Singleton Variants Dominate the Genetic Architecture of Human Gene Expression

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

Affiliations:

1) Bioengineering & Therapeutic Sciences, UCSF, San Francisco, CA;

2) Institute for Human Genetics, UCSF, San Francisco, CA;

3) Institute for Quantitative Biosciences, UCSF, San Francisco, CA;

4) Institute for Computational Health Sciences, UCSF, San Francisco, CA;

5) Department of Biology, Stanford University, Stanford, CA;

6) Biological and Medical Informatics Graduate Program, UCSF, San Francisco, CA;

7) Epidemiology & Biostatistics, UCSF, San Francisco, CA;

8) Department of Medicine Lung Biology Center, UCSF, San Francisco, CA.

*Correspondence To:

Ryan D. Hernandez Ryan.Hernandez@ucsf.edu,

Noah Zaitlen Noah.Zaitlen@ucsf.edu

In Brief: The vast majority of variants so far discovered in humans are rare, and together they have a substantial impact on gene regulation.

Highlights:

- Singleton variants are by far the largest contributor to gene expression heritability
- Globally rare variants explain over half of the *cis*-heritability of gene expression
- Heritability estimation can be downwardly biased by excluding rare variants
- Rampant purifying selection acts on variants that contribute to transcriptional regulation

SUMMARY

The vast majority of human mutations have minor allele frequencies (MAF) under 1%, with the plurality observed only once (i.e., “singletons”). While Mendelian diseases are predominantly caused by rare alleles, their cumulative contribution to complex phenotypes remains largely unknown. We develop and rigorously validate an approach to jointly estimate the contribution of alleles with different frequencies, 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 RNA sequencing data from 360 European individuals, we find that singletons alone contribute ~23% of all *cis*-heritability across genes (dwarfing the contributions of other frequencies). Strikingly, we find that the vast majority (67-90%) of singleton heritability derives from ultra rare variants that are absent from thousands of additional samples. Further, over half of all *cis*-heritability is contributed by globally rare variants (MAF<0.1%), which we show is the result of rampant purifying selection shaping the regulatory architecture of most human genes.

INTRODUCTION

The recent explosive growth of human populations has produced an abundance of genetic variants with minor allele frequencies (MAF) less than 1% (Keinan and Clark, 2012). While many rare variants underlying Mendelian diseases have been found (Bamshad et al., 2011), their role in complex disease remains unknown (Fuchsberger et al., 2016; Gaugler et al., 2014; Li et al., 2017; Montgomery et al., 2011; Yang et al., 2015; Zhao et al., 2016). Evolutionary models predict that the contribution of rare variants depends highly on selection strength (Eyre-Walker, 2010; Uricchio et al., 2016), and that population growth can magnify their impact (Simons et al., 2014; Uricchio et al., 2016). Recent methodological breakthroughs (Das et al., 2016; McCarthy et al., 2016) 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 likely driven by natural selection (Gazal et al., 2017; Mancuso et al., 2016; Schoech et al., 2017; Yang et al., 2015; Zeng et al., 2017). However, these studies excluded the rarest variants (Mancuso et al., 2016) or included only well-imputed variants (Yang et al., 2015). 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, direct genetic research efforts, and clarify how natural selection has shaped human phenotypes.

In this work, we develop, validate, and apply an approach for inferring the relative phenotypic contributions of all variants, from singletons to high frequency. 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 (Bulik-Sullivan et al., 2015; Gusev et al., 2013; Maurano et al., 2012), and recent examinations have identified significant rare variant effects on transcription (Li et al., 2017). To characterize the genetic architecture of gene expression, we analyze 360 unrelated individuals of European ancestry with paired whole genome DNA (1000 Genomes Project Consortium et al., 2015) and RNA (Lappalainen et al., 2013) sequencing 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 (see Supplemental Information).

