A more accurate method for colocalisation analysis allowing for multiple causal variants

Chris Wallace[1,2]

[1]Cambridge Institute for Therapeutic Immunology and Infectious Disease and
[2]MRC Biostatistics Unit, University of Cambridge

Abstract

In genome-wide association studies (GWAS) it is now common to search for, and find, multiple causal variants located in close proximity. It has also become standard to ask whether different traits share the same causal variants, but one of the popular methods to answer this question, coloc, makes the simplifying assumption that only a single causal variant exists for any given trait in any genomic region. Here, we examine the potential of the recently proposed Sum of Single Effects (SuSiE) regression framework, which can be used for fine-mapping genetic signals, for use with coloc. SuSiE is a novel approach that allows evidence for association at multiple causal variants to be evaluated simultaneously, whilst separating the statistical support for each variant conditional on the causal signal being considered. We show this results in more accurate coloc inference than other proposals to adapt coloc for multiple causal variants based on conditioning or masking. We therefore recommend that coloc be used in combination with SuSiE to optimise accuracy of colocalisation analyses when multiple causal variants exist.

Introduction

Colocalisation is a technique used for assessing whether two traits share a causal variant in a region of the genome, typically limited by LD. In its original form, it made the simplifying assumption that the region harboured at most one causal variant per trait[1]. The approach proceeds by enumerating all *variant-level hypotheses* - the possible pairs of causal variants (or none) for the two traits - and the relative support for each in terms of Bayes factors. Each one of these combinations is associated to exactly one *global hypothesis*

H_0 : no association with either trait in the region

H_1 : association with trait 1 only

H_2 : association with trait 2 only

H_3 : both traits are associated, but have different causal variants

H_4 : both traits are associated and share the same causal variant

The Bayes factors for each of these global hypotheses may be calculated by summing the related Bayes factors for each variant-level hypothesis, and simple combination with prior probabilities of each hypothesis allows us to calculate posterior probabilities.

This simple summation is enabled by the single causal variant assumption, which implies that each pair of variants being causal are mutually exclusive events. However, the assumption is unrealistic, as multiple causal variants may exist in proximity. A suite of Bayesian fine mapping methods have been developed recently which calculate posterior probabilities of sets of causal variants for a given trait[2, 3, 4]. However, the marginal posterior probabilities calculated from these are no longer mutually exclusive events, so they could not be easily adapted to the colocalisation framework. Instead, all possible combinations of models between two traits could be considered, but this combinatorial problem is computationally expensive[5].

The single causal variant assumption was recently relaxed instead by adopting an approach of repeatedly conditioning on the top signal to identify secondary signals, and attempting to colocalise each pair of signals between the traits[6]. This allows the simple combination of Bayes factors through summation, but explicitly assumes that the causal variants are known in the conditioning. This is both untrue and runs counter to the goal of colocalisation which is to assess the chance of causal variant sharing without needing to specify causal variant identity - after all, if we knew the identity of the causal variant(s), no analysis would be required. Further, the stepwise regression approach upon which conditioning is based is also known generally to produce potentially unreliable results[7], a phenomenon that can be exacerbated by the extensive correlation between genetic variants caused by linkage disequilibrium (LD)[5]. Thus, this solution remains unsatisfactory.

A related approach, masking, has also been proposed, which instead of conditioning on the most likely causal variant, simply hides that variant and those in some LD with it[6]. This approach was developed to enable coloc to be applied to multiple causal variant datasets without necessitating aligning alleles in the GWAS datasets with those in the reference dataset used to estimate an LD matrix, but suffers from the same issues as conditioning.

Recently, the Sum of Single Effects (SuSiE) regression framework[8] was developed which reformulates the multivariate regression and variable selection problem as the sum of individual regressions each representing one causal variant of unknown identity. Conditional on the regression being considered, the variant-level hypotheses are again mutually exclusive. This allows the distinct signals in a region to be considered simulataneously, and enables quantification of the strength of evidence for each variant being responsible for that signal. Here we describe the adaptation of coloc to use the SuSiE framework and demonstrate improved efficacy over the previously proposed approaches, conditioning and masking.

