

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

(19) World Intellectual Property

Organization

International Bureau

(43) International Publication Date

06 August 2020 (06.08.2020)



(10) International Publication Number

WO 2020/159828 A1

(51) International Patent Classification:

C07K 7/04 (2006.01)

C40B 40/10 (2006.01)

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))
- with sequence listing part of description (Rule 5.2(a))

(21) International Application Number:

PCT/US2020/015033

(22) International Filing Date:

24 January 2020 (24.01.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/797,844

28 January 2019 (28.01.2019)

US

(71) Applicant: THE BOARD OF TRUSTEES OF THE LE-
LAND STANFORD JUNIOR UNIVERSITY [US/US];

Office of the General Counsel, Building 170, 3rd Floor,
Main Quad, P.O. Box 20386, Stanford, California 94305
(US).

(72) Inventors: GREENLEAF, William J.; 279 Campus Drive West, Beckman Center, B-257A/259, Stanford, California 94305 (US). LAYTON, Curtis J.; 279 Campus Drive West, Beckman Center, B-257A/259, Stanford, California 94305 (US).

(74) Agent: KEDDIE, James S.; Bozicevic, Field & Francis LLP, 201 Redwood Shores Parkway, Suite 200, Redwood City, California 94065 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: VARIANT FLAG PEPTIDES AND THEIR USE IN EPITOPE TAGGING

(57) Abstract: Variant FLAG peptides and methods of using them in epitope tagging are disclosed. In particular, the invention relates to variant FLAG peptides that bind with high affinity to the M2 anti-FLAG monoclonal antibody and their use in detection, enrichment, purification, and proteomic profiling of proteins and peptides tagged with the variant FLAG peptides.

WO 2020/159828 A1

VARIANT FLAG PEPTIDES AND THEIR USE IN EPITOPE TAGGING

CROSS-REFERENCING

This application claims the benefit of U.S. provisional application serial no.
5 62/797,844, filed on Jan 28, 2019, which application is incorporated by reference herein for all purposes.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

This invention was made with Government support under contract GM111990
10 awarded by the National Institutes of Health. The Government has certain rights in this invention.

TECHNICAL FIELD

The present invention pertains generally to epitope tagging. In particular, the
15 invention relates to variant FLAG peptides and their use in detection, enrichment, purification, and proteomic profiling of proteins and peptides tagged with the variant FLAG peptides.

BACKGROUND

20 Epitope tagging is commonly used to facilitate detection and purification of recombinant proteins. An epitope for an antibody is fused to a target protein using recombinant DNA techniques. The fusion protein can then be readily detected or purified using an antibody specific for the epitope tag. The epitope can be fused, for example, to the N- or C- terminus of a protein or internally in a loop region. Preferably the epitope is
25 fused to a region of a protein that is accessible to binding by the antibody and that does not disrupt the structure or interfere with the function of the target protein.

One commonly used tag is the FLAG peptide (Brizzard et al. (2001) Curr. Protoc. Neurosci. Chapter 5:Unit 5.8, Johnson et al. (2002) Biotechniques 32(6):1270, 1272-6, 1278, 1280). The FLAG tag is recognized by the anti-FLAG M2 monoclonal antibody,
30 which can be used to detect and purify FLAG-tagged proteins. The FLAG tag has been widely used for detection and purification of proteins e.g., in *E. coli* (Brizzard et al.,

BioTechniques, 16:730-735 (1994)), *Saccharomyces cerevisiae* (Lee et al., Nature, 372:739-746 (1994); Prickett et al., BioTechniques, 7:580-589 (1989)), *Drosophila* (Xu et al., Development, 117:1223-1237 (1993)), Baculovirus (Dent et al., Mol. Cell Biol., 15:4125-4135 (1995); Ritchie et al., Biochem. Journal, 338:305-10 (1999)), and
5 mammalian systems (Overholt et al., Clin. Cancer Res., 3:185-191 (1997); Schulte am Esch et al., Biochemistry, 38:2248-2258 (1999)).

Detection of proteins using current methods of epitope tagging can be problematic when proteins are present at low concentrations. Thus, there remains a need for better methods of epitope tagging to improve the sensitivity of detection and purification of
10 tagged proteins, particularly of expressed proteins present at low abundance.

SUMMARY

The invention relates to variant FLAG peptides having greater affinity than the wild-type FLAG tag for the M2 monoclonal antibody and methods of using such variant
15 FLAG peptides for detection, enrichment, purification, and proteomic profiling of proteins and peptides tagged with the variant FLAG peptides.

In one aspect, the invention includes a variant FLAG peptide comprising an amino acid sequence selected from the group consisting of SEQ ID NOS:3-184 (i.e., a peptide sequence shown in Table 1), or a variant thereof comprising a sequence having at
20 least about 70-100% sequence identity thereto, including any percent identity within this range, such as 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, or 99% sequence identity thereto, wherein the variant FLAG peptide is capable of binding to an anti-FLAG antibody (e.g., M2 monoclonal antibody) with greater affinity than the wild-type FLAG epitope tag (SEQ ID NO:1). In
25 one embodiment, the variant FLAG peptide consists of the sequence of SEQ ID NO:3.

In another aspect, the invention includes a fusion protein comprising a variant FLAG peptide described herein connected to a target protein.

In certain embodiments, the fusion protein further comprises a linker (e.g., located between the variant FLAG peptide and the target protein). Linker amino acid sequences
30 will typically be short, e.g., 20 or fewer amino acids (i.e., 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1). Examples include short peptide sequences which

facilitate cloning, poly-glycine linkers (Gly_n where n = 2, 3, 4, 5, 6, 7, 8, 9, 10 or more), histidine tags (His_n where n = 3, 4, 5, 6, 7, 8, 9, 10 or more), linkers composed of glycine and serine residues, wherein n = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or more), GSAT, SEG, and Z-EGFR linkers. Linkers may include restriction sites, which aid
5 cloning and manipulation.

In certain embodiments, the fusion protein comprises more than one variant FLAG peptide described herein, such as 1, 2, 3, 4, 5, or more variant FLAG peptides.

Fusion proteins can be used in combination with an anti-FLAG antibody (e.g., M2 monoclonal anti-FLAG antibody) in various methods, for example, in Western blotting,
10 immunoprecipitation, or affinity chromatography for detection, identification, enrichment, purification, or proteomic profiling of a target protein tagged with a variant FLAG peptide as described herein.

In another embodiment, the invention includes a method of detecting a fusion protein comprising a variant FLAG peptide described herein connected to a target
15 protein, the method comprising contacting the fusion protein with a detectably labeled anti-FLAG antibody (e.g., M2 monoclonal anti-FLAG antibody), wherein the anti-FLAG antibody specifically binds to the variant FLAG peptide of the fusion protein; and detecting the bound fusion protein. The fusion protein can be detected, for example, by performing a Western, immunofluorescence, or immunoprecipitation. The detectable
20 label can be any molecule or substance capable of detection. In certain embodiments, the detectable label is a fluorescent label, chemiluminescent label, or radioactive label.

In another embodiment, the invention includes a method of isolating a fusion protein described herein, the method comprising immunoprecipitating the fusion protein with an anti-FLAG antibody (e.g., M2 monoclonal anti-FLAG antibody) that specifically
25 binds to the variant FLAG peptide of the fusion protein. In certain embodiments, the anti-FLAG antibody is immobilized on a solid support (e.g., a magnetic bead, nonmagnetic bead, membrane, slide, or plate).

In another embodiment, the invention includes a method of purifying a fusion protein described herein from a mixture using an affinity chromatography matrix
30 comprising an anti-FLAG antibody (e.g., M2 monoclonal anti-FLAG antibody) that specifically binds to the variant FLAG peptide of the fusion protein, the method

comprising: a) contacting the mixture with the affinity chromatography matrix under a first set of conditions, such that the fusion protein binds to the anti-FLAG antibody attached to the affinity chromatography matrix; and b) eluting the fusion protein from the affinity chromatography matrix under a second set of conditions thereby purifying the fusion protein from the mixture. Affinity chromatography may be performed in a column or in batch.

In another aspect, the invention includes a polynucleotide encoding a fusion protein described herein. In one embodiment, the polynucleotide is a recombinant polynucleotide comprising a polynucleotide encoding the fusion protein operably linked to a promoter. In another embodiment, the recombinant polynucleotide comprises a vector such as a viral vector or plasmid.

In another aspect, the invention includes a host cell comprising a recombinant polynucleotide encoding the fusion protein.

In another aspect, the invention includes a method of producing a fusion protein described herein, the method comprising: a) culturing a host cell comprising a recombinant polynucleotide encoding the fusion protein under conditions suitable for the expression of the fusion protein; and b) recovering the fusion protein from the host cell culture using an affinity agent comprising an anti-FLAG antibody.

In another embodiment, the invention includes a vector for producing a fusion protein (i.e., comprising a variant FLAG peptide connected to a target protein), the vector comprising: a) a polynucleotide encoding a variant FLAG peptide described herein; b) a cloning site suitable for in-frame insertion of a polynucleotide encoding a target protein, wherein the cloning site is positioned before or after the polynucleotide encoding the variant FLAG peptide (i.e., to produce a fusion protein comprising the variant FLAG peptide connected at the C-terminus or N-terminus of the target protein, respectively); and c) a promoter operably linked to sequences encoding the fusion protein.

In another aspect, the invention includes a kit for tagging a target protein with a variant FLAG peptide described herein. In certain embodiments, the kit comprises a vector comprising a site for insertion of a polynucleotide encoding the target protein as described herein. Such kits may further comprise an anti-FLAG antibody (e.g., M2 monoclonal anti-FLAG antibody). In certain embodiments, the kit comprises an affinity

chromatography matrix comprising an immobilized anti-FLAG antibody or an anti-FLAG antibody immobilized on some other solid support (e.g., a magnetic bead, nonmagnetic bead, membrane, slide, or plate). Kits may further comprise reagents for performing a Western, immunoprecipitation, immunofluorescence, or affinity chromatography.

Also provided is a method comprising: (a) on a planar substrate (e.g., using an Illumina substrate), sequencing a library of engineered DNA constructs that encode for polypeptides of interest; and (b) performing an *in vitro* transcription/translation on the substrate after (a) such that both the transcript and nascent peptide remain associated with their DNA template, thereby producing a tethered protein array.

These and other embodiments of the subject invention will readily occur to those of skill in the art in view of the disclosure herein.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1. A high-throughput protein array on a DNA sequencing flow cell (A) 1.) A polypeptide-encoding DNA library is clustered and sequenced on a MiSeq, then the chip is transferred to an imaging station where 2.) roadblocked transcription and stalled translation takes place *in vitro*, producing bound polypeptides. A primary antibody and fluorescently-labeled secondary antibody are then bound to the array and imaged. 3.) Fluorescence images are registered to the sequence information, then quantified and fit. (B) A representative flow cell image shows fiducial marks and cluster-of-interest positions before the binding assay (top). Experimental images show fluorescent secondary antibody detection of binding across increasing concentrations of M2 anti-FLAG primary antibody (bottom). (C) Quantified fluorescence medians that rise above a background threshold (grey dashed line) are extrapolated (solid) or interpolated (dashed) to estimate the limit of detection (LoD, open squares) for each FLAG variant, including WT FLAG (DYKDDDDK, SEQ ID NO:1) and negative control (AAADDDDK, SEQ ID NO:2) (top) as well as superFLAG variant (DYKDEDLL, SEQ ID NO:3), which, like 2xFLAG (DYKDDDDKGDYKDHD, SEQ ID NO:4), gives much higher signal than WT (bottom, note scale). Open symbols indicate points below background, closed symbols are above.

FIG. 2. Binding landscape of FLAG peptide variants to M2 monoclonal antibody. (A) 6 amino acid mutations, each representing a physicochemical category, are represented by color. LoD for M2 anti-FLAG antibody binding to (B) single mutants and (C) double mutants of the canonical (WT) peptide, DYKDDDDK (SEQ ID NO:1), across all six amino acids at each position (if the WT identity is the same, the mutation in parenthesis is made). (D) LoD (bottom left) and cooperativity (top right) of double mutants of the D4L base mutant (triple mutants from the WT). Cooperativity = $-\log(\text{LoD}_{\text{DM}}/\text{LoD}_{\text{WT}}) - (-\log(\text{LoD}_{\text{SM1}}/\text{LoD}_{\text{WT}}) + -\log(\text{LoD}_{\text{SM2}}/\text{LoD}_{\text{WT}}))$, where DM refers to the double mutant and SM1 and SM2 to the two corresponding single mutants. (E) Orthogonal investigation of individual peptide variants with a fluorescence-based plate assay. Antibodies bound to variant peptides (see legend) were challenged for 2 hours with varying concentrations of 3xFLAG competitor peptide. Fluorescence values report residual bound M2 antibody.

FIG. 3. High-throughput, on-chip characterization of activity across the mutational landscape of a full-length enzyme. (A) On-chip SNAP-tag-catalyzed fluorescent self-labeling (B) 7 target amino acid positions (red) and benzylguanine substrate (green) on the structure of SNAP-tag (PDB ID: 3KZZ) (C) Library coverage for single, double, and triple mutants of SNAP-tag enzyme. (D) Quantified fluorescence medians for selected single mutants. Variants that rose above the background threshold (grey dashed) were fit with a single exponential. (E) Fit k_{obs} values for single mutants and (F) double mutants across the 20 amino acids. WT (unmutated) variants are boxed in black. (G) Position 153 is in a loop connecting two domains. (H) Correlation of activity with backbone flexibility [33] for all 19 amino acid substitutions at position 153. $R^2=0.43$ (p-value = 1.6×10^{-3} by permutations). Proline is the least flexible and glycine the most. (I) Comparison of double mutant effects on activity (y-axis) with the sum of single mutant effects (x-axis). Only a few residue position pairs (orange) demonstrate high positive cooperativity ("highly cooperative" refers hereafter to cooperativity < -1.69 (see 3K) where cooperativity = $-\log(k_{\text{obs,DM}}/k_{\text{obs,WT}}) - (-\log(k_{\text{obs,SM1}}/k_{\text{obs,WT}}) + -\log(k_{\text{obs,SM2}}/k_{\text{obs,WT}}))$). (J) Cumulative number of pairs with C α -C α distance within a given distance for highly cooperative (orange), and all target residue pairs (blue). Most highly cooperative interactions occur within 13Å (grey dashed) (K) cooperativity values for all

target residue pairs with C α -C α distance < 13Å (orange) and > 13Å (blue) (top) and by amino acid (counted for either or both partners) (bottom). Sum of highly cooperative pairs (grey dashed) per amino acid (left). (I) Number of highly cooperative interactions across all amino acid pairs.

FIG. 4. DNA constructs for protein array generation—strategies for mutational library generation and PCR assembly for (A) FLAG and (B) SNAP-tag studies

FIG. 5. CJL-PPP is the engineered ribosomal stall sequence for E. coli ribosomes developed and used in this study. To generate CJL-PPP, the PPP motif was added with maximally stalling upstream context residues (purple) [38]. Critical residues from SecM [40] and TnaC [41] (red) were transplanted where possible. SEQ ID NOS 186, 186, and 187. From top to bottom, SEQ ID NOS:185-190:

FIG. 6. F_{max} threshold for fitting quality control (QC). SNAP-tag variants were discarded from further analysis if the F_{max} parameter determined by an unconstrained nonlinear least squares fit was above threshold F_{max} (orange dashed; see methods).

DETAILED DESCRIPTION

The practice of the present invention will employ, unless otherwise indicated, conventional methods of proteomics, chemistry, immunology, biochemistry, molecular biology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, e.g., A.K. Mallia, P.K. Smith, G.T. Hermanson *Immobilized Affinity Ligand Techniques* (Academic Press; 1st edition, 1992); *Affinity Chromatography: Methods and Protocols* (Methods in Molecular Biology, P. Bailon ed., Humana Press; 1st edition, 2000); *Separation Methods In Proteomics* (G.B. Smejkal and A. Lazarev eds., CRC Press; 1st edition, 2005); A.L. Lehninger, *Biochemistry* (Worth Publishers, Inc., current addition); Sambrook, et al., *Molecular Cloning: A Laboratory Manual* (3rd Edition, 2001); *Handbook of Experimental Immunology*, Vols. I-IV (D.M. Weir and C.C. Blackwell eds., Blackwell Scientific Publications); *Methods In Enzymology* (S. Colowick and N. Kaplan eds., Academic Press, Inc.).

All publications, patents and patent applications cited herein, whether *supra* or *infra*, are hereby incorporated by reference in their entireties.

I. DEFINITIONS

In describing the present invention, the following terms will be employed, and are intended to be defined as indicated below.

It must be noted that, as used in this specification and the appended claims, the singular forms "a", "an" and "the" include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to "a peptide" includes a mixture of two or more peptides, and the like.

