

CLAIMS

What is claimed is:

1. A system for generating a computer graphic representing sequence reads aligned to haplotypes of a genomic region, the system comprising one or more processors and system memory, wherein the one or more processors are configured to:

(a) align a plurality of sequence reads to a set of alignment positions on a plurality of haplotype sequences corresponding to a plurality of haplotypes of the genomic region, wherein the plurality of sequence reads is obtained from a genomic region of a nucleic acid sample;

(b) estimate an alignment score for the set of alignment positions;

(c) repeat (a)-(b) for multiple iterations to obtain a plurality of alignment scores for a plurality of different sets of alignment positions;

(d) select a set of alignment positions from the plurality of different sets of alignment positions based on the plurality of alignment scores; and

(e) generate a computer graphic representing the plurality of sequence reads and the plurality of haplotypes, wherein the plurality of sequence reads is aligned to the plurality of haplotypes at the set of alignment positions selected in (d).

2. The system of claim 1, wherein the alignment score indicates how evenly the plurality of sequence reads is distributed on the plurality of haplotype sequences.

3. The system of any one of claims 1-2, wherein the genomic region comprises one or more tandem repeats.

4. The system of any one of claims 1-3, wherein at least one haplotype of the plurality of haplotypes comprises a repeat expansion.

5. The system of any one of claims 1-4, wherein each haplotype comprises an allele.

6. The system of any one of claims 1-5, wherein the plurality of haplotypes comprises two haplotypes.

7. The system of any one of claims 1-6, wherein the selected set of alignment positions has a best alignment score among the plurality sets of different alignment positions.

8. The system of any one of claims 1-7, wherein the selected set of alignment positions has an alignment score exceeding a selection criterion.

9. The system of any one of claims 1-8, wherein at least one haplotype of the plurality of haplotypes comprises a structural variant.

10. The system of claim 9, wherein the structural variant is longer than 50 bp and is selected from the group consisting of: deletions, duplications, copy-number variants, insertions, inversions, translocations, and any combinations thereof.

11. The system of claim 9, wherein the structural variant comprises a variant shorter than 50 bp.

12. The system of claim 11, wherein the variant shorter than 50 bp comprises a single nucleotide polymorphism (SNP).

13. The system of any one of claims 1-12, wherein (a) comprises:

(i) determining possible alignment positions of each read to each haplotype, wherein the plurality of sequence reads comprises read pairs obtained by paired-end sequencing;

(ii) creating constrained alignment positions for each read pair from alignment positions of constituent reads in such a way that (A) both reads of the read pair align to the same haplotype, and (B) the corresponding fragment length of the read pair is as close as possible to a mean fragment length; and

(iii) randomly choosing an alignment position for each read pair from the constrained alignment positions.

14. The system of any one of claims 1-13, wherein the alignment score comprises a root mean squared difference from the mean of distance between starting positions of two consecutive reads.

15. The system of any one of claims 1-14, wherein the alignment score is estimated using a probabilistic model assuming read pairs are uniformly distributed on the plurality of haplotype sequences.

16. The system of claim 15, wherein the alignment score comprises a probability of the plurality of sequence reads being derived from the set of alignment positions given the probabilistic model.

17. The system of claim 16, wherein the plurality of sequence reads comprises pair-end reads obtained from nucleic acid fragments and the probabilistic model is configured to receive a mean fragment length as an input.

18. The system of claim 16, wherein the probabilistic model is configured to receive a length of haplotype as an input.

19. The system of claim 16, wherein a probability of an individual alignment position x of the k^{th} read pair from the beginning of the haplotype, denoted by $p_{(k)}$ is modeled as:

$$\mathcal{P}(p_{(k)} = x) = \sum_{j=0}^{n_i-k} \binom{n_i}{j} \left((1 - F(x))^j (F(x))^{n_i-j} - (1 - F(x) + f(x))^j (F(x) - f(x))^{n_i-j} \right)$$

wherein

i is the haplotype to which the read pair was aligned,

$$F(x) = \frac{x}{H_i - L - 1},$$

$$f(x) = \frac{1}{H_i - L - 1},$$

H_i is the length of haplotype i ,

L is the mean fragment length, and

n_i is the number of read pairs aligned to haplotype i .

