Supplementary Material of the manuscript

"Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays"

June 20, 2014

1 Details of priors for Bayes Factor calculations, and quantile transformation

We use the Bayes Factor calculations from Servin and Stephens (2007), which are based on the following model and priors. The model is:

$$y^i_{sl} = \mu_{sl} + \gamma_{sl}\beta_{sl}g^i + \epsilon^i_{sl} \quad \text{with} \quad \epsilon^i_{sl} \sim \mathcal{N}(0, \sigma^2_{sl}). \tag{1}$$

The priors are:

$$\begin{aligned} \sigma^2_{sl} &\sim \Gamma^{-1}(\kappa^a_{sl}, \kappa^b_{sl}), & (2)\\ \mu_{sl}|\sigma^2_{sl} &\sim N(0, \sigma^2_{\mu,sl}\sigma^2_{sl}), & (3)\\ \beta_{sl}|\sigma^2_{sl} &\sim N(0, \sigma^2_{\beta,sl}\sigma^2_{sl}) & (4) \end{aligned}$$

and a discrete uniform prior on $\sigma_{\beta,sl} \in \{0.05, 0.1, 0.2, 0.4\}$ (values which were chosen to span a wide range of potential effect sizes from very small to moderately large). As in Servin and Stephens (2007), we use the limiting Bayes Factor obtained in the limits $\kappa^a_{sl}, \kappa^b_{sl} \to 0$, and $\sigma^2_{\mu,sl} \to \infty$. In addition, when estimating effect sizes, we use the posterior distribution on β_{sl} in the same limit. All posteriors we compute here are proper in these limits; see Servin and Stephens (2007) for discussion.

To quantile transform the vector y_{sl} to the quantiles of a normal distribution, with ties broken at random, we use the following R function:

```
QT_randomTie <- function(x) {
x.rank = rank(x, ties.method="random")
return(qqnorm(x.rank,plot.it = F)$x)
}
```

2 Posterior distribution on effect sizes

2.1 Posterior distribution of effect size on WC at scale s and location l when $\gamma_{sl} = 0$, $\mathsf{P}(\beta_{sl} \mid \gamma_{sl} = 0, y_{sl}, g)$

In this section we drop the subscript sl for convenience. As derived in Protocol S1 in Supporting Information of Servin and Stephens (2007),

$$\begin{aligned} \sigma^2 \mid y, g &\sim \Gamma^{-1}(\frac{N+2\kappa^a}{2}, \frac{M}{2}), & (5)\\ (\mu, \beta) \mid y, g, \sigma^2 &\sim N(\mathbf{B}, \sigma^2\mathbf{\Omega}), & (6) \end{aligned}$$

where

$$\begin{aligned} M &= y^t y - \mathbf{B}^t\mathbf{\Omega}^{-1}\mathbf{B} + 2\kappa^b & (7)\\ \mathbf{B} &= \mathbf{\Omega}\mathbf{X}^t y & (8)\\ \mathbf{\Omega} &= (\mathbf{D}^{-1} + \mathbf{X}^t\mathbf{X})^{-1} & (9)\\ \mathbf{D} &= diag(\sigma_\mu^2, \sigma_\beta^2) & (10) \end{aligned}$$

and $\mathbf{X}$ has two columns, the first column being a vector of all ones and the second column being a vector of genotypes, g. Then,

$$\beta \mid y, g, \sigma^2 \sim N(\mathbf{B}_2, \sigma^2\mathbf{\Omega}_{22}), \quad (11)$$

where $\mathbf{B}_2$ denotes the second element of $\mathbf{B}$ and $\mathbf{\Omega}_{22}$ denotes (2,2)th element of $\mathbf{\Omega}$.

