CLAIMS

1. A computer-implemented method of analysing genetic data about an organism, the method comprising:

receiving a plurality of input units, wherein each input unit comprises information about the association between a plurality of genetic variants in a region of interest of the genome of the organism and a target phenotype of the organism;

carrying out one or more iterations comprising, for each of the plurality of genetic variants:

determining whether the genetic variant is causal for the target phenotype based on the plurality of input units; and

if the genetic variant is determined to be causal, determining a sampled effect size of the genetic variant on the target phenotype for each of the input units based on the plurality of input units and information about correlations between the plurality of genetic variants in the region of interest, the sampled effect size of the genetic variant on the target phenotype being non-zero for all of the input units; and

for each genetic variant, determining a prediction effect size of the genetic variant on the target phenotype for each of the input units based on an average across at least a subset of the iterations of the sampled effect sizes of the genetic variant for the input unit or of posterior effect sizes of the genetic variant for the input unit calculated using the sampled effect sizes.

2. The method of claim 1, wherein determining whether the genetic variant is causal comprises calculating the probability of the information from the plurality of input units, assuming that the genetic variant is causal and a probability of the information from the plurality of input units assuming that the genetic variant is not causal, and stochastically determining the genetic variant to be causal with a probability dependent on a ratio of the probability of the input data assuming that the genetic variant is causal and the probability of the input data assuming that the genetic variant is not casual.

3. The method of claim 2, wherein the probability of the information from the

plurality of input units assuming that the genetic variant is causal is dependent on:

a proportion of the plurality of genetic variants expected to be causal;

the plurality of input units; and

a correlation between the effect sizes of the genetic variant on the target phenotype for each of the input units.

4. The method of claim 2 or 3, wherein the probability of the information from the plurality of input units assuming that the genetic variant is not causal is dependent on:

a proportion of the plurality of genetic variants expected to be causal; and

the plurality of input units.

5. The method of claim 3 or 4, wherein the proportion of the plurality of genetic variants expected to be causal is predetermined.

6. The method of any of claims 3 to 5, wherein the correlation between the effect sizes of the genetic variant on the target phenotype for each of the input units is predetermined.

7. The method of claim 3, 4, or 6, wherein the proportion of the plurality of genetic variants expected to be causal is updated at each iteration.

8. The method of any of claims 3 to 5, or 7, wherein the correlation between the effect sizes of the genetic variant on the target phenotype for each of the input units is updated at each iteration.

9. The method of any of claims 2 to 8, wherein the input units are determined from respective groups of individuals, and the probability of the information from the plurality of input units assuming that the genetic variant being causal is dependent on one or more parameters quantifying an overlap in the groups of individuals between respective pairs of input units.

10. The method of any preceding claim, wherein determining the sampled effect size of

the genetic variant comprises calculating a probability distribution, for example a multivariate normal distribution, of effect sizes of the genetic variant on the target phenotype for the input units, and sampling values of the effect sizes for the input units from the probability distribution.

11. The method of claim 10, wherein the sampling of values of the effect size in each iteration is dependent on the sampled effect sizes from one or more previous iterations.

12. The method of claim 10 or 11, wherein the sampling of values of the effect size is performed using a Monte-Carlo Gibbs sampler.

13. The method of any of claims 10 to 12, wherein the probability distribution is dependent on a correlation between the effect sizes of the genetic variant on the target phenotype for each of the input units.

14. The method of claim 13, wherein the correlation between the effect sizes of the genetic variant on the target phenotype for each of the input units is either predetermined or updated at each iteration.

15. The method of any preceding claim, wherein:

each of the one or more iterations further comprises, for each genetic variant determined to be causal, subtracting weighted effect sizes from the information about the association between each other genetic variant and the target phenotype of each input unit;

the weighted effect sizes are the sampled effect size of the genetic variant on the target phenotype for the input unit weighted by respective correlation factors between the genetic variant and each other genetic variant; and

the correlation factors are determined based on the information about correlations between the plurality of genetic variants in the region of interest.

16. The method of claim 15, wherein the input units are determined from respective groups of individuals, and the correlation factors between the genetic variant and each

other genetic variant depend on an ancestry of the group of individuals of the input unit.

17. The method of claim 16, wherein the group of individuals of at least one of the input units comprises individuals having a common ancestry, the correlation factors being determined based on correlations between genetic variants in the region of interest for individuals having the common ancestry.

18. The method of claim 16 or 17, wherein the group of individuals of at least one of the input units comprises individuals having different ancestries, the correlation factors being determined based on an average of correlations between genetic variants in the region of interest for individuals having each of the different ancestries.

19. The method of any preceding claim, wherein the group of individuals of at least one of the input units comprises one or both of individuals having the same value of a characteristic and individuals having different values of a characteristic.

20. The method of claim 19, wherein the characteristic is one of sex, age, weight, a molecular biomarker, or a behavioural characteristic.

21. The method of any preceding claim, wherein carrying out one or more iterations comprises carrying out a predetermined number of iterations.

22. The method of any preceding claim, wherein each of the one or more iterations further comprises a step of evaluating a convergence parameter, and carrying out one or more iterations comprises carrying out iterations until a predetermined condition on the convergence parameter is met.

23. The method of any preceding claim, wherein the information about the association between the plurality of genetic variants and the target phenotype comprises, for each of the plurality of genetic variants, an estimate of a strength of association between the genetic variant and the target phenotype and an error in the estimate of the strength of

association.

24. A method of determining a polygenic risk score for a target phenotype for a target individual comprising:

receiving genetic information about a region of interest of the genome of the target individual;

receiving prediction effect sizes on the target phenotype of a plurality of genetic variants in the region of interest determined using the method of analysing genetic data of any preceding claim; and

determining the polygenic risk score based on the genetic information for the target individual and the prediction effect sizes.

25. The method of claim 24, wherein the input units received in the method of analysing genetic data are determined from respective groups of individuals, and the polygenic risk score for the individual is determined using the prediction effect sizes for the input unit determined from a group of individuals most similar to the target individual.

26. An apparatus for analysing genetic data about an organism, the apparatus comprising:

a receiving unit configured to receive a plurality of input units, wherein each input unit comprises information about the association between a plurality of genetic variants in a region of interest of the genome of the organism and a target phenotype of the organism; and

a data processing unit configured to:

carry out one or more iterations comprising, for each of the plurality of genetic variants:

determining whether the genetic variant is causal for the target phenotype based on the plurality of input units; and

if the genetic variant is determined to be causal, determining a sampled effect size of the genetic variant on the target phenotype for each of the input units based on the plurality of input units and information about correlations between the plurality of

genetic variants in the region of interest, the sampled effect size of the genetic variant on the target phenotype being non-zero for all of the input units; and

for each genetic variant, determine a prediction effect size of the genetic variant on the target phenotype for each of the input units based on an average across at least a subset of the iterations of the sampled effect sizes of the genetic variant for the input unit or of posterior effect sizes of the genetic variant for the input unit calculated using the sampled effect sizes.

27. A computer program or a computer-readable medium comprising instructions which, when executed by a computer, cause the computer to carry out the method of any of claims 1 to 25.