nature genetics

Meta-analysis of gene-level tests for rare variant association

Dajiang J Liu[1,19], Gina M Peloso[2–4,19], Xiaowei Zhan[1,19], Oddgeir L Holmen[5,6,19], Matthew Zawistowski[1], Shuang Feng[1], Majid Nikpay[7], Paul L Auer[8,9], Anuj Goel[10,11], He Zhang[12,13], Ulrike Peters[8,14], Martin Farrall[10,11], Marju Orho-Melander[11,15], Charles Kooperberg[8,16], Ruth McPherson[7], Hugh Watkins[10,11], Cristen J Willer[12,13], Kristian Hveem[5,17], Olle Melander[11,15], Sekar Kathiresan[2–4,18,20] & Gonçalo R Abecasis[1,20]

The majority of reported complex disease associations for common genetic variants have been identified through meta-analysis, a powerful approach that enables the use of large sample sizes while protecting against common artifacts due to population structure and repeated small-sample analyses sharing individual-level data. As the focus of genetic association studies shifts to rare variants, genes and other functional units are becoming the focus of analysis. Here we propose and evaluate new approaches for performing meta-analysis of rare variant association tests, including burden tests, weighted burden tests, variable-threshold tests and tests that allow variants with opposite effects to be grouped together. We show that our approach retains useful features from single-variant meta-analysis approaches and demonstrate its use in a study of blood lipid levels in ~18,500 individuals genotyped with exome arrays.

Proceeding from the discovery of a genetic association signal to mechanistic insight into human biology should be much easier for alleles with a clear functional consequence, including nonsynonymous, splice-altering and protein-truncating alleles. Most of these alleles are very rare, with only one such allele expected to reach minor allele frequency (MAF) of >5% in the average human gene[1]. Recent advances in exome sequencing and the development of exome genotyping arrays are enabling explorations of the very large reservoir of rare coding variants in humans and are expected to accelerate the pace of discovery in human genetics[2].

Rare variants can be examined using association tests that group alleles in a gene or another functional unit[3]. Compared to tests of individual alleles, tests with this grouping can have increased power, especially when applied to large samples where several rare variants are observed in the same functional unit[4]. The simplest rare variant tests consider the number of potentially functional alleles in each individual[5], but these tests can be refined to weigh variants according to their likely functional impact[6], to allow for imputed or uncertain genotypes[7,8] or to allow variants that increase and decrease risk to reside in the same gene[9–11] (a feature that is important when the same gene harbors hypermorphic and hypomorphic alleles[12]). The optimal strategy for grouping and weighting rare variants—ranging from a focus on protein-truncating alleles to consideration of all nonsynonymous variants and encompassing strategies that examine all variants with a frequency of <5% as well as alternative strategies that examine only singletons—depends on the unknown genetic architecture of each trait and each locus[13].

Here we describe practical approaches for the meta-analysis of rare variants. Our approach starts with simple statistics that can be calculated in an individual study (single-site score statistics and their covariance matrix, which summarizes linkage disequilibrium information and relatedness among sampled individuals). We then show that, when these statistics are shared, a wide variety of gene-level association tests can be executed centrally—including both weighted and unweighted burden tests with a fixed[5] or variable[6] frequency threshold and the sequence kernel association test (SKAT) that accommodates alleles with opposite effects in a gene[9]. Our approach generates comparable results to sharing individual-level data (and, in fact, identical results when allowing for between-study heterogeneity in nuisance parameters, such as trait means, variances and covariate effects). To demonstrate our approach, we analyze blood lipid levels in >18,500 individuals genotyped with exome genotyping arrays. Our analysis of blood lipid levels provides examples of loci where the signal for gene-level association tests exceeds the signal for single-variant tests and shows that our approach can recover signals driven by very rare variants (with a frequency of <0.05%).

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [5]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway. [6]St. Olav Hospital, Trondheim University Hospital, Trondheim, Norway. [7]University of Ottawa Heart Institute, Ottawa, Ontario, Canada. [8]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [9]School of Public Health, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA. [10]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [11]Department of Cardiovascular Medicine, University of Oxford, Oxford, UK. [12]Division of Cardiology, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA. [13]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, USA. [14]Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA. [15]Department of Clinical Sciences, Lund University, Malmö, Sweden. [16]Department of Biostatistics, University of Washington School of Public Health, Seattle, Washington, USA. [17]Department of Medicine, Levanger Hospital, Nord-Trøndelag Health Trust, Levanger, Norway. [18]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [19]These authors contributed equally to this work. [20]These authors jointly directed this work. Correspondence should be addressed to D.J.L. (dajiang@umich.edu) or G.R.A. (goncalo@umich.edu).

Received 18 March; accepted 20 November; published online 15 December 2013; doi:10.1038/ng.2852

Given that very large sample sizes are required for successful rare variant association studies, we expect that our methods (and refined versions thereof) will be widely useful.

