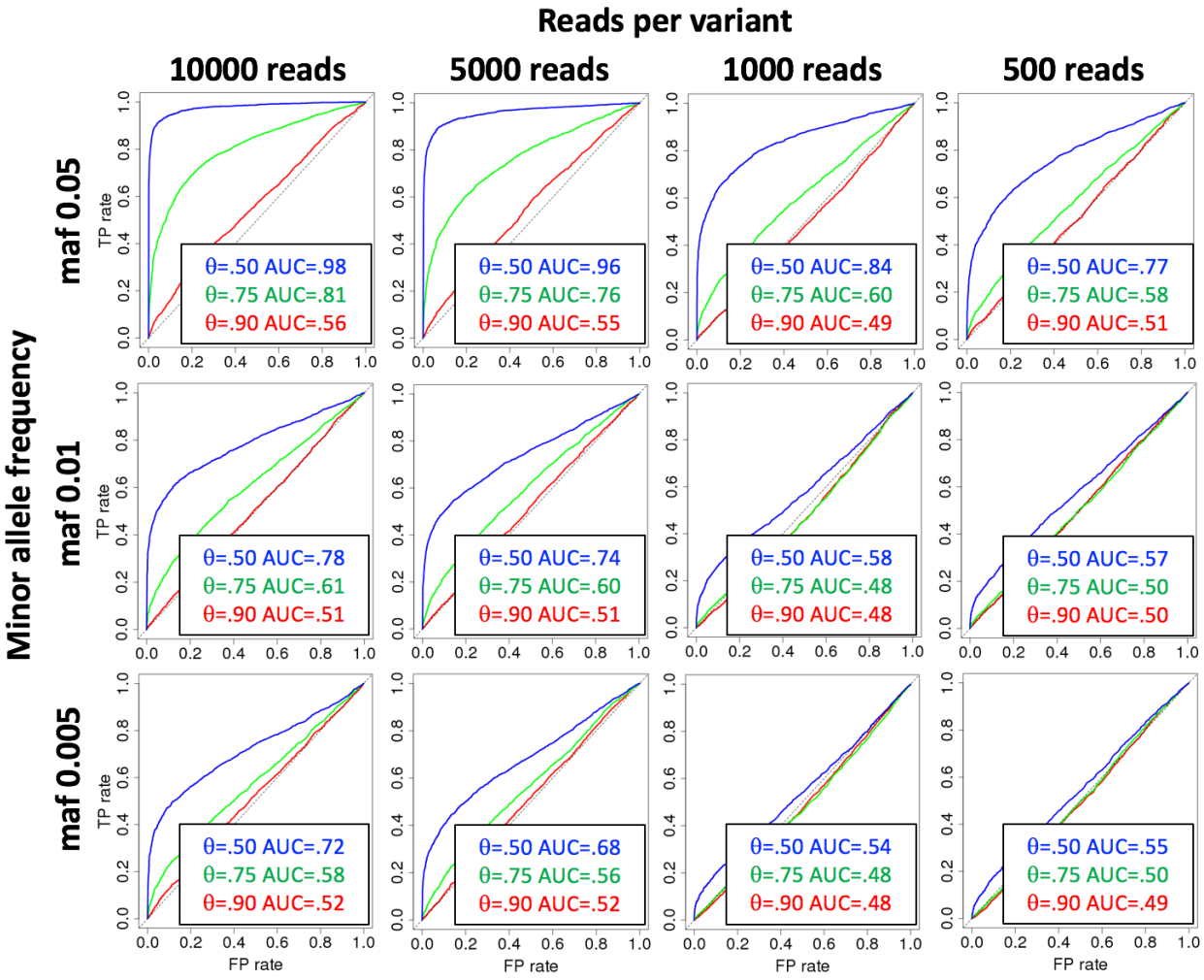
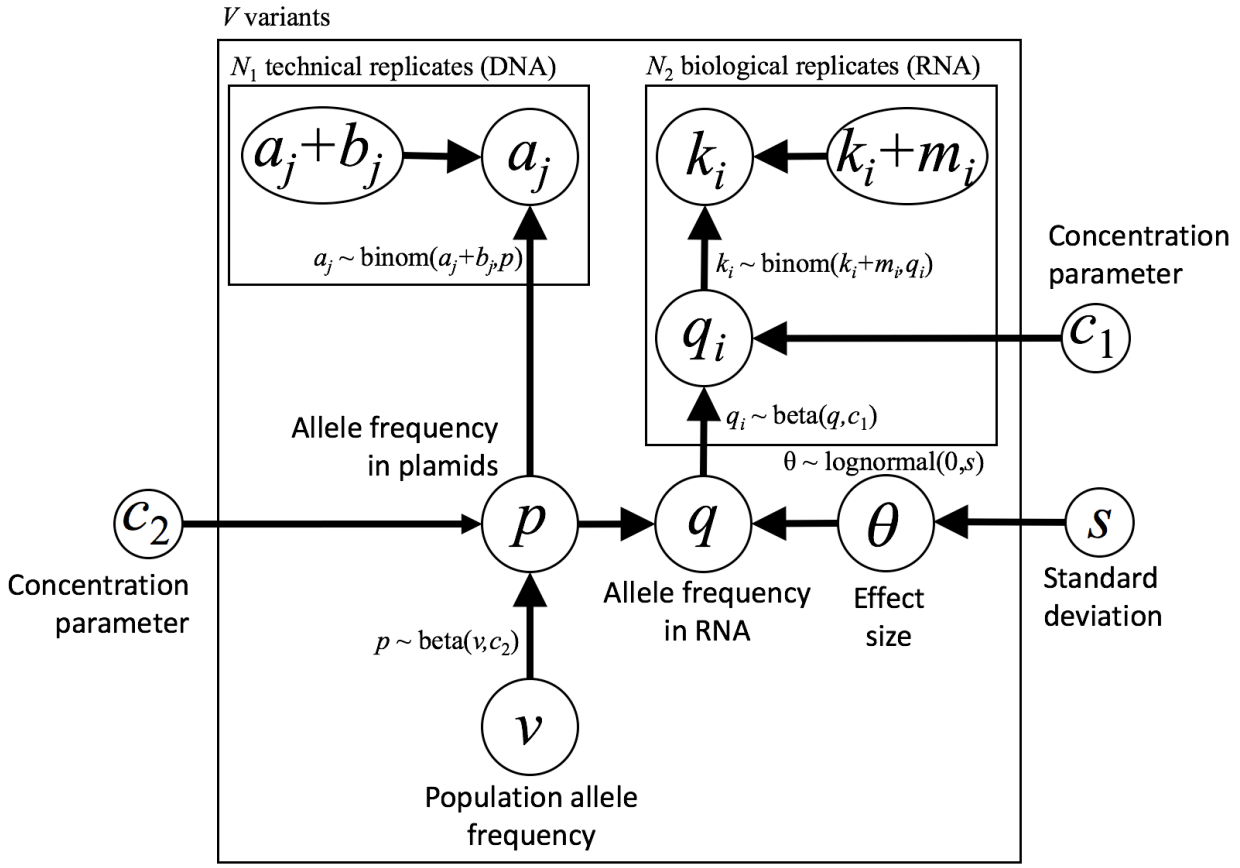
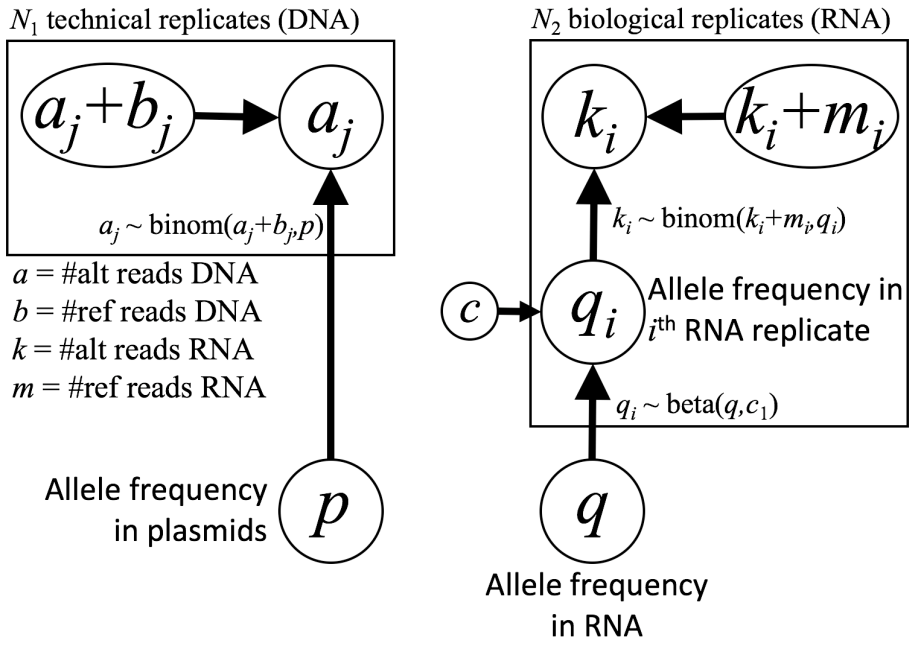
Supplementary Figures and Tables for Majoros et al. (2019), Bayesian Estimation of Genetic Regulatory Effects in High-throughput Reporter Assays



