

CLAIMS

1. A computer-implemented method of analysing genetic data about an organism to obtain information about the organism, the method comprising:
 - receiving input data comprising strengths of association between one or more phenotypes including a target phenotype and a plurality of genetic variants in a region of interest of the genome of the organism;
 - applying a fine-mapping algorithm to all or a subset of the input data to identify one or more independent phenotype-variant associations within the region of interest, comprising identifying for each association a set of one or more fine-mapped variants from the plurality of genetic variants, and determining for each fine-mapped variant an estimated probability of being causal for the phenotype-variant association, the sum of the probabilities for the fine-mapped variants within the set adding to one;
 - calculating, on the basis of the input data and the set of fine-mapped variants, a fine-mapping predictive model quantifying an effect on the target phenotype of the set of fine-mapped variants;
 - subtracting from the input data, using the fine-mapping predictive model, the effect on the target phenotype of the set of fine-mapped variants to obtain residual association data; and
 - applying a machine learning algorithm to the residual association data to identify further predictive correlations between the target phenotype and the plurality of genetic variants.
2. The method of claim 1, wherein the strengths of association comprise an estimated effect size of each of the plurality of genetic variants on the target phenotype, and a standard error of each of the estimated effect sizes.
3. The method of claim 1 or 2, wherein the step of receiving input data comprises:
 - receiving individual level data comprising genotypes and corresponding phenotypes for each of a plurality of individuals; and
 - determining using the individual level data an estimated effect size of each of the

plurality of genetic variants on the target phenotype and a standard error of each of the estimated effect sizes.

4. The method of any preceding claim, wherein the identifying of the set of fine-mapped variants is performed using an iterative method, wherein each iteration comprises:

- identifying, on the basis of the input data, a fine-mapped variant within the region of the genome different from any previously identified fine-mapped variant;
- updating the input data to account for the effect on the target phenotype of the fine-mapped variants already identified, using a matrix of correlations between the genetic variants within the region of the genome; and
- determining whether to perform a further iteration on the basis of the updated input data.

5. The method of any preceding claim, wherein the identifying of the set of fine-mapped variants comprises using a plurality of instrument traits known to affect the target phenotype, the use of the instrument traits comprising:

- determining a set of fine-mapped variants for the instrument traits; and
- determining whether to include each of one or more of the fine-mapped variants for the instrument traits in the set of fine-mapped variants for the target phenotype on the basis of a relationship between the plurality of instrument traits and the target phenotype.

6. The method of any preceding claim, wherein the identifying of the set of fine-mapped variants comprises identifying a set of fine-mapped variants for one or more directly causal instrument traits known to affect the target phenotype.

7. The method of claim 5 or 6, wherein the calculating of the fine-mapping predictive model comprises:

- determining effect sizes on the one or more instrument traits of the set of fine-mapped variants for the one or more instrument traits, and
- determining an effect size for the target phenotype of each of the fine-mapped variants for the instrument traits included in the set of fine-mapped variants for the target

phenotype on the basis of a predetermined relationship between effect sizes for the instrument traits and effect sizes for the target phenotype.

8. The method of any preceding claim, wherein:

the strengths of association comprise an estimated effect size of each of the plurality of genetic variants on the target phenotype, and a standard error of each of the estimated effect sizes; and

the fine-mapping predictive model comprises a fine-mapped effect size on the target phenotype for each of the fine-mapped variants, the fine-mapped effect size being calculated from the estimated effect size of the fine-mapped variants taking account of the estimated probability of the fine-mapped variants being causal for the phenotype-variant association.

9. The method of any preceding claim, wherein the effect on the target phenotype of the set of fine-mapped variants is inferred using a machine learning algorithm.

10. The method of claim 9, wherein the set of fine-mapped variants further comprises one or more variants known to have a high likelihood of being causal for the target phenotype.

11. The method of any preceding claim, wherein:

the strengths of association comprise an estimated effect size of each of the plurality of genetic variants on the target phenotype, and a standard error of each of the estimated effect sizes; and

the step of subtracting from the input data the effect on the target phenotype of the set of fine-mapped variants comprises obtaining a residual effect size for each of a plurality of the genetic variants in the input data, the residual association data comprising the residual effect sizes,

wherein, after appropriate renormalisation of the effect sizes to ensure equal variance, the residual effect size $\hat{\beta}_i$ for genetic variant i is given by:

$$\hat{\beta}_i = \beta_i - \sum_{j=1}^N p_j r_{ij} \tilde{\beta}_j$$

where β_i is the estimated marginal effect size of genetic variant i , N is the number of fine-mapped variants, p_j is the probability that variant j is causal, $\tilde{\beta}_j$ is the fine-mapped effect size of the j^{th} fine-mapped variant on the target phenotype, and r_{ij} is a correlation between the j^{th} fine-mapped variant and genetic variant i .