Our approach is based on the insight that analogs of most gene-level association tests can be constructed using single-variant test statistics and knowledge of their correlation structures. As shown in the Online Methods, simple[14] and weighted[10,15] burden tests, variable-threshold tests[6] and tests allowing for variants with opposite effects[9] can be constructed in this manner. We perform meta-analysis of single-variant statistics using the Cochran-Mantel-Haenszel method, calculate variance-covariance matrices for these statistics and construct gene-level association tests by combining the two. In the **Supplementary Note**, we show that the rare variant statistics generated in this way are identical to those obtained by sharing individual-level data and allowing for heterogeneity in nuisance parameters, with no loss of power. Notably, rare variant statistics calculated with our approach are less vulnerable to artifacts due to population stratification than statistics generated by naively pooling individual-level data. As in other meta-analysis settings, sharing summary statistics accelerates the overall analysis process, mitigates concerns about participant confidentiality and reduces the risk that data will be used for unapproved analyses (as always, to avoid violating the trust of research subjects, we strongly recommend that investigators sharing summary statistics agree that these will not be used to identify research subjects). To evaluate significance, we propose methods for calculating *P* values using asymptotics and also using Monte-Carlo methods that apply knowledge of linkage disequilibrium relationships to sample plausible combinations of single-variant statistics and then generate empirical distributions for gene-level statistics. Because evaluating asymptotic *P* values can be numerically unstable, Monte-Carlo methods can be used to verify interesting *P* values.

RESULTS

We first evaluated our method using simulations. Genes were simulated as stretches of 5,000 bp in length using coalescence[16] and a demographic model (including an ancient bottleneck, recent exponential growth, differentiation and migration) calibrated to mimic a sample of multiple European populations[17,18] (**Supplementary Fig. 1** and **Supplementary Note**). F_{ST}, which measures population differentiation, averaged 0.004 between simulated populations, as expected when a distribution of rare variants is geographically restricted[19]. The simulations produced samples of 1,000 individuals, each drawn from one of several related populations, typically including a few shared variants and many population-specific variants. Half of the simulated variants were randomly set to increase trait values by 0.125 s.d. (**Supplementary Fig. 2**; see **Supplementary Figs. 3** and **4** for similar results using alternative trait models).

We analyzed each simulated sample with a series of gene-level association tests. Results obtained for 10,000 simulated genes using our meta-analysis approach compared to a combined analysis of individual-level data across studies are shown in **Supplementary Figures 2–4**. In variable-threshold tests, we found that the *P* values were sometimes slightly different ($r^2 = 0.995$ between the two sets of log-transformed *P* values); in the other two tests, *P* values and test statistics were indistinguishable. Calculation of analytical *P* values for variable-threshold tests requires the evaluation of high-dimensional integrals that can be numerically unstable and is thus very sensitive to small differences in the variance-covariance matrix. In practice, it will often be a good idea to confirm significant *P* values using our Monte-Carlo approach.

To evaluate our Monte-Carlo approach, we compared its empirical *P* values to those obtained by permuting phenotypes between individuals within each study. We implemented adaptive versions of both algorithms[20], with more simulations carried out when the *P* value was small and fewer simulations carried out when the *P* value was large. Log-transformed *P* values for the two approaches were highly concordant ($r^2 = 0.996$). When small *P* values were estimated, increasing the number of simulations improved the precision of the estimated *P* values (**Supplementary Fig. 5**).

We next verified that type I error was well controlled (**Supplementary Table 1**). In all analyses, we first applied an inverse normal transformation to trait residuals (which helped ensure that our statistics could be modeled using a normal distribution, even for very rare variants, as in **Supplementary Fig. 6**). Reassured that type I error was well controlled, we next explored power to detect associated variants in several scenarios (**Fig. 1a–c** and **Supplementary Fig. 7a–c**). It is clear that, for the effect sizes simulated here, very large samples may be required. In some settings, power only reached ~60% in analyses of ~100,000 individuals. We did not find a method that was most powerful in all situations, emphasizing the value of implementing a diverse set of test statistics (see also Ladouceur *et al.*[13]). Because meta-analysis methods that combine *P* values are popular for common variants and can also be implemented for rare variants, we compared power for our method with that for analyses based on Fisher's method or on the minimal *P*-value approach for combining *P* values (**Fig. 1** and **Supplementary Fig. 7**). In all the simulation scenarios considered, our method outperformed these alternative methods, especially when information was combined across a large number of samples. In addition to power, our approach provides three useful features. First, it provides great flexibility in the choice of rare variant association test (definition of functional units, choice of variants to be grouped and frequency thresholds for analysis); approaches based on Fisher's method would likely require every contributing study to reanalyze their data when any of these parameters were changed. Second, because our approach provides, in addition to *P* values, estimates of effect size (in all cases) and allele frequency thresholds for candidate variants (in the variable-threshold test), our method provides rich information that helps in interpretation. Third, our approach allows the relationship between multiple association signals in a region to be dissected through conditional analysis, as detailed below.

We proceeded to a meta-analysis of blood lipid levels in 18,699 individuals of European ancestry genotyped with Illumina exome arrays and drawn from 7 studies: the Women's Health Initiative (WHI)[21], the Ottawa Heart Study[22], the Malmö Diet and Cancer Study–Cardiovascular Cohort (MDC)[23], the Precocious Coronary Artery Disease (PROCARDIS) Case Series, the PROCARDIS Control series[24] and the Nord-Trøndelag Health Study (HUNT) of myocardial infraction cases and matched controls[25] (see **Supplementary Tables 2** and **3** for summary statistics for each of these samples, including basic demographics, summaries of lipid levels, number of nonsynonymous and loss-of-function variants per individual and number of variant sites shared across different studies). Overall, 171,193 variants were polymorphic in at least one individual. Of these variants, 125,702—the vast majority—had a frequency of <1%.

To verify the soundness of our approach, we repeated our power and type I error simulations using real genotype data from the HUNT and MDC studies but simulated phenotypes. These additional experiments confirmed that our method produces well-calibrated statistics and is more robust to stratification than analyses that directly pool individual-level data and treat the complete data set as a single study without modeling heterogeneity between studies (**Supplementary Fig. 8**). In addition, the power for our method continued to exceed that for alternative methods that directly combined *P* values from individual studies (**Supplementary Fig. 9**).

