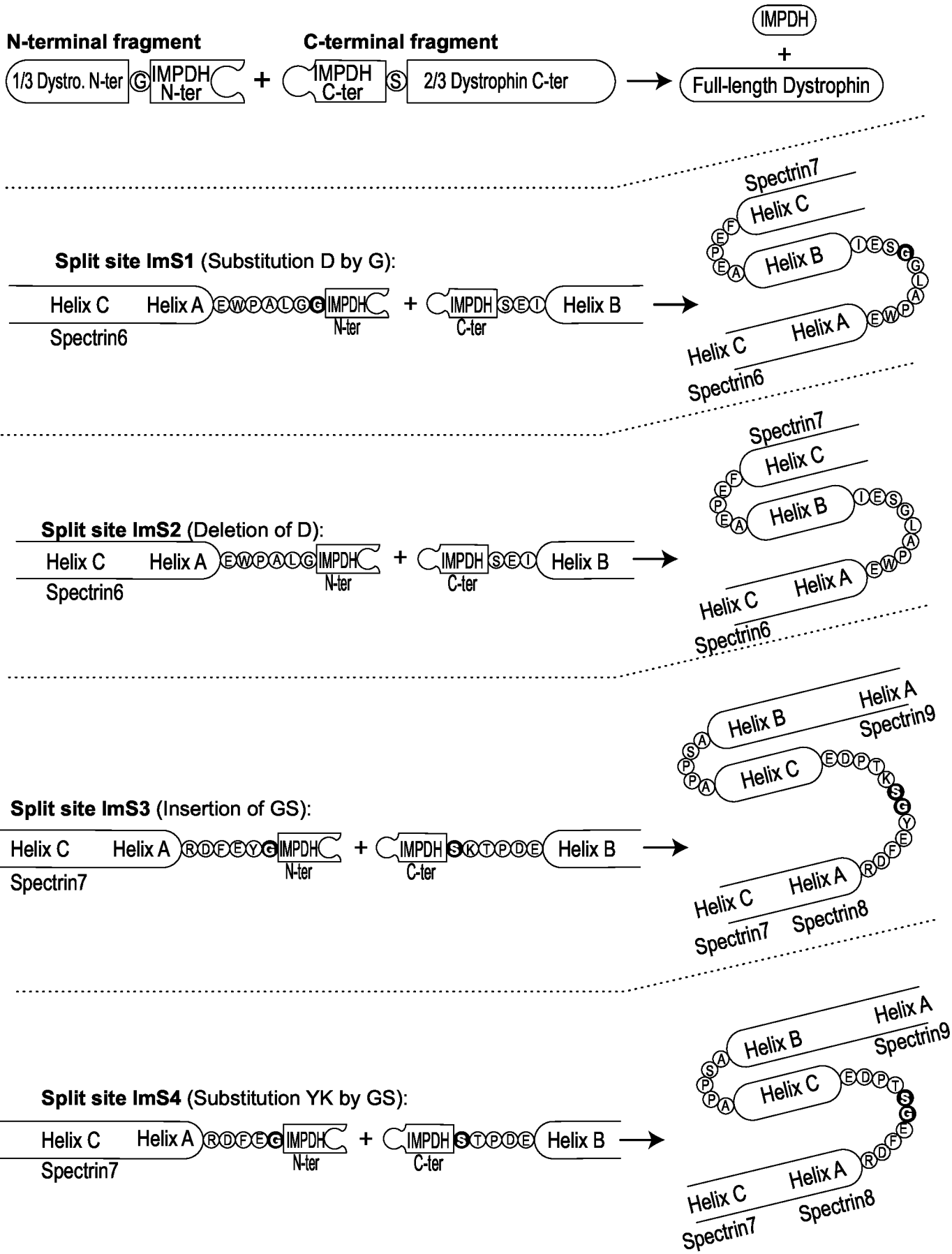

N-terminal fragment
C-terminal fragment
1/3 Dystro. N-ter
IMPDH N-ter
IMPDH C-ter
2/3 Dystrophin C-ter
IMPDH
+
Full-length Dystrophin
Split site ImS1 (Substitution D by G):
Helix C
Helix A
Spectrin6
Helix B
Spectrin7
Split site ImS2 (Deletion of D):
Split site ImS3 (Insertion of GS):
Spectrin8
Split site ImS4 (Substitution YK by GS):