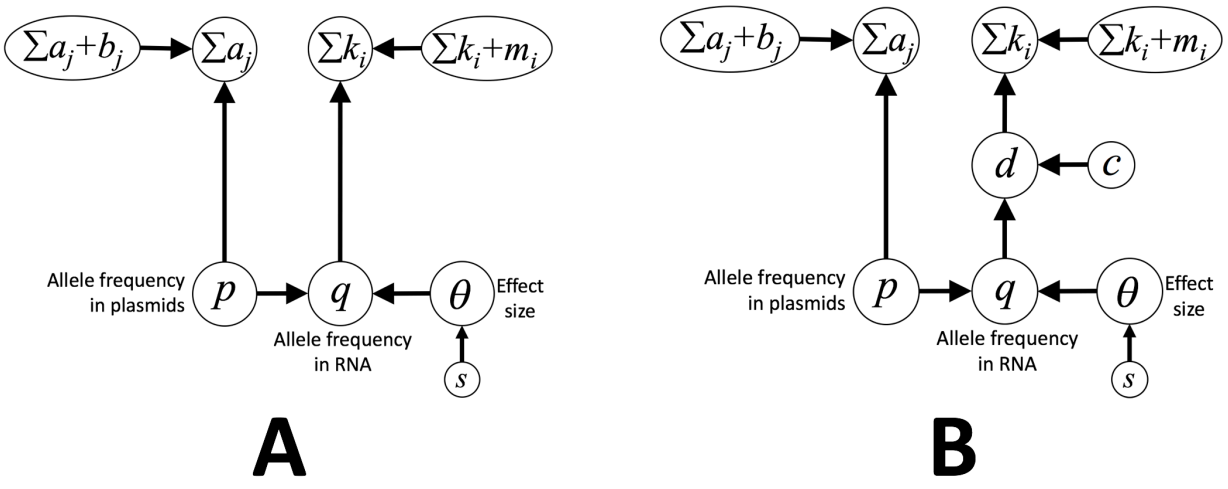
Suppl. Fig. S1: ROC curves for the BIRD model on simulated data, for different combinations of variant read coverage (columns), minor allele frequency (rows), and effect size (color: θ), where effect size represents the transcriptional output of the alternate (minor) allele divided by the transcriptional output of the reference allele. **Blue**: θ = 0.5 (50% reduction in transcriptional rate). **Green**: θ = 0.75 (25% reduction in transcriptional rate). **Red**: θ = 0.90 (10% reduction in transcriptional rate). Simulations: 10,000 variants, 1 DNA replicate, 10 RNA replicates.



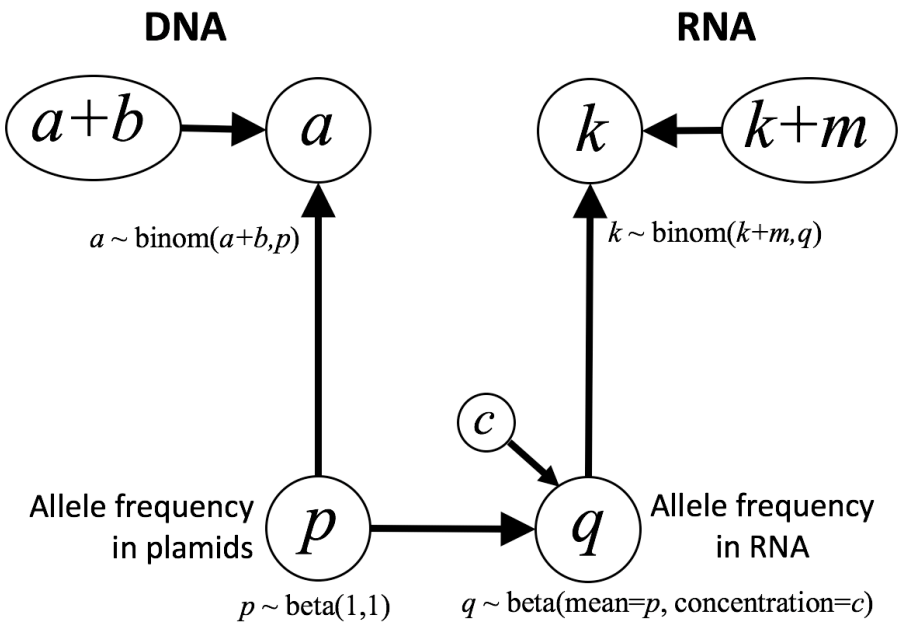
Suppl. Fig. S2: Multi-site model used to estimate dispersion parameters from real data, for use in the simulator. Variables are as defined in Table 1, with the addition of: c_1 = concentration parameter for beta prior on q_i; c_2 = concentration parameter of beta prior on p; v = allele frequency estimated from prior data. Beta priors are parameterized by mode and concentration. All variables inside the outermost box have implicit subscripts (not shown) denoting which site they pertain to. Variables outside the outermost box are shared across sites. This model is used to estimate those shared parameters (c_1 and c_2).



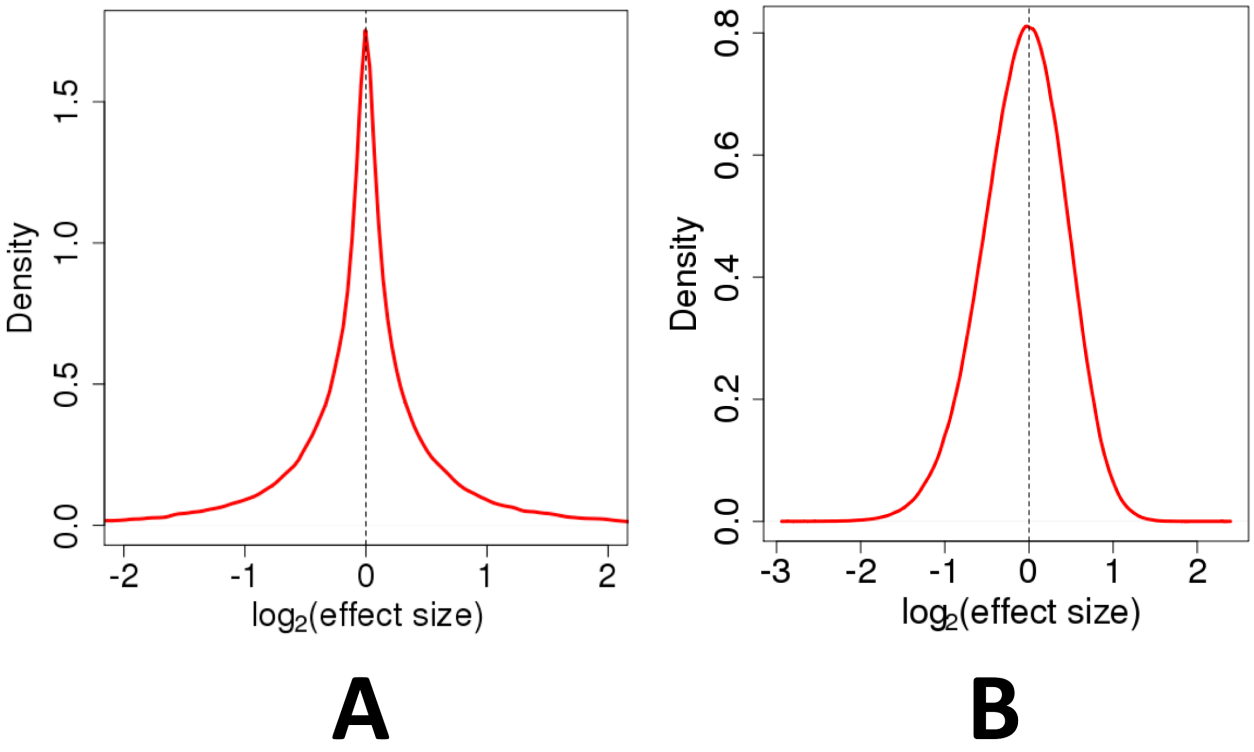
Suppl. Fig. S3: Modified version of BIRD lacking a prior on effect size. Variables are as defined in Table 1.