RESULTS

Building and testing our model

Before analyzing 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 use real genotype data [all variants within 1 megabase (Mb) of the transcription start or end sites of genes] and generate gene expression phenotypes across individuals while varying the number of causal variants contributing to the phenotype (from 1 to

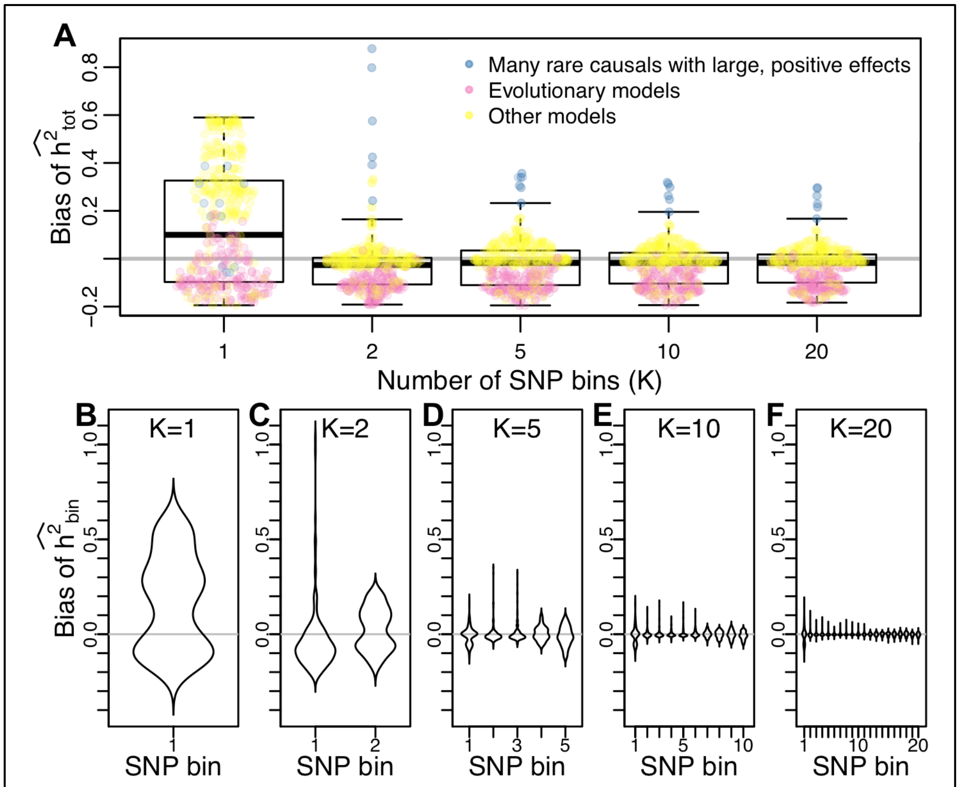

Fig 1. Simulation results showing that across a broad range of parameters, the accuracy of heritability inference improves as the number of SNP bins (partitioned by MAF) increases. (**A**) Mean bias of total heritability (inferred-true) for different numbers of SNP bins (K), where each point represents the mean of 500 simulations for different parameters, and a box plot summarizing the bias distribution across all parameters. (**B-F**) The distribution of average bias across simulated parameters for each SNP bin, showing that both mean and variance of the bias decrease as K increases.

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 Supplemental Information). In total, we simulated 440 different genotype-phenotype models that span beyond the range of genetic architectures that could plausibly underlie complex phenotypes such as gene expression, and analyzed each simulated dataset multiple ways. A common approach for estimating heritability in unrelated samples is to fit a linear mixed model (LMM) via restricted maximum likelihood [REML (Golan et al., 2014; Yang et al., 2014)]. However, Haseman-Elston (H-E) regression [an alternative approach based on regressing phenotypic covariance on genotypic covariance (Golan et al., 2014)] is more robust in small samples (see Supplemental Information).

Similar to previous work (Speed et al., 2012), we found that for many simulation settings, jointly analyzing all variants together can result in a substantial over- or underestimate of heritability (Figure 1A, which shows results when true heritability is 0.2). One common solution is to partition sites by frequency (Mancuso et al., 2016; Speed et al., 2017; Yang et al., 2015). Simply isolating rare (MAF<=1%) from common variants using two partitions and performing joint inference (Mancuso et al., 2016) can 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 by MAF (Yang et al., 2015) alleviates this problem. Remarkably, not only does the overall heritability bias decrease as the number of allele frequency categories increases, but Figure 1B-F shows that the bias of the heritability for each MAF bin also decreases substantially across all models (see Supplemental Information). These simulations suggest that with our sample size, partitioning SNPs into 20 MAF bins results in the smallest bias in our estimate of total heritability as well as the smallest bias for each bin across all simulated parameters.