The posterior on effect size, $\mathsf{P}(\beta \mid \gamma = 0, y, g)$, can be written as

$$\begin{aligned} &\int\int \mathsf{P}(\mu, \beta \mid y, g, \sigma^2)\mathsf{P}(\sigma^2 \mid y, g)\mathrm{d}\mu\mathrm{d}\sigma^2 & (12)\\ &= \int \mathsf{P}(\beta \mid y, g, \sigma^2)\mathsf{P}(\sigma^2 \mid y, g)\mathrm{d}\sigma^2. & (13) \end{aligned}$$

and

$$\begin{aligned} &\propto \int [\sigma^2]^{-\frac{1}{2}} \exp[-\frac{(\beta - \mathbf{B}_2)^2}{2\sigma^2\mathbf{\Omega}_{22}}][\sigma^2]^{-\frac{N+2\kappa^a}{2}-1} \exp[-\frac{M}{2\sigma^2}]\mathrm{d}\sigma^2 & (14)\\ &\propto [\frac{(\beta - \mathbf{B}_2)^2}{\mathbf{\Omega}_{22}} + M]^{-\frac{N+2\kappa^a+1}{2}} & (15)\\ &\propto [1 + \frac{(N+2\kappa^a)(\beta - \mathbf{B}_2)^2}{(N+2\kappa^a)\mathbf{\Omega}_{22}M}]^{-\frac{N+2\kappa^a+1}{2}}. & (16) \end{aligned}$$

Taking the limit $\kappa^a, \kappa^b \to 0$, and $\sigma_\mu^2 \to \infty$ yields the limiting posterior

$$\propto [1 + \frac{N(\beta - \mathbf{B}_2^*)^2}{N\mathbf{\Omega}_{22}^*(y^t y - (\mathbf{B}^*)^t(\mathbf{\Omega}^*)^{-1}\mathbf{B}^*)}]^{-\frac{N+1}{2}}, \quad (17)$$

where $\mathbf{B}^*$ and $\mathbf{\Omega}^*$ are obtained in the limit. Then, the limiting posterior, $\mathsf{P}(\beta \mid \gamma = 0, y, g)$ is a three parameter version of a t distribution (Jackman, 2009) with density

$$p(x|\nu, a, b) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu b}} \left(1 + \frac{1}{\nu}\frac{(x-a)^2}{b}\right)^{-\frac{\nu+1}{2}} \tag{18}$$

where

$$\begin{aligned} a &= \mathbf{B}_2^* && (19)\\ b &= \frac{\mathbf{\Omega}_{22}^*(y^t y - (\mathbf{B}^*)^t(\mathbf{\Omega}^*)^{-1}\mathbf{B}^*)}{N} && (20)\\ \nu &= N && (21) \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}(x) &= a \quad \text{for} \quad \nu > 1 && (22)\\ \mathrm{Var}(x) &= \frac{b\nu}{\nu - 2} \quad \text{for} \quad \nu > 2. && (23) \end{aligned}$$

2.2 Posterior distribution of effect size on WC at scale s and location l, $\mathsf{P}(\beta_{sl} \mid y_{sl}, g)$

The posterior distribution of the effect size on WC at scale s and location l, $\mathsf{P}(\beta_{sl} \mid y_{sl}, g)$, can be written as

$$\mathsf{P}(\beta_{sl} \mid \gamma_{sl} = 1, y_{sl}, g)\phi_{sl} + \mathsf{P}(\beta_{sl} \mid \gamma_{sl} = 0, y_{sl}, g)(1 - \phi_{sl}), \tag{24}$$

where $\mathsf{P}(\beta_{sl} \mid \gamma_{sl} = 0, y_{sl}, g)$ is a point mass at zero and ϕ_{sl} is

$$\begin{aligned} \phi_{sl} &:= \mathsf{P}(\gamma_{sl} = 1 \mid \hat{\pi}, y_{sl}, g) && (25)\\ &= \frac{\mathsf{P}(y_{sl} \mid g, \gamma_{sl} = 1)\mathsf{P}(\gamma_{sl} = 1 \mid \hat{\pi})}{\mathsf{P}(y_{sl} \mid g, \gamma_{sl} = 1)\mathsf{P}(\gamma_{sl} = 1 \mid \hat{\pi}) + \mathsf{P}(y_{sl} \mid g, \gamma_{sl} = 0)\mathsf{P}(\gamma_{sl} = 0 \mid \hat{\pi})} && (26)\\ &= \frac{\hat{\pi}_s \mathrm{BF}_{\mathrm{sl}}}{\hat{\pi}_s \mathrm{BF}_{\mathrm{sl}} + (1 - \hat{\pi}_{\mathrm{s}})}. && (27) \end{aligned}$$