FIG. 17

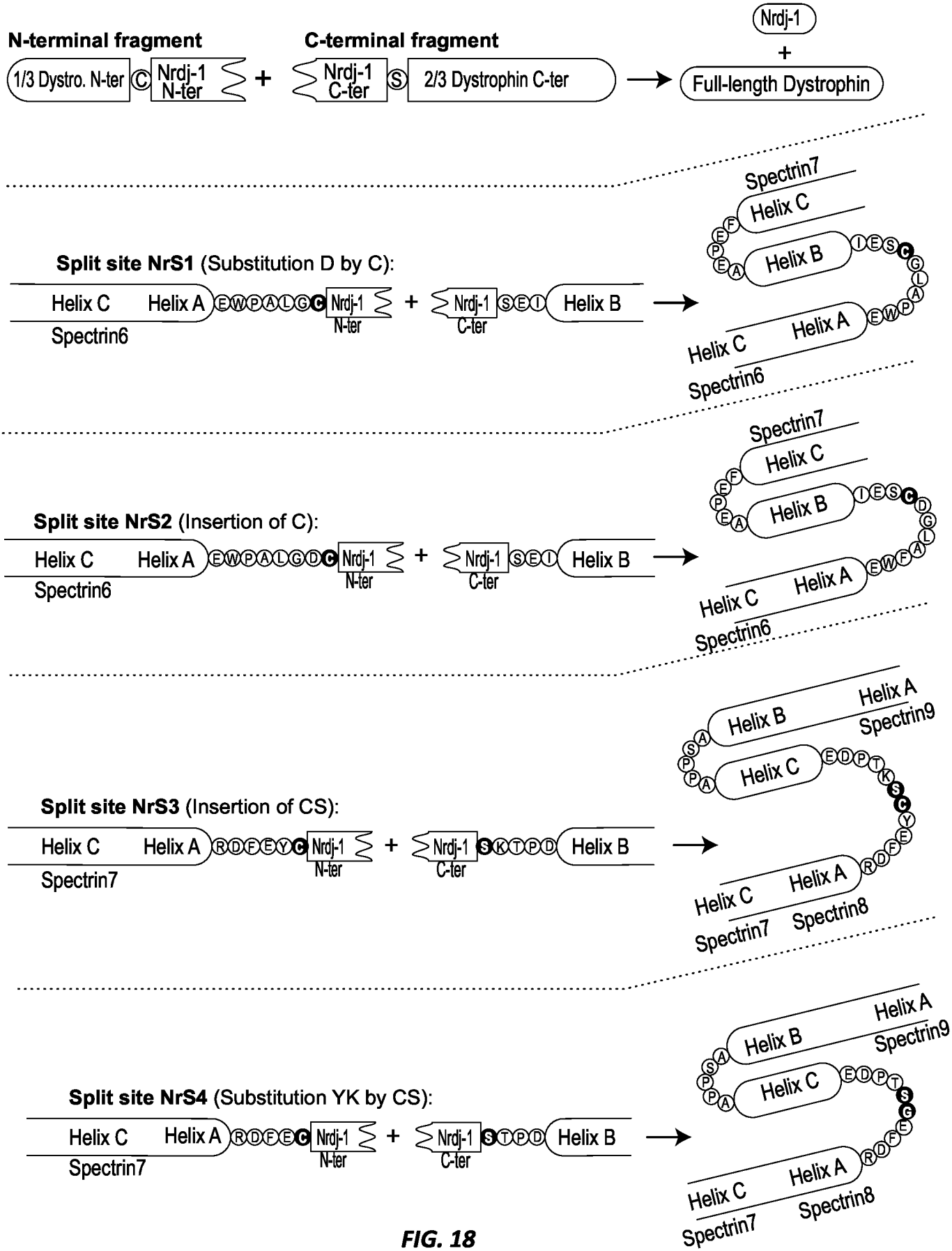

N-terminal fragment
C-terminal fragment
1/3 Dystro. N-ter
Nrdj-1 N-ter
Nrdj-1 C-ter
2/3 Dystrophin C-ter
Nrdj-1
+
Full-length Dystrophin
Split site NrS1 (Substitution D by C):
Helix C
Helix A
Nrdj-1
N-ter
C-ter
Helix B
Spectrin6
Spectrin7
Split site NrS2 (Insertion of C):
Split site NrS3 (Insertion of CS):
Spectrin9
Spectrin8
Split site NrS4 (Substitution YK by CS):

