Article

Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues

BROWN, Andrew Anand, *et al*.

BROWN, Andrew Anand, *et al*. Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics*, 2017, vol. 49, no. 12, p. 1747-1751

PMID : 29058714
DOI : 10.1038/ng.3979

LETTERS

nature genetics

Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues

Andrew Anand Brown[1–4], Ana Viñuela[1–3], Olivier Delaneau[1–3], Tim D Spector[5], Kerrin S Small[5] & Emmanouil T Dermitzakis[1–3]

Genetic association mapping produces statistical links between phenotypes and genomic regions, but identifying causal variants remains difficult. Whole-genome sequencing (WGS) can help by providing complete knowledge of all genetic variants, but it is financially prohibitive for well-powered GWAS studies. We performed mapping of expression quantitative trait loci (eQTLs) with WGS and RNA-seq, and found that lead eQTL variants called with WGS were more likely to be causal. Through simulations, we derived properties of causal variants and used them to develop a method for identifying likely causal SNPs. We estimated that 25–70% of causal variants were located in open-chromatin regions, depending on the tissue and experiment. Finally, we identified a set of high-confidence causal variants and showed that these were more enriched in GWAS associations than other eQTLs. Of those, we found 65 associations with GWAS traits and provide examples in which genes implicated by expression are functionally validated as being relevant for complex traits.

Genome-wide association studies (GWAS) have uncovered thousands of genetic associations between regions of the genome and complex traits[1], but moving from the associations to identifying the underlying mechanisms has proven complicated[2]. Statistical associations between traits and genomic regions indicate a variant with a causal effect on the trait, because reverse causation or unmeasured confounders modifying DNA can be ruled out (i.e., causal effects are interpreted in the probabilistic sense, in which a direct intervention modifying one factor has consequences on another). A first step for understanding the mechanism would be to identify the exact variant, because knowing the exact localization would allow for exploration of the transcription-factor-binding sites and regulatory elements affected. However, such efforts are complicated because most loci tested in GWAS are not directly measured but instead are imperfectly imputed[3]. Although WGS does directly ascertain all genotype calls, in spite of falling costs, it remains very expensive to perform for the sample sizes used in modern GWAS (**Supplementary Table 1**). In contrast, eQTL studies linking variants and gene expression have discovered thousands of associations by using several hundreds of samples, a scale at which collecting WGS data is feasible[4].

Here, we describe analyses combining two previously published data sets derived from individuals in the TwinsUK cohort: RNA-seq from four tissues[5,6] and WGS from the UK10K project[7]. (Previously, gene expression quantified with microarrays[8] has been combined with the same WGS data set to corroborate specific GWAS associations[9,10].) We explored the properties of causal variants by using simulations, and we propose the 'causal-variant evidence mapping using nonparametric resampling' (CaVEMaN) method to estimate the probability that a variant most associated with the expression trait is causal for that association. We successfully used this method to produce a robust set of likely causal SNPs. Hence, CaVEMaN may provide an important resource for developing methods to call personalized regulatory variants from WGS and sequence annotations.

With WGS, genotypes are directly measured at far more sites than are available on current genotyping chip arrays (although sites on a genotyping chip are typically measured with more accuracy). The 1000 Genomes Project has estimated that >99% of SNPs are observed with minor allele frequencies >1%. For low-coverage sequencing and genotyping arrays, imputation methods are frequently used to impute better-quality calls at sites with no coverage on the arrays and low or no coverage with sequence data. The degree, if any, to which sequence information at more sites can decrease imputation noise and increase power to map eQTLs is currently unknown. For a simple comparison, we mapped independent eQTLs within 1 Mb of the transcription start site for protein-coding genes and long noncoding RNAs in four tissues (fat, lymphoblastoid cell lines (LCLs), skin, and whole blood), using individuals for whom expression, sequence and genotype array data were all available (n from 242 (whole blood) to 506 (LCLs)). We identified 27,659 independent autosomal eQTLs affecting 11,865 genes by using WGS (8,690,715 variants), and 26,351 affecting 11,642 genes by using genotypes called from arrays and imputed into the

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [2]Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland. [3]Swiss Institute of Bioinformatics, Geneva, Switzerland. [4]NORMENT, KG Jebsen Centre for Psychosis Research, Oslo University Hospital, Oslo, Norway. [5]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. Correspondence should be addressed to A.A.B. (andrew.brown@unige.ch) or E.T.D. (emmanouil.dermitzakis@unige.ch).

Received 21 November 2016; accepted 27 September 2017; published online 23 October 2017; doi:10.1038/ng.3979

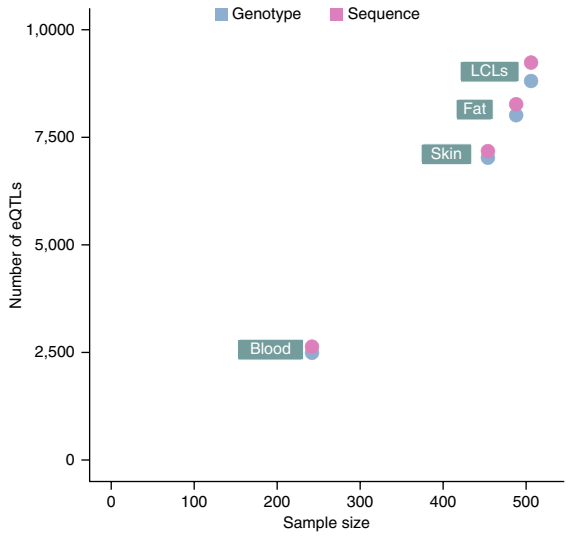

Figure 1 eQTL discovery with different genotyping technologies. Number of autosomal eQTLs discovered in each tissue when genotype information is provided by arrays imputed into a reference panel (genotype) and by WGS (sequence). There is a modest (3.7%) increase in the total number of eQTLs discovered with WGS over all tissues.