Therefore, the posterior distribution of the effect size on WC at scale s and location l is a mixture of a point mass at zero and a three parameter version of a t distribution with a location parameter a_{sl}, a squared of scale parameter b_{sl}, and the number of degrees of freedom ν_{sl} (see Section 2.1). The three parameter t distribution mixture proportion is ϕ_{sl}.

2.3 Effect size in the original space

From $\alpha = W^{-1}\beta$, and noting that W is orthogonal so $W^{-1} = W'$, the posterior mean and variance of α are given by

$$\mathrm{E}(\alpha_b) = \sum_{s,l} a_{sl}\phi_{sl}w_{sl,b}, \tag{28}$$

$$\mathrm{Var}(\alpha_b) = \sum_{s,l}[v_{sl} + a_{sl}^2 - a_{sl}^2\phi_{sl}]\phi_{sl}w_{sl,b}^2, \tag{29}$$

$$v_{sl} = \frac{b_{sl}\nu_{sl}}{\nu_{sl} - 2}. \tag{30}$$

where $w_{sl,b}$ is the element of the DWT matrix in row corresponding to scale s and location l, and the column corresponding to base b.

3 Selection of the top 1% of 1024bp sites with the highest DNase I sensitivity

We focus our association analysis on the top 1% of 1024bp sites with the highest DNase I sensitivity (146,435 sites) that are selected by the following procedure. As in Degner *et al.* (2012), we divide the whole genome into non-overlapping 100bp windows and rank them according to a DNase I sensitivity (for the definition of DNase I sensitivity, see Supplementary Material of Degner *et al.* (2012)). Then, we select the top 1% of 100bp windows with the highest DNase I sensitivity. We merge those 100bp windows if they are adjacent to each other and if the length of the merged window is less than 1024bp, leading to 146,435 windows. The 146,435 1024bp sites used in our analysis are centered at those windows.

4 Sensitivity analysis for low count threshold and site size

To assess robustness of our method to choice of low count threshold L and site size B, we reanalyzed the 50,000 sites considered for the analysis in Section 4.3 of the main text using different values of L and B. For each choice of L and B we computed the number of significant sites at a given FDR (as in Section 4.3 in the main text).

4.1 Sensitivity to low count threshold L

We considered $L = 0, 1, 2, 3, 5, 10$, where $L = 0$ filtered only WCs that have no reads in the region being contrasted for all individuals. Other settings (e.g., site size, number of