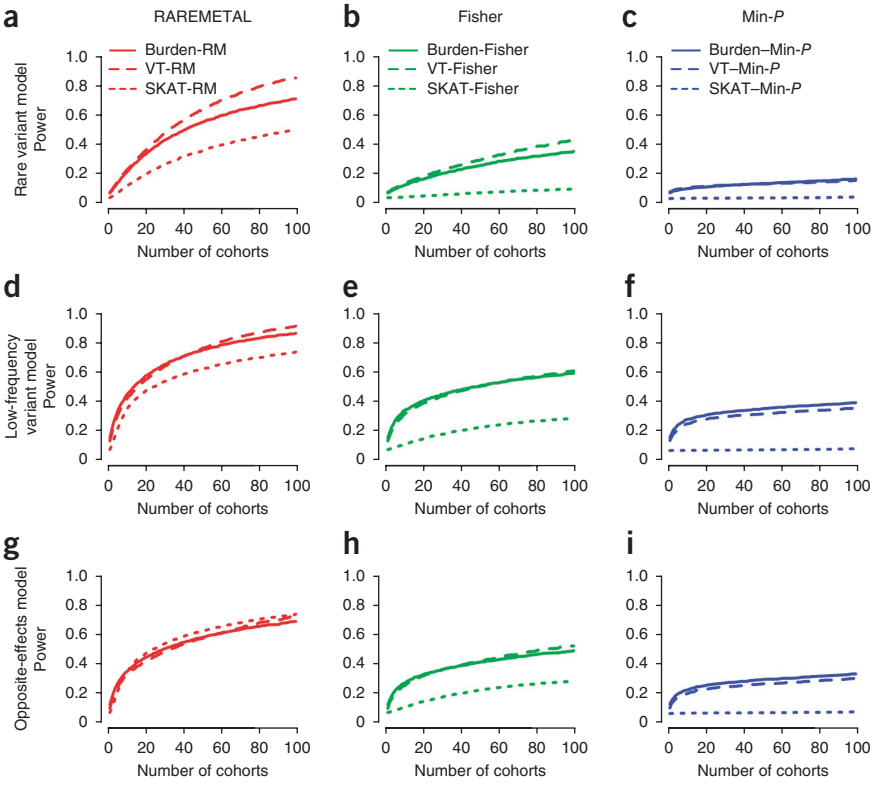

Figure 1 Power comparison for our approach, Fisher's method and the minimal *P*-value approach. Three phenotype models were simulated: (i) 50% of low-frequency variants with MAF < 0.5% are causal, each increasing expected trait values by 0.25 s.d.; (ii) 50% of all variants are causal, irrespective of frequency, and increase trait values by 0.25 s.d.; and (iii) 50% of variants are casual, irrespective of frequency, and 80% of these increase expected trait values by 0.25 s.d., while the remaining 20% decrease trait values by the same amount. A total of 2–100 samples of size 1,000 were simulated for each model, with each sample drawn from a randomly chosen population. Meta-analysis was performed using our approach: RAREMETAL (RM for short) or using Fisher's method or the minimal *P*-value (min-*P*) approach to combine burden test, variable-threshold (VT) and SKAT test statistics for variants with MAF < 5%. Power was evaluated at the significance threshold of $\alpha = 2.5 \times 10^{-6}$ using 10,000 replicates. (**a**–**c**) Power using model (i) for our approach (**a**), the Fisher approach (**b**) and the minimal *P*-value approach (**c**). (**d**–**f**) Power using model (ii) for our approach (**d**), the Fisher approach (**e**) and the minimal *P*-value approach (**f**). (**g**–**i**) Power using model (iii) for our approach (**g**), the Fisher approach (**h**) and the minimal *P*-value approach (**i**). Note that differences between our approach and these alternatives become more marked when more studies are used for meta-analysis.

We then performed meta-analysis of single-variant association test results. The resulting test statistics appeared well calibrated, with a genomic control value of <1.05 for all three traits, both for common and rare variants (**Supplementary Fig. 10**). At a significance threshold of $P < 3 \times 10^{-7}$ (corresponding to 0.05/171,193), we found significantly associated variants (with MAF < 5%) at *LPL*[26], *ANGPTL4* (ref. 26), *LIPG*[26], *CD300LG*[27], *LIPC*[26], *APOB*[26] and *HNF4A*[26] for high-density lipoprotein (HDL) levels, at *PCSK9* (ref. 26), *BCAM-CBLC-PVR* (neighboring *APOE*)[26] and *APOB*[26] for low-density lipoprotein (LDL) levels and at *ANGPTL4* (ref. 26), *LPL*[26] and *APOB*[26] for triglyceride levels (**Supplementary Table 4**). Except for the variants in *LIPC* and *APOB*, all significantly associated variants had a frequency of >1%, reflecting the limited power of single-variant association tests for rare alleles.

