

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
07 May 2020 (07.05.2020)



(10) International Publication Number
WO 2020/092309 A1

- (51) **International Patent Classification:**
G16B 30/20 (2019.01)
- (21) **International Application Number:**
PCT/US2019/058476
- (22) **International Filing Date:**
29 October 2019 (29.10.2019)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/753,786 31 October 2018 (31.10.2018) US
- (71) **Applicant:** **ILLUMINA, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US).
- (72) **Inventors:** **ZHAO, Chen**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US). **WU, Kevin Eric**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US).
- (74) **Agent:** **FULLER, Michael L.**; 2040 Main Street, Fourteenth Floor, Irvine, California 92614 (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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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,

(54) **Title:** SYSTEMS AND METHODS FOR GROUPING AND COLLAPSING SEQUENCING READS

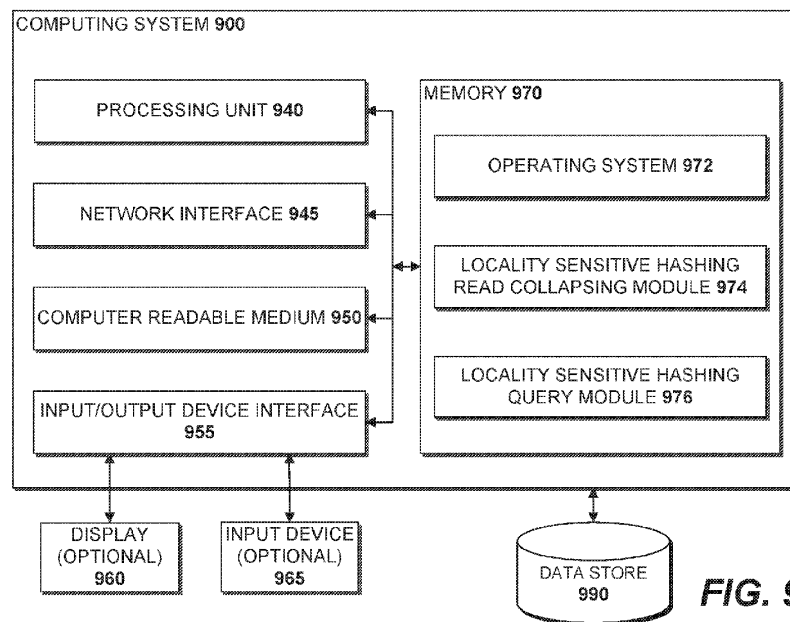


FIG. 9

(57) **Abstract:** Disclosed herein are systems and methods for collapsing sequencing reads and identifying similar sequencing reads. In one example, a method includes generating a plurality of first identifier subsequences from a first identifier sequence of each nucleotide sequencing read and generating a first signature for the nucleotide sequencing read by applying hashing to the plurality of first identifier subsequences. The method may include assigning the nucleotide sequencing read to a first particular bin of a first data structure based on the first signature and determining a nucleotide sequence for each first particular bin of the first data structure with one or more nucleotide sequencing reads assigned.

[Continued on next page]

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).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*