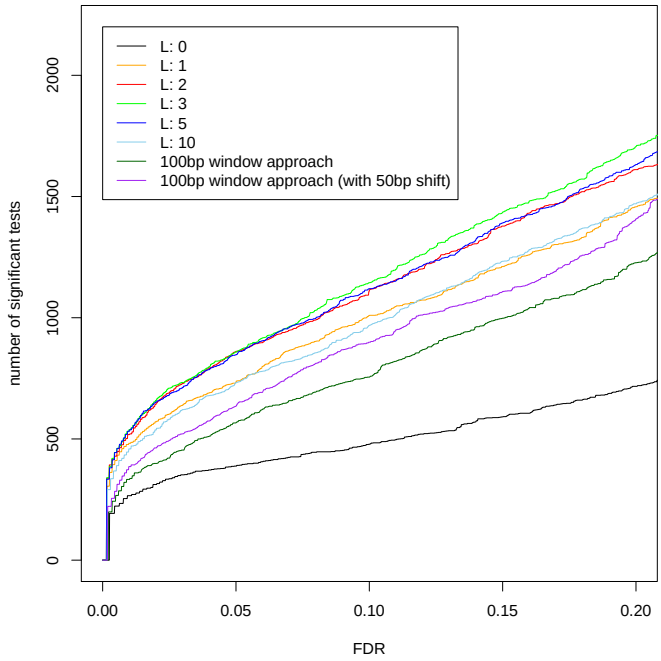


Figure 1: **Sensitivity analysis for low count threshold**

PCs) are as in the main text, so $L = 2$ corresponds to the analysis in the main text. Figure 1 shows the number of significant sites at a given FDR for different values of L. Results are almost identical for $L = 2, 3, 5$ (results from $L = 3$ being very slightly better). Results for $L = 1, 10$ are noticeably less good, but still better than the 100bp window approach (with or without 50bp shift). Results for $L = 0$ show poor performance, worse than the 100bp window approach. This analysis shows that some filtering is necessary to remove noise, but that results are relatively robust to choice of that threshold.

4.2 Effect of choice of site size, B

In our main text we analyzed sites of size $B = 1024$. This decision was based on observations in Degner *et al.* (2012) that typical dsQTLs affect chromatin accessibility over roughly 200-300bp, and so we chose a slightly bigger region for our analysis. Here we assess the effect of using sites that are half the size ($B = 512$) or twice the size ($B = 2048$). For each of the 50,000 1024bp sites, we selected a 512bp site and a 2048bp site with the same center as the 1024bp site. We then repeated the analyses for each value of $B = 512, 1024, 2048$ using the same settings (and the same set of candidate SNPs) as we used in the main text, so the analysis with 1024bp is the same as the analysis in the main text.

Figure 2 shows the number of significant sites at a given FDR for the three analyses.

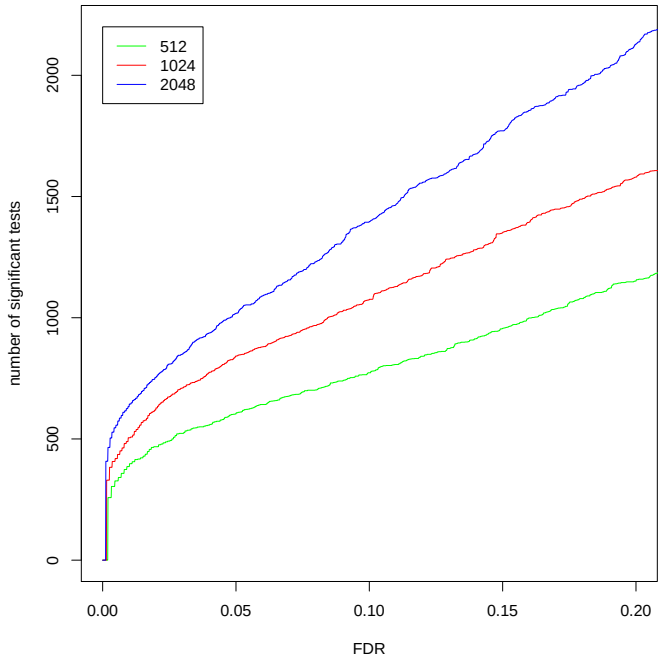


Figure 2: **Sensitivity analysis for site size**

Results shows that larger sites increased power to detect dsQTLs. At FDR=1%, the analyses with 512bp, 1024bp, and 2048bp detect 386, 505, and 644 dsQTLs, respectively. Among the 386 dsQTLs identified by 512bp analysis, 354 (92%) and 340 (88%) are detected by 1024bp and 2048bp analyses, respectively. Among the 505 dsQTLs identified by 1024bp analysis, 462 (91%) are detected by 2048bp analysis. The increased power for larger B presumably reflects the fact that larger sites contain additional signals, while at the same time the increased site size does not substantially "dilute" the signals that are also contained in the smaller sites. This reflects the multi-scale nature of the wavelet method: even when a site of $B = 2048$ is used, it does not substantially reduce power to detect signals that are contained in the central 512bp section compared with using $B = 512$.

Although the larger site size detected more signals in this analysis, choice of larger site size also involves some trade-offs. First, analysis with larger sites is more computationally intensive (computation grows approximately linearly with site size, for a given number of permutations). Second, a significant result with a larger site size provides less information: the larger the site one tests, the larger the uncertainty in the location of any detected effect. At the extreme one could test the whole genome as a "site", but a significant results would leave you wondering which part of the genome was driving the result. Thus choice of site size in a particular application may depend on how the analyst weights these trade-offs, and not only on which choice provides the greatest power.