The term "about", particularly in reference to a given quantity, is meant to encompass deviations of plus or minus five percent.

The terms "polypeptide," "peptide" and "protein" are used interchangeably to refer to a polymer of amino acid residues. The terms also apply to amino acid polymers in which one or more amino acids are chemical analogues or modified derivatives of a corresponding naturally-occurring amino acids. Amino acid derivatives can include modifications to the native sequence, such as deletions, additions and substitutions (generally conservative in nature), so long as the polypeptide or peptide maintains the desired activity (e.g., binds to anti-FLAG antibody). These modifications may be deliberate, as through site-directed mutagenesis, or may be accidental, such as through mutations of hosts that produce the proteins or errors due to PCR amplification. Furthermore, modifications may be made that have one or more of the following effects: increasing affinity and/or specificity for an anti-FLAG antibody. Polypeptides described herein can be made recombinantly, synthetically, or in tissue culture.

The terms "fusion protein" and "fusion polypeptide," as used herein refer to a fusion comprising at least one variant FLAG peptide in combination with a selected target protein of interest as part of a single continuous chain of amino acids, which chain does not occur in nature. The variant FLAG peptide and target protein may be connected directly to each other by peptide bonds or may be separated by an intervening amino acid sequence (i.e., linker). The fusion protein may comprises one or more variant FLAG peptides (e.g., 1, 2, 3, 4, 5, or more variant FLAG peptides). The target protein in the fusion may include an entire protein or a fragment thereof (e.g., a biologically active domain) of a protein of interest. The fusion protein may also contain sequences exogenous to the variant FLAG peptide and target protein. For example, the fusion

protein may include targeting or localization sequences, tag sequences, or detectable labels.

"Substantially purified" generally refers to isolation of a substance (compound, polynucleotide, protein, polypeptide, peptide composition) such that the substance comprises the majority percent of the sample in which it resides. Typically in a sample, a substantially purified component comprises 50%, preferably 80%-85%, more preferably 90-95% of the sample. Techniques for purifying polynucleotides and polypeptides of interest are well-known in the art and include, for example, ion-exchange chromatography, affinity chromatography and sedimentation according to density.

By "isolated" is meant, when referring to a polypeptide or peptide, that the indicated molecule is separate and discrete from the whole organism with which the molecule is found in nature or is present in the substantial absence of other biological macro molecules of the same type. The term "isolated" with respect to a polynucleotide is a nucleic acid molecule devoid, in whole or part, of sequences normally associated with it in nature; or a sequence, as it exists in nature, but having heterologous sequences in association therewith; or a molecule disassociated from the chromosome.

The terms "label" and "detectable label" refer to a molecule or substance capable of detection, including, but not limited to, fluorescers, chemiluminescers, chromophores, bioluminescent proteins, enzymes, enzyme substrates, enzyme cofactors, enzyme inhibitors, isotopic labels, semiconductor nanoparticles, dyes, metal ions, metal sols, ligands (e.g., biotin, streptavidin or haptens) and the like. The term "fluorescer" refers to a substance or a portion thereof which is capable of exhibiting fluorescence in the detectable range. Particular examples of labels which may be used in the practice of the invention include, but are not limited to, SYBR green, SYBR gold, a CAL Fluor dye such as CAL Fluor Gold 540, CAL Fluor Orange 560, CAL Fluor Red 590, CAL Fluor Red 610, and CAL Fluor Red 635, a Quasar dye such as Quasar 570, Quasar 670, and Quasar 705, an Alexa Fluor such as Alexa Fluor 350, Alexa Fluor 488, Alexa Fluor 546, Alexa Fluor 555, Alexa Fluor 594, Alexa Fluor 647, and Alexa Fluor 784, a cyanine dye such as Cy 3, Cy3.5, Cy5, Cy5.5, and Cy7, fluorescein, 2', 4', 5', 7'-tetrachloro-4-7-dichlorofluorescein (TET), carboxyfluorescein (FAM), 6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein (JOE), hexachlorofluorescein (HEX), rhodamine, carboxy-X-

rhodamine (ROX), tetramethyl rhodamine (TAMRA), FITC, dansyl, umbelliferone, dimethyl acridinium ester (DMAE), Texas red, luminol, and quantum dots, enzymes such as alkaline phosphatase (AP), beta-lactamase, chloramphenicol acetyltransferase (CAT), adenosine deaminase (ADA), aminoglycoside phosphotransferase (neo^r, G418^r)

5 dihydrofolate reductase (DHFR), hygromycin-B-phosphotransferase (HPH), thymidine kinase (TK), β -galactosidase (lacZ), and xanthine guanine phosphoribosyltransferase (XGPRT), beta-glucuronidase (gus), placental alkaline phosphatase (PLAP), and secreted embryonic alkaline phosphatase (SEAP). Enzyme tags are used with their cognate substrate. The terms also include chemiluminescent labels such as luminol, isoluminol,

10 acridinium esters, and peroxyoxalate and bioluminescent proteins such as firefly luciferase, bacterial luciferase, Renilla luciferase, and aequorin. The terms also include isotopic labels, including radioactive and non-radioactive isotopes, such as, ³H, ²H, ¹²⁰I, ¹²³I, ¹²⁴I, ¹²⁵I, ¹³¹I, ³⁵S, ¹¹C, ¹³C, ¹⁴C, ³²P, ¹⁵N, ¹³N, ¹¹⁰In, ¹¹¹In, ¹⁷⁷Lu, ¹⁸F, ⁵²Fe, ⁶²Cu, ⁶⁴Cu, ⁶⁷Cu, ⁶⁷Ga, ⁶⁸Ga, ⁸⁶Y, ⁹⁰Y, ⁸⁹Zr, ^{94m}Tc, ⁹⁴Tc, ^{99m}Tc, ¹⁵⁴Gd, ¹⁵⁵Gd, ¹⁵⁶Gd, ¹⁵⁷Gd, ¹⁵⁸Gd, ¹⁵O,

15 ¹⁸⁶Re, ¹⁸⁸Re, ⁵¹M, ^{52m}Mn, ⁵⁵Co, ⁷²As, ⁷⁵Br, ⁷⁶Br, ^{82m}Rb, and ⁸³Sr. The terms also include color-coded microspheres of known fluorescent light intensities (see e.g., microspheres with xMAP technology produced by Luminex (Austin, TX); microspheres containing quantum dot nanocrystals, for example, containing different ratios and combinations of quantum dot colors (e.g., Qdot nanocrystals produced by Life Technologies (Carlsbad,

20 CA); glass coated metal nanoparticles (see e.g., SERS nanotags produced by Nanoplex Technologies, Inc. (Mountain View, CA); barcode materials (see e.g., sub-micron sized striped metallic rods such as Nanobarcodes produced by Nanoplex Technologies, Inc.), encoded microparticles with colored bar codes (see e.g., CellCard produced by Vitra Bioscience, vitrabio.com), glass microparticles with digital holographic code images (see

25 e.g., CyVera microbeads produced by Illumina (San Diego, CA), near infrared (NIR) probes, and nanoshells. The terms also include contrast agents such as ultrasound contrast agents (e.g. SonoVue microbubbles comprising sulfur hexafluoride, Optison microbubbles comprising an albumin shell and octafluoropropane gas core, Levovist microbubbles comprising a lipid/galactose shell and an air core, Perflexane lipid

30 microspheres comprising perfluorocarbon microbubbles, and Perflutren lipid microspheres comprising octafluoropropane encapsulated in an outer lipid shell),

magnetic resonance imaging (MRI) contrast agents (e.g., gadodiamide, gadobenamic acid, gadopentetic acid, gadoteridol, gadofosveset, gadoversetamide, gadoxetic acid), and radiocontrast agents, such as for computed tomography (CT), radiography, or fluoroscopy (e.g., diatrizoic acid, metrizoic acid, iodamide, iotalamic acid, ioxitalamic acid, ioglicic acid, acetrizoic acid, iocarmic acid, methiodal, diodone, metrizamide, 5 iohexol, ioxaglic acid, iopamidol, iopromide, iotrolan, ioversol, iopentol, iodixanol, iomeprol, iobitridol, ioxilan, iodoxamic acid, iotroxic acid, ioglycamic acid, adipiodone, iobenzamic acid, iopanoic acid, iocetamic acid, sodium iopodate, tyropanoic acid, and calcium iopodate). As with many of the standard procedures associated with the practice 10 of the invention, skilled artisans will be aware of additional labels that can be used.

The phrase "specifically (or selectively) binds" to an antibody or "specifically (or selectively) immunoreactive with," when referring to a protein or peptide fused to an epitope tag (e.g., variant FLAG tag), refers to a binding reaction that is determinative of the presence of the tagged protein or peptide in a heterogeneous population of proteins 15 and other biologics. Thus, under designated immunoassay conditions, the specified antibodies bind to tagged proteins or peptides carrying a particular epitope tag at least two times the background and do not substantially bind in a significant amount to untagged proteins present in the sample. Specific binding to an antibody under such conditions may require an antibody that is selected for its specificity for a particular 20 epitope tag. For example, polyclonal antibodies raised to an epitope tag can be selected to obtain only those polyclonal antibodies that are specifically immunoreactive with the epitope tag and not with other peptides. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular peptide or epitope tag. For example, solid-phase ELISA immunoassays are routinely used to select antibodies 25 specifically immunoreactive with a protein (see, e.g., Harlow & Lane. Antibodies, A Laboratory Manual (1988), for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity). Typically a specific or selective reaction will be at least twice background signal or noise and more typically more than 10 to 100 times background.

30 "Homology" refers to the percent identity between two polynucleotide or two polypeptide molecules. Two nucleic acid, or two peptide sequences are "substantially

homologous" to each other when the sequences exhibit at least about 50% sequence identity, preferably at least about 75% sequence identity, more preferably at least about 80% 85% sequence identity, more preferably at least about 90% sequence identity, and most preferably at least about 95% 98% sequence identity over a defined length of the molecules. As used herein, substantially homologous also refers to sequences showing complete identity to the specified sequence.

In general, "identity" refers to an exact nucleotide to nucleotide or amino acid to amino acid correspondence of two polynucleotides or peptide sequences, respectively. Percent identity can be determined by a direct comparison of the sequence information between two molecules by aligning the sequences, counting the exact number of matches between the two aligned sequences, dividing by the length of the shorter sequence, and multiplying the result by 100. Readily available computer programs can be used to aid in the analysis, such as ALIGN, Dayhoff, M.O. in Atlas of Protein Sequence and Structure M.O. Dayhoff ed., 5 Suppl. 3:353 358, National biomedical Research Foundation, Washington, DC, which adapts the local homology algorithm of Smith and Waterman Advances in Appl. Math. 2:482 489, 1981 for peptide analysis. Programs for determining nucleotide sequence identity are available in the Wisconsin Sequence Analysis Package, Version 8 (available from Genetics Computer Group, Madison, WI) for example, the BESTFIT, FASTA and GAP programs, which also rely on the Smith and Waterman algorithm. These programs are readily utilized with the default parameters recommended by the manufacturer and described in the Wisconsin Sequence Analysis Package referred to above. For example, percent identity of a particular nucleotide sequence to a reference sequence can be determined using the homology algorithm of Smith and Waterman with a default scoring table and a gap penalty of six nucleotide positions.

Another method of establishing percent identity in the context of the present invention is to use the MPSRCH package of programs copyrighted by the University of Edinburgh, developed by John F. Collins and Shane S. Sturrok, and distributed by IntelliGenetics, Inc. (Mountain View, CA). From this suite of packages the Smith Waterman algorithm can be employed where default parameters are used for the scoring table (for example, gap open penalty of 12, gap extension penalty of one, and a gap of

six). From the data generated the "Match" value reflects "sequence identity." Other suitable programs for calculating the percent identity or similarity between sequences are generally known in the art, for example, another alignment program is BLAST, used with default parameters. For example, BLASTN and BLASTP can be used using the
5 following default parameters: genetic code = standard; filter = none; strand = both; cutoff = 60; expect = 10; Matrix = BLOSUM62; Descriptions = 50 sequences; sort by = HIGH SCORE; Databases = non redundant, GenBank + EMBL + DDBJ + PDB + GenBank CDS translations + Swiss protein + Spupdate + PIR. Details of these programs are readily available.

10 Alternatively, homology can be determined by hybridization of polynucleotides under conditions which form stable duplexes between homologous regions, followed by digestion with single stranded specific nuclease(s), and size determination of the digested fragments. DNA sequences that are substantially homologous can be identified in a Southern hybridization experiment under, for example, stringent conditions, as defined
15 for that particular system. Defining appropriate hybridization conditions is within the skill of the art. See, e.g., Sambrook et al., *supra*; DNA Cloning, *supra*; Nucleic Acid Hybridization, *supra*.

"Recombinant" as used herein to describe a nucleic acid molecule means a polynucleotide of genomic, cDNA, viral, semisynthetic, or synthetic origin which, by
20 virtue of its origin or manipulation, is not associated with all or a portion of the polynucleotide with which it is associated in nature. The term "recombinant" as used with respect to a protein (e.g., fusion protein comprising a target protein fused to an epitope tag) or peptide (e.g., epitope tag) means a protein or peptide produced by expression of a recombinant polynucleotide. In general, the gene of interest is cloned and
25 then expressed in transformed organisms, as described further below. The host organism expresses the foreign gene to produce the protein under expression conditions.

The term "transformation" refers to the insertion of an exogenous polynucleotide into a host cell, irrespective of the method used for the insertion. For example, direct uptake, transduction or f-mating are included. The exogenous polynucleotide may be
30 maintained as a non-integrated vector, for example, a plasmid, or alternatively, may be integrated into the host genome.

"Recombinant host cells," "host cells," "cells," "cell lines," "cell cultures," and other such terms denoting microorganisms or higher eukaryotic cell lines cultured as unicellular entities refer to cells which can be, or have been, used as recipients for recombinant vector or other transferred DNA, and include the original progeny of the original cell which has been transfected.

A "coding sequence" or a sequence which "encodes" a selected protein or peptide, is a nucleic acid molecule which is transcribed (in the case of DNA) and translated (in the case of mRNA) into a protein or peptide *in vivo* when placed under the control of appropriate regulatory sequences (or "control elements"). The boundaries of the coding sequence can be determined by a start codon at the 5' (amino) terminus and a translation stop codon at the 3' (carboxy) terminus. A coding sequence can include, but is not limited to, cDNA from viral, prokaryotic or eukaryotic mRNA, genomic DNA sequences from viral or prokaryotic DNA, and even synthetic DNA sequences. A transcription termination sequence may be located 3' to the coding sequence.

Typical "control elements" include, but are not limited to, transcription promoters, transcription enhancer elements, transcription termination signals, polyadenylation sequences (located 3' to the translation stop codon), sequences for optimization of initiation of translation (located 5' to the coding sequence), and translation termination sequences.

"Operably linked" refers to an arrangement of elements wherein the components so described are configured so as to perform their usual function. Thus, a given promoter operably linked to a coding sequence is capable of effecting the expression of the coding sequence when the proper enzymes are present. The promoter need not be contiguous with the coding sequence, so long as it functions to direct the expression thereof. Thus, for example, intervening untranslated yet transcribed sequences can be present between the promoter sequence and the coding sequence and the promoter sequence can still be considered "operably linked" to the coding sequence.

"Encoded by" refers to a nucleic acid sequence which codes for a polypeptide sequence, wherein the polypeptide sequence or a portion thereof contains an amino acid sequence of at least 3 to 5 amino acids, more preferably at least 8 to 10 amino acids, and

even more preferably at least 15 to 20 amino acids from a polypeptide encoded by the nucleic acid sequence.

"Expression cassette" or "expression construct" refers to an assembly which is capable of directing the expression of the sequence(s) or gene(s) of interest. An expression cassette generally includes control elements, as described above, such as a promoter which is operably linked to (so as to direct transcription of) the sequence(s) or gene(s) of interest, and often includes a polyadenylation sequence as well. Within certain embodiments of the invention, the expression cassette described herein may be contained within a plasmid construct. In addition to the components of the expression cassette, the plasmid construct may also include, one or more selectable markers, a signal which allows the plasmid construct to exist as single stranded DNA (e.g., a M13 origin of replication), at least one multiple cloning site, and a "mammalian" origin of replication (e.g., a SV40 or adenovirus origin of replication).

"Purified polynucleotide" refers to a polynucleotide of interest or fragment thereof which is essentially free, e.g., contains less than about 50%, preferably less than about 70%, and more preferably less than about at least 90%, of the protein with which the polynucleotide is naturally associated. Techniques for purifying polynucleotides of interest are well-known in the art and include, for example, disruption of the cell containing the polynucleotide with a chaotropic agent and separation of the polynucleotide(s) and proteins by ion-exchange chromatography, affinity chromatography and sedimentation according to density.