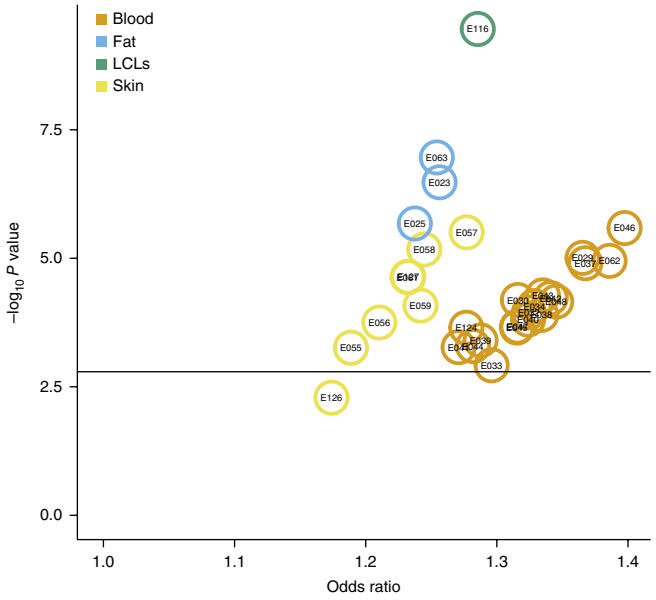

Figure 2 Relative enrichment in eQTLs discovered with different genotyping technologies in functional regions. Odds ratios and *P* values for enrichment (two-tailed Fisher's exact test) of LEVs called from sequences located in DHSs[11] relative to LEVs called from array-derived genotypes. A total of 31 experiments related to the tissue from which RNA-seq data were collected were analyzed. The code given relates to the Roadmap Epigenomics code; **Supplementary Table 2** shows the original experiments. All but enrichment of skin eQTLs in DHSs assayed in NHDF-Ad adult dermal fibroblast primary cells were Bonferroni significant (two-tailed Fisher's test, $P<0.05$).

1000 Genomes Project Phase 1 reference panel (6,263,243 variants) (**Fig. 1**; analysis of all individuals with expression and WGS data (*n* from 246 to 523) and including the X chromosome identified 28,141 eQTLs affecting 12,243 genes). This result corresponded to only a 3.7% increase in discovered eQTLs by using WGS. Given that the cost of collecting the data is at least tenfold higher with WGS, this procedure does not currently seem worthwhile. This demonstrates the ability of imputation approaches to accurately assay common variation, particularly because the denser genotyping arrays and larger reference panels now available would decrease and possibly even remove this difference (more details on imputation accuracy in Online Methods).

We frequently observed that the lead eQTL variant (LEV, the variant most associated with the trait) differs between the two data sets. Because genotypic uncertainty should be lower for WGS, we presumed that the WGS LEVs should be the causal variant more frequently than LEVs from genotype arrays. To test this hypothesis, we searched for enrichment of WGS-derived LEVs relative to array-genotype -derived LEVs in biochemically active regions of the genome. Indeed, for 30 out of 31 experiments carried out by the Roadmap Epigenomics Consortium[11] in relevant tissues, there was significant enrichment of sequence LEVs compared with genotype LEVs in DNase I–hypersensitive sites (DHSs) (odds ratio 1.17–1.40; **Fig. 2**). From this result, we inferred that the LEVs called with WGS are more likely to be causal variants.

To better understand the properties of causal variants, we simulated expression data sets in which the causal variant was known and whose effect size, distance to the transcription start site, and minor allele frequency were matched to those of the LEVs from the original eQTL mapping with sequence genotypes. Repeating the eQTL mapping on these simulated data sets, we found that the causal variant was the LEV in 45% of cases. This number was consistent across tissues, despite the sample size and power to map eQTLs being much lower for whole blood (**Supplementary Fig. 1**). This number was also similar to that obtained from the analysis of the Geuvadis data (55%) through a different methodology[4]. We also observed a rapid decline for lower-ranked candidate variants: the tenth most associated SNP was causal in only 1% of cases.

Our simulations showed that, across all genes, the LEV was a strong candidate for the causal variant. However, for specific LEVs, causality depends on the linkage-disequilibrium structure around the true causal variant and phenotypic uncertainty in expression of the particular gene. For these reasons, we developed the CaVEMaN method, which uses bootstrap methods similar to those previously proposed by others[12,13] to estimate the probability that the LEV is the causal variant (details in Online Methods).

We applied the CaVEMaN method to all four tissues and the Geuvadis LCL RNA-seq data ($n = 445$; results in **Supplementary Data Set 1**). The distributions of probabilities of LEVs being causal were similar across tissues and studies (**Fig. 3**). For 7.5% of the eQTLs, the LEV had $P>0.8$ of being the causal variant; we refer to those as high-confidence causal variants (HCCVs). For comparison, we applied the CAVIAR method[14] to the largest data set (TwinsUK LCLs) and applied dap-g[15] to simulated data (details in Online Methods).