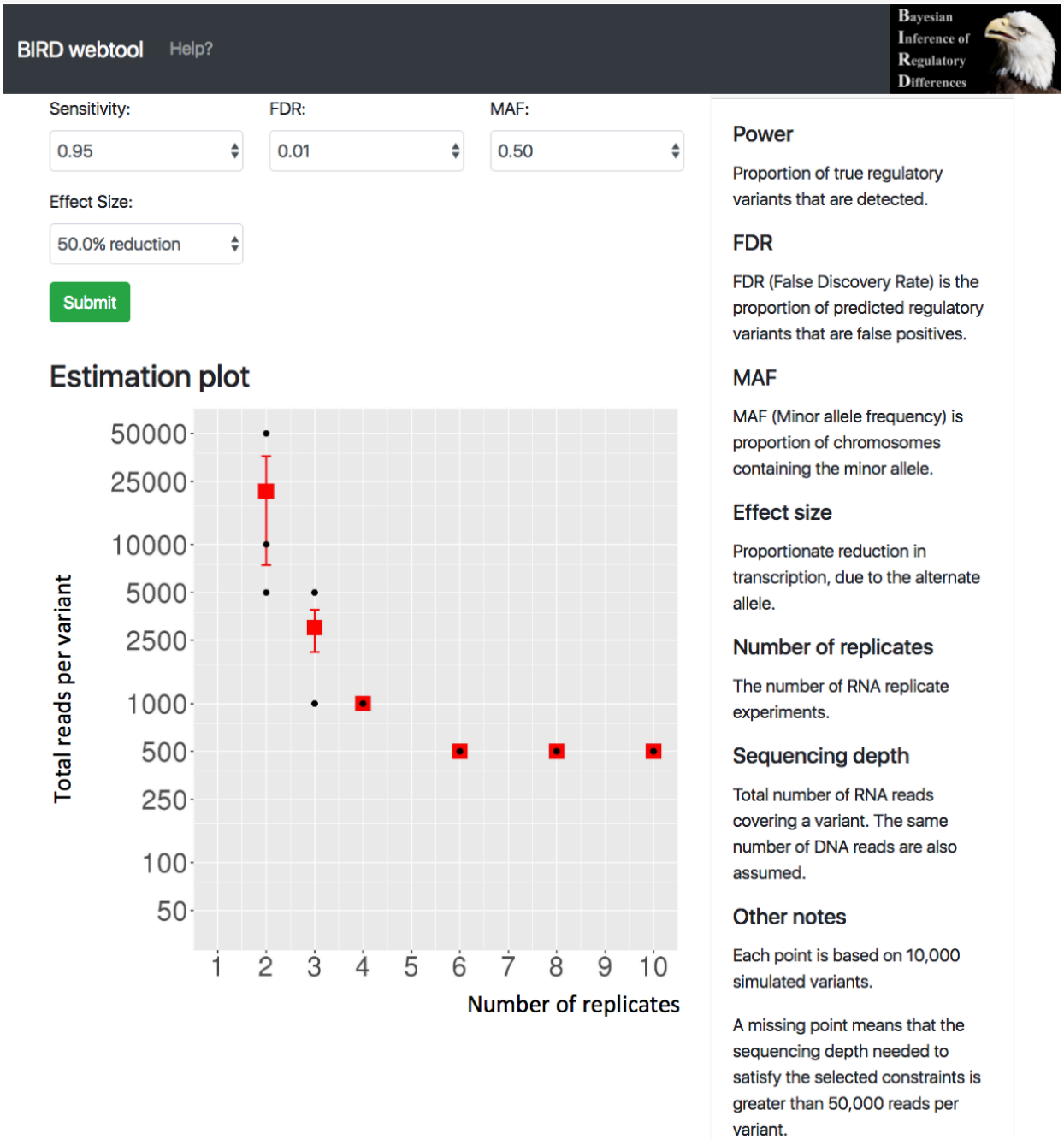
Suppl. Fig. S4: **A.** Model NR1, a modified version of BIRD in which counts are pooled across replicates. **B**. Model NR2, another modified version of BIRD in which counts are pooled across replicates; the additional concentration parameter c is intended to accommodate greater variance, due to replicates. All other variables are as defined in Table 1.



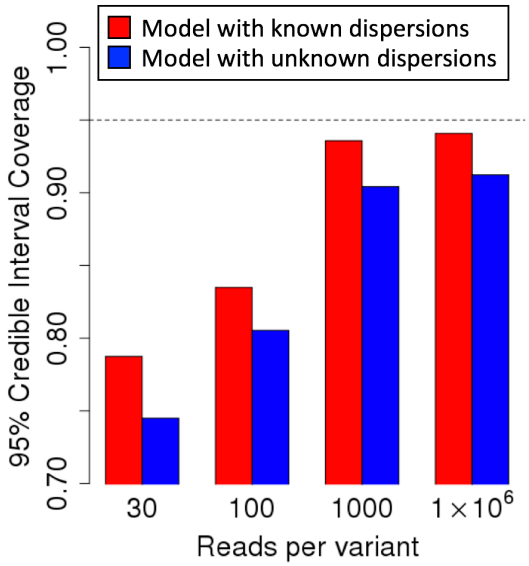
Suppl. Fig. S5: The Swift model. Variable p is sampled conditional on a and b only, and then q is sampled conditional on p, k, and m (and c) Read counts are pooled across replicates, so that $a = \sum_i a_i$, $b = \sum_i b_i$, $k = \sum_i k_i$, and $m = \sum_i m_i$. Variables are otherwise defined as in Table 1.



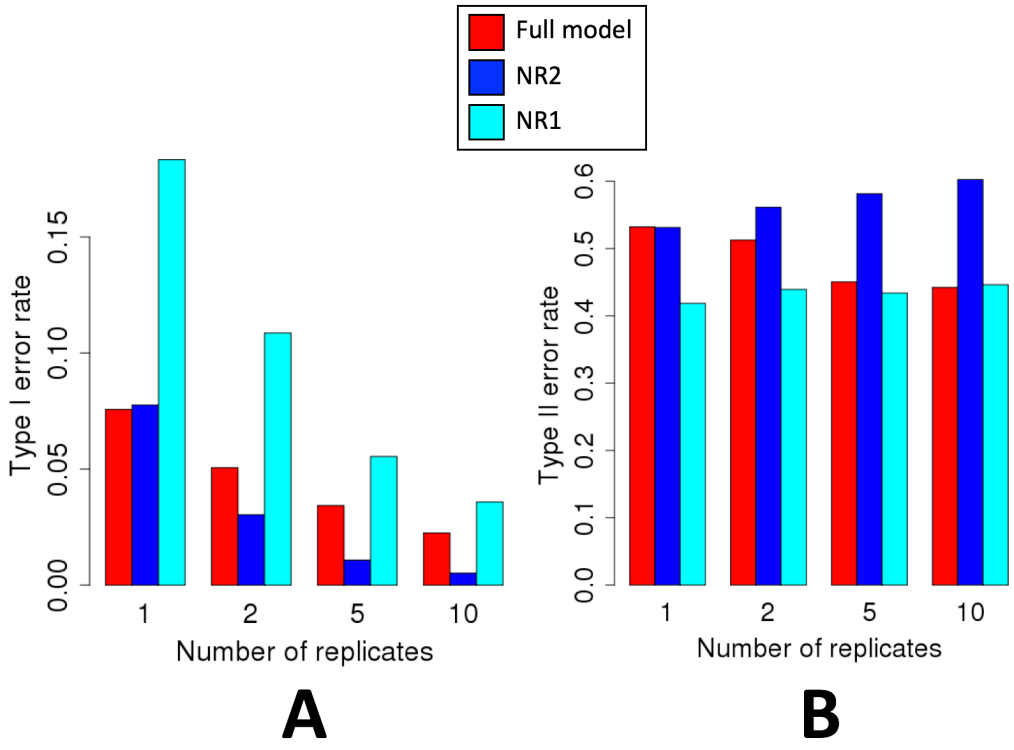
Suppl. Fig. S6: **A.** Informative prior for effect size used in the full BIRD model: $\theta \sim \text{lognormal}(0, s)$, $s \sim \text{gamma}(1.1,3)$. On a non-log scale, this prior shrinks effect sizes toward the null value of 1 (no effect), and places most of the mass between a halving and a doubling of transcriptional rate. **B.** The implicit prior on θ that is induced by the beta prior on q in the Swift model. This prior also shrinks effect sizes toward the null value of 1 and places most of the mass between a halving and a doubling of transcriptional rate.



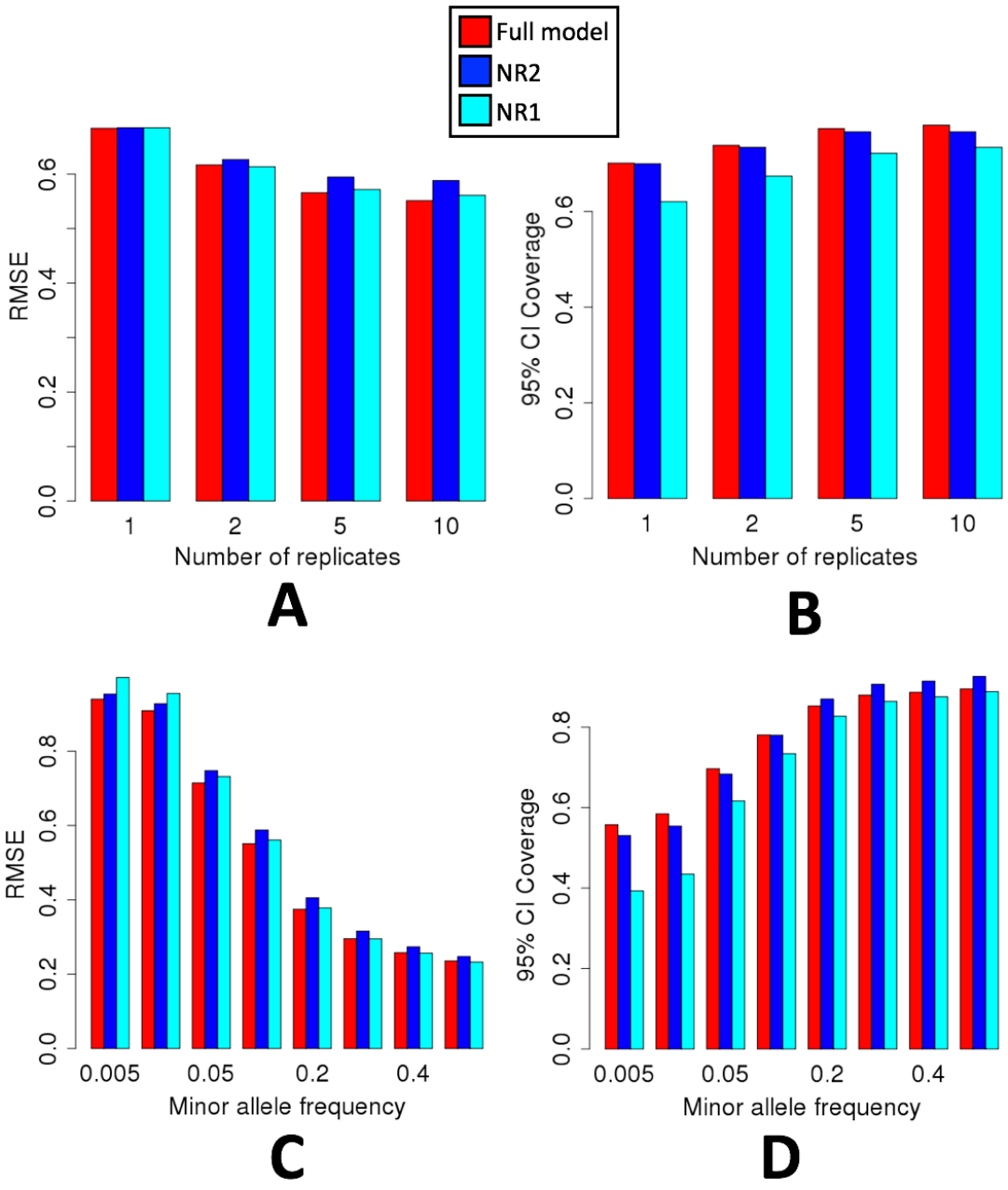
Suppl. Fig. S7: An example output of the web tool (http://67.159.92.22:8080/) that estimates required variant read coverage and number of replicates to achieve a given sensitivity and false discovery rate. Each black dot represents one or more simulations at the given number of replicates and read coverage, and for the specified effect size and minor allele frequency (maf). Red squares are median read counts required at a given number of replicates to achieve a given sensitivity and false discovery rate. Error bars denote standard errors.