The term "transfection" is used to refer to the uptake of foreign DNA by a cell. A cell has been "transfected" when exogenous DNA has been introduced inside the cell membrane. A number of transfection techniques are generally known in the art. See, e.g., Graham et al. (1973) Virology, 52:456, Sambrook et al. (2001) Molecular Cloning, a laboratory manual, 3rd edition, Cold Spring Harbor Laboratories, New York, Davis et al. (1995) Basic Methods in Molecular Biology, 2nd edition, McGraw-Hill, and Chu et al. (1981) Gene 13:197. Such techniques can be used to introduce one or more exogenous DNA moieties into suitable host cells. The term refers to both stable and transient uptake of the genetic material, and includes uptake of peptide- or antibody-linked DNAs.

A "vector" is capable of transferring nucleic acid sequences to target cells (e.g., viral vectors, non-viral vectors, particulate carriers, and liposomes). Typically, "vector construct," "expression vector," and "gene transfer vector," mean any nucleic acid construct capable of directing the expression of a nucleic acid of interest and which can transfer nucleic acid sequences to target cells. Thus, the term includes cloning and expression vehicles, as well as viral vectors.

The terms "variant," "analog" and "mutein" refer to biologically active derivatives of the reference molecule that retain desired activity, such as the ability to bind to an anti-FLAG antibody. In general, the terms "variant" and "analog" refer to compounds having a native peptide sequence and structure with one or more amino acid additions, substitutions (generally conservative in nature) and/or deletions, relative to the native molecule, so long as the modifications do not destroy biological activity and which are "substantially homologous" to the reference molecule as defined below. In general, the amino acid sequences of such analogs will have a high degree of sequence homology to the reference sequence, e.g., amino acid sequence homology of more than 50%, generally more than 60%-70%, even more particularly 80%-85% or more, such as at least 90%-95% or more, when the two sequences are aligned. Often, the analogs will include the same number of amino acids but will include substitutions, as explained herein. The term "mutein" further includes polypeptides having one or more amino acid-like molecules including but not limited to compounds comprising only amino and/or imino molecules, polypeptides containing one or more analogs of an amino acid (including, for example, unnatural amino acids, etc.), peptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring (e.g., synthetic), cyclized, branched molecules and the like. The term also includes molecules comprising one or more N-substituted glycine residues (a "peptoid") and other synthetic amino acids or peptides. (See, e.g., U.S. Patent Nos. 5,831,005; 5,877,278; and 5,977,301; Nguyen et al., Chem. Biol. (2000) 7:463-473; and Simon et al., Proc. Natl. Acad. Sci. USA (1992) 89:9367-9371 for descriptions of peptoids). Methods for making peptide analogs and muteins are known in the art and are described further below.

As explained above, analogs generally include substitutions that are conservative in nature, i.e., those substitutions that take place within a family of amino acids that are

related in their side chains. Specifically, amino acids are generally divided into four families: (1) acidic -- aspartate and glutamate; (2) basic -- lysine, arginine, histidine; (3) non-polar -- alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan; and (4) uncharged polar -- glycine, asparagine, glutamine, cysteine, serine, threonine, and tyrosine. Phenylalanine, tryptophan, and tyrosine are sometimes classified as aromatic amino acids. For example, it is reasonably predictable that an isolated replacement of leucine with isoleucine or valine, an aspartate with a glutamate, a threonine with a serine, or a similar conservative replacement of an amino acid with a structurally related amino acid, will not have a major effect on the biological activity. For example, the polypeptide of interest may include up to about 5-10 conservative or non-conservative amino acid substitutions, or even up to about 15-25 conservative or non-conservative amino acid substitutions, or any integer between 5-25, so long as the desired function of the molecule remains intact. One of skill in the art may readily determine regions of the molecule of interest that can tolerate change by reference to Hopp/Woods and Kyte-Doolittle plots, well known in the art.

The term "derived from" is used herein to identify the original source of a molecule but is not meant to limit the method by which the molecule is made which can be, for example, by chemical synthesis or recombinant means.

A polynucleotide "derived from" a designated sequence refers to a polynucleotide sequence which comprises a contiguous sequence of approximately at least about 6 nucleotides, preferably at least about 8 nucleotides, more preferably at least about 10-12 nucleotides, and even more preferably at least about 15-20 nucleotides corresponding, i.e., identical or complementary to, a region of the designated nucleotide sequence. The derived polynucleotide will not necessarily be derived physically from the nucleotide sequence of interest, but may be generated in any manner, including, but not limited to, chemical synthesis, replication, reverse transcription or transcription, which is based on the information provided by the sequence of bases in the region(s) from which the polynucleotide is derived. As such, it may represent either a sense or an antisense orientation of the original polynucleotide.

II. Modes of Carrying Out the Invention

Before describing the present invention in detail, it is to be understood that this invention is not limited to particular formulations or process parameters as such may, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments of the invention only, and is not intended to be limiting.

Although a number of methods and materials similar or equivalent to those described herein can be used in the practice of the present invention, the preferred materials and methods are described herein.

The present invention relates to variant FLAG peptides having greater affinity than the wild-type FLAG epitope tag for the M2 monoclonal antibody and methods of using such variant FLAG peptides for detection, enrichment, purification, and proteomic profiling of proteins and peptides tagged with the variant FLAG peptides. In order to further an understanding of the invention, a more detailed discussion is provided below regarding the variant FLAG peptides and methods of using them in detection, purification, and proteomic profiling of tagged proteins and peptides.

Embodiments

Embodiment 1. A variant FLAG peptide comprising an amino acid sequence selected from the group consisting of:

a) an amino acid sequence selected from the group consisting of SEQ ID NOS:3-184; and

b) an amino acid sequence having at least 80% identity to an amino acid sequence selected from the group consisting of SEQ ID NOS:3-184, wherein the variant FLAG peptide is capable of binding to an anti-FLAG antibody with greater affinity than a FLAG peptide comprising the amino acid sequence of SEQ ID NO:1.

Embodiment 2. The variant FLAG peptide of embodiment 1, wherein the variant FLAG peptide consists of the sequence of SEQ ID NO:3.

Embodiment 3. A fusion protein comprising one or more variant FLAG peptides of any prior embodiment connected to a target protein.

Embodiment 4. The fusion protein of embodiment 3, further comprising a linker.

Embodiment 5. The fusion protein of embodiment 3 or 4, wherein the fusion protein comprises 1-5 variant FLAG peptides.

Embodiment 6. A method of detecting the fusion protein of any of embodiments 3-5, the method comprising:

- a) contacting the fusion protein with a detectably labeled anti-FLAG antibody, wherein the anti-FLAG antibody specifically binds to the variant FLAG peptide of the fusion protein; and
- b) detecting the bound fusion protein.

Embodiment 7. The method of embodiment 6, wherein said detecting comprises performing a Western, immunofluorescence, or immunoprecipitation.

Embodiment 8. The method of embodiment 6 or 7, wherein the detectably labeled anti-FLAG antibody comprises a fluorescent label, a chemiluminescent label, or a radioactive label.

Embodiment 9. The method of any of embodiments 6-8, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.

Embodiment 10. A method of isolating the fusion protein described of any of embodiments 6-9, the method comprising immunoprecipitating the fusion protein with an anti-FLAG antibody that specifically binds to the variant FLAG peptide of the fusion protein.

Embodiment 11. The method of embodiment 10, wherein the anti-FLAG antibody is immobilized on a solid support.

Embodiment 12. The method of embodiment 11, wherein the solid support is a magnetic bead or a nonmagnetic bead.

Embodiment 13. The method of any of embodiments 10-12, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.

Embodiment 14. A method of purifying the fusion protein of any of embodiments 6-9 from a mixture using an affinity chromatography matrix comprising an anti-FLAG antibody, the method comprising:

- a) contacting the mixture with the affinity chromatography matrix under a first set of conditions, such that the fusion protein binds to the anti-FLAG antibody attached to the

affinity chromatography matrix; and b) eluting the fusion protein from the affinity chromatography matrix under a second set of conditions thereby purifying the fusion protein from the mixture.

wherein the affinity chromatography may be performed in a column or in batch.

5 Embodiment 15. The affinity reagent of embodiment 14, wherein the solid support is selected from the group consisting of a bead, a particle, a rod, a membrane, a gel, a resin, a column, a chip, a slide, a plate, and a microarray.

 Embodiment 16. The affinity reagent of embodiment 15, wherein the bead is an agarose bead.

10 Embodiment 17. The affinity reagent of embodiment 15, wherein the bead is a magnetic bead or a nonmagnetic bead.

 Embodiment 18. The affinity reagent of embodiment 15, wherein the bead is a microbead.

 Embodiment 19. The affinity reagent of embodiment 15, wherein the plate is a
15 microtiter plate.

 Embodiment 20. The affinity reagent of embodiment 15, wherein the membrane is a nitrocellulose membrane or a polyvinylidene difluoride (PVDF) membrane.

 Embodiment 21. The method of any of embodiments 14-20, wherein the anti-FLAG antibody specifically binds to the variant FLAG peptide of the fusion protein.

20 Embodiment 22. The method of any of embodiments 14-21, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.

 Embodiment 23. A polynucleotide encoding the fusion protein of any of embodiments 3-9.

 Embodiment 24. A recombinant polynucleotide comprising the polynucleotide of
25 embodiment 23 operably linked to a promoter.

 Embodiment 25. A host cell comprising the recombinant polynucleotide of embodiment 24.

 Embodiment 26. A method for producing a fusion protein, the method comprising the steps of:

30 a) culturing a host cell comprising the recombinant polynucleotide of embodiment 24 under conditions suitable for the expression of the fusion protein; and

b) recovering the fusion protein from the host cell culture using an affinity agent comprising an anti-FLAG antibody.

Embodiment 27. A vector for producing a tagged target protein, the vector comprising:

- a) a polynucleotide encoding the variant FLAG peptide of embodiment 1;
- b) a cloning site suitable for in-frame insertion of a polynucleotide encoding the target protein, wherein the cloning site is positioned before or after the polynucleotide encoding the variant FLAG peptide; and
- c) a promoter operably linked to sequences encoding the fusion protein.

Embodiment 28. The vector of embodiment 27, wherein the cloning site comprises one or more restriction sites.

Embodiment 29. The vector of embodiment 27, further comprising a polynucleotide encoding a linker, wherein said polynucleotide encoding the linker is positioned in between the cloning site and the polynucleotide encoding the variant FLAG peptide.

Embodiment 30. The vector of embodiment 27, further comprising a selectable marker.

Embodiment 31. A kit for tagging a target protein with a variant FLAG peptide, the kit comprising the vector of embodiment 27 and instructions for preparing and detecting the fusion protein.

Embodiment 32. The kit of embodiment 31, further comprising an anti-FLAG antibody.

Embodiment 33. The kit of embodiment 32, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.

Embodiment 34. The kit of embodiment 32, wherein the anti-FLAG antibody is immobilized on a solid support.

Embodiment 35. The kit of embodiment 34, wherein the solid support is selected from the group consisting of a bead, a particle, a rod, a membrane, a gel, a resin, a column, a chip, a slide, a plate, and a microarray.

Embodiment 36. The kit of embodiment 35, wherein the bead is a magnetic bead or a nonmagnetic bead.

Embodiment 37. The kit of embodiment 35, wherein the plate is a microtiter plate.

Embodiment 38. The kit of embodiment 35, wherein the membrane is a nitrocellulose membrane or a polyvinylidene difluoride (PVDF) membrane.

Embodiment 39. The kit of embodiment 31, further comprising reagents for performing a Western, immunoprecipitation, immunofluorescence, or affinity chromatography.

Embodiment 40. A screening method comprising:

(a) on a planar substrate (e.g., using an Illumina substrate), sequencing a library of engineered DNA constructs that encode for polypeptides of interest; and

(b) performing an in vitro transcription/translation on the substrate after (a) such that both the transcript and nascent peptide remain associated with their DNA template, thereby producing a tethered protein array.

Embodiment 41. The method of embodiment 40, wherein each member of the sequencing library contains DNA sequence elements that allow for 1) prokaryotic in vitro transcription, 2) immobilization of the resulting RNA transcript, 3) translation, and 4) ribosome stall.

Embodiment 42. The method of embodiments 40-41, further comprising performing a fluorescence-based functional assay (e.g. quantifying binding of fluorescently-labeled binding partners or incorporation of fluorescent substrates) tethered protein array on the substrate.

Embodiment 43. The method of embodiments 42, further comprising reading the substrate.

Embodiment 44. The method of embodiment 43, further comprising matching the results to the sequence of the engineered DNA constructs.

III. Experimental

Below are examples of specific embodiments for carrying out the present invention. The examples are offered for illustrative purposes only, and are not intended to limit the scope of the present invention in any way.

Efforts have been made to ensure accuracy with respect to numbers used (e.g., amounts, temperatures, etc.), but some experimental error and deviation should, of course, be allowed for.

5

Example 1

Large-Scale, Quantitative Protein Assays on a High-Throughput DNA Sequencing Chip

10

High-throughput DNA sequencing technologies have enabled the investigation of diverse biological processes wherein the functional consequences of nucleic acid variation can be linked directly to the abundance and sequence of DNA fragments that are quantified at scale. These applications (e.g. ChIP-seq, ATAC-seq, bisulfite sequencing, Hi-C, bind-n-seq, etc. [1-5]) largely define the contemporary methodological foundations of modern functional genomics. In contrast, methods for directly assaying the influence of *protein* sequence variation on function has remained challenging to similarly scale and disseminate [6]. Approaches for high-throughput protein functional measurements have included the quantification of selective enrichment [7-9] of protein particles linked to their encoding nucleic-acids [10, 11] and parallelized binding assays on protein and peptide microarrays [12], including arrays generated with *in vitro* protein translation [13-15] and *in situ* peptide synthesis [16-18]. However, no existing implementation has provided the scalability, simplicity, automation, or accessibility sufficient for widespread application. Furthermore, these approaches often do not allow for a direct and quantitative assay of protein function, restricting our ability to develop and test a predictive understanding of the functional impact of coding mutations, or identify and characterize amino acid interaction networks and dependencies. These limitations likewise hamper protein engineering efforts, where facile high-throughput methods for assaying diverse protein functions could provide large-scale data sets for learning new principles and paradigms for rational design of protein function.

30

To enable facile, widely deployable, high-throughput, and quantitative protein characterization, we sought to leverage the capabilities and widespread adoption of the now-ubiquitous DNA sequencing flow cells to directly assay protein function at scale. High-throughput flow cell DNA arrays, the core of Illumina sequencing [19], have recently been repurposed for quantitative high-throughput investigation of DNA-protein, RNA-protein, and RNA-RNA interactions across nucleic-acid sequence space. Building on this and other work [9, 20-24] we have reengineered high-throughput DNA sequencing methods to assay protein function across a vast polypeptide sequence space. This approach aims to bring quantitative protein functional investigation to DNA sequencing-scale throughput using a hardware platform and microfluidic chip compatible with fluorescence-based sequencing by synthesis (SBS) methods, demonstrating that a single, widely-available high-throughput platform can, in principle, perform both sequencing-by-synthesis and protein function assays.

To develop methods capable of generating and quantifying our protein array on SBS-compatible platform, we constructed a flexible, programmable workstation capable of microfluidically interfacing with and imaging sequencing flow cells [23, 25]. The resulting TIRF microscopy platform, based on the automated fluidics and fluorescence microscopy components of an Illumina GAIIx sequencer, interfaces with previously-sequenced (and therefore sequence- and position-indexed) Illumina MiSeq flow cell arrays [23]. To generate a protein array, we created a library of engineered DNA constructs that encode for polypeptides of interest (FIG. 4), then sequenced this library on an Illumina MiSeq. After moving the sequenced chip to our assay platform, we re-registered the cluster positions to their sequences (see methods). We next performed *in vitro* transcription/translation on chip such that both the transcript and nascent peptide remain associated with their DNA template, producing a tethered protein array. To enable this tethered display, each member of the sequencing library contained DNA sequence elements that allowed for 1) prokaryotic *in vitro* transcription, 2) immobilization of the resulting RNA transcript, 3) translation, and 4) ribosome stall, similar to ribosome display [26] (FIG. 1A). Fluorescence-based functional assays (e.g. quantifying binding of fluorescently-labeled binding partners or incorporation of fluorescent substrates) were conducted directly on this array. Fluorescence images, paired to cluster DNA sequences

generated from the sequencing run, were then quantified (see methods) to assay polypeptide binding or other protein function. We call this method **Protein display on a Massively Parallel Array**, or Prot-MaP.

