



- (51) International Patent Classification:
G06F 19/18 (2011.01) *C12Q 1/68* (2006.01)
- (21) International Application Number:
PCT/US2012/059123
- (22) International Filing Date:
5 October 2012 (05.10.2012)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
61/544,251 6 October 2011 (06.10.2011) US
61/663,477 22 June 2012 (22.06.2012) US
61/709,899 4 October 2012 (04.10.2012) US
- (71) Applicant: **SEQUENOM, INC.** [US/US]; 3595 John Hopkins Court, San Diego, CA 92121 (US).
- (72) Inventors: **DECIU, Cosmin**; 10545 Sea Mist Way, San Diego, CA 92121 (US). **DZAKULA, Zeljko**; 12830 Sundance Avenue, San Diego, CA 92129 (US). **EHRICH, Mathias**; 1434 Law Street, San Diego, CA 92109 (US). **KIM, Sung, Kyun**; 662 Glenmore Boulevard, Glendale, CA 91206 (US).
- (74) Agents: **FORCE, Walker, R.** et al.; Grant Anderson LLP, c/o PortfolioIP, P.O. Box 52050, Minneapolis, MN 55402 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- with amended claims (Art. 19(1))

- (88) Date of publication of the international search report:
7 November 2013

Date of publication of the amended claims: 27 December 2013



WO 2013/052913 A4

(54) Title: METHODS AND PROCESSES FOR NON-INVASIVE ASSESSMENT OF GENETIC VARIATIONS

(57) Abstract: Provided herein are methods, processes and apparatuses for non-invasive assessment of genetic variations.

received by the International Bureau on 25 October 2013 (25.10.2013)

1. A system comprising memory and one or more microprocessors, which one or more microprocessors are configured to perform, according to instructions in the memory, a process for calculating with reduced bias genomic section levels for a test sample, which process comprises:
 - (a) obtaining counts of sequence reads mapped to portions of a reference genome, which sequence reads are reads of circulating cell-free nucleic acid from a test sample;
 - (b) determining a guanine and cytosine (GC) bias coefficient for the test sample based on a fitted relation between (i) the counts of the sequence reads mapped to each of the portions and (ii) GC content for each of the portions; and
 - (c) calculating a genomic section level for each of the portions based on the counts of (a), the GC bias coefficient of (b) and a fitted relation, for each of the portions, between (i) the GC bias coefficient for each of multiple samples and (ii) the counts of the sequence reads mapped to each of the portions for the multiple samples, thereby providing calculated genomic section levels, whereby bias in the counts of the sequence reads mapped to each of the portions of the reference genome is reduced in the calculated genomic section levels.
2. The system of claim 1, wherein the GC bias coefficient is a slope for a linear fitted relation or a curvature estimation for a non-linear fitted relation.
3. The system of claim 1 or 2, wherein the fitted relation of (b) and the fitted relation of (c) are linear.
4. The system of any one of claims 1 to 3, wherein each or the fitted relation of (b) and the fitted relation of (c) independently are fitted by a linear regression.
5. The system of any one of claims 1 to 4, wherein the GC bias coefficient for each of the multiple samples in (c)(i) is the slope of a fitted linear relation, for each of the multiple samples, between (i) the counts of the sequence reads mapped to each of the portions and (ii) GC content for each of the portions.

6. The system of any one of claims 1 to 5, wherein the calculated genomic section level L is determined for the test sample for each portion of the reference genome according to Equation B:

$$L = (M - GS)/I \quad \text{Equation B}$$

wherein M is the counts of the sequence reads mapped to the portion for the test sample, G is the GC bias coefficient for the test sample, I is an intercept of the fitted linear relation of (c) for the portion, S is a slope of the fitted linear relationship of (c) for the portion.

7. The system of claim 1 or 2, wherein the fitted relation of (b) is non-linear.

8. The system of any one of claims 1 to 7, wherein each of the portions of the reference genome comprises a nucleotide sequence of a predetermined length.

9. The system of any one of claims 1 to 8, which process comprises, prior to (a), determining the sequence reads by sequencing circulating cell-free nucleic acid from the test sample.

10. The system of any one of claims 1 to 9, which process comprises, prior to (a), mapping the sequence reads to the portions of the reference genome.

11. The system of any one of claims 1 to 10, wherein the test sample is from a human pregnant female and which process comprises determining the presence or absence of a fetal chromosome aneuploidy for the test sample according to the calculated genomic section levels.