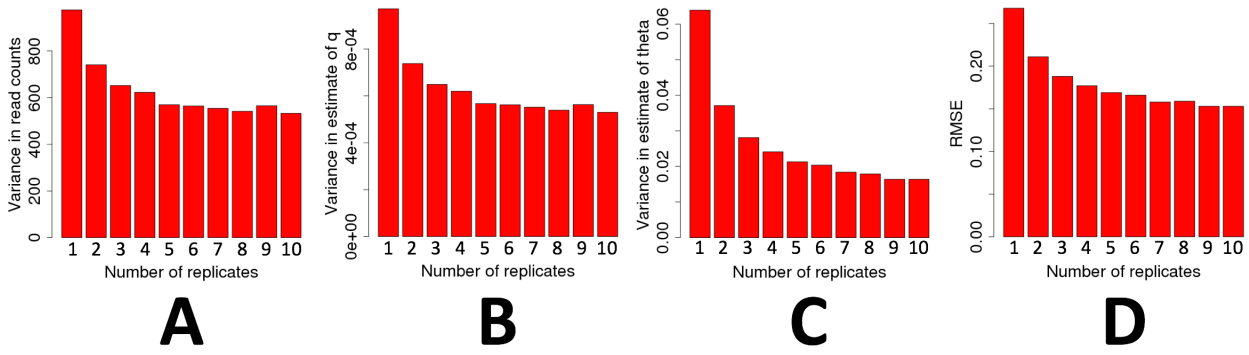
Suppl. Fig. S8: Coverage of 95% credible intervals on 10,000 simulated variants, for the BIRD model (blue) and for a version of BIRD in which simulated dispersions are fixed in the model (red). The 95% credible intervals for the model with known dispersions converge toward a coverage of approximately 0.94. Variants were simulated to have an effect size of 0.5 and an allele frequency of 0.5, with 10 RNA replicates having the given total number of reads per variant.



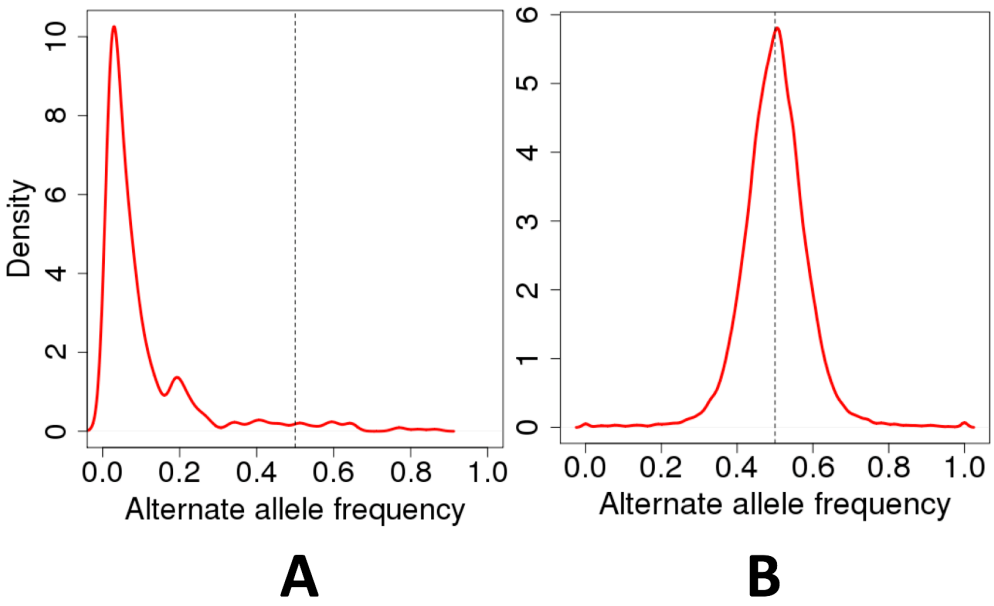
Suppl. Fig. S9: **A.** Type I error of the BIRD model (red) and the models in which read counts are pooled across replicates (NR2: blue; NR1: cyan). Simulation parameters: θ = 1, total number of DNA reads = 500, total number of RNA reads = 500, number of DNA replicates = 1, alternate allele frequency = 0.1, number of variants = 10,000. **B.** Type II error of the BIRD model (red) and the models in which read counts are pooled across replicates (NR2: blue; NR1: cyan). Simulation parameters: θ = 0.5, total number of DNA reads = 500, total number of RNA reads = 500, number of DNA replicates = 1, alternate allele frequency = 0.1, number of variants = 10,000.



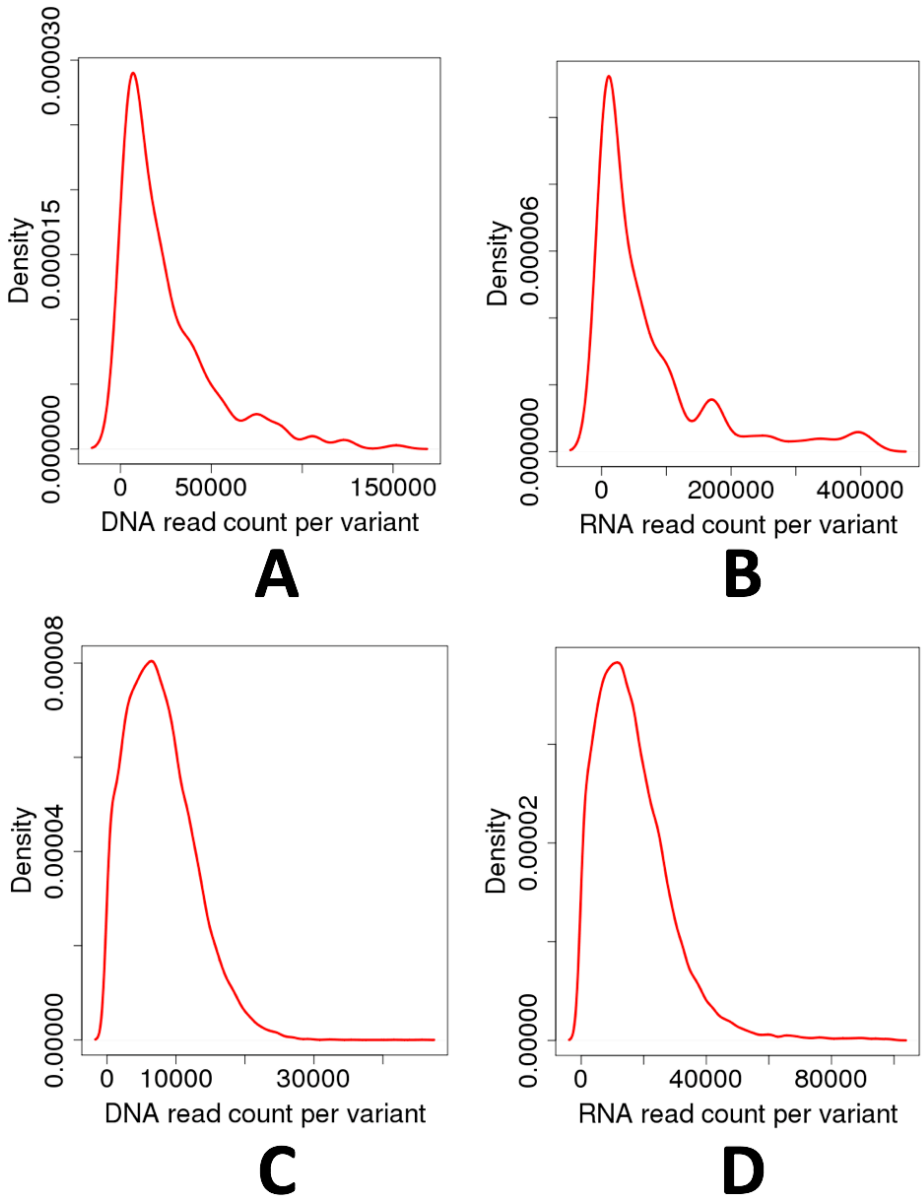
Suppl. Fig. S10: RMSE and coverage of credible intervals on simulated variants, for the BIRD model (red) and versions of BIRD that pool read counts across replicates (NR2: blue; NR1: cyan). **A.** RMSE as a function of number of replicates. **B.** Coverage of 95% credible intervals as a function of number of replicates. **C.** RMSE as a function of minor allele frequency. **D.** Coverage of 95% credible intervals as a function of minor allele frequency. **Simulation parameters for panels A and B**: θ = 0.5, total number of DNA reads = 500, total number of RNA reads = 500, number of DNA replicates = 1, minor allele frequency = 0.01, number of variants = 10,000. **Simulation parameters for panels C and D**: θ = 0.5, total number of DNA reads = 500, total number of RNA reads = 500, number of DNA replicates = 1, number of RNA replicates = 10, number of variants = 10,000.



