Supplementary Materials for

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

Michelle P. Harwood, Isabel Alves, Hilary Edgington, Mawusse Agbessi, Vanessa Bruat, David Soave, Fabien C. Lamaze, Marie-Julie Favé, Philip Awadalla*

*Corresponding author. Email: Philip.awadalla@oicr.on.ca

Published 13 May 2022, Sci. Adv. 8, eabl3819 (2022)
DOI: 10.1126/sciadv.abl3819

This PDF file includes:

Figs. S1 to S10
Tables S1 to S5
Fig S1. Modeling the proportion of the derived allele expression as a function of CADD score for all SNPs. Sites were scored based on scaled Combined Annotation Dependent Depletion (CADD) scores, where larger CADD signifies deleteriousness. We modelled the proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression (Methods).
Fig. S2. Comparison of ASE relationships in missense and synonymous SNPs. Sites were annotated based on mutation type and filtered to include missense and synonymous SNPs for analyses. (Methods). (A) Proportion of significant ASE sites that are missense compared to synonymous. (B) Model of proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression for synonymous SNPs only.
Fig. S3. Modeling ASE as a function of recombination class for all three recombination bins. We modelled ASE significance (ASE/ no ASE) as a function of recombination class (CS/HRR/Normal), accounting for random effects due to individuals (Methods). CS are regions >50 Kb with recombination rates <0.5 cM/Mb, and HRR are regions of high-density hotspots (Methods). OR > 1 means increased odds that variants in a recombination class exhibit significant ASE compared to no ASE. Error bars represent 95% confidence. Downstream analyses grouped normal and HRR based on the same direction of effects.
Fig. S4. Modelling the proportion of the derived allele expression as a function of CADD score for GTEx tissues. Sites were scored based on the scaled Combined Annotation Dependent Depletion (CADD) score, which signifies deleteriousness. (A) Model of proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression, stratified by HRR ($B=0.0018263, p=2.67 \times 10^{-14}$) and CSs ($B=-1.73 \times 10^{-03}, p<2 \times 10^{-16}$) for Brain tissue in GTEx. (B) Model of proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression (Methods), stratified by HRR ($B=0.001969, p=0.00646$) and CSs ($B=-0.0003877, p=0.0491$) for ovarian tissue in GTEx. (C) Model of proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression, stratified by HRR ($B=-0.0040676, p=1.23 \times 10^{-06}$) and CSs ($B=-0.002909, p<2 \times 10^{-16}$) for Lung tissue in GTEx. (D) Model of proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression, stratified by HRR ($B=-0.0019219, p<4.43 \times 10^{-11}$) and CSs ($B=-4.64 \times 10^{-04}, p=2.74 \times 10^{-07}$) for Muscle tissue in GTEx.
**Fig. S5. Random sampling procedure to test total expression versus read coverage.** Random sampling procedure was performed across all sites to test if the observed differences in total expression was an artifact of read coverage. All sites regardless of expression class were randomly sampled for 250 reads, with replacement, and we modeled the proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression, stratified by expression bin for HRR and CS. This random sampling procedure was performed 25 times, and the mean beta, intercept, and 95% confidence intervals were plotted. Red lines indicate models fit from sites in high recombination regions (HRR), black are models fit from sites in cold spots (CS), and grey lines indicate 95% confidence.
Fig. S6. Modelling recombination class as a function of ASE for GTEx tissues, stratified by total expression. We modelled recombination class (CS versus HRR/Normal) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals, stratified by per-site total expression (Methods). High per-site total expression categorizes sites above the third quartile, and low expression categorizes sites below the first quartile (Methods). OR > 1 means that variants in an expression class have increased odds of exhibiting ASE in HRR/Normal compared to CS. Error bars represent 95% confidence. Data used from GTEx (A) whole blood, (B) muscle, (C) brain, (D) ovarian, (E) lung, and (F) liver tissue.
Fig. S7. Modelling the proportion of the derived allele expression as a function of CADD score for GTEx tissues, stratified by total expression. We modeled the proportion of the derived allele expression to total expression as a function of CADD score using binomial mixed effects regression, stratified by expression bin for HRR and CS. Data used from GTEx (A) whole blood, (B) muscle, (C) brain, (D) ovarian, (E) lung, and (F) liver tissue.
Fig. S8. Modeling recombination class as a function of ASE, stratified by populations using equal sample sizes. All populations were subsampled randomly to have equal number of samples being tested (n=19). Sites were separated into recombination cold spots (CS) (regions >50 Kb with recombination rates <0.5 cM/Mb), and high-to-normal recombination (HRR/Normal) (all other regions, including high recombination regions and recombination hot spots). We modeled recombination class (CS versus HRR/Normal) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals, stratified by ancestral populations (Methods). OR > 1 means that sites have increased odds of exhibiting ASE in HRR/Normal compared to CS. Error bars represent 95% confidence.
Fig. S9. Modeling recombination class as a function of ASE, stratified by populations for GTEx tissues. Sites were separated into recombination cold spots (CS) (regions >50 Kb with recombination rates <0.5 cM/Mb), and high-to-normal recombination (HRR/Normal) (all other regions, including high recombination regions and recombination hot spots). We modeled recombination class (CS versus HRR/Normal) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals, stratified by ancestral populations (Methods). OR > 1 means that sites have increased odds of exhibiting ASE in HRR/Normal compared to CS. Error bars represent 95% confidence. Data used from GTEx (A) brain, (B) ovarian, (C) muscle, (D) lung, and (E) liver tissue.
Fig. S10. Modeling recombination class as a function of ASE, stratified by populations including East Asian population. Sites were separated into recombination cold spots (CS) (regions >50 Kb with recombination rates <0.5 cM/Mb), and high-to-normal recombination (HRR/Normal) (all other regions, including high recombination regions and recombination hot spots). We modeled recombination class (CS versus HRR/Normal) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals, stratified by ancestral populations (Methods). OR > 1 means that sites have increased odds of exhibiting ASE in HRR/Normal compared to CS. Error bars represent 95% confidence. Data used from GTEx and CARTaGENE whole blood.
Table S1. Summary of ASE significance genome-wide by population in CARTaGENE after filtering. ASE significance determined by binomial tests (FDR<0.05).

