Some tools to build PRSs are designed to take advantage of summary statistics data. One approach, popularised by the LDpred software (Vilhjálmsson et al 2015, https://github.com/bvilhjal/ldpred), iterates through multiple random selections of plausible variants genome-wide based on a single GWAS and, as variants are picked or removed, estimates the residual signal.

Existing methods to deal with this issue are based upon creating PRS using training datasets from individuals exhibiting the trait (or phenotype) or combination of traits of interest. However, the amount of data that is available for particular phenotypes can vary greatly, both in quantity and quality. For example, where the trait of interest is chance of stroke, this can be difficult to quantify in a robust and consistent way. This affects in turn the usefulness of PRS calculated from studies of stroke risk. It would be advantageous to be able to analyse data from multiple studies in a way which improves the calculation of PRS for phenotypes of this kind.

It is an object of the invention to improve analysis of genetic data about an organism and/or allow more robust and/or accurate PRSs to be obtained for individuals.

According to a~~n~~ first aspect of the ~~invention~~disclosure, there is provided a computer-implemented method of analysing genetic data about an organism. The method comprises 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 one of a plurality of phenotypes or phenotype combinations of the organism, carrying out one or more iterations comprising, for each of the plurality of genetic variants: determining for which of the plurality of phenotypes or phenotype combinations the genetic variant is causal based on the plurality of input units; and if the genetic variant is determined to be causal for one or more of the phenotypes or phenotype combinations, determining a sampled effect size of the genetic variant on each of the one or more phenotypes or phenotype combinations based on the plurality of input units and information about correlations between the plurality of genetic variants in the region of interest, and for each genetic variant, determining a prediction effect size of the genetic variant on one or more of the phenotypes or phenotype combinations based on an average across at least a subset of the iterations of the sampled effect sizes of the genetic variant on the one or more phenotypes or phenotype combinations or of posterior effect

sizes of the

In some embodiments, 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 phenotype or phenotype combination of each input unit; the weighted effect sizes being the sampled effect size of the genetic variant on the phenotype or phenotype combination of 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. Subtracting the effect of a variant determined to be causal from linked variants ensures that multiple causal variants are not erroneously identified based on a single causal relationship. Using input-unit specific correlation factors allows the method to account for variations in genetic correlations between subpopulations.

In some embodiments, carrying out one or more iterations comprises carrying out a predetermined number of iterations. Carrying out a predetermined number of iterations may provide adequate results for a known type of problem while remaining computationally efficient.

In some embodiments, 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. Calculating a convergence parameter may be advantageous where an appropriate number of iterations is uncertain.

In some embodiments, the information about the association between the plurality of genetic variants and each of the phenotypes or phenotype combinations comprises, for each of the plurality of genetic variants, an estimate of a strength of association between the genetic variant and the phenotype or phenotype combination and an error in the estimate of the strength of association. As mentioned above, using this type of summary statistic data has advantages in the availability of large quantities of data.

According to ~~another~~ a second aspect, there is provided a method of determining a polygenic risk score for a target phenotype or target phenotype combination for a target individual. The method comprises: receiving genetic information about a region of interest of the genome of the target individual; receiving prediction effect sizes on the target

phenotype or

target phenotype combination of a plurality of genetic variants in the region of interest determined using the method of analysing genetic data of ~~any preceding claim~~the first aspect; and determining the polygenic risk score based on the genetic information for the target individual and the prediction effect sizes. As mentioned above, calculating polygenic risk scores is a particularly desirable use of the prediction effect sizes determined for genetic variants, and can be used for a variety of clinical applications.

According to ~~another~~ a third aspect of the ~~invention~~disclosure, there is provided an apparatus for analysing genetic data about an organism. The apparatus comprises 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 one of a plurality of phenotypes or phenotype combinations 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 for which of the plurality of phenotypes or phenotype combinations the genetic variant is causal based on the plurality of input units; and if the genetic variant is determined to be causal for one or more of the phenotypes or phenotype combinations, determining a sampled effect size of the genetic variant on the one or more phenotypes or phenotype combinations based on the plurality of input units and information about correlations between the plurality of genetic variants in the region of interest; and for each genetic variant, determining a prediction effect size of the genetic variant on one or more of the phenotypes or phenotype combinations based on an average across at least a subset of the iterations of the sampled effect sizes of the genetic variant on the one or more phenotypes or phenotype combinations or of posterior effect sizes of the genetic variant for the input unit calculated using the sampled effect sizes.

~~The invention may also be embodied in~~According to a fourth aspect, the present disclosure provides a computer program or a computer-readable medium comprising instructions which~~ cause the computer to carry out the method, or a computer-readable medium comprising instructions which~~, when executed by a computer, cause the computer to carry out the method of the first aspect.

