Supplemental Text, Figures & Tables for the Research Article:

"Inferring the Allelic Series at QTL in Multiparental Populations"

by
Wesley L. Crouse, Samir N.P. Kelada and William Valdar

October 15, 2020

Supplemental Text: Alternative Prior Distributions for the Allelic Series

Alternatives Considered

In the main text, we focused on a specific prior distribution for the CRP concentration parameter, comparing it via simulation with the full haplotype-based association approach and evaluating the relative benefit of including tree information. In this supplement, we consider an alternative prior distribution for the concentration parameter, as well as an alternative prior distribution for the allelic series. The prior distributions considered in this supplement are:

- the one used in the main text, which is the CRP-based model with a relatively-conservative exponential prior distribution on the concentration parameter (termed the "Exponential" model here),

$$\begin{aligned} \mathbf{M}|\alpha &\sim \text{CRP}(\alpha) \\ \alpha &\sim \text{Ga}(a_\alpha = 1, b_\alpha \approx 2.33). \end{aligned}$$

  With $J = 8$ possible haplotypes, this prior distribution corresponds to a 50% probability of $K = 1$ functional allele, and monotonically favors smaller numbers of functional alleles *a priori*.

- the CRP-based model, but with a weakly informative gamma prior distribution on the concentration parameter (the "Gamma" model):

$$\begin{aligned} \mathbf{M}|\alpha &\sim \text{CRP}(\alpha) \\ \alpha &\sim \text{Ga}(a_\alpha \approx 2.30, b_\alpha \approx 0.75), \end{aligned}$$

  With $J = 8$ possible haplotypes, this prior distribution corresponds to a 5% probability of the null model with $K = 1$ functional allele and a 1% probability of the full model with $K = 8$ functional alleles.

- and a uniform prior which assumes that all configurations of the allelic-series are *a priori* equally likely (the "Uniform" model):

$$\text{p}(\mathbf{M}) \propto 1.$$

  This is implemented as a non-exchangeable uniform process prior (Wallach *et al.* 2008).

These alternative prior distributions are shown in **Figure S1**. We evaluated these alternatives using the simulation procedure described in the main text.

Simulation Results

Figure S2 shows the 0-1 accuracy of the MAP allelic series under the Uniform, Gamma and Exponential alternatives, for different numbers of true functional alleles and effect sizes. In the low power scenario (**A**), the ability to detect multiallelic series is low, but the two CRP-based approaches, Exponential and Gamma, have high accuracy when the QTL has only $K = 1$ functional allele (the null model), or when the QTL is biallelic. In the high power scenario (**B**), the CRP-based approaches have reasonably high accuracy (approximately 80%) for up to $K = 4$ functional alleles, with the Exponential outperforming the Gamma through this range. The more-diffuse Gamma prior maintains some limited accuracy through $K = 8$ alleles, outperforming the Exponential, although accuracy for highly-multiallelic series is generally low. In both scenarios, the Uniform approach is worse than the Exponential and Gamma, except when there is an intermediate number of alleles.

Figure S3 is similar to the previous figure but shows the posterior probability of the correct allelic series, rather than the accuracy of the MAP allelic series. The Gamma and Uniform priors have relatively low certainty across both power scenarios and for all true numbers of functional alleles. In contrast, the Exponential prior is decisive when the true number of alleles is low, but at the expense of accuracy when the true number of alleles is high. Notably, in the high power scenario (**B**), the Gamma prior has reduced certainty when the QTL is null relative to biallelic, suggesting it has a tendency to overestimate the number of alleles under the null.

Figure S4 is also similar to the previous figures but shows the posterior expectation of the number of alleles. In the low power scenario (**A**), the expectations for both the Gamma and Uniform are insensitive to the true number of alleles, consistently reporting an intermediate number of functional alleles. In contrast, the Exponential is accurate when the true number of alleles is one or two, but it reports approximately three alleles when the true number of alleles is higher. In the high power scenario (**B**), all the alternatives are more sensitive to the true number of alleles. In particular, the Exponential approaches the correct expectation for as many as $K = 5$ alleles. Notably, both the Exponential and Uniform show a tendency to overestimate when the QTL is null. We hypothesize that this is due to their relatively fat tails with respect to prior number of functional alleles, and that this acts in combination with a prior for QTL effect size that can accommodate small effects. When the true number of alleles is one, the QTL effect is necessarily zero, and it becomes "easier" for these permissive allelic series priors to estimate many effects, each of very small size.

