nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0487-7

Ultrarare variants drive substantial *cis* heritability of human gene expression

Ryan D. Hernandez[1,2,3,4,5,6*], Lawrence H. Uricchio[7], Kevin Hartman[8], Chun Ye[2,9], Andrew Dahl[2,3] and Noah Zaitlen[2,3,10*]

The vast majority of human mutations have minor allele frequencies under 1%, with the plurality observed only once (that is, 'singletons'). While Mendelian diseases are predominantly caused by rare alleles, their cumulative contribution to complex phenotypes is largely unknown. We develop and rigorously validate an approach to jointly estimate the contribution of all alleles, including singletons, to phenotypic variation. We apply our approach to transcriptional regulation, an intermediate between genetic variation and complex disease. Using whole-genome DNA and lymphoblastoid cell line RNA sequencing data from 360 European individuals, we conservatively estimate that singletons contribute approximately 25% of *cis* heritability across genes (dwarfing the contributions of other frequencies). The majority (approximately 76%) of singleton heritability derives from ultrarare variants absent from thousands of additional samples. We develop an inference procedure to demonstrate that our results are consistent with pervasive purifying selection shaping the regulatory architecture of most human genes.

The recent explosive growth of human populations has produced an abundance of genetic variants with minor allele frequencies (MAFs) less than 1% (ref. [1]). While many rare variants underlying Mendelian diseases have been found[2], their role in complex disease is unknown[3–8]. Evolutionary models predict that the contribution of rare variants to complex disease is highly dependent on selection strength[9,10] and that population growth can magnify their impact[10–12]. Recent methodological breakthroughs[13,14] have enabled researchers to jointly estimate the independent contributions of low- and high-frequency alleles to complex traits, often demonstrating a large rare variant contribution probably driven by natural selection[5,15–18]. However, these studies excluded the rarest variants[15] or included only well-imputed variants[5]. This is a problematic limitation given that some plausible evolutionary models predict that the largest contributions to phenotypic variance could be from the rarest variants[9–11,19]. Directly querying the role of all variants with large-scale sequencing and sensitive statistical tests has the potential to reveal important sources of missing heritability, inform strategies to increase the success rate of association studies and clarify how natural selection has shaped human phenotypes.

In this study, we develop, validate and apply an approach for inferring the relative phenotypic contributions of all variants, from singletons to high-frequency variants. We focus on the narrow-sense heritability (h^2) of gene expression because a growing body of literature suggests that genetic variants primarily affect disease by modifying gene regulatory programs[20–23], and recent examinations have identified significant rare variant effects on transcription[8]. To characterize the genetic architecture of gene expression, we analyzed 360 unrelated individuals of European ancestry with paired whole-genome DNA[24] and RNA[25] sequencing (RNA-seq) of lymphoblastoid cell lines (LCLs). We evaluate the robustness of our approach to genotyping errors, read mapping errors, population structure, rare variant stratification and a wide range of possible genetic architectures.

Results

Building and testing our model. We developed a method to estimate the effect of rare alleles on trait variance and validated our approach with an extensive set of simulations. Before analyzing real expression data, we performed a rigorous series of simulations to identify an approach for estimating heritability that is robust to possible confounding factors. In our simulations, we used real genotype data (all variants within 1 megabase (Mb) of the transcription start or end sites of genes) and generated gene expression phenotypes across individuals while varying the number of causal variants contributing to the phenotype (from 1 to 1,000), the distribution of effect sizes (including uniform, frequency-dependent and an evolutionary-based model) and the distribution of causal allele frequencies (ranging from predominantly rare to predominantly common; see Supplementary Note). In total, we simulated 440 different genotype–phenotype models that span the range of genetic architectures that are likely to underlie complex phenotypes such as gene expression, and analyzed each simulated dataset using multiple distinct methods. These include fitting a linear mixed model via restricted maximum likelihood[26,27] and Haseman–Elston regression, an alternative approach based on regressing phenotypic covariance on genotypic covariance[26], which is more robust in small samples (see Supplementary Note).

Similar to previous work[28], we found that for many simulation settings, jointly analyzing all variants together can result in a substantial over- or underestimate of heritability (Fig. 1a; it shows results when true $h^2 = 0.2$). One common solution is to partition sites by frequency[5,15,29]. We found that simply isolating rare (MAF ≤ 1%) from common variants using two partitions and performing joint

[1]Bioengineering & Therapeutic Sciences, UCSF, San Francisco, CA, USA. [2]Institute for Human Genetics, UCSF, San Francisco, CA, USA. [3]Institute for Quantitative Biosciences, UCSF, San Francisco, CA, USA. [4]Institute for Computational Health Sciences, UCSF, San Francisco, CA, USA. [5]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [6]McGill University and the Genome Quebec Innovation Center, Montreal, Quebec, Canada. [7]Department of Biology, Stanford University, Stanford, CA, USA. [8]Biological and Medical Informatics Graduate Program, UCSF, San Francisco, CA, USA. [9]Epidemiology & Biostatistics, UCSF, San Francisco, CA, USA. [10]Department of Medicine Lung Biology Center, UCSF, San Francisco, CA, USA. *e-mail: Ryan.Hernandez@me.com; Noah.Zaitlen@ucsf.edu