FIG. 18

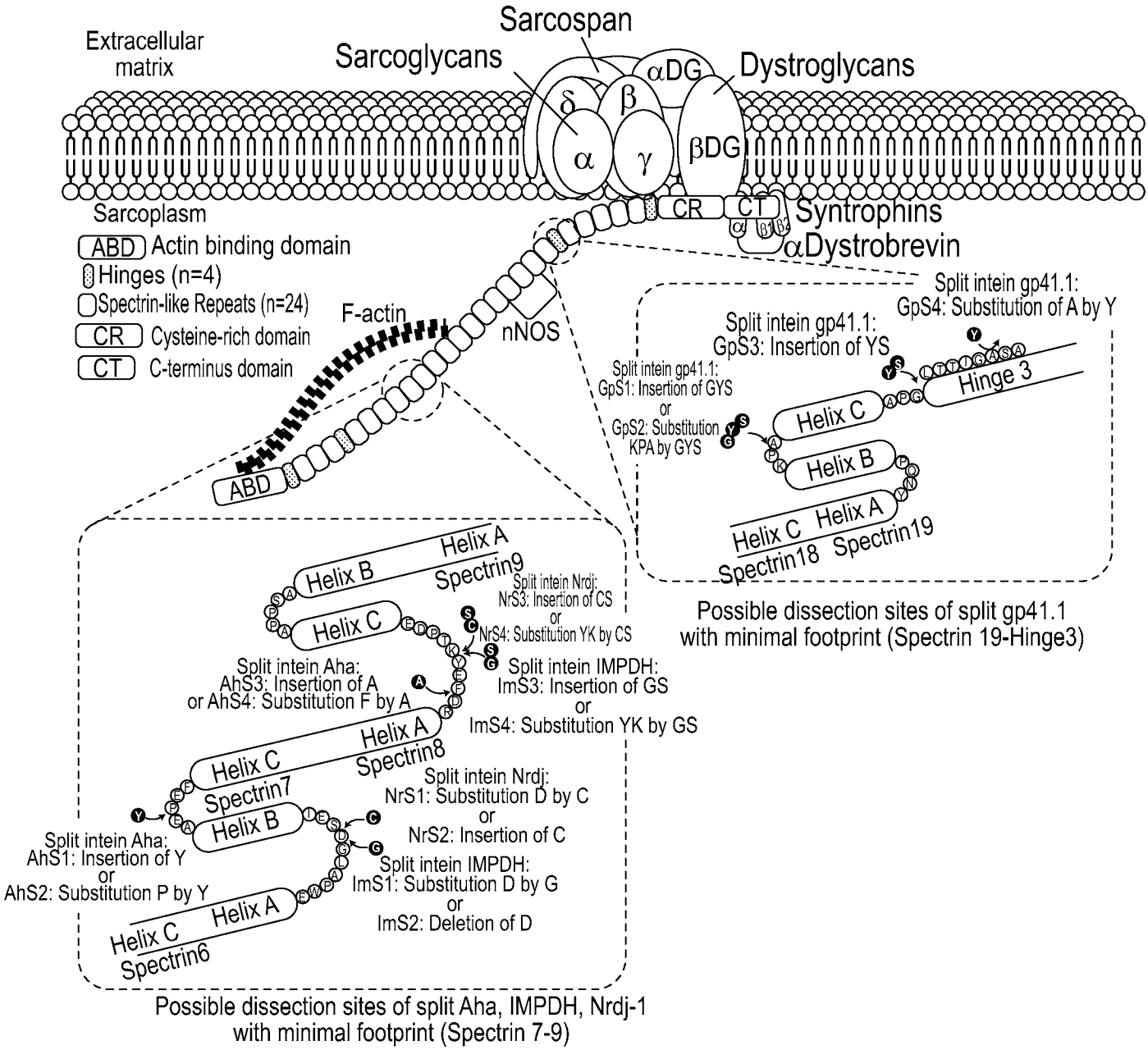

Extracellular matrix
Sarcospan
Sarcoglycans
αDG
Dystroglycans
δ
β
α
γ
βDG
CR
CT
Syntrophins
αDystrobrevin
Sarcoplasm
ABD Actin binding domain
Hinges (n=4)
Spectrin-like Repeats (n=24)
CR Cysteine-rich domain
CT C-terminus domain
F-actin
nNOS
ABD
Split intein gp41.1:
GpS4: Substitution of A by Y
Split intein gp41.1:
GpS3: Insertion of YS
Split intein gp41.1:
GpS1: Insertion of GYS
or
GpS2: Substitution KPA by GYS
Hinge 3
Helix C
Helix B
Helix C Helix A
Spectrin18 Spectrin19
Possible dissection sites of split gp41.1
with minimal footprint (Spectrin 19-Hinge3)
Helix B
Helix A
Spectrin9
Split intein Nrdj:
NrS3: Insertion of CS
or
NrS4: Substitution YK by CS
Helix C
Split intein Aha:
AhS3: Insertion of A
or AhS4: Substitution F by A
Split intein IMPDH:
ImS3: Insertion of GS
or
ImS4: Substitution YK by GS
Helix C
Helix A
Spectrin8
Spectrin7
Split intein Nrdj:
NrS1: Substitution D by C
or
NrS2: Insertion of C
Helix B
Split intein Aha:
AhS1: Insertion of Y
or
AhS2: Substitution P by Y
Split intein IMPDH:
ImS1: Substitution D by G
or
ImS2: Deletion of D
Helix C
Helix A
Spectrin6
Possible dissection sites of split Aha, IMPDH, Nrdj-1
with minimal footprint (Spectrin 7-9)