5 Comparisons of methods on simulated data

To supplement our empirical comparisons of methods on real data in the main text, we also performed comparisons on simulated data.

To try to make the simulations as realistic as possible, we based our simulations on estimated effect sizes from the 578 dsQTLs identified by either or both of the wavelet-based or 100bp window approach at FDR=0.01 (from analyses in Section 4.3 of the main text). Each dsQTL consists of DNase-seq count data over 1024bp and genotypes at the most strongly associated genetic variant on 70 individuals. (When the most strongly associated genetic variants from the wavelet-based and window-based approaches were different, we used the one from the analysis with the stronger p-value.) To avoid biasing the estimated effect sizes towards one approach or other we used a third method, based on a sliding window, to estimate the effect sizes for simulation (details in Section 5.2.1).

For each dsQTL, we simulated two data sets (null and alternative) with two groups of individuals (40 vs 30), with the procedure described in Section 5.2.3. We then applied each method (wavelet-based and 100bp window-based methods) to the 1156 data sets (578 null and 578 alternative data sets), obtained a p-value by permutation for each data set from each method, and assessed the performance of each method by plotting the receiver-operating characteristic (ROC) curve for each method as the p value threshold varies. Because these simulations were based on the 578 strongest dsQTLs we also conducted a second set of simulations with a 50% reduction in the simulated effect size (details in Section 5.2.2), to compare performances on slightly weaker effects. The results (Figure 3) showed that, similar to our empirical comparisons, the wavelet-based method has higher average power than the 100bp window based at any given p value threshold (for both the stronger and weaker effects).

5.1 Simulations stratified by different types of effect

To provide more insights into which settings the wavelet-based approach produces the largest gain in power, we conducted further simulations, stratified by three different types of dsQTL effect identified in the main text (Figure 2, Figure 3(a), and Figure 4 of the main text). The first of these represents a typical dsQTL identified by both methods, whereas the latter two represent dsQTLs that are found by the wavelet-based approach, but not by the window-based approach.

For each of the three dsQTLs, we estimated effect size using a sliding window approach (Section 5.2.1) and simulated 200 null and 200 alternative data sets with two groups of individuals (40 vs 30) (see Section 5.2.3). We then applied the wavelet-based and 100bp-window-based methods to the 400 simulated data sets (200 null and 200 alternative data sets) of each type, computed a p value from each method by permutation, and compared performance using an ROC curve.

Figure 4 (a)-(c) show the results of the three simulations. In the first simulation,

(a) Original Effects

(b) Weakened Effects (reduced by 50%)