As a testbed for this methodology, we characterized the widely-used FLAG peptide/M2 antibody interaction. FLAG peptide is commonly used for protein labeling, purification, and as a general affinity reagent [27]. The consensus sequence profile of the linear peptide epitope of M2 has previously been characterized as DYKxxDxx based on 5 [28] or 18 [29] clones selected from random peptide libraries. We aimed to generate a library of target peptides that would comprehensively probe the contribution of residues in the FLAG peptide to the M2 antibody interaction with a much larger library of variants. We engineered a variant library of 13,154 sequences that encoded all single, double, and triple-combinations of mutant positions, with each position substituted to each of 6 different amino acids – A,K,D,S,F, and L – that represent small, positive, negative, polar, aromatic, and aliphatic substitutions, respectively. The peptide-coding mutant library was produced with microarray synthesis and assembled into a sequenceable library construct with elements enabling transcription, translation, and stable peptide display (FIG. 1A, FIG. 4, see methods).

After DNA sequencing, 623,075 total clusters encoding FLAG library members (~2.5% of the total flow cell was needed to cover our library) were produced with 12,739 of 13,154 (96.8%) programmed variants represented by 8 or more peptide clusters. After peptide generation, M2 antibody was introduced at increasing concentrations and allowed to bind to the peptide, with each concentration followed by detection with a fluorescently-labeled secondary antibody, then imaging (FIG. 1B). After image registration and fluorescence quantification, the limit of detection (LoD) for antibody binding was determined for each variant (FIG. 1C, see methods), representing the lowest concentration of antibody that produces detectable binding. This Prot-MaP LoD assay is analogous to a massively multiplexed ELISA whereby primary antibody interactions are probed by secondary detection steps that produce measurable signal.

The single and double mutant affinity landscapes (FIGS. 2A-2B) of the canonical sequence (DYKDDDDK (SEQ ID NO:1), "WT") largely recapitulate the previously reported motif pattern, DYKxxDxx [28, 29]. However, we observe substantial additional

constraint at position D4, with 5 of 6 mutations exhibiting no detectable binding. To investigate mutational constraints at this position further, we asked if detrimental mutations at position 4 could be rescued by mutations at other positions in our higher order mutants. Looking at all double mutants of the only measurable D4 mutant (D4L; 5 triple mutants from WT; FIG. 2C), we observe that several combinations of mutations that include D5E and D7K partially rescue D4L, lowering LoD to near-WT levels. Many of these compensating combinations exhibit significant positively-cooperativity (e.g. D4L, D5E, D7K), demonstrating that, even with short peptides, antibody affinities can substantially deviate from simple additive interaction models (FIG. 2D).

10 Interestingly, we also found several FLAG variant sequences that exhibited significantly lower LoD than WT, including a variant we term "superFLAG" (DYKDEDLL, SEQ ID NO:3) with an extrapolated LoD of 0.14 nM, 7.9-fold lower than WT (FIG. 1C). To confirm these observations, we performed a fluorescence-based plate assay to measure M2 binding competition between immobilized variant peptides and 15 solution of 3xFLAG high-affinity peptide (FIG. 2E). SuperFLAG displays higher avidity to the M2 antibody than all other assayed peptides, including the original DYKDDDDK (SEQ ID NO:1) immunogen to which M2 was raised.

Unlike chemically synthesized peptide arrays, our *in vitro* transcribed and translated array enables the possibility of generating full-length protein features [13-15].

20 To demonstrate this capability, we investigated the mutational landscape of SNAP-tag, an engineered version of *O*⁶-alkylguanine-DNA alkyltransferase, a 181 AA, 20kD human DNA repair enzyme [30]. The SNAP-tag protein transfers a benzyl group from a derivatized benzylguanine substrate (often substituted with a fluorophore) to its own Cys 145, thereby covalently labeling itself. The simplicity and specificity of this fluorogenic 25 self-labeling reaction has led SNAP-tag to become widely used as a protein labeling tag, and here it provides an elegant model system for investigation of the sequence-function relationship of an enzyme with Prot-MaP (FIG. 3A).

To explore the relationship between protein sequence variation and catalysis in high throughput, we examined the functional interrelationship of seven residues 30 previously identified to modulate function without entirely abolishing catalytic activity (Y114, A121, K131, S135, L153, G157, and E159) [31] (FIG. 3B). The DNA sequence

encoding this mutated region (residues 114-159, the "mutant region"), was combinatorially assembled from oligos with the aim of generating single-, double-, and triple-mutant combinations of the 7 target positions across all possible 20 amino acid substitutions (FIG. 4). Once assembled, the DNA fragment population was bottlenecked
5 (by diluting to a target number of molecules) and reamplified with PCR to obtain multiple identical copies of molecular variants on the array.

After sequencing this library, RNA synthesis and *in situ* protein generation was performed on the MiSeq flow cell. We then introduced a solution of 200 nM substrate (SNAP-Surface 549) continuously onto the resulting protein array, then measured
10 substrate incorporation over 160 minutes (see methods). A total of 6,751,654 clusters were successfully quantified across all timepoints, and signals were averaged across clusters displaying identical protein variants (> 8 clusters per variant). A total of 156,140 unique variants were assayed, including all possible single mutants (133), 7570 of 7581 possible double mutants, 125,076 of 240,065 possible triple mutants (FIG. 3C), as well as
15 23,360 other variants. The majority of mutants (118,025/156,140; 75.59%) exhibited no detectable catalytic activity above background levels, while 32,072 variants exhibited detectable activity that could be well-fit by a single exponential to obtain k_{obs} (FIG. 3D, see methods).

We observed large variation in the mutational constraints for each of the 7
20 targeted amino acids. At one extreme, Y114 is fully constrained to tyrosine across all single and double mutants (FIGS. 3E, 3F). Even Y114F, a conservative substitution that deletes only a hydroxyl moiety, is inactive [32] (FIG. 3D), suggesting that the hydrogen bonds that Y114 forms with the benzylguanine substrate are absolutely required for function. While E159 also forms a hydrogen bond with the substrate, a number of
25 mutations to E159 retain measurable activity, suggesting that less stringent physicochemical requirements on substrate interactions at this position.

In contrast to these constrained residues, A121 is tolerant to all single mutations. L153 is tolerant to nearly all substitutions, except for proline, which is rotationally constrained in psi and phi Ramachandran angles. By exploring our catalogue of double
30 mutants, we observed that proline is consistently the most detrimental of all the 20 amino acids at position 153 across 119/120 other single mutant backgrounds. These

observations led to the hypothesis that L153, which is positioned in the turn of a loop region connecting helix-loop domain 153-176 (including the critical residue E159) with the rest of the protein, may serve as a "hinge" region that connects domains (FIG. 3G).

We investigated this possibility by examining the relationship between the observed enzymatic activity and amino acid backbone flexibility [33] of amino acid substitutions at position 153. We observed a strong relationship ($R^2=0.43$; $p\text{-value}=1.6\times 10^{-3}$; FIG. 3H), supporting this hypothesis and highlighting the utility of comprehensive mutational substitutions for testing mechanistic hypotheses.

We next aimed to characterize cooperativity in the interactions between the AAs we perturbed by analyzing double mutant cycles found within our mutational dataset (FIG. 3I). We observed a strong enrichment for physical proximity between highly positively cooperative amino acid pairs, with virtually all strong positively cooperative interactions occurring at $C\alpha$ - $C\alpha$ distances less than 13 Å (FIG. 3J). For example, many of the most positively cooperative double mutants are between positions 135 and 159 – two AAs that directly physically interact in the crystal structure (PDB ID: 3KZZ). The most cooperative pair is S135D-E159R, a double mutant that produces a favorable charge-based sidechain interaction.

Interestingly, histidine appears particularly amenable to cooperative interactions. Across all positional combinations, histidine was far more likely to form a cooperative interaction than any other amino acid (FIG. 3K), and did so fairly uniformly with all possible amino acid partners (FIG. 3L). Histidine can be either positively charged or neutral in different contexts, can have aromatic character, and can function as both a hydrogen bonding donor and acceptor. We hypothesized that histidine may thus function as a "Swiss army knife," pairing promiscuously with diverse partners to produce favorable interactions and positive cooperativity. We anticipate that further investigation with high-throughput methods will provide the opportunity to explore this hypothesis in a variety of protein backgrounds.

Prot-MaP enables the generation of immense mutational datasets for both peptides and full-length functional proteins directly on a sequencing flow cell, allowing high-throughput analysis of mutational effects based on direct biophysical observations of protein function. Large-scale quantitative measurements of peptide-protein interaction

can demonstrate and quantify nonadditivity in the affinity landscapes and allow the identification of enhanced-affinity interactions (such as superFLAG) and the large-scale functional analysis of full-length proteins allows data-driven characterization of functional networks and cooperative interactions of individual amino acids, geometric constraints that enforce amino acid preferences, as well as global observations and hypothesis testing regarding the functionality of specific amino acid species (e.g. histidine's capacity for widespread cooperative interactions).

ProtMAP does have limitations, including a practical limit of about ~1-2 kb on the protein-encoding DNA constructs that can be efficiently clustered and sequenced.

Additionally, many proteins will require different *in vitro* expression conditions and components for efficient production of functional protein (including post-translational modifications). Finally, while binding studies are broadly compatible with fluorescence-based assays, other assays (e.g. for arbitrary enzymatic activities, protein folding, or conformational changes) may necessitate the development of fluorescence-compatible measurement strategies. However, analogous to the many methods built on the foundations of high-throughput sequencing [34], we anticipate that the compatibility of Prot-MaP with widely-available high-throughput SBS platforms may serve as a catalyst for further development of specialized applications.

The implementation of facile high-throughput protein functional analysis on an already widely-used sequencing flow cell demonstrates a potentially straightforward path to bring protein functional assay into the high-throughput era, using automated fluorescence imaging apparatus consistent with that required for sequencing by synthesis. Given the demonstrated power and ubiquity of high-throughput DNA sequencing methods (as applied through functional genomic methods such as ChIP-seq, Hi-C etc.) to provide insights into the relationship between DNA sequence and regulatory function, we anticipate that a similarly ubiquitous platform for high-throughput protein assays could have analogous impact on our ability to dissect biologically relevant relationships between protein sequence, structure, and function.

Materials and Methods

Custom Fluorescence Microscopy Instrument to Interface with MiSeq flow cells
("Imaging Station")

We developed a custom instrument with all the core functionality of a sequencer that interfaces with a MiSeq flow cell, as well as software methods to align sequence reads to image features (see supplement). This platform has been described previously [23] but briefly, the camera, lasers, z-stage, x-y stage, syringe pump, and objective lens salvaged from Illumina GAIIx instruments were combined with custom laser control electronics, retrofitted temperature control, and a fluidic interface designed to be compatible with the MiSeq flow cell.

Protein Array constructs

To generate a protein array, engineered DNA constructs (FIG. 1A, FIG. 4) were constructed to have the elements necessary for sequencing on an unmodified Illumina MiSeq, principally the sequencing adapters P5 and P7 and sequencing primer hybridization sites. After clustering and sequencing on the MiSeq, the sequenced flow cell was moved to our custom platform, where transcription/translation was performed on chip such that both the transcript and nascent peptide remain associated with their DNA template, producing a tethered protein array. Library constructs therefore contain the elements requisite for prokaryotic *in vitro* transcription and translation, in addition to those that enable tethered display. These include an RNAP promoter and stall sequence that allow generation of a tethered RNA substrate with one-RNAP-per-template as previously described [22]. To allow generation of polypeptide, constructs contain a Shine-Dalgarno ribosomal binding site to initiate translation followed by a start codon. The polypeptide library of interest is encoded at the N-terminal end of the translated sequence. After this coding region, the open reading frame (ORF) continues with no stop codon via a short linker to a spacer polypeptide (2x ybbR tag [35]). A spacer is necessary to display the protein/peptide of interest (POI) away from the ribosome where it can be accessible to interaction and, in the case of longer proteins, fold correctly [36]. Sufficient linker sequence is present to ensure that the nascent POI and spacer become accessible outside the ribosome exit tunnel [37]. Next, a ribosome stall sequence that we engineered for maximum stalling efficiency (see FIG. 5) stalls ribosomal progress, allowing robust

display of the nascent peptide (FIG. 1A). Fluorescence-based functional assays (here, measuring binding of a fluorescently-labeled partner or the enzymatic incorporation of fluorescent substrates) may then be carried out directly on this generated array.

5 PCR assembly of constructs

All PCR assembly was performed with NEBNext High-Fidelity 2X PCR Master Mix (NEB) using the recommended reaction setup and thermocycling conditions.

Ribosomal Stall Sequence

10 Highly-stable, efficient ribosomal stalling is critical for generating high-quality peptide arrays. We achieve stalling with a ribosome stall sequence (CJL-PPP) that we rationally designed based on polyproline translational pausing in a maximally efficient context [38]. This sequence produces a strong pausing signal in our *in vitro* translation mix that does not contain Elongation Factor P (EF-P) (which alleviates stalling at the PPP motif) [39] nor Release Factor 1 (RF1), which would allow release at the Amber stop
15 codon that immediately follows the polyproline stretch. To further stabilize this display, critical residues from SecM and TnaC that are presumed to promote stalling via direct interaction with the exit tunnel (F150, W155, I156 from SecM [40], and W12, D16 from TnaC [41]) were also transplanted into the CJL-PPP sequence (FIG. 5).

20

FLAG Library

Each residue in the 8-amino acid WT FLAG peptide sequence (DYKDDDDK) is substituted with 6 different amino acids (A,K,D,S,F, and L, representing small, positive, negative, polar, aromatic, and aliphatic substitutions, respectively). If the wild-type
25 identity of the residue at that position is already one of the mutation set, a contingency substitution (V,R,E,T,Y, or I) was made in lieu of a redundant mutation. The library consists of all combinations of single, double, and triple substitutions of these 6 mutations at each position using optimal codons for *E. coli*. 2xFLAG (DYKDDDDKGDYKDHD, the first two FLAG motif repeats of the commonly-used
30 3xFLAG sequence) was also included in the library. Designs were synthesized on

microarray (CustomArray) with flanking primer regions, then assembled by PCR into the full-length protein array construct (FIG. 4)

SNAP-tag Library

5 Seven amino acid positions in the SNAP-tag sequence that are involved in substrate specificity (Y114, A121, K131, S135, L153, G157, and E159) [31] were targeted for mutation. The 138-base mutant region from 114-159, as well as the sequencing primer regions 31 bases upstream and 31 bases downstream, were covered with three overlapping oligos that could be combinatorially assembled. Oligos were
10 obtained (IDT) with all combinations of wild-type and NNK codons at each mutant position. The first primer covered positions 114 and 121, the second 131 and 135, and the third 153, 157, and 159. Assembly reactions for each single (7 reactions), double (21 reactions), and triple (35 reactions) combination of positions were individually assembled (FIG. 4), quantified, and pooled so that the final library representation was 1% single
15 mutants, 5% double mutants, and 94% triple mutants.

Additional constructs spanning the mutant region were also included in the final sequencing library. One encodes the WT region with no mutations, and another NNK codons at all 7 targeted positions. These were spiked in downstream at low percentage, along with the fiducial marks.

Bottlenecking

20 In order to have each molecular variant represented multiple times on chip, protein array libraries were bottlenecked down to 33fM (FLAG library) and 300fM (SNAP library) in 0.1% tween-20, then reamplified. The ~1kb SNAP library constructs
25 were gel extracted from a 1.5% agarose gel to ensure length homogeneity.

Sequencing Diversity

As with all Illumina sequencing, it is necessary to ensure sufficient sequence diversity in the sequenced region. FLAG peptide constructs were sequenced together with
30 other, unrelated libraries that added sequencing diversity (e.g. ATAC-seq samples). However, because the SNAP-tag constructs are ~1kb in size, a diversity element of

matched length (and therefore clustering efficiency) was specifically constructed. This diversity element has randomized N78mer sequence across the entire mutant region, but is otherwise identical to the protein array constructs. Because the SNAP-tag library contains stretches that are essentially homogenous, this diversity element was spiked in at up to 40%.

On-flow cell Polypeptide Array Generation

Libraries were prepared for sequencing on a MiSeq by first precisely quantifying all concentrations with qPCR and/or TapeStation (Agilent), then spiking in fiducial mark constructs into protein array libraries at 0.5-1%. Fiducial marks are sparse library members that serve as hybridization targets for a fluorescent probe; the resulting pattern is used to orient the chip relative to the sequencing data on the imaging station platform. Other library constructs, including the specifically-designed SNAP-tag diversity elements, were also present in the pooled sequencing sample to add sufficient sequence diversity for high-quality sequencing.

Standard amplicon sequencing was performed for FLAG tag peptide variants. For the ~1kb SNAP-tag variant library, a modified "Long Amplicon" sequencing recipe with more clustering cycles was used. In order to make these modifications, the "Amplicon" XML recipe in the MiSeq control software was modified to increase the bridge PCR "Repeat" cycles in "Amplification 1" during initial "OnBoardClusterGeneration" from 24 to 32, as well as the "Repeat" cycles in "Resynthesis" during "PairedEndTurnaround" from 12 to 16.) Sequencing was then performed with custom construct-specific sequencing primers (the read1 primer is the 31-base sequence immediately preceding the mutant region, and the read2 primer is the reverse complement of the 31-base sequence immediately following the mutant region). After sequencing, the sequenced MiSeq flow cell is moved to our custom imaging station.