One possible confounding factor is the effect of genotyping error on heritability estimation (Chen et al., 2016). 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 considered a range of genotyping error models, and found that all investigated forms of genotyping error eroded efficiency of heritability estimation, but did not induce a detectable upward bias (see Supplemental Information).

When partitioning variants into multiple MAF bins, singletons are quickly isolated into their own category. Intuitively, if some fraction of singletons is causal, then individuals with higher singleton load may be more likely to be phenotypic outliers. It is therefore reasonable to ask what contribution singletons make to patterning phenotypic variation across a population. We therefore investigated the theoretical properties of heritability estimation from singleton variants, and show analytically that when genotypic covariance is estimated using singletons alone, H-E regression is equivalent to regressing squared phenotypes against singleton counts (see Supplemental Information).

A direct implication of our derivation is that H-E regression is unbiased unless singletons have large non-zero mean effect sizes (violating an explicit assumption of both H-E regression and LMMs). Interestingly, these are precisely the simulation scenarios in Figure 1A where heritability estimates remain upwardly biased (blue points). We develop an alternative approach that produces unbiased estimates of both heritability and mean effect size in all examined cases (see Supplemental Information), but because H-E regression is well understood and flexible, we recommend its use when mean effect sizes are near zero.

Singletons drive the genetic architecture of human gene expression

In order to characterize the genetic architecture of human gene regulation, we partitioned the heritability of gene expression by frequency. We used *n*=360 unrelated individuals of European descent with RNA sequencing data from GEUVADIS (Lappalainen et al., 2013) and whole genome sequencing data from 1000 Genomes Project [TGP (1000 Genomes Project Consortium et al., 2015)]. After extensive quality control to remove genes not expressed in LCLs, our data set includes 10,203 autosomal genes (see Supplemental Information). For each gene, we extracted all variants within 1Mb of the transcription start or end sites; we do not consider *trans*-effects because of the small sample size. To control for non-normality, population structure, and batch effects, we quantile normalize expression values and include the first 10 principal components from both the genetic and phenotypic data in all analyses (see Supplemental Information). We estimate heritability using H-E regression because we estimate a mean singleton effect size that is statistically indistinguishable from zero (see Supplemental Information). We focus on 20 MAF bins because this was the most robust approach across simulated scenarios (Figure 1 and discussion in Supplemental Information), and present average heritability across genes to characterize the genetic architecture of human gene regulation.

