HUMAN GENETICS

Recombination affects allele-specific expression of deleterious variants in human populations

Michelle P. Harwood[1,2], Isabel Alves[3], Hilary Edgington[4], Mawusse Agbessi[1], Vanessa Bruat[1], David Soave[1,5], Fabien C. Lamaze[1,6], Marie-Julie Favé[1], Philip Awadalla[1,2,7]*

How the genetic composition of a population changes through stochastic processes, such as genetic drift, in combination with deterministic processes, such as selection, is critical to understanding how phenotypes vary in space and time. Here, we show how evolutionary forces affecting selection, including recombination and effective population size, drive genomic patterns of allele-specific expression (ASE). Integrating tissue-specific genotypic and transcriptomic data from 1500 individuals from two different cohorts, we demonstrate that ASE is less often observed in regions of low recombination, and loci in high or normal recombination regions are more efficient at using ASE to underexpress harmful mutations. By tracking genetic ancestry, we discriminate between ASE variability due to past demographic effects, including subsequent bottlenecks, versus local environment. We observe that ASE is not randomly distributed along the genome and that population parameters influencing the efficacy of natural selection alter ASE levels genome wide.

INTRODUCTION

The consequences of mutations on individual health are influenced by the regulation of gene expression, causing individuals with the same genotype to display highly variable phenotypes. Allele-specific approaches are used to investigate gene regulation by detecting differences in expression of each allele at a particular locus (*1*). While previous studies have investigated the specific genomic location, tissue specificity, and disease phenotype impact of allele-specific expression (ASE) (*2*–*7*), few have documented how evolutionary forces, such as natural selection and genetic drift, regulate gene expression evolution genome wide (*8*, *9*). In addition to evolutionary forces and population demographics, variability in gene expression explained by genic interactions with the environment has been limited to only a few instances where external environmental exposures have been measured (*10*–*14*).

Natural selection acts on phenotypes, causing underlying genotype and allele frequencies to shift over time; however, the strength of selection in a population is affected by many factors, including effective population size (N_e) (*15*, *16*). In general, purifying selection depletes haplotypes that are likely pathogenic (*15*–*17*). Castel *et al.* (*18*) observed that the overexpression of haplotypes with disease-associated variants was enriched in cohorts of patients with autism and cancer. Despite evidence of selection acting on cis-regulation (*18*, *19*), little is known about the degree to which population and genomic parameters influence selection efficiency in relation to ASE. Selection efficiency varies across the genome based on local recombination rates (*20*–*24*). At regions with low levels of recombination, genetic variation and effective population size are greatly reduced, making the population less able to respond to selection and eliminate deleterious mutations (*20*, *23*). Without recombination, deleterious mutations accumulate irreversibly, where a mutation-free haplotype cannot be regenerated once lost—a process termed as "Muller's ratchet" (*25*). In addition, functional diversity is affected when mutations compete with each other to become fixed in a population, known as Hill-Robertson (HR) interference (*23*, *26*). We have previously shown that recombination heterogeneity can have profound consequences on how genomic variation is structured across genomes, as selection appears more efficient at removing harmful mutations in recombination hotspots and elicit an enrichment of damaging mutations in cold spots (CSs) (*15*). Consistent with this model is the observation that, in larger populations, selection is more effective at removing damaging mutations in recombination hotspots (*15*, *27*–*30*). Here, we demonstrate how recombination and N_e collectively affect the expression of deleterious variants, specifically through the regulation of ASE. To improve our understanding of the role of evolution on gene expression variability, we investigated how selection efficiency in different demographic populations and environments alters levels of ASE genome wide. We demonstrate that selection efficiency, through recombination and population histories influencing N_e, contributes to ASE evolution and maintenance by reducing the functional effects of deleterious variants.

RESULTS

ASE controls the expression of deleterious mutations

We confirm that ASE within individuals can diminish the relative expression of pathogenic mutations at heterozygous loci (*18*). We tested whether derived alleles that accumulated in human genomes were under stronger ASE, reflecting negative selection compared to ancestral alleles. We quantified ASE by combining RNA sequencing and genotyping of 844 individuals from three regions in Quebec of multiple ancestries from the CARTaGENE cohort (*12*, *13*, *31*) and 752 individuals from the Genotype-Tissue Expression (GTEx) project (*32*). We quantified ASE at missense and synonymous single-nucleotide polymorphisms (SNPs), which included 3083 SNPs from CARTaGENE whole blood and 5974, 5738, 4431, 6795, 9035, and 5091 from whole blood, muscle, brain, ovarian, lung, and liver tissue

[1]Ontario Institute for Cancer Research, Toronto, ON, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [3]Université de Nantes, CHU Nantes, CNRS, INSERM, L'Institut du thorax, F-44000 Nantes, France. [4]Department of Biology, College of Wooster, Wooster, OH, USA. [5]Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada. [6]Institut universitaire de cardiologie et de pneumologie de Québec, Université Laval, Québec, QC, Canada. [7]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada.
*Corresponding author. Email: philip.awadalla@oicr.on.ca

from GTEx, respectively (Materials and Methods and Fig. 1) (*32*, *33*). In CARTaGENE, we identified 23,589 (4.9%) individual-SNP pairs with significant ASE, as determined by binomial tests after correcting for multiple testing and mapping bias using WASP software (Materials and Methods and table S1) (*34*). We investigated the magnitude of ASE relative to the derived allele (Materials and Methods). Among the 23,589 individual-SNP pairs with significant ASE, 12,133 pairs (51.4%) underexpress the derived allele, which were relatively consistent across populations and tissues (tables S1 and S2). Near-equal proportions of directional ASE also aligned with previous observations in humans (*35*).

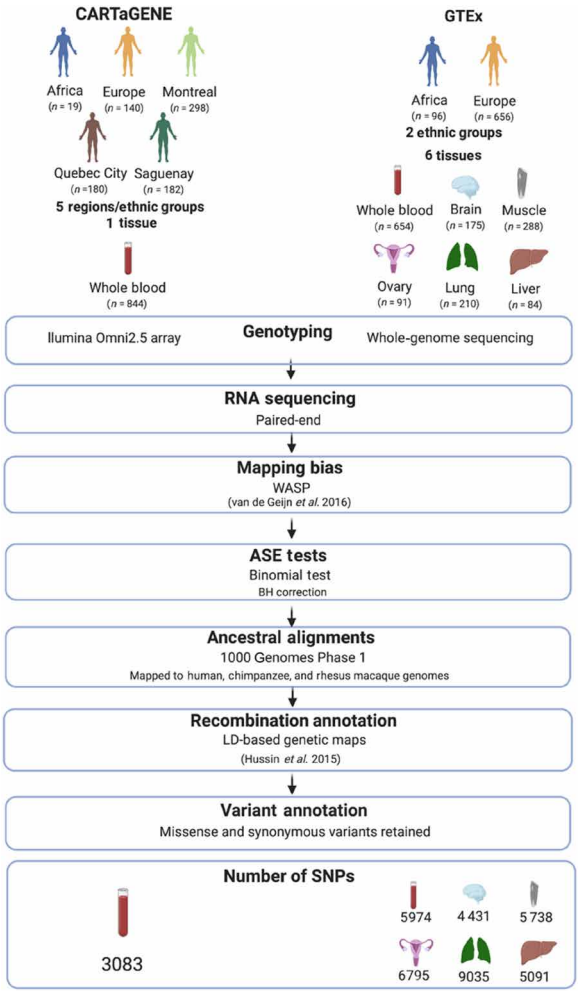


Fig. 1. The ASE framework: Investigating genome-wide ASE across two ethnic groups, three French-Canadian regions, and six tissues. A total of 844 individuals were genotyped and RNA sequenced from the whole blood from individuals in Quebec in the CARTaGENE cohort, with ancestry from Africa ($n = 19$), Europe ($n = 140$), Quebec City ($n = 180$), Montreal ($n = 298$), and Saguenay ($n = 182$) (*12*, *13*). We also use the whole blood ($n = 654$), muscle ($n = 288$), brain ($n = 175$), ovarian ($n = 91$), lung ($n = 210$), and liver ($n = 84$) tissue from GTEx v8 release (*32*, *33*). The ASE pipeline for the two cohorts differs in their methods for genotyping, but the downstream approaches are consistent (Materials and Methods). ASE is tested using binomial tests with Benjamini-Hochberg (BH) multiple testing correction, and ancestral alignments from 1000 Genomes Phase 1 are used to identify the ancestral and derived alleles (Materials and Methods). Recombination annotations are from population-specific linkage disequilibrium (LD) genetic maps by Hussin *et al.* (*15*). Missense and synonymous variants were retained for both cohorts (Materials and Methods), resulting in 3083 SNPs used from CARTaGENE and 5974, 4431, 5738, 6795, 9035, and 5091 used from GTEx blood, brain, muscle, ovarian, lung, and liver tissue, respectively.

We confirmed that deleterious mutations tend to underexpress the derived allele in general human populations, as demonstrated by Castel *et al.* (fig. S1) (*18*). We tested whether putatively deleterious mutations were more commonly over- or underexpressed by scoring variants using the Combined Annotation Dependent Depletion (CADD) score (*36*). The CADD score estimates the level of deleteriousness of an SNP using a combination of functional predictions and conservation measurements (*36*). Using a binomial mixed-effects regression treating individuals as a random effect (Materials and Methods), we observe that sites underexpressing the derived allele have larger CADD scores compared to sites overexpressing the derived allele ($\beta = -4.190 \times 10^{-3}$, $P < 2 \times 10^{-16}$; fig. S1) despite having comparable proportions of synonymous and missense variants (fig. S2). These findings suggest that ASE can reduce the exposure of putatively deleterious variants to natural selection by reducing their allelic expression.