20. The system of claim 19, wherein the alignment score for the set of alignment positions is estimated as a product of probabilities of individual alignment positions.

21. The system of any one of claims 1-2, wherein the one or more processors are further configured to estimate one or more sequencing metrics for the plurality of sequence reads aligned to the plurality of haplotype sequences at the set of selected alignment positions.

22. The system of claim 21, wherein the one or more sequencing metrics comprise a sequence coverage.

23. The system of claim 22, wherein the one or more sequencing metrics comprise a sequence coverage for each alignment position.

24. The system of claim 21, wherein the one or more sequencing metrics comprise an alignment quality score.

25. The system of claim 24, wherein the one or more sequencing metrics comprise an alignment quality score for each alignment position.

26. The system of claim 21, wherein the one or more sequencing metrics comprise a mapping quality score.

27. The system of any one of claims 1-26, wherein the plurality of sequence reads comprises at least 100 sequence reads.

28. The system of any one of claims 1-27, wherein the one or more processors are configured to perform operation (a) for different genomic regions using different sets of sequence reads.

29. The system of claim 28, wherein the different genomic regions comprise at least 100 different genomic regions.

30. The system of any one of claims 1-29, wherein the one or more processors are configured to align, before operation (a), a first number of sequence reads to one or more sequence graphs corresponding to the genomic region to obtain the plurality of sequence reads and/or the plurality of haplotypes.

31. The system of claim 30, wherein aligning the first number of sequence reads to the sequence graph comprises:

providing the first number of sequence reads of the nucleic acid sample;

aligning the first number of sequence reads to one or more repeat sequences each represented by a sequence graph, wherein the sequence graph has a data structure of a directed graph with vertices representing nucleic acid sequences and directed edges connecting the vertices, and wherein the sequence graph comprises one or more self-loops, each self-loop representing a repeat sub-sequence, each repeat sub-sequence comprising repeats of a repeat unit of one or more nucleotides;

determining one or more genotypes of the one or more repeat sequences; and

providing the first number of sequence reads as the plurality of sequence reads of (a) and/or the one or more genotypes of the one or more repeat sequences.

32. The system of claim 31, wherein the one or more processors are configured to phase the one or more genotypes of the one or more repeat sequences to determine the plurality of haplotypes of (b).

33. The system of claim 31, wherein the one or more processors are configured to initially align a second number of sequence reads to a genome to provide the first number of sequence reads, wherein the second number of sequence reads comprise at least 10,000 sequence reads.

34. A method, implemented using a computer comprising one or more processors and system memory, for generating computer graphics the method comprising:

(a) aligning, using the one or more processors, a plurality of sequence reads to a set of alignment positions on a plurality of haplotype sequences corresponding to a plurality of haplotypes of the genomic region, wherein the plurality of sequence reads is obtained from a genomic region of a nucleic acid sample;

(b) estimating, by the one or more processors, an alignment score for the set of alignment positions;

(c) repeating (a)-(b) for multiple iterations to obtain a plurality of alignment scores for a plurality of different sets of alignment positions;

(d) selecting, by the one or more processors, a set of alignment positions from the plurality of different sets of alignment positions based on the plurality of alignment scores; and

(e) generating, using the one or more processors, a computer graphic representing the plurality of sequence reads and the plurality of haplotypes, wherein the plurality of sequence reads is aligned to the plurality of haplotypes at the set of alignment positions selected in (d).

35. The method of claim 34, wherein the alignment score indicates how evenly the plurality of sequence reads is distributed on the plurality of haplotype sequences.

36. The method of claim 34, wherein the genomic region comprises one or more tandem repeats.

37. The method of any one of claims 34-36, wherein at least one haplotype of the plurality of haplotypes comprises a repeat expansion.

38. The method of any one of claims 34-37, wherein each haplotype comprises an allele.

39. The method of any one of claims 34-39, wherein the plurality of haplotypes comprises two haplotypes.

40. The method of any one of claims 34-39, wherein the selected set of alignment positions has a best alignment score among the plurality sets of different alignment positions.