Embodiments of the invention will be further described by way of example only with reference to the accompanying drawings, in which:

Fig. 1 is a flowchart of a method of analysing genetic data about an organism

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 one of a plurality of phenotypes or phenotype combinations of the organism;

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

determining for which of the plurality of phenotypes or phenotype combinations the genetic variant is causal based on the plurality of input units; and

if the genetic variant is determined to be causal for one or more of the phenotypes or phenotype combinations, determining a sampled effect size of the genetic variant on each of the one or more phenotypes or phenotype combinations based on the plurality of input units and information about correlations between the plurality of genetic variants in the region of interest; and

for each genetic variant, determining a prediction effect size of the genetic variant on one or more of the phenotypes or phenotype combinations based on an average across at least a subset of the iterations of the sampled effect sizes of the genetic variant on the one or more phenotypes or phenotype combinations 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 for which of the plurality of phenotypes or phenotype combinations the genetic variant is causal comprises calculating a plurality of probabilities comprising:

a probability of the information from the plurality of input units assuming that the genetic variant is not causal for any of the phenotypes or phenotype combinations;

a probability of the information from the plurality of input units assuming that the genetic variant is causal for all of the phenotypes or phenotype combinations; and

for one or more subsets of the phenotypes or phenotype combinations, a probability

of the information from the plurality of input units assuming that the genetic variant is causal for the subset of phenotypes or phenotype combinations, and

stochastically determining for which of the plurality of phenotypes or phenotype combinations the genetic variant is causal with a probability based on the plurality of probabilities.

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 for one or more of the phenotypes or phenotype combinations 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 phenotypes or phenotype combinations.

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 for any of the phenotypes or phenotype combinations 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 any of claims 2 to 4, wherein, for each of the one or more subsets of the phenotypes or phenotype combinations, the probability of the information from the plurality of input units assuming that the genetic variant is causal for the subset of phenotypes or phenotype combinations is dependent on:

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

a subset of input units comprising the input units comprising information about the association between the plurality of genetic variants and one of the subset of phenotypes or phenotype combinations; and

a correlation between the effect sizes of the genetic variant on the phenotypes or phenotype combinations.

6. The method of any of claims 3 to 5, wherein the proportion of the plurality of genetic variants expected to be causal is predetermined.

7. The method of any of claims 3 to 6, wherein the correlation between the effect sizes of the genetic variant on the phenotypes or phenotype combinations is predetermined.

8. The method of any of claims 3 to 5, or 7, wherein the proportion of the plurality of genetic variants expected to be causal is updated at each iteration.

9. The method of any of claims 3 to 6, or 8, wherein the correlation between the effect sizes of the genetic variant on the phenotypes is updated at each iteration.

10. The method of any of claims 2 to 9, wherein the input units are determined from respective groups of individuals, and each of the plurality of probabilities is dependent on one or more parameters quantifying an overlap in the groups of individuals between respective pairs of input units.

11. 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 one or more phenotypes or phenotype combinations, and sampling values of the effect sizes from the probability distribution.

~~12. The method of claim 11, wherein the probability distribution is a multivariate normal distribution.~~

~~13~~12. The method of claim 11 ~~or 12~~, wherein the sampling of values of the effect size is performed using a Monte-Carlo Gibbs sampler.

~~14~~13. The method of ~~any of~~ claim~~s~~ 11 ~~to~~or ~~13~~12, 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.

~~15~~14. The method of any of claims 11 to ~~14~~13, wherein the probability distribution is dependent on a correlation between the effect sizes of the genetic variant on the phenotype or phenotype combinations.

~~16~~15. The method of claim ~~15~~14, wherein the correlation between the effect sizes of the genetic variant on the phenotypes or phenotype combinations is either predetermined or updated at each iteration.

~~17. The method of claim 15, wherein the correlation between the effect sizes of the genetic variant on the phenotypes or phenotype combinations is updated at each iteration.~~

~~18~~16. The method of any preceding claim, wherein determining the sampled effect sizes comprises using a model of causal relationships between the plurality of phenotypes or phenotype combinations.

~~19~~17. 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 phenotype or phenotype combination of each input unit;

the weighted effect sizes being the sampled effect size of the genetic variant on the phenotype or phenotype combination of 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.

~~20~~18. The method of any preceding claim, wherein carrying out one or more iterations comprises carrying out a predetermined number of iterations.

~~21~~19. 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.

~~22~~20. The method of any preceding claim, wherein the information about the association between the plurality of genetic variants and each of the phenotypes or phenotype combinations comprises, for each of the plurality of genetic variants, an estimate of a strength of association between the genetic variant and the phenotype or phenotype combination and an error in the estimate of the strength of association.

~~23~~21. A method of determining a polygenic risk score for a target phenotype or target phenotype combination 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 or target phenotype combination 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.

~~24~~22. 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 one of a plurality of phenotypes or phenotype combinations 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 for which of the plurality of phenotypes or phenotype

combinations the genetic variant is causal based on the plurality of input units; and

if the genetic variant is determined to be causal for one or more of the phenotypes or phenotype combinations, determining a sampled effect size of the genetic variant on the one or more phenotypes or phenotype combinations based on the plurality of input units and information about correlations between the plurality of genetic variants in the region of interest; and

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

~~25~~23. A computer program or a computer-readable medium comprising instructions which, when ~~the program is~~ executed by a computer, causes the computer to carry out the method of any of claims 1 to ~~23~~21.

~~26. A computer-readable medium comprising instructions which, when executed by a computer, causes the computer to carry out the method of any of claims 1 to 23.~~