High recombination regions are more efficient at regulating the expression of deleterious mutations

Next, we exposed how natural selection acts on ASE by testing how different evolutionary forces that contribute to selection efficiency—specifically recombination, per-site expression level, and population histories—are associated with ASE directionality. Regions of the genome with higher recombination rates have more efficient selection, because recombination both increases local N_e (*20*) and reduces HR interference (*23*). SNPs were classified as either being located in low recombination, referred to as CS, normal recombination, or high recombination region (HRR) using population-specific linkage disequilibrium-based genetic maps (*15*, *37*). We grouped normal recombination with HRR in several downstream analyses (referred to as HRR/Normal) due to the relatively low number of segregating sites in HRR (6.7%) and the similar direction of odds ratio and β estimates observed between normal and HRR (fig. S3). We modeled recombination class as a function of derived allele expression (Materials and Methods) and identified that SNPs with statistical evidence of ASE are more likely to be observed in HRR/Normal compared to CS (Fig. 2A). Recombination rates in human genomes are known to correlate with genomic features such as guanine-cytosine (GC) content, average exon expression, and exon size (*15*, *38*, *39*); however, it is unlikely that these genomic features explain the enrichment of ASE in HRR/Normal regions, as (i) models stratified by low, medium, and high values of these features had comparable odds ratios to original estimates (table S3); (ii) odds ratio estimates were comparable with models including these genomic features (table S4); and (iii) models tested with the various genomic features alone were significantly improved when adding recombination (table S5). Enrichment of ASE in HRR/Normal regions was observed across the whole blood, brain, muscle, ovarian, lung, and liver tissue (Fig. 2A). These results suggest that ASE is not randomly distributed along the genome and that varying recombination rates and its impact on the efficiency of selection contributes to genomic variation in ASE evolution.

The effect of ASE and recombination rate was observed for both over- and underexpression of the derived allele (Fig. 2A); however, on the basis of the results from Castel *et al.* (*18*), we would expect to have a stronger effect of underexpression of the derived allele in

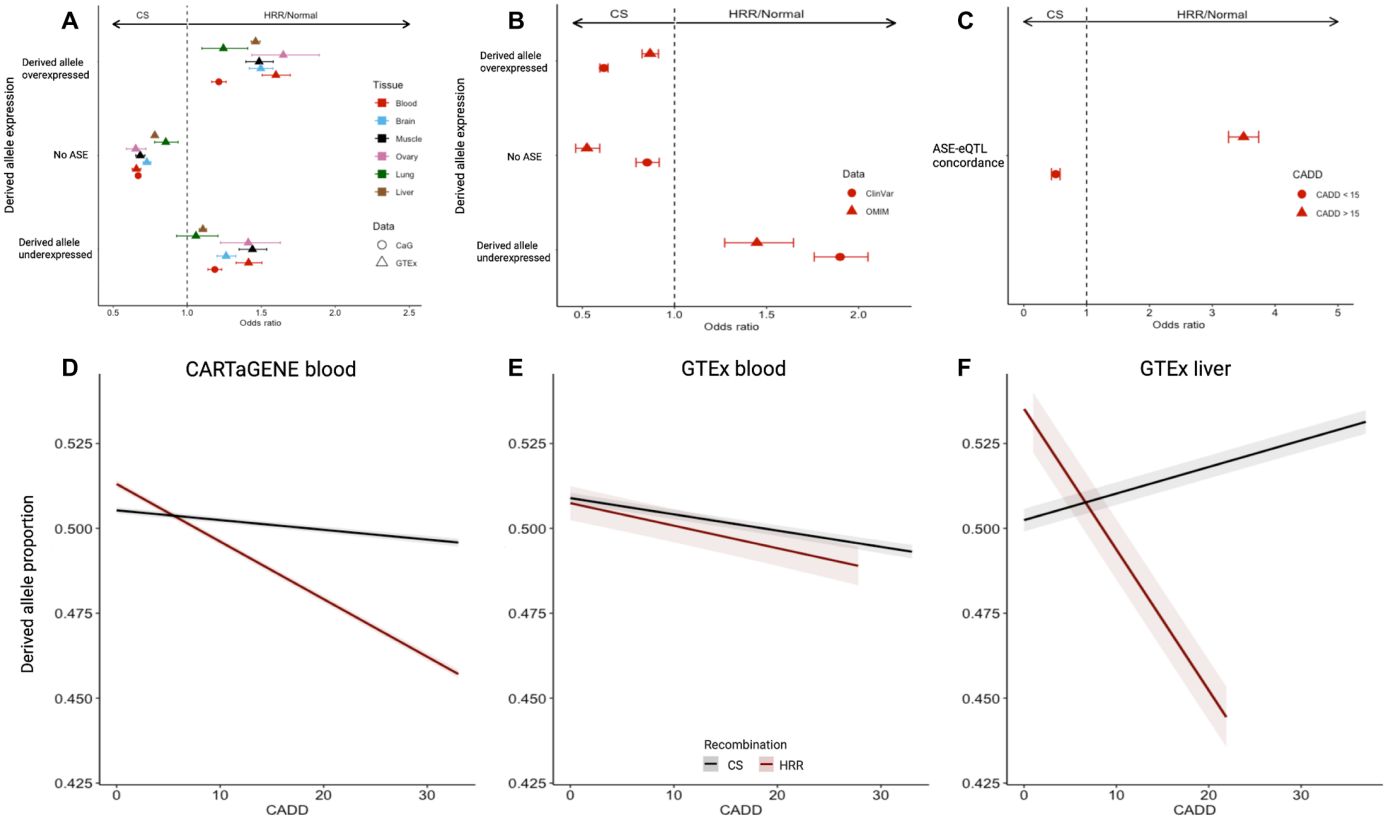


Fig. 2. ASE is enriched in regions of high recombination, regulating the expression of deleterious and disease associated variants. (**A**) We modeled recombination class as a function of derived allele expression for CARTaGENE whole blood and GTEx whole blood, brain, muscle, ovarian, lung, and liver tissue (Materials and Methods). Odds ratio (OR) > 1 means increased odds that variants exhibit the ASE behavior in HRR/Normal compared to CS. Error bars represent 95% confidence intervals. (**B**) Equivalent models of recombination class as a function of derived allele expression for variants found in ClinVar and OMIM databases. (**C**) ORs modeling recombination class as a function of ASE-eQTL concordance, stratified by CADD score, where CADD > 15 represents putatively deleterious variants. ASE-eQTL concordance is defined as haplotype combinations, where a cis-regulatory variant underexpresses the ASE SNP (Materials and Methods). OR > 1 means haplotype combinations with ASE-eQTL concordance are more likely to be in HRR/Normal compared to CS. Error bars represent 95% confidence intervals.(**D**) We modeled the proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed-effects regression (Materials and Methods), stratified by HRR ($\beta = -0.049725$, $P < 2 \times 10^{-16}$) and CS ($\beta = -0.001245$, $P < 2 \times 10^{-16}$), with 95% confidence bands. Equivalent models of the proportion of the derived allele as a function of CADD score on (**E**) GTEx whole blood, stratified by HRR ($\beta = -0.0026557$, $P < 2 \times 10^{-16}$) and CS ($\beta = -0.0019145$, $P < 2 \times 10^{-16}$) and (**F**) GTEx liver, stratified by HRR ($\beta = -0.0166216$, $P < 2 \times 10^{-16}$) and CS ($\beta = -3.13 \times 10^{-3}$, $P < 2 \times 10^{-16}$). Equivalent models for the brain, ovarian, lung, and muscle tissue are found in fig S4.

disease-associated sites. Therefore, we filtered the data to only include variants with putative disease associations in ClinVar (*40*) and Online Mendelian Inheritance in Man (OMIM) (*41*). Here, we found that variants with putative disease associations had significant odds of underexpressing the derived allele in HRR/Normal regions; however, sites overexpressing the derived allele were more likely to be observed in CS compared to HRR/Normal (Fig. 2B).

We next tested the hypothesis that protective haplotype combinations would be observed at higher frequencies in HRR/Normal regions because of more efficient selection removing haplotype combinations that overexpress deleterious alleles in these regions (*18*). To test this hypothesis, we identified ASE SNPs that are in genes associated with expression quantitative trait loci (eQTLs) (Materials and Methods). Using haplotype phasing information, we identified ASE-eQTL concordant sites, where an eQTL cis-regulatory variant that decreases expression is on the same haplotype as the derived SNP under ASE and therefore underexpressing the derived allele at the ASE site (Materials and Methods). We observed that concordant haplotypes that underexpress deleterious variants (CADD > 15) are more likely to be observed in HRR/Normal regions compared to CS (Fig. 2C). This result supports the hypothesis that protective haplotype combinations of ASE underexpressing a deleterious variant is more common in HRR/Normal regions, likely due to the higher efficiency of selection acting on both the cis-regulatory variant and the ASE variant.

We modeled the proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed-effects regression and tested for interaction between CADD score and recombination region (Materials and Methods). We observe that HRRs underexpress putatively deleterious mutations more efficiently ($\beta = -0.049725$) compared to CS ($\beta = -0.001245$, $P < 2 \times 10^{-16}$) (Fig. 2D). This finding of HRR underexpressing deleterious mutations more efficiently was replicated in the whole blood, liver, (Fig. 2, E and F), lung, and muscle tissue from GTEx (fig. S4, C and D) but not observed in the brain and ovarian tissue (fig. S4, A and B). Together, these results suggest that regions of high recombination are more efficient at minimizing the expression of transcripts harboring deleterious and/or disease-associated variants.

