



(51) International Patent Classification:
C12Q 1/68 (2018.01)

(21) International Application Number:
PCT/IB2021/055058

(22) International Filing Date:
09 June 2021 (09.06.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16/918,012 01 July 2020 (01.07.2020) US

(71) Applicant: **INTERNATIONAL BUSINESS
MACHINES CORPORATION** [US/US]; New Orchard
Road, Armonk, New York, New York 10504 (US).

(71) Applicants (for MG only): **IBM ISRAEL SCIENCE AND
TECHNOLOGY LTD.** [IL/IL]; Haifa University Cam-
pus, 165 Aba Khoushy Ave, Mount Carmel, 3498825 Haifa
(IL). **IBM (CHINA) INVESTMENT COMPANY LTD.**
[CN/CN]; 25/F, Pangu Plaza, No.27, Central North 4th Ring
Road, Chaoyang District, Beijing 100101 (CN).

(72) Inventors: **MAPIYE, Darlington Shingirirai**; IBM -
WITS RESEARCH LAB, 47Juta Street, Braamfontein,
2001 Johannesburg (ZA). **MASHIYANE, James Junior**;
IBM - WITS RESEARCH LAB, 47Juta Street, Braam-
fontein, 2001 Johannesburg (ZA). **MULLER, Stephanie**;
IBM - WITS RESEARCH LAB, 47Juta Street, Braam-
fontein, 2001 Johannesburg (ZA). **MOKOATLE, Mpho**;
IBM - WITS RESEARCH LAB, 47Juta Street, Braam-
fontein, 2001 Johannesburg (ZA). **DLAMINI, Gciniwe**;
IBM - WITS RESEARCH LAB, 47Juta Street, Braam-
fontein, 2001 Johannesburg (ZA).

(54) Title: DOWNSAMPLING GENOMIC SEQUENCE DATA

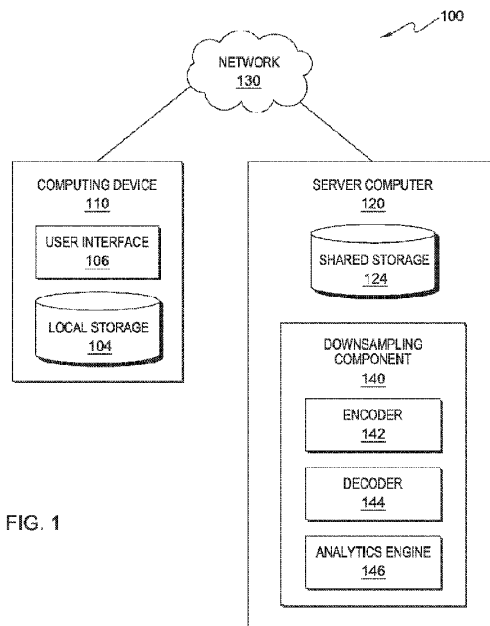


FIG. 1

(57) **Abstract:** A method for automatically downsampling DNA sequence data using variational autoencoders and preserving genomic integrity of an original file is disclosed. The method comprises: executing, by an encoder(142), bootstrapping on genomic sequence data to produce resamples; assessing, by the encoder(142), unrepresentativeness and self-inconsistency of the resamples and selecting a representative resample according to the assessment; building, by a modified encoder(142), vector representations from genotype likelihoods based on the selected representative sample; integrating, by an analytics engine, mapping positional information and the genotype likelihoods to identify an optimum vector representation of a resample; and decoding, by a modified decoder(144), the identified optimum vector representation of the resample to obtain a down-sampled read file that resembles and maintains the genomic integrity of the original file.

[Continued on next page]

(74) **Agent: GILBOA, Eyal;** IBM ISRAEL SCIENCE AND TECHNOLOGY LTD., Haifa University Campus, 165 Aba Khoushy Ave, Mount Carmel, 3498825 Haifa (IL).

(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))