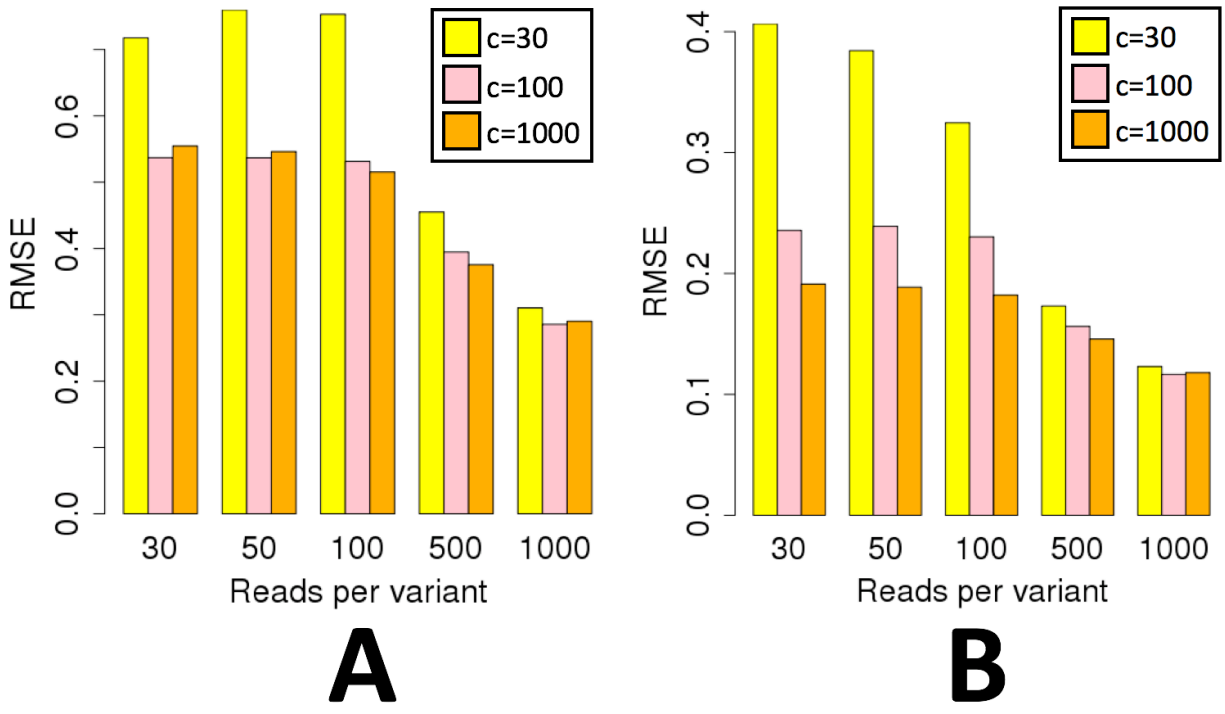
Suppl. Fig. S11: **A.** Sample variance in alternate allele read count for simulated variants, as a function of the number of RNA replicates. **B.** Variance in maximum likelihood estimate of q (alternate allele frequency in RNA) from simulated variants, as a function of the number of RNA replicates. **C.** Variance in *ad hoc* estimate of θ from simulated variants, as a function of the number of RNA replicates. **D.** Variance in RMSE of *ad hoc* estimate of θ from simulated variants, as a function of the number of RNA replicates. **Simulation parameters for all panels**: θ = 0.5, total number of DNA reads = 1000, total number of RNA reads = 1000, number of DNA replicates = 1, minor allele frequency = 0.1, number of variants = 10,000.



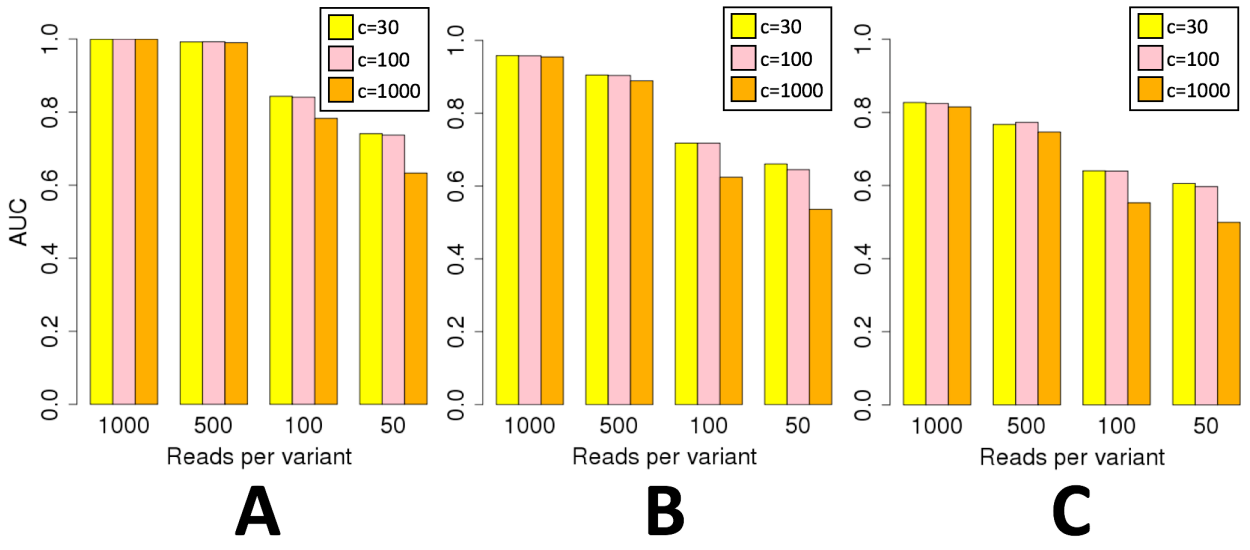
Suppl. Fig. S12: **A.** Alternate allele frequency distribution in 760 human donors from fetal adiposity study (Urbanek, et al. 2013). **B.** Alternate allele frequency distribution for Tewhey et al. (2016) LCL data.



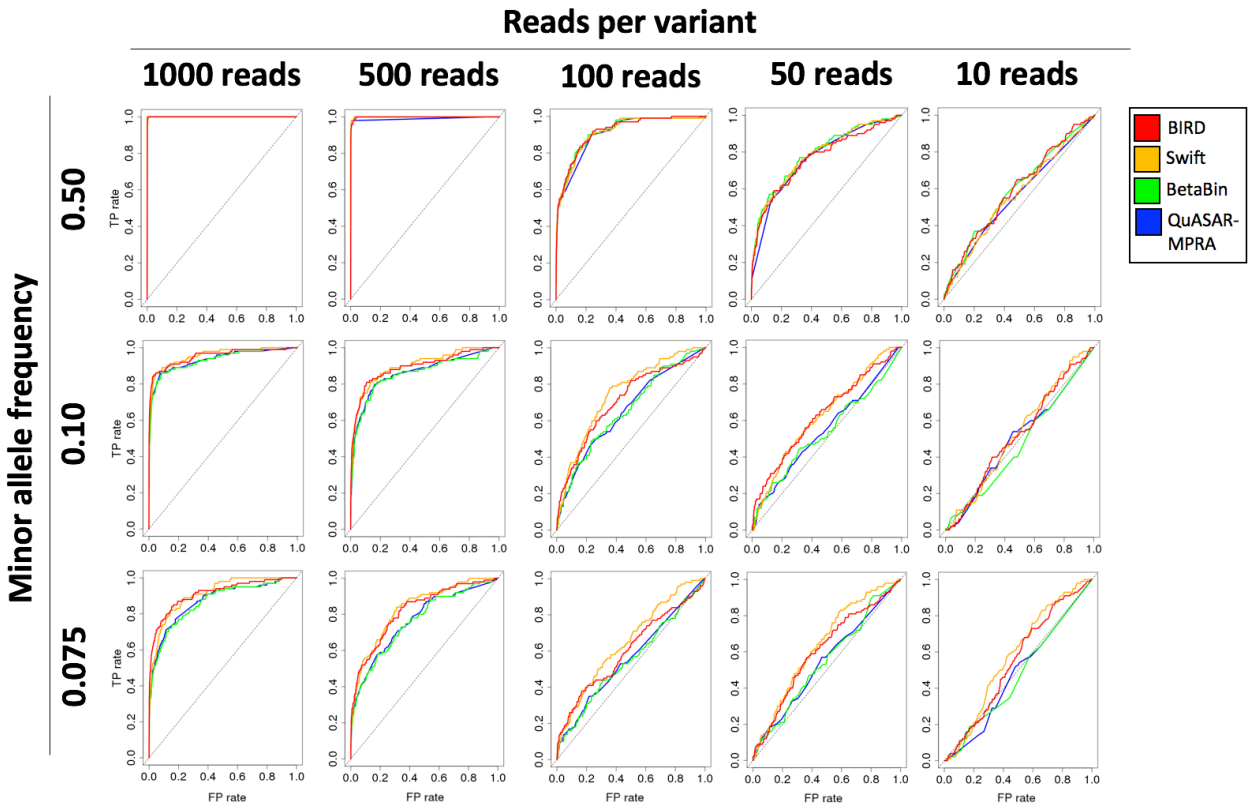
Suppl. Fig. S13: Distributions of per-variant read coverage. **A.** DNA read coverage in fetal adiposity data from HepG2 cells. **B.** RNA read coverage in fetal adiposity data from HepG2 cells. **C.** DNA read coverage in Tewhey et al. (2016) LCL data. **D.** RNA read coverage in Tewhey et al. (2016) LCL data.