On the Imaging Station platform, residual primers or read fragments from sequencing are stripped off with 100% formamide at 55°C. Cy3-labeled fiducial mark probes in hybridization buffer (5x SSC, 5mM EDTA, 0.01% Tween-20) are then hybridized to the fiducial marks (60°C for 15 min, 40°C for 12 min). Fiducial mark probes remain hybridized as reference points throughout the experiment.

Though the MiSeq flow cell has clusters along the length of its continuous channel, during the course of sequencing the MiSeq only collects data at a series of tiled locations. Experimental images must be taken the same tile positions in order to have sequence data match the imaged clusters. In order to discover these tile locations and program their relative positions on a new instrument, images from the imaging station are registered to the fiducial mark construct locations identified from the sequencing data with cross-correlation. MiSeq flow cells have two imaging surfaces (top and bottom) with a typical throughput of 20-25 million reads total. Because the current implementation of our imaging station only images the bottom of the flow cell and obstructs the objective lens from imaging the last (19th) tile, we typically make 10-12 million individual measurements across bottom tiles 1-18.

Once registered to the sequencing data, tethered RNA transcript is generated on chip as described previously [22] by stalling *E. coli* RNA polymerase on a DNA templates (clusters) with a terminal streptavidin roadblock. Constructs are washed with PBST +MgCl₂ (137 mM NaCl, 2.7 mM KCl, 8 mM Na₂HPO₄, 2 mM KH₂PO₄, 0.05% Tween-20, 7mM MgCl₂), then prokaryotic ribosomes are initiated on the transcript, which contains a Shine-Dalgarno sequence, with the PURExpress [42] *in vitro* translation mix. The PURExpress system is reconstituted from purified components and lacks EF-P [39] allowing ribosomes to stall at a polyproline-containing ribosome stall sequence that we engineered into the library, stably displaying the nascent peptide. We used a custom PURExpress formulation that omits Release Factors 1, 2, and 3 as well as T7 RNA polymerase, CTP, and UTP. Not only are these transcriptional components unnecessary (as the RNA template is already generated and displayed before adding the translation reagents) they interfered with stable nascent peptide display in this system despite the lack of a T7 promoter on our constructs.

Translation was performed for 1 hour at 37°C. After peptide display, the flow cell was washed with PBST +MgCl₂. All buffers post-RNA generation, including the translation mix, contain 0.8 U/μL Suprase-In RNase inhibitor (Thermo AM2696).

M2 Anti-FLAG Antibody Binding Study

After generation of the peptide array, nonspecific binding sites were blocked with bind buffer (0.8% BSA and 500nM goat IgG in PBST +MgCl₂) for 30min at 25°C. All subsequent binding steps were performed at 25°C. 35nM Alexa555-goat anti-mouse secondary antibody (Abcam ab150086) in bind buffer was then incubated for 37 min so that the initial, zero-primary-antibody-concentration point would show any background binding of secondary antibody as baseline. Each experimental concentration of primary antibody (1nM, 3.2nM, 10nM, 32nM, 100nM) was incubated for 45 mins, followed by a 25 min wash with bind buffer, then detected with a 37 min incubation with 35nM Alexa555-goat anti-mouse secondary antibody in bind buffer, and a 25 min wash with bind buffer. After each experimental concentration, tiles on the bottom side of the flow cell were TIRF imaged at 532nm excitation with a 590/104 nm bandpass (Semrock FF01-590/104-25) emission filter, using the same laser power and camera exposure settings across all concentrations.

Continuous flow at 12.5 µL/min was used in lieu of static incubation for low-nM primary Ab concentrations to mitigate depletion.

SNAP-tag Enzyme Activity Assay

SNAP-tag substrate SNAP-surface 549 (NEB #S9112) was dissolved in DMSO to a stock concentration of 1µM, then was diluted immediately before use to a working concentration of 200nM in PBST +MgCl₂. After generation of the protein array, the working solution of SNAP substrate was flowed continuously into the flow cell at 12.5µL/min for the experimental incubation times to minimize substrate depletion or product inhibition. At each imaging timepoint, fresh PBST +MgCl₂ was flowed in as an imaging buffer, and tiles on the bottom side of the flow cell were TIRF imaged at 532nm excitation and a 590/104 nm bandpass (Semrock FF01-590/104-25) emission filter. (Cumulative reactions times include only incubation time in the presence of substrate, not imaging buffer).

Quantification

Fluorescence images, paired with cluster sequence identities and positions from the sequencing run, are quantified to measure fluorescent signal, and therefore assay the fluorescence-reported polypeptide or protein function at each cluster.

After imaging, fine registration of the images to the sequencing data is necessary before image quantification due to microscope stage position inaccuracy and, more onerously, potentially non-affine optical aberrations between the sequencing platform and the imaging station. To address this, each image is divided into a grid of sub-images that are individually registered to the data, followed by a fit of the offset positions to a 2D quadratic surface. The resulting continuous offset map is then applied to the data to achieve fine registration of cluster positions to the images [23].

Images are then quantified by fitting small image subtiles to a summation of 2D gaussians centered at the finely registered positions of each cluster. Edge effects in quantification are mitigated by discarding fit values close enough to an edge to be affected by out-of-bounds clusters, and using overlapping subtiles to completely cover the image [22].

To correct for small variances in quantified intensity caused by inconsistent focus and/or illumination power fluctuation, median fluorescence values were normalized to the median intensity of the fiducial marks.

Data from quantified images are consolidated across concentrations on a per-cluster basis. The resulting experimental series are then grouped into molecular variants with identical amino acid sequences. Grouping into variants was handled slightly differently for the FLAG and SNAP experiments. For FLAG variants, a 16-base random unique molecular identifier (UMI) was read (as index read 2) and the consensus sequence of the amino acid-coding region determined for each UMI was used. SNAP library constructs did not have a UMI sequence and the raw coding sequence was used. For all variants represented by some minimum number of clusters (here, 8), median fluorescence values were calculated at each experimental timepoint, which were then used for downstream analysis.

M2 Anti-FLAG Analysis

A threshold of detection was defined by a set of negative control peptide sequences. The

negative set consists of all variants represented by 8-or-more clusters where residues D1, Y2, and K3 (each individually critical for binding) were all mutated in combination. The threshold was then defined as 3σ above the mean of the fluorescence medians of this negative set. Encouragingly, fluorescent signal from the negative set did not appreciably increase with increasing concentration, so the median value of the negative set across all concentration points was used as a concentration-independent background threshold.

This threshold of detection served as the basis for quantification of variant LoD. Any variant that did not cross the threshold was classified as "negative" (LoD greater than the highest the assay can measure).

For variants with 4-or-more points above threshold, the points above threshold (only) were fit to a line and the intersection of the line with the threshold value was reported as the LoD. For variants with fewer than 4 points above threshold, a linear interpolation was done between the point before and the point after crossing the threshold (FIG. 1C).

SNAP-tag Analysis

A threshold of detection was defined by a set of negative control enzyme variants. The negative set consists of all well-represented variants where all seven targeted residues (Y114, A121, K131, S135, L153, G157, E159) were mutated in combination. The threshold was then defined as 3.3σ above the mean of the negative set medians across concentration points.

All variant means that rose above threshold by the second-to-last timepoint were fit by unconstrained nonlinear least squares to a single exponential:

$$F = F_{min} + F_{max}(1 - e^{-kt})$$

Quality control (QC) of fits was evaluated by three criteria: (1) $R^2 > 0.70$, (2) F_{max} above threshold at the highest concentration (3) $F_{min} < F_{max} < \text{threshold} F_{max}$ (see FIG. 6). Variants that did not meet these criteria were excluded from downstream analysis.

Orthogonal investigation of M2/FLAG peptide variant interactions with a fluorescent plate assay

To test M2 antibody binding to these variant peptides orthogonally from Prot-MaP, an *in vitro* plate assay was performed. First, several 8-amino acid variants of the FLAG peptide were individually synthesized followed by the context bases DHDGS (identical to the 5 downstream bases of the on-chip constructs) and a C-terminal PEG-biotin. A neutravidin-coated plate (Pierce 15117) was washed, then blocked with StartingBlock PBS (Thermo 37538). Biotinylated peptides were then bound to the plate by incubating for 30 min at a concentration of 5 nM. After washing, primary antibody (M2 anti-FLAG, Sigma F1804) was incubated in the plate for 30 minutes, then washed, then fluorescently-labeled (DyLight 550) goat Anti-Mouse secondary antibody (Abcam ab98713) was incubated for 1 hour, then washed again. In order to observe differential binding, a concentration gradient of 3xFLAG peptide (Sigma F4799) competitor in PBST +MgCl₂ was added and allowed to equilibrate for 2 hours, after which the plate was washed and read at 535(25) nm excitation and 580(30) nm emission. (Antibodies and peptides were diluted in StartingBlock PBS + 0.05% Tween-20, all washes and fluorescence measurements were performed with PBST +MgCl₂, and all incubations were performed at 4°C with shaking.)

Table 1: FLAG Peptide Variants

AA sequence	LoD (nM)
DYKDEDLL (SEQ ID NO:3)	0.13718
DYKDEDKL (SEQ ID NO:4)	0.17039
DYKDEDSA (SEQ ID NO:5)	0.18977
DYKDEDLA (SEQ ID NO:6)	0.20199
DYKDEDSL (SEQ ID NO:7)	0.21084
DYKDDDKA (SEQ ID NO:8)	0.21400
DYKDEDLF (SEQ ID NO:9)	0.21498
DYKDEDKA (SEQ ID NO:10)	0.21683
DYKDDDLL (SEQ ID NO:11)	0.22373
DYKDDDKS (SEQ ID NO:12)	0.22766
DYKDDDKF (SEQ ID NO:13)	0.23451

DYKDADLF (SEQ ID NO:14)	0.23606
DYKDEDDFF (SEQ ID NO:15)	0.23861
DYKDEEDAS (SEQ ID NO:16)	0.23980
DYKDEEDDS (SEQ ID NO:17)	0.24136
DYKDDDLF (SEQ ID NO:18)	0.24739
DYKDEEDLS (SEQ ID NO:19)	0.24832
DYKDFDKF (SEQ ID NO:20)	0.25084
DYKDEEDKS (SEQ ID NO:21)	0.25223
DYKDEEDFL (SEQ ID NO:22)	0.25310
DYKDEEDSF (SEQ ID NO:23)	0.25586
DYKDFDLL (SEQ ID NO:24)	0.25603
DYKDDDLIS (SEQ ID NO:25)	0.25608
DYKDADKK (SEQ ID NO:26)	0.25628
DYKDDDKD (SEQ ID NO:27)	0.26360
DYKDDDAF (SEQ ID NO:28)	0.26659
DYKDFDFS (SEQ ID NO:29)	0.27736
DYKDEEAA (SEQ ID NO:30)	0.28229
DYKDEEDLR (SEQ ID NO:31)	0.28250
DYKDDDFS (SEQ ID NO:32)	0.28437
DYKDADKS (SEQ ID NO:33)	0.28445
DYKDADKF (SEQ ID NO:34)	0.29054
DYKDADLA (SEQ ID NO:35)	0.29187
DYKDEEDAL (SEQ ID NO:36)	0.29328
DYKDFDLI (SEQ ID NO:37)	0.29676
DYKDEDEL (SEQ ID NO:38)	0.29681
DYKDADKA (SEQ ID NO:39)	0.29933
DYKDEEDSR (SEQ ID NO:40)	0.30346
DYKDDDLA (SEQ ID NO:41)	0.30419
DYKDADSF (SEQ ID NO:42)	0.30474
DYKDEEDFD (SEQ ID NO:43)	0.30475
DYKDFDLS (SEQ ID NO:44)	0.30934
DYKDEEDFS (SEQ ID NO:45)	0.31040
DYKDEDED (SEQ ID NO:46)	0.31135
DYKDDDLK (SEQ ID NO:47)	0.31495
DYKDLDLF (SEQ ID NO:48)	0.31628
DYKDFDKS (SEQ ID NO:49)	0.31658
DYKDEDEF (SEQ ID NO:50)	0.31698
DYKDEDAF (SEQ ID NO:51)	0.31758
DYKDLDFL (SEQ ID NO:52)	0.31792

DYKDEDKR (SEQ ID NO:53)	0.32063
DYKDEDFK (SEQ ID NO:54)	0.32339
DYKDEDDA (SEQ ID NO:55)	0.32888
DYKDFDLR (SEQ ID NO:56)	0.33108
DYKDDDDS (SEQ ID NO:57)	0.33120
DYKDLDFS (SEQ ID NO:58)	0.33280
DYKDFDLA (SEQ ID NO:59)	0.34252
DYKDADKR (SEQ ID NO:60)	0.34308
DYKDDDKR (SEQ ID NO:61)	0.34430
DYKDADAF (SEQ ID NO:62)	0.34767
DYKDADLK (SEQ ID NO:63)	0.35190
DYKDFDFA (SEQ ID NO:64)	0.35292
DYKDEDDFA (SEQ ID NO:65)	0.35322
DYKDFDFL (SEQ ID NO:66)	0.35703
DYKDLDLL (SEQ ID NO:67)	0.35910
DYKDDDEF (SEQ ID NO:68)	0.36444
DYKDDDL D (SEQ ID NO:69)	0.36502
DYKDDDSF (SEQ ID NO:70)	0.36550
DYKDLDKF (SEQ ID NO:71)	0.36653
DYKDEDFR (SEQ ID NO:72)	0.36756
DYKDLDKS (SEQ ID NO:73)	0.36818
DYKDFDLF (SEQ ID NO:74)	0.37021
DYKDDDLR (SEQ ID NO:75)	0.37262
DYKDDDKL (SEQ ID NO:76)	0.37538
DYKDDDF A (SEQ ID NO:77)	0.37747
DYKDFDKA (SEQ ID NO:78)	0.37890
DYKDEDES (SEQ ID NO:79)	0.38189
DYKDEDAR (SEQ ID NO:80)	0.38450
DYKDLDLA (SEQ ID NO:81)	0.38592
DYKDFDFD (SEQ ID NO:82)	0.39141
DYKDFDFF (SEQ ID NO:83)	0.39710
DYKDLDKR (SEQ ID NO:84)	0.39917
DYKDFDLK (SEQ ID NO:85)	0.40338
DYKDESD (SEQ ID NO:86)	0.40580
DYKDEDS S (SEQ ID NO:87)	0.40794
DYKDEDSK (SEQ ID NO:88)	0.41061
DYKDDDAS (SEQ ID NO:89)	0.41125
DYKDDDA A (SEQ ID NO:90)	0.41478
DYKDLDFR (SEQ ID NO:91)	0.42117

DYKDLCLS (SEQ ID NO:92)	0.42276
DYKDLDDS (SEQ ID NO:93)	0.42441
DYKDFDDS (SEQ ID NO:94)	0.42627
DYKDADFL (SEQ ID NO:95)	0.42712
DYKDFDKD (SEQ ID NO:96)	0.42812
DYKDLCLK (SEQ ID NO:97)	0.42926
DYKDADFA (SEQ ID NO:98)	0.42934
DYKDEDDF (SEQ ID NO:99)	0.43227
DYKDADLL (SEQ ID NO:100)	0.43252
DYKDEDEK (SEQ ID NO:101)	0.43645
DYKDDDFR (SEQ ID NO:102)	0.43708
DYKDFDSL (SEQ ID NO:103)	0.43896
DYKDFDFR (SEQ ID NO:104)	0.43993
DYKDFDES (SEQ ID NO:105)	0.44063
DYKDDDAL (SEQ ID NO:106)	0.44305
DYKDDDES (SEQ ID NO:107)	0.44789
DYKDDDFL (SEQ ID NO:108)	0.45404
DYKDEDDR (SEQ ID NO:100)	0.45465
DYKDLKA (SEQ ID NO:110)	0.45476
DYKDFDSS (SEQ ID NO:111)	0.45545
DYKDFDFK (SEQ ID NO:112)	0.45630
DYKDFDSA (SEQ ID NO:113)	0.46285
DYKDFDSF (SEQ ID NO:114)	0.46671
DYKDLCLR (SEQ ID NO:115)	0.46720
DYKDADSS (SEQ ID NO:116)	0.46749
DYKDLDDF (SEQ ID NO:117)	0.46814
DYKDLDSA (SEQ ID NO:118)	0.46956
DYKDL DFA (SEQ ID NO:119)	0.47156
DYKDADFS (SEQ ID NO:120)	0.47602
DYKDLDEF (SEQ ID NO:121)	0.47754
DYKDADLS (SEQ ID NO:122)	0.48498
DYKDFDKL (SEQ ID NO:123)	0.48837
DYKDEDDD (SEQ ID NO:124)	0.48845
DYKDFDSR (SEQ ID NO:125)	0.49137
DYKDADFF (SEQ ID NO:126)	0.49138
DYKDADLD (SEQ ID NO:127)	0.49306
DYKDFDEF (SEQ ID NO:128)	0.49440
DYKDLDSR (SEQ ID NO:129)	0.49820
DYKDDDF (SEQ ID NO:130)	0.50031