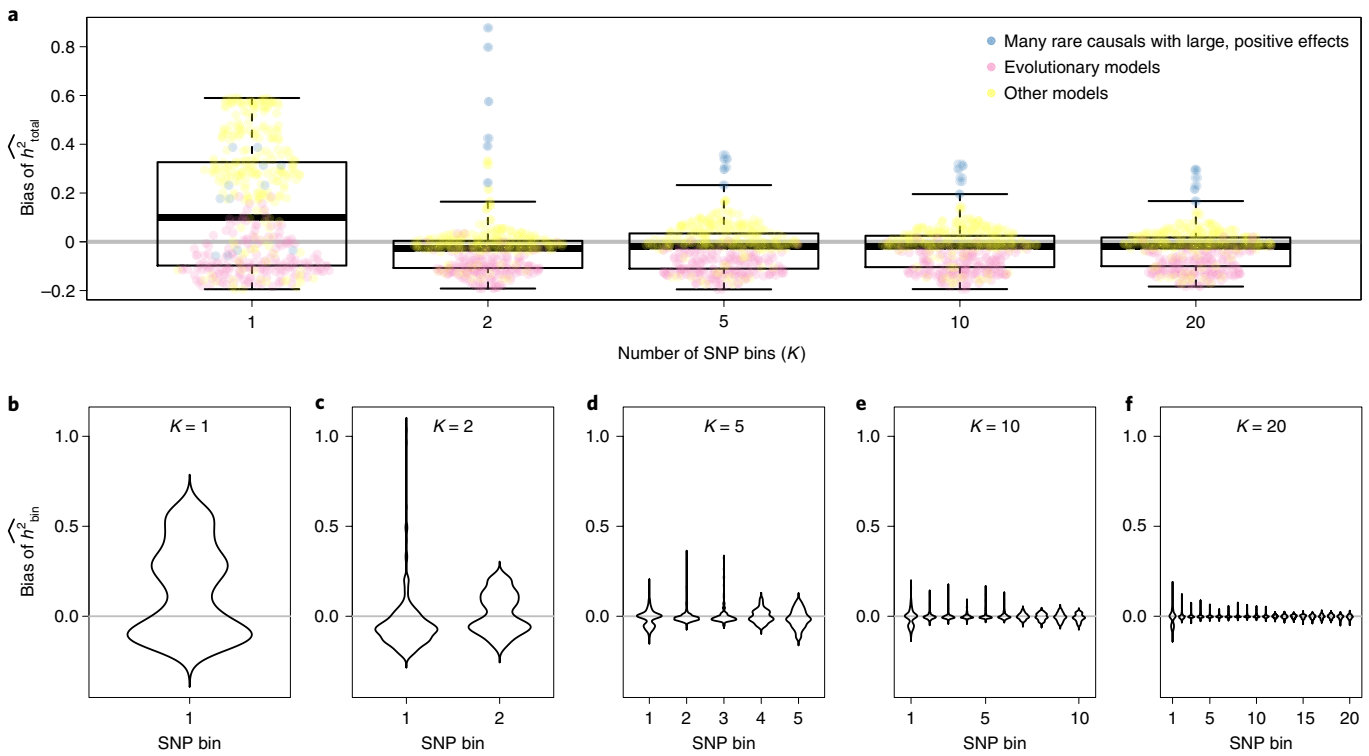

Fig. 1 | Simulation results. Across a broad range of parameters, the accuracy of heritability inference improves as the number of SNP bins (partitioned by MAF) increases. **a**, Mean bias of total heritability (inferred-true) for different numbers of SNP bins (*K*), where each point represents the mean of 500 simulations for different parameters, and box plot summarizing the bias distribution across all parameters (indicating the median, upper and lower quartile and twice the interquartile range). **b**–**f**, Distribution of average bias across simulated parameters for each SNP bin, showing that both mean and variance of the bias decrease as *K* increases ($n = 500$ simulations in each plot).

inference[15] could improve the accuracy for most models. However, when there are many causal rare variants, the estimator remains upwardly biased. Partitioning alleles into five or more categories using MAF[5] alleviates this problem. Notably, not only does the overall bias decrease as the number of MAF categories increases, but the bias for each MAF bin also decreases substantially across all models (see Fig. 1b–f and Supplementary Note). These simulations suggested that with our sample size, partitioning SNPs into 20 MAF bins resulted in the smallest bias in our estimate of total heritability (h^2_{total}) and the smallest bias for each bin across all simulated parameters. (However, see the Supplementary Note for further discussion of models that can induce bias.) Notably, further partitioning can improve results even further (see Supplementary Note); however, variance will probably increase unless prior knowledge about causal variation exists.

When partitioning variants into multiple MAF bins, singletons are inevitably isolated into their own category. Intuitively, if some fraction of singletons are causal, then individuals with a higher singleton load will probably be phenotypic outliers. (Indeed, individuals with outlier expression patterns have been observed to have an enrichment of nearby rare variants[8].) Therefore, it is reasonable to ask what contribution singletons make to patterning phenotypic variation across a population. We investigated the theoretical properties of heritability estimation from singleton variants and show analytically that when genotypic covariance is estimated using singletons alone, Haseman–Elston regression is equivalent to regressing squared standardized phenotypes against singleton counts (see Supplementary Note).

A direct implication of our derivation is that Haseman–Elston regression is unbiased unless singletons have large nonzero mean effect sizes (violating an explicit assumption of standard linear mixed models), which are the only simulation scenarios where heritability estimates remain upwardly biased (Fig. 1a, blue points). We developed an alternative approach that produces unbiased estimates of both heritability and mean effect size in all examined cases. Intuitively, the Singleton Heritability inference with REML (SingHer) method conditions on total singleton count (per *cis* window) to (1) appropriately estimate total *cis* heritability and (2) partition singleton heritability into directional and random components (see Supplementary Note). However, because Haseman–Elston regression is well understood and flexible, we recommend its use when mean effect sizes are near zero. For the data we analyze in this article, the SingHer method estimated that mean effect sizes were near zero; therefore, we proceeded with Haseman–Elston regression.

Singletons drive the genetic architecture of gene expression. To characterize the genetic architecture of human gene regulation, we partitioned the heritability of gene expression into 20 MAF bins. We used $n = 360$ unrelated individuals of European descent with both RNA-seq data from the GEUVADIS[25] project (Genetic European Variation in Disease—European Medical Sequencing Consortium) and whole-genome sequencing (WGS) data from the 1000 Genomes Project (1KGP)[24]. After extensive quality control to remove genes not expressed in LCLs, our dataset included 10,203 autosomal genes (see Supplementary Note). For each gene, we extracted all variants within 1 Mb of the transcription start or end sites (corresponding to an average of 13,839 variants per gene; 35.2% are singletons); we did not consider *trans* effects because of the small sample size (though we do analyze the effects of varying the window size in the Supplementary Note).