To understand more about the relationship between causal regulatory variation and active genomic regions found by chromatin immunoprecipitation coupled with DNA sequencing (ChIP–seq) in single individuals, we integrated our causal probabilities with DHSs from the Roadmap Epigenomics Consortium. We observed a simple linear relationship between the causal probability of the LEV and the probability of the LEV being located in a DHS (**Fig. 4**) (although low-probability blood eQTLs ($P<0.25$) were found less often in DHSs than expected by the linear model, possibly because these LEVs were less reliable because of the smaller sample size). We exploited the linear relationship to estimate the proportion of regulatory variants

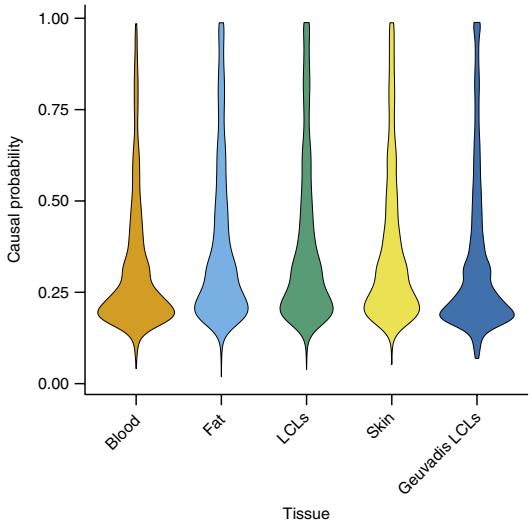

Figure 3 Distribution of the CaVEMaN estimated causal probabilities for LEVs. The distribution of causal probabilities for all LEVs discovered in each of the tissues is shown.

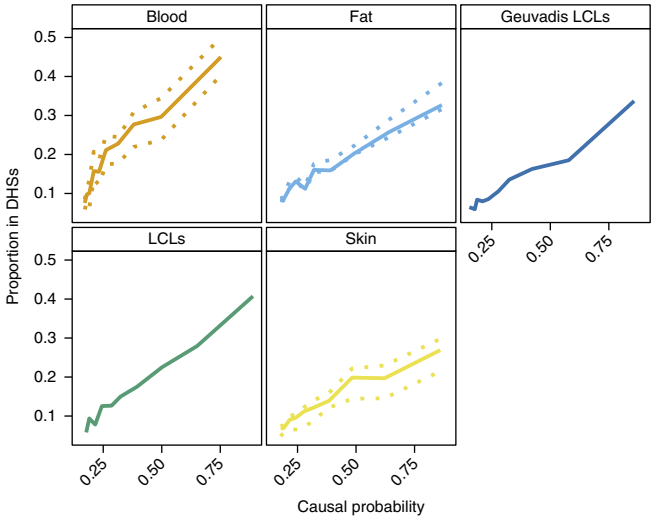

Figure 4 Proportion of LEVs in DHS regions, plotted against causal probability. LEVs were divided into ten equally sized groups on the basis of causal probability, and the proportion in DHS regions was calculated for each group and each experiment. The complete line represents the median result across experiments; when there was more than one experiment for a given tissue, the dotted lines show the maximum and minimum across experiments. We observed a linear relationship between the two probabilities. A full list of experiments can be found in **Supplementary Table 2**.

with a causal probability of 1 that were located within DHSs identified by particular experiments. For all tissues except blood, only a minority of regulatory variants were within DHSs called by specific experiments (**Fig. 5**). Blood eQTLs, discovered in a smaller sample size than those in the other tissues, had larger effect sizes and thus were more likely to affect promoter activity, thus providing a possible explanation for the observed greater enrichment. If CaVEMaN were applied to larger eQTL data sets with the power to discover eQTLs with more subtle effects, the proportion of causal regulatory variants in DHSs might possibly be even lower, thus implying limited utility of regulatory annotations for interpretation of enhancer and weaker regulatory variants.

It is widely known that associations with whole-organism traits, as discovered by GWAS, are enriched in eQTLs[16]; by defining a set of eQTLs for which the causal variant is known with high probability, these eQTLs may show greater enrichment (a shared GWAS–eQTL signal would not be diluted by linkage). In addition, by providing both a mediating gene and a variant causative for the expression signal, these results may provide a more mechanistic understanding of GWAS signal. We extracted *P* values for association for all of the LEVs from 16 GWAS studies with publicly available summary statistics (Online Methods) and observed greater enrichment of small *P* values for HCCVs compared with all other eQTLs (proportion of alternative hypotheses (π_1) = 16.2 compared with π_1 = 14.0, estimated with qvalue[17]). We also observed greater enrichment when considering the proportion of shared signals between GWAS associations with $P < 5 \times 10^{-8}$ listed in the National Human Genome Research Institute (NHGRI)–European Bioinformatics Institute (EBI) Catalog and eQTLs located in the same recombination hotspot (16.0% of proximal HCCVs and GWAS associations shared, and 2.49% for all other eQTLs, as estimated with the regulatory trait concordance method[18,19] (RTC)). We also found Bonferroni-significant GWAS associations between 53 HCCVs and 65 GWAS traits ($P < 3 \times 10^{-6}$; **Fig. 6** and **Supplementary Data Set 2**). Applying the coloc method to test whether the eQTL and GWAS trait were affected by the same causal variant[20], we observed 18 cases showing strong evidence of common genetic effects (coloc probability >0.95) and 29 cases with at least moderate evidence (coloc probability >0.7).