Total expression level contributes to ASE variation at sites with deleterious mutations

Next, we found that per-site total expression is positively associated with ASE, potentially because genes with higher expression are under stronger evolutionary pressures (*42*). Mean expression can influence the efficiency of selection, because evolvability by natural selection is associated with increased variance, as described in Fisher's fundamental theorem (*43*). We tested whether the previous result of HRR efficiently minimizing the expression of deleterious variants was consistent across SNPs with different per-site total expression levels (Materials and Methods). We observed an increased odds of exhibiting ASE in HRR/Normal compared to CS for SNPs in the high-expression bin (above third quartile) compared to medium (between first and third quartiles) and low-expression (below first quartile) bins (Fig. 3A). Variants with lower per-site total expression do not efficiently underexpress deleterious mutations, regardless of recombination region (HRR: $\beta = 0.008657$, $P = 1.33 \times 10^{-12}$; CS: $\beta = 0.0008008$, $P = 0.008466$); however, sites with higher per-site expression more efficiently underexpress the derived allele at highly deleterious mutations in HRR ($\beta = -0.0509201$, $P < 2.2 \times 10^{-16}$) compared to CS ($\beta = -1.71 \times 10^{-3}$, $P < 2 \times 10^{-16}$) (Fig. 3B). Since it is possible that the difference in expression observed is related to read coverage, we used a sampling approach to randomly sample reads, with replacement, from each position, such that all positions have 250 reads (Materials and Methods). We fit the mixed-effects regression models, stratified by expression, on the randomly sampled reads for 25 iterations and took the mean β, intercept, and 95% confidence intervals from all iterations (Materials and Methods). This random sampling approach observed the same relationship of HRR more efficiently underexpressing the derived allele at deleterious variants for sites with high expression, supporting the notion that the observed relationship is due to total expression and likely not an artifact of read coverage (fig. S5). This relationship with HRR more efficiently underexpressing highly deleterious mutations at sites with high expression was replicated in the blood, muscle, brain, lung, and liver tissue, stratified by per-site total expression (fig. S7). These results suggest that sites with higher expression and recombination regulate ASE more efficiently, likely because selection is more efficient at these sites, resulting in the

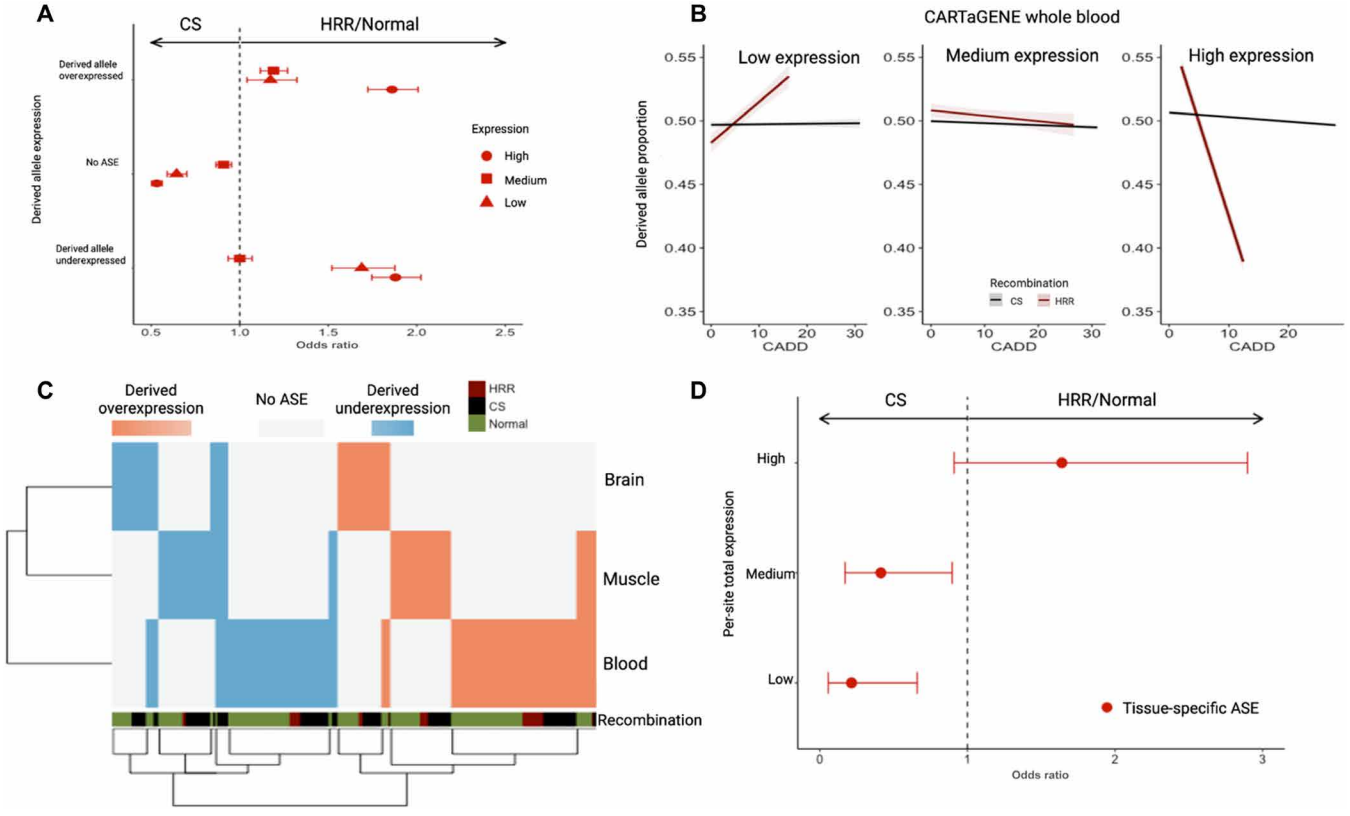


Fig. 3. Per-site total expression contributes to ASE variability and tissue specificity. (**A**) We modeled recombination class as a function of derived allele expression, stratified by total expression (Materials and Methods). "High" expression categorizes SNPs above the third quartile, and "low" expression categorizes below the first quartile. OR > 1 indicates that variants in an expression class have increased odds of exhibiting ASE in HRR/Normal compared to CS. Error bars represent 95% confidence intervals. Data used from CARTaGENE whole blood. Equivalent models for GTEx whole blood, brain, muscle, ovarian, lung, and liver tissue are found in fig S6. (**B**) We modeled the proportion of the derived allele expression as a function of CADD score, stratified by expression for HRR (high: $\beta = -0.0509201$, $P < 2.2 \times 10^{-16}$; medium: $\beta = -5.82 \times 10^{-6}$, $P = 0.993$; and low: $\beta = 0.008657$, $P = 1.33 \times 10^{-12}$) and CS (high: $\beta = -1.71 \times 10^{-3}$, $P < 2 \times 10^{-16}$; medium: $\beta = -8.99 \times 10^{-5}$, $P = 0.492$; and low: $\beta = 0.0008008$, $P = 0.008466$). Data used from CARTaGENE whole blood. Equivalent models for GTEx blood, brain, muscle, ovarian, lung, and liver tissue are found in fig S7. (**C**) We identified 79 individuals with measurements in the blood, muscle, and brain tissue at 448 sites with significant ASE in at least one tissue. Heatmap demonstrates each individual-SNP pair in columns and ASE levels for each tissue in rows. (**D**) Using the same individual-SNP pairs from (C), ORs were calculated after modeling recombination class as a function of the total expression for tissue-specific ASE (Materials and Methods). OR > 1 indicates that variants in an expression class have increased odds of having tissue-specific ASE in HRR/Normal compared to CS. Error bars represent 95% confidence intervals.

removal of haplotype combinations that overexpress deleterious mutations.

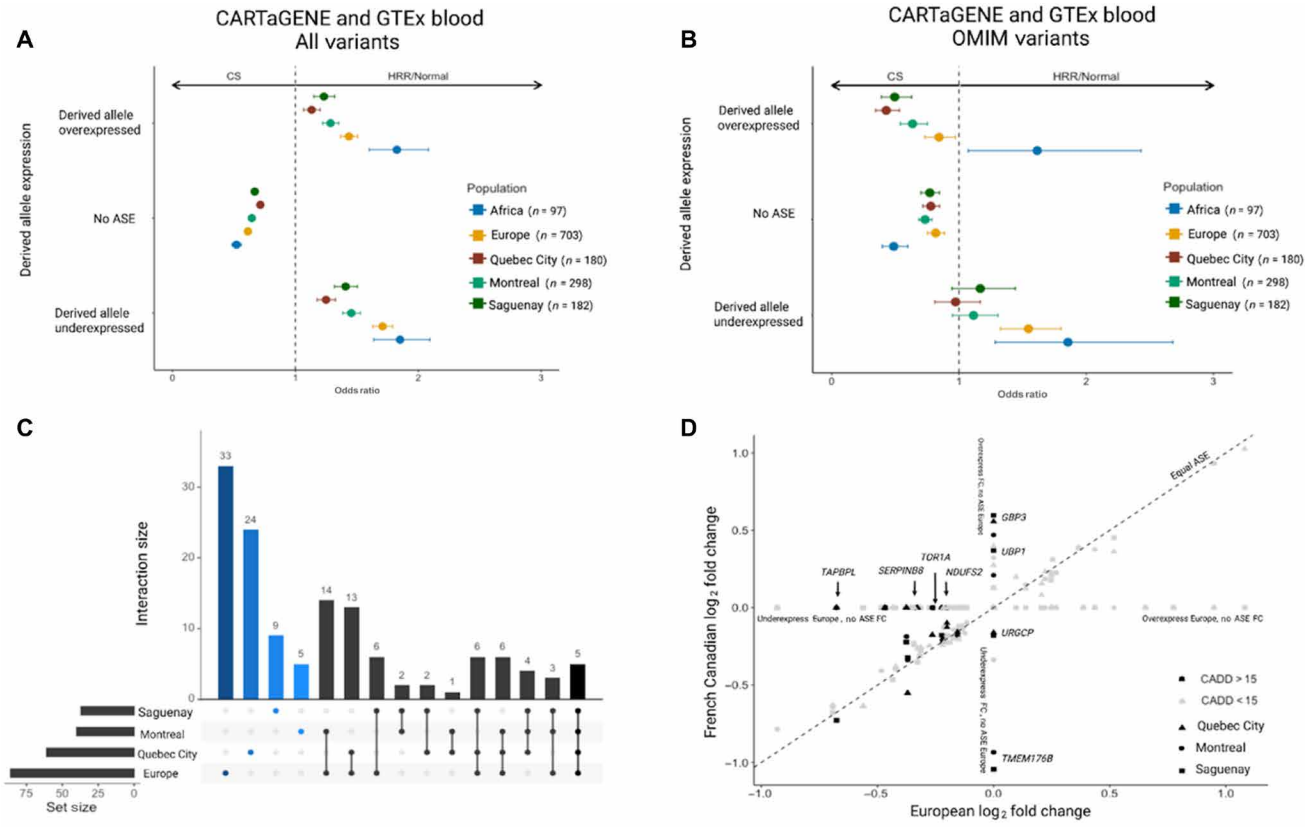
Since tissues vary in expression level, we next tested whether expression and recombination contributes to tissue-specific ASE. To test this, we identified 79 individuals with ASE calculations in the blood, muscle, and brain tissues at 448 sites with ASE in at least one tissue (Materials and Methods). We confirm that ASE is highly variable across tissues within the same individual (Fig. 3C) (*18*). Variable tissue-specific ASE did not appear to be directly linked to recombination alone (Fig. 3C); however, sites that had tissue-specific ASE and were highly expressed had increased odds of being in HRR/Normal recombination compared to CS, which was not observed for medium and lowly expressed sites (Fig. 3D). This result suggests that sites under stronger selective pressures, specifically higher expression and recombination, are more likely to exhibit tissue-specific ASE.