Methods

Adaptation of coloc approach

SuSiE returns a matrix of posterior probabilities, with rows corresponding to regressions, and columns to variants. Variant-level Bayes factors for each detected signal can be back-

calculated by noting that

$$\text{prior odds for variant } i \text{ to be causally associated} = \frac{\pi_i}{1-\sum \pi_i}$$

$$\text{posterior odds for variant } i \text{ to be causally associated} = \frac{P_i}{P_0}$$

$$\text{Bayes factor comparing hypotheses that variant } i \text{ is causally associated to not causally associated} = \frac{\text{posterior odds for variant } i}{\text{prior odds for variant } i}$$

where π_i is the prior probability of causality for variant i which may be fixed or estimated internally by SuSiE, P_i is the posterior probability that variant i is causal, calculated by SuSiE, and P_0 is the posterior probability that there is no association for this regression (H_0). Typically we do not distinguish *a priori* between variants, and $\pi_i = \pi$ for all i. The vectors of Bayes factors thus defined (one vector per regression for which P_0 is small) can then be analysed in the standard coloc approach, for every pair of regressions across traits. The new `coloc.susie` function in the coloc package (`https://github.com/chr1swallace/coloc/tree/susie`) takes a pair of summary datasets in the form expected by other coloc functions, runs SuSiE on each (by calling functions in the susieR package) and performs colocalisation as described.

Decreasing the computational burden

While SuSiE has been shown to have greater accuracy than other fine mapping approaches[8], it becomes computationally expensive as the number of variants in a region increases because the number of potential models increases exponentially. Both coloc and fine mapping require dense genotyping data to make an accurate assessment, so computational complexity can become a considerable burden in larger genomic regions. In such regions, we propose that the fine mapping result can be approximated by running SuSiE only on a subset of variants with some weak evidence for association (eg excluding those with p values above 0.5), and setting the Bayes factors at those variants not considered to the minimum Bayes factor over all other variants. We use the term "trimming" to describe this approach hereafter.

Simulation strategy

We examined the performance of the approximation described above to decrease the computational burden, and of using SuSiE for colocalisation by simulation. To investigate the validity of the approximation, we repeatedly simulated GWAS summary data for a single trait in small genomic regions (1000 SNPs) by resampling haplotypes from public 1000 Genomes data. We simulated 1000 such examples, and analysed each dataset twice, once with the approximation and once using all SNPs, and calculated the difference between posterior inclusion probabilities (PIP, the probability that the variant is included in the true multi-SNP causal model) at the (simulated) causal variants from each run.

For colocalisation, we simulated data for two traits in the same way, such that each trait had one or two causal variants and each pair of traits shared zero, one or two causal variants. We simulated 10,000 examples from each collection, with each example analysed independently. Analysis compared different approaches:

1. single causal variant coloc analysis of every pair of traits

2. multiple causal variant coloc analysis using a conditioning approach to allow for multiple causal variants
3. multiple causal variant coloc analysis using a masking approach to allow for multiple causal variants
4. multiple causal variant coloc analysis using SuSiE to allow for multiple causal variants, including data trimming based on $|Z|$ score

Recall that coloc does not directly identify the causal variant, but it does provide posterior probabilities for each variant to be considered causal, given that H_4 is true. We labelled each comparison considered by coloc according to which pair of causal variants most closely corresponded to according to the r^2 between the causal variants and the lead trait variant reported by coloc (that with the largest Bayes factor). If r^2 between the reported variant and a specific causal variant j was > 0.5 *and* it was higher than the r^2 between the reported variant and any other causal variant, the reported variant was labelled "j", otherwise "unknown". We compared the average posterior probability profiles between methods, stratified according to this labelling scheme.

Results

First we assessed the impact of trimming data on the accuracy and speed of SuSiE. We found that trimming had a very minimal effect on PIP estimates at the causal variants in single causal variant data sets when SNPs with $|Z| < 1$ or lower were trimmed, but that errors begain to accumulate on larger regions (3000 SNPs) with 2 causal variants, where a threshold of $|Z| < 0.5$ might be preferred (Fig 1). Either of these thresholds reduced the median time for a SuSiE run per region more than ten fold (Fig 2).

The results of the coloc simulation study are given in Supplementary Table 1, and presented graphically in Fig 3. We found that inference with SuSiE coloc was broadly equivalent to that with other approaches when both traits really did contain only a single causal variant (top two row sets of Fig 3). When either one or both traits had two causal variants (bottom two row sets of Fig 3), all methods apart from single coloc were broadly similar in terms of favouring H_4 when comparing truly colocalising signals ("AA" or "BB" comparisons). Single coloc tended to equivocate between H_3 and H_4 when testing AB-like signals (ie where the peak signals in each trait related to distinct causal variants) which should be inferred H_3. This is a known feature of coloc, which may detect the colocalising signal even when additional non-colocalising signals are present,[1] and as such is not strictly an error, but does produce inconclusive results with posterior support split between H_3 and H_4. Either conditioning or masking also tends to have this pattern, either because the AB-like signal is tested first and the colocalising signal has not been conditioned out, or because the colocalising signal was conditioned out but used an imperfect tag of the true causal variant to condition on, leaving a shadow of the truly colocalising signal.