41. The method of any one of claims 34-40, wherein the selected set of alignment positions has an alignment score exceeding a selection criterion.

42. The method of any one of claims 34-41, wherein at least one haplotype of the plurality of haplotypes comprises a structural variant.

43. The method of claim 42, wherein the structural variant is longer than 50 bp and is selected from the group consisting of: deletions, duplications, copy-number variants, insertions, inversions, translocations, and any combinations thereof.

44. The method of claim 42, wherein the structural variant comprises a variant shorter than 50 bp.

45. The method of claim 44, wherein the variant shorter than 50 bp comprises a single nucleotide polymorphism (SNP).

46. The method of any one of claims 34-45, wherein (a) comprises:

(i) determining possible alignment positions of each read to each haplotype, wherein the plurality of sequence reads comprises read pairs obtained by paired-end sequencing;

(ii) creating constrained alignment positions for each read pair from alignment positions of constituent reads in such a way that (A) both reads of the read pair align to the same haplotype, and (B) the corresponding fragment length of the read pair is as close as possible to a mean fragment length; and

(iii) randomly choosing an alignment position for each read pair from the constrained alignment positions.

47. The method of any one of claims 34-47, wherein the alignment score comprises a root mean squared difference from the mean of distance between starting positions of two consecutive reads.

48. The method of any one of claims 34-47, wherein the alignment score is estimated using a probabilistic model assuming read pairs are uniformly distributed on the plurality of haplotype sequences.

49. The method of claim 48, wherein the alignment score comprises a probability of the plurality of sequence reads being derived from the set of alignment positions given the probabilistic model.

50. The method of claim 49, wherein the plurality of sequence reads comprises pair-end reads obtained from nucleic acid fragments and the probabilistic model is configured to receive a mean fragment length as an input.

51. The method of claim 49, wherein the probabilistic model is configured to receive a length of haplotype as an input.

52. The method of claim 49, wherein a probability of an individual alignment position x of the k^{th} read pair from the beginning of the haplotype, denoted by $p_{(k)}$ is modeled as:

$$\mathcal{P}(p_{(k)} = x) = \sum_{j=0}^{n_i-k} \binom{n_i}{j} \left((1 - F(x))^j (F(x))^{n_i-j} - (1 - F(x) + f(x))^j (F(x) - f(x))^{n_i-j} \right)$$

wherein

i is the haplotype to which the read pair was aligned,

$$F(x) = \frac{x}{H_i - L - 1},$$

$$f(x) = \frac{1}{H_i - L - 1},$$

H_i is the length of haplotype i ,

L is the mean fragment length, and

n_i is the number of read pairs aligned to haplotype i .

53. The method of claim 52, wherein the alignment score for the set of alignment positions is estimated as a product of probabilities of individual alignment positions.

54. The method of any one of claims 34-53, further comprising estimating one or more sequencing metrics for the plurality of sequence reads aligned to the plurality of haplotype sequences at the set of selected alignment positions.

55. The method of claim 54, wherein the one or more sequencing metrics comprise a sequence coverage.

56. The method of claim 55, wherein the one or more sequencing metrics comprise a sequence coverage for each alignment position.

57. The method of claim 54, wherein the one or more sequencing metrics comprise an alignment quality score.

58. The method of claim 57, wherein the one or more sequencing metrics comprise an alignment quality score for each alignment position.

59. The method of claim 54, wherein the one or more sequencing metrics comprise a mapping quality score.

60. The method of claim 34, wherein the plurality of sequence reads comprises at least 100 sequence reads.

61. The method of any one of claims 34-61, further comprising performing operation (a) for different genomic regions using different sets of sequence reads.

62. The method of claim 61, wherein the different genomic regions comprise at least 100 different genomic regions.

63. The method of any one of claims 34-63, further comprising aligning, before operation (a), a first number of sequence reads to one or more sequence graphs corresponding to the genomic region to obtain the plurality of sequence reads and/or the plurality of haplotypes.

64. The method of claim 63, wherein aligning the first number of sequence reads to the sequence graph comprises:

providing the first number of sequence reads of the nucleic acid sample;

aligning the first number of sequence reads to one or more repeat sequences each represented by a sequence graph, wherein the sequence graph has a data structure of a directed graph with vertices representing nucleic acid sequences and directed edges connecting the vertices, and wherein the sequence graph comprises one or more self-loops, each self-loop representing a repeat sub-sequence, each repeat sub-sequence comprising repeats of a repeat unit of one or more nucleotides;

determining one or more genotypes of the one or more repeat sequences; and

providing the first number of sequence reads as the plurality of sequence reads of (a) and/or the one or more genotypes of the one or more repeat sequences.

65. The method of claim 64, further comprising phasing the one or more genotypes of the one or more repeat sequences to determine the plurality of haplotypes of (b).

66. The method of claim 64, further comprising initially aligning a second number of sequence reads to a genome to provide the first number of sequence reads, wherein the second number of sequence reads comprise at least 10,000 sequence reads.

67. A computer product comprising one or more computer-readable non-transitory storage media having stored thereon computer-executable instructions that, when executed by one or more processors of a computer system, cause the computer system to:

(a) align a plurality of sequence reads to a set of alignment positions on a plurality of haplotype sequences corresponding to a plurality of haplotypes of the genomic region, wherein the plurality of sequence reads is obtained from a genomic region of a nucleic acid sample;

(b) estimate an alignment score for the set of alignment positions;

(c) repeat (a)-(b) for multiple iterations to obtain a plurality of alignment scores for a plurality of different sets of alignment positions;

(d) select a set of alignment positions from the plurality of different sets of alignment positions based on the plurality of alignment scores; and

(e) generate a computer graphic representing the plurality of sequence reads and the plurality of haplotypes, wherein the plurality of sequence reads is aligned to the plurality of haplotypes at the set of alignment positions selected in (d).

68. A method for generating computer graphics, implemented using a system comprising one or more processors and system memory, the method comprising:

(a) determining a plurality of sets of alignment positions of a plurality of sequence reads aligned to a plurality of haplotype sequences corresponding to a plurality of haplotypes of a genomic region, wherein the plurality of sequence reads is obtained from the genomic region of a nucleic acid sample;

(b) selecting a set of alignment positions that is more evenly distributed on the plurality of haplotypes than other sets of alignment positions in the plurality of sets of alignment positions; and

(c) generating a computer graphic representing the plurality of sequence reads and the plurality of haplotypes, wherein the plurality of sequence reads is aligned to the plurality of haplotypes at the set of alignment positions selected in (b)

69. The method of claim 68, wherein the plurality of sequence reads comprise pair-end reads obtained from nucleic acid fragments.

70. The method of any one of claims 68-69, wherein the plurality of haplotypes are two haplotypes.

71. The method of any one of claims 68-70, the genomic region comprises one or more tandem repeats.

72. The method of claim 71, at least one haplotype of the plurality of haplotypes comprises a repeat expansion.

73. The method of any one of claims 68-72, at least one haplotype of the plurality of haplotypes comprises a structural variant.

74. The method of claim 73, the structural variant is longer than 50 bp and is selected from the group consisting of: deletions, duplications, copy-number variants, insertions, inversions, translocations, and any combinations thereof.

75. A method for aligning sequence reads to a genomic region, implemented using a system comprising one or more processors and system memory, the method comprising:

(a) aligning a plurality of sequence reads to a set of alignment positions on a plurality of haplotype sequences corresponding to a plurality of haplotypes of the genomic region, wherein the plurality of sequence reads is obtained from a genomic region of a nucleic acid sample;

(b) estimating an alignment score for the set of alignment positions;

(c) repeating (a)-(b) for multiple iterations to obtain a plurality of alignment scores for a plurality of different sets of alignment positions;

(d) selecting a set of alignment positions from the plurality of different sets of alignment positions based on the plurality of alignment scores; and

(e) determining final alignment positions of the plurality of sequence reads to be the set of alignment positions selected in (d).

76. A system comprising one or more processors and system memory, wherein the one or more processors are configured to perform any one of the methods of claims 68-75.

77. A computer product comprising one or more computer-readable non-transitory storage media having stored thereon computer-executable instructions that, when executed by one or more processors of a computer system, cause the computer system to perform any one of the methods of claims 68-75.