Given these examples of variants with high-confidence causal effects on expression and statistical associations with GWAS traits, functional evidence connecting the expression of the gene with the trait would also implicate a causal link between the variant and trait. For example, an HCCV (rs10274367; all rs IDs are as defined in dbSNP, build 148, GRCh37) associated with *GPER1* was also associated with levels of high-density-lipoprotein cholesterol (coloc estimate of shared causal variant = 0.999). Female knockout mice for the gene have lower high-density-lipoprotein levels than those in wild-type[21]. We also found rs1805081 to be a HCCV for *NPC1*, and it has been found to be the lead variant associated with body mass index in a large GWAS study[22] (coloc probability = 0.722). Heterozygous mouse models (*Npc1*$^{+/-}$), in which the gene is expressed at half normal levels, exhibit high weight gain when fed high-fat diets but not low-fat diets[23,24], and higher levels of *NPC1* in human adipose tissue have been found to normalize after bariatric surgery and behavioral modification[25]. In this example, the expression of *NPC1* is modified by rs1805081 and is hypothesized to be a response to changes in body mass index. Expression changes in *NPC1* appear to be part of a compensatory mechanism to modify weight gain due to dietary excess and result from diet–genotype interactions. Finally, we observed rs4702 as an HCCV for the *FURIN* gene in our analysis, and it has been found to be the lead variant in a GWAS study of schizophrenia[26], coloc probability = 0.999). Altering expression of *FURIN* produces neuroanatomical deficits in zebrafish and abnormal neural migration in human induced pluripotent stem cells[27].

This study produced a method for identifying causal variants influencing gene expression. Notably, association of an HCCV with a GWAS trait does not necessarily mean that both share a common causal variant or that the causal mechanism acts in the tissue under study. However, combining fine-mapping by using CaVEMaN with colocalization methods that formally test whether genetic variants affecting multiple traits are shared[18,20] and methods that aim to predict

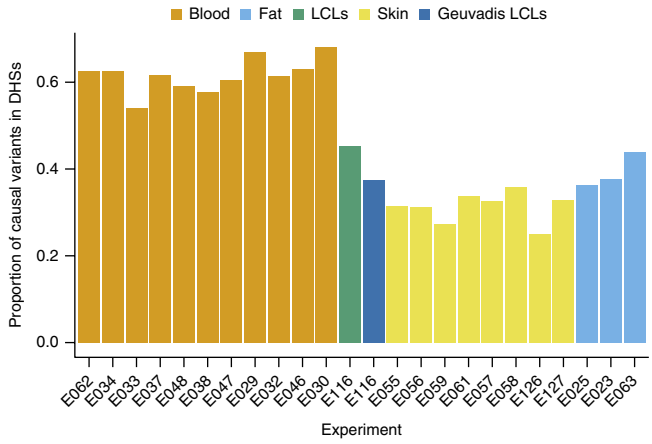

Figure 5 Proportion of functional variants in regions identified by single ChIP–seq experiments. We estimated the proportion of LEVs with a causal probability of 1 being located in functional regions defined by the Roadmap Epigenomics Consortium by extrapolating from the relationship observed in **Figure 4**. Blood showed the highest proportion of causal variants in annotated regions; for all other tissues, we estimated that only a minority of causal variants were in DHS regions.

causal tissues[19,28] may pinpoint precise variants, genes, and tissues underlying GWAS traits. In addition, methods for fine-mapping and for testing for colocalization share common features. Similarly to how a fine-mapping method (CAVIAR[14]) has been extended to test for colocalization (eCaviar[28]), CaVEMaN could also be extended to test for colocalization.

In summary, we produced a method to estimate the probability that a lead eQTL variant is the causal variant. We used this method to estimate the effectiveness of ChIP–seq experiments from a single individual in predicting regions containing regulatory variation, as well as to suggest variants that might be causal for GWAS associations. This method could also be applied to GWAS data to identify candidate causal variants for whole-organism traits. Pinpointing the causal variant in such studies should facilitate the integration of these association signals with mechanistic regulatory interactions and likely upstream regulators, and should also allow for the development of interpretation methods from genome sequence alone, after a large number of representative causal variants have been discovered.

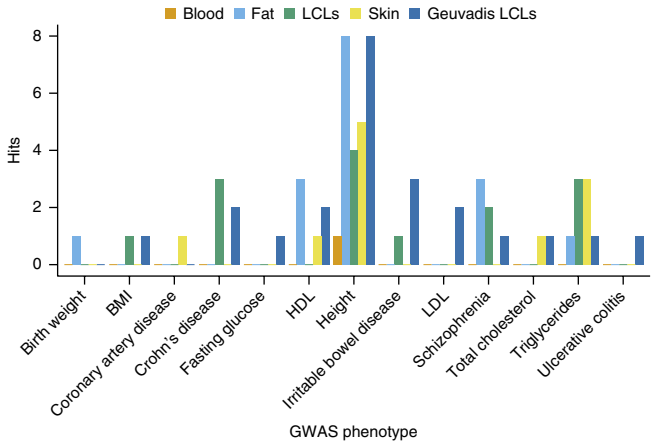

Figure 6 HCCVs statistically associated with GWAS traits. Numbers of Bonferroni-significant associations between HCCVs (causal probability >0.8) and GWAS traits, divided by tissue type. HCCVs showed more statistical associations with GWAS traits than other eQTLs, because cosegregating signals are not weakened by imperfectly captured markers.