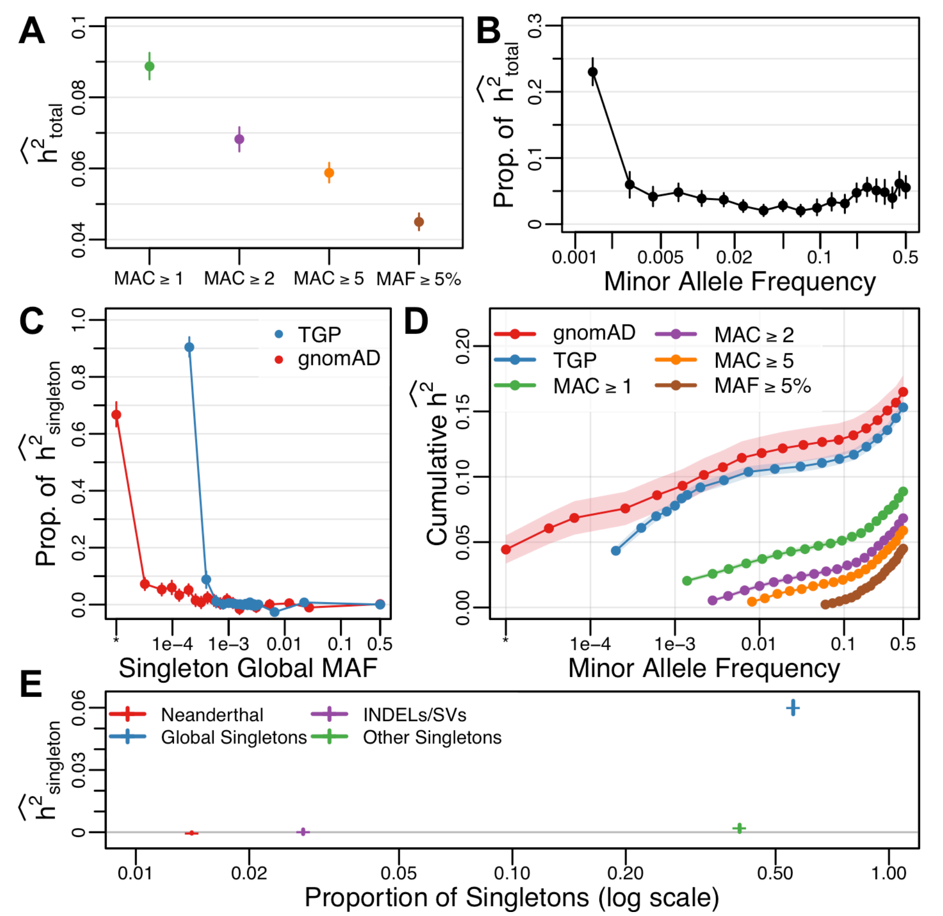

Figure 2. Characterizing the genetic architecture of human gene expression. (**A**) Average total heritability inferred across genes for different frequency filters (based on minor allele count, MAC, or minor allele frequency, MAF). Excluding low frequency variants (MAF<5%) results in substantial "missing heritability". (**B**) The proportion of heritability attributed to each MAF bin. Singletons represent >20% of the total inferred heritability. (**C**) Partitioning singletons in our data (*n*=360) by global MAF (based on TGP, *n*=2504, or gnomAD, *n*>15k) demonstrates that the vast majority of singleton heritability is due to sites that are globally very rare (i.e. singletons in all of TGP or not reported in gnomAD, indicated by *). Note that all bins exclusively contain singletons. (**D**) Cumulative heritability inferred as a function of observed MAF for different frequency filter thresholds (brown, orange, purple, green) and as a function of global MAF (based on gnomAD, red, or TGP, blue). Including all SNPs and partitioning by global MAF (instead of observed MAF) results in a substantially increased level of heritability. (**E**) Global (i.e. TGP-wide) singleton SNPs represent 56% of all singletons, but contribute the vast majority of singleton heritability. Structural variants (SVs) and alleles likely introgressed from Neanderthal contribute negligibly to average heritability across genes. In all panels, points represent mean across genes, with whiskers and envelops representing 99% quantile range for 10,000 bootstrap samples.

Early studies of heritability filtered out SNPs with MAF<5% prior to their analysis (Yang et al., 2010), and more recent studies only remove the rarest variants (Mancuso et al., 2016; Yang et al., 2015). We show that the process of removing any SNPs based on MAF has a direct impact on the estimate of heritability. In Figure 2A, we show the total heritability inferred for different minor allele count (MAC) thresholds (averaged over all genes). We find that by adding progressively rarer variants to the analysis, there is a monotonic increase in the inferred heritability. Indeed, including all variants down to singletons nearly doubles the total heritability inferred ($\widehat{h^2}_{total} = 0.089$) compared to the case when only common variants (MAF≥5%) are analyzed ($\widehat{h^2}_{common} = 0.045$). Most of the increased heritability derives from singletons, which alone contribute ~23%, dwarfing the contribution of all other frequency bins (Figure 2B).

However, not all singletons contribute equally to heritability, and finding the source of large-effect rare variants is of utmost importance (Li et al., 2017). Evolutionary modeling suggests that rare variants will only contribute a substantial amount to heritability when causal alleles are deleterious (Eyre-Walker, 2010; Lohmueller, 2014; Simons et al., 2017; Uricchio et al., 2016). Under such models, natural selection should restrain the frequency of large-effect alleles. We therefore hypothesized that the singletons that were contributing most to heritability would also be rare in much larger multi-ethnic cohorts, i.e. globally rare. We tested this hypothesis by partitioning our singletons into 20 bins based on their global allele frequencies observed across the entire worldwide sample of 2504 individuals in TGP, and using H-E regression to jointly infer the heritability contributed by each class of singletons. Strikingly, 90% of all singleton heritability is contributed by those alleles that are actually singletons across all 2504 samples in TGP (MAF<0.02%; Figure 2C). Pushing this result further, we partitioned our singletons based on the global frequency observed in >15,000 individuals with high coverage whole genome sequencing (WGS) in the gnomAD data set (Lek et al., 2016). We found that 31% of our singletons were not reported in gnomAD, but this subset of variants nonetheless explains 67% of our singleton heritability (indicated by * in Figure 2C). We confirm that the majority of this signal derives from true-positive singletons by analyzing a subset of 58 individuals with high coverage whole genome sequencing, and show that 88% of the singleton heritability derives from variants that validate (Supplemental Information). Previous work has shown that additionally partitioning common variants by LD resulted in minimal change after partitioning by MAF (Yang et al., 2015).