|     | Number of Unique Sites | Number of Tests | Number of Significant Over-expressed Derived Sites | Number of Significant Under-expressed Derived Sites | % Significant | % of Significant Sites that Under-express Derived Alleles |
|-----|------------------------|-----------------|---------------------------------------------------|---------------------------------------------------|---------------|--------------------------------------------------------|
| Total | 3,083                  | 480,333         | 11,456                                            | 12,133                                            | 4.9%         | 51.4%                                                  |
| Africa | 2,347                  | 18,638          | 236                                               | 296                                               | 2.8%         | 55.6%                                                  |
| Europe | 2,365                  | 73,375          | 1,747                                             | 1,846                                             | 4.8%         | 51.4%                                                  |
| Quebec City | 2,311               | 89,618          | 2,561                                             | 2,624                                             | 5.8%         | 49.4%                                                  |
| Montreal | 2,395                 | 153,394         | 3,486                                             | 3,753                                             | 4.7%         | 48.2%                                                  |
| Saguenay | 1,968                 | 89,423          | 1,968                                             | 2,072                                             | 4.5%         | 48.7%                                                  |

Table S2. Summary of ASE significance genome-wide by tissue in GTEx after filtering. ASE significance determined by binomial tests (FDR<0.05).

|     | Number of Unique Sites | Number of Tests | Number of Significant Over-expressed Derived Sites | Number of Significant Under-expressed Derived Sites | % Significant | % of Significant Sites that Under-express Derived Alleles |
|-----|------------------------|-----------------|---------------------------------------------------|---------------------------------------------------|---------------|--------------------------------------------------------|
| Whole Blood | 5,976                  | 130,222         | 8,056                                            | 7,097                                            | 11.6%         | 53.2%                                                  |
| Muscle      | 5,738                  | 169,355         | 5,807                                             | 5,199                                             | 6.5%         | 52.8%                                                  |
| Brain       | 4,431                  | 353,946         | 6,276                                             | 6,739                                             | 3.7%         | 48.2%                                                  |
| Ovarian     | 6,795                  | 75,680          | 1,507                                             | 1,395                                             | 3.8%         | 51.9%                                                  |
| Lung        | 9,035                  | 70,327          | 1,401                                             | 1,217                                             | 3.7%         | 53.5%                                                  |
| Liver       | 5,091                  | 1,706,134       | 39,518                                            | 33,904                                            | 4.3%         | 53.8%                                                  |
Table S3. Modelling recombination class as a function of ASE for low, medium, and high values of GC content, average exon expression, and exon size. We modelled recombination class (CS versus HRR/Normal) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals for CARTaGENE whole blood. We fit separate models after stratifying by low, medium, and high values for various genomic features (GC content, exon size, and average exon expression). “Low” category for all features are sites below the 1<sup>st</sup> quartile, “medium” includes sites between the 1<sup>st</sup> and 3<sup>rd</sup> quartile, and “high” includes above the 3<sup>rd</sup> quartile. Significant odds ratio estimates are reported. NS indicates not significant (p>0.05).