Populations with smaller N_e have weaker enrichment of ASE in recombination regions

As recombination influences N_e, we next investigated whether ancestral populations with smaller N_e had a reduced enrichment of ASE in HRR/Normal. Higher recombination rates and larger N_e together increase the number of haplotype combinations, making natural selection more efficient (*15*). We made use of the data from historically founding populations and contrasted other populations with smaller N_e to explore the relationship of N_e in association with ASE variation (*44*). The French-Canadian population originates from a founder event 400 years ago from the French population, and smaller subpopulations were established in remote areas through subsequent founder events, such as in the Saguenay–Lac-Saint-Jean region (*44*). The Saguenay region remained relatively isolated for many decades and now shows higher levels of relatedness and homozygosity (*17*, *44*). In addition, we examined samples of African and European descent in the analysis to reveal the impact of N_e on ASE profiles. Larger and older populations, such as Africa (*45*), have more efficient selection and reduced genetic drift, whereas founder populations with small N_e, such as the three Quebec subpopulations, have less efficient selection and stronger effects of genetic drift (*15*, *46*, *47*). Using data from CARTaGENE and GTEx blood, we found increased odds of significant ASE in HRR/Normal regions compared to CS, which showed greater estimates in populations with larger N_e, such as Africa (N_e = 13,900) (*45*, *48*–*50*), compared to Europe (N_e = 10,427), Montreal (N_e = 11,401), Quebec City (N_e = 11,202), and Saguenay (N_e = 9123), which have smaller N_e (Fig. 4A) (*15*, *17*). The three subpopulations of Quebec have overlapping confidence intervals for N_e estimates, and hence, the differences that we observed are among continents, as we do not have the resolution to observe differences within subpopulations studied here. To test the impact of sample size, we subsampled the CARTaGENE data, such that each population had equal sample size (n = 19) and observed similar signatures of African individuals having increased odds in HRR/Normal for overexpressing the derived allele (fig. S8). The signature of African individuals having increased odds of ASE in HRR/Normal compared to CS was also demonstrated in GTEx muscle, brain, ovarian, lung, and liver tissue (fig. S9). We tested individuals from East Asia from GTEx (n = 9) and CARTaGENE (n = 2) and observed consistent results of significant ASE in HRR/Normal regions with odds ratio estimates between European and African populations, which is consistent with N_e expectations, although the large confidence intervals are very likely due to small sample sizes (fig. S10). In addition, when only including pathogenic variants in OMIM, the observed pattern of increased odds of ASE in HRR/Normal compared to CS was specifically captured for sites that underexpress the derived allele, although population differences consistent with expected difference in N_e among populations were still observed (Fig. 4B). The signature of increased odds of overexpressing the derived allele for pathogenic variants in CS was observed for all populations except for Africa, which may be due to an overall larger effective population size in Africa genome wide (Fig. 4B). These findings additionally support the notion that recombination can influence the evolution of gene expression regulation by affecting N_e at genomic regions in some populations and implicate variation in demographic histories at a global scale of a population contributing to differences in ASE.

Next, we explored population-specific ASE patterns using differential expression (DE) analyses. DE analyses were run separately on each ancestral region, comparing the derived and ancestral allele counts at each gene (Materials and Methods). Significant DE suggests population-wide ASE at the gene, allowing us to compare differential ASE genes across four of the geographic regions. Africa was removed from the analyses due to a small sample size, with concerns of observing large quantities of false negatives and incorrectly identifying population-specific ASE (Materials and Methods). We investigated 1045 genes that had expression data across all four regions, 133 of which demonstrated significant DE [false discovery rate (FDR) < 0.05] and an absolute $\log_2$ fold change above the threshold of 0.25 for at least one population. We detected 86, 61, 40, and 37 significant ASE genes in Europe, Quebec City, Montreal, and Saguenay, respectively (Fig. 4C). Of those significant ASE genes, 33, 24, 9, and 5 of the genes demonstrate population-specific ASE in Europe, Quebec City, Saguenay, and Montreal, respectively (Fig. 4C).

To compare DE genes between populations, we analyze differences in significance and effect size ($\log_2$ fold change values) between regions. Positive $\log_2$ fold change values demonstrate overexpression of the derived allele, negative values demonstrate underexpression, and a value of 0 was reassigned for genes that were not significant. We observed that the European population had more variation in ASE effect size compared to the three French-Canadian regions, as there are more genes that are not significant in the French-Canadian regions but are significant in Europe (Fig. 4D). Gene ontology analysis did not demonstrate significant over- or underrepresentation of a particular gene category for genes that are significant for ASE in European individuals but not French-Canadians. To investigate genes with differential ASE across populations, we annotated the genes based on CADD scores and explored genes with putatively deleterious variants (CADD > 15), including *NDUFS2*, *SERPINB8*, *TOR1A*, *TMEM176B*, *URGCP*, *GBP3*, *TAPBPL*, and *UBP1* (Fig. 4D). *NDUFS2* encodes a protein that is a core subunit of the mitochondrial membrane respiratory chain, *SERPINB8* belongs to a family of serine protease inhibitors, *TOR1A* is in a family of adenosine triphosphatases, *TMEM176B* likely plays a role in maturation of dendritic cells, *URGCP* may promote hepatocellular growth and survival, *GBP3* encodes a member of the guanylate-binding protein (GBP) family, *TAPBPL* is a member of the immunoglobulin superfamily that links major histocompatibility complex class I to the transporter for antigen processing, and *UBP1* encodes a transcriptional activator in a promoter context-dependent manner (*51*). Some genes with differential ASE across populations were also located at genomic regions with differing estimates of recombination rate across populations (Fig. 4E). For example, genes *SAMM50* and *TRAF3IP3* both demonstrate significant



E

Gene	CADD	ASE Europe	ASE Quebec	ASE Montreal	ASE Saguenay	Recombination EU	Recombination FC
NDUFS2	25.4	Under ($P = 1.33 \times 10^{-12}$)	Under ($P = 6.8 \times 10^{-5}$)	None ($P = 0.99$)	None ($P = 0.99$)	CS	Normal
SERPINB8	23.4	Under ($P = 0.02$)	Under ($P = 0.019$)	None ($P = 0.99$)	None ($P = 0.99$)	HRR	Normal
TOR1A	24.8	Under ($P = 0.0003$)	None ($P = 0.99$)	None ($P = 0.99$)	Under ($P = 0.0014$)	Normal	CS
TMEM176B	21.9	None ($P = 0.37$)	None ($P = 0.28$)	Under ($P = 0.029$)	Under ($P = 0.038$)	Normal	HRR
URGCP	21.9	None ($P = 0.26$)	Under ($P = 0.017$)	Under ($P = 0.0081$)	Under ($P = 0.024$)	Normal	CS
GBP3	20.5	None ($P = 0.22$)	Over ($P = 2.7 \times 10^{-4}$)	Over ($P = 3.8 \times 10^{-6}$)	Over (4.8×10^{-4})	CS	CS
UBP1	17.3	None ($P = 0.20$)	None ($P = 0.99$)	Over ($P = 0.99$)	Over ($P = 0.033$)	CS	CS
TAPBPL	17.2	Under ($P = 1.2 \times 10^{-35}$)	None ($P = 0.99$)	None ($P = 0.99$)	Under ($P = 3.5 \times 10^{-55}$)	Normal	CS

Fig. 4. Larger effective populations are more efficient at regulating ASE. (**A**) We modeled recombination class as a function of the derived allele expression, stratified by ancestry (Materials and Methods). OR > 1 indicates increased odds of exhibiting ASE in HRR/Normal compared to CS. Error bars represent 95% confidence intervals. Data are from CARTaGENE and GTEx blood. Equivalent models for GTEx brain, ovarian, lung, and liver are found in fig. S9. (**B**) Equivalent model of recombination class as a function of derived allele expression, stratified by population for OMIM variants. (**C**) Differential expression (DE) analyses were computed to determine population-varying ASE sites at 133 genes (Materials and Methods). Upset plot demonstrates the number of genes that vary in significance (FDR < 0.05) across populations, indicated by the connecting dots. (**D**) Population-level ASE was compared between Europe and French-Canadian subpopulations using the $\log_2$ fold change values from the DE analyses. Each dot represents a gene. The $\log_2$ fold change values from European individuals are on the *x* axis, and $\log_2$ fold change values from one of the three French-Canadian populations are on the *y* axis. Positive $\log_2$ fold change demonstrates overexpression of the derived allele, negative values demonstrate underexpression, and 0 was assigned for no significance. Dashed line shows 1:1 line where the $\log_2$ fold change values of the populations would be theoretically equal. Genes were scored on the basis of CADD scores at ASE sites, where CADD > 15 signifies putatively deleterious variants. (**E**) Comparison of genes with CADD > 15 that demonstrated different ASE between European (EU) and French-Canadians (FC) based on DE analysis.