Figure S5 shows the MSE of haplotype effect estimates for the alternative allelic series prior distributions. In the low power scenario (**A**), the allele-based approaches outperform the Full haplotype-based approach when the true number of alleles is small. When power is high (**B**), this is also true when there are intermediate numbers of alleles. However, in both scenarios, when the true number of alleles is high, the Full model is better than the allele-based approach. As with accuracy, the Exponential is relatively better than Gamma and Uniform when the true number of alleles is low, and it is relatively worse when the true number of is high.

Discussion

On the basis of these results, we recommend using the Exponential prior distribution for the allelic series. In many applied cases it will be reasonable to expect that QTL have only a few functional alleles, and when this is the case, the Exponential consistently outperforms than the other alternatives. In particular, when the number of alleles is small, the Exponential is less likely to overestimate the number of alleles than the Gamma or Uniform, and haplotype effect estimates are improved. Additionally, the Exponential prior typically provides a decisive posterior in the case of biallelic QTL. For these reasons, we focused on the Exponential prior distribution for the simulations in the main text.

Supplemental Figures

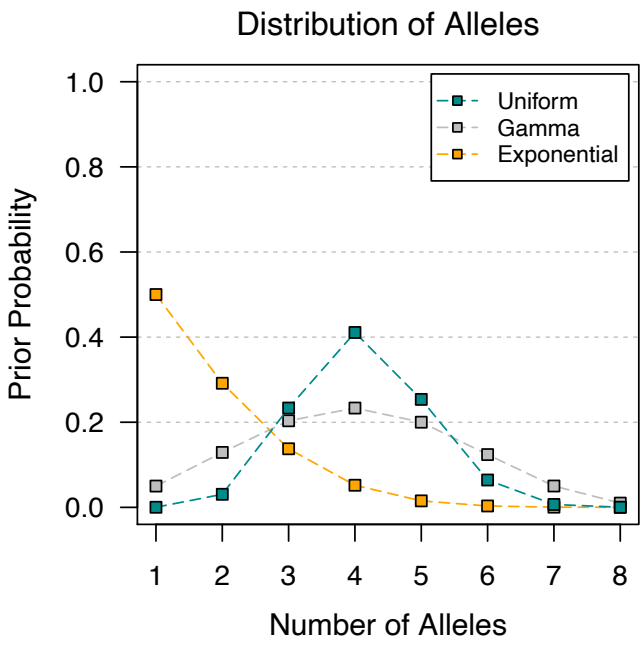

Figure S1 Prior distribution of number of functional alleles for the Uniform, Gamma, and Exponential prior distributions. Points are connected for clarity. Uniform places high prior weight on an intermediate number of functional alleles. Gamma is less informative than the Uniform with fatter tails. Exponential favors smaller numbers of alleles than the others.

.