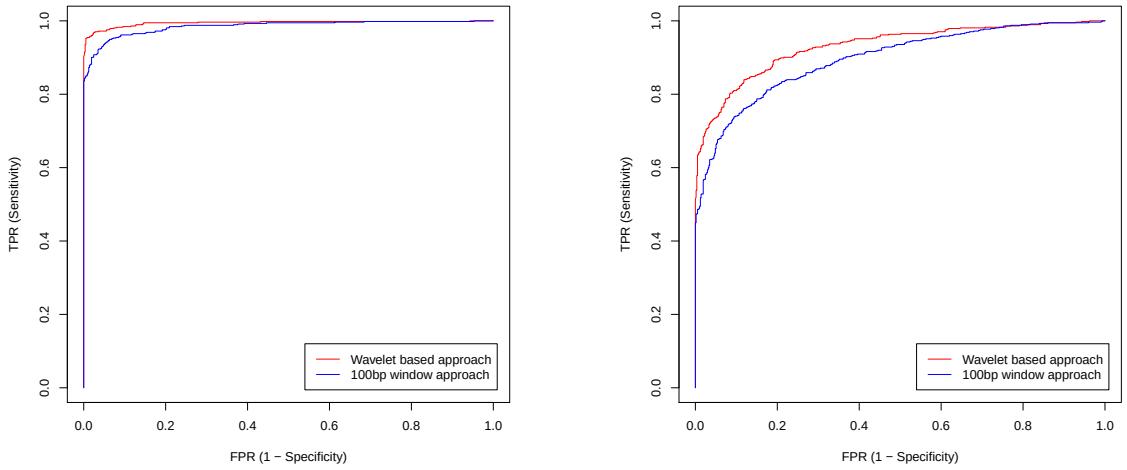


Figure 3: **Results for simulations based on top 578 dsQTLs: the wavelet-based approach shows higher power than the 100bp window based approach at all thresholds for both the original effects (a) and the weakened effects (b).**

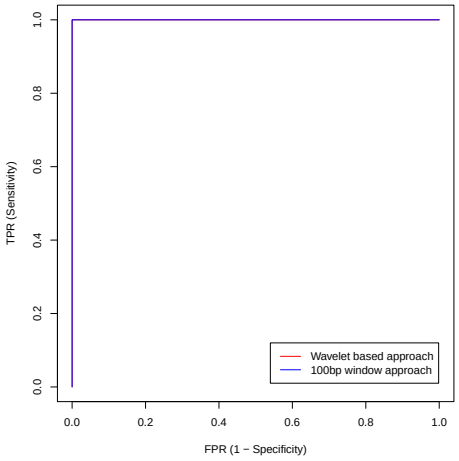
which was based on a dsQTL identified by both methods, both methods perform perfectly (100% power). In the other settings, as expected, the wavelet-based method outperforms the window approach.

In the main text, we observed that a pattern of effect, qualitatively similar to the typical dsQTL identified by both methods, can be missed by the 100bp window approach if signals are modest (see Figure 3(b) in the main text). To further assess this we repeated the first set of simulations described above, based on the dsQTL found by both methods, but with a reduced effect size (reduced to 50% of the original estimated effect size; see Section 5.2.1). The ROC curves for this simulation, in Figure 4(d), demonstrate that even for this dsQTL that was detected by both methods, the wavelet based method has higher power than the window based methods to detect a more modest effect with the same shape.

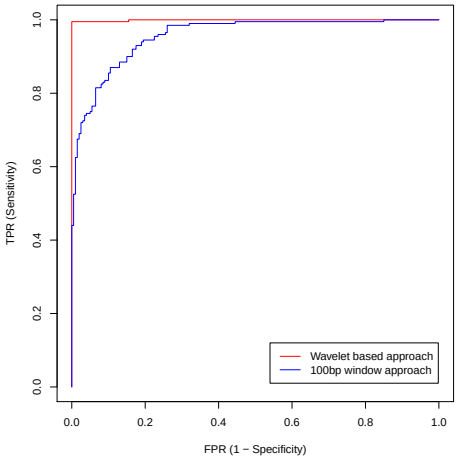
5.2 Simulation Details

To simplify the simulations we consider only two groups of individuals (major homozygotes and heterozygotes). To estimate the effect size, in brief we use a sliding window to estimate a smooth mean curve for each group and then estimate an effect size at each base as the ratio of these two means. We then simulate read data at each base by thinning the empirical data by an amount determined by the effect size, so that the expected ratio of the means in the simulated data matches the estimated effect size. The two procedures are detailed fully in the next subsections.

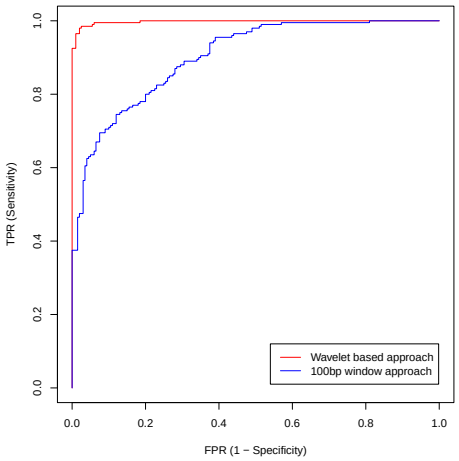
(a) chr17:10160989-10162012



(b) chr12:6264339-6265362



(c) chr2:110326889-110327912



(d) chr17:10160989-10162012 (50% reduced strength)