ASE in the European population but have no statistical evidence of ASE in the three French-Canadian populations. *SAMM50* and *TRAF3IP3* are in a normal recombination region in Europeans, and, in a CS for French-Canadians, perhaps demonstrating how changes in population recombination histories could influence population-specific ASE patterns. The variation in ASE and recombination rates at these genes across populations suggests that N_e and recombination influence the efficiency of genome regulation at a continental scale. Overall, these

findings highlight genes with population-specific ASE, which may be explained by recombination and N_e variability within and across genomes.

Environmental modifiers affect the ASE regulation at heterozygous loci containing a deleterious mutation

Population differences in ASE may not only be attributable to the haplotypic genetic variation as a consequence of N_e and recombination but may also be modulated by epigenetic modifiers through environmental exposure. Evolutionary change through natural selection is a driver of genetic adaptation in many populations, and these evolutionary forces, in combination with the interaction between genetic and environmental variables, contribute to the epigenetic nature of disease (*52*).

To test for potential environmental effects on ASE regulation at specific genes, we investigated ASE differences between French-Canadian individuals whose regional ancestry is different from their current residential region, in which we previously identified environmental variables associated with profiles of gene expression, including pollutant levels (*13*). We computed DE analysis using subpopulations of Quebec with differing demographic bottleneck and migratory histories using the method previously described for ancestral comparisons, separating individuals as either population locals, those whose ancestry are identical to the current region where they reside, or population migrants, whose ancestry and current region differ (Materials and Methods). We investigated 92 genes that had expression data across all regions and environments. We identified genes with differential ASE across individuals having the same ancestry but living in different environments by comparing individuals with ancestry from Saguenay but living in Montreal, Quebec City, or Saguenay (Fig. 5A). We captured genes with deleterious variants that show different ASE profiles between individuals with the same ancestry but living in different geographic environments, indicating a possible effect of the environment on ASE regulation. These genes include *GBP3*, *PTPRA*, *UBE3B*, and *NDUFS2* (Fig. 5, C and D). *GBP3* encodes a member of the GBP family, *PTPRA* encodes a protein tyrosine phosphatase, *UBE3B* is involved in protein degradation, and *NDUFS2* encodes a protein that is a core subunit of the mitochondrial membrane respiratory chain (*51*). Conversely, we also identified genes with differential ASE between individuals living in the same regional environment but having different genetic ancestries, indicating a possible effect of recent ancestry itself on ASE regulation (Fig. 5B). These genes include *CHPT1*, *GBP3*, *CTSB*, *KIAA0922*, and *MCM3AP* (Fig. 5D). *CHPT1* plays a role in formation and maintenance of vesicular membrane, *CTSB* is involved in intracellular degradation and turnover of proteins, *KIAA0922* antagonizes canonical Wnt signaling, and *MCM3AP* is essential for the initiation of DNA replication (*51*). Similarly with gene ontology analyses, we did not observe significant overrepresentation of a particular gene category. *GBP3* was observed to have variable ASE in both the ancestrality and environmental analyses (Figs. 4D and 5, C and D). *GBP3* exhibits antiviral activity against influenza virus, which might suggest that ASE changes may be in response to changing antiviral activity (*51*). Together, these results suggest that environmental modifiers also contribute to ASE variability in addition to the genetic ancestry of an individual.

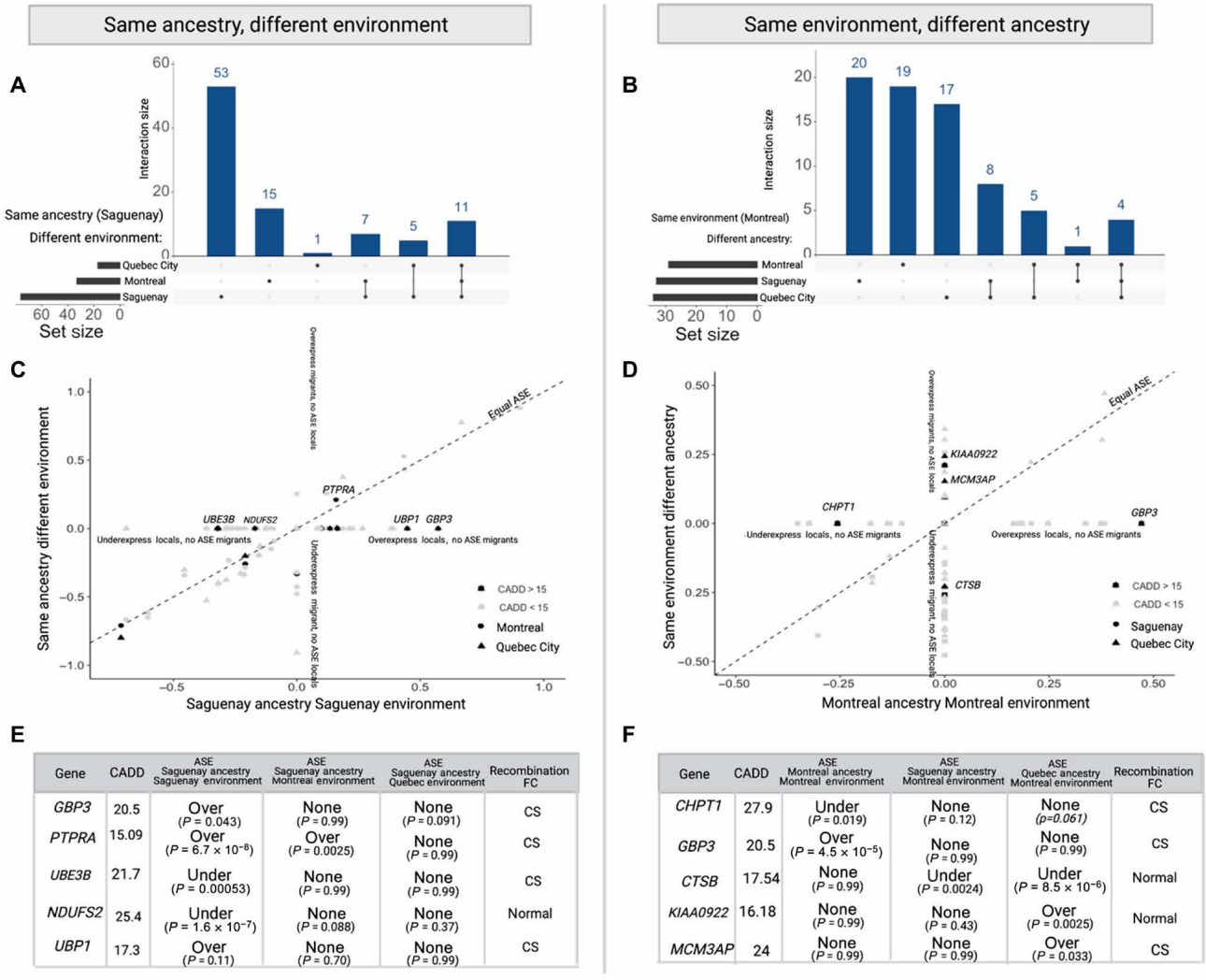
DISCUSSION

The regulation of gene expression has a leading role in the link between genotypes and phenotypes. In diploid organisms, the independent regulation of two alleles of a gene can have profound implications on gene expression, ultimately affecting the phenotypic penetrance. Regions of low recombination are enriched for deleterious variants in human populations due to consequences of HR interference (*15*); however, the effects of HR interference and recombination rate on the regulation of gene expression have remained largely unexplored. Recently, it has been documented that haplotype combinations that reduce the expression of deleterious alleles are depleted in healthy individuals, providing evidence of purifying selection on gene expression regulation (*18*). Here, we asked how natural selection influences the genomic distribution of ASE. We demonstrate that ASE is not randomly distributed along the genome and that population parameters such as varying recombination rate, per-site and tissue-specific expression level, and population histories may facilitate ASE evolution due to their impact on the efficiency of natural selection.

In this work, we demonstrate that patterns of ASE are also affected by HR effects, as we observe less ASE regulation in regions of recombination CS. In CS, fewer recombination events are expected between each ASE variant and their regulatory variants, leading to a faster loss of diversity in regulatory regions and thus a smaller N_e (*53*). A possible mechanism based on our results may be that the loss of diversity in regulatory regions in CS can lead to an equal ratio of allelic expression in these regions versus HRR because of regulatory polymorphisms eventually becoming fixed in the population. In addition, we demonstrate that protective haplotype combinations, where a cis-regulatory variant is linked to the underexpression of a deleterious coding variant, are more commonly observed in HRR/Normal regions compared to CS. This result supports and extends previous findings by Castel *et al.* (*18*), suggesting that selection removes haplotype combinations that are more damaging, as HRR/Normal regions have more efficient selection and have more haplotype combinations that underexpress a deleterious allele. These results show how purifying selection and recombination have acted together on the combination of cis-regulatory variation with coding variation, which influences the penetrance of pathogenic variants in human populations.

Within the CARTaGENE population cohort and replicated across several tissues in GTEx, we show that deleterious derived alleles are generally underexpressed, which likely tampers downstream effects on phenotypes. We observe a higher prevalence of ASE in genomic regions with HRR/Normal recombination, specifically underexpressing the derived allele for sites with pathogenic variants, and consistent with HRR/Normal regions of the genome having higher N_e and more efficiently underexpressing pathogenic variants. This pattern is observed at variants with high total expression levels, as this is an additional factor that contributes to selection efficiency. We also observe more ASE in HRR/Normal regions across a number of human populations.