FIG. 19

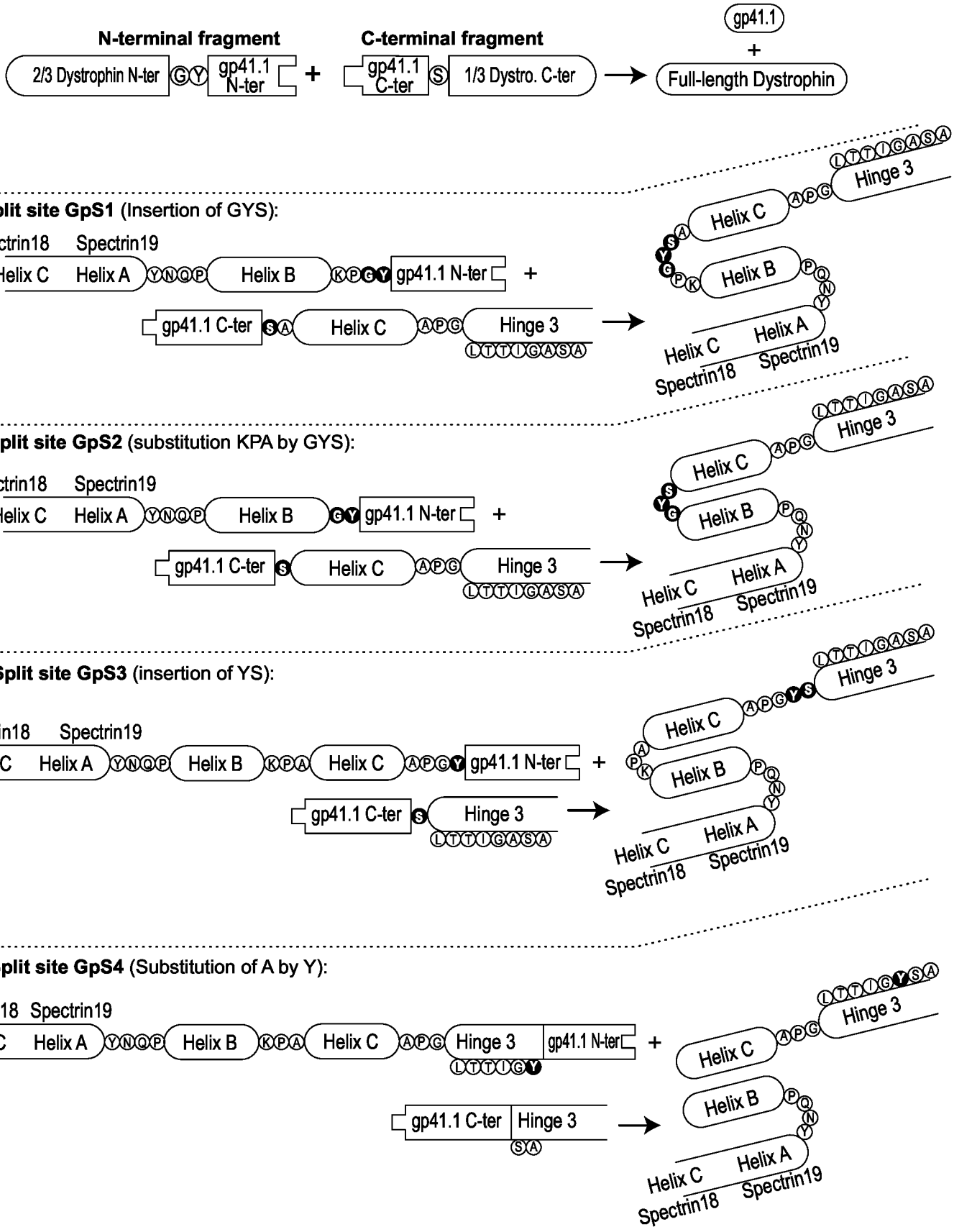

N-terminal fragment
C-terminal fragment
gp41.1
+
2/3 Dystrophin N-ter
GY
gp41.1 N-ter
+
gp41.1 C-ter
S
1/3 Dystro. C-ter
Full-length Dystrophin
Split site GpS1 (Insertion of GYS):
Spectrin18
Spectrin19
Helix C
Helix A
YNQP
Helix B
KPGY
gp41.1 N-ter
+
gp41.1 C-ter
SA
Helix C
APG
Hinge 3
LTTIGASA
Helix C
Helix A
Helix B
Hinge 3
Split site GpS2 (substitution KPA by GYS):
Split site GpS3 (insertion of YS):
Split site GpS4 (Substitution of A by Y):

FIG. 20

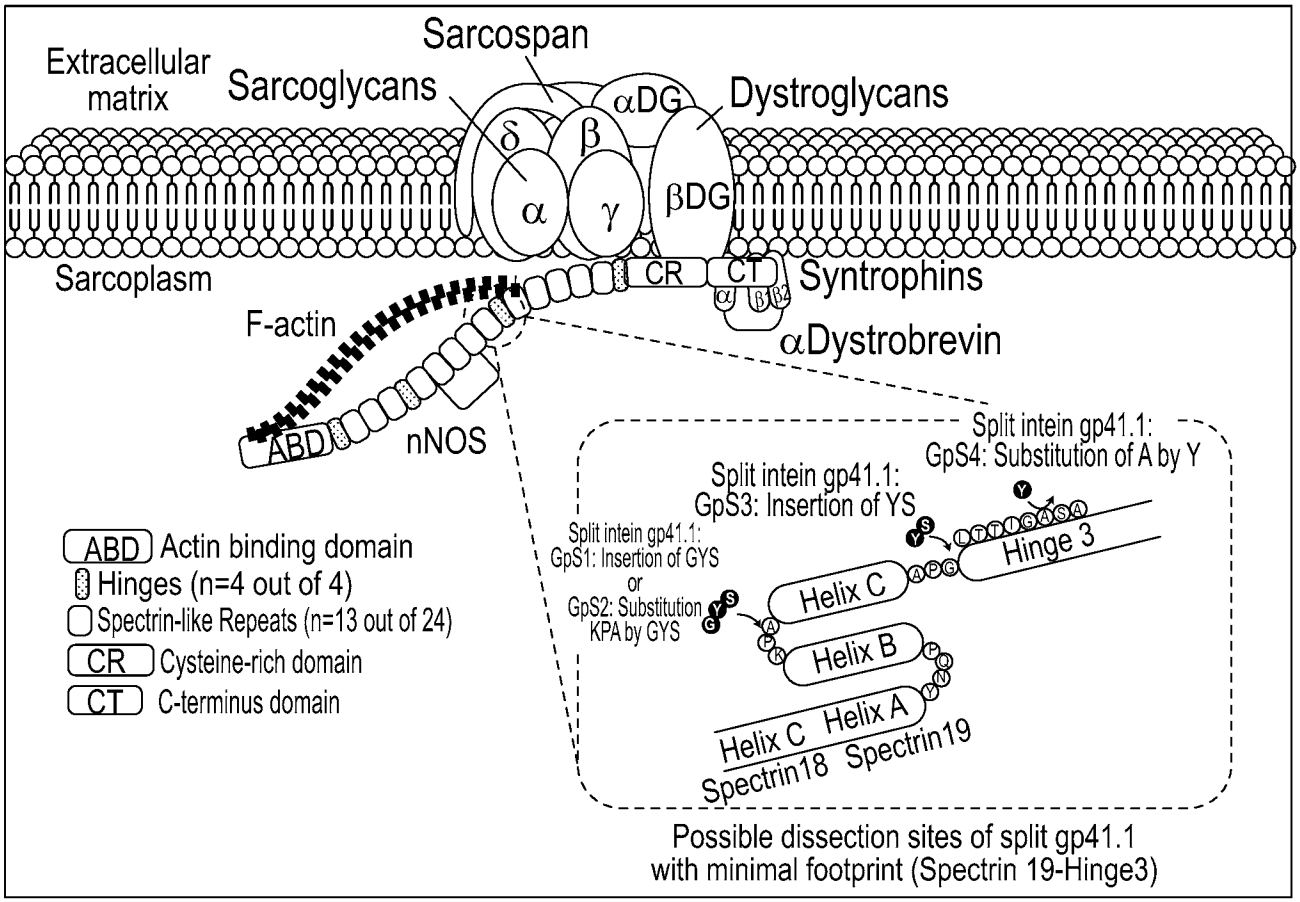

Extracellular matrix
Sarcoglycans
Sarcospan
αDG
Dystroglycans
δ
β
α
γ
βDG
Sarcoplasm
CR
CT
Syntrophins
α
β1
β2
F-actin
αDystrobrevin
ABD
nNOS
Split intein gp41.1:
GpS4: Substitution of A by Y
Split intein gp41.1:
GpS3: Insertion of YS
Split intein gp41.1:
GpS1: Insertion of GYS
or
GpS2: Substitution
KPA by GYS
Helix C
Hinge 3
Helix B
Helix C
Helix A
Spectrin18
Spectrin19
ABD Actin binding domain
Hinges (n=4 out of 4)
Spectrin-like Repeats (n=13 out of 24)
CR Cysteine-rich domain
CT C-terminus domain
Possible dissection sites of split gp41.1
with minimal footprint (Spectrin 19-Hinge3)