12. The system of claim 11, wherein the fetal chromosome aneuploidy is a trisomy.

13. The system of claim 12, wherein the trisomy is chosen from a trisomy of chromosome 21, chromosome 18, chromosome 13 or combination thereof.

14. The system of claim 12 or 13, wherein the presence or absence of the trisomy is determined with a sensitivity of 96% or greater or a specificity of 96% or greater, or a sensitivity of 96% or greater and a specificity of 96% or greater.
15. The system of any one of claims 1 to 14, which process comprises, prior to (b), calculating a measure of error for the counts of sequence reads mapped to some or all of the portions of the reference genome and removing or weighting the counts of sequence reads for certain portions of the reference genome according to a threshold of the measure of error.
16. The system of claim 15, wherein the threshold is selected according to a standard deviation gap between a first genomic section level and a second genomic section level of 3.5 or greater.
17. The system of claim 15 or 16, wherein the measure of error is an R factor.
18. The system of claim 17, wherein the counts of sequence reads for a portion of the reference genome having an R factor of about 7% or greater are removed prior to (b).
19. A method for calculating with reduced bias genomic section levels for a test sample, comprising:
- (a) obtaining counts of sequence reads mapped to portions of a reference genome, which sequence reads are reads of circulating cell-free nucleic acid from a test sample;
 - (b) determining a guanine and cytosine (GC) bias coefficient for the test sample based on a fitted relation between (i) the counts of the sequence reads mapped to each of the portions and (ii) GC content for each of the portions; and
 - (c) calculating, using a microprocessor, a genomic section level for each of the portions based on the counts of (a), the GC bias coefficient of (b) and a fitted relation, for each of the portions, between (i) the GC bias coefficient for each of multiple samples and (ii) the counts of the sequence reads mapped to each of the portions for the multiple samples, thereby providing calculated genomic section levels, whereby bias in the counts

of the sequence reads mapped to each of the portions of the reference genome is reduced in the calculated genomic section levels.

20. The method of claim 19, wherein the GC bias coefficient is a slope for a linear fitted relation or a curvature estimation for a non-linear fitted relation.

21. The method of claim 19 or 20, wherein the fitted relation of (b) and the fitted relation of (c) are linear.

22. The method of any one of claims 19 to 21, wherein each or the fitted relation of (b) and the fitted relation of (c) independently are fitted by a linear regression.

23. The method of any one of claims 19 to 22, wherein the GC bias coefficient for each of the multiple samples in (c)(i) is the slope of a fitted linear relation, for each of the multiple samples, between (i) the counts of the sequence reads mapped to each of the portions and (ii) GC content for each of the portions.

24. The method of any one of claims 19 to 23, wherein the calculated genomic section level L is determined for the test sample for each portion of the reference genome according to Equation B:

$$L = (M - GS)/I$$

Equation B

wherein M is the counts of the sequence reads mapped to the portion for the test sample, G is the GC bias coefficient for the test sample, I is an intercept of the fitted linear relation of (c) for the portion, S is a slope of the fitted linear relationship of (c) for the portion.

25. The method of claim 19 or 20, wherein the fitted relation of (b) is non-linear.

26. The method of any one of claims 19 to 25, wherein each of the portions of the reference genome comprises a nucleotide sequence of a predetermined length.

27. The method of any one of claims 19 to 26, comprising, prior to (a), determining the sequence reads by sequencing circulating cell-free nucleic acid from the test sample.
28. The method of any one of claims 19 to 27, comprising, prior to (a), mapping the sequence reads to the portions of the reference genome.
29. The method of any one of claims 19 to 28, wherein the test sample is from a human pregnant female and which method comprises determining the presence or absence of a fetal chromosome aneuploidy for the test sample according to the calculated genomic section levels.
30. The method of claim 29, wherein the fetal chromosome aneuploidy is a trisomy.
31. The method of claim 30, wherein the trisomy is chosen from a trisomy of chromosome 21, chromosome 18, chromosome 13 or combination thereof.
32. The method of claim 30 or 31, wherein the presence or absence of the trisomy is determined with a sensitivity of 96% or greater or a specificity of 96% or greater, or a sensitivity of 96% or greater and a specificity of 96% or greater.
33. The method of any one of claims 19 to 32, which comprises, prior to (b), calculating a measure of error for the counts of sequence reads mapped to some or all of the portions of the reference genome and removing or weighting the counts of sequence reads for certain portions of the reference genome according to a threshold of the measure of error.
34. The method of claim 33, wherein the threshold is selected according to a standard deviation gap between a first genomic section level and a second genomic section level of 3.5 or greater.
35. The method of claim 33 or 34, wherein the measure of error is an R factor.

