



(51) International Patent Classification:

C12Q 1/6869 (2018.01) *G16B 20/00* (2019.01)
C12N 15/10 (2006.01) *G16B 20/20* (2019.01)
G16B 10/00 (2019.01) *G16B 40/00* (2019.01)

(21) International Application Number:

PCT/US2021/050182

(22) International Filing Date:

14 September 2021 (14.09.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/078,148 14 September 2020 (14.09.2020) US

(71) Applicants: CHAN ZUCKERBERG BIOHUB, INC.

[US/US]; 499 Illinois Street, San Francisco, California
94158 (US). THE BOARD OF TRUSTEES OF THE LE-

LAND STANFORD JUNIOR UNIVERSITY [US/US];
Office of the General Counsel, Bldg. 170, Third Floor, Main
Quad, P. O. Box 20386, Stanford, California 94305-2038
(US).

(72) Inventors: MONTERRAT, Daniel Mas; c/o The Board
of Trustees of the Leland Stanford Junior University, Of-
fice of the General Counsel, Bldg. 170, Third Floor, Main
Quad, P.O. Box 20386, Stanford, California 94305-2038
(US). IOANNIDIS, Alexander; c/o The Board of Trustees
of the Leland Stanford Junior University, Office of the Gen-
eral Counsel, Bldg. 170, Third Floor, Main Quad, P.O. Box
20386, Stanford, California 94305-2038 (US). BUSTA-
MANTE, Carlos; c/o The Board of Trustees of the Leland
Stanford Junior University, Office of the General Counsel,
Bldg. 170, Third Floor, Main Quad, P.O. Box 20386, Stan-
ford, California 94305-2038 (US).

(54) Title: GENOMIC SEQUENCE DATASET GENERATION

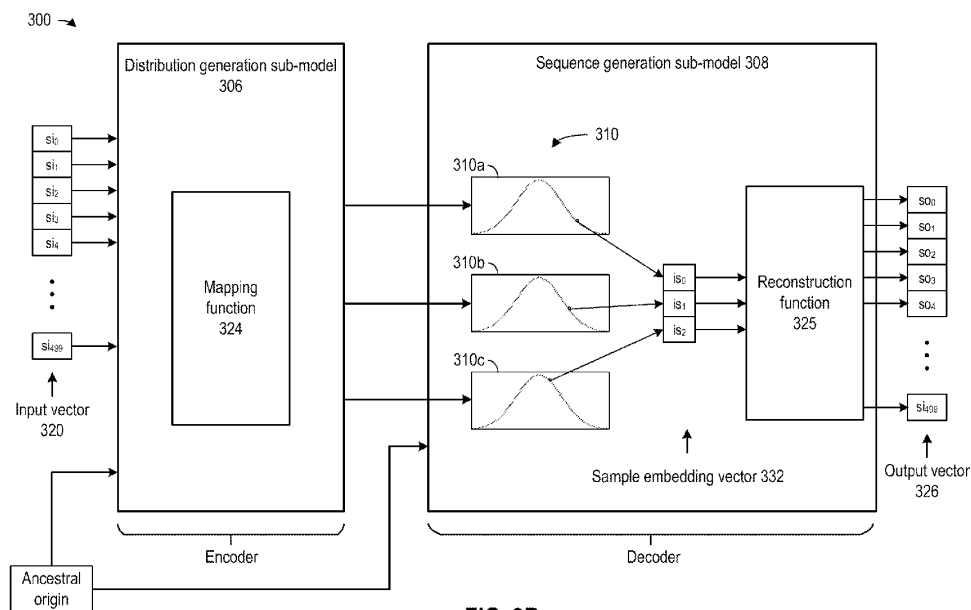


FIG. 3B

(57) Abstract: In one example, a method comprises: receiving a trait indicator; obtaining, based on the trait indicator, a probability distribution of embedding vectors in a latent space, the probability distribution being generated by a distribution generation sub-model of a trained generative machine learning model from an input vector representing a variant segment associated with the trait indicator, the input vector being defined in a variant segment space having a larger number of dimensions than the latent space; obtaining a sample vector by sampling the probability distribution; reconstructing, by a sequence generation sub-model of the trained generative machine learning model and based on the trait indicator, an output vector from the sample vector, the output vector being defined in the variant segment space; and generating a simulated genome sequence based on the output vector.

(74) **Agent: RACZKOWSKI, David B.**; Kilpatrick Townsend & Stockton LLP, Mailstop: IP Docketing - 22, 1100 Peachtree Street, Suite 2800, Atlanta, Georgia 30309 (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, IT, 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).

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))