We next carried out gene-level tests. Again, test statistics appeared well calibrated, with a genomic control value of <1.05 (**Supplementary Fig. 11**). At a significance threshold of $P < 3.1 \times 10^{-6}$ (corresponding to 0.05/16,153 and allowing for the number of genes tested), we observed association at *LIPC*, *LPL*, *ANGPTL4*, *LIPG*, *HNF4A* and *CD300LG* for HDL levels, at the *PCSK9*, *APOE* locus (as well as at nearby genes *PVR*, *BCAM* and *CBLC*) and at *LDLR* for LDL levels and at *ANGPTL4* and *LPL* for triglyceride levels (**Table 1**). At these loci, much stronger signals were identified in the meta-analysis than in any component study (**Supplementary Table 5**). Reassuringly, these signals corresponded with the loci identified in previous genome-wide association studies and/or resequencing studies. Notably, our approach was able to appropriately identify the signal in *LDLR*, which is driven by several very rare variants (each with a frequency of <0.00052) that nearly always increase blood LDL-cholesterol levels. Furthermore, at several other loci, gene-level *P* values exceeded the best single-variant *P* value in the gene, illustrating the value of aggregating information across variants (**Supplementary Table 6**). We again compared our method with conventional methods such as a minimal *P*-value approach, Fisher's method and an extended Fisher's method taking into account unequal sample sizes (Online Methods). Our method identified a larger number of loci (**Supplementary Tables 7–9**), all known to be associated with lipid levels in humans. We also compared the results obtained from our meta-analysis method with the results from directly pooling a subset of the data (after normal transformation of trait values in each sample to avoid artifacts due to stratification). Reassuringly, *P* values from our approach and from joint analysis of pooled data were highly concordant, with $r^2 > 0.99$ (**Supplementary Fig. 12**), in accordance with the results obtained using coalescent simulations.

An added convenience of sharing single-variant statistics together with their covariance matrices, as we propose, is that this facilitates conditional analyses, extending an idea used by Yang *et al.*[28] for the analysis of common variants by genome-wide association study meta-analysis. We demonstrate in **Supplementary Figure 13** how, in simulations, common variants can generate shadow rare variant association signals at nearby genes and how our method for conditional analysis resolves this problem. Using real data, we reexamined two of the LDL-associated loci in detail, *LDLR* and *APOE-BCAM-CBLC-PVR*. For *LDLR*, we examined the relationship between rare variant signals and three nearby common variants[26]. Specifically, we conditioned on genotypes for three common variants (rs6511720, rs2228671 and rs72658855) exhibiting significant association in the region and found that the association of the *LDLR* rare variant remained significant ($P = 4.6 \times 10^{-7}$) (**Supplementary Table 10**). For the *APOE-BCAM-CBLC-PVR* locus, after conditioning on the common variant showing the strongest association in the region (rs7412), gene-level associations at *BCAM*, *CLBC* and *PVR* became non-significant, suggesting that these rare variant signals were the result of regional linkage disequilibrium with more common and well-described variants in *APOE* (**Supplementary Table 11**). We also analyzed top single-association

Table 1 Results for meta-analysis of gene-level rare variant association tests

Gene	Gene position[a]	Burden-1	Burden-5	SKAT-1	SKAT-5	VT	MAF cutoff	Direction of single-variant association statistics[b]	Estimates of genetic average effect (s.d.) for rare variants under different MAFs thresholds		
									0.01	0.05	VT
HDL											
LIPC[c]	Chr. 15: 58.7 Mb	**1.4×10^{-12}**	**3.5×10^{-7}**	**1.8×10^{-9}**	1.4×10^{-2}	**4.5×10^{-12}**	3.7×10^{-3}	– + + + + – – + –	0.5	0.1	0.5
LPL[c]	Chr. 8: 19.8 Mb	9.7×10^{-1}	**2.5×10^{-24}**	3.5×10^{-1}	**5.0×10^{-13}**	**1.5×10^{-23}**	2.5×10^{-2}	(–) – (–) + – + +	–	–0.3	–0.3
ANGPTL4[c]	Chr. 19: 8.4 Mb	2.2×10^{-2}	**2.9×10^{-19}**	2.2×10^{-2}	**3.0×10^{-19}**	**1.8×10^{-18}**	2.6×10^{-2}	(+) – – + + – + + +	–	0.3	0.3
LIPG[c]	Chr. 18: 47.1 Mb	2.2×10^{-5}	**6.4×10^{-19}**	2.1×10^{-5}	**2.9×10^{-9}**	**4.4×10^{-18}**	1.3×10^{-2}	– + + – – – – (+) +	–	0.4	0.4
HNF4A	Chr. 20: 43.0 Mb	7.5×10^{-1}	**2.8×10^{-7}**	6.8×10^{-1}	**2.5×10^{-7}**	**1.5×10^{-6}**	4.1×10^{-2}	(–) – – + – +	–	–0.1	–0.1
CD300LG	Chr. 17: 41.9 Mb	4.9×10^{-1}	**8.5×10^{-7}**	5.2×10^{-1}	1.0×10^{-5}	**3.1×10^{-6}**	3.3×10^{-2}	(–) + – (+)	–	–0.1	–
LDL											
PCSK9[c]	Chr. 1: 55.5 Mb	1.8×10^{-2}	**7.4×10^{-19}**	8.1×10^{-2}	**5.5×10^{-17}**	**2.0×10^{-28}**	1.3×10^{-2}	(–) – – (–) – – + –+ + –	–	–0.3	–0.5
BCAM	Chr. 19: 45.3 Mb	1.7×10^{-1}	**1.6×10^{-18}**	1.5×10^{-1}	3.0×10^{-5}	**2.6×10^{-17}**	3.6×10^{-2}	(–) + + + (–) + – + + + – – – + (–) + – – + – – + +	–	–0.1	–0.1
CBLC	Chr. 19: 45.3 Mb	9.4×10^{-1}	**2.0×10^{-15}**	4.4×10^{-1}	1.5×10^{-4}	**1.0×10^{-14}**	4.4×10^{-2}	– (–) – – + – (–) (+)	–	–0.1	–0.1
PVR	Chr. 19: 45.2 Mb	6.1×10^{-2}	**3.0×10^{-10}**	4.8×10^{-2}	6.3×10^{-2}	**1.1×10^{-9}**	4.9×10^{-2}	(–) + + – – +	–	–0.1	–0.1
LDLR[c]	Chr. 19: 11.2 Mb	1.8×10^{-3}	4.7×10^{-5}	3.8×10^{-2}	2.5×10^{-1}	**2.4×10^{-7}**	5.2×10^{-4}	+ + + + + + + + + –+ + + + – – +	–	–	0.8
Triglycerides											
ANGPTL4[c]	Chr. 19: 8.4 Mb	2.6×10^{-2}	**1.2×10^{-24}**	3.7×10^{-2}	**3.9×10^{-25}**	**7.1×10^{-24}**	2.6×10^{-2}	(–) + – – – + – – –	–	–0.3	–0.2
LPL[c]	Chr. 8: 19.8 Mb	6.8×10^{-1}	**7.7×10^{-20}**	2.6×10^{-1}	**1.8×10^{-11}**	**4.6×10^{-19}**	2.5×10^{-2}	(+) + (+) – – + –	–	0.2	0.2