SuSiE seems to resolve this issue, with AB-like comparisons clearly having strongest posterior support for H_4. Interestingly, results are similar across the range of different possible thresholds for trimming, with trimming producing a slight improvement in coloc accuracy, particularly in regions with more SNPs.

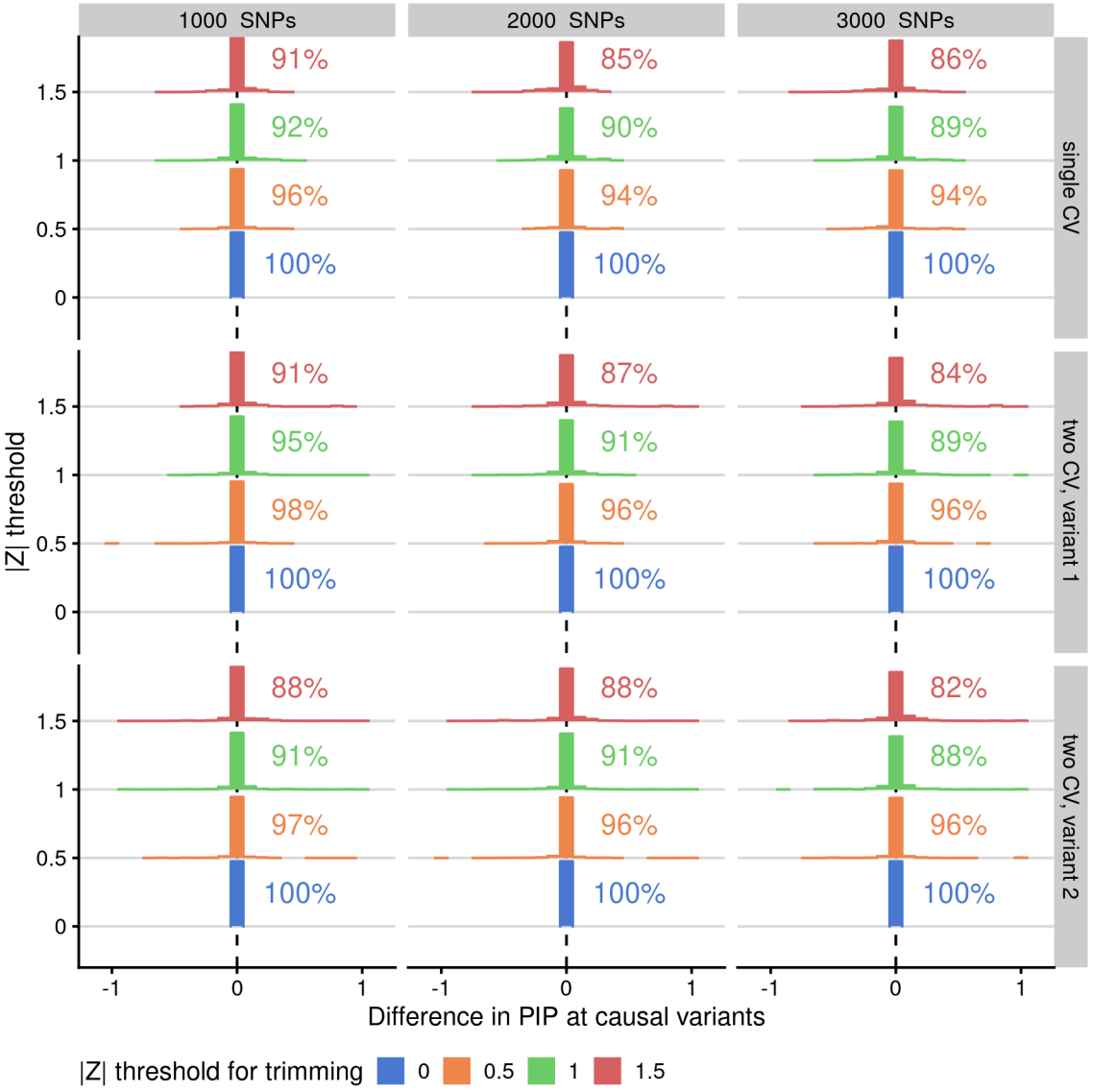

Figure 1: Histograms of the difference in PIP estimates at the causal variants between analysis with the full model and data trimmed to $|Z|$ above some threshold. The percentage of observations falling in the central column corresponding to the smallest PIP difference of $-0.05 < \text{PIP} < 0.05$ is shown. We include a threshold of 0 to demonstrate that results are consistent across multiple runs with the same full data. Datasets differ in the number of SNPs in a region (1000, 2000, 3000), and the number of causal variants ("CV", 1 or 2). "variant" indicates which of the two causal variants the PIP is estimated for in the case of 2 causal variants.