To control for possible non-normality, population structure and batch effects, we quantile-normalized expression values and

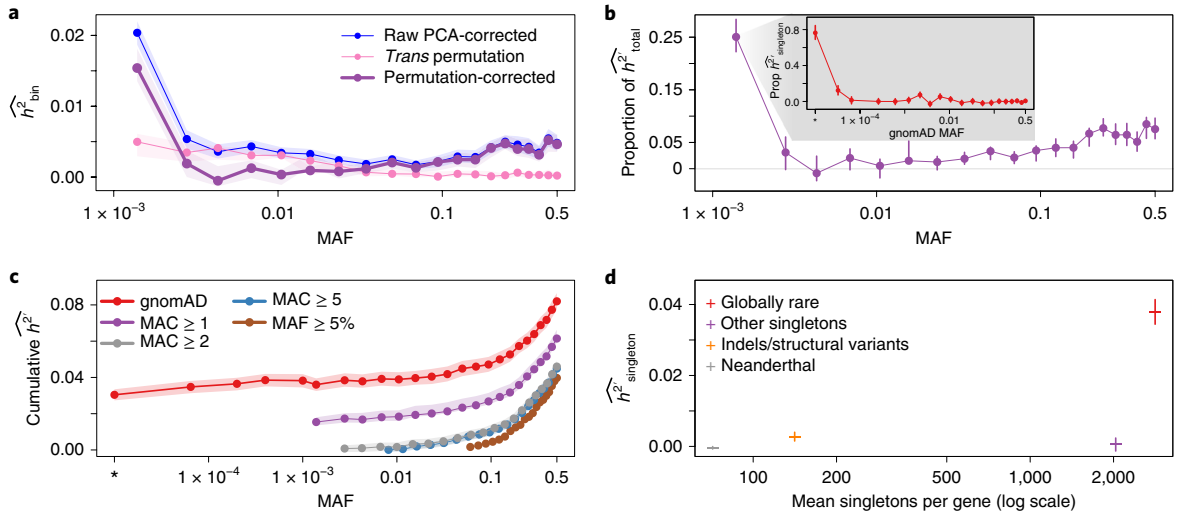

Fig. 2 | Partitioning heritability. Rare variants dominate the genetic architecture of human gene expression. **a**, Average heritability estimates across genes, partitioned across MAF bins ($h^{2'}$, purple) after correcting for population structure using PCA (blue) and eliminating residual rare variant structure identified using a *trans* permutation (pink). **b**, The proportion of heritability attributed to each MAF bin. Singletons represent approximately 25% of the total inferred heritability, the vast majority of which is due to variants that are extremely rare in the population (inset, partitioning singletons in our data by the MAF observed in gnomAD, $n > 15$ k; singletons not reported in gnomAD are indicated with an asterisk). **c**, Cumulative $h^{2'}$ inferred as a function of MAF for different frequency filter thresholds (purple, green, blue, brown), and when singletons are partitioned by population MAF (based on gnomAD, red). Including all SNPs and partitioning singletons by population MAF (instead of observed MAF) results in a substantially increased level of $h^{2'}$. **d**, Globally rare singletons represent 56% of all singletons, but contribute 92% of $h^{2'}_{singleton}$. Rare indels and structural variants also have enriched contributions to heritability (2.8% of singletons but 7.8% of $h^{2'}_{singleton}$). However, singletons inferred to derive from Neanderthal introgression or having gnomAD MAF $\geq 10^{-4}$ make negligible contributions to $h^{2'}_{singleton}$. In all cases, confidence intervals/envelopes are based on the 95% quantile range of 1,000 bootstrap simulations.

included the first ten principal components from both genetic and phenotypic data in all analyses; we present the average h^2 estimate across genes in each MAF bin in Fig. 2a (blue curve). We found that h^2 is highest for the first MAF bin (singletons). However, using a new *trans* permutation procedure, we detected evidence for residual population stratification in low-frequency (but not high-frequency) SNPs that could not be accounted for using principal components (pink curve; see Supplementary Note). Note that differential population structure among common and rare variants is a documented, although understudied, phenomenon in human genetics[30]. We corrected for this population stratification bias by subtracting the permutation-based estimate from the raw principal component-corrected h^2 estimate, shown in purple and henceforth indicated as $h^{2'}$. We found that the plurality of $h^{2'}$ comes from singletons, but common variants also contribute a substantial amount toward $h^{2'}$. Low- and intermediate-frequency SNPs make a minimal contribution to $h^{2'}$. Note that this is a conservative correction because our *trans* permutations capture both the effect of stratification and true *trans* heritability.

Figure 2b shows the proportion of $h^{2'}$ explained by each MAF bin, showing that singletons represent approximately 25% of the total $h^{2'}$, dominating the estimates from other MAF bins. Based on population genetic theory[9,10,12,31], we hypothesized that purifying selection has constrained causal regulatory alleles to low frequency. To test this hypothesis, we sorted our singletons by their population MAF, as inferred from a large external database (gnomAD). We reasoned that some of the singletons in our dataset would be evolutionarily neutral and have an intermediate population frequency, whereas the most deleterious singletons would almost always be constrained to a low population frequency. Therefore, we partitioned the singletons observed in our data by their MAF observed in the Genome Aggregation Database (gnomAD) dataset (representing high-coverage WGS on > 15,000 individuals) and performed Haseman–Elston inference of $h^{2'}$ across 20 singleton bins based on their MAF observed in gnomAD. (We also partitioned by functional predictions and evolutionary conservation; see Supplementary Note.) The inset in Fig. 2b shows that the vast majority (>90%) of singleton $h^{2'}$ is derived from variants that have a gnomAD MAF < 0.01%. This is strong evidence that natural selection constrains alleles with the largest effects on gene regulation to very low frequency. Notably, we found that 31% of our singletons were not reported in gnomAD, but this subset of variants (indicated by an asterisk in Fig. 2b) nonetheless explains approximately 80% of $h^{2'}_{singleton}$. We confirmed that the majority of this signal is derived from true-positive singletons by analyzing a subset of 58 individuals with high-coverage WGS and estimated that 88% of $h^{2'}_{singleton}$ is derived from variants that validate (Supplementary Note). Previous work has shown that additionally partitioning common variants by linkage disequilibrium resulted in minimal change after partitioning by MAF[5].