Figure 2D shows how heritability accumulates as a function of MAF for different filtering schemes (with coloring as in Figure 2A) as well as when we partition all alleles by global MAF (based on either all of TGP or gnomAD). Surprisingly, partitioning variants by global MAF nearly doubles the inferred total heritability compared to cohort MAF ($\widehat{h^2}_{total} = 0.165$ and 0.153 for gnomAD and TGP, respectively, versus $\widehat{h^2}_{total} = 0.089$ for GEUVADIS), and that a majority of heritability (52.1% and 50.9% for gnomAD and TGP, respectively) is due to globally rare variants (MAF<0.1%). We show analytically and with simulations that these results are consistent with a “singleton-LD” effect (see Supplemental Information), which previously has only been reported for common variants (Speed et al., 2012; Yang et al., 2015).

To investigate the ability of rare variants to capture heritability of common variants (and vice-versa), we refit H-E regression removing MAF bins from rarest to most common (and vice-versa). We found that while rare variants could capture some of the heritability of more common variants, common variants could not capture the heritability derived from singleton variants (see Supplemental Information). This suggests that rare variants have not been indirectly captured in any published heritability estimates through “synthetic association” tagging (Dickson et al., 2010).

Recent studies of gene expression variation in humans have suggested that one-quarter of Neanderthal-introgressed haplotypes have *cis*-regulatory effects (McCoy et al., 2017), and that expression outliers are enriched for having nearby rare structural variants (SVs) compared to non-outliers (Li et al., 2017). However, the overall contribution of these classes of variants to expression variation across genes remains unknown. We therefore performed H-E regression on four categories of singletons (Neanderthal-introgressed, indels/SVs, global singletons, and other singletons), and found that global singletons (i.e., singletons in our data that are also singletons across all 2504 samples in TGP) contribute the vast majority (97%) of singleton heritability (Figure 2E). Neanderthal-introgressed variants and indels/SVs represent 1.4% and 2.8% of all singleton variants, but their average contribution to heritability across genes is indistinguishable from 0.

Genotype quality does not drive our inference of heritability

We performed several analyses to examine possible confounding effects in these data. First, we ranked singletons by their reported genotype likelihood as reported for the individual carrying the singleton allele in TGP (1000 Genomes Project Consortium et al., 2015), and partitioned them into four equal groups (quartiles). We then ran H-E regression with these four groups of singletons (along with 10 PCs). Strikingly, we find that only those singletons with high SNP quality contribute positively to our inference of heritability (see Supplemental Information). Second, since both the DNA and RNA sequencing is based on lymphoblastoid cell lines, 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 TGP Strict Mask (1000 Genomes Project Consortium et al., 2015), 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 Supplemental Information). Finally, a subset of n=58 samples were sequenced at high coverage by Complete Genomics Inc (CGI) as part of the 1000 Genomes Project Consortium (2015). 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 CGI or gnomAD datasets. Running H-E regression on this subset of individuals shows that singleton heritability is predominantly driven by singletons that replicate in the CGI data but are not reported gnomAD (consistent with Figure 2), and that singletons that are absent from CGI (and are therefore more likely to be false-positives) contribute negligibly to heritability (2.6% versus -0.14%, respectively).

Purifying selection drives causal variants

We find that rare variants are a major source of heritability of gene expression patterns, which we hypothesized was due rampant purifying selection acting to restrain the frequencies of large-effect alleles. To test this hypothesis, we performed extensive simulations of human evolutionary history (Hernandez, 2008; Uricchio et al., 2015), and developed a novel model to infer the parameters of an evolutionary model for complex traits (see Supplemental Information). Our three-parameter phenotype model was previously described (Uricchio et al., 2016), and captures the pleiotropy of causal variation (through ρ), the scaling relationship between effect sizes and selection coefficients (through τ), and the overall strength of selection (which we capture with ϕ, a mixture parameter between strong and weak selection distributions, where ϕ=1 corresponds to strong selection). We inferred approximate posterior distributions for each of these parameters by rejection sampling (Tavaré et al., 1997), which compares a set of informative summary statistics from genetic data simulated under a model of European demography (Tennessen et al., 2012) and selection (Boyko et al., 2008; Torgerson et al., 2009) to the observed data (see Supplemental Information). Note that our inference procedure allows each parameter to vary across genes, but we only seek to infer the mean of ρ, τ, and ϕ across genes. We rigorously evaluated the performance of this inference procedure with simulations, and found that we can infer ρ and τ with fairly high accuracy, but ϕ (while broadly unbiased) is less informative (Figure 3A-C).