URLs. Early Growth Genetics (EGG) Consortium, http://www.egg-consortium.org/; NHGRI–EBI GWAS Catalog, https://www.ebi.ac.uk/gwas/; 1000 Genomes Project, http://www.internationalgenome.org/data/; National Center for Biotechnology Information (NCBI) ftp site, ftp://ftp.ncbi.nlm.nih.gov/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank N. Lykoskoufis for assistance with the enrichment analysis. T.S. is supported as an NIHR Senior Research Fellow. This project was supported by a Helse Sør-Øst grant (2011060) to A.B. and an MRC Project Grant (L01999X/1) to K.S., and by grants from the NIH-NIMH (NIH-R01MH101814-GTEx), an IMI-Joint Undertaking of the European Commission (UE7-DIRECT-115317-1), the European Commission (UE7-EUROBATS-259749), the European Research Council (UE7-POPRNASEQ-260927), the Louis Jeantet Foundation, the Swiss National Science Foundation (31003A-149984 and 31003A-170096), and SystemsX (2012/201-SysGenetix) to E.T.D. The TwinsUK study was funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013) and the Medical Research Council. The study also received support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre, based at Guy's and St Thomas' NHS Foundation Trust, in partnership with King's College London. SNP genotyping was performed by The Wellcome Trust Sanger Institute and National Eye Institute via NIH-CIDR. This study used data generated by the UK10K Consortium. Funding for UK10K was provided by the Wellcome Trust under award WT091310. A full list of the investigators who contributed to the generation of the UK10K data is available at http://www.UK10K.org/. This research was supported by grants from the European Research Council. Computation was performed at the Vital-IT Center (http://www.vital-it.ch/) for high-performance computing of the SIB Swiss Institute of Bioinformatics.

AUTHOR CONTRIBUTIONS

A.A.B. and E.T.D. designed the study. A.A.B. ran the analyses. A.A.B., A.V., and E.T.D. interpreted the results. A.A.B., A.V., and E.T.D. wrote the manuscript. O.D. provided methodological suggestions. K.S.S. and T.D.S. contributed data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Spain, S.L. & Barrett, J.C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111–R119 (2015).
3. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
4. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
5. Brown, A.A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3**, e01381 (2014).
6. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
7. UK10K Consortium. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
8. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
9. Timpson, N.J. *et al.* A rare variant in *APOC3* is associated with plasma triglyceride and VLDL levels in Europeans. *Nat. Commun.* **5**, 4871 (2014).

10. Iotchkova, V. *et al.* Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat. Genet.* **48**, 1303–1312 (2016).
11. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
12. Lebreton, C.M. & Visscher, P.M. Empirical nonparametric bootstrap strategies in quantitative trait loci mapping: conditioning on the genetic model. *Genetics* **148**, 525–535 (1998).
13. Visscher, P.M., Thompson, R. & Haley, C.S. Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020 (1996).
14. 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).
15. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
16. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
17. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
18. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
19. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* http://dx.doi.org/10.1038/ng.3981 2017).
20. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
21. Sharma, G. *et al.* GPER deficiency in male mice results in insulin resistance, dyslipidemia, and a proinflammatory state. *Endocrinology* **154**, 4136–4145 (2013).
22. Meyre, D. *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* **41**, 157–159 (2009).
23. Jelinek, D., Heidenreich, R.A., Erickson, R.P. & Garver, W.S. Decreased *Npc1* gene dosage in mice is associated with weight gain. *Obesity (Silver Spring)* **18**, 1457–1459 (2010).
24. Jelinek, D. *et al. Npc1* haploinsufficiency promotes weight gain and metabolic features associated with insulin resistance. *Hum. Mol. Genet.* **20**, 312–321 (2011).
25. Bambace, C., Dahlman, I., Arner, P. & Kulyté, A. NPC1 in human white adipose tissue and obesity. *BMC Endocr. Disord.* **13**, 5 (2013).
26. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
27. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
28. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

ONLINE METHODS

TwinsUK data. *Expression and genotype data from arrays.* RPKM expression quantification and array genotype data used in this paper have been previously analyzed[5,6], and the production of these data is described in full in the **Supplementary Note**.

Genotypes called from sequencing. The vcf files produced by the UK10K Consortium[7] were downloaded from the European Genome-Phenome Archive. When one monozygotic twin in the sample had been sequenced, the same data were used for the genetically identical sibling. Of 856 individuals with expression, 552 had available sequence data (246 individuals had expression quantified in whole blood, 505 in adipose tissue, 523 in LCLs and 471 in skin). For multiallelic variants, dosage was calculated as two times the number of copies of the most common allele. Variants were filtered if the major allele had a frequency >0.99.

Ethics statement. The St. Thomas' Research Ethics Committee (REC) approved (on 20 September 2007) the protocol for dissemination of data, including DNA, with REC reference number RE04/015. On 12 March 2008, the REC confirmed that this approval extended to expression data. Volunteers provided informed consent and signed an approved consent form before the biopsy procedure. In addition, before the biopsy date, volunteers were mailed an appropriate detailed information sheet regarding the research project and biopsy procedure. Consent to link the RNA-seq data with the WGS data was approved by the TwinsUK Resource Executive Committee (TREC) on 22 April 2015.