FIG. 21

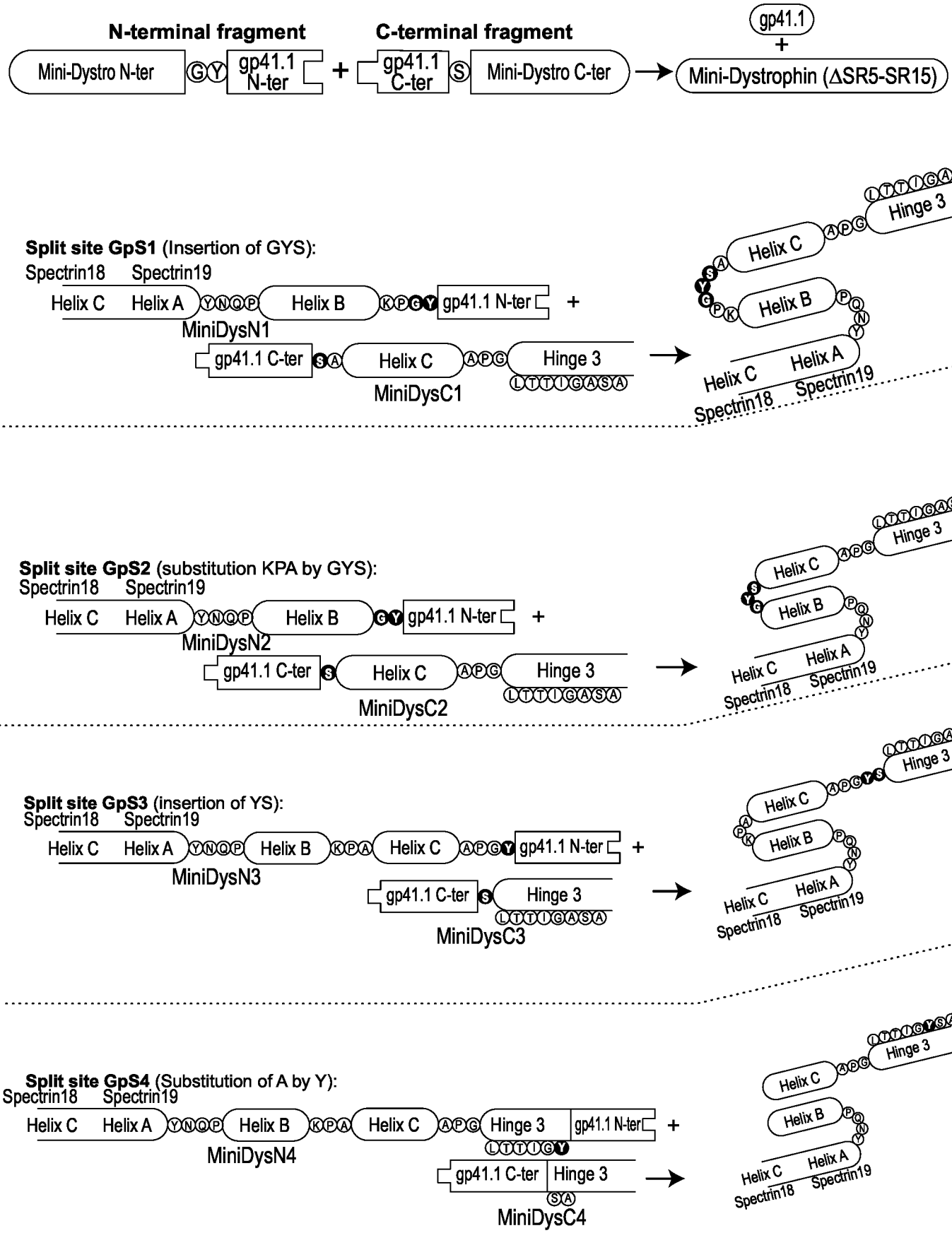

N-terminal fragment
C-terminal fragment
gp41.1
Mini-Dystro N-ter
GY
gp41.1 N-ter
gp41.1 C-ter
S
Mini-Dystro C-ter
Mini-Dystrophin (ΔSR5-SR15)
Split site GpS1 (Insertion of GYS):
Spectrin18 Spectrin19
Helix C Helix A YNQP Helix B KPGY gp41.1 N-ter
MiniDysN1
gp41.1 C-ter SA Helix C APG Hinge 3
LTTIGASA
MiniDysC1
Split site GpS2 (substitution KPA by GYS):
Helix C Helix A YNQP Helix B GY gp41.1 N-ter
MiniDysN2
gp41.1 C-ter S Helix C APG Hinge 3
LTTIGASA
MiniDysC2
Split site GpS3 (insertion of YS):
Helix C Helix A YNQP Helix B KPA Helix C APGY gp41.1 N-ter
MiniDysN3
gp41.1 C-ter S Hinge 3
LTTIGASA
MiniDysC3
Split site GpS4 (Substitution of A by Y):
Helix C Helix A YNQP Helix B KPA Helix C APG Hinge 3 gp41.1 N-ter
LTTIGY
MiniDysN4
gp41.1 C-ter Hinge 3
SA
MiniDysC4