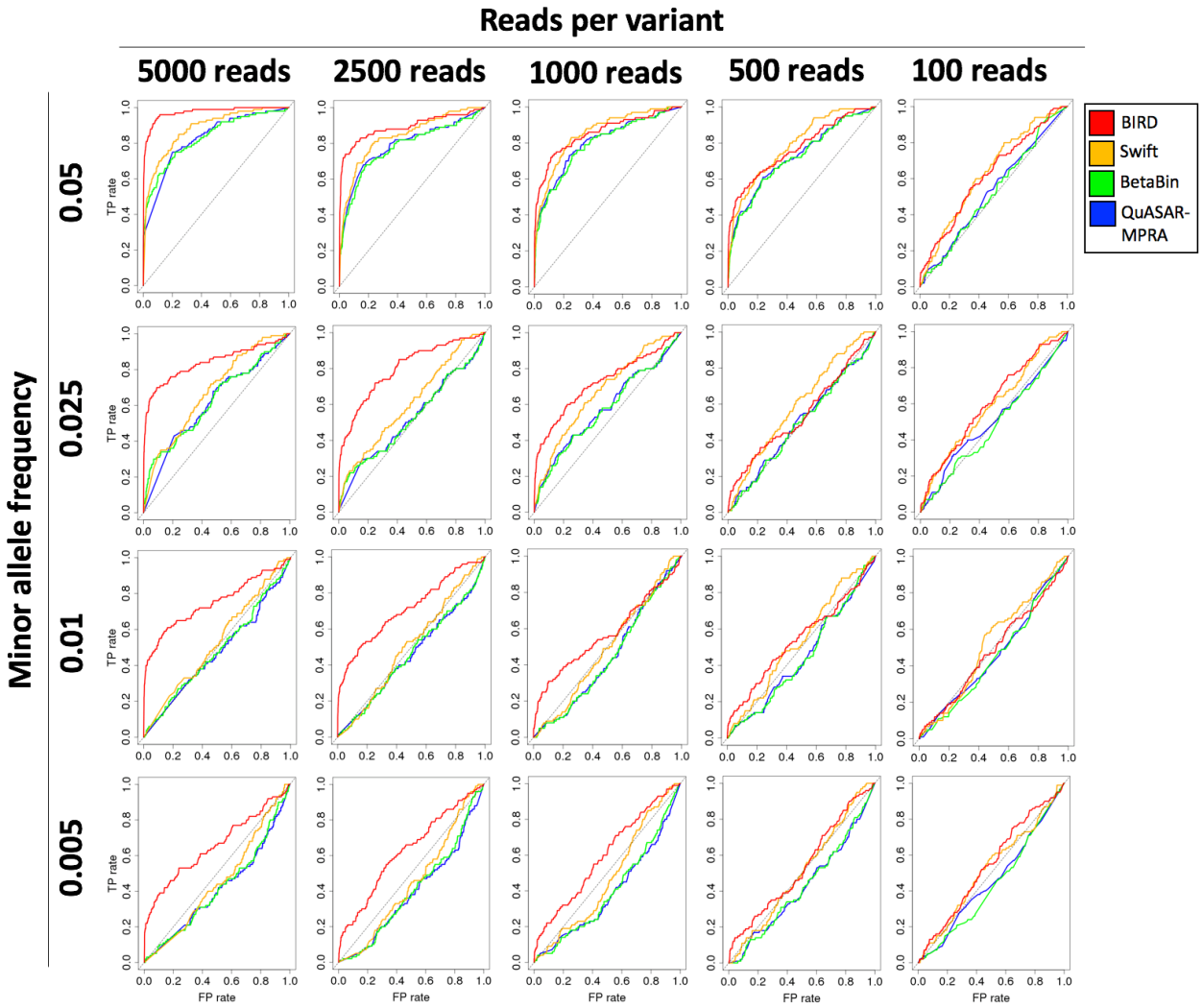
Suppl. Fig. S14: **A.** RMSE of the Swift model on fetal adiposity data in HepG2 cells, for different values of concentration parameter c, which controls the strength of the shrinkage prior on q and on the estimated effect size. **B.** RMSE of the Swift model on Tewhey et al. (2016) LCL data, for different values of concentration parameter c, which controls the strength of the shrinkage prior on q and on the estimated effect size.



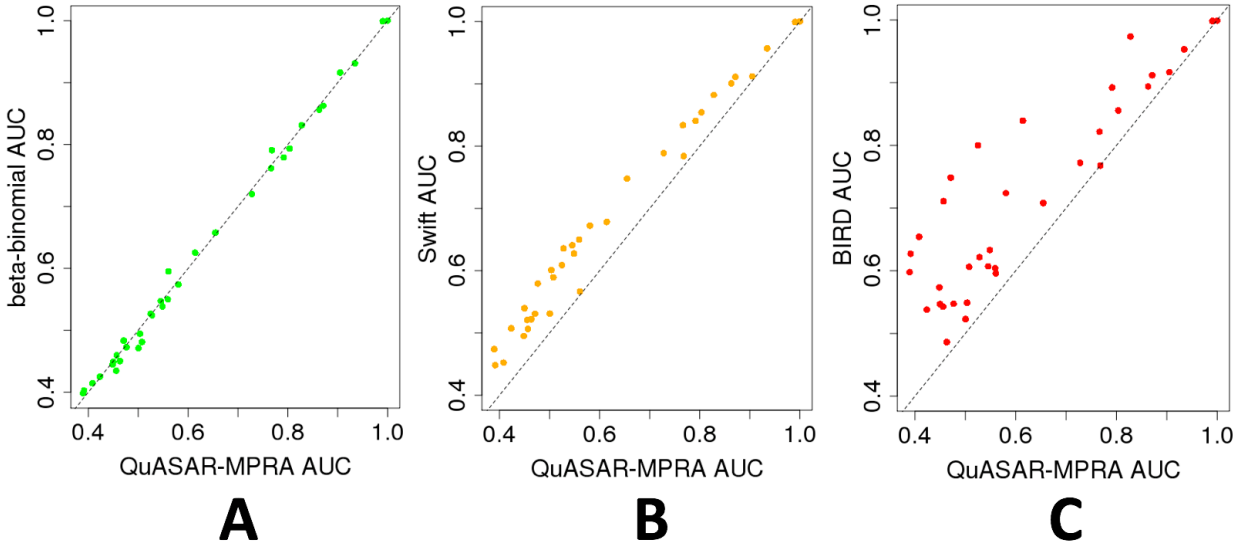
Suppl. Fig. S15: Area under ROC curves for Swift model on simulated variants. All simulations consisted of one DNA replicate and 10 RNA replicates with 10,000 null variants (θ=1) and 10,000 regulatory variants (θ=0.5). Parameter c controls the strength of the shrinkage prior on q and on the estimated effect size. **A.** AUC for simulations with maf=0.25. **B.** AUC for simulations with maf=0.1. **C.** AUC for simulations with maf=0.05.



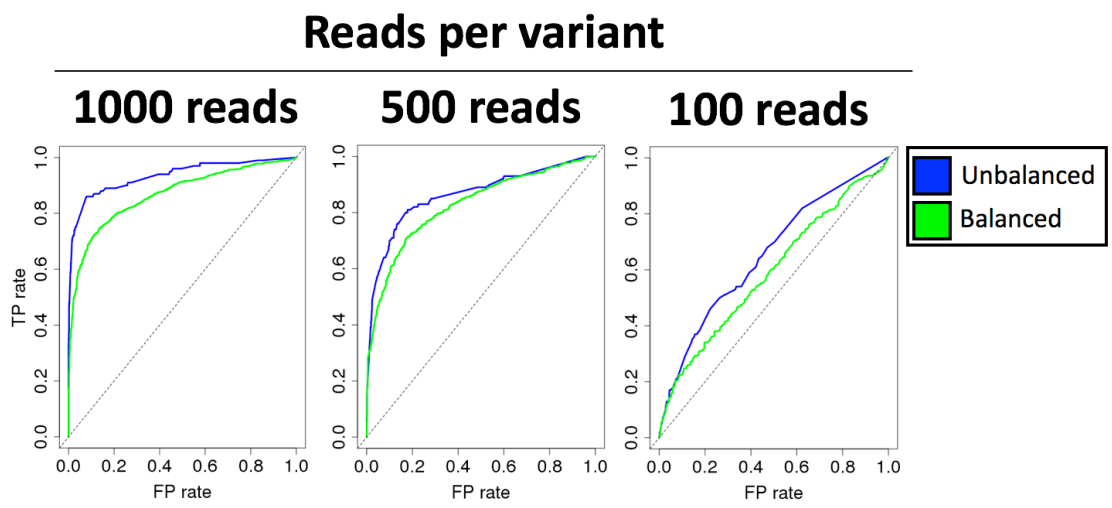
Suppl Fig. S16: ROC curves for BIRD, Swift, the beta-binomial test ("BetaBin"), and QuASAR-MPRA on 20,000 simulated variants at moderate-to-high minor allele frequency (100 regulatory variants, θ=0.5; 19,900 neutral variants, θ=1). AUC values are given in Suppl. Tables S1, S2.