Geuvadis data. BAM files for RNA-seq were downloaded from EBI ArrayExpress (accession code E-GEUV-3). The data were mapped to the GRCh37 reference genome[29] with GEM (version 1.7.1)[30], and protein-coding and long noncoding RNAs were quantified with GENCODE v19 annotations[31]. The population group was regressed out of RPKM values by using a linear model; values were centered and scaled to mean 0, variance 1; and 50 principal components were removed. Genotype vcf files from phase 3 of the 1000 Genomes Project[32] were downloaded from the 1000 Genomes Project website (URLs). In non-pseudo-autosomal regions of the X chromosome, male dosage was calculated as two times the number of copies of the alternative allele. A minor-allele-frequency cutoff of 0.01 was applied.

eQTL mapping. eQTLs were mapped with fastQTL, which tests for association between expression and genotype with a two-tailed Wald test[33]. To discover multiple independent eQTLs, a stepwise regression procedure was applied. First, for each tissue, fastQTL was run with 10,000 permutations to discover a set of eGenes (FDR <0.01). Then, the maximum beta-adjusted *P* value (with correction for multiple testing across SNPs) over these genes was taken as the gene-level threshold. The next stage proceeded iteratively for each gene. At each iteration, a *cis* scan of the window was performed, using 10,000 permutations and correcting for all previously discovered SNPs. If the beta-adjusted *P* value for the LEV was not significant at the gene-level threshold, the procedure moved on to the backward step. If this *P* value was significant, the LEV was added to the list of discovered eQTLs as an independent signal, and the forward step proceeded to the next iteration.

After the forward stage was completed for a given gene, a list of associated SNPs was produced, which we refer to as forward signals. The backward stage consisted of testing each forward signal separately, controlling for all other discovered signals. For each forward signal, we ran a *cis* scan over all variants in the window by using fastQTL, fitting all other discovered signals as covariates. If no SNP was significant at the gene-level threshold, the signal being tested was dropped; otherwise, the LEV from the scan was chosen as the variant that best represented the signal in the full model.

Properties of LEVs estimated with sequence and genotyping arrays. We investigated the differences between LEVs identified with sequence data and data from genotyping arrays to better understand the slight increase in power that we observed when using sequence data. The minor allele frequency of eQTLs called with sequence data was slightly lower than those identified with genotype data (median minor allele frequency of 26.0% compared with 27.4%; two-tailed Mann–Whitney *U*-test $P = 5.52 \times 10^{-21}$; **Supplementary Fig. 2**). We found that 3,383 out of 22,656 LEVs called on the basis of sequence were removed from the array data, owing to INFO scores <0.8; most of these LEVs failed imputation criteria based on the HumanHap300 array (3,334 failed on this array, 2,290 failed on the HumanHap610Q, and 2,241 failed on both; **Supplementary Fig. 3**). Finally, for the remaining 19,273 sequence LEVs for which the genotype imputation passed the quality filters, we observed good agreement between calls made with the two technologies, with a median proportion of different calls of only 0.94%. However, a small minority of LEVs (0.93%) showed a larger discrepancy between the two call sets, with more than 10% of individuals showing differences. Together, these results suggested that both genotyping arrays with more SNPs and larger reference panels that enumerate more haplotypes would further decrease the power differences between studies using sequencing and those using genotyping arrays.

Enrichment analysis. Bed files listing DHSs, produced by the Roadmap Epigenomics Consortium[11], were downloaded from the NCBI ftp site (URLs). Experiments were linked to tissues for which RNA-seq was available, as shown in **Supplementary Table 2**. Over each ChIP–seq/RNA-seq combination, the odds ratio for enrichment was calculated by using the number of LEVs that were called on the basis of sequence or array-based genotypes and located within regions called in the experiment and the total numbers of eQTLs. A two-tailed Fisher's exact test was performed to test the hypothesis that equal proportions of sequence and genotype LEVs were located in these regions.

Simulations. For all discovered eQTLs, the LEV for association was identified, and its minor allele frequency and distance to the transcription start site were calculated. Beta and sigma coefficients from a regression of expression on the LEV were also estimated. Then, a matched SNP was chosen with a distance to the transcription start site of a gene within 1 kb of the original and minor allele frequency within 0.025. Simulated expression was produced by multiplying the SNP genotype by beta and adding a random normally distributed term with a standard error of sigma. Five simulated data sets were produced for each TwinsUK tissue; eQTL mapping was applied to each during searching for only primary eQTLs; and the rank of the nominal *P* value for the causal variant was collected.

CaVEMaN. *A frequentist definition of causal probability.* Several methods have been proposed that use Bayesian methodologies to estimate the probability that a variant is causal for an effect on expression, combining prior distributions with likelihoods to estimate posterior probabilities[15,34,35]. We, however, used a frequentist definition of the probability of being causal. Causal probabilities were assigned to LEVs with the following property: if an eQTL was sampled randomly from the set of all eQTLs having a causal probability equal to a number x, the probability that a causal variant was chosen was equal to x. In this way, the results match the intuitive understanding of what a causal probability is: if a LEV is chosen at random, the probability that a causal variant will be chosen is equal to the estimate from CaVEMaN.

Learning parameter estimates from simulations. First, we used simulations in which a specific variant was chosen to act as the causal variant to estimate the probability that the causal variant would be the ith-ranked SNP in eQTL mapping. To do so, we calculated the proportion of times that this phenomenon occurred across all tissues and simulations (this quantity is denoted p_i; **Supplementary Fig. 1**). Because CaVEMaN focuses on the top ten ranked variants from an eQTL analysis, p_i values, with i from 1 to 10, were normalized to sum up to 1.