The influence of natural selection through recombination on ASE was documented across tissue types; however, we did not observe HRR underexpressing deleterious variants more efficiently in brain or ovarian tissue. The divergence of pattern for some tissues is likely attributed to the tissue specificity of gene expression at many loci, where brain tissue often has the largest deviation from other tissue types (*54*). In addition, brain and ovarian tissues had the lowest proportion of sites with statistically significant ASE, with 3.7 and 3.8%, respectively (table S2), which may also contribute to altered results. We also hypothesize that tissue specificity observed in the blood may be due to allele-specific changes during immune response (*55*),



E

Gene	CADD	ASE Saguenay ancestry Saguenay environment	ASE Saguenay ancestry Montreal environment	ASE Saguenay ancestry Quebec environment	Recombination FC
GBP3	20.5	Over ($P = 0.043$)	None ($P = 0.99$)	None ($P = 0.091$)	CS
PTPRA	15.09	Over ($P = 6.7 \times 10^{-8}$)	Over ($P = 0.0025$)	None ($P = 0.99$)	CS
UBE3B	21.7	Under ($P = 0.00053$)	None ($P = 0.99$)	None ($P = 0.99$)	CS
NDUFS2	25.4	Under ($P = 1.6 \times 10^{-7}$)	None ($P = 0.088$)	None ($P = 0.37$)	Normal
UBP1	17.3	Over ($P = 0.11$)	None ($P = 0.70$)	None ($P = 0.99$)	CS

F

Gene	CADD	ASE Montreal ancestry Montreal environment	ASE Saguenay ancestry Montreal environment	ASE Quebec ancestry Montreal environment	Recombination FC
CHPT1	27.9	Under ($P = 0.019$)	None ($P = 0.12$)	None *(p=0.061)*	CS
GBP3	20.5	Over ($P = 4.5 \times 10^{-5}$)	None ($P = 0.99$)	None ($P = 0.99$)	CS
CTSB	17.54	None ($P = 0.99$)	Under ($P = 0.0024$)	Under ($P = 8.5 \times 10^{-6}$)	Normal
KIAA0922	16.18	None ($P = 0.99$)	None ($P = 0.43$)	Over ($P = 0.0025$)	Normal
MCM3AP	24	None ($P = 0.99$)	None ($P = 0.99$)	Over ($P = 0.033$)	CS

Fig. 5. Environment contributes to differential ASE across populations. DE analyses were computed at 92 genes (Materials and Methods). Ancestry-environment relationships were evaluated using migrant French-Canadian individuals. (**A**) Upset plot demonstrating the number of overlapping significant DE genes (FDR < 0.05) from individuals with Saguenay ancestry but varying environments. (**B**) Upset plot demonstrating the number of overlapping significant DE genes (FDR < 0.05) from individuals living in Montreal with different ancestry. (**C**) Log_2 fold change from individuals with Saguenay ancestry and environment compared to individuals with Saguenay ancestry living in Montreal or Quebec City. Each dot represents a gene. Log_2 fold change on the *x* axis is based on individuals with ancestry and environment in Saguenay, whereas the *y* axis includes individuals with ancestry from Saguenay with Montreal or Quebec City environments. Positive log_2 fold change demonstrates overexpression of the derived allele, negative demonstrates underexpression, and zero is not significant. Dashed line shows if the values were theoretically equal. (**D**) Log_2 fold change from individuals with Montreal ancestry and environment compared to individuals with Montreal environment with ancestry from Quebec or Saguenay. Each dot represents a gene. Log_2 fold change on the *x* axis is based on individuals whose ancestry and environment is Montreal, whereas the *y* axis includes individuals with Montreal environment with Saguenay or Quebec City ancestry. Dashed line demonstrates if the population values were theoretically equal. (**E**) Genes with deleterious variants (CADD > 15) that demonstrated different ASE from (C). (**F**) Genes with deleterious variants (CADD > 15) that demonstrated different ASE from (D).

as immunity is strongly targeted by natural selection. We observed that *GBP3*, a gene involved in antiviral activity against influenza virus (*51*), had variable ASE across ancestral populations and environments, further supporting the notion that immune-related genes may demonstrate variable ASE during immune response.

We further highlight numerous genes with population-specific ASE based on genetic ancestry or geographic environment. Population-specific ASE may be partially explained by past demographic histories, such as recent bottlenecks, or different exposures (*13*). Strong bottlenecks lead to reduced genetic diversity, and any disease-associated alleles that are present in the founder individuals may rise in frequency in the descending population due to strong forces of genetic drift and weaker selection, which consequently explains the higher prevalence of many genetic diseases in these founder populations (*56*). Similarly to low recombination regions of the genome, small post-bottleneck populations have a smaller N_e, and thus the reduced genetic diversity in regulatory regions leads to increased probability of equal allelic expression.

There is a long-standing investigation in evolutionary biology about how past demographic changes have affected modern phenotypes and disease risk (*57*, *58*). The impact of evolutionary factors, such as natural selection and genetic drift, on gene expression variation

has been understudied, likely due to the lack of empirical data previously available (*59*). Using large population cohorts, we have assessed ASE variation across the genome and global populations with different population sizes. Gene expression is a key intermediate step in translating genotypes into phenotypes, and thus understanding how gene expression is regulated and evolves is critical for deconvoluting the relationship between phenotypic variation and disease penetrance across human populations.

MATERIALS AND METHODS

Data

The CARTaGENE biobank comprises more than 40,000 participants aged between 40 and 60 years, recruited at random among three urban centers in the province of Quebec: Montreal, Quebec City, and Saguenay. Data used in this study have been previously published by Favé *et al.* (*13*). The study protocol was approved by the University of Toronto, and all participants provided informed consent. In total, 1000 samples were genotyped on the Illumina Omni2.5 array to obtain high-density SNP genotyping data. A total of 1,213,103 SNPs were retained after filtering and quality control (Hardy-Weinberg P value > 0.001, minor allele frequency > 5%, and percent of missing data < 1%). Paired-end RNA sequencing was performed on a HiSeq 2000 platform at the Genome Quebec Innovation Center (Montreal, Canada). Sequencing was performed for freeze 1 (708 samples) using three samples per lane and using six samples per lane for freeze 2 (292 samples), yielding about 60 million reads per sample. Mapping bias was corrected using WASP v0.3.4 (*34*), which also corrects for GC content affecting read depth inconsistencies by fitting polynomials to the read counts and calculating a corrected read depth per region (*34*). Possible CNVs were identified from the Database of Genomic Variants (*60*), and genes overlapping with these variants were removed. Missense or synonymous SNPs with >30 total reads were retained for ASE calculations (n = 3083 SNPs).

Genetic ancestry

Favé *et al.* (*13*) determined genetic ancestry from CARTaGENE participants using ChromoPainter v0.04 (*61*) to detect fine-scale genetic structure, followed by the fineSTRUCTURE algorithm. Regional ancestry for each French-Canadian was determined on the basis of the three clusters obtained from the fineSTRUCTURE tree, which was in agreement with Quebec settlement history revealing subpopulations of individuals that follow a North-South structure.

Ancestral allele identification

Ancestral allele annotations were captured from the 1000 Genomes Phase 1 (*62*). Ancestral alleles were defined as the allele found in human, chimpanzee, and rhesus macaque genomes. Only alleles that overlapped all three genomes were used, resulting in 289,212 individual-SNP pairs retained, where 51.3% had the reference allele match the ancestral allele, and 48.6% had the alternative allele match the ancestral allele.

ASE calculations

ASE estimates were calculated for germline heterozygous positions from CARTaGENE. Significant ASE sites were determined using two-tailed binomial tests with Benjamini-Hochberg multiple testing correction (FDR < 0.05) to test the null hypothesis that $P = 0.5$, where P is the proportion of reads with the derived allele divided by the total read count for the site. For the 289,212 individual-SNP pairs with ancestral allele annotations, significant ASE was separated into overexpression and underexpression of the derived allele by dividing the derived allele count by the total read count. A derived allele proportion of >0.5 demonstrates overexpression of the derived allele and <0.5 demonstrates underexpression of the derived allele.

Tissue replication with GTEx

GTEx was used for replication of the results observed in CARTaGENE. We used the whole blood (n = 654), muscle (n = 288), brain (n = 175), ovarian (n = 91), lung (n = 210), and liver (n = 84) tissue for 752 individuals from GTEx v8 release (*33*) using ASE calculations published by Castel *et al.* (*32*). GTEx also uses WASP for mapping bias correction. The total number of sites tested for ASE significance using binomial tests and after filtering were 5974, 5738, 4431, 6795, 9035, and 5091 SNPs for the whole blood, muscle, brain, ovarian, lung, and liver tissue, respectively. All individuals included had ancestry from Europe or Africa.

Recombination mapping

Recombination maps were built by Hussin *et al.* (*15*), who built linkage disequilibrium-based genetic maps using genotyping data from French-Canadian populations from CARTaGENE and CEU (Northern Europeans from Utah) and YRI (Yorubans from Ibadan, Nigeria) populations from HapMap 3 (*37*). Maps were used to locate CSs and hotspots of recombination. CSs are defined as regions longer than 50 kb, with a recombination rate below 0.5 cM/Mb. Hotspots are defined as short segments (<15 kb), with recombination rates above 5 cM/Mb. HRRs are regions with high density of hotspots, where the distance between neighboring hotspots is less than 50 kb. CSs and HRR were identified for each population independently. We identified 1560 CSs and 165 HRR in French-Canadians, 1140 CS and 81 HRR in Europeans, and 288 CSs and 16 HRR in Africans. Many analyses grouped normal recombination with HRR (referred to as HRR/Normal) due to the low number of sites in HRR and the similar direction of effect observed between normal and HRRs (fig. S2). Estimates of effective population size were also obtained by Hussin *et al.* (*15*) and McEvoy *et al.* (*50*).

Variant annotation

We annotated SNPs based on function using SeattleSeq Annotation v9.10 (*63*). We retained missense and synonymous variants for analysis. Nonsense and splicing variants can also result in ASE through nonsense-mediated decay and alternative splicing; however, it may have different selective pressures compared to ASE as a result of regulatory mechanisms. To remove potential noise from nonsense mediated decay and alternative splicing, we removed nonsense and splice variants from downstream analyses. We annotated SNPs using CADD v1.6 for a measure of deleteriousness, where a deleterious mutation was classified as CADD > 15, as previously reported (*36*). Variants with potential and validated disease associations were retrieved from the ClinVar (*40*) and OMIM databases (*41*). We retrieved 653 variants in the CARTaGENE dataset in ClinVar and 340 sites in OMIM genes. Absolute gene expression for each SNP is determined by the total read count, correcting for library size. Per-site total gene expression was calculated on the basis of tissue-specific quartiles, where low expression is observed within the first quartile and high expression is observed above the third quartile. GC content, exon size, average exon expression, and CpG islands were obtained from Hussin *et al.* (*15*),

which determined these metrics based on the same CARTaGENE French-Canadian samples.

