S3 Appendix: Mediation analysis

For this study, simple mediation models consisting of three variables were used: the independent variable X, the mediator M, and the dependent variable Y [1, 2]. Superscripts are used to indicate whether a variable is the QTL, chromatin accessibility at site k (C_k), or expression of gene j (G_j) in the model. When an eQTL is detected, the following relationship is supported:

$$X^{QTL} \quad \rightarrow \quad Y^{G_j} \tag{1}$$

Candidate mediators of eQTL, either nearby chromatin regions or genes, must also possess their own local-QTL, represented similarly:

$$X^{QTL} \quad \rightarrow \quad M \tag{2}$$

where M is some Y^{G_k} or Y^{C_k}, a candidate mediator of the eQTL of Y^{G_j}. Here we consider two models of mediation: 1) the effect of an eQTL is mediated through local chromatin state (M^{C_k}) and 2) the effect of the distal-eQTL is mediated through a nearby gene (M^{G_k}), depicted in Fig 6. Co-localizing QTL are consistent with the co-occurrence of $X \rightarrow Y^{G_j}$ and $X \rightarrow Y^{C_k}$ at a locus. If X is fully independent of Y given M, denoted as:

$$X^{QTL} \quad \perp\!\!\!\perp \quad Y^{G_j} \quad | \quad M \tag{3}$$

where "A $\perp\!\!\!\perp$ B" indicates that A and B are independent of each other and "A | B" is A conditioned on B, then X acts on Y completely through M, which is commonly referred to as full mediation. Partial mediatory relationships are also possible in which X affects Y both directly and through M, traditionally identified by testing for:

$$M \quad \rightarrow \quad Y^{G_j} \quad | \quad X^{QTL} \tag{4}$$

In practice, the true relationships among the variables can be complex and obscured by noise, making it hard to completely distinguish full mediation, partial mediation, and X separately affecting M and Y. Furthermore, genome-wide data present a significant challenge to systematically testing these relationships because of the presence of many candidate mediators, potential correlation structure due to linkage disequilibrium (LD), and error heterogeneity across the levels of genomic data. For these reasons, we do not use the traditional mediation tests [1], but adapt and simplify them for genomic data.

Scans

We perform mediation scans, described in Eq 3 of the Methods. These scans test for significant reduction in the relationship $X \rightarrow Y|M$, akin to a full mediator (relationship 5), though without attempting to define a formal threshold for defining whether variables are weakly associated or independent. The relationship is tested for all mediators within the local region of the eQTL (± 10Mb), thus could plausibly explain the eQTL.

Permutations

Empirical p-values for the mediators can be defined through permutations, similarly to as is done with the QTL. Instead of permuting the trait outcome, the pairing of trait to QTL is maintained, and the mediator is permuted. GEVs are then fit for potential mediators from the minimum logP from the permutation scans, instead of the maximum logP in QTL scans. FWER-controlled mediation p-values are calculated from the cumulative distribution function of the GEV: $\text{permP}^m = F_{\text{GEV}}(\text{logP})$. An additional FDR adjustment is not used with mediation because it is only evaluated for detected QTL, which we found likely to violate the uniform null distribution expectations of FDR. Notably, we use the genome-wide and chromosome-wide measurement of molecular traits, the vast majority of which are ineligible to be mediators of the eQTL due to location, to characterize empirical error distributions for the mediator test, and thus derive FWER-controlling thresholds. This approach is statistically stringent, given that mediators must be local to the eQTL, which is appealing in terms of controlling false positives that could result in biologically-complex data.

Mediation criterion

Mediation of local-eQTL through chromatin state.

Mediation through chromatin accessibility was evaluated for all local-eQTL detected through the lenient Analysis L, leveraging prior knowledge of the predominance of local genetic regulation and the effects of proximal chromatin state. Candidate mediators are detected at genome and chromosome-wide significance through 1,000 permutations. Genome-wide mediation is detected if $\text{permP}_{\text{G}}^m < 0.05$ and chromosome-wide if $\text{permP}_{\text{C}}^m < 0.05$.

Mediation of distal-eQTL through local gene expression.

Mediation of the effect of genome-wide significant distal-eQTL (Analysis G) through the expression of nearby genes was evaluated. Genome-wide significance is required for distal-eQTL because there is not additional strong biological information to support reduced stringency as is the case with local-eQTL. As with mediation through chromatin state, 1,000 permutations are used to derive genome-wide mediation p-values, $\text{permP}_{\text{G}}^m$.

  Mediation of cQTL through other chromatin sites and gene expression were also considered, but did not result in the detection of any candidate mediators other than *Zfp985* expression as a mediator of a distal-cQTL on chromosome 4 for a chromatin region located chromosome 13 near *Akr1e1*.

Mediation steps

There are important caveats and pitfalls for mediation analysis compared to QTL mapping, in part due to the fact that the relationships 1 and 2 could be complex, whereas the directionality of the relationships for QTL is fixed because the genetic state at the QTL cannot change because of the trait. This does not hold for the mediation model, where Y could in fact causally influence M (Y → M). False mediation can be detected when the association between X and M are particularly strong, whereby M acts as a surrogate for X [3]. Based on concerns over fitting relatively complex models to data comprising 47 mice, we use stringent requirements for declaring evidence for mediation:

1. Detection of a significant eQTL (relationship 1).

   (a) Mediation through chromatin: $\text{permP}_{\text{C}}^m < 0.05$ from Analysis L
   (b) Mediation through expression: $\text{permP}_{\text{G}}^m < 0.05$ from Analysis G

2. Detection of significant mediator within a ±10Mb window around the eQTL from previous step.

3. Detection of a significant QTL for the mediator from previous step.

   (a) Mediation through chromatin: $\text{permP}_{\text{C}}^m < 0.05$ from Analysis L and within ±10Mb of gene TSS
   (b) Mediation through expression: $\text{permP}_{\text{G}}^m < 0.05$ from Analysis G and within ±10Mb of the eQTL position

4. It is not possible to distinguish M → Y from M ← Y given these data. If M is an intermediary of X on Y, the expectation is that the effect of X → M is greater than X → Y. We check that the effect size of the QTL for step 1 (X → Y) is less than the effect size of the QTL from step 3 (X → M). Failure of this step does not disprove the proposed mediation model, but it could suggest that the relationship between M and Y is more complicated than the simple models tested here, or that greater noise is incurred in the measurement of M than Y.

  Note: Step 4 is an approximation of testing whether M provides more information about Y than X, which in theory, could be more directly tested with the partial mediation test (relationship 6). However, we found this test was inconclusive (partial mediation p-value < 0.1) for the clear example of *Zfp985* expression mediating the distal-eQTL of *Akr1e1*, suggesting the error heterogeneity across the measurements of these molecular phenotypes may be too great to allow for full use of traditional mediation tests, such as distinguishing partial mediation, full mediation, and X affecting M and Y separately.

References

[1] Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986;51(6):1173–1182. doi:10.1037/0022-3514.51.6.1173.

[2] MacKinnon DP, Fairchild AJ, Fritz MS. Mediation Analysis. Annual Review of Psychology. 2007;58(1):593–614. doi:10.1146/annurev.psych.58.110405.085542.

[3] Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Statistical Methods in Medical Research. 2007;16(4):309–30. doi:10.1177/0962280206077743.

[4] Hamilton-Williams EE, Rainbow DB, Cheung J, Christensen M, Lyons PA, Peterson LB, et al. Fine mapping of type 1 diabetes regions Idd9.1 and Idd9.2 reveals genetic complexity. Mammalian Genome. 2013;24(9-10):358–75. doi:10.1007/s00335-013-9466-y.