Studies of heritability typically filter out rare variants[5,15,32]. We showed that removing any SNPs based on MAF has a direct impact on the estimate of heritability. Figure 2c shows the cumulative h^2 inferred as a function of MAF for different minor allele count (MAC) thresholds (averaged over all genes). We found that adding progressively rarer variants to the analysis resulted in a monotonic increase in inferred heritability. Including all variants down to singletons (purple curve) increases $\widehat{h^{2\prime}}_{total}$ by approximately 50% ($\widehat{h^{2\prime}}_{total} = 0.061$) compared to the case when only common variants (MAF ≥ 5%) are analyzed (brown curve, $\widehat{h^{2\prime}}_{common} = 0.04$), indicating that common variants cannot tag heritability from lower-frequency variants (that is, 'synthetic association' tagging[33] is minimal, although rare variants can tag some common variant heritability; see Supplementary Note). However, not all singletons contribute equally to heritability and pooling them together can deflate $h^{2'}$ estimates (a 'singleton linkage disequilibrium' effect previously only reported for common variants[5,28]; see Supplementary Note). Partitioning singletons into 6 bins based on their observed MAF in gnomAD (red curve) increased our $\widehat{h^{2\prime}}_{total}$ estimate to 0.082 and showed that nearly half of the total heritability (46.6%) is explained by the 27.6% of variants that are globally rare (with $MAF_{gnomAD} < 0.1\%$).

Recent studies of gene expression variation in humans have suggested that one-quarter of Neanderthal-introgressed haplotypes have *cis*-regulatory effects[34] and that expression outliers are enriched for having nearby rare structural variants compared to nonoutliers[8]. However, the overall contribution of these classes of variants to expression variation had not been characterized. We performed Haseman–Elston regression on four disjoint categories of singletons (Neanderthal-introgressed, indels/structural variants, globally rare singletons and other singletons) and found that globally rare singletons (that is, singletons in our data that are also singletons across all 2,504 samples in the 1KGP[24]) contribute the vast majority (92%) of singleton heritability (Fig. 2d). Rare indels/structural variants also have an enriched contribution to gene expression variation (representing 2.8% of singletons, but 6.8% of $h^2_{\text{singleton}}$), but Neanderthal-introgressed singletons and other singletons make a negligible contribution to $h^{2'}_{\text{singleton}}$.

Genotype quality does not drive the inference of heritability. One possible confounding factor is the effect of genotyping error on heritability estimation[35]. If heritability is biased by genotyping error and genotyping error also varies as a function of MAF, there could be differential bias across frequency bins when analyzing real data. We simulated a range of genotyping error models and found that all investigated forms of genotyping error increased the variance of heritability estimation, but did not induce a detectable upward bias (Supplementary Note).

We also performed several analyses to examine the possible confounding effects in these data (Supplementary Note). First, we ranked singletons by their reported genotype likelihood as reported for the individual carrying the singleton allele in 1KGP[24] and partitioned them into four equal groups (quartiles). We then ran Haseman–Elston regression with these four groups of singletons (along with ten principal components). Notably, we found that only those singletons with high SNP quality contributed positively to our inference of heritability (see Supplementary Note). Second, since both DNA sequencing and RNA-seq are based on LCLs, it is conceivable that difficult-to-sequence regions of the genome could result in correlated errors that confound our inference. To test this, we restricted our analysis to regions of the genome passing the 1KGP strict mask[24] and found that our inference of heritability was unchanged. We further ranked genes based on the number of exon bases passing the strict mask and found no difference in the genetic architecture of genes having high versus low overlap with the strict mask (see Supplementary Note). Finally, a subset of $n = 58$ samples were sequenced at high coverage by Complete Genomics as part of the 1KGP[24]. We identified the singletons carried by these individuals and partitioned them into four groups by cross-classifying them as being present or absent in the Complete Genomics or gnomAD datasets. Running Haseman–Elston regression on this subset of individuals shows that $h^{2'}_{\text{singleton}}$ is predominantly driven by singletons that replicate in the Complete Genomics data but are not reported in gnomAD (consistent with Fig. 2), and that singletons that are absent from Complete Genomics (and therefore more probably false positives) contribute negligibly to $h^{2'}_{\text{singleton}}$ (Supplementary Note).

Selection drives the genetic architecture of gene expression. We found that rare variants are a major source of heritability of gene expression, which we hypothesized was due to purifying selection constraining the frequencies of large-effect alleles. To test this hypothesis, we performed extensive simulations of human evolutionary history[36,37] and developed a method to infer the parameters of an evolutionary model for complex traits (see Supplementary Note). Our three-parameter phenotype model extends a previously described model of the pleiotropy of causal variation[11]—captured by ρ, where increasing values indicate higher correlations among expression effect sizes and the fitness effects acting on causal variants—and the scaling relationship between expression effect sizes and selection coefficients[9] (τ, where increasing values indicate that the distribution of effect sizes has a longer tail toward strong effects), to include the overall strength of selection (ϕ), a mixture parameter between strong and weak selection distributions, where $\phi = 1$ corresponds to strong selection. We inferred the approximate posterior distributions for each of these parameters using rejection sampling[38], which compares a set of informative summary statistics from genetic data simulated under a model of European demography[39] and selection[40,41] to the observed data (see Supplementary Note). Note that our inference procedure allows each parameter to vary across genes, but we only sought to infer the distribution of the average values of ρ, τ and ϕ across genes because we did not have the statistical power to infer ρ and τ for each gene. We rigorously evaluated the performance of this inference procedure with simulations and found that we could infer ρ and τ with fairly high accuracy; however, ϕ (while broadly unbiased) is less informative (Supplementary Note).