Recombination class as a function of derived allele expression class

We modeled the recombination class as a function of the observed behavior of derived allele expression (overexpressed/no ASE/underexpressed). Specifically, three separate models were fit by regressing the binary outcome of recombination class (CS versus HRR/Normal) separately on each of three binary covariates for derived allele expression behavior: (i) overexpressed versus no ASE, (ii) no ASE versus ASE, and (iii) under expressed versus no ASE. Random effects due to individual were included in the models to account for repeated measurements using the glmer() function with "binomial" link function (*64*), with confidence intervals calculated using the Wald method. Significance (*P* values) of the estimated main effect (odds ratio) of derived allele expression behavior on recombination class was estimated using likelihood ratio tests. These models were fit to the datasets stratified by population/tissue type and data subsets for putatively disease-associated variants (defined by ClinVar and OMIM).

Recombination class as a function of ASE-eQTL concordance

We tested for eQTLs using an additive model with Matrix eQTL v2.1.0 (*65*) with a Bonferroni correction to identify eQTL sites and their associated eGenes. We removed the effects of hidden covariates using surrogate variable analysis, as described by Favé *et al.* (*13*), where five surrogate variables (technical and biological) and variance associated to batch (technical) were used for correction. One significant eQTL per eGene was selected on the basis of having the smallest *P* value (FDR < 0.05), and we identified ASE SNPs within each eGenes. Ancestral alignments for the eQTL site were used to determine ancestral and derived alleles using the same method described above. Haplotype phasing was performed using SHAPEIT (v2.r64410) (*66*) using default parameters, as described in Favé *et al.* (*13*). A total of 333 ASE SNPs and their associated eQTLs were used to identify ASE-eQTL concordance. ASE-eQTL haplotype concordance were defined as combinations that either (i) ASE-derived allele and eQTL-derived allele were on the same haplotype, the eQTL β value was negative, and ASE was underexpressing the derived allele or (ii) ASE-derived allele and eQTL-derived allele were on different haplotypes, the eQTL β value was positive, and ASE was underexpressing the derived allele. All other combinations were considered discordant. We modeled recombination class (CS versus HRR/Normal) as a function of concordance, stratified by CADD score, where CADD of >15 signifies likely deleterious variants. Random effects due to individual were included in the models to account for repeated measurements using the glmer() function with binomial link function in the lme4 package (*64*) in the R Project software (*67*), with confidence intervals calculated using the Wald method. Significance (*P* values) of the estimated main effect (odds ratio) of derived allele expression behavior on recombination class was estimated using likelihood ratio tests.

Proportion of derived allele expression and CADD score

We modeled the proportion of the derived allele expression to the total expression as a function of CADD score using binomial mixed-effects regression. The primary model regressed the bivariate vector of the derived and total allele expression on CADD score with the inclusion of random effect due to individual, thus accounting for repeated measurements using the generalized linear mixed-effects regression function glmer() function with binomial link function from the lme4 package v1.1-27.1 (*64*) in the R Project software (*67*). In addition to the primary model with the fixed main effect of interest (CADD score), we additionally stratified by population. Figures visualize the predictions from the models using ggpredict() in the ggeffects package v1.1.0 (*68*) in R, with 95% confidence bands. Significance (*P* values) of the estimated main effect (odds ratio) of CADD score in the primary model was estimated using a likelihood ratio test comparing "full" models with nested submodels, where the covariate of interest was removed.

Proportion of derived allele expression and CADD score with total expression as an interaction effect

In addition to the primary model with the fixed main effect of interest (CADD score), we additionally tested for the presence of an interaction effect between CADD score and the level of total expression (coded as a factor variable with three levels: low, medium, and high). These models were fit to the datasets stratified by recombination region (CS/HRR). Figures visualize the predictions from the models using ggpredict() in the ggeffects package v1.1.0 (*68*) in R, with 95% confidence bands. In addition, to determine whether the results observed with total expression as an interaction effect was explained by read coverage, we randomly sampled 250 reads, with replacement, from each position, regardless of expression class, and fit the mixed-effects regression models stratified by expression on the randomly sampled read proportions. The random sampling procedure was performed for 25 iterations, with the mean β, intercept, and 95% confidence intervals being compared after all iterations.

DE by ancestry and environment

DE per gene was computed using DESeq2 v1.32.0 (*69*) using the model ~Derived/Ancestral + ID, where Derived/Ancestral represents the read counts for each allele and ID represents the individual. Genes were retained only if (i) there were only one SNP per gene for each individual or multiple SNPs with ASE directionality concordance where the read counts were summed and (ii) with expression data from all populations. These restrictions resulted in a total of 1045 genes tested. The model was fit after stratifying for population ancestry or environment. Genes with FDR < 0.05 and where the absolute[$\log_2$(Derived) − $\log_2$(Ancestral)] > 0.25 were considered differentially expressed. Genes with DE in one or more populations were compared for a total of 133 genes.

Africa was removed from analyses due to the small sample size causing concerns regarding large quantities of false negatives. Since the DE analysis calculates across all individuals in the population, we would only be comparing 19 data points, as that is the sample size that we have for African individuals. We would therefore observe less significant genes from African individuals due to low power at most genes. Thus, when comparing significant genes directly across the populations, we may incorrectly identify certain genes with ASE in other populations but not Africa, which may be an artifact of sample size and not biological. This was not a concern for previous analyses, because we calculated ASE within each individual; therefore, we are more confident with the ASE estimates because the read coverage that we have per individual has passed quality checks and thresholds, providing enough power to accurately test for ASE, regardless of sample size.

DE analyses were also computed to compare ancestry and environmental effects on ASE using migrant French-Canadian individuals. DE analysis on ancestry-environment effects used the same model but stratified by ancestry-environment groups. We compared Saguenay ancestry living in Saguenay ($n = 93$), Saguenay ancestry living in Montreal ($n = 27$), and Saguenay ancestry living in Quebec City ($n = 38$) in addition to Montreal ancestry living in Montreal ($n = 215$), Montreal ancestry living in Saguenay ($n = 4$), and Montreal ancestry living in Quebec ($n = 53$). A total of 93 genes with expression across all geographic regions were used.

DE analyses were compared investigating the effect size ($\log_2$ fold change) for each group, where a positive effect size signifies overexpression of the derived allele and negative values signify underexpression of the derived allele, and genes that were not significant were reassigned a value of 0. CADD scores for variants in the genes identified were also investigated to identify genes with highly deleterious variants (CADD > 15) that had altered ASE across populations.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abl3819