Associations that attain exome-wide significance ($P < 3.1 \times 10^{-6}$) are shown. Five gene-level association tests were used to analyze the data: simple burden tests with a frequency cutoff of 1% or 5% (Burden-1 and Burden-5), SKAT tests with a frequency cutoff of 1% or 5% (SKAT-1 and SKAT-5) and variable-threshold (VT) tests that analyze variants with MAF < 5%. Significant *P* values for each test are shown in bold. For the associations that are significant, estimates of average genetic effect are also shown.

[a]Gene position is defined according to hg19, GRCh37 (Genome Reference Consortium Human Reference 37. [b]Direction of single-site statistics for variants with MAF< 5%. Variants in parentheses have frequency >1%. [c]Loci with one or more gene-level association signal exceeding the top single-variant signal.

signal conditional on the genotypes of rare variants (with MAF ≤ 5%) that were included in the burden tests. We showed that the top single-variant signals for both the *APOE* gene and the *LDLR* gene remained significant (**Supplementary Table 12**). For completeness, we showed that conditional analyses using individual-level data in a subset of samples and conditional analyses using our meta-analysis–based approach gave highly concordant *P* values ($r^2 > 0.99$) (**Supplementary Figs. 14** and **15**).

DISCUSSION

In the analysis of each sample, when population stratification is of concern, we recommend that principal components of the genotype matrix be incorporated in the regression model as covariates[29] or that linear mixed models with empirically estimated kinship matrices be used[30]. Linear mixed models can also be used to account for relatedness in family studies or other samples that include cryptically related individuals. Our software implementation readily allows for both these options, including the correct calculation of kinship matrices to allow family samples to be included in meta-analyses (Online Methods).

Although we only presented applications of our method to quantitative trait meta-analysis, our methods and tools can be applied to binary traits as well (Online Methods). For binary traits, assumptions about normality of test statistics may be less reliable. Deviations from normality could affect the performance of our resampling method for empirical *P* values, meta-analysis results for the rarest variants and conditional analysis statistics (see also the work of Lin and Tang[9] and of Lee *et al.*[31]). Because the performance of our method (and other similar approaches) for binary traits will depend on factors such as sample size and the balance of cases and controls in each sample, we recommend careful quality control of the results for such studies, including, for example, review of quantile-quantile plots for variants of different frequency. Our method is implemented as freely available software, including programs for calculating summary statistics, annotating the resulting summary statistics, performing meta-analysis, calculating gene-level statistics and executing conditional analyses. Our tools work with standard VCF files[32] for genotype data and with Merlin[33] or PLINK[34] files for phenotype data.

Meta-analysis has facilitated many discoveries in common variant association studies. Here we describe a powerful framework for the meta-analysis of rare variants at the level of genes or other functional units. Using simulation and empirical evaluation, we demonstrate that our approach is well calibrated and provides comparable power to more cumbersome analyses that require that all individual-level data be pooled. Through the analysis of blood lipid levels across seven studies, we show that our approach can detect rare variant association signals at known candidate loci. Our method has a variety of unique features, which include support for a variety of rare variant association tests, allowing for the analysis of family samples and the calculation of empirical *P* values and for conditional analysis that can distinguish truly novel rare variant signals from shadows of other nearby common or rare associations. We envision that this approach (and continued development of related approaches[35–37]) will facilitate the large sample sizes required to accelerate discoveries in complex trait genetics.

URLs. The authors' website for Rare Variant Analysis and Meta-Analysis is available at http://genome.sph.umich.edu/wiki/RAREMETAL-SOFTWARE.

METHODS

Methods and any associated 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

The authors would like to thank M. Boehnke, X. Wen and S. Zoellner for helpful discussions. This work was supported by research grants R01HG007022 from the National Human Genome Research Institute, R01EY022005 from the National Eye Institute and R01HL117626 from the National Heart, Lung, and Blood Institute. G.M.P. was supported by award T32HL007208 from the National Heart, Lung, and

Blood Institute. S.K. is supported by a Research Scholar award from Massachusetts General Hospital (MGH), the Howard Goodman Fellowship from MGH, the Donovan Family Foundation and grant R01HL107816 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the US National Institutes of Health. The WHI program is funded by the National Heart, Lung, and Blood Institute, US National Institutes of Health, US Department of Health and Human Services through contracts N01WH22110, N01WH24152, N01WH32100-2, N01WH32105-6, N01WH32108-9, N01WH32111-13, N01WH32115, N01WH32118-32119, N01WH32122, N01WH42107-26, N01WH42129-32 and N01WH44221. This manuscript was prepared in collaboration with investigators from the WHI and has been approved by the WHI. WHI investigators are listed at https://cleo.whi.org/researchers/SitePages/WHI%20Investigators.aspx. The full list of PROCARDIS acknowledgments is available at http://www.procardis.org/. The Ottawa Heart Genomics Study was supported by Canadian Institutes of Health Research (CIHR) grants MOP-82810, MOP-77682 and MOP-2380941 and Canada Foundation for Innovation (CFI) grant 11966. The studies for the Malmö Diet and Cancer cohort were supported by grants from the Swedish Research Council, the Swedish Heart and Lung Foundation, the Påhlsson Foundation, the Novo Nordic Foundation and European Research Council starting grant StG-282255.