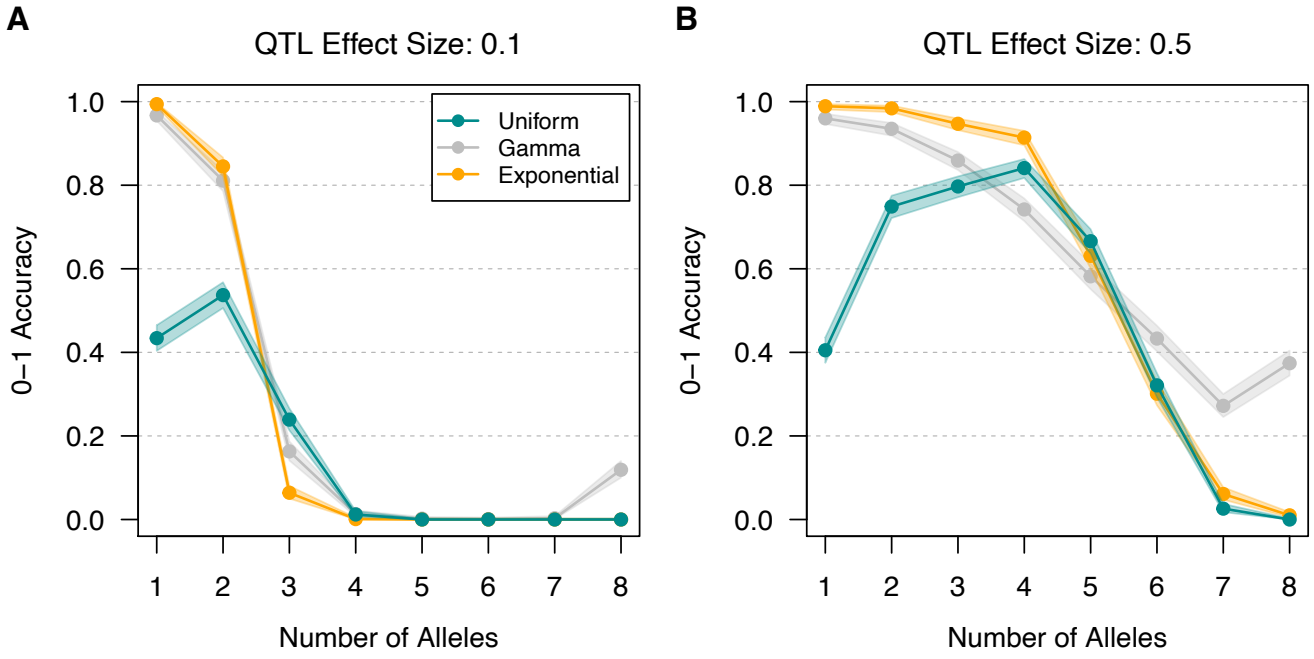

Figure S2 0-1 accuracy of posterior allelic series inference using the Uniform, Gamma, and Exponential prior distributions, for varying numbers of true functional alleles, across two effect sizes. Points are connected for clarity. Shading denotes 95% confidence intervals. In the low power scenario (**A**), accuracy for Exponential and Full is high when the QTL has two or fewer alleles but low when it is multiallelic. In the high power scenario (**B**), Exponential and Full have reasonable accuracy for an intermediate number of alleles, with the Exponential outperforming the Gamma through this range. Gamma maintains some accuracy for highly multiallelic series, outperforming the Exponential, although accuracy for highly-multiallelic series is generally low. Across both scenarios, Uniform is worse than Exponential and Gamma, except for an intermediate number of alleles.