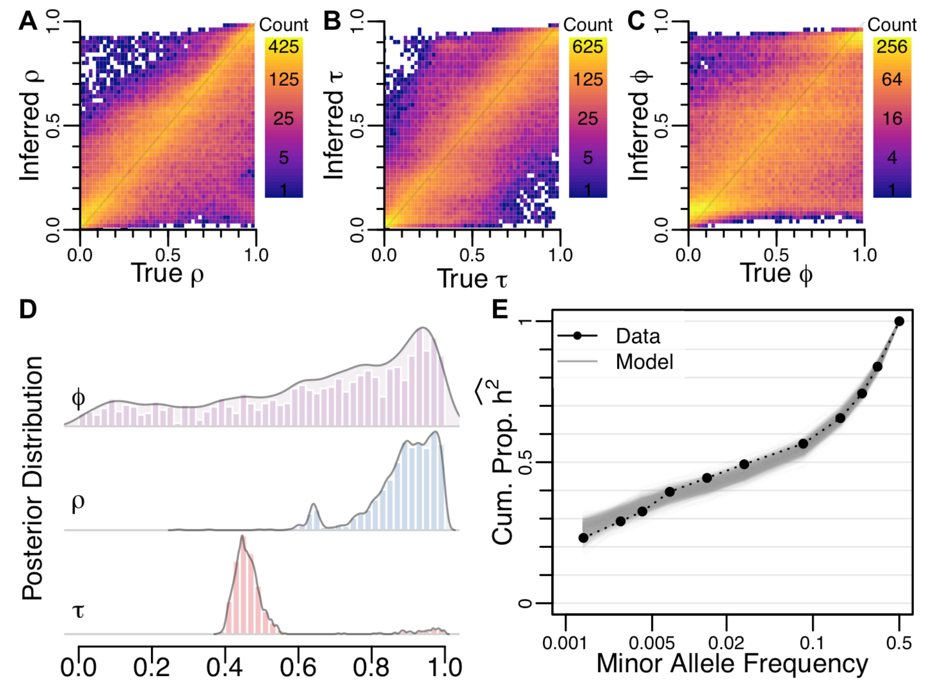

Figure 3. Inferring an evolutionary model for the genetic architecture of human gene regulation. (**A-C**) Simulation results showing the accuracy of our rejection algorithm for inferring each of our model parameters. (**D**) The posterior distributions for each model parameter from observed data, suggesting that the effect size that a causal variant has on gene expression correlates (inferred ρ≈0.9) with ~square root of the deleteriousness of its fitness effect in the population (inferred τ≈0.5). (**E**) The cumulative proportion of heritability inferred from the gene expression data (dots, dashed line) compared to the expected cumulative proportion based on 1000 draws from the posterior distribution (grey shades).

Applying this model to our data, we find that natural selection has had a major impact on the genetic architecture of human gene expression. In Figure 3D, we plot the posterior distributions of the mean values of ρ, τ, and ϕ, which suggest that the effect size of causal variants is highly correlated ($\hat{\rho} \approx 0.9$) with the square root of their selection coefficients ($\hat{\tau} \approx 0.5$), implying that larger causal effects tend to be more deleterious (Simons et al., 2017; Uricchio et al., 2017). Moreover, while a wide range of mixture coefficients (ϕ) are consistent with our observed data, much more probability mass is centered in the strong selection regime, suggesting that the selective pressure acting on most causal variants is likely to be just as strong as selection acting on nonsynonymous variants in coding regions. Consistent with this prediction, sites with increased evolutionary constraint exhibit higher heritability estimates (see Supplemental Information).