Applying this model to our data, we found that purifying selection had a major impact on the genetic architecture of human gene expression and that a range of previously explored evolutionary models can plausibly explain the empirical data. In Fig. 3a, we plotted the posterior distributions of the mean values of ϕ, ρ and τ. This suggested that, on average, the fitness effects acting on causal variants tend to follow the distribution inferred from conserved noncoding loci ($\bar{\phi} \approx 0$), but selection is pervasive in the sense that gene expression effect sizes are highly correlated with the fitness effects acting on causal variants. Figure 3b shows that our data are consistent with a ridge of evolutionary scenarios that connect models where causal alleles are highly modular (for example, effect sizes are correlated with dampened fitness effects, as in the model of Eyre-Walker[9], which assumes $\rho = 1$ with intermediate τ) and models with highly pleiotropic causal alleles and more extreme effect sizes (for example, the Simons et al.[11] model, which assumes $\tau = 1$, but a more moderate ρ). This observation could only be identified using our integrated model and suggests highly heterogeneous processes acting on individual genes. Our parameter inference suggests that while mean ρ, τ and ϕ can vary substantially among the best-fitting models, individual genes tend to have extreme values (that is, either 0 or 1) for all three parameters (Fig. 3a). Figure 3c shows the cumulative proportion of h^2 as a function of MAF from 1,000 bootstrap draws from our posterior distribution, along with the cumulative proportion of $h^{2'}$ inferred from our data. Compared to a neutral evolutionary model (pink), the posterior draws (gray, representing points along the ridge of evolutionary phenotype models show in Fig. 2b) are all highly concordant with our data.

Discussion

There is great interest in characterizing the genetic basis for complex traits to improve our understanding of human health and disease and substantial resources are being spent to collect ever-larger cohorts to investigate the role of rare variants. Such studies will clarify what we have learned from our relatively small study of just 360 individuals. We developed, tested and applied a technique for interrogating the role of rare variants in gene regulation using a relatively small cohort of $n = 360$ individuals who had whole-genome DNA sequencing and RNA-seq performed on their derived LCLs. We estimated that the total narrow-sense heritability of LCL gene expression is approximately 8.2% and that the largest contributors to gene expression heritability are the rarest of variants in our data, that is, singletons where just one copy of the allele has been observed in our sample of 720 chromosomes (MAF = 0.0014). Globally rare variants ($\text{MAF}_{\text{gnomAD}} < 0.01\%$) explain approximately 90% of $h^2_{\text{singleton}}$, implying that many of these causal variants would remain singletons even if tens of thousands more samples were sequenced and many more

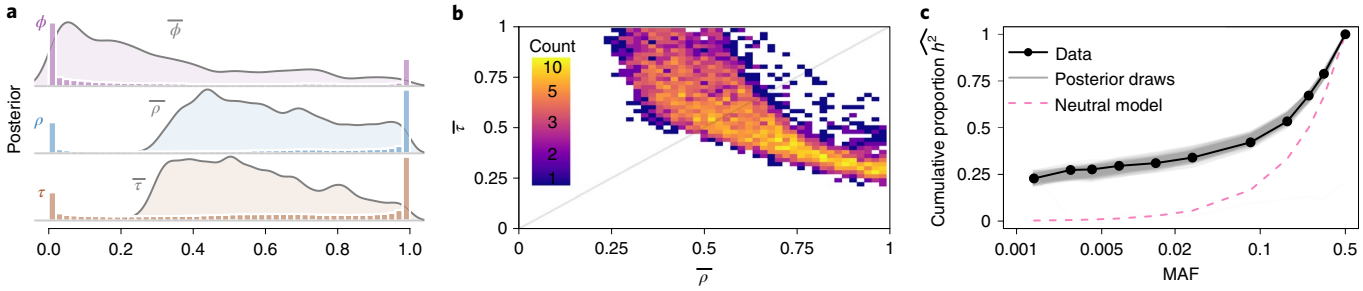

Fig. 3 | Pervasive purifying selection drives the genetic architecture of gene expression. Our model infers the strength of purifying selection acting on causal variants (ϕ), the correlation between the fitness and the effect size of causal variants (ρ) and a scaling factor that transforms fitness into effect sizes (τ). **a**, The posterior distribution of the mean of each parameter across genes (curves) and a histogram of the posterior parameter estimates for each gene. **b**, The joint posterior distribution of the average ρ and τ across genes shows an evolutionary trade-off between the correlation and scaling of fitness and effect sizes. **c**, The cumulative proportion of heritability inferred from the gene expression data (dots) compared to the expected patterns from 1,000 draws from the posterior distribution (gray) and neutral expectation (pink).

singletons would be discovered. However, given that the plurality of variants is ultrarare, do we infer more heritability than would be expected given the fraction of variants observed at these frequencies? In the Supplementary Note, we show that heritability enrichment is 'U'-shaped as a function of MAF (on a logarithmic scale), suggesting that both rare and common alleles contribute more than twofold excess of heritability, while intermediate/low-frequency variants (MAF = 0.1–5%) constitute a dearth of heritability. This does not give us direct insight into the underlying distribution of regulatory effect sizes per causal variant, but it would be reasonable to speculate that the distribution of effect sizes for rare causal variants may be considerably larger (in absolute values) than common variants.

This excess of heritability due to ultrarare variants is best explained by pervasive purifying selection, where most *cis*-acting regulatory variants are deleterious. We inferred the parameters of an evolutionary model that are consistent with these data and found that for approximately two-thirds of genes, the effect sizes of *cis*-regulatory variants are highly correlated with how deleterious the fitness effects are on causal variants. Further, for the majority of genes, the fitness effects are more consistent with broadly defined conserved noncoding regions of the genome[40] than the strongly selected nonsynonymous[41] or ultraconserved regions of the genome[42]. However, while these parameters allow us to generate simulated data consistent with our observations, they remain post hoc parametric models that do not necessarily represent a generative model of how the genetic architecture of *cis*-regulatory variation evolved, and do not incorporate potentially important contributions from other modes of natural selection (such as positive or balancing selection, which may be rare but can have substantial impact on gene expression when they act[43]).