Multiple variants affecting expression of one gene. Previous fine-mapping approaches can be categorized into two classes: those that assume that only one genetic signal affects the phenotype[34] and those that map multiple genetic signals simultaneously[15,35]. CaVEMaN takes a different approach, in that the procedure is separated into two steps: first, a stepwise regression approach is used to estimate the number of eQTLs affecting the expression of the gene, and then each independent eQTL is mapped separately. The advantage of this method is that it provides a well-grounded statistical methodology for answering questions regarding multiple independent variables affecting expression and for addressing issues of multiple testing and significance.

After a set of eGenes and the independent eQTLs affecting them were identified, we created new 'single signal' expression phenotypes. For each eQTL, these were made by regressing out all other eQTLs discovered

doi:10.1038/ng.3979

for the gene, thus producing an expression phenotype reflecting the signal from only one eQTL.

Calculating the CaVEMaN score. This new matrix of expression data was sampled with replacement 10,000 times to create 10,000 new data sets of the same size. A *cis* eQTL mapping testing association using a two-tailed test for significant correlation was run on each of these data sets, and the proportion of times that a given SNP was ranked *i*, with *i* from 1 to 10, was calculated (denoted F_i, an estimate of the probability that the SNP would be the rank *i*th most associated SNP). The CaVEMaN score was defined as $\sum_{i=1}^{10} p_i F_i$; i.e., the sum of the product of the probability that the SNP was ranked *i* in an eQTL analysis with the probability that the *i*th-ranked SNP was causal for the association.

Calibrating CaVEMaN score for LEVs by using simulation. Finally, we further exploited the simulations to calibrate the CaVEMaN score of the LEV. CaVEMaN was run on all simulated data. Then, across all simulated data sets (with blood removed because it was an outlier resulting in less conservative estimates of causal probabilities), we divided the CaVEMaN scores of the LEVs into 20 quantiles. Within each quantile, we calculated the proportion of times that the lead SNP was the causal SNP and then drew a monotonically increasing smooth spline from the origin, through the 20 quantiles, to the point (1, 1), by using the gsl interpolate functions with the Steffen method (gsl-2.1; **Supplementary Fig. 4**). This function allowed us to map the CaVEMaN score of the lead SNP onto causal probabilities, and we applied this function to the CaVEMaN scores of the LEV to estimate their causal probabilities.

Validating the method with simulations in Geuvadis data. The CaVEMaN method uses parameters estimated from simulations based on UK10K expression data (primarily the distribution of ranks of causal eQTLs and the relationship between the CaVEMaN score and causal probability); hence, these simulations cannot later be used to validate the CaVEMaN estimates. We ran further simulations using the Geuvadis data to demonstrate that the estimates of the causal probability for the LEVs were well calibrated when parameters were estimated separately from the analyzed data set. We ran a total of five simulations, again using effect size and residual variance estimated from the original data. We plotted binned estimates of the estimated causal probabilities against the proportion of times that the LEV was the causal variant (**Supplementary Fig. 5**) and observed good agreement between our estimates and the true causal probabilities for these bins: the minimum, median, and maximum difference between the estimates and the true values were 0.0056, 0.036, and 0.071, respectively.

In addition, we ran a simulation to test the behavior of the model when there were weaker eQTL effects that were not detected by the original multiple-eQTL-mapping strategy. As before, we simulated a primary eQTL with minor allele frequency, effect size, and distance to the transcription start site matched to those of an eQTL discovered in the original analysis. Then, we randomly chose a second variant in the *cis* window, with minor allele frequency >0.05, and used this variant to simulate an extra eQTL effect on the phenotype, with an effect size one-half that of the primary eQTL. Then, a residual noise term was generated such that the primary eQTL explained the same proportion of variance as the original matched eQTL. We found that in estimating the causal probabilities, there was still good agreement between the primary eQTL and the known ground truth (**Supplementary Fig. 5**).

Comparing results from CaVEMaN with results from CAVIAR for TwinsUK LCL data. CAVIAR and equivalent Bayesian methods[36–39] have previously been suggested as fine-mapping methods for estimating credible sets of SNPs with a given probability of containing the causal variant. For genes with an eQTL in LCLs, we used CAVIAR[14] to produce another estimate of causal-variant probability for comparison. Because CAVIAR is limited in the number of SNPs that it can analyze, we first extracted all variants with $P < 0.01$, up to the first 50. The *Z* scores for these variants were produced with the correlation matrix of these SNPs, and CAVIAR was run with the default settings. There was good agreement in the causal probabilities of the LEV (Spearman $\rho = 0.856$, $P < 10^{-216}$; **Supplementary Fig. 6**), but the CAVIAR method produced estimates of the causal probabilities that were more conservative (median probability 0.12 versus 0.29). Because the CaVEMaN estimates were calibrated by using simulations, the CAVIAR estimates appeared to be, on average, underestimates of the true probabilities, possibly because of a combination of the priors not reflecting the true regulatory landscape and the sample size being insufficient to overcome this effect. CAVIAR does not suggest adjusting the priors when studying expression rather than GWAS trait associations, despite the fundamentally different genetic architectures and sample sizes between these types of studies. The approach of calibrating estimates of probabilities by using simulations could also be easily extended to other fine-mapping methods such as CAVIAR.