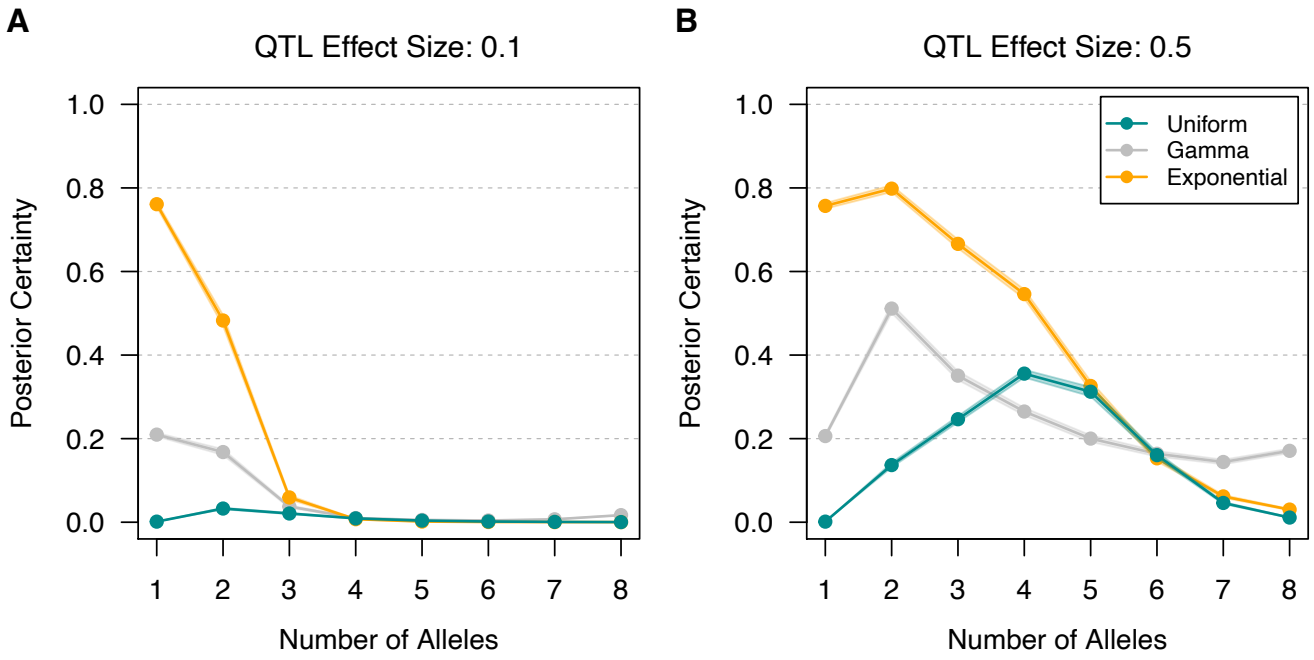

Figure S3 Posterior certainty of the correct allelic allelic series using the Uniform, Gamma, and Exponential prior distributions, for varying numbers of true functional alleles, across two effect sizes. Points are connected for clarity. Shading denotes 95% confidence intervals. In the low power scenario (**A**), posterior certainty for Exponential is high when the QTL has two or fewer alleles but low when it is multiallelic. Posterior certainty for Gamma and Uniform is low for any number of alleles. In the high power scenario (**B**), Exponential has higher posterior certainty for an intermediate number of alleles than Gamma and Uniform. Gamma has the highest posterior certainty when the QTL is highly multiallelic, but posterior certainty is low.

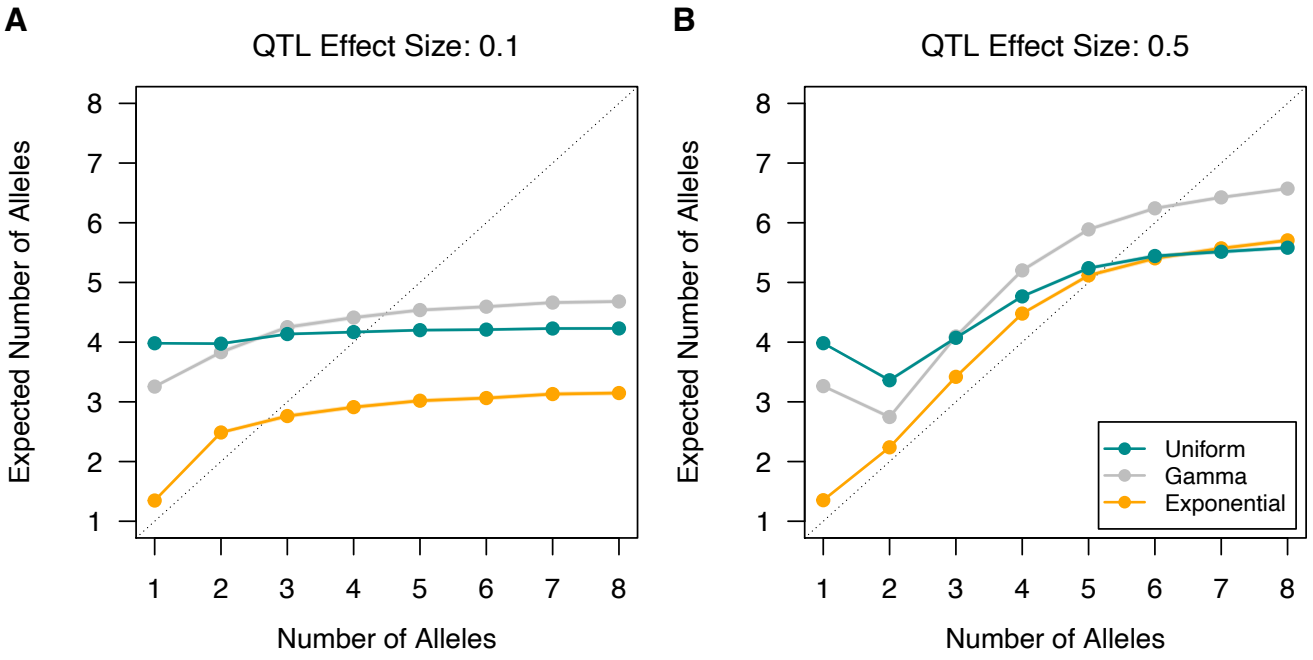

Figure S4 Posterior expectation of number of alleles using the Uniform, Gamma, and Exponential prior distributions, for varying numbers of true functional alleles, across two effect sizes. The dotted line indicates the correct expectation. Points are connected for clarity. Shading denotes 95% confidence intervals. In the low power scenario (**A**), Uniform and Gamma are relatively invariant and always expect an intermediate number of alleles. Exponential is better when the QTL has two or fewer alleles, but it underestimates the number of alleles for multiallelic series. In the high power scenario (**B**), Exponential is close to the correct expectation for an intermediate number of alleles but underestimates when the QTL is highly multiallelic. Gamma and Uniform are similar to Exponential but estimate more alleles, and they substantially overestimate when there is only one allele.