Our estimate of total *cis* heritability is slightly larger than the previous estimates of $h^2_{cis} = 0.057$ and 0.055 in blood and adipose tissue, respectively[44], but lower than recent twin-based estimates of overall narrow-sense heritability $h^2 = 0.26$, 0.21 and 0.16 in adipose tissue, LCLs and skin, respectively[45] as well overall broad-sense heritability $H^2 = 0.38$ and 0.32 for LCLs and whole blood[46]. Therefore, it is plausible that rare variants account for some 'missing heritability' in human gene expression; however, differences in population, tissue and/or technology could also explain some of these patterns and there could also be differences between the genetic architecture of *cis* and *trans* regulation.

A concurrent examination of rare variant heritability via an allele-specific expression approach[47] reported a lower, yet substantial contribution to heritability from rare variation. However, there are fundamental differences between our analyses that probably contribute to the difference in estimates. First, the work by Glassberg et al.[47] examined a much narrower window around genes. This leads to differences if selection has acted differently in promoters compared to more distal regulatory regions[48] (Supplementary Note). Second, their work used a smaller sample size; thus, their definition of rare is less stringent than ours. Finally, they did not reclassify rare variants according to external reference panels, which greatly increased our estimates of rare variant heritability.

While it might at first seem logical to genotype some (or all) of these singletons in a larger panel of individuals to statistically identify the causal ones, our analysis uncovered a major challenge with this approach. Our results can only be explained if the causal alleles driving heritability are evolutionarily deleterious, with effect sizes often scaling with the strength of selection acting on them. This means that the alleles that have the greatest impact on gene expression are probably extremely rare in the broader population and are unlikely to exist in more than a few unrelated individuals in any given population. Indeed, our analysis shows that 90% of the singleton heritability is derived from alleles that are either not reported or have a MAF < 0.01% in the $n > 15{,}000$ samples in gnomAD. Therefore, we conclude that identifying causal alleles for transcriptional variation probably requires the incorporation of new biological information, possibly including large-scale experimental testing of singleton variants to improve functional predictions.

As the number of samples with detailed phenotype data and WGS data increases, it will be possible to apply the approach we have developed in the present study to characterize the genetic architecture of additional complex traits. Indeed, in a recent WGS study of height and body mass index, we found that rare variants comprise essentially the entirety of 'missing heritability' for these traits[49]. By integrating such methods with functional genomic data, we may also learn more about the biology of causal variants, which could enable improved identification of clinically actionable variants in some cases. However, it is not clear that risk prediction from genomic data for most diseases will be feasible for otherwise healthy individuals with limited family history information. Population genetic theory tells us that rare variants only contribute a substantial source of heritability when causal alleles are evolutionarily deleterious. But the biology of human health and disease is complex. While not all human diseases themselves impart a strong fitness effect, extensive pleiotropy resulting from tightly interconnected networks of interacting proteins experiencing cell-specific regulatory mechanisms could. Indeed, under the omnigenic model of disease, variants that affect any one of these components could contribute to an individual's risk for any disease involving any downstream pathway[23].

We developed an approach to examine the heritability of singleton variants and the results have important implications for future genetic studies. We rigorously evaluated the performance of our

inference procedure using extensive simulations and multiple types of permutations (see Supplementary Note). While we employed several approaches to test for the presence of confounders from population structure, genotyping/mapping error and cell line artifacts, there may be other unknown confounders that have biased the results of this study. We conservatively used quantile normalization on the expression phenotypes to enforce normality and this often reduces the overall heritability estimates (see Supplementary Note) by diminishing the impact of outliers[8,50]. There are several other contributors to broad-sense heritability that we have not attempted to model; they may also account for some of the heritability estimated in family-based studies, such as gene–gene interactions, gene–environment interactions and other nonadditive components.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0487-7.

Received: 16 December 2018; Accepted: 8 July 2019;
Published online: 2 September 2019

References

1. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
2. Bamshad, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
3. Zhao, J. et al. A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **98**, 299–309 (2016).
4. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7**, e1002144 (2011).
5. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
6. Gaugler, T. et al. Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
7. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
8. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
9. Eyre-Walker, A. Evolution in health and medicine Sackler colloquium: genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl Acad. Sci. USA* **107**, 1752–1756 (2010).
10. Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S. & Hernandez, R. D. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res.* **26**, 863–873 (2016).
11. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
12. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).
13. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
14. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
15. Mancuso, N. et al. The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**, 30–35 (2016).
16. Schoech, A. P. et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).
17. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
18. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
19. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
20. Gusev, A. et al. Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).
21. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
22. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
23. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
24. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
25. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
26. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl Acad. Sci. USA* **111**, E5272–E5281 (2014).
27. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
28. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
29. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
30. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
31. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, e2002985 (2018).
32. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
33. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
34. McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* **168**, 916–927.e12 (2017).
35. Chen, L., Liu, P., Evans, T. C.Jr. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**, 752–756 (2017).
36. Uricchio, L. H., Torres, R., Witte, J. S. & Hernandez, R. D. Population genetic simulations of complex phenotypes with implications for rare variant association tests. *Genet. Epidemiol.* **39**, 35–44 (2015).
37. Hernandez, R. D. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**, 2786–2787 (2008).
38. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
39. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
40. Torgerson, D. G. et al. Evolutionary processes acting on candidate *cis*-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet.* **5**, e1000592 (2009).
41. Boyko, A. R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008).
42. Katzman, S. et al. Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
43. Andrés, A. M. et al. Balancing selection maintains a form of *ERAP2* that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* **6**, e1001157 (2010).
44. Price, A. L. et al. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).
45. Grundberg, E. et al. Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
46. Powell, J. E. et al. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* **22**, 456–466 (2012).
47. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Measurement of selective constraint on human gene expression. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/345801v1 (2018).
48. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
49. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/563866v1 (2019).
50. Schweiger, R. et al. Detecting heritable phenotypes without a model using fast permutation testing for heritability and set-tests. *Nat. Commun.* **9**, 4919 (2018).

Acknowledgements

We thank H. M. Kang, A. Auton and S. Gusev for discussions about possible confounders that improved our analysis; members of the Pritchard laboratory for comments on rejection sampling; J. Barrett and K. Karczewski for peer-reviewing our preprint; R. Torres for assistance with data analysis; J. Wall for assistance with the Neanderthal-introgressed alleles;

and A. Hernandez for discussions on figure colors. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health (award no. R01HG007644 to R.D.H. and award no. K25HL121295). L.H.U. was supported by an Institutional Research and Academic Career Development Award (National Institute of General Medical Sciences, grant no. K12GM088033). K.H. was supported by a Gilliam Fellowship for Advanced Study; A.D. was supported by NIH (award nos. U01HG009080 and R01HG006399).