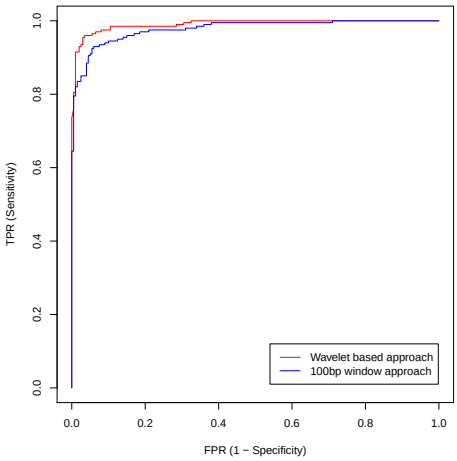


Figure 4: **Comparisons on simulations based on three different types of dsQTL identified in main text: the wavelet-based approach outperforms the 100bp window based approach in all cases except (a) where both methods have 100% power to detect effects.**

5.2.1 Estimation of effect size over a 1024bp region from dsQTL data

For the major homozygote individuals (M), we calculate the average read count across those individuals at each position b, as

$$\bar{c_b}^M = \max\left(\frac{1}{70}, \quad \frac{\sum_{j\in M} c_b^j}{\#\{j : j \in M\}}\right) \quad \text{for} \quad b = 1, \ldots, 1024, \tag{31}$$

where c_b^j is read count at position b for individual j. To prevent the estimated effect size from being infinite, we use $1/70$ for positions with no reads for all major homozygotes. Then, we smooth the average read counts using a simple sliding window approach (21bp as a window size), leading to a smoothed mean curve,

$$s_b^M = \frac{\sum_{t\in W_b} \bar{c_t}^M}{\#\{t : t \in W_b\}} \quad \text{for} \quad b = 1, \ldots, 1024, \tag{32}$$

where $W_b = \{w : \max(1, b-10) \leq w \leq \min(1024, b+10)\}$.

Similarly, we also compute a smoothed mean curve for heterozygotes individuals (H) s_b^H using the same procedure. We then define the effect size at position b as the ratio $r_b = s_b^H / s_b^M$, and this ratio is used in simulating data below.

In general we would not advocate using the simple sliding window to estimate the effect size - we would expect other more sophisticated methods, including our wavelet-based method, to provide a more accurate estimate. However, we used a sliding window here to avoid using a "multi-scale" method for effect estimation that might produce an advantage to the wavelet-based approach in subsequent comparisons.

5.2.2 Decrease/increase the estimated effect size

We can control (decrease or increase) the strength of the estimated effect size while maintaining the pattern of the effect as follows. For each b, the estimated effect size, r_b, can be written as

$$r_b = 1 + (r_b - 1) \tag{33}$$

$$= \frac{s_b^M + (s_b^H - s_b^M)}{s_b^M}, \tag{34}$$

where $s_b^H - s_b^M$ indicates a difference in signal between two groups. We can rescale this difference without changing its sign by multiplying by a positive number c (decrease strength if $c < 1$), yielding a rescaled effect size

$$r_b^c = \frac{s_b^M + c(s_b^H - s_b^M)}{s_b^M}. \tag{35}$$

$$= 1 + c(r_b - 1) \tag{36}$$

$$= 1 + cr_b - c. \tag{37}$$

We used this approach (with $c = 0.5$) to obtain the simulation results in Figure 3(b) and Figure 4(d).