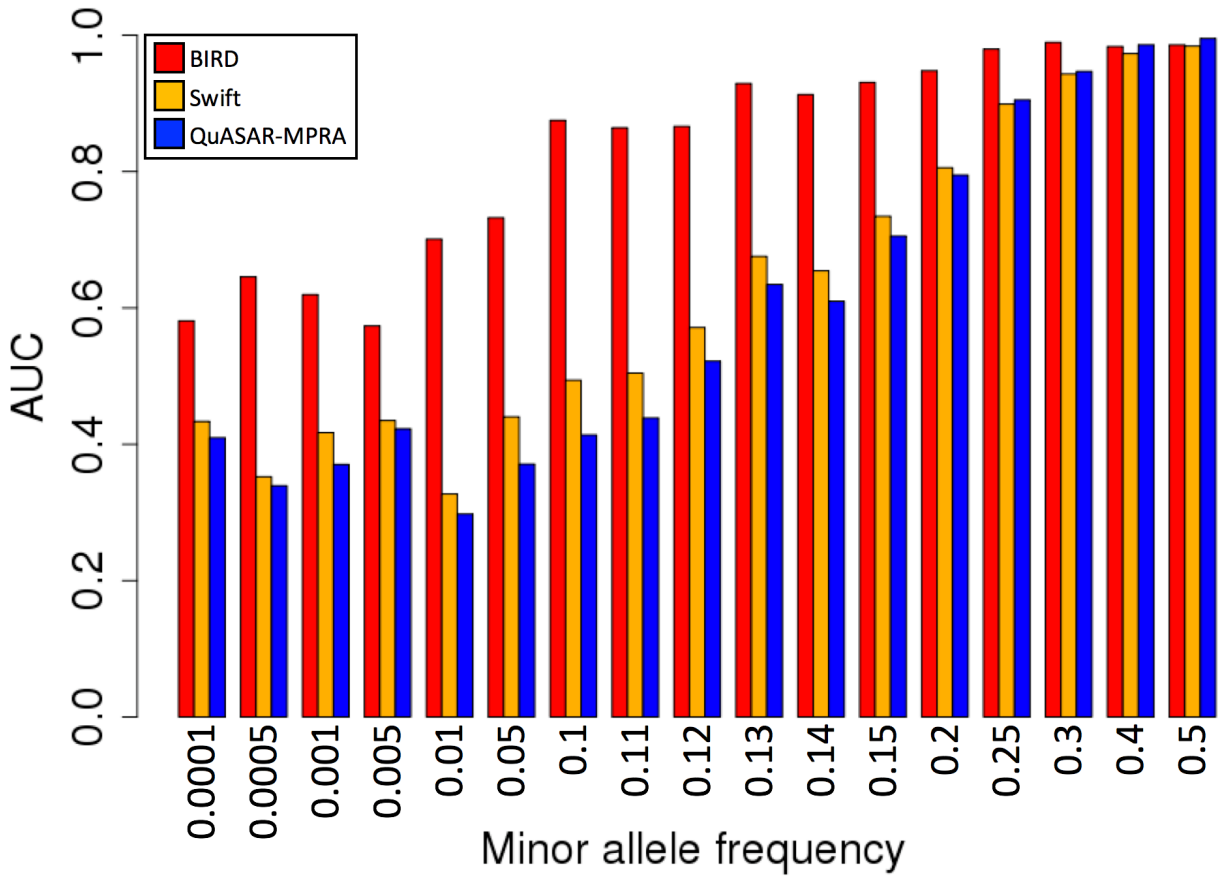
Suppl Fig. S17: ROC curves for BIRD, Swift, the beta-binomial test ("BetaBin"), and QuASAR-MPRA on 20,000 simulated variants at low minor allele frequency (100 regulatory variants, θ=0.5; 19,900 neutral variants, θ=1). AUC values are given in Suppl. Tables S1, S2.



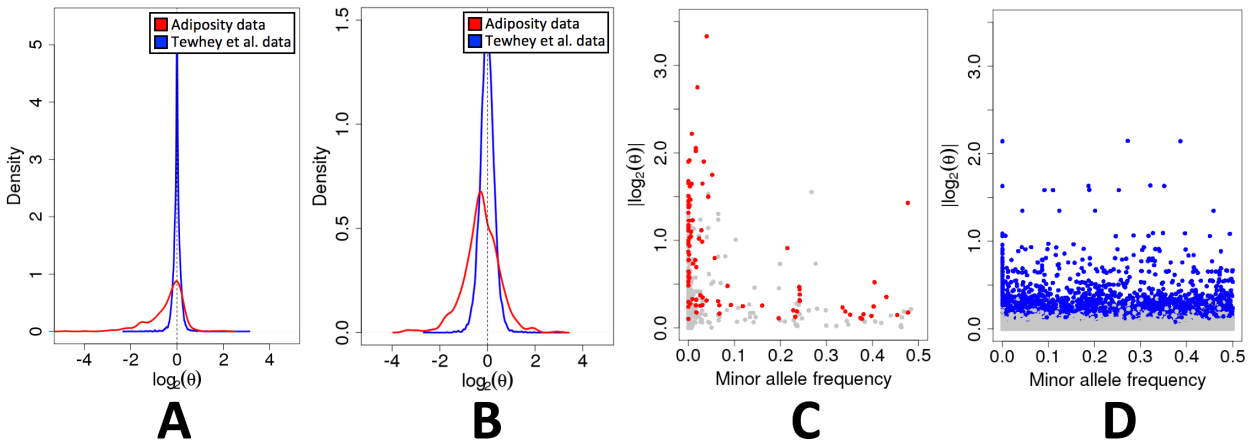
Suppl Fig. S18: **A.** Scatterplot of AUC values (across all MAFs and read coverages) from Suppl. Figs. S16, S17, for QuASAR-MPRA versus beta-binomial test. **B.** Scatterplot of AUC values (across all MAFs and read coverages) from Suppl. Figs. S16, S17, for QuASAR-MPRA versus Swift. **C.** Scatterplot of AUC values (across all MAFs and read coverages) for Suppl. Figs. S16, S17, for QuASAR-MPRA versus BIRD.



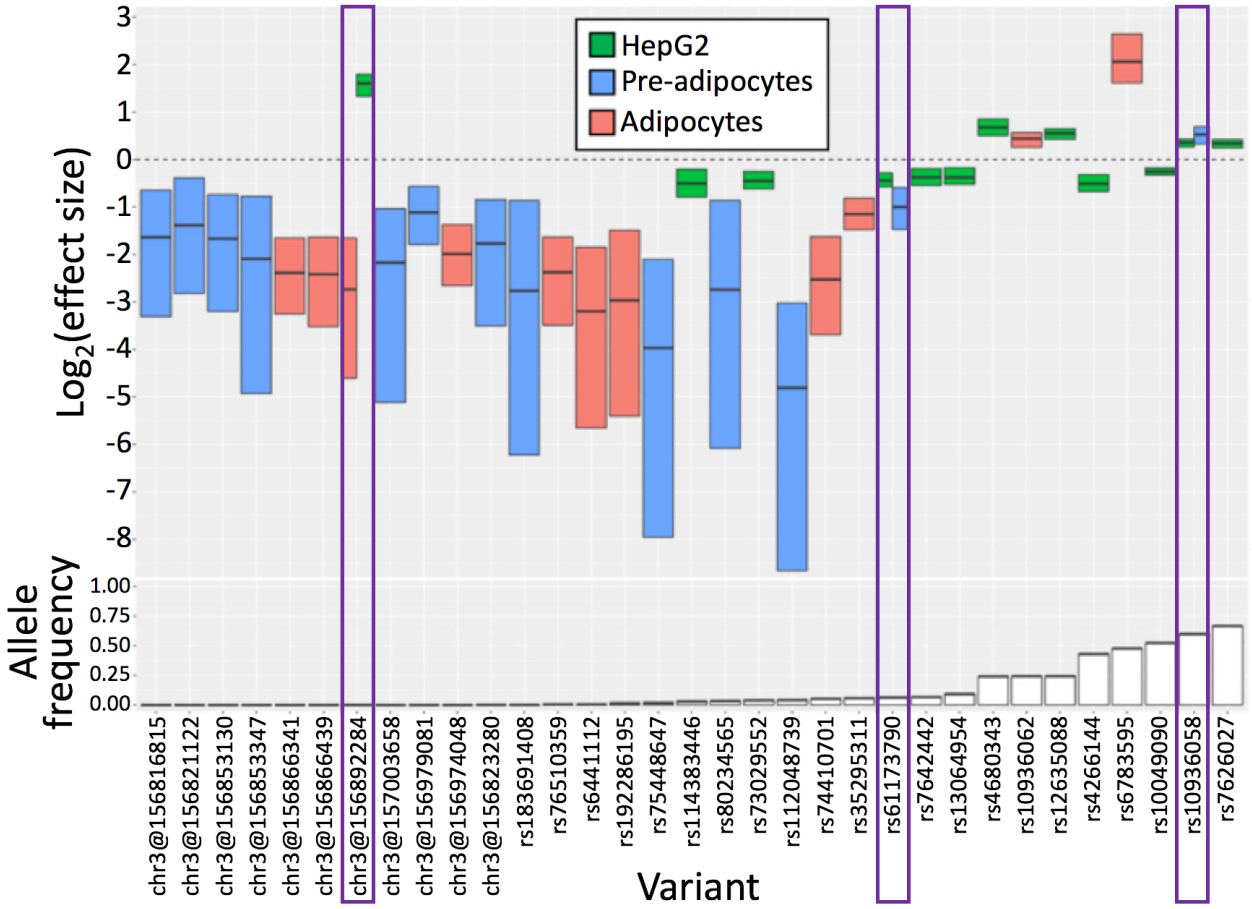
Suppl Fig. S19: ROC curves forQuASAR-MPRA, comparing performance on balanced test data (1000 regulatory variants, θ=0.5; 1000 neutral variants, θ=1) (green) and unbalanced test data (100 regulatory variants, θ=0.5; 19,900 neutral variants, θ=1) (blue) at maf=0.1.