AUTHOR CONTRIBUTIONS

D.J.L., S.K. and G.R.A. conceived and designed the study. D.J.L., G.M.P. and X.Z. carried out primary data analysis. D.J.L., X.Z. and S.F. wrote the software package implementing the proposed methodologies. O.L.H., M.N., P.L.A., A.G., H.Z., U.P., M.F., M.O.-M., C.K., R.M., H.W., C.J.W., K.H. and O.M. contributed phenotypes, exome array genotypes and analyses for the study. M.Z. conducted population genetics simulation analysis. D.J.L. and G.R.A. wrote the first version of the manuscript. All authors critically reviewed and approved the manuscript. S.K. and G.R.A. jointly supervised the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
2. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623–630 (2012).
3. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
4. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. USA* **106**, 3871–3876 (2009).
5. Morris, A.P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
6. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
7. Liu, D.J. & Leal, S.M. Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am. J. Hum. Genet.* **87**, 790–801 (2010).
8. Zawistowski, M. *et al.* Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.* **87**, 604–617 (2010).
9. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
10. Lin, D.Y. & Tang, Z.Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–367 (2011).
11. Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309–326 (1997).
12. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
13. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M. & Richards, J.B. The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genet.* **8**, e1002496 (2012).
14. Morris, A.P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
15. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
16. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
17. Adams, A.M. & Hudson, R.R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699–1712 (2004).
18. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
19. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
20. Besag, J. & Clifford, P. Sequential Monte Carlo *p*-values. *Biometrika* **78**, 301–304 (1991).
21. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
22. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
23. Kathiresan, S. *et al.* Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.* **358**, 1240–1249 (2008).
24. Clarke, R. *et al.* Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* **361**, 2518–2528 (2009).
25. Krokstad, S. *et al.* Cohort Profile: the HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).
26. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
27. Albrechtsen, A. *et al.* Exome sequencing–driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298–310 (2013).
28. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
29. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
30. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
31. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
32. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
33. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
35. Hu, Y.J. *et al.* Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am. J. Hum. Genet.* **93**, 236–248 (2013).
36. Lee, S., Teslovich, T.M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
37. Tang, Z.Z. & Lin, D.Y. MASS: meta-analysis of score statistics for sequencing studies. *Bioinformatics* **29**, 1803–1805 (2013).

ONLINE METHODS

This section starts with a summary of notation and proceeds to describe the statistics to be shared between studies and methods for single-variant meta-analysis. We then show that the statistics for different gene-level tests can be calculated using summary-level data, enabling efficient meta-analysis. In the **Supplementary Note**, we provide additional details and summarize how each of the test statistics used here can be derived as a score test using likelihood functions that allow for per-sample nuisance parameters.

Notation. For simplicity, we describe our strategy for the analysis of a single gene. We let J be the number of variant nucleotide sites genotyped in at least one study. For study k, we let n_k denote the number of samples phenotyped and genotyped, and we let the vector $\mathbf{y_k} = (Y_{1,k},\cdots,Y_{N_k,k})^T$ denote the quantitative trait residuals (after adjustment for any covariates), with variance of σ_k^2. Within each study k, we encoded genotype information in matrix $\boldsymbol{X_k}$, where each entry $X_{i,j,k}$ represented the genotype for individual i at site j, coded as the number of alternative alleles. We encoded missing genotypes in the data set as the average number of minor alleles in individuals who were genotyped for that marker. The multisite genotype for individual i was denoted by the row vector $\mathbf{x}_{i,\bullet,k}$, and the genotypes for all N_k individuals at site j were given by the column vector $\mathbf{x}_{\bullet,j,k}$. For ease of presentation, we define the mean genotype matrix $\bar{\boldsymbol{X}}_{\boldsymbol{k}}$, where the (i,j)th element was $(\Sigma_i X_{i,j,k})/N_k$.

Summary statistics to be shared. For each study, we first calculated and shared a vector of score statistics $\mathbf{u_k} = (\boldsymbol{X_k} - \bar{\boldsymbol{X}}_{\boldsymbol{k}})^T \mathbf{y_k}$, a corresponding variance-covariance matrix $\boldsymbol{V_k} = \hat{\sigma}_k^2 N_k \operatorname{cov}(\boldsymbol{X_k}) = \hat{\sigma}_k^2 (\boldsymbol{X_k} - \bar{\boldsymbol{X}}_{\boldsymbol{k}})^T (\boldsymbol{X_k} - \bar{\boldsymbol{X}}_{\boldsymbol{k}})$ and allele frequencies for each marker $p_{j,k} = \Sigma_i X_{i,j,k}/2N_k$. Note that $\boldsymbol{V_k}$ effectively describes linkage disequilibrium relationships between the variants being examined. To perform quality control, we also shared mean and variance values for the quantitative trait residuals, genotype call rates and Hardy-Weinberg equilibrium P values at each variant site.