FIG. 21, cont'd

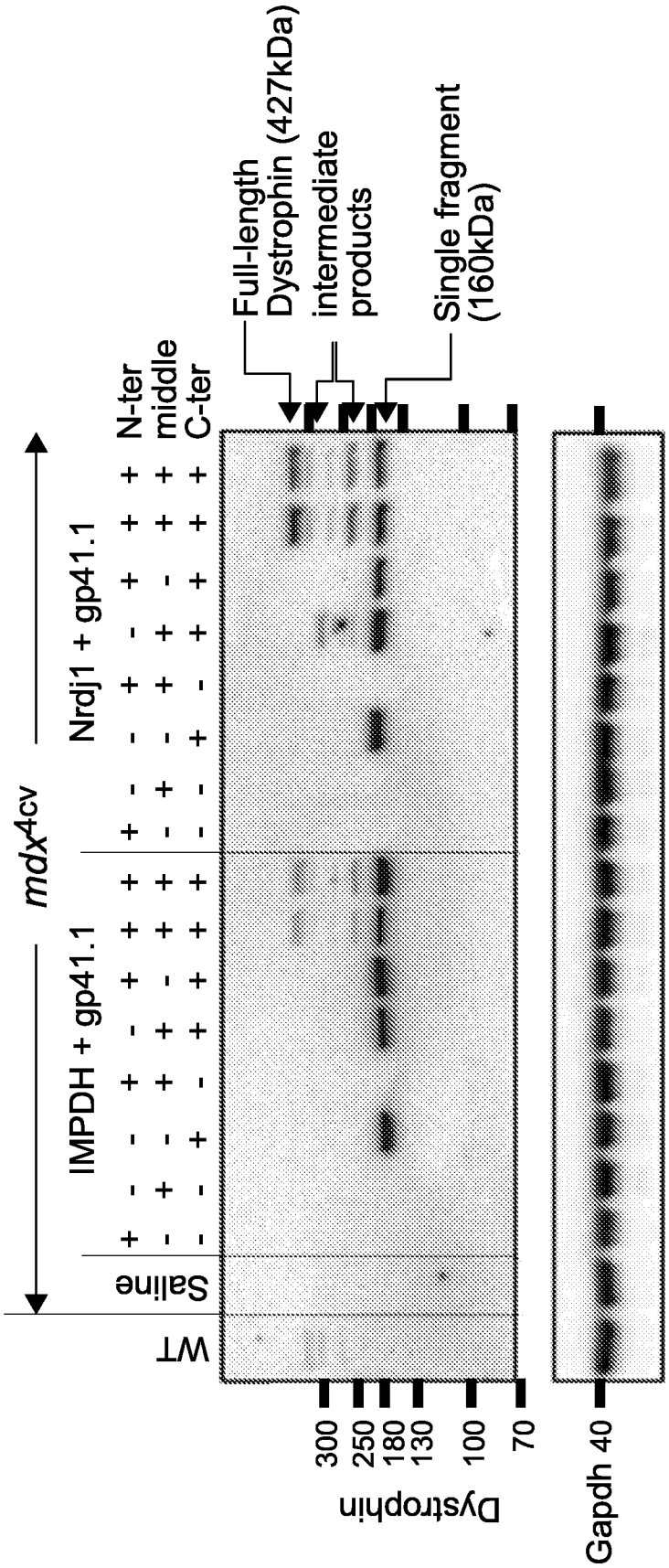

Full-length Dystrophin (427kDa)
intermediate products
Single fragment (160kDa)
N-ter
middle
C-ter
Nrdj1 + gp41.1
IMPDH + gp41.1
mdx4cv
WT
Saline
Dystrophin
300
250
180
130
100
70
Gapdh 40