Suppl Fig. S20: AUC values for BIRD, Swift, and QuASAR-MPRA on simulated variants in which the simulator was parameterized using concentration parameters estimated from HepG2 data previously published by Guo et al. (2017), instead of the concentration parameters estimated from Tewhey et al. (2016) LCL data that were used for all other simulations. AUC values are given in Suppl. Tables S3, S4. Simulations: 100 regulatory variants (θ=0.5) and 19,900 neutral variants (θ=1) were simulated to have total variant coverage of 5000 DNA reads per variant (1 replicate), and 5000 RNA reads per variant across 10 replicates. AUC values are given in Suppl. Tables S3, S4



Suppl Fig. S21: **A.** Distribution of $\log_2$ effect sizes predicted by BIRD on adiposity data (red) and Tewhey et al. LCL data (blue). **B.** Distribution of $\log_2$ effect sizes predicted by QuASAR-MPRA on adiposity data (red) and Tewhey et al. LCL data (blue); θ was computed from QuASAR-MPRA's allele skew $s=\text{logit}(q)=\log(q/(1-q))$, via $\theta=[q/p]/[(1-q)/(1-p)]$ for p the allele frequency in DNA. **C.** BIRD $|\log_2|$ effect sizes for adiposity data, as a function of minor allele frequency in 1000 Genomes Project data (shown are the largest effect sizes across the three cell types); red points correspond to posterior > 0.99. **D.** BIRD $|\log_2|$ effect sizes for Tewhey et al. data, as a function of minor allele frequency in 1000 Genomes Project data; blue points correspond to posterior > 0.99.



Suppl Fig. S22: Higher-resolution version of Fig. 3B. Log_2 effect sizes of high-confidence BIRD predictions on variants in a GWAS locus (top panel), and allele frequencies in Thousand Genomes Project samples (The 100 Genomes Project Consortium et al., 2015) (bottom panel). Segmented colored boxes indicate posterior median and 95% credible interval under the BIRD model. Gray boxes highlight the variants having an effect in multiple cell types.

MAF	COVERAGE	QuASAR-MPRA	BIRD	Swift	BetaBin	p-value	padj
0.50	1000	1.000000	0.998865	1.000000	1.000000	1.122e-09	4.363333e-09
0.50	500	0.990320	0.998310	0.999340	0.999340	0.1962	0.2145938
0.50	100	0.904910	0.916740	0.912030	0.916155	0.03369	0.04535192
0.50	50	0.767985	0.767615	0.783795	0.790855	0.9795	0.9795
0.50	10	0.560360	0.595795	0.566450	0.595285	0.0737	0.08320968
0.10	1000	0.934500	0.953245	0.956910	0.931250	0.001242	0.00207
0.10	500	0.862910	0.893845	0.900810	0.856300	6.679e-07	1.798192e-06
0.10	100	0.654740	0.707920	0.747820	0.657580	0.0003849	0.0007924412
0.10	50	0.548615	0.632925	0.627570	0.538250	4.294e-05	0.0001001933
0.10	10	0.500450	0.522845	0.531025	0.471315	0.5591	0.5929848
0.075	1000	0.871150	0.911855	0.911305	0.862705	9.705e-07	2.42625e-06
0.075	500	0.766430	0.821715	0.833675	0.761615	2.295e-08	7.302273e-08
0.075	100	0.559150	0.604015	0.650010	0.550040	0.04286	0.0544375
0.075	50	0.545335	0.607030	0.640810	0.547140	0.04355	0.0544375
0.075	10	0.476410	0.547415	0.579340	0.472765	0.04754	0.05737586
0.05	5000	0.827815	0.973540	0.882180	0.831190	6.351e-16	1.111425e-14
0.05	2500	0.791495	0.892155	0.840565	0.779215	1.098e-10	6.405e-10
0.05	1000	0.803460	0.855570	0.854225	0.793605	0.000418	0.0008127778
0.05	500	0.728015	0.772265	0.788755	0.720215	0.005651	0.008241042
0.05	100	0.527855	0.621705	0.635885	0.524030	0.0005698	0.00099715
0.025	5000	0.614115	0.839415	0.678220	0.625265	1.98e-13	1.6345e-12
0.025	2500	0.524750	0.799960	0.608920	0.526550	3.504582e-27	1.226604e-25
0.025	1000	0.580505	0.723970	0.672300	0.573820	3.26e-14	3.803333e-13
0.025	500	0.503510	0.548955	0.600860	0.494310	0.06745	0.07869167
0.025	100	0.507400	0.606275	0.589290	0.481065	0.004891	0.007442826
0.01	5000	0.470860	0.748370	0.530840	0.483300	1.33e-10	6.65e-10
0.01	2500	0.456705	0.710895	0.506350	0.459515	2.335e-13	1.6345e-12
0.01	1000	0.448600	0.573405	0.495050	0.445115	5.717e-05	0.0001250594
0.01	500	0.449860	0.546815	0.539520	0.448920	0.001831	0.002912955
0.01	100	0.463210	0.486160	0.522150	0.450275	0.5807	0.5977794
0.005	5000	0.408105	0.654190	0.452340	0.414525	1.461e-07	4.26125e-07
0.005	2500	0.391300	0.627120	0.448230	0.402575	4.329e-09	1.51515e-08
0.005	1000	0.389350	0.597745	0.473960	0.398240	6.279e-10	2.747062e-09
0.005	500	0.423305	0.538075	0.507365	0.424925	0.0004935	0.0009090789
0.005	100	0.455695	0.542720	0.521040	0.434660	0.02877	0.040278

Suppl Table S1: AUC values for the ROC curves in Suppl. Figs. S16, S17. "BetaBin" is the standard beta-binomial test. The p-value column gives two-tailed p-values for the difference between BIRD and QuASAR-MPRA AUC values, as reported by pAUC (Robin et al., 2011). The padj column gives FDR-adjusted p-values (Benjamini and Hochberg, 1995) for BIRD versus QuASAR-MPRA.

	Min	1stQu	Median	Mean	3rdQu	Max
BIRD	0.48616	0.5967700	0.707920	0.7182697	0.8474925	0.998865
Swift	0.44823	0.5309325	0.635885	0.6796839	0.8371200	1.000000
BetaBin	0.39824	0.4654150	0.547140	0.6189117	0.7850350	1.000000
QuASAR-MPRA	0.38935	0.4670350	0.548615	0.6202621	0.7797400	1.000000

Suppl Table S2: Summary statistics for AUC values reported in Suppl. Table S1. "BetaBin" is the standard beta-binomial test.

MAF	QuASAR-MPRA	BIRD	Swift	p-value	padj
0.5	0.995250	0.985765	0.984040	0.002099	0.002548786
0.4	0.985900	0.983385	0.973015	0.5018	0.5018
0.3	0.947040	0.989435	0.943285	0.0002166	0.0002832462
0.25	0.905130	0.979550	0.899055	2.045e-07	3.4765e-07
0.2	0.794740	0.947915	0.805530	8.138e-11	1.537178e-10
0.15	0.705355	0.930805	0.734415	4.057e-12	9.852714e-12
0.14	0.609780	0.912870	0.654560	3.451158e-22	1.173394e-21
0.13	0.634340	0.929220	0.675595	4.250433e-25	1.806434e-24
0.12	0.521920	0.866315	0.571285	1.847216e-26	1.046756e-25
0.11	0.438610	0.864245	0.504425	8.112443e-34	6.895577e-33
0.1	0.413630	0.875075	0.494060	1.552553e-35	2.63934e-34
0.05	0.370870	0.732265	0.440160	2.707e-11	5.752375e-11
0.01	0.297745	0.700930	0.327165	1.817e-12	5.148167e-12
0.005	0.422640	0.573735	0.434865	0.01226	0.01302625
0.001	0.370325	0.619425	0.417100	7.54e-06	1.068167e-05
0.0005	0.339190	0.645600	0.352285	3.429e-06	5.299364e-06
0.0001	0.409600	0.581085	0.433395	0.004607	0.005221267

Suppl Table S3: AUC values for the plot in Suppl. Fig. S20. The p-value column gives two-tailed p-values for the difference between BIRD and QuASAR-MPRA AUC values, as reported by pAUC (Robin et al., 2011). The padj column gives FDR-adjusted p-values (Benjamini and Hochberg, 1995) for BIRD versus QuASAR-MPRA.

	Min	1stQu	Median	Mean	3rdQu	Max
BIRD	0.5737	0.7009	0.8751	0.8304	0.9479	0.9894
Swift	0.3272	0.4349	0.5713	0.6261	0.8055	0.9840
QuASAR-MPRA	0.2977	0.4096	0.5219	0.5978	0.7947	0.9952

Suppl Table S4: Summary statistics for AUC values reported in Suppl. Table S3.