Meta-analysis of single-variant association test statistics. We first combined single-variant association test statistics across studies using the Cochran-Mantel-Haenszel method. Specifically, we calculated a score statistic at each site as

$$t_{j,\bullet} = U_{j,\bullet} \Big/ \sqrt{V_{j,j,\bullet}} \tag{1}$$

where $U_{j,\bullet} = \Sigma_k U_{j,k}$ and $V_{j,j,\bullet} = \Sigma_k V_{j,j,k}$. For ease of presentation, we denote the vector of single-variant association tests after meta-analysis as $\mathbf{u} = \Sigma_k \mathbf{u_k}$. Under the null hypothesis of no gene-phenotype associations, this vector was distributed as multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_k V_k$.

Burden tests that assume variants have similar effect sizes. For a simple burden test in study k, the impact of multiple rare variants in a region can be modeled using a shared regression coefficient in a model that takes the form

$$Y_{i,k} = \beta_{0,k} + \beta_{\text{burden}} C_{\text{burden}}(\mathbf{x_{i,\bullet,k}}) + \varepsilon_{i,k} \tag{2}$$

where $\varepsilon_{i,k} \sim \mathrm{N}(0,\sigma_k^2)$. $C_{\text{burden}}(\mathbf{x_{i,\bullet,k}})$ is a function that takes genotypes for a single individual as input and returns the count of rare alleles (the 'rare variant burden') in the gene being examined. When individual-level data are available and nuisance parameters $\beta_{0,k}$ and σ_k^2 are allowed to vary between studies, the score statistic for a rare variant burden test becomes

$$U_{\text{burden}} = \sum_k U_{\text{burden},k} = \sum_k \boldsymbol{\omega}^{\mathbf{T}} \mathbf{u_k} = \boldsymbol{\omega}^{\mathbf{T}} \mathbf{u} \tag{3}$$

which is equal to a linear sum of (weighted) single-variant score statistics.

Under the null, this statistic is approximately normally distributed with mean of 0 and variance of $V_{\text{burden}} = \boldsymbol{\omega}^T (\Sigma_k \boldsymbol{V_k}) \boldsymbol{\omega}$, enabling significance tests. Here $\boldsymbol{\omega}$ is the vector of weights, which is $\boldsymbol{\omega} = (\omega_1,\cdots,\omega_J)$, with each element ω_j representing the weight assigned to variant j according to its allele frequency or its computationally predicted functional impact[10,15]. The formula above makes it clear that, when nuisance parameters are allowed to vary between studies, the same burden score statistics that could be calculated by sharing individual-level data can be equivalently calculated using shared summary statistics.

Variable-threshold tests with an adaptive frequency threshold. In the variable-threshold test, rare variant burden statistics were calculated for each observed variant MAF threshold, and significance was evaluated for the maximum of these statistics. Given a specific variant frequency threshold F, we defined the resulting burden score statistic as

$$U_{\text{burden}(F)} = \mathbf{v_F^T} \vec{U} \tag{4}$$

Here $\mathbf{V_F}$ was a vector of indicators, where the jth element was equal to 1 if the pooled MAF at variant site j was less than F and 0 otherwise. For convenience, we also defined a matrix of indicators for MAF thresholds $\boldsymbol{\Phi} = (\mathbf{v_{F_1}}, \mathbf{v_{F_2}},\cdots,\mathbf{v_{F_J}})$. After a burden statistic was calculated for each potential frequency threshold, these were standardized, dividing each statistic by its corresponding variance, and the maximum statistic was identified as

$$T_{\text{VT}} = \max_F \left\{ T_{\text{burden}(F)} \right\} \tag{5}$$

where

$$T_{\text{burden}(F)} = U_{\text{burden}(F)} \Big/ \sqrt{\mathbf{v}_F^T \sum_k \boldsymbol{V_k} \mathbf{v}_F}$$

Significance for this statistic can be evaluated using the cumulative distribution function for the multivariate normal distribution[38]. Specifically, given the definition of the covariance between burden statistics calculated using different allele frequency thresholds, we generated

$$\left(T_{\text{burden}(F_1)},\cdots,T_{\text{burden}(F_M)}\right) \sim \text{MVN}\left(\mathbf{0}, \boldsymbol{\Phi}\left(\sum_k \boldsymbol{V_k}\right)\boldsymbol{\Phi}^{\boldsymbol{T}}\right) \tag{6}$$

The P value for the variable-threshold test statistic was given by

$$\begin{aligned} p &= 1 - \Pr\left(T_{\text{VT}} \le t_{\text{VT}}\right) \\ &= 1 - \Pr\left(T_{\text{burden}(F_1)} \le t,\cdots,T_{\text{burden}(F_M)} \le t\right) \\ &= 1 - F_{\text{MVN}}(t,\cdots,t) \end{aligned} \tag{7}$$

where F_{MVN} was the distribution function for the multivariate normal distribution $\text{MVN}(\mathbf{0}, \boldsymbol{\Phi}(\Sigma_k \boldsymbol{V_k})\boldsymbol{\Phi}^{\boldsymbol{T}})$.

Burden tests that assume a distribution of variant effect sizes (for example, SKAT tests). The simple burden test and variable-threshold test described above can be underpowered when variants with opposite phenotypic effects reside in the same gene and are grouped together, as the shared regression coefficient can average close to zero in that situation[9–12]. To accommodate this situation, we considered an underlying distribution of rare variance effect sizes with a mean of zero and tested whether the variance of this distribution τ was greater than zero.