Comparison of simulation results from CaVEMaN with those from dap-g. We compared the results from CaVEMaN applied to one of the simulation data sets with the results from dap-g[15], a method recently proposed for fine-mapping. For each simulated gene expression, all SNPs in the *cis* window were extracted, and dap-g was run, specifying the option -ld_control 0.25. Then, for a comparable estimate of the posterior probability of the LEV, we extracted the highest posterior probability of any single-variant model and conditioned this probability on only one genetic signal by dividing it by the sum of the posterior probabilities of all single-SNP models. The two methods identified exactly the same sets of LEVs, and there was good agreement between the estimates of causal probabilities (Spearman $\rho = 0.95$, $P < 10^{-216}$). However, plotting the causal probabilities against the proportion of LEVs that were the causal variants indicated that dap-g underestimates this quantity (**Supplementary Fig. 5**).

Application of simulations to other data sets. The Geuvadis data set differs in many aspects from the TwinsUK data on which the CaVEMaN method was trained. Geuvadis samples were sequenced in multiple laboratories rather than just one; Geuvadis uses a multiethnic cohort, thus implying a different linkage structure in the genome; a different mapper (a splice-aware mapper) was used to quantify the data; and the tissue type, sample size, and ability to map eQTLs were all different from those of three out of four TwinsUK tissues. Our results thus indicated that the parameters estimated in TwinsUK were robust to a range of factors. However, in the future, similar data sets with thousands of samples are expected, and it is possible that our proposed method may not generalize to that case. For this reason, we provide methods to repeat these simulations in new data sets, as described on our accompanying website ('Code availability' section).

Statistical associations between eQTLs and GWAS traits from summary statistics. We downloaded the GWAS summary statistics for 16 different GWAS traits: autism[40], birth weight[41], body mass index (analyzing all ancestries)[42], coronary artery disease[43], Crohn's disease[44], diabetes[45], fasting glucose[46], fasting insulin[46], height[47], high-density lipoprotein[48], irritable bowel disease[44], low-density lipoprotein[48], schizophrenia[26], total cholesterol[48], triglycerides[48], and ulcerative colitis[44]. Data on the birth weights were contributed by the EGG Consortium using the UK Biobank Resource (URLs). For all LEVs, the *P* value for each trait was extracted (if available), and the qvalue package[17] was used to estimate π_1, the proportion of alternative hypotheses (i.e., association between variant and GWAS trait). Finally, Bonferroni-significant GWAS associations for HCCVs were reported, with controlling for multiple testing across all phenotypes and variants.

Testing HCCVs associated with GWAS traits for cosegregation with coloc. For HCCVs significantly associated with GWAS traits, we used the coloc method[20] to test the hypothesis of a shared causal mechanism. *P* values for association, available for both expression and GWAS associations, were extracted in a 200,000-bp region around the eQTL. Minor allele frequencies for the variants were extracted from the 1000 Genomes Phase 3 release[32]. After running coloc, we reported the probability of a shared causal variant for both associations, conditional on genuine associations existing for both traits (P(H3)/(P(H3) + P(H4)) reported by coloc).

Regulatory trait concordance (RTC) method for testing for cosegregation with NHGRI–EBI Catalog GWAS associations. We downloaded the NHGRI–EBI Catalog of reported genome-wide-significant associations (URLs) in September 2016 and removed all SNPs with $P > 5 \times 10^{-8}$ and those for which the variant was not listed in dbSNP (build 148)[49], thus leaving 11,636 reported associations. RTC, implemented in QTLtools[50], was applied with the default settings to assess sharing of these GWAS variants with eQTLs. Because the RTC statistic is uniformly distributed under the null hypothesis of two separate causal loci independently located within the hotspot, 1 – RTC can

be interpreted as a *P* value for a shared causal variant. The qvalue package[17] estimated π_1, the proportion of GWAS/eQTLs signals in the same recombination interval with the same causal variant.

Code availability. Code for correcting the expression data sets for multiple eQTLs, running the CaVEMaN method, converting the CaVEMaN score to a causal probability, and repeating simulations on new data sets can be found at https://github.com/funpopgen/CaVEMaN/.

Accession codes. BAM files for the RNA-seq are available from EBI ArrayExpress (accession code E-GEUV-3; Geuvadis cohort) and the European Genome-Phenome Archive (study ID EGAS00001000805; TwinsUK cohort). WGS data are available from the European Genome-Phenome Archive (study ID EGAS00001000108; TwinsUK) and the 1000 Genomes Project (URLs).

Data availability. Data are available from the corresponding authors upon reasonable request.

A **Life Sciences Reporting Summary is available.**

29. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
30. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
31. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
32. 1000 Genomes Project Consortium. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
33. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
34. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).
35. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).
36. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).
37. The International Multiple Sclerosis Genetics Consortium. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
38. Chen, W. *et al.* Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).
39. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
40. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552–555 (2016).
41. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).
42. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
43. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
44. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
45. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
46. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
47. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
48. Willer, C.J. *et al.*; Global Lipid Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
49. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
50. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

doi:10.1038/ng.3979

nature research

Corresponding Author: Andrew Brown

Date: Jul 14, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined. Sample size is larger than in comparable studies.

2. Data exclusions

   Describe any data exclusions. No data were excluded.

3. Replication

   Describe whether the experimental findings were reliably reproduced. n/a

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. n/a

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. n/a

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☒	☐	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

   Describe the software used to analyze the data in this study. Code is available on github: https://github.com/funpopgen/CaVEMaN

Nature Genetics: doi:10.1038/ng.3979

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used.
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used.
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used.
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Relevant information on the cohort is listed in Online methods.

Nature Genetics: doi:10.1038/ng.3979