FIG.22A

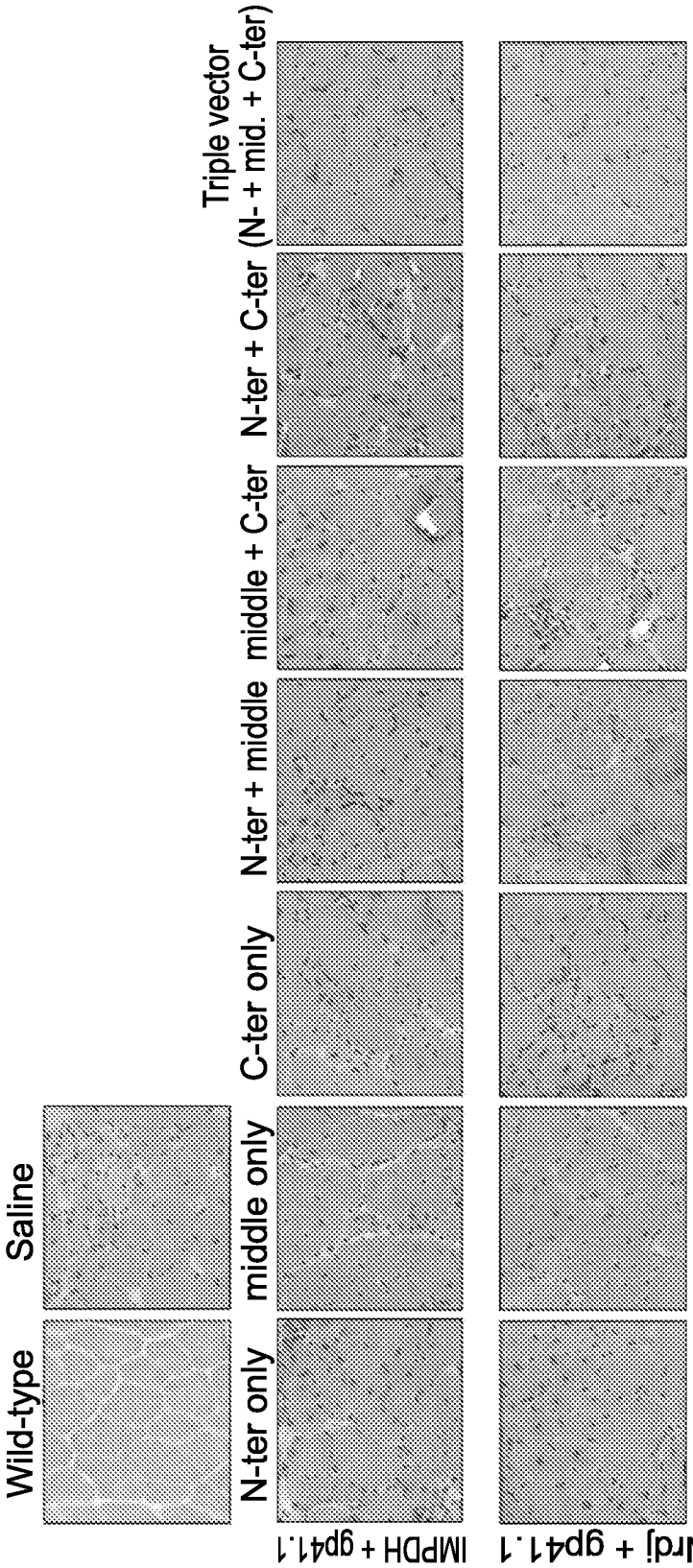

Wild-type
Saline
N-ter only
middle only
C-ter only
N-ter + middle
middle + C-ter
N-ter + C-ter
Triple vector (N- + mid. + C-ter)
IMPDH + gp41.1
Nrdj + gp41.1

FIG.22B

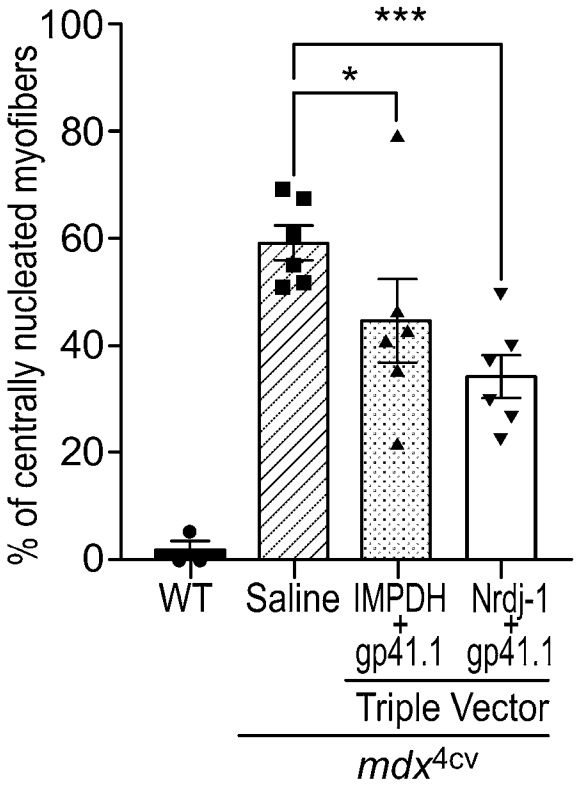

% of centrally nucleated myofibers
100
80
60
40
20
0
***
*
WT
Saline
IMPDH + gp41.1
Nrdj-1 + gp41.1
Triple Vector
mdx4cv

FIG. 22C

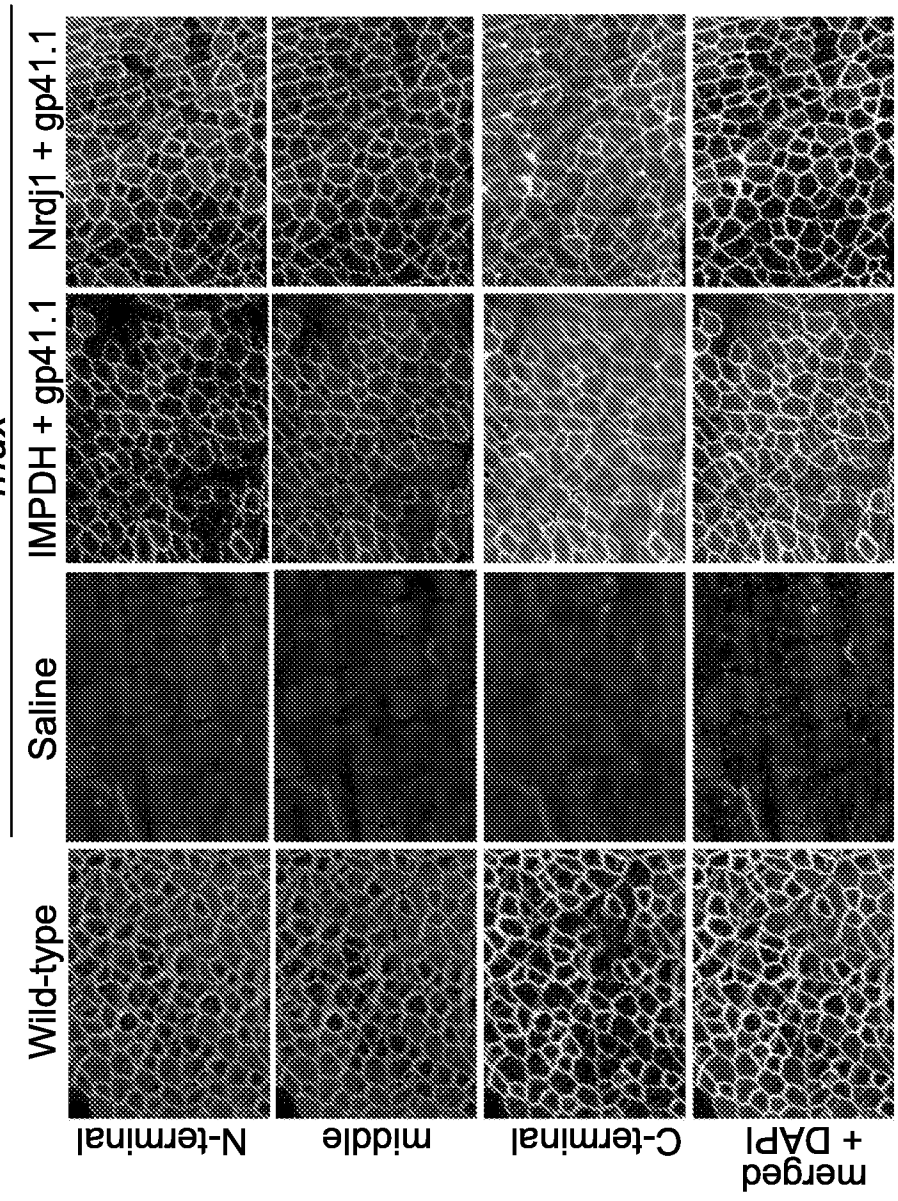

mdx4cv
Wild-type
Saline
IMPDH + gp41.1
Nrdj1 + gp41.1
N-terminal
middle
C-terminal
merged + DAPI

FIG.22D

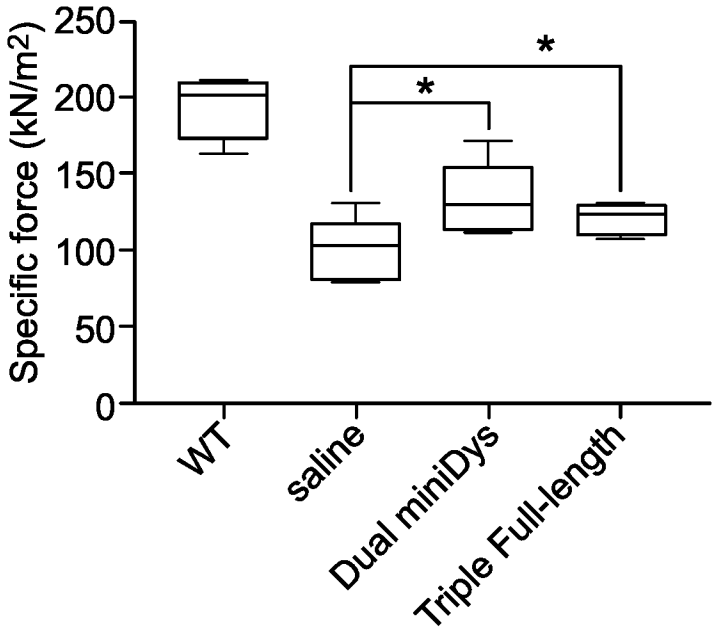


FIG. 23A

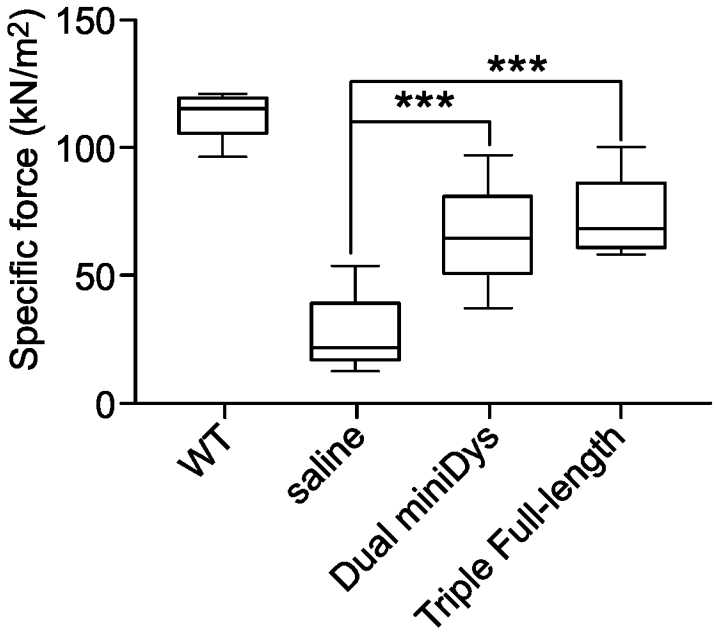


FIG. 23B

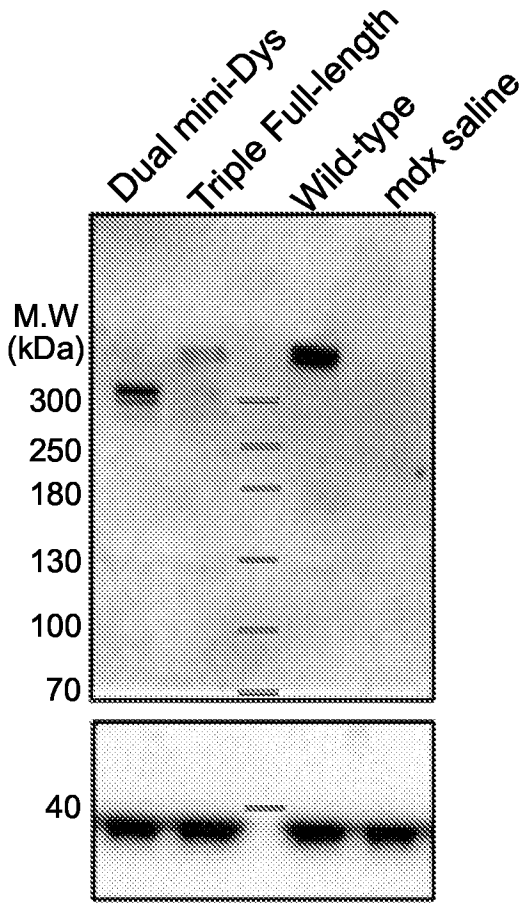

Dual mini-Dys
Triple Full-length
Wild-type
mdx saline
M.W
(kDa)
300
250
180
130
100
70
40

FIG. 23C