Author contributions

R.D.H. and N.Z. conceived and designed the study. L.H.U. and A.D. developed methods. R.D.H., L.H.U., K.H., C.Y., A.D. and N.Z. contributed to data analysis or simulations. R.D.H. and N.Z. wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0487-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.D.H. or N.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

The full methodological details can be found in the Supplementary Note accompanying this manuscript (along with the simulation results, testing robustness of model assumptions and evolutionary modeling). In this article, we provide details of the primary methods used for data analysis.

Frequency-binned Haseman–Elston regression. Given genotypes at M SNPs over N individuals we considered additive phenotypic models such that the phenotype of individual i is: $y_i = \sum_{j=1}^{M} g_{ij}\beta_j + \epsilon_i; \epsilon_i \sim N(0, \sigma_e^2)$, where g_{ij} is the genotype of individual i at SNP j, β_j is the effect size of SNP j and ϵ_i is the residual, independent and identically distributed, normally distributed noise of individual i. We partitioned the SNPs into K disjoint sets determined by the MAF of each SNP (or some other characteristic indicated in the text) and estimated the contribution of SNPs in the k^{th} set to the heritability of y: $h_k^2 = \sigma_k^2/\sigma_y^2$, where σ_j^2 is the genetic variance contributed by all of the SNPs in the kth partition, $\sigma_g^2 = \sum_{k=1}^{K} \sigma_k^2$ is the total genetic variance and $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$ is the total phenotypic variance, assumed to be equal to 1 going forward.

To infer the heritability of gene expression levels across individuals, we primarily relied on Haseman–Elston regression[26]. The premise of Haseman–Elston regression is that heritability can be estimated by the correlation between the phenotypic covariance across individuals and the genotypic covariance across individuals. In practice, for a single gene, we estimate the phenotypic covariance (P) as the upper triangle of the outer product of quantile-normalized $\log_2$(fragments per kilobase of transcript per million mapped reads) across our sample. For each of the K partitions, we estimated genotypic covariance with the upper triangle of a kinship matrix generated from all SNPs in the partition. Given a standardized genotype matrix of SNPs in the kth partition (G_k, with N rows and M_k columns, where each column has mean 0 and unit variance), the kth kinship matrix is $R_k = G_k G_k'/M_k$. Haseman–Elston regression is then performed using the lm() function in R:

$$P \approx R_1 + \ldots + R_K$$

Specifically, the regression is ordinary least squares applied to the (vectorized) strict upper triangles of these matrices, which, for N individuals, has $\binom{N}{2}$ elements. In Haseman–Elston regression, with scaled and centered genotypes and phenotypes, the effect size for the kth term represents the genetic variance explained by the kth SNP partition $(\beta_k = \sigma_k^2)$, with the total genetic variance explained by all SNPs given by $\sigma_g^2 = \sum_{k=1}^{K} \sigma_k^2$. In the absence of other genetic contributions to phenotypic variation, heritability is equal to the total additive genetic variance explained by SNPs, $h^2 = \sigma_g^2$. Therefore, in most instances in this article we simply refer to the genetic variance explained as heritability.

In general, we included the first ten principal components generated from our genome-wide genotype matrix as well as the first ten principal components generated from our transcriptome-wide expression matrix (described later). We show in our Supplementary Note that the number of principal components included does not qualitatively impact our results. Formally, we include the jth principal component (or an arbitrary numerical covariate) by adding the upper triangle of the principal component (or covariate) outer product with itself to our symbolic regression equation outlined earlier. Our results suggest that inclusion of principal components and other covariates did not completely account for population structure, especially in the low-frequency bins. Therefore, we relied on a *trans* sampling approach (see Supplementary Note) to account for residual population structure. Importantly, these results suggest that other investigations into rare variant heritability may not be completely accounting for population structure by simply including principal component covariates.

GEUVADIS dataset and quality control. RNA-seq gene expression data were downloaded from http://www.internationalgenome.org/data-portal/data-collection/geuvadis. This dataset contains 375 individuals of European descent from 4 locations. Each of these individuals are contained in the 1KGP and genome sequence data were downloaded from www.1000genomes.org (ref. [24]).

The GEUVADIS data consists of RNA-seq data for 464 LCL samples from 5 populations in the 1KGP. Of these, 375 are of European ancestry (CEU, FIN, GBR, TSI) and 89 are of African ancestry (YRI). In these analyses, we considered only the European ancestry samples. Some individuals were previously identified as having cryptic relatedness by the 1KGP[24] using identity by state analyses and were therefore pruned. Our resulting dataset contains 360 unrelated individuals of European descent from 4 populations. Raw RNA-seq reads obtained from the European Nucleotide Archive (www.ebi.ac.uk/ena) were aligned to the transcriptome using University of California Santa Cruz annotations matching hg19 coordinates. RNA-seq by expectation-maximization[51] was used to estimate the abundances of each annotated isoform; total gene abundance was calculated as the sum of all isoform abundances normalized to one million total counts or transcripts per million. For each population, one million total counts or transcripts per million were $\log_2$-transformed and median normalized to account for differences in sequencing depth in each sample. The genotype data was obtained from the 1KGP phase 3 V5 dataset[24]. To remove potential confounders, such as population structure and batch effects, we performed principal component analysis (PCA).

PCA analyses. PCA was performed on both genome-wide genotype data and transcriptome-wide expression data. We obtained expression principal components from http://www.internationalgenome.org/data-portal/data-collection/geuvadis and ran PCA on the WGS data as follows.

1KGP phase 3 V5 variant call files. VCFtools v.0.1.14 (ref. [52]) was used to filter out related individuals, exclude singletons sites, remove indels and filter out all nonbiallelic sites.