DISCUSSION

There is substantial 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. In this study, we take a different approach. We developed, tested, and applied a novel 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 and RNA sequencing performed on their derived lymphoblastoid cell lines. We estimate that the total narrow sense heritability of LCL gene expression is 15-16%, and that an average of nearly a quarter of all heritability of gene expression can be explained by the rarest of variants in our data: singletons, where just one copy of the allele has been observed in our sample of 720 chromosomes (MAF=0.0014). Globally rare variants (global MAF<0.02%) explain 85-90% of this singleton heritability, and we argue that the vast majority of heritability would be contributed singletons independent of what sample size is analyzed. Our estimate of total *cis*-heritability is larger than the previous estimates of h_{cis}^2=0.057 and h_{cis}^2=0.055 in blood and adipose respectively (Price et al., 2011), but lower than recent twin-based estimates of overall narrow-sense heritability h^2=0.26, 0.21, and 0.16 in adipose, LCLs, and skin respectively (Grundberg et al., 2012) as well overall broad-sense heritability H^2=0.38 and 0.32 for LCLs and whole blood (Powell et al., 2012). It is therefore possible that rare variants account for substantial "missing heritability" in human gene expression, but differences in population, tissue, and/or technology could also explain some of these patterns.

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 scaling with the square root of the strength of selection acting on them. This means that the alleles that have the greatest impact on gene expression are likely to be extremely rare in the broader population, and may be unlikely to exist in more than a few unrelated individuals across the world. This is consistent with a recent finding that a large fraction of individuals with outlier expression for a gene also tend to have a globally rare variant in the vicinity (Li et al., 2017). We push this result further to quantify the overall impact that rare variants have on gene expression across a population. Indeed, we find that globally rare variants are the predominant source of heritability for gene expression. Our analysis shows that 85-90% of the singleton heritability derives from alleles that are not carried by any of the other 2504 individuals in TGP (and are either not reported or have MAF<0.02% in the n>15,000 samples in gnomAD). We therefore conclude that identifying causal variation for transcriptional variation will likely require the incorporation of new biological information, possibly including large-scale experimental testing of singleton variants to improve functional predictions.

Our results suggest that one cannot capture the heritability of rare or low frequency alleles by analyzing additional common alleles. This implies that "synthetic associations" (Dickson et al., 2010; Wray et al., 2011) are uncommon for gene expression data. A broader consequence is that, when rare variants matter, approaches that rely on genotyping large samples followed by imputing missing genotypes from reference populations may not successfully reconstruct the true impact of rare variants (especially when the reference panel is smaller than the test sample). This is because both genotyping and imputation require the variant to be present at a reasonable frequency in the reference population, which is highly unlikely for strongly deleterious alleles (indeed, we found that 67% of our singleton heritability was attributable to variants not reported in gnomAD). Instead, WGS of large cohorts may be necessary (though the actual sample size required will depend on several factors that have not yet been elucidated).

As the number of samples with detailed phenotype data and WGS data increases, it will be possible to apply the approach we have developed here to characterize the genetic architecture of additional complex traits. 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 the hope of *a priori* risk prediction from genomic data for a most diseases will be feasible for an otherwise healthy individual with limited family history information. Population genetic theory tells us that rare variants will only be a significant source of heritability when causal alleles are evolutionarily deleterious. But the biology of human health and disease is complex. While not all human diseases will 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 (Boyle et al., 2017).

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 Supplemental Information). 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 (see Supplemental Information). We conservatively used quantile normalization on the expression phenotypes to enforce normality, and this often reduces the overall heritability estimates (see Supplemental Information) by diminishing the impact of outliers (Li et al., 2017). There are several other contributors to broad sense heritability that we have not attempted to model and may also account for some of the heritability estimated in family-based studies, such as gene-gene interactions, gene-environment interactions, and other non-additive components.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Methods and Procedures, 24 figures, and three tables.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

We thank Hyun Min Kang, Adam Auton, and Sasha Gusev for discussions about possible confounders that improved our analysis; Jeffrey Barret and Konrad Karczewski for peer review of our preprint; Raul Torres for assistance with data analysis; Jeffrey Wall for assistance with Neanderthal-introgressed alleles; and Aria Hernandez for discussions on figure colors. Research reported in this publication was supported by National Human Genome Research Institute of the National Institutes of Health under award number R01HG007644 to RDH and K25HL121295. LHU was supported by IRACDA NIGMS grant K12GM088033, KH was supported by a Gilliam Fellowship, and AD was supported by U01HG009080 and R01HG006399.