DYKDLDFE (SEQ ID NO:131)	0.50049
DYKDLDSL (SEQ ID NO:132)	0.50659
DYKDFDDF (SEQ ID NO:133)	0.51483
DYKDDDSR (SEQ ID NO:134)	0.51498
DYKDDDFD (SEQ ID NO:135)	0.51612
DYKDEDAD (SEQ ID NO:136)	0.51859
DYKDFDSD (SEQ ID NO:137)	0.51870
DYKDADSL (SEQ ID NO:138)	0.52067
DYKDDDEL (SEQ ID NO:139)	0.52594
DYKDDDSA (SEQ ID NO:140)	0.53001
DYKDFDAR (SEQ ID NO:141)	0.53319
DYKDLDS (SEQ ID NO:142)	0.53415
DYKDLDFD (SEQ ID NO:143)	0.53752
DYKDADEA (SEQ ID NO:144)	0.54072
DYKDADAR (SEQ ID NO:145)	0.54695
DYKDEDER (SEQ ID NO:146)	0.56211
DYKDFDSK (SEQ ID NO:147)	0.57179
DYKDDDER (SEQ ID NO:148)	0.57197
DYKDEDDL (SEQ ID NO:149)	0.57346
DYKDLDFK (SEQ ID NO:150)	0.57357
DYKDADSK (SEQ ID NO:151)	0.57853
DYKDLDSF (SEQ ID NO:152)	0.57948
DYKDDDDF (SEQ ID NO:153)	0.57983
DYKDFDER (SEQ ID NO:154)	0.58536
DYKDDDAR (SEQ ID NO:155)	0.58589
DYKDDDSL (SEQ ID NO:156)	0.58594
DYKDADKD (SEQ ID NO:157)	0.59029
DYKDFDAS (SEQ ID NO:158)	0.60022
DYKDFDDD (SEQ ID NO:159)	0.60496
DYKDADLR (SEQ ID NO:160)	0.61532
DYKDADDS (SEQ ID NO:161)	0.61557
DYKDADSA (SEQ ID NO:162)	0.61685
DYKDADEF (SEQ ID NO:163)	0.61908
DYKDFDDA (SEQ ID NO:164)	0.62052
DYKDFDAA (SEQ ID NO:165)	0.62491
DYKDLDKD (SEQ ID NO:166)	0.62762
DYKDADSR (SEQ ID NO:167)	0.63341
DYKDADAA (SEQ ID NO:168)	0.63606
DYKDFDAF (SEQ ID NO:169)	0.63946

DYKDLDES (SEQ ID NO:170)	0.63946
DYKDLDAK (SEQ ID NO:171)	0.64398
DYKDADAL (SEQ ID NO:172)	0.64624
DYKDDDFK (SEQ ID NO:173)	0.65028
DYKDLDEA (SEQ ID NO:174)	0.65449
DYKDLDL (SEQ ID NO:175)	0.65975
DYKDLDS (SEQ ID NO:176)	0.66328
DYKDADFR (SEQ ID NO:177)	0.66465
DYKDLDA (SEQ ID NO:178)	0.66969
DYKDADDF (SEQ ID NO:179)	0.68294
DYKDDDS (SEQ ID NO:180)	0.68952
DYKDDDEK (SEQ ID NO:181)	0.69516
DYKDDDAK (SEQ ID NO:182)	0.69991
DYKDADAK (SEQ ID NO:183)	0.70666
DYKDSDFL (SEQ ID NO:184)	0.71906

References and Notes

1. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology*. Nature Reviews Genetics, 2009. **10**(10): p. 669.
2. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position*. Nature methods, 2013. **10**(12): p. 1213.
3. Krueger, F., et al., *DNA methylome analysis using short bisulfite sequencing data*. Nature methods, 2012. **9**(2): p. 145.
4. Lieberman-Aiden, E., et al., *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome*. Science, 2009. **326**(5950): p. 289293.
5. Zykovich, A., I. Korf, and D.J. Segal, *Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing*. Nucleic acids research, 2009. **37**(22): p. e151-e151.
6. Kingsmore, S.F., *Multiplexed protein measurement: technologies and applications of protein and antibody arrays*. Nature reviews Drug discovery, 2006. **5**(4): p. 310-321.

7. Larman, H.B., et al., *Autoantigen discovery with a synthetic human peptidome*. Nat Biotech, 2011. **29**(6): p. 535-541.
8. Fowler, D.M., et al., *High-resolution mapping of protein sequence-function relationships*. Nature methods, 2010. **7**(9): p. 741-746.
- 5 9. Gu, L., et al., *Multiplex single-molecule interaction profiling of DNA-barcoded proteins*. Nature, 2014. **515**(7528): p. 554-557.
10. Levin, A. and G. Weiss, *Optimizing the affinity and specificity of proteins with molecular display*. Molecular Biosystems, 2006. **2**(1): p. 49-57.
11. Odegrip, R., et al., *CIS display: in vitro selection of peptides from*
10 *libraries of protein– DNA complexes*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(9): p. 2806-2810.
12. Chandra, H., P.J. Reddy, and S. Srivastava, *Protein microarrays and novel detection platforms*. Expert review of proteomics, 2011. **8**(1): p. 61-79.
13. He, M., et al., *Printing protein arrays from DNA arrays*. Nat Meth, 2008.
15 **5**(2): p. 175177.
14. Ramachandran, N., et al., *Self-Assembling Protein Microarrays*. Science, 2004. **305**(5680): p. 86-90.
15. Tao, S.-C. and H. Zhu, *Protein chip fabrication by capture of nascent polypeptides*. Nat Biotech, 2006. **24**(10): p. 1253-1254.
- 20 16. Hilpert, K., D.F. Winkler, and R.E. Hancock, *Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion*. Nature protocols, 2007. **2**(6): p. 1333-1349.
17. Legutki, J.B., et al., *Scalable high-density peptide arrays for*
25 *comprehensive health monitoring*. 2014. **5**: p. 4785.
18. Forsström, B., et al., *Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays*. Molecular & Cellular Proteomics, 2014. **13**(6): p. 1585-1597.
19. Bentley, D.R., et al., *Accurate whole human genome sequencing using*
30 *reversible terminator chemistry*. nature, 2008. **456**(7218): p. 53-59.

20. Nutiu, R., et al., *Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument*. Nat Biotech, 2011. **29**(7): p. 659-664.
21. Tome, J.M., et al., *Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling*. Nature methods, 2014. **11**(6): p. 683-688.
22. Buenrostro, J.D., et al., *Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes*. Nat Biotech, 2014. **32**(6): p. 562-568.
23. She, R., et al., *Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome*. Proceedings of the National Academy of Sciences, 2017. **114**(14): p. 3619-3624.
24. Svensen, N., O.B. Peersen, and S.R. Jaffrey, *Peptide Synthesis on a Next-Generation DNA Sequencing Platform*. ChemBioChem, 2016. **17**(17): p. 1628-1635.
25. Jung, C., et al., *Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips*. Cell, 2017. **170**(1): p. 35-47. e13.
26. Lipovsek, D. and A. Plückthun, *In-vitro protein evolution by ribosome display and mRNA display*. Journal of Immunological Methods, 2004. **290**(1): p. 51-67.
27. Einhauer, A. and A. Jungbauer, *The FLAGTM peptide, a versatile fusion tag for the purification of recombinant proteins*. Journal of biochemical and biophysical methods, 2001. **49**(1): p. 455-465.
28. Srila, W. and M. Yamabhai, *Identification of Amino Acid Residues Responsible for the Binding to Anti-FLAGTM M2 Antibody Using a Phage Display Combinatorial Peptide Library*. Applied Biochemistry and Biotechnology, 2013. **171**(3): p. 583-589.
29. Osada, E., et al., *Epitope mapping using ribosome display in a reconstituted cell-free protein synthesis system*. Journal of biochemistry, 2009. **145**(5): p. 693-700.

30. Sun, X., et al., *Development of SNAP-Tag Fluorogenic Probes for Wash-Free Fluorescence Imaging*. ChemBioChem, 2011. **12**(14): p. 2217-2226.
31. Gautier, A., et al., *An engineered protein tag for multiprotein labeling in living cells*. Chemistry & biology, 2008. **15**(2): p. 128-136.
- 5 32. Mollwitz, B., et al., *Directed evolution of the suicide protein O 6-alkylguanine-DNA alkyltransferase for increased reactivity results in an alkylated protein with exceptional stability*. Biochemistry, 2012. **51**(5): p. 986-994.
33. Fuchs, J.E., et al., *Independent Metrics for Protein Backbone and Side-Chain Flexibility: Time Scales and Effects of Ligand Binding*. Journal of chemical
10 theory and computation, 2015. **11**(3): p. 851-860.
34. Morozova, O. and M.A. Marra, *Applications of next-generation sequencing technologies in functional genomics*. Genomics, 2008. **92**(5): p. 255-264.
35. Yin, J., et al., *Genetically encoded short peptide tag for versatile protein labeling by Sfp phosphopantetheinyl transferase*. Proceedings of the National
15 Academy of Sciences of the United States of America, 2005. **102**(44): p. 15815-15820.
36. Schaffitzel, C., et al., *Ribosome display: an in vitro method for selection and evolution of antibodies from libraries*. Journal of immunological methods,
20 1999. **231**(1): p. 119-135.
37. Hardesty, B. and G. Kramer, *Folding of a nascent peptide on the ribosome*. Progress in nucleic acid research and molecular biology, 2000. **66**: p. 41-66.
38. Starosta, A.L., et al., *Translational stalling at polyproline stretches is modulated by the sequence context upstream of the stall site*. Nucleic acids
25 research, 2014. **42**(16): p. 10711-10719.
39. Ude, S., et al., *Translation Elongation Factor EF-P Alleviates Ribosome Stalling at Polyproline Stretches*. Science, 2013. **339**(6115): p. 82-85.
40. Nakatogawa, H. and K. Ito, *The ribosomal exit tunnel functions as a discriminating gate*. Cell, 2002. **108**(5): p. 629-636.
30

41. Seidelt, B., et al., Structural insight into nascent polypeptide chain–mediated translational stalling. *Science*, 2009. 326(5958): p. 1412-1415.
42. Shimizu, Y., et al., The PURE System for Protein Production, in *Cell-Free Protein Synthesis: Methods and Protocols*, K. Alexandrov and W.A. Johnston, Editors. 2014, Humana Press: Totowa, NJ. p. 275-284.

While the preferred embodiments of the invention have been illustrated and described, it will be appreciated that various changes can be made therein without departing from the spirit and scope of the invention.

Claims

What is claimed is:

1. A variant FLAG peptide comprising an amino acid sequence selected from the group consisting of:
 - a) an amino acid sequence selected from the group consisting of SEQ ID NOS:3-184; and
 - b) an amino acid sequence having at least 80% identity to an amino acid sequence selected from the group consisting of SEQ ID NOS:3-184, wherein the variant FLAG peptide is capable of binding to an anti-FLAG antibody with greater affinity than a FLAG peptide comprising the amino acid sequence of SEQ ID NO:1.
2. The variant FLAG peptide of claim 1, wherein the variant FLAG peptide consists of the sequence of SEQ ID NO:3.
3. A fusion protein comprising one or more variant FLAG peptides of any prior claim connected to a target protein.
4. The fusion protein of claim 3, further comprising a linker.
5. The fusion protein of claim 3 or 4, wherein the fusion protein comprises 1-5 variant FLAG peptides.
6. A method of detecting the fusion protein of any of claims 3-5, the method comprising:
 - a) contacting the fusion protein with a detectably labeled anti-FLAG antibody, wherein the anti-FLAG antibody specifically binds to the variant FLAG peptide of the fusion protein; and
 - b) detecting the bound fusion protein.

7. The method of claim 6, wherein said detecting comprises performing a Western, immunofluorescence, or immunoprecipitation.
8. The method of claim 6 or 7, wherein the detectably labeled anti-FLAG antibody comprises a fluorescent label, a chemiluminescent label, or a radioactive label.
9. The method of any of claims 6-8, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.
10. A method of isolating the fusion protein described of any of claims 6-9, the method comprising immunoprecipitating the fusion protein with an anti-FLAG antibody that specifically binds to the variant FLAG peptide of the fusion protein.
11. The method of claim 10, wherein the anti-FLAG antibody is immobilized on a solid support.
12. The method of claim 11, wherein the solid support is a magnetic bead or a nonmagnetic bead.
13. The method of any of claims 10-12, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.
14. A method of purifying the fusion protein of any of claims 6-9 from a mixture using an affinity chromatography matrix comprising an anti-FLAG antibody, the method comprising:
 - a) contacting the mixture with the affinity chromatography matrix under a first set of conditions, such that the fusion protein binds to the anti-FLAG antibody attached to the affinity chromatography matrix; and b) eluting the fusion protein from the affinity chromatography matrix under a second set of conditions thereby purifying the fusion protein from the mixture.

wherein the affinity chromatography may be performed in a column or in batch.

15. The affinity reagent of claim 14, wherein the solid support is selected from the group consisting of a bead, a particle, a rod, a membrane, a gel, a resin, a column, a chip, a slide, a plate, and a microarray.
16. The affinity reagent of claim 15, wherein the bead is an agarose bead.
17. The affinity reagent of claim 15, wherein the bead is a magnetic bead or a nonmagnetic bead.
18. The affinity reagent of claim 15, wherein the bead is a microbead.
19. The affinity reagent of claim 15, wherein the plate is a microtiter plate.
20. The affinity reagent of claim 15, wherein the membrane is a nitrocellulose membrane or a polyvinylidene difluoride (PVDF) membrane.
21. The method of any of claims 14-20, wherein the anti-FLAG antibody specifically binds to the variant FLAG peptide of the fusion protein.
22. The method of any of claims 14-21, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.
23. A polynucleotide encoding the fusion protein of any of claims 3-9.
24. A recombinant polynucleotide comprising the polynucleotide of claim 23 operably linked to a promoter.
25. A host cell comprising the recombinant polynucleotide of claim 24.

26. A method for producing a fusion protein, the method comprising the steps of:
- a) culturing a host cell comprising the recombinant polynucleotide of claim 24 under conditions suitable for the expression of the fusion protein; and
 - b) recovering the fusion protein from the host cell culture using an affinity agent comprising an anti-FLAG antibody.
27. A vector for producing a tagged target protein, the vector comprising:
- a) a polynucleotide encoding the variant FLAG peptide of claim 1;
 - b) a cloning site suitable for in-frame insertion of a polynucleotide encoding the target protein, wherein the cloning site is positioned before or after the polynucleotide encoding the variant FLAG peptide; and
 - c) a promoter operably linked to sequences encoding the fusion protein.
28. The vector of claim 27, wherein the cloning site comprises one or more restriction sites.
29. The vector of claim 27, further comprising a polynucleotide encoding a linker, wherein said polynucleotide encoding the linker is positioned in between the cloning site and the polynucleotide encoding the variant FLAG peptide.
30. The vector of claim 27, further comprising a selectable marker.
31. A kit for tagging a target protein with a variant FLAG peptide, the kit comprising the vector of claim 27 and instructions for preparing and detecting the fusion protein.
32. The kit of claim 31, further comprising an anti-FLAG antibody.
33. The kit of claim 32, wherein the anti-FLAG antibody is an M2 monoclonal anti-FLAG antibody.
34. The kit of claim 32, wherein the anti-FLAG antibody is immobilized on a solid

support.

35. The kit of claim 34, wherein the solid support is selected from the group consisting of a bead, a particle, a rod, a membrane, a gel, a resin, a column, a chip, a slide, a plate, and a microarray.

36. The kit of claim 35, wherein the bead is a magnetic bead or a nonmagnetic bead.

37. The kit of claim 35, wherein the plate is a microtiter plate.

38. The kit of claim 35, wherein the membrane is a nitrocellulose membrane or a polyvinylidene difluoride (PVDF) membrane.

39. The kit of claim 31, further comprising reagents for performing a Western, immunoprecipitation, immunofluorescence, or affinity chromatography.

40. A screening method comprising:

(a) on a planar substrate (e.g., using an Illumina substrate), sequencing a library of engineered DNA constructs that encode for polypeptides of interest; and

(b) performing an *in vitro* transcription/translation on the substrate after (a) such that both the transcript and nascent peptide remain associated with their DNA template, thereby producing a tethered protein array.

41. The method of claim 40, wherein each member of the sequencing library contains DNA sequence elements that allow for 1) prokaryotic *in vitro* transcription, 2) immobilization of the resulting RNA transcript, 3) translation, and 4) ribosome stall.

