

## CLAIMS

What is claimed is:

1. A computer-implemented method of generating a synthetic digital genomic dataset, comprising:  
obtaining a reference genome and introducing a plurality of SNPs at a predetermined frequency and distribution into at least one autosome and an X-chromosome of the reference genome to so prepare a synthetic maternal genome;  
introducing a plurality of SNPs at a predetermined frequency and distribution into at least one autosome and an X- or Y-chromosome of the reference genome to so prepare a synthetic paternal genome; and  
merging the maternal and paternal synthetic genomes into a combined dataset.
2. The method of claim 1 further comprising a step of sampling the combined dataset to thereby produce a plurality of simulated reads.
3. The method of claim 2 wherein the step of sampling is performed to simulate a read coverage of at least 25x.
4. The method of claim 2 or 3 wherein the step of sampling is performed using a read error and base quality profile representative of a frozen tissue sample.
5. The method of claim 2, 3, or 4 wherein the step of sampling is performed to produce simulated reads having a length of between 100 and 400 bases.
6. The method of any of the preceding claims wherein the predetermined frequency and distribution is representative of a naturally occurring frequency and distribution of SNPs.
7. The method of any of the preceding claims further comprising a step of including into the combined dataset a list identifying type and position of the SNPs relative to the reference genome.
8. The method of any of the preceding claims further comprising a step of including into the combined dataset a plurality of simulated reads from the combined dataset.
9. The method of any of the preceding claims further comprising a step of introducing into at least one of the synthetic maternal and paternal genome a further genomic change selected from the

group consisting of a SNV, an indel, and a copy number alteration to thereby produce a synthetic tumor data set.

10. The method of claim 9 wherein the synthetic tumor data set further comprises a list identifying type and position of the further genomic change relative to the at least one of the synthetic maternal and paternal genome.
11. The method of claim 9 or 10 wherein the synthetic tumor data set further comprises a plurality of simulated reads from the synthetic tumor data set.
12. The method of any one of claims 9-11 wherein the SNVs are based on at least one of COSMIC mutations, somatic TCGA mutations, and random locations in the genome.
13. The method of claim 12 wherein between 2,500 and 25,000 SNVs are introduced into the synthetic tumor data set.
14. The method of any one of claims 9-13 wherein the indels have a length of between 1 and 10 bases.
15. The method of claim 14 wherein between 100 and 1,000 indels are present in the synthetic tumor data set.
16. The method of any one of claims 9-15 wherein the copy number alteration is selected from the group consisting of (i) 25 small deletions, each with a size randomly chosen from 5,000 bp to 500,000 bp; (ii) 25 small tandem amplifications, each with a size randomly chosen between 5,000 bp to 500,000 bp. These amplifications have copy number between 2 and 5, randomly chosen; (iii) 10 small tandem hyperamplifications, with size 5,000 to 500,000 bp, and copy number randomly chosen between 15 and 30; (iv) Large arm/chromosome deletions, each with a size between 30% and 100% of a chromosome, anchored to a telomere.
17. The method of claim 2 or 3 wherein the step of sampling is performed to produce simulated reads having a length of between 100 and 400 bases.
18. The method of claim 1 wherein the predetermined frequency and distribution is representative of a naturally occurring frequency and distribution of SNPs.

19. The method of claim 1 further comprising a step of including into the combined dataset a list identifying type and position of the SNPs relative to the reference genome.
20. The method of claim 1 further comprising a step of including into the combined dataset a plurality of simulated reads from the combined dataset.
21. The method of claim 1 further comprising a step of introducing into at least one of the synthetic maternal and paternal genome a further genomic change selected from the group consisting of a SNV, an indel, and a copy number alteration to thereby produce a synthetic tumor data set.
22. The method of claim 21 wherein the synthetic tumor data set further comprises a list identifying type and position of the further genomic change relative to the at least one of the synthetic maternal and paternal genome.
23. The method of claim 21 or 22 wherein the synthetic tumor data set further comprises a plurality of simulated reads from the synthetic tumor data set.
24. The method of claim 21 or 22 wherein the SNVs are based on at least one of COSMIC mutations, somatic TCGA mutations, and random locations in the genome.
25. The method of claim 24 wherein between 2,500 and 25,000 SNVs are introduced into the synthetic tumor data set.
26. The method of claim 21 wherein the indels have a length of between 1 and 10 bases.
27. The method of claim 26 wherein between 100 and 1,000 indels are present in the synthetic tumor data set.
28. The method of claim 21 wherein the copy number alteration is selected from the group consisting of (i) 25 small deletions, each with a size randomly chosen from 5,000 bp to 500,000 bp; (ii) 25 small tandem amplifications, each with a size randomly chosen between 5,000 bp to 500,000 bp. These amplifications have copy number between 2 and 5, randomly chosen; (iii) 10 small tandem hyperamplifications, with size 5,000 to 500,000 bp, and copy number randomly chosen between 15 and 30; (iv) Large arm/chromosome deletions, each with a size between 30% and 100% of a chromosome, anchored to a telomere.

29. A method of testing or validating an algorithm associated with genomic analysis comprising a step of using the synthetic genomic dataset of any one of claims 17-28 as input for the algorithm, and preparing a performance report listing deviations from the synthetic genomic dataset in an output of the algorithm.
30. The method of claim 29 wherein the algorithm is any one of the following: an algorithm that groups a plurality of read files, an algorithm that annotates a plurality or group of read files, an algorithm that transports a plurality of read files between a sequencing device and a downstream device, an algorithm that assembles and indexes a plurality of read files, and a variant calling algorithm.
31. A method of validating operation of a plurality of computing devices that are informationally coupled to each other, comprising a step of using the synthetic genomic dataset of any one of claims 17-28 as an input into a first of the devices, and using an output of the first of the devices as input into a second of the devices.