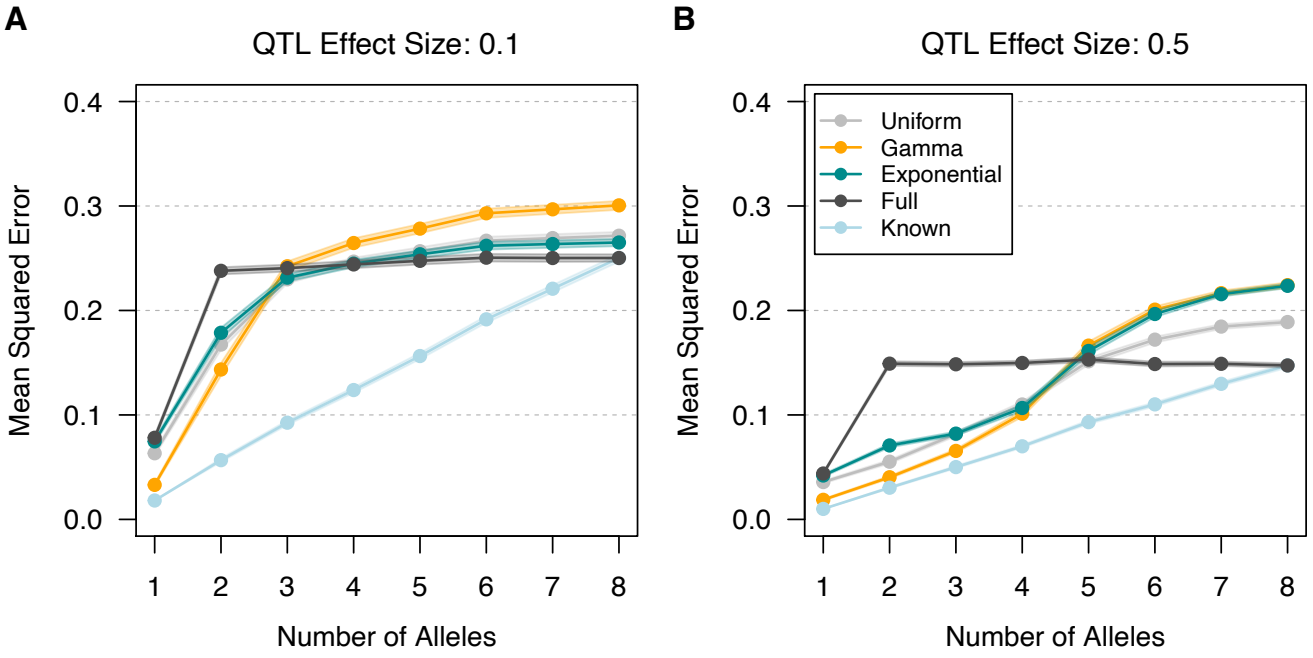

Figure S5 Mean squared error (MSE) of haplotype effect estimates using the Uniform, Gamma, and Exponential prior distributions, for varying numbers of true functional alleles, across two effect sizes. Full is the haplotype-based approach where all haplotypes are functionally distinct, and Known is an oracle prior in which the correct allelic series is known. Points are connected for clarity. Shading denotes 95% confidence intervals. In the low power scenario (**A**), Uniform, Gamma, and Exponential outperform Full when the true number of alleles is small. Exponential is better than Uniform and Gamma when the QTL has two or fewer alleles, but it is worse when the QTL is multiallelic. In the high power scenario (**B**), the Uniform, Gamma, and Exponential outperform Full for an intermediate number of alleles.

Supplemental Tables

Table S1 Computation time in minutes for selected analyses.

Analysis	# of Haplotypes	# of Samples	Duration (min)
PreCC MCV Full	8	100,000	3.61
PreCC MCV CRP	8	100,000	8.01
PreCC MCV Tree (prior)	8	(1000 trees)	395.97
PreCC MCV Tree (posterior)	8	100,000	8.19
DSPR CG4086 Full	15	100,000	28.21
DSPR CG4086 CRP	15	1,000,000	410.99
DSPR CG10245 Full	15	100,000	29.67
DSPR CG10245 CRP	15	1,000,000	646.38

Table S2 Width of the 95% highest posterior density for MCV (fL) haplotype effects using the Full, CRP, and Tree approaches.