42. The method of claims 40-41, further comprising performing a fluorescence-based functional assay (e.g. quantifying binding of fluorescently-labeled binding partners or incorporation of fluorescent substrates) tethered protein array on the substrate.

43. The method of claims 42, further comprising reading the substrate.
44. The method of claim 43, further comprising matching the results to the sequence of the engineered DNA constructs.

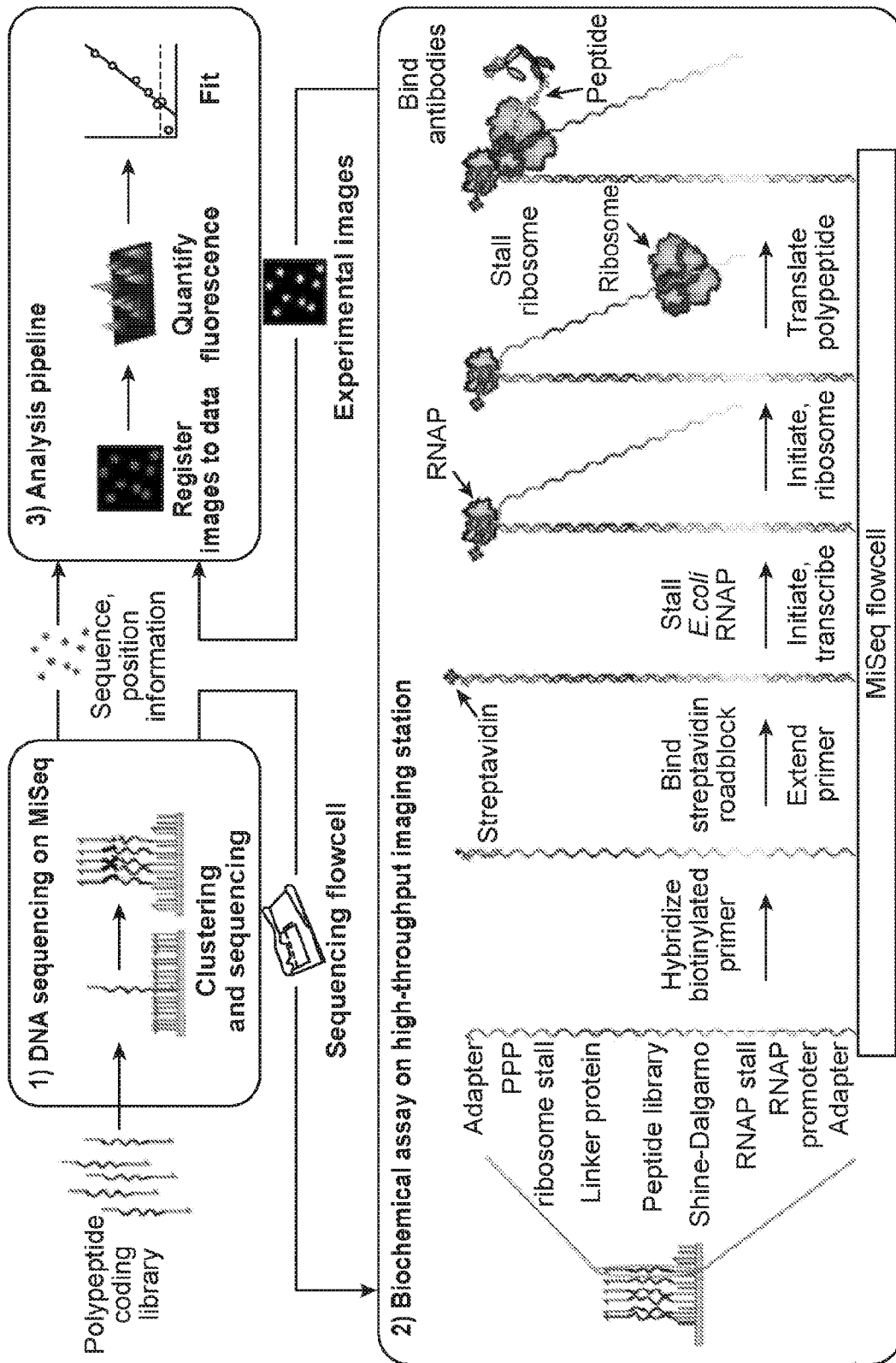


FIG. 1

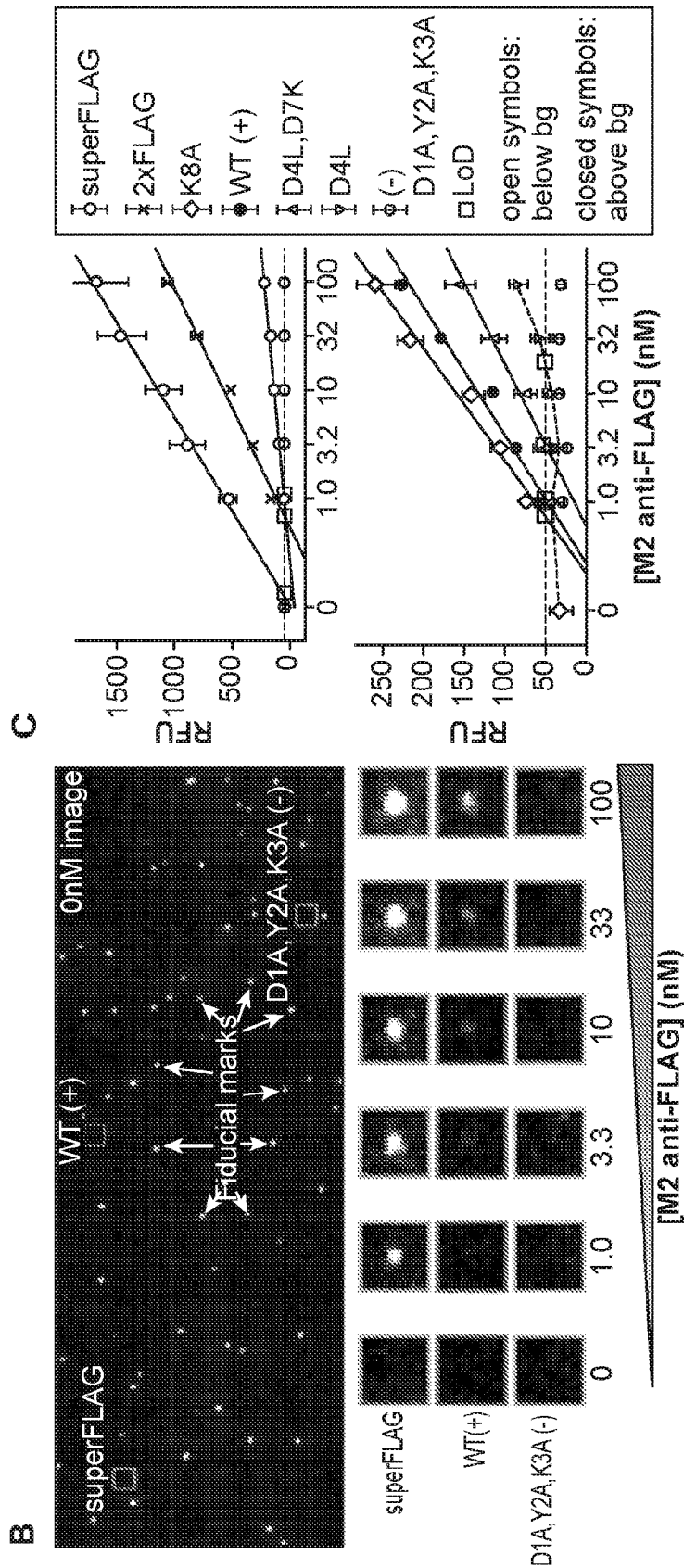


FIG. 1 (Cont.)

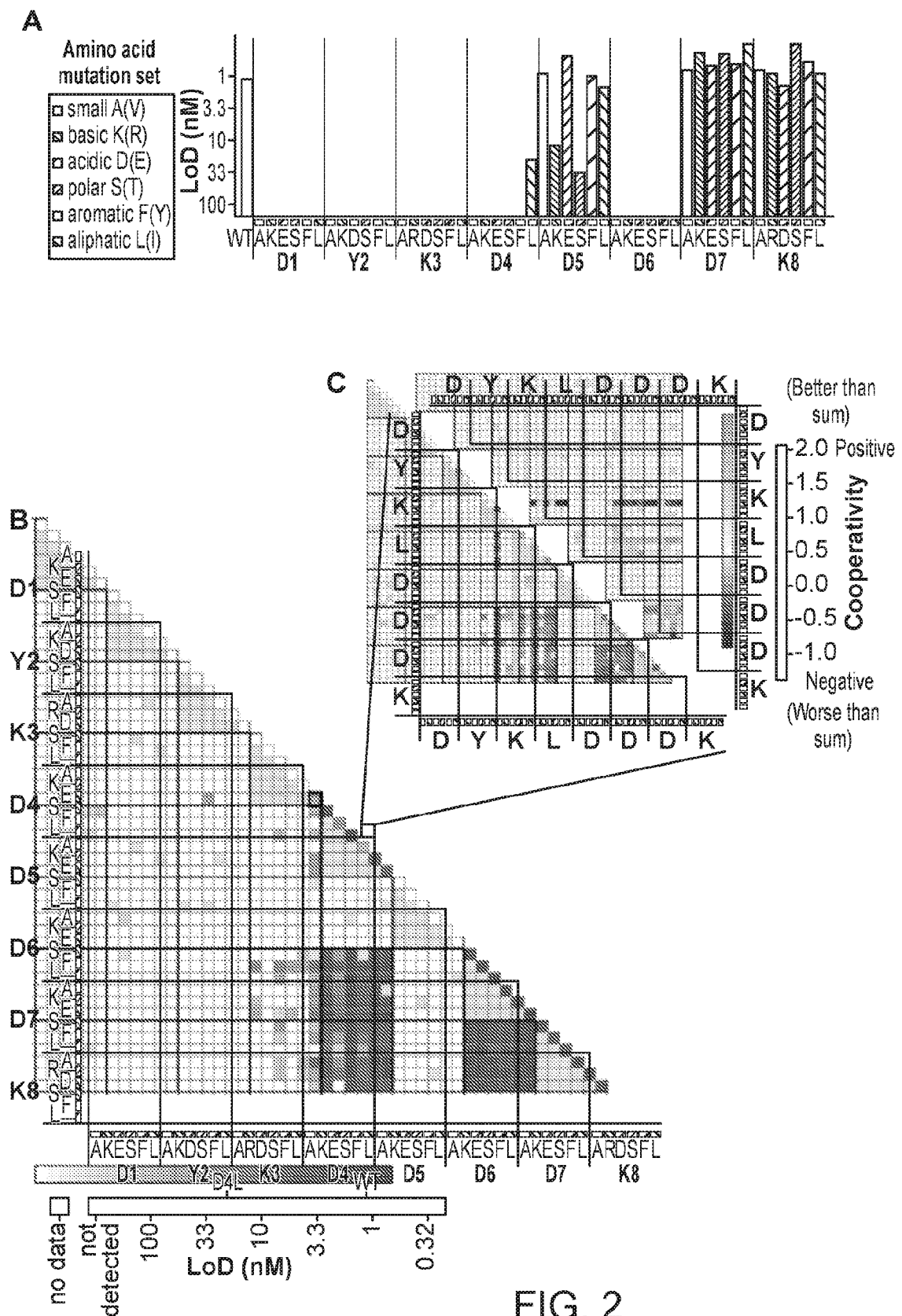


FIG. 2

4/9

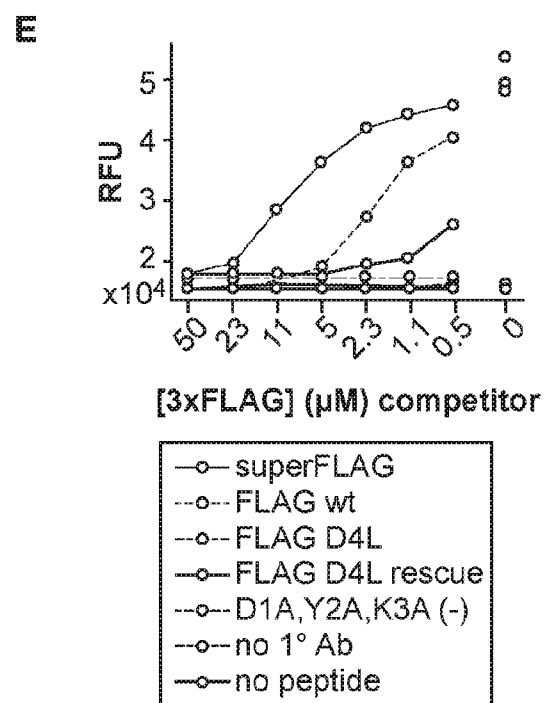
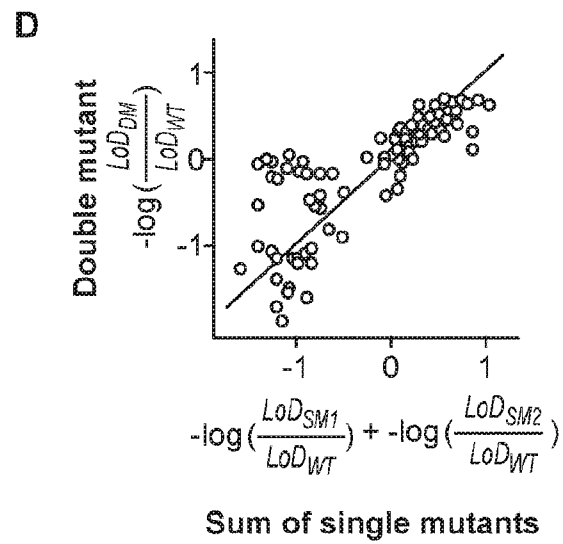


FIG. 2 (Cont.)

5/9

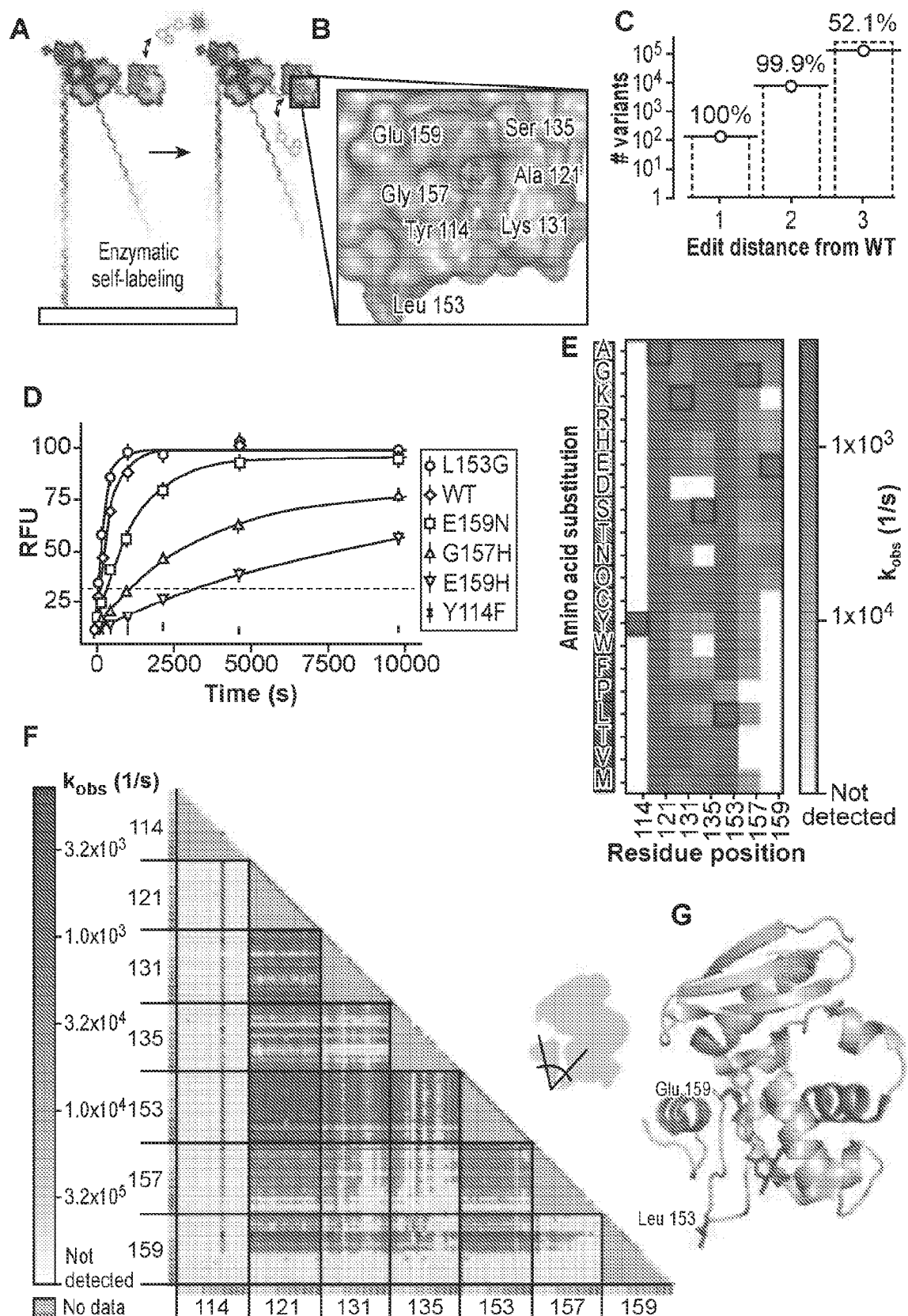


FIG. 3

SUBSTITUTE SHEET (RULE 26)