12. The method of any preceding claim, wherein the input data are derived from a plurality of different genetic studies, and the step of applying a machine learning algorithm to the residual association data comprises using a prior probability for each of the plurality of genetic variants of being causal for the target phenotype that is dependent on the consistency of the strength of association between each genetic variant and the target phenotype between the different genetic studies.

13. The method of any preceding claim, wherein the step of applying a machine learning algorithm to the residual association data comprises using a prior probability for each of the plurality of genetic variants of being causal for the target phenotype that is dependent on genomic annotations of the plurality of genetic variants in the region of interest.

14. The method of any preceding claim, further comprising a step of calculating a polygenic risk score for an individual for the target phenotype by applying to genetic data from the individual the fine-mapping predictive model and the further predictive correlations identified by the machine learning algorithm.

15. The method of claim 14, wherein the polygenic risk score is given by the followed weighted sum:

$$PRS = \sum_{l=1}^L \alpha_l x_l$$

where L is the number of variants that contribute to the PRS, each variant being included either in the fine-mapping predictive model or in the further predictive correlations from

the machine learning algorithm, α_l quantifies a strength of association of variant l on the target phenotype, the strength of association being specified by the fine-mapping predictive model or by the further predictive correlations from the machine learning algorithm, and x_l is the genotype for variant l .

16. The method of claim 14 or 15, wherein the polygenic risk score for the individual is derived from a combination of a first partial polygenic risk score provided by applying the fine-mapping predictive model to genetic data from the individual and a second partial polygenic risk score provided by applying the further predictive correlations of the machine learning algorithm to the genetic data from the individual.

17. The method of any preceding claim, wherein the input data are derived from a plurality of different populations of the organism, and either or both of the following is satisfied:

the calculating of the fine-mapping predictive model is performed separately for portions of the input data corresponding to different populations to obtain multiple respective population-matched fine-mapping predictive models; and

the applying of the machine learning algorithm to the residual association data is performed separately for portions of the input data corresponding to different populations to obtain multiple respective sets of population-matched further predictive correlations.

18. The method of claim 17, further comprising:

receiving input data from an individual having genes from a mixture of the different populations; and

calculating a polygenic risk score for the individual by performing either or both of:
matching each of multiple population-matched fine-mapping predictive models to a corresponding portion of the input data that matches the population of the population-matched fine-mapping predictive model and applying each matched fine-mapping predictive model to the corresponding portion of the input data; and

matching each of multiple sets of population-matched further predictive correlations to a corresponding portion of the input data that matches the population of the

set of population-matched further predictive correlations and applying each matched set of further predictive correlations to the corresponding portion of the input data.

19. The method of claim 18, wherein the matching of each of multiple sets of population-matched further predictive correlations is performed and the matching of each of multiple population-matched fine-mapping predictive models is not performed, the calculation of the polygenic risk score comprising applying a shared population-consistent fine-mapping predictive model to the input data from the individual.

20. The method of claim 17, further comprising:
receiving input data from an individual having genes predominantly from one of the different populations; and
calculating a polygenic risk score for the individual by performing either or both of:
applying a population-matched fine-mapping predictive model to all of the input data from the individual, the population-matched fine-mapping predictive model being matched to the population of the individual; and
applying a set of population-matched further predictive correlations to all of the input data from the individual, the set of population-matched further predictive correlations being matched to the population of the individual.

21. The method of claim 20, wherein the applying of the set of population-matched further predictive correlations is performed and the applying of the population-matched fine-mapping predictive model is not performed, the calculation of the polygenic risk score comprising applying a shared population-consistent fine-mapping predictive model to the input data from the individual.

22. The method of any preceding claim, wherein the identifying of the one or more fine-mapped variants by the fine-mapping algorithm takes account of associations between the plurality of genetic variants and phenotypes other than the target phenotype.

23. The method of any preceding claim, wherein the organism is a human.

24. An apparatus for analysing genetic data about an organism to obtain information about the organism, the apparatus comprising:

a receiving unit configured to receive input data comprising strengths of association between one or more phenotypes including a target phenotype and a plurality of genetic variants in a region of interest of the genome of the organism; and

a data processing unit configured to:

apply a fine-mapping algorithm to all or a subset of the input data to identify one or more independent phenotype-variant associations within the region of interest, by identifying for each association a set of one or more fine-mapped variants from the plurality of genetic variants, and determining for each fine-mapped variant an estimated probability of being causal for the phenotype-variant association, the sum of the probabilities for the fine-mapped variants within the set adding to one;

calculate, on the basis of the input data and the set of fine-mapped variants, a fine-mapping predictive model quantifying an effect on the target phenotype of the set of fine-mapped variants;

subtract, using the fine-mapping predictive model, the effect on the target phenotype of the set of fine-mapped variants from the input data to obtain residual association data; and

apply a machine learning algorithm to the residual association data to identify further predictive correlations between the target phenotype and the plurality of genetic variants.

25. A computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out the method of any of claims 1 to 23.

26. 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 23.