5.2.3 Simulating reads by thinning real data

Given the effect size, $r_1, \ldots, r_B$, we simulate null/alternative data sets with N individuals and two groups (group M and group H) as follows. Let X_b^i denote the number of reads in sample i at base b; we wish to simulate X_b^i for $i = 1, \ldots, N$ and $b = 1, \ldots, B$. We perform this simulation by "thinning" the real data reads at base b, totaled across all individuals (c_b^+) by an individual-specific and base-specific thinning probability p_b^i. That is,

$$X_b^i \sim \text{Binom}(c_b^+, p_b^i) \quad \text{where } c_b^+ := \sum_{j=1}^{70} c_b^j, \quad \text{for} \quad b = 1, \ldots, B, \tag{38}$$

where c_b^j denotes read count at position b for individual j in the DNase-seq data. We simulate alternative data with an effect using different thinning probabilities for the two groups M and H: p^M for group M and p^H for group H. Then, X^i, for individual i has expectation

$$= \begin{cases} c_b^+ p_b^M & \text{if individual } i \text{ is in group } M, \\ c_b^+ p_b^H & \text{if individual } i \text{ is in group } H. \end{cases} \tag{39}$$

Here, we used $p_b^M = 1/70$, and $p_b^H = r_b p_b^M$. This ensures that the expectation for group M matches the empirical average from the real data, and that the ratio of the expectations for the two groups at base b is the effect size r_b as desired. (In the rare cases that $r_b p_b^M > 1$, we use $p_b^H = 1$.) To simulate null data we follow the same procedure but with effect $r_b = 1$ for all b.

6 Permutation procedure

To reduce computing time for permutation analysis we adapted the sequential procedure from BESAG and CLIFFORD (1991). The algorithm is provided in Algorithm 1. The key idea is that the procedure performs permutations until the number of permuted data sets with test statistic greater than or equal to the observed test statistic reaches a given value (10 in our analysis) [or the number of permutations reaches a predefined maximum – 100,000 in our analysis]. This procedure ensures that, for data sets with no signal, relatively few permutations are performed (because we do not need extremely accurate estimates of p values for non-significant data sets). The only additional issue this raises is that, for insignificant p values, there is some "discretization effect" due to the small number of permutations. This can cause problems for the `qvalue` package, which uses the insignificant p values to estimate the proportion of nulls that are true. To solve this problem we use randomization to produce continuous p values that are

uniformly distributed under the null. For example, if the first 10 permutations are all more significant than the observed data, the algorithm generates a p value uniformly between 11/12 and 1, rather than simply using 1 every time (see Algorithm for details).

Algorithm 1 A sequential permutation procedure to obtain p-value (D, $\hat{\Lambda}_{\max}$, max.numPerm, max.numSig)

Input: data set (D); observed test statistic ($\hat{\Lambda}_{\max}$); the maximum number of permutations (max.numPerm); the maximum number of permuted data sets with test statistic greater than or equal to the observed test statistic (max.numSig).
Output: p-value (p-value); the number of permutations performed (numPerm); the number of permuted data sets with test statistic greater than or equal to the observed test statistic when the procedure stops (numSig).

$i \leftarrow 0$.
numPerm $\leftarrow$ NA.
for $p \leftarrow 1$ to max.numPerm **do**
  Permute individual labels in genotypes of data set D.
  Compute a test statistic $\hat{\Lambda}^p_{\max}$.
  if $\hat{\Lambda}^p_{\max} \geq \hat{\Lambda}_{\max}$ **then**
    $i \leftarrow i + 1$.
  end if
  if $i =$ max.numSig **then**
    numSig $\leftarrow i$.
    numPerm $\leftarrow p$.
    Sample p-value from Uniform($\frac{\text{numSig} + 1}{\text{numPerm}+2}$, $\frac{\text{numSig} + 1}{\text{numPerm}+1}$).
    break.
  end if
end for
if numPerm = NA **then**
  numSig $\leftarrow$ i.
  numPerm $\leftarrow$ max.numPerm.
  p-value $\leftarrow \frac{\text{numSig} + 1}{\text{numPerm}+1}$.
end if
return p-value, numPerm, and numSig.

References

BESAG, J. and CLIFFORD, P. (1991) Sequential Monte Carlo p -values. *Biometrika*, **78**, 301–304.

Degner, J. F., Pai, A. a., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y. and Pritchard, J. K. (2012) DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–4.

Jackman, S. (2009) *Bayesian Analysis for the Social Sciences*. Wiley.

Servin, B. and Stephens, M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, **3**, e114.