Haplotype	Full	CRP	Tree
A	6.60	4.06	2.99
B	5.03	3.36	3.23
C	7.14	5.04	3.04
D	5.87	3.44	3.23
E	5.96	3.43	3.23
F	5.96	5.57	5.72
G	6.08	3.96	4.43
H	5.07	3.74	2.99

Table S3 Top ten prior allelic series for MCV QTL in the PreCC using the Tree approach.

	Allelic Series	# of Alleles	Posterior Probability
1	0,0,0,0,0,0,0,0	1	0.4645
2	0,1,0,1,1,0,0,0	2	0.1612
3	0,1,0,1,1,0,2,0	3	0.0635
4	0,1,0,1,1,2,3,0	4	0.0521
5	0,0,0,0,0,0,1,0	2	0.0450
6	0,0,0,0,0,1,0,0	2	0.0290
7	0,1,0,1,1,2,0,0	3	0.0263
8	0,1,0,1,1,1,1,0	2	0.0257
9	0,1,0,1,1,2,2,0	3	0.0232
10	0,1,0,1,1,0,1,0	2	0.0141

Table S4 Top ten posterior allelic series for CG4086 cis-eQTL in the DSPR using the CRP approach.

	Allelic Series	# of Alleles	Posterior Probability
1	0,0,0,0,0,1,1,0,0,2,0,0,0,0,0	3	0.2536
2	0,0,0,0,0,1,2,0,0,3,0,0,0,0,0	4	0.0503
3	0,0,0,0,0,1,1,0,0,2,1,0,0,0,0	3	0.0482
4	0,0,0,0,1,1,1,0,0,2,0,0,0,0,0	3	0.0459
5	0,0,0,0,0,1,1,0,0,1,0,0,0,0,0	2	0.0361
6	0,1,1,1,1,0,0,1,1,2,1,1,1,1,1	3	0.0297
7	0,1,1,1,1,2,2,1,1,0,1,1,1,1,1	3	0.0253
8	0,0,0,0,0,1,1,0,0,2,2,0,0,0,0	3	0.0238
9	0,0,0,0,1,2,2,0,0,1,0,0,0,0,0	3	0.0229
10	0,0,0,0,1,1,1,0,0,2,1,0,0,0,0	3	0.0147

Table S5 Width of the 95% highest posterior density interval for CG4086 haplotype effects in the DSPR using the Full and CRP approaches.

Haplotype	Full	CRP
A1	3.56	2.99
A2	1.19	0.17
A3	1.03	0.17
A4	0.99	0.17
A5	5.11	3.11
A6	1.81	1.31
A7	1.07	0.68
AB8	1.22	0.19
B1	2.03	0.19
B2	1.02	0.56
B3	5.11	3.11
B4	1.09	0.17
B5	1.19	0.18
B6	1.04	0.17
B7	1.06	0.17

Table S6 Top ten posterior allelic series for CG10245 cis-eQTL in the DSPR using the CRP approach.

	Allelic Series	# of Alleles	Posterior Probability
1	0,1,2,1,2,1,3,4,1,1,5,6,1,7,8	9	0.002380
2	0,1,2,3,2,1,3,3,3,3,4,1,3,0,2	5	0.001602
3	0,1,2,3,2,4,5,3,3,3,1,6,3,5,7	8	0.001206
4	0,1,2,3,2,1,4,3,3,3,1,5,3,6,0	7	0.001068
5	0,1,2,3,2,1,4,5,1,1,6,7,1,8,9	10	0.001064
6	0,1,2,1,2,1,3,4,4,1,1,5,4,6,0	7	0.001050
7	0,1,2,2,2,1,3,3,3,3,4,1,3,0,2	5	0.001040
8	0,1,2,3,2,4,5,3,3,3,1,6,3,7,5	8	0.001034
9	0,1,2,1,2,1,3,4,1,5,6,7,1,8,9	10	0.001030
10	0,1,2,2,2,1,3,4,1,1,5,6,1,7,8	9	0.001021

Table S7 Width of the 95% highest posterior density interval for CG10245 haplotype effects in the DSPR using the Full and CRP approaches.

Haplotype	Full	CRP
A1	1.10	1.56
A2	2.42	2.61
A3	1.33	1.83
A4	9.95	7.76
A5	1.11	1.72
A6	1.01	1.59
A7	0.88	1.53
AB8	0.90	1.12
B1	0.98	1.55
B2	9.98	7.80
B3	0.89	1.52
B4	1.80	2.53
B5	0.92	1.55
B6	0.91	1.53
B7	1.09	1.88