REFERENCES:

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet *12*, 745–755.

Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet *4*, e1000083.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet *47*, 291–295.

Chen, L., Liu, P., Evans, T.C., and Ettwiller, L.M. (2016). DNA damage is a major cause of sequencing errors, directly confounding variant identification. BioRxiv.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat Genet *48*, 1284–1287.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol *8*, e1000294.

Eyre-Walker, A. (2010). 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 U S A *107 Suppl 1*, 1752–1756.

Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature *536*, 41–47.

Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat Genet *46*, 881–885.

Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., et al. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat Genet *49*, 1421–1427.

Golan, D., Lander, E.S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. Proc Natl Acad Sci U S A *111*, E5272-81.

Grundberg, E., Small, K.S., Hedman, Å.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet *44*, 1084–1089.

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., et al. (2013). Quantifying missing heritability at known GWAS loci. PLoS Genet *9*, e1003993.

Hernandez, R.D. (2008). A flexible forward simulator for populations subject to selection and demography. Bioinformatics *24*, 2786–2787.

Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

Lohmueller, K.E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet *10*, e1004379.

Mancuso, N., Rohland, N., Rand, K.A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., et al. (2016). The contribution of rare variation to prostate cancer heritability. Nat Genet *48*, 30–35.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet *48*, 1279–1283.

McCoy, R.C., Wakefield, J., and Akey, J.M. (2017). Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell *168*, 916–927.e12.

Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E.T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet *7*, e1002144.

Powell, J.E., Henders, A.K., McRae, A.F., Wright, M.J., Martin, N.G., Dermitzakis, E.T., Montgomery, G.W., and Visscher, P.M. (2012). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res *22*, 456–466.

Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet *7*, e1001317.

Schoech, A., Jordan, D., Loh, P.-R., Gazal, S., O'Connor, L., Balick, D.J., Palamara, P.F., Finucane, H., Sunyaev, S.R., and Price, A.L. (2017). Quantification of frequency-dependent genetic architectures and action of negative selection in 25 UK Biobank traits. BioRxiv.

Simons, Y.B., Turchin, M.C., Pritchard, J.K., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. Nat Genet *46*, 220–224.

Simons, Y.B., Bullaughey, K., Hudson, R.R., and Sella, G. (2017). A model for the genetic architecture of quantitative traits under stabilizing selection.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am J Hum Genet *91*, 1011–1021.

Speed, D., Cai, N., UCLEB Consortium, Johnson, M.R., Nejentsev, S., and Balding, D.J. (2017). Reevaluation of SNP heritability in complex human traits. Nat Genet *49*, 986–992.

Tavaré, S., Balding, D.J., Griffiths, R.C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. Genetics *145*, 505–518.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Torgerson, D.G., Boyko, A.R., Hernandez, R.D., Indap, A., Hu, X., White, T.J., Sninsky, J.J., Cargill, M., Adams, M.D., Bustamante, C.D., et al. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet *5*, e1000592.

Uricchio, L.H., Torres, R., Witte, J.S., and Hernandez, R.D. (2015). Population genetic simulations of complex phenotypes with implications for rare variant association tests. Genet Epidemiol *39*, 35–44.

Uricchio, L.H., Zaitlen, N.A., Ye, C.J., Witte, J.S., and Hernandez, R.D. (2016). Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res *26*, 863–873.

Uricchio, L.H., Kitano, H.C., Gusev, A., and Zaitlen, N.A. (2017). Evidence for evolutionary shifts in the fitness landscape of human complex traits. BioRxiv.

Wray, N.R., Purcell, S.M., and Visscher, P.M. (2011). Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol *9*, e1000579.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet *42*, 565–569.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat Genet *46*, 100–106.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet *47*, 1114–1120.

Zeng, J., de Vlaming, R., Wu, Y., Robinson, M., Lloyd-Jones, L., Yengo, L., Yap, C., Xue, A., Sidorenko, J., McRae, A., et al. (2017). Widespread signatures of negative selection in the genetic architecture of human complex traits. BioRxiv.

Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T.J., Lee, C.M., Banskota, S., Marigorta, U.M., Bao, G., and Gibson, G. (2016). A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. Am J Hum Genet *98*, 299–309.