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. T. Pastinen, Genome-wide allele-specific analysis: Insights into regulatory variation. *Nat. Rev. Genet.* **11**, 533–538 (2010).
2. M. Pirinen, T. Lappalainen, N. A. Zaitlen; GTEx Consortium, E. T. Dermitzakis, P. Donnelly, M. I. McCarthy, M. A. Rivas, Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics* **31**, 2497–2504 (2015).
3. L. Valle, T. Serena-Acedo, S. Liyanarachchi, H. Hampel, I. Comeras, Z. Li, Q. Zeng, H. T. Zhang, M. J. Pennison, M. Sadim, B. Pasche, S. M. Tanner, A. Chapelle, Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science* **321**, 1361–1365 (2008).
4. Q.-X. Wei, R. Claus, T. Hielscher, D. Mertens, A. Raval, C. C. Oakes, S. M. Tanner, A. de la Chapelle, J. C. Byrd, S. Stilgenbauer, C. Plass, Germline allele-specific expression of DAPK1 in chronic lymphocytic leukemia. *PLOS ONE* **8**, e55261 (2013).
5. D. M. McKean, J. Homsy, H. Wakimoto, N. Patel, J. Gorham, S. R. DePalma, J. S. Ware, S. Zaidi, W. Ma, N. Patel, R. P. Lifton, W. K. Chung, R. Kim, Y. Shen, M. Brueckner, E. Goldmuntz, A. J. Sharp, C. E. Seidman, B. D. Gelb, J. G. Seidman, Loss of RNA expression and allele-specific expression associated with congenital heart disease. *Nat. Commun.* **7**, 12824 (2016).
6. X. Rao, K. S. Thapa, A. B. Chen, H. Lin, H. Gao, J. L. Reiter, K. A. Hargreaves, J. Ipe, D. Lai, X. Xuei, Y. Wang, H. Gu, M. Kapoor, S. P. Farris, J. Tischfield, T. Foroud, A. M. Goate, T. C. Skaar, R. D. Mayfield, H. J. Edenberg, Y. Liu, Allele-specific expression and high-throughput reporter assay reveal functional genetic variants associated with alcohol use disorders. *Mol. Psychiatry* **26**, 1142–1151 (2021).
7. M. Langmyhr, S. P. Henriksen, C. Cappelletti, W. D. J. van de Berg, L. Pihlstrøm, M. Toft, Allele-specific expression of Parkinson's disease susceptibility genes in human brain. *Sci. Rep.* **11**, 504 (2021).
8. T. Lappalainen, E. Salmela, P. M. Andersen, K. Dahlman-Wright, P. Sistonen, M. L. Savontaus, S. Schreiber, P. Lahermo, J. Kere, Genomic landscape of positive natural selection in Northern European populations. *Eur. J. Hum. Genet.* **18**, 471–478 (2010).
9. S. Kudaravalli, J. B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, J. K. Pritchard, Gene expression levels are a target of recent natural selection in the human genome. *Mol. Genet. Evol.* **3**, 649–658 (2009).
10. I. Idaghdour, W. Czika, K. V. Shianna, S. H. Lee, P. M. Visscher, H. C. Martin, K. Miclaus, J. Jadallah, D. B. Goldstein, R. D. Wolfinger, G. Gibson, Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* **42**, 62–67 (2010).
11. I. Idaghdour, P. Awadalla, Exploiting gene expression variation to capture gene-environment interactions for disease. *Front. Genet.* **3**, 228 (2012).
12. P. Awadalla, C. Boileau, Y. Payette, Y. Idaghdour, J.-P. Goulet, B. Knoppers, P. Hamet, C. Laberge; CARTaGENE Project, Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *Int. J. Epidemiol.* **42**, 1285–1299 (2013).
13. M.-J. Favé, F. C. Lamaze, D. Soave, A. Hodgkinson, H. Gauvin, V. Bruat, J.-C. Grenier, E. Gbeha, K. Skead, A. Smargiassi, M. Johnson, Y. Idaghdour, P. Awadalla, Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nat. Commun.* **9**, 827 (2018).
14. H. M. Natri, K. S. Bobowik, P. Kusuma, C. C. Darusallam, G. S. Jacobs, G. Hudjashov, J. S. Lansing, S. Sudoyo, N. E. Banovich, M. P. Cox, I. G. Romero, Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the Indonesian archipelago. *PLOS Genet.* **16**, e1008749 (2020).
15. J. G. Hussin, A. Hodgkinson, Y. Idaghdour, J.-C. Grenier, J.-P. Goulet, E. Gbeha, E. Hip-Ki, P. Awadalla, Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat. Genet.* **47**, 400–404 (2015).
16. F. Casals, A. Hodgkinson, J. Hussin, Y. Idaghdour, V. Bruat, T. Millard, J. C. Grenier, E. Gbeha, F. F. Hamdan, S. Girard, J. F. Spinella, M. Larivier, V. Saillour, P. Awadalla, Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLOS Genet.* **10**, e1004260 (2013).
17. S. Peischl, I. Dupanloup, A. Foucal, M. Jomphe, V. Bruat, J. C. Grenieer, A. Gouy, K. J. Gilbert, E. Gbeha, L. Bosshard, E. Hip-Ki, M. Agbessi, A. Hodgkinson, H. Vezina, P. Awadalla, L. Excoffier, Relaxed selection during a recent human expansion. *Genetics* **208**, 763–777 (2018).
18. S. E. Castel, A. Cervera, P. Mohammadi, F. Aguet, F. Reverter, A. Wolman, R. Guigo, I. Iossifov, A. Vasileva, T. Lappalainen, Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
19. E. C. Glassberg, Z. Gao, A. Harpak, X. Lan, J. K. Pritchard, Evidence for weak selective constraint on human gene expression. *Genetics* **211**, 757–772 (2019).
20. M. Nordborg, B. Charlesworth, D. Charlesworth, The effect of recombination on background selection. *Genet. Res.* **67**, 159–174 (1996).
21. A. Eyre-Walker, P. D. Keightley, N. G. C. Smith, D. Gaffney, Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* **19**, 2142–2149 (2002).
22. D. C. Presgraves, Recombination enhances protein adaptation in Drosophila melanogaster. *Curr. Biol.* **15**, 1651–1656 (2005).
23. P. Keightley, S. Otto, Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* **443**, 89–92 (2006).
24. M. Hartfield, S. P. Otto, P. D. Keightley, The role of advantageous mutations in enhancing the evolution of a recombination modifier. *Genetics* **184**, 1153–1164 (2010).
25. H. J. Muller, The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2–9 (1964).
26. W. G. Hill, A. Robertson, The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
27. R. M. Kliman, J. Hey, Reduced natural selection associated with low recombination in Drosophila melanogaster. *Mol. Biol. Evol.* **10**, 1239–1258 (1993).
28. A. Caballero, E. Santiago, Response to selection from new mutation and effective size of partially inbred populations. I. Theoretical results. *Genet. Res.* **66**, 213–225 (1995).
29. J. Wang, W. G. Hill, D. Charlesworth, B. Charlesworth, Dynamics of inbreeding depression due to deleterious mutations in small populations: Mutation parameters and inbreeding rate. *Genet. Res.* **74**, 165–178 (1999).
30. J. M. Comeron, A. Williford, A. Kliman, The Hill-Robertson effect: Evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**, 19–31 (2008).
31. A. Hodgkinson, Y. Idaghdour, E. Gbeha, J. C. Grenier, E. Hip-Ki, V. Bruat, J. P. Goulet, T. de Malliard, P. Awadalla, High-resolution of genomic analysis of human mitochondrial RNA sequence variation. *Science* **25**, 413–415 (2014).
32. S. E. Castel, F. Aguet, P. Mohammadi; GTEx Consortium, K. G. Ardlie, T. Lappalainen, A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (2020).
33. GTEx Consortium, The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
34. D. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
35. Z. Liu, X. Dong, Y. Li, A genome-wide study of allele-specific expression in colorectal cancer. *Front. Genet.* **9**, 570 (2018).
36. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
37. HapMap Consortium, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
38. C. C. Spencer, P. Deloukas, S. Hunt, J. Mulikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, G. McVean, The influence of recombination on human genetic diversity. *PLOS Genet.* **2**, e148 (2006).
39. S. M. Fullerton, A. B. Carvalho, A. G. Clark, Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**, 1139–1142 (2001).
40. M. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, D. R. Maglott, ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, 980–985 (2014).
41. V. A. McKusick, *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders* (Johns Hopkins Univ. Press, ed. 12, 1998) vol. 4, pp. 588–604.

42. J. F. Gout, D. Khan, L. Duret, The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLOS Genet.* **6**, e1000944 (2010).
43. R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).
44. M. H. Roy-Gagnon, C. Moreau, C. Bherer, P. St-Onge, P. D. Sinnett, C. Laprise, H. Vézina, D. Labuda, Genomic and genealogical investigation of the French-Canadian founder population structure. *Hum. Genet.* **129**, 521–531 (2011).
45. H. Li, R. Durbin, Inference of human population history from individual whole genome sequences. *Nature* **475**, 493–496 (2011).
46. M. Kimura, T. Ohta, The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61**, 763–771 (1969).
47. B. M. Henn, L. R. Botigué, S. Peischl, I. Dupanloup, M. Lipatov, B. K. Maples, A. R. Martin, S. Musharoff, H. Cann, M. P. Snyder, L. Excoffier, J. M. Kidd, C. D. Bustamante, Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E449–E449 (2016).
48. A. Tenesa, P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, P. M. Visscher, Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* **17**, 520–526 (2007).
49. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
50. B. P. McEvoy, J. E. Powell, M. E. Goddard, P. M. Visscher, Human population dispersal "out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res.* **21**, 821–829 (2011).
51. G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran, D. Lancet, The GeneCards Suite: From gene data mining to disease genome sequence analyses. *Bioinformatics* **54**, 1.30.1–1.30.33 (2016).
52. C. H. Waddington, Epigenetics and evolution. *Symp. Soc. Exp. Biol* **7**, 186–199 (1953).
53. F. Fyon, A. Cailleau, T. Lenormand, Enhancer runaway and the evolution of diploid gene expression. *PLOS Genet.* **11**, e1005665 (2015).
54. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
55. M. Gutierrez-Arcelus, Y. Baglaenko, J. Arora, S. Hannes, Y. Luo, T. Amariuta, N. Teslovich, D. A. Rao, J. Ermann, A. H. Jonsson; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, C. Navarrete, S. S. Rich, K. D. Taylor, J. I. Rotter, P. K. Gregersen, T. Esko, M. B. Brenner, S. Raychaudhuri, Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat. Genet.* **52**, 247–253 (2020).
56. A. Prohaska, F. Racimo, A. J. Schork, M. Sikora, A. J. Stern, M. Ilardo, M. E. Allentoft, L. Folkersen, A. Buil, J. V. Moreno-Mayar, T. Korneliussen, D. Geschwind, A. Ingason, T. Werge, R. Nielsen, E. Willerslev, Human disease variation in the light of population genomics. *Cell* **117**, 115–131 (2019).
57. S. Ramachandran, O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, L. Cavalli-Sforza, Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15942–15947 (2005).
58. G. Coop, J. K. Pickrell, J. Novembre, S. Kudaravailli, J. Li, D. Absher, R. M. Myers, L. L. Cavalli-Sforza, M. W. Felman, J. K. Pritchard, The role of geography in human adaptation. *PLOS Genet.* **5**, e1000500 (2009).
59. M. S. Hill, P. V. Zande, P. J. Wittkopp, Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* **22**, 203–215 (2020).
60. J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, S. W. Scherer, The database of genomic variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, 986–992 (2014).
61. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLOS Genet.* **8**, e1002453 (2012).
62. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
63. S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, J. Shendure, Targeted capture and massively parallele sequencing of 12 human exomes. *Nature* **461**, 272–276 (2010).
64. D. Bates, M. Machler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2014).
65. A. Shabalin, Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
66. O. Delaneau, J. F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetics studies. *Nat. Med.* **10**, 5–6 (2013).
67. R Core Development Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2021); https://www.R-project.org/.
68. D. Luüdecke, ggeffects: Tidy data frames of marginal effects from regression models. *J. Open Source Softw.* **3**, 772 (2018).
69. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgments: We thank members of the P.A. lab, H. Gibling, A. A. Houle, T. Oullette, and E. Bader for comments on the manuscript. **Funding:** This work was supported by the Canadian Institutes of Health Research award no. EC3-144623 (to P.A.), Ontario Graduate Scholarship (to M.P.H.), and the Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship-Doctoral (to M.P.H.). **Author contributions:** Conceptualization: M.P.H., I.A., H.E., M.-J.F., F.C.L., and P.A. Data preparation: M.-J.F., M.A., and V.B. Statistical analysis: M.P.H., I.A., H.E., M.-J.F., F.C.L., and D.S. Visualization: M.P.H. Writing: M.P.H., I.A., H.E., M.-J.F., F.C.L., D.S., and P.A. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper, the Supplementary Materials, and external data repositories. Requests for data published here should be submitted to the CARTaGENE program at access@cartagene.qc.ca, citing this study. GTEx ASE data used in this study are available to authorized users through dbGaP (accession no. phs000424.v8).

Submitted 9 July 2021
Accepted 29 March 2022
Published 13 May 2022
10.1126/sciadv.abl3819