Figure 2: Time to run SuSiE per region in relation to the number of SNPs in the region (1000, 2000 or 3000), the number of causal variants (1 or 2) and the threshold used to trim SNPs by their —Z— scores. The point shows the median time over 1000 simulations, and the vertical range its interquartile range.

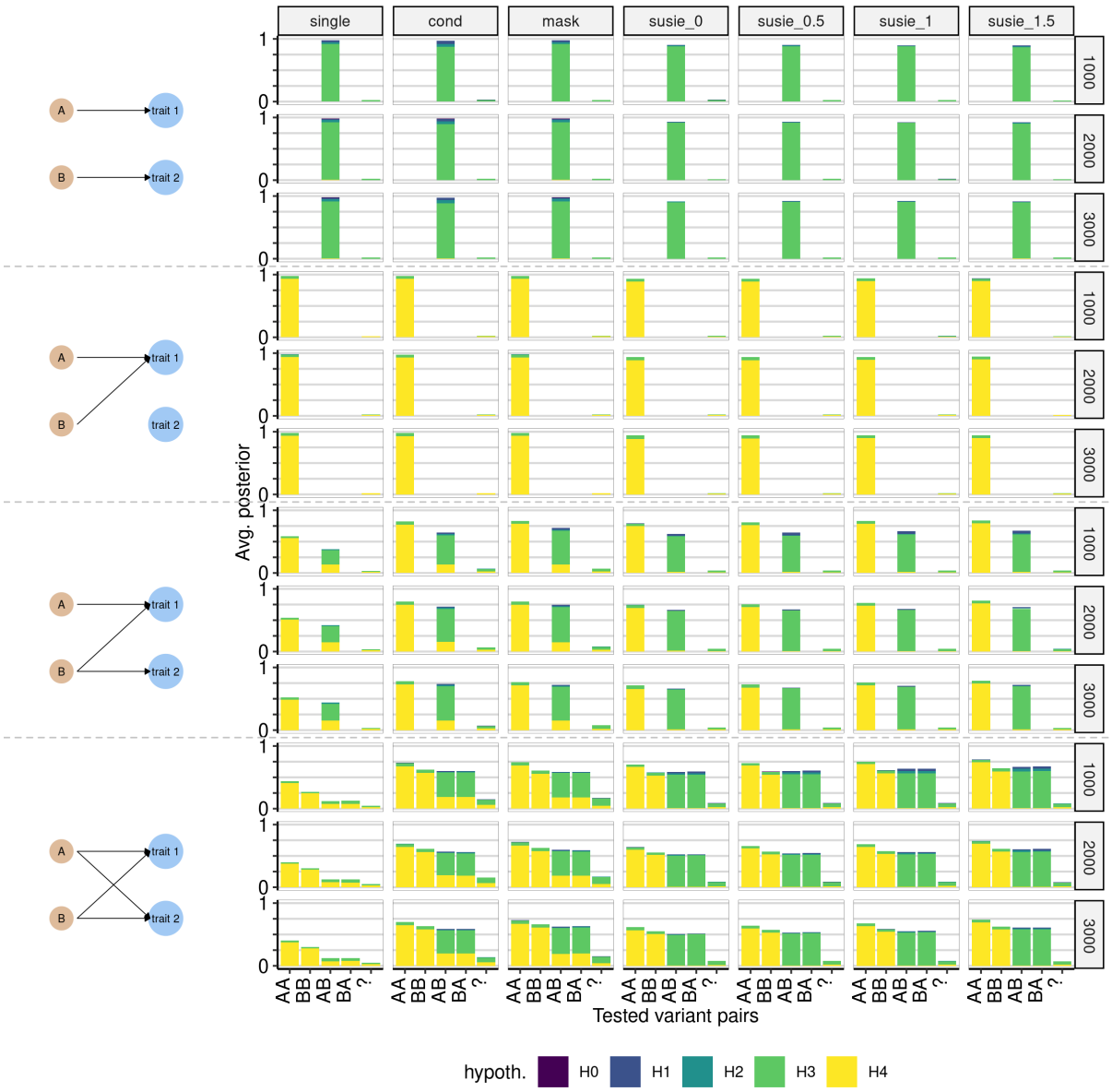

Figure 3: Average posterior probability distributions in simulated data. The four classes of simulated datasets are shown in three rows, with the scenario indicated in the left hand column. For example, the top row shows a scenario where traits 1 and 2 have distinct causal variants A and B. Within each scenario, there are three rows, corresponding to 1000, 2000 or 3000 SNP regions. Columns indicate the different analysis methods, with susie_x indicating that SuSiE was run with data trimmed at $|Z| < \text{x}$. For any method except "single", more than one colocalisation test may be performed. We estimated which pair of variants were being tested according to the LD between the variant with highest fine mapping posterior probability of causality for each trait and the true causal variants A and B. If $r^2 > 0.8$ between the fine mapped variant and true causal variant A, and r^2 with A was higher than r^2 with B, we labeled the test variant A, and vice versa for B. Where at least one test variant could not be unambigously assigned, we labelled the pair "?". The shaded proportion of each bar corresponds to the average posterior for the indicated hypothesis, and the total height of each bar has been scaled to the proportion of comparisons that were run, out of those that could have been conducted, and typically does not reach 1 because there is not always power to perform all possible tests.

Discussion

While coloc has been a popular method for identifying sharing of causal variants between traits, the common simplifying assumption of a single causal variant has been criticised, because it does not accord with findings that causal variants for the same trait may cluster in location (e.g. because they act via the same gene)[9]. Using the new SuSiE framework appears to resolve this issue better than the previously proposed conditional approach. It allows multiple signals to be distinguished, and then colocalisation analysis conducted on all possible pairs of signals between the traits.

Despite the adoption of a novel iterative procedure to fit the SuSiE model, the procedure is still slow for large regions with many SNPs, which can be a barrier to its adoption for a technique like coloc which has always boasted speed as an advantage. As only SNPs with some posterior support for causality can contribute to colocalisation comparisons, we considered approximating the SuSiE posterior by using a trimmed set of data, discarding SNPs with $|Z|$ scores below some small threshold, on the assumption that a causal SNP with detectable association should produce a Z score of reasonable magnitude. (For reference, whilst we only consider discarding