| Genomic Feature          | Category | Odds Ratio Derived Over-expression | Odds Ratio no ASE | Odds Ratio Derived Under-expression |
|--------------------------|----------|------------------------------------|------------------|-------------------------------------|
| GC content (% GC per exon) | Low      | NS                                 | 0.79             | 1.61                                |
|                          | Medium   | 1.13                               | 0.63             | 1.26                                |
|                          | High     | 1.94                               | 0.77             | 1.38                                |
| Exon Size (Kb)           | Low      | 1.62                               | 0.71             | 1.43                                |
|                          | Medium   | 1.13                               | 0.66             | 1.40                                |
|                          | High     | 1.14                               | 0.66             | 1.31                                |
| Average Exon Expression   | Low      | 1.46                               | 0.78             | 1.18                                |
|                          | Medium   | 0.84                               | 0.72             | 1.27                                |
|                          | High     | 2.10                               | 0.56             | 1.68                                |
Table S4. Modelling recombination class as a function of ASE including factors of GC content, average exon expression, exon size, and CpG islands. We modelled recombination class (CS versus HRR/Normal) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals for CARTaGENE whole blood. The first row of the table demonstrates the odds ratio estimates presented in the main text, and the subsequent rows demonstrate the odds ratio estimates with various genomic features (GC content, exon size (Kb), average exon expression, and CpG islands) included in the model. 95% confidence estimates are presented in square brackets next to the odds ratio estimates.

| Description                              | Odds Ratio Derived Over-expression | Odds Ratio no ASE [0.63 - 0.66] | Odds Ratio Derived Under-expression |
|------------------------------------------|-----------------------------------|---------------------------------|-----------------------------------|
| ASE~ Recombination                       | 1.24 [1.20 - 1.27]                | 0.65 [0.66]                     | 1.38 [1.34 - 1.42]                |
| ASE~ Recombination + GC Content          | 1.24 [1.20 - 1.27]                | 0.67 [0.66 - 0.68]              | 1.37 [1.33 - 1.41]                |
| ASE~ Recombination + Exon Size (Kb)      | 1.23 [1.19 - 1.26]                | 0.65 [0.64 - 0.65]              | 1.39 [1.35 - 1.43]                |
| ASE~ Recombination + Average Exon Expression | 1.24 [1.21 - 1.28]              | 0.65 [0.64 - 0.66]              | 1.38 [1.34 - 1.42]                |
| ASE~ Recombination + CpG Islands         | 1.16 [1.13 - 1.19]                | 0.66 [0.65 - 0.67]              | 1.37 [1.33 - 1.41]                |
Table S5. Recombination significantly improves models of ASE as a function of GC content, exon size, average exon expression, and CpG Islands. We modelled the magnitude of various genomic features (high versus low/medium) as a function of the observed behaviour of the derived allele expression (over-expressed/ no ASE/ under-expressed), accounting for random effects due to individuals for CARTaGENE whole blood. “High” category was defined for each genomic features as sites with values above the 3rd quartile. A second model was fit for each genomic feature, which included recombination class (CS versus HRR/Normal) into the model. The two models were compared by calculating the difference in R² values and an ANOVA to test if the addition of recombination significantly improves the model. Asterisks indicate level of significance where NS indicates p>0.05, * indicates p<0.05, ** indicates p<0.01, *** indicates p<0.001.

|                      | GC Content | Exon Size | Average Exon Expression | CpG Islands |
|----------------------|------------|-----------|-------------------------|-------------|
| **R²** ASE ~ Genomic Feature | Derived Over-expression | 0.00011 ** | 0.0020 *** | 0.018 *** | 0.0010 *** |
|                      | No ASE     | 0.0017 *** | 0.00034 *** | 0.0064 *** | 0.000070 *** |
|                      | Derived Under-expression | 0.0011 *** | 0.00028 *** | 0.0051 *** | 0.00019 *** |
| **R²** ASE ~ Genomic Feature + Recombination | Derived Over-expression | 0.0033 *** | 0.0046 *** | 0.021 *** | 0.0024 *** |
|                      | No ASE     | 0.0123 *** | 0.012 *** | 0.019 *** | 0.012 *** |
|                      | Derived Under-expression | 0.0064 *** | 0.0073 *** | 0.012 *** | 0.0065 *** |
| Difference in R²     | Derived Over-expression | -0.0032 | -0.0026 | -0.003 | -0.0014 |
|                      | No ASE     | -0.012 | -0.01166 | -0.013 | -0.01193 |
|                      | Derived Under-expression | -0.0053 | -0.0070 | -0.0069 | -0.00631 |
| ANOVA model comparison p-value | Derived Over-expression | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** |
|                      | No ASE     | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** |
|                      | Derived Under-expression | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** | < 2.2x 10 -16 *** |