36. The method of claim 35, wherein the counts of sequence reads for a portion of the reference genome having an R factor of about 7% or greater are removed prior to (b).

37. A system comprising a sequencing apparatus and one or more computing apparatus, which sequencing apparatus is configured to produce signals corresponding to nucleotide bases of a nucleic acid loaded in the sequencing apparatus, which nucleic acid is circulating cell-free nucleic acid from a test sample from a pregnant human female bearing a fetus, or which circulating cell-free nucleic acid nucleic acid loaded in the sequencing apparatus is processed or modified; and

which one or more computing apparatus comprise memory and one or more processors, which memory comprises instructions executable by the one or more processors and which instructions executable by the one or more processors are configured to:

(a) produce sequence reads from the signals and map the sequence reads to a reference genome;

(b) obtain counts of sequence reads mapped to the portions of the reference genome;

(c) determine a guanine and cytosine (GC) bias coefficient for the test sample based on a fitted relation between (i) the counts of the sequence reads mapped to each of the portions and (ii) GC content for each of the portions; and

(d) calculate a genomic section level for each of the portions based on the counts of (b), the GC bias coefficient of (c) and a fitted relation, for each of the portions, between (i) the GC bias coefficient for each of multiple samples and (ii) the counts of the sequence reads mapped to each of the portions for the multiple samples, thereby providing calculated genomic section levels, whereby bias in the counts of the sequence reads mapped to each of the portions of the reference genome is reduced in the calculated genomic section levels.

38. The system of claim 37, wherein the GC bias coefficient is a slope for a linear fitted relation or a curvature estimation for a non-linear fitted relation.

39. The system of claim 37 or 38, wherein the fitted relation of (c) and the fitted relation of (d) are linear.

40. The system of any one of claims 37 to 39, wherein each of the fitted relation of (c) and the fitted relation of (d) independently are fitted by a linear regression.

41. The system of any one of claims 37 to 40, wherein the GC bias coefficient for each of the multiple samples in (d)(i) is the slope of a fitted linear relation, for each of the multiple samples, between (i) the counts of the sequence reads mapped to each of the portions and (ii) GC content for each of the portions.

42. The system of any one of claims 37 to 41, wherein the calculated genomic section level L is determined for the test sample for each portion of the reference genome according to Equation B:

$$L = (M - GS)/I$$

Equation B

wherein M is the counts of the sequence reads mapped to the portion for the test sample, G is the GC bias coefficient for the test sample, I is an intercept of the fitted linear relation of (d) for the portion, S is a slope of the fitted linear relationship of (d) for the portion.

43. The system of claim 37 or 38, wherein the fitted relation of (c) is non-linear.

44. The system of any one of claims 37 to 43, wherein each of the portions of the reference genome comprises a nucleotide sequence of a predetermined length.

45. The system of any one of claims 37 to 44, which memory comprises instructions configured to determine the presence or absence of a fetal chromosome aneuploidy for the test sample according to the calculated genomic section levels.

46. The system of claim 45, wherein the fetal chromosome aneuploidy is a trisomy.

47. The system of claim 46, wherein the trisomy is chosen from a trisomy of chromosome 21, chromosome 18, chromosome 13 or combination thereof.

48. The system of claim 46 or 47, wherein the presence or absence of the trisomy is determined with a sensitivity of 96% or greater or a specificity of 96% or greater, or a sensitivity of 96% or greater and a specificity of 96% or greater.

49. The system of any one of claims 37 to 48, which memory comprises instructions configured to, prior to (c), calculate a measure of error for the counts of sequence reads mapped to some or all of the portions of the reference genome and remove or weight the counts of sequence reads for certain portions of the reference genome according to a threshold of the measure of error.

50. The system of claim 49, wherein the threshold is selected according to a standard deviation gap between a first genomic section level and a second genomic section level of 3.5 or greater.

51. The system of claim 49 or 50, wherein the measure of error is an R factor.

52. The system of claim 51, wherein the counts of sequence reads for a portion of the reference genome having an R factor of about 7% or greater are removed prior to (c).