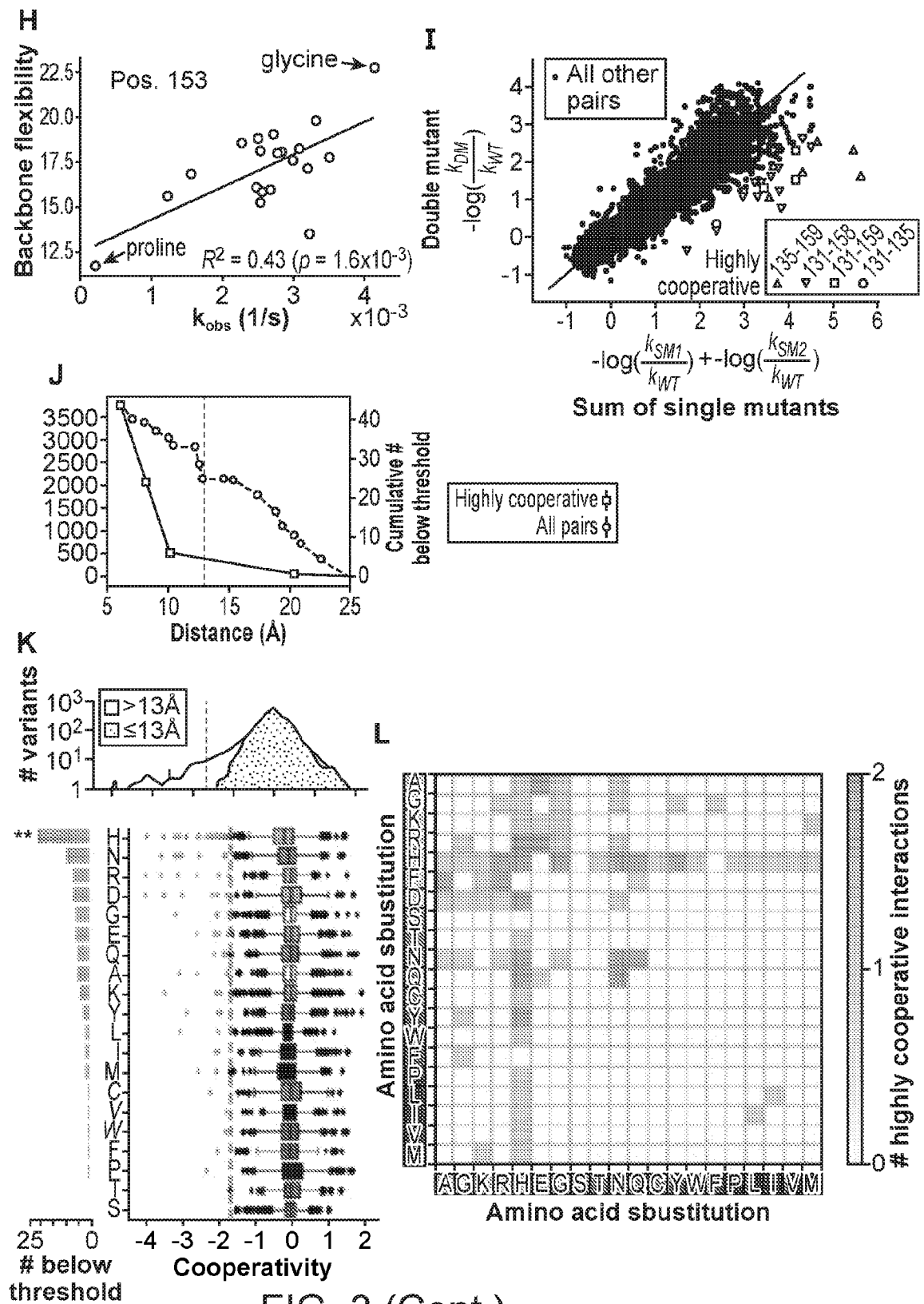
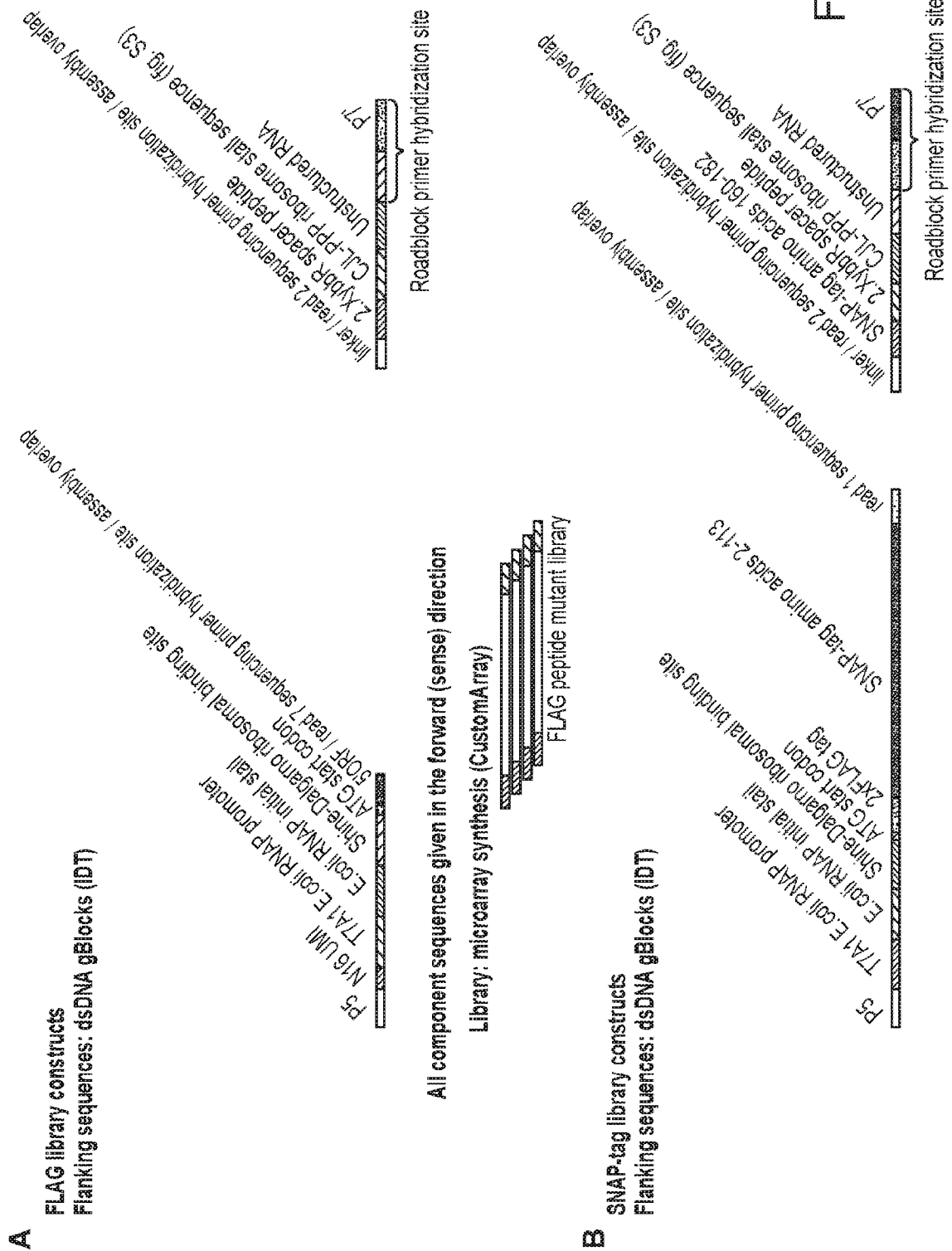


FIG. 3 (Cont.)


 KARNATAKA
 GOVT
 DEPT. OF
 TOURISM



SecM-WT																											
TTTACCCACAAAGCAAAATTTCAGCACGCCCGTCTGGATAAGCCAGGCGCAAGGCATCCGTGCTGGCCCTCAACGCCCTCACCTAA																											
F	T	P	Q	A	K	F	S	T	P	V	W	I	S	Q	A	Q	G	I	R	A	G	P	Q	R	L	T	STOP
TnaC-WT																											
AATATCTTACATATATGTGTGACCTCAAAATGGTTCAATATTGACAAACAAAAATTGTCGATCACCGCCCTTAGTTGCCCTTTAA																											
N	I	L	H	I	C	V	T	S	K	W	F	N	I	D	N	K	I	V	D	H	R	P	Am	F	A	L	STOP
CJL-PPP																											
AAAGAAGCAGCTGCAAAAGTTCAGCACGAAATGGTGGATAATTGACAAATGGCGCCACCGTCCGCCACCTTAGCGCCTCACC																											
K	E	A	A	A	K	F	S	T	K	W	I	I	D	K	W	R	H	R	P	P	P	P	Am	R	L	T	

FIG. 5

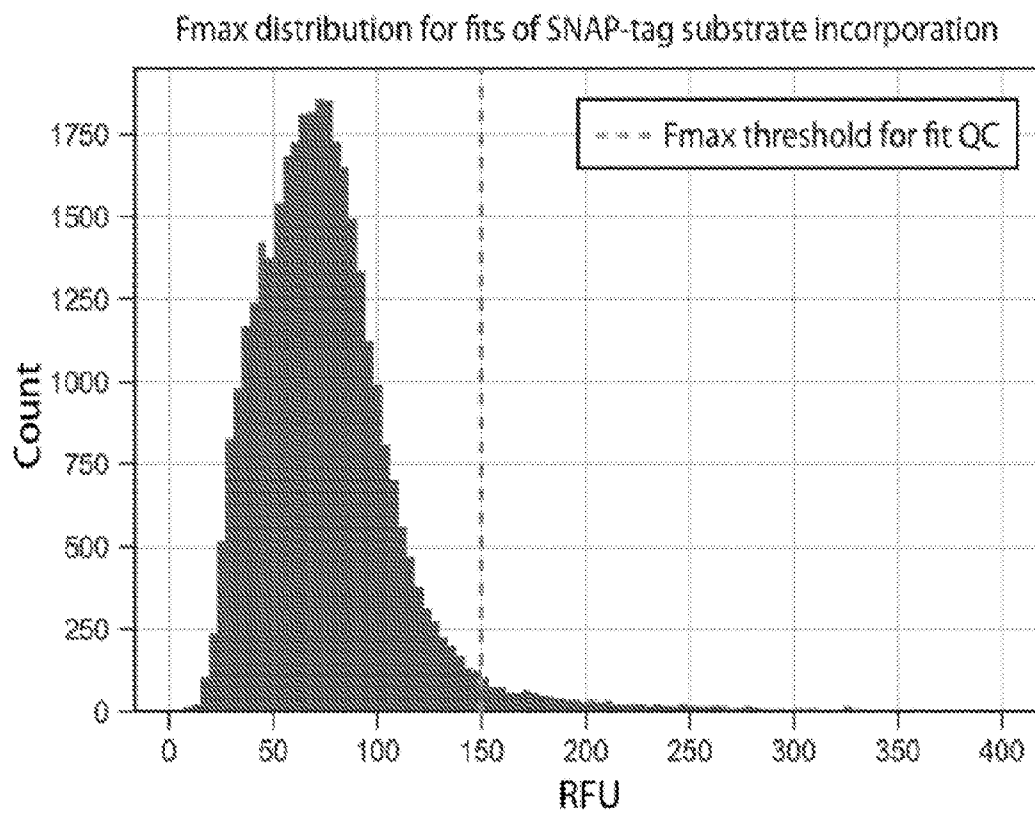


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/15033

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C07K 7/04; C40B 40/10 (2020.01)

CPC - C07K 7/04; C07K 2319/40; C40B 40/10

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LAYTON et al. Large-scale, quantitative protein assays on a high-throughput DNA sequencing chip. BioRxiv, 04 October 2018 [online]. [Retrieved 6 May 2020]. Retrieved from the internet: <URL: https://www.biorxiv.org/content/10.1101/342808v4 >. Especially abstract, PDF pg 10 fig 1.	1-4, 27-39
A	US 9,963,554 B2 (BOLT THREADS, INC.) 08 May 2018 (08.05.2018). Especially SEQ ID NO: 1409.	1-4, 27-39

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

06 May 2020

Date of mailing of the international search report

10 JUN 2020

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/15033

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
 - a. ☒ forming part of the international application as filed:
 - ☒ in the form of an Annex C/ST.25 text file.
 - ☐ on paper or in the form of an image file.
 - b. ☐ furnished together with the international application under PCT Rule 13ter.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
 - c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
 - ☐ in the form of an Annex C/ST.25 text file (Rule 13ter.1(a)).
 - ☐ on paper or in the form of an image file (Rule 13ter.1(b) and Administrative Instructions, Section 713).
2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

GenCore ver 6.4.1 SEQ ID NOs: 1 and 3 were searched.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/15033

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 5-26, 42-44
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
-----Go to Extra Sheet for continuation-----

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
Claims 1-4, 27-39, limited to SEQ ID NO: 3

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US 20/15033

Box III: Observations where Unity of Invention is lacking

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I+: Claims 1-4, 27-39, drawn to a variant FLAG peptide.

The variant FLAG peptide will be searched to the extent that the amino acid sequence is the first named, SEQ ID NO: 3. It is believed that claims 1-4, 27-39 read on this first named invention and thus these claims will be searched without fee to the extent that they encompass SEQ ID NOs: 3. Additional flag variant peptides will be searched upon payment of additional fees. Applicant must specify the claims that encompass any additional elected peptide variants. Applicants must further indicate, if applicable, the claims which read on the first named invention if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined. An exemplary election would be: SEQ ID NO: 4 (claims 1, (3,4)(in part), 27-39).

Group II: Claims 40-41, drawn to a method of screening.

The inventions listed as Groups I+ and II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Technical Features:

Group I+ has the special technical feature of a variant FLAG peptide amino acid sequence, not required by Group II.

Group II has the special technical feature of a method involving the specific step of performing an in vitro transcription /translation procedure, not required by Group I+.

No technical features are shared between the peptide sequences of Group I+ and accordingly, this group lacks unity a priori.

Additionally, even if Groups I+ inventions were considered to share the technical features of claim 1, this shared technical feature is previously disclosed by the publication titled "Large-scale, quantitative protein assays on a high-throughput DNA sequencing chip" by Layton et al. (hereinafter "Layton") [published in BioRxiv 4 October 2018 and available on the internet: <URL: <https://www.biorxiv.org/content/10.1101/342808v4> >].

As to claim 1, discloses a variant FLAG peptide comprising an amino acid sequence wherein the variant FLAG peptide is capable of binding to an anti-FLAG antibody with greater affinity than a FLAG peptide (abstract; "We quantified the binding of the M2 anti-FLAG antibody to a library of 1.3x10⁴ variant FLAG peptides, discovering non-additive effects of combinations of mutations, as well as a "superFLAG"; PDF pg 10 fig 1; "Quantified fluorescence medians that rise above a background threshold (grey dashed line) are extrapolated (solid) or interpolated (dashed) to estimate the limit of detection (LoD, open squares) for each FLAG variant, including WT FLAG (DYKDDDDK) and negative control (AAADDDDK) (top) as well as superFLAG variant, which, like 2xFLAG (DYKDDDDKGDYKDHD), gives much higher signal than WT (bottom, note scale). Open symbols indicate points below background, closed symbols are above. epitope variant") comprising the amino acid sequence of SEQ ID NO:1 (PDF pg 10 fig 1; "WT FLAG (DYKDDDDK)").

As the shared technical feature was known in the art at the time of the invention, this cannot be considered a shared special technical feature that would otherwise unify the groups. The inventions lack unity with one another.

Therefore, Groups I+ and II lack unity of invention under PCT Rule 13 because they do not share a same or corresponding special technical feature.

Item 4 (cont.): Claims 5-26, 42-44 are multiple dependent claims and are not drafted according to the second and third sentences of PCT Rule 6.4(a).