

CLAIMS

1. A computer-implemented method for identifying somatic variants in a plurality of variants, comprising:
 - (a) applying a database filter to the plurality of variants comprising somatic variants and germline variants by the computer, comprising:
 - determining first germline variants in the plurality of variants, wherein the first germline variants each have an allele count in a first reference set of variants greater than or equal to a threshold allele count;
 - (b) applying a proximity filter to the plurality of variants by the computer, comprising:
 - (i) binning variants of the plurality of variants into a plurality of bins, wherein variants located in the same region of a genome are binned into the same bin,
 - (ii) determining database variants in the plurality of variants, wherein a database variant is present in a second reference set of variants, and
 - (iii) determining second germline variants in the plurality of variants, wherein the second germline variants each have an allele frequency within a proximate range of an allele frequency of at least one database variant in the same bin as the second germline variant, wherein the proximate range is a range having a maximum and a minimum of 0.05 from the allele frequency of a second germline variant or wherein the proximate range is a range having a maximum and a minimum of two standard deviations from a binomial distribution of an allele frequency of a second germline variant, and centered from the allele frequency of a second germline variant; and
 - (c) determining somatic variants in the plurality of variants by removing the identified first and second germline variants from the plurality of variants by the computer.
2. The method of claim 1, wherein (a) and (b) are performed by the computer consecutively.
3. The method of claim 1, wherein (b) is performed by the computer before (a).
4. The method of any one of claims 1-3, wherein the threshold allele count is 5.
5. The method of claim 4, wherein the threshold allele count is 10.

6. The method of any one of claims 1-5, wherein the first and second reference set of variants are the same reference set.

7. The method of any one of claims 1-6, wherein the first or second reference set of variants comprises a database of variants for a plurality of individuals.

8. The method of any one of claims 1-7, wherein the first or second reference set of variants comprises at least one database selected from a genome aggregation database (gnomAD), and a 1000 genome database.

9. The method of any one of claims 1-8, wherein the same region of a genome is within the same chromosome.

10. The method of any one of claims 1-9, wherein the same region of a genome is within the same chromosomal arm.

11. The method of any one of claims 1-10, wherein the same region of a genome is within the same chromosomal cytoband.

12. The method of any one of claims 1-11, wherein the same region of a genome is within a 10 Mb region.

13. The method of any one of claims 1-12, wherein the applying a proximity filter further comprises identifying a second germline variant having an allele frequency greater than or equal to 0.9 by the computer.

14. The method of any one of claims 1-13, wherein the applying a proximity filter further comprises identifying a second germline variant in the plurality of variants by the computer, wherein the second germline variant is a database variant present in the second reference set of variants.

15. The method of any one of claims 1-14, wherein the proximate range is a range having a maximum and a minimum of 0.05 from the allele frequency of a second germline variant.

16. The method of any one of claims 1-15, wherein the proximate range is a range having a maximum and a minimum of two standard deviations from a binomial distribution of an allele frequency of a second germline variant, and centered from the allele frequency of a second germline variant.

17. The method of any one of claims 1-16, wherein the second germline variants have an allele frequency within a threshold proximity to an allele frequency of at least five database variants in the same bin as the second germline variant.

18. The method of any one of claims 1-17, wherein the second germline variants have an allele frequency within a threshold proximity to an allele frequency of at least ten database variants in the same bin as the second germline variant.

19. The method of any one of claims 1-18, wherein the plurality of variants is identified by obtaining sequence data from a biological sample comprising a tumor cell by the computer.

20. The method of claim 19, wherein the plurality of variants is identified by aligning the sequence data with a reference sequence, and identifying variants in the sequence data by the computer.

21. The method of claim 19 or 20, wherein the biological sample comprising a tumor cell is selected from a serum sample, a stool sample, a blood sample, a tumor sample.

22. The method of claim 21, wherein the tumor sample is fixed.

23. A computer-implemented method of determining a tumor mutation burden of a tumor, comprising:

obtaining sequence data from a biological sample comprising a tumor cell by the computer;

determining a plurality of variants from the sequence data by the computer; and

determining the number of somatic variants in a plurality of variants by the computer according to the method of any one of claims 1-22, wherein the number of somatic variants is the tumor mutation burden of the tumor.

24. An electronic system for analyzing genetic variation data, comprising:

an informatics module running on a processor and adapted to identify a plurality of variants from sequence data from a biological sample comprising a tumor cell, wherein the plurality of variants comprises somatic variants and germline variants;

a database filter module adapted to remove first germline variants from the plurality of variants, wherein the first germline variants each have an allele count in a first reference set of variants greater than or equal to a threshold allele count;

a proximity filter module adapted to remove second germline variants from the plurality of variants, the proximity filter module comprising:

a binning sub-module adapted to return a plurality of bins, each bin containing variants of the plurality of variants located in the same region of a genome,

an identification sub-module adapted to return database variants in the plurality of variants, wherein a database variant is present in a second reference set of variants, and

a removal sub-module adapted to remove second germline variants from the plurality of variants, wherein the second germline variants each have an allele frequency within a proximate range of an allele frequency of at least one database variant in the same bin as the second germline variant, wherein the proximate range is a range having a maximum and a minimum of 0.05 from the allele frequency of a second germline variant or wherein the proximate range is a range having a maximum and a minimum of two standard deviations from a binomial distribution of an allele frequency of a second germline variant, and centered from the allele frequency of a second germline variant; and

a display module adapted to return variants not removed from the plurality of variants.

25. The system of claim 24, wherein informatics module comprises a variant annotation tool.

26. The system of claim 24 or 25, wherein the threshold allele count is 5.

27. The system of claim 26, wherein the threshold allele count is 10.

28. The system of any one of claims 24-27, wherein the first and second reference set of variants are the same reference set.

