CLAIMS

1. A system for determining genotype of a gene comprising at least one repeat region, the system comprising:

at least one memory storing computer-executable instructions; and

at least one processor in communication with the at least one memory, wherein the at least one processor is configured to execute the computer-executable instructions to:

receive an input set of genetic sequencing data;

measure at least one repeat structure of at least one sequence read of the genetic sequencing data and remove sequence reads with abnormal repeat structures to generate a first dataset, the first dataset excluding sequence reads with sequence artifacts;

filter sequence reads of the first dataset to generate a second dataset, the second dataset excluding sequences with structural artifacts;

apply input data to a model to generate a third dataset, the input data comprising sequence reads of the second dataset, and the third dataset including sequencing reads categorized by allele;

generate a consensus sequence for each allele in the third dataset to generate a fourth dataset of consensus reads for all alleles in the third dataset;

translate each allele of the fourth dataset to generate a fifth dataset, the fifth dataset including a corresponding amino-acid sequence of each allele; and

determine a protein-altering variant status based on the amino-acid sequences of the fifth dataset.

2. The system of claim 1, wherein the at least one processor is further configured to execute the computer-executable instructions to compare amino-acid sequence lengths of the fifth dataset to an expected length of an amino-acid reference sequence, to generate a sixth dataset of protein-truncating variants.

3. The system of claim 2, wherein the at least one processor is further configured to execute the computer-executable instructions to classify the protein-truncating variants of the sixth dataset as stop-gain variants or frameshift variants.

4. The system of any of claims 1-3, wherein the at least one processor is further configured to execute the computer-executable instructions to exclude sequencing artifacts by detecting read structure by alignment, grouping sequencing reads by structure, sorting read groups by frequency, and identifying novel alleles with previously unknown copy number.

5. The system of claim 4, wherein an expected length is a size range as determined by a range of lengths of variants of the gene known to exist.

6. The system of any of claims 1-3, wherein the at least one processor is further configured to execute the computer-executable instructions to exclude structural artifacts by mapping sequences from known references of each individual repeat to the sequencing reads to determine the structure of the reads, and exclude reads that include repeat structures that are not representative of known variants of the gene, and are thus deemed to be sequencing artifacts.

7. A computer-implemented method for determining a genotype of a repeat rich gene, the method implemented using a system including a computing device including a processor communicatively coupled to memory device, the method comprising:

receiving an input set of genetic sequencing data;

measuring at least one repeat structure of at least one sequence read of the genetic sequencing data and remove sequence reads with abnormal repeat structures to generate a first dataset, the first dataset excluding sequence reads with sequence artifacts;

filtering sequence reads of the first dataset to generate a second dataset, the second dataset excluding sequences with structural artifacts;

applying input data to a model to generate a third dataset, the input data comprising sequence reads of the second dataset, and the third dataset including sequencing reads categorized by allele;

generating a consensus sequence for each allele in the third dataset to generate a fourth dataset of consensus reads for all alleles in the third dataset;

translating each allele of the fourth dataset to generate a fifth dataset, the fifth dataset including a corresponding amino-acid sequence of each allele; and

determining a protein-altering variant status based on the amino-acid sequences of the fifth dataset.

8. The computer-implemented method of claim 7, further comprising comparing amino-acid sequence lengths of the fifth dataset to an expected length of an amino-acid reference sequence, to generate a sixth dataset of protein-truncating variants.

9. The computer-implemented method of claim 8, further comprising classifying the protein-truncating variants of the sixth dataset as stop-gain variants or frameshift variants.

10. The computer-implemented method of any of claims 7-9, further comprising excluding sequencing artifacts by detecting read structure by alignment, grouping sequencing reads by structure, sorting read groups by frequency, and identifying novel alleles with previously unknown copy number.

11. The computer-implemented method of claim 10, wherein an expected length is a size range as determined by a range of lengths of variants of the gene known to exist.

12. The computer-implemented method of any of claims 7-9, exclude structural artifacts by mapping sequences from known references of each individual repeat to the sequencing reads to determine the structure of the reads, and excluding reads that include repeat structures that are not representative of known variants of the gene, and are thus deemed to be sequencing artifacts.

13. At least one non-transitory computer-readable storage medium having computer-executable instructions embodied thereon, wherein when executed by at least one processor, the computer-executable instructions cause the at least one processor to:

receive an input set of genetic sequencing data;

measure at least one repeat structure of at least one sequence read of the genetic sequencing data and remove sequence reads with abnormal repeat structures to generate a first dataset, the first dataset excluding sequence reads with sequence artifacts;

filter sequence reads of the first dataset to generate a second dataset, the second dataset excluding sequences with structural artifacts;

apply input data to a model to generate a third dataset, the input data comprising sequence reads of the second dataset, and the third dataset including sequencing reads categorized by allele;

generate a consensus sequence for each allele in the third dataset to generate a fourth dataset of consensus reads for all alleles in the third dataset;

translate each allele of the fourth dataset to generate a fifth dataset, the fifth dataset including a corresponding amino-acid sequence of each allele; and

determine a protein-altering variant status based on the amino-acid sequences of the fifth dataset.