PLINK v.1.90b3x (ref. [53]) was used to identify sites approximately in linkage equilibrium $r^2 < 0.2$ examining 50 kb windows in 5 site increments, extract these sites and recode in an additive model (0, 1, 2).

R (https://www.r-project.org/) was used to concatenate chromosomes and run PCA on the centered and scaled genotype matrix.

We also ran PCA on the genotype data with a higher MAF filter (MAF ≥ 5%) and got highly correlated results. However, because our analysis is based on rare variants, we wanted to include signals of population structure that manifest primarily in rare variants, hence including all variants seen at least twice.

We then checked for residual population structure using permutations. We first applied the standard permutation style, whereby phenotypes are shuffled among individuals before running the Haseman–Elston regression. We found that this removed all signals in the data and gives $\widehat{h^2} = 0$. We then developed another permutation, which we refer to as a *trans* permutation. In this case, we maintain the order of gene expression and genotypes among individuals, but we perform Haseman–Elston regression on the SNPs in a window around one gene with the expression values of a random autosomal gene (that is, a gene in *trans*). We show the results of this permutation in Fig. 1a and in several supplementary figures. We found some degree of residual population structure for rare variants, but not common variants (despite the fact that we included rare variants in our PCA analysis). The main caveat with this approach is that we are unable to distinguish population structure from pervasive true *trans* effects, but we argue that removing the residual $\widehat{h^2}$ from the *trans* permutation is conservative.

Constructing bootstrap confidence intervals. In Fig. 2 and in the Supplementary Note, we compare heritability estimates in many ways. Our primary approach to estimating uncertainty was based on rigorous bootstrapping. Except where noted, all error bars (sometimes plotted as envelopes encompassing the mean) were calculated from the 95% interquartile range of 1,000 bootstrap samples. This is an appropriate method for estimating uncertainty in independent and identically distributed data, and has previously been shown to work well in far broader settings[50]. Further, bootstrapping is a statistically appropriate way to estimate uncertainty when analyzing functions of correlated parameter estimates, for example, when estimating total h^2, which is the sum across h^2 estimates per bin. These bootstrap intervals represent uncertainty in the across-gene average heritability estimates per category (indeed, the single-gene uncertainties are much larger) and retain any across-gene correlations that are present in the real data. Hence, our s.e.m. estimates naturally account for correlated expression.

Evolutionary modeling and parameter inference. We suppose that gene expression is evolutionarily optimized, such that mutations that affect gene expression levels are deleterious. While many different evolutionary models can encode this qualitative behavior, we chose a previously studied theoretical model that allows for variation in pleiotropy and selection strength across genes[10].

We used rejection sampling to infer evolutionary parameters. Rejection sampling compares a set of informative summary statistics computed on the output of model-based simulations to observed genomic and phenotypic data. The simulations that generate summary statistics that are most similar to the observed data are retained and the parameter values from the retained simulations are used to generate a posterior distribution over the true parameter values. In the present study, we took the proportion of variance explained by alleles up to minor allele count x as summary statistics, for x in {1,2,5,10,20,60,120,180,240,360}. We focused on inferring the mean strength of selection ($2Ns$), the correlation between selection strength and effect size (ρ), the mean shape of the effect size distribution (τ) and the selection strength on *cis*-regulatory variants (ϕ, representing the proportion of regulatory variants under strong negative selection). We inferred the posterior distribution of the mean of each of these parameters across genes as opposed to the parameter values for individual genes because single-gene estimates proved too noisy to be reliably computed.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

RNA-seq gene expression data were downloaded from http://www.internationalgenome.org/data-portal/data-collection/geuvadis. This dataset contains 375 individuals of European descent from 4 locations. Each of these

individuals are contained in the 1KGP and genome sequence data were downloaded from www.1000genomes.org (ref. [24]).

Code availability

Three open source software tools are being made available as part of this study; all are available on GitHub: (1) HEh2.R—R code that performs all the Haseman–Elston analyses and simulations discussed in this paper. It also implements the artificial intelligence algorithm for parameter inference of linear mixed models. It is available from https://github.com/hernrya/HEh2; (2) SingHer R package discussed in the Supplementary Note, with performance statistics and available from https://github.com/andywdahl/SingHer; and (3) rejection sampling: scripts demonstrating how we used rejection sampling to infer parameters of the phenotype model are available from https://github.com/uricchio/HE_scripts.

References

51. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
52. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
53. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

nature research

Corresponding author(s): *Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data collection was performed for this analysis. No software were used for data collection.
Data analysis	Three open source software tools are being made available as part of this study, all available on GitHub: HEh2.R – R code that performs all H-E analyses and simulations discussed in this paper. Also implements AI algorithm for parameter inference of LMM. Available here: https://github.com/hernrya/HEh2. Contact: Ryan Hernandez <ryan.hernandez@me.com>. For data analysis, we used version posted on April 25, 2019. SingHer R package – Singleton Heritability inference with REML, discussed in Section 2.4, with performance statistics in Table S2, and available here: https://github.com/andywdahl/SingHer. Contact: Andrew Dahl <andywdahl@gmail.com>, Noah Zaitlen <noahaz@gmail.com>. For data analysis, we used version posted August 28, 2018. Rejection sampling: Scripts demonstrating how we used rejection sampling to infer parameters of the phenotype model are available here https://github.com/uricchio/HE_scripts. Contact: Lawrence Uricchio <uricchil@gmail.com>. For data analysis, we used version posted April 24, 2019.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are publicly available with no restrictions.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	360 unrelated European individuals were available with deep RNA sequencing (through GEUVADIS) and whole genome sequencing (through 1000 Genomes Project).
Data exclusions	Related individuals can bias results, and therefore individuals with a relationship closer than 3rd cousin were removed. This decision was made prior to any analysis.
Replication	There was no data collection, and therefore there were no steps taken to ensure reproducibility of experimental findings. To ensure reliability of our inference, we used an alternative subset of genotype data from high coverage whole genome sequencing.
Randomization	We performed many analyses to dissect stratification.
Blinding	Our analysis is based on the inference of heritability of gene expression, as such knowledge of each individual's expression values and genotypes were essential, and therefore blinding was not possible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging