CLAIMS

What is claimed is:

1. A method, implemented using a computer comprising one or more processors and system memory, for genotyping one or more repeat sequences each including one or more repeat sub-sequences, the method comprising:

(a) collecting, using the one or more processors, sequence reads of a test sample from a database;

(b) aligning, by the one or more processors, the sequence reads to the 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; and

(c) determining, by the one or more processors, one or more genotypes for the one or more repeat sequences using the sequence reads aligned to the one or more repeat sequences.

2. The method of claim 1, wherein a repeat sequence of the one or more repeat sequences comprises a particular repeat unit comprising at least one incompletely specified nucleotide.

3. The method of claim 2, wherein the particular repeat unit comprises degenerate codons.

4. The method of any of the preceding claims, wherein the one or more self- loops comprise two or more self-loops representing two or more repeat sub- sequences.

5. The method of any of the preceding claims, wherein the sequence graph further comprises two or more alternative paths for two or more alleles.

6. The method of claim 5, wherein the two or more alleles comprise an indel or a substitution.

7. The method of claim 6, wherein the substitution comprises a single nucleotide variant (SNV) or a single nucleotide polymorphism (SNP).

8. The method of claim 5, further comprising genotyping the two or more alleles using sequence reads aligned to the two or more alternative paths.

9. The method of claim 8, wherein genotyping the two or more alleles comprises providing coverages of the two or more alternative paths to a probabilistic model to determine the probabilities of the two or more alleles.

10. The method of claim 9, wherein the probabilistic model simulates a probability of an allele as a function of the coverage of the allele, the function being selected from a Poisson distribution, negative-binomial distribution, a binomial distribution, or a beta-binomial distribution.

11. The method of claim 10, wherein the Poisson distribution's rate parameter is estimated from read length and mean depth observed at the genomic locus.

12. The method of any of the preceding claims, further comprising, aligning, before (b), the sequence reads to a reference genome to determine genomic coordinates of the sequence reads, and selecting a subset of sequence reads as the sequence reads to be aligned to the one or more repeat sequences each represented by a sequence graph.

13. The method of claim 12, wherein, the subset of sequence reads comprises reads aligned to or near a region represented by a sequence graph.

14. The method of claim 12, wherein, the subset of sequence reads comprises unaligned reads whose mates map to or near a region represented by a sequence graph.

15. The method of claim 12, wherein, the subset of sequence reads comprises reads aligned to one or more off-target regions that are known hot spots for misaligning reads.

16. The method of any of the preceding claims, wherein aligning a sequence read to the sequence graph comprises:

finding a kmer match between the sequence read and a path of the sequence graph; and

extending the kmer match to a full alignment of nodes and edges of the sequence graph including one or more self-loops.

17. The method of any of the preceding claims, wherein aligning a sequence read to the sequence graph comprises graph shrinking by removing low confidence ends of the alignments.

18. The method of any of the preceding claims, wherein aligning a sequence read to the sequence graph comprises alignment merging by:

aligning subsequences of the read to a sequence graph; and

merging alignments of the subsequences to form a full alignment of the sequence read.

19. The method of any of the preceding claims, further comprising generating the sequence graph based on locus specification comprising a locus structure of the genomic locus.

20. The method of any of the preceding claims, wherein the sequence reads comprise paired end reads, and operation (c) comprises:

(i) identifying anchor and anchored reads in the paired end reads, wherein the anchor reads are reads aligned to or near the one or more repeat sequences, and the anchored reads are unaligned reads that are paired with the anchor reads; and

(ii) determining the one or more genotypes for the one or more repeat sequences using at least the anchored reads.

21. The method of claim 20, wherein operation (ii) comprises determining the one or more genotypes for the one or more repeat sequences using the anchor reads as well as the anchored reads.

22. The method of claim 20 or claim 21, wherein the anchor reads are aligned to within about 5 kb of a repeat sequence.

23. The method of any of claims 20-22, wherein the unaligned reads comprise reads that cannot be aligned or are poorly aligned to the sequence graph.

24. The method of any of the preceding claims, wherein the one or more repeat sequences comprise a short tandem repeat (STR) sequence.

25. The method of claim 24, wherein an expansion of the STR is associated with Fragile X syndrome, amyotrophic lateral sclerosis (ALS), Huntington's disease, Friedreich's ataxia, spinocerebellar ataxia, spino-bulbar muscular atrophy, myotonic dystrophy, Machado-Joseph disease, or dentatorubral pallidoluysian atrophy.

26. The method of any of the preceding claims, further comprising using a sequencer to generate paired end reads from the test sample.

27. The method of any of the preceding claims, further comprising extracting the test sample from an individual.

28. The method of any of the preceding claims, wherein the test sample is a blood sample, a urine sample, a saliva sample, or a tissue sample.

29. The method of any of the preceding claims, wherein the repeat unit comprises 1 to 50 nucleotides.

30. The method of any of the preceding claims, wherein the reads are shorter than at least one of the one or more repeat sequences.

31. A system comprising:

system memory; and

one or more processors configured to:

(a) collect sequence reads of a test sample from a database;

(b) align the sequence reads to the 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; and

(c) determine one or more genotypes for the one or more repeat sequences using the sequence reads aligned to the one or more repeat sequences.

32. A computer program product comprising a non-transitory machine readable medium storing program code that, when executed by one or more processors of a computer system, causes the computer system to implement a method for genotyping a repeat sequence in a test sample comprising nucleic acids, said program code comprising:

(a) code for collecting sequence reads of a test sample from a database;

(b) code for aligning the sequence reads to the 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; and

(c) code for determining one or more genotypes for the one or more repeat sequences using the sequence reads aligned to the one or more repeat sequences.