SNPs with $|Z| < 1.5$ at most, the standard genome-wide significance threshold of $p < 5 \times 10{-}8$ corresponds to $|Z| > 5.45$). Thus, this approximation makes the assumption that true causal variants will have at least some weak marginal evidence of association, and we note that it is possible to construct examples which will violate this assumption, for example if two causal variants in strong LD but with opposite directions of effects exist. Given the trend for more limited differences in PIP estimates with smaller trimming thresholds and the slightly improved performance of coloc with larger thresholds, we suggest a threshold of $|Z| < 1$ is acceptable to allow SuSiE coloc to run at speed in larger regions, but leave the threshold as a user-set parameter which we recommend should be reported along with any results. Coloc benefits from comparing posterior probabilities across SNPs for two traits which may enable some robustness against inaccuracies. However, this does apply to single trait fine mapping and therefore our results do not imply that in general SuSiE fine mapping studies will benefit from trimming data.

Availability

Code to perform the simulations may be found at `https://github.com/chr1swallace/coloc-susie-paper`.

A version of coloc including SuSiE is at `https://github.com/chr1swallace/coloc/tree/susie`.

Acknowledgements

We thank Stasia Grinberg and Anna Hutchinson for comments on an earlier version of this manuscript.

CW is funded by the Wellcome Trust (WT107881) and the MRC (MC_UU_00002/4) and supported by the NIHR Cambridge BRC (BRC-1215-20014). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

This research was funded in whole, or in part, by the Wellcome Trust [WT107881]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014). URL `http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004383`.

[2] Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016). URL `http://dx.doi.org/10.1093/bioinformatics/btw018`.

[3] Newcombe, P. J., Conti, D. V. & Richardson, S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet. Epidemiol.* **40**, 188–201 (2016). URL `http://dx.doi.org/10.1002/gepi.21953`.

[4] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014). URL `http://dx.doi.org/10.1534/genetics.114.167908`.

[5] Asimit, J. L. *et al.* Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nature Communications* **10**, 3216 (2019). URL `https://www.nature.com/articles/s41467-019-11271-0`.

[6] Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics* **16**, e1008720 (2020). URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008720`.

[7] Miller, A. J. Selection of Subsets of Regression Variables. *J. R. Stat. Soc. Ser. A* **147**, 389–425 (1984). URL `http://www.jstor.org/stable/2981576`.

[8] Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273–1300 (2020). URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12388`.

[9] Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012). URL `http://dx.doi.org/10.1038/ng.2213`.