29. The system of any one of claims 24-28, wherein the first or second reference set of variants comprises a database of variants for a plurality of individuals.

30. The system of any one of claims 24-29, wherein the first or second reference set of variants comprises at least one database selected from a genome aggregation database (gnomAD), and a 1000 genome database.

31. The system of any one of claims 24-30, wherein the same region of a genome is within the same chromosome.

32. The system of any one of claims 24-31, wherein the same region of a genome is within the same chromosomal arm.

33. The system of any one of claims 24-32, wherein the same region of a genome is within the same chromosomal cytoband.

34. The system of any one of claims 24-33, wherein the same region of a genome is within a 10 Mb region.

35. The system of any one of claims 24-34, wherein the removal sub-module is adapted to remove a variant having an allele frequency greater than or equal to 0.9 from the plurality of variants.

36. The system of any one of claims 24-35, wherein the removal sub-module is adapted to remove a database variant present in the second reference set of variants from the plurality of variants.

37. The system of any one of claims 24-36, wherein the proximate range is a range having a maximum and a minimum of 0.05 from the allele frequency of a second germline variant.

38. The system of any one of claims 24-37, wherein the proximate range is a range having a maximum and a minimum of two standard deviations from a binomial distribution of an allele frequency of a second germline variant, and centered from the allele frequency of a second germline variant.

39. The system of any one of claims 24-38, wherein the second germline variants have an allele frequency within a threshold proximity to an allele frequency of at least five database variants in the same bin as the second germline variant.

40. The system of any one of claims 24-39, wherein the second germline variants have an allele frequency within a threshold proximity to an allele frequency of at least ten database variants in the same bin as the second germline variant.

41. The system of any one of claims 24-40, wherein the biological sample comprising a tumor cell is selected from a serum sample, a stool sample, a blood sample, a tumor sample.

42. The system of claim 41, wherein the tumor sample is fixed.

43. A computer-implemented method for identifying somatic variants in a plurality of variants, comprising: performing the method of any one of claims 1-22 by the computer.

44. A computer-implemented method for identifying somatic variants in a plurality of variants, comprising:

(a) receiving a plurality of variants from sequence data from a biological sample comprising a tumor cell by the computer, wherein the plurality of variants comprising somatic variants and germline variants;

(b) applying a database filter to the plurality of variants by the computer, comprising:

creating an index of documents for the plurality of variants,

searching a first reference set of variants with the index to identify first germline variants in the index, wherein the first germline variants each have an allele count in the first reference set of variants greater than or equal to a threshold allele count, and

removing the identified first germline variants from the index to create an index of first filtered variants;

(c) applying a proximity filter to the index of first filtered variants by the computer, comprising:

(i) creating a plurality of bins for different regions of a genome,

(ii) binning variants of the index of first filtered variants, wherein variants located in the same region of a genome are binned into the same bin,

(iii) searching a second reference set of variants with the index of first filtered variants to identify database variants in the index of first filtered variants,

(iv) generating an index of second germline variants from the index of first filtered variants by identifying second germline variants, wherein the second germline variants each have an allele frequency within a proximate range of an allele frequency of at least one database variant in the same bin as the second germline variant, wherein the proximate range is a range having a maximum and a minimum of 0.05 from the allele frequency of a second germline variant or wherein the proximate range is a range having a maximum and a minimum of two standard deviations from a binomial distribution of an allele frequency of a second germline variant, and centered from the allele frequency of a second germline variant, and

(v) removing the identified second germline variants from the index of first filtered variants to create an index of somatic variants, thereby identifying somatic variants in the plurality of variants.

45. The method of claim 44, wherein the threshold allele count is 5.

46. The method of claim 45, wherein the threshold allele count is 10.

47. The method of any one of claims 44-46, wherein the first and second reference set of variants are the same reference set.

48. The method of any one of claims 44-47, wherein the first or second reference set of variants comprises a database of variants for a plurality of individuals.

49. The method of any one of claims 44-48, wherein the first or second reference set of variants comprises at least one database selected from a genome aggregation database (gnomAD), and a 1000 genome database.

50. The method of any one of claims 44-49, wherein the same region of a genome is within the same chromosome.

51. The method of any one of claims 44-50, wherein the same region of a genome is within the same chromosomal arm.

52. The method of any one of claims 44-51, wherein the same region of a genome is within the same chromosomal cytoband.

53. The method of any one of claims 44-52, wherein the same region of a genome is within a 10 Mb region.

54. The method of any one of claims 44-53, wherein the generating an index of second filtered variants further comprises identifying a second germline variant having an allele frequency greater than or equal to 0.9 by the computer.

55. The method of any one of claims 44-54, wherein the generating an index of second filtered variants further comprises identifying a second germline variant in the plurality of variants by the computer, wherein the second germline variant is a database variant present in the second reference set of variants.

56. The method of any one of claims 44-55, wherein the proximate range is a range having a maximum and a minimum of 0.05 from the allele frequency of a second germline variant.

57. The method of any one of claims 44-56, wherein the proximate range is a range having a maximum and a minimum of two standard deviations from a binomial distribution of an allele frequency of a second germline variant, and centered from the allele frequency of a second germline variant.

58. The method of any one of claims 44-57, wherein the second germline variants have an allele frequency within a threshold proximity to an allele frequency of at least five database variants in the same bin as the second germline variant.

59. The method of any one of claims 44-58, wherein the second germline variants have an allele frequency within a threshold proximity to an allele frequency of at least ten database variants in the same bin as the second germline variant.

60. The method of any one of claims 44-58, wherein the biological sample comprising a tumor cell is selected from a serum sample, a stool sample, a blood sample, a tumor sample.

61. The method of claim 60, wherein the tumor sample is fixed.