When individual-level data were available, association analysis in study k was performed using the following model

$$Y_{i,k} = \beta_{0,k} + \sum_j \beta_j X_{i,j,k} + \varepsilon_{i,k} \tag{8}$$

where $\varepsilon_{i,k} \sim \mathrm{N}(0,\sigma_k^2)$. We made inferences about rare variant effect sizes $\boldsymbol{\beta} = (\beta_1,\beta_2,\cdots,\beta_J)$ by assuming these followed a common distribution with mean of zero and variance of τ. Under the null, τ was zero. Following the example of Wu *et al.*[9], in the **Supplementary Note** we derive the score statistic for this model and show that it can be calculated on the basis of per-study summary statistics as

$$Q = \left(\sum_k \mathbf{u_k}\right)^T \boldsymbol{K} \left(\sum_k \mathbf{u_k}\right) \tag{9}$$

Here, $\boldsymbol{K}$ is the kernel matrix that compares multisite genotypes. A default choice[9] is a diagonal matrix $\boldsymbol{K} = \operatorname{diag}(\omega_1,\omega_2,\cdots,\omega_J)$, with ω_j being the weight assigned to variant site j. The statistic Q follows a mixture χ^2 distribution[31], which means that Q is equivalent in distribution to a weighted sum of

independent χ^2 random variables. The weights (or mixture proportions) are given by the eigenvalues for the matrix $(\Sigma_k \boldsymbol{V_k})^{1/2} \boldsymbol{K} (\Sigma_k \boldsymbol{V_k})^{1/2}$.

Monte-Carlo method for empirical assessment of significance. The previous sections describe how a series of gene-level test statistics can be calculated and, for each one, propose a strategy for evaluating significance using asymptotic distributions. In practice, evaluating the required numerical integrals can be challenging because variance-covariance matrices are sometimes singular or nearly singular.

Note that single-variant test statistics are distributed as

$$\sum_k \mathbf{u_k} = \sum_k \mathbf{y_k^T} \left(\boldsymbol{X_k} - \bar{\boldsymbol{X}}_{\boldsymbol{k}} \right) \sim \mathrm{MVN}\left(\mathbf{0}, \sum_k \boldsymbol{V_k} \right) \quad (10)$$

Then, to evaluate significance empirically, one can sample random vectors from the distribution $\mathrm{MVN}(\mathbf{0}, \Sigma_k \boldsymbol{V_k})$ and calculate gene-level rare variant test statistics for each of these sampled random vectors, resulting in an empirical distribution for any gene-level statistic[39]. As usual, *P* values can then be evaluated by comparing the test statistics for the original data with those in the empirical distribution. For computational efficiency, we used an adaptive algorithm where a larger number of vectors were sampled when assessing small *P* values and fewer vectors were sampled when assessing larger *P* values[20].

Conditional analyses. It is well known that, owing to linkage disequilibrium, one or more common causal variants can result in shadow association signals at other nearby common variants. For common variants, Yang *et al.*[28] have shown that linkage disequilibrium relationships between variants, estimated from external reference panels, can be used to enable conditional analysis in meta-analysis settings. For rare variants and gene-level tests, accurately describing relationships between variants is crucial, and we advise against the use of external reference panels. Instead, in the **Supplementary Note**, we describe how conditional analysis statistics can be derived for different gene-level tests in our meta-analysis setting.

Analysis of samples of known or hidden relatedness. Our methods and tools can also be used when samples in a study are related to each other. Detailed formulae for the score statistics and their covariance matrices when linear mixed models are used to account for relatedness are described in the **Supplementary Note**.

Analysis of dichotomous traits. Our approach extends naturally to the analysis of binary traits. Specifically, when single-variant score statistics and their covariance matrices are shared, meta-analysis test statistics can be calculated in the same manner as for continuous traits. Detailed definitions of test statistics for binary traits are given in the **Supplementary Note**. A limitation is that, when variant counts in a gene or analysis unit are very small or the number of cases and controls in each study is very unbalanced, the asymptotic distributions for burden statistics may not hold, and *P* values obtained using our approach may not be accurate. In practice, we recommend careful review of quantile-quantile plots for meta-analysis statistics (as is standard in genome-wide association studies).

Weighted Fisher's methods, incorporating unequal sample sizes. To accommodate the scenario where meta-analysis is performed on samples of different size, we used a modified version of Fisher's method that incorporates sample sizes as weights for each study. Specifically, our test statistic was defined by $T_{\text{weighted-Fisher}} = -2\Sigma_k N_k \log p_k$. The weighted Fisher's test statistic follows a mixture χ^2 distribution with mixture proportions given by N_1, N_1, N_2, N_2, …, N_k, N_k.

Simulation of population genetic data. We simulated haplotypes using a coalescent model and the program *ms*[16]. We chose a demographic model consistent with European demographic history[4], including an ancestral bottleneck followed by more recent population differentiation and exponential growth. Model parameters were based on estimates from large-scale sequencing studies[40], as detailed in the **Supplementary Note**.

Meta-analysis of lipid traits. Summary statistics were calculated for each participating study and shared to enable a central meta-analysis. In single-variant and gene-based rare variant association analysis, age, age^2, sex and cohort-specific covariates, such as principal components of ancestry, were included in the analysis. Trait residuals were standardized using inverse normal transformation. More detailed descriptions for each participating cohort are given in the **Supplementary Note**. This research was approved by the institutional review boards of the University of Michigan and the Broad Institute. Informed consent was obtained from all study subjects. In addition, all participating studies received approvals from their local ethics committees.

38. Genz, A. Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* **1**, 141–149 (1992).
39. Zou, F., Fine, J.P., Hu, J. & Lin, D.Y. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* **168**, 2307–2316 (2004).
40. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1**, 131 (2010).