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

S1. Model Description

The BIRD model (Fig. 1B) includes two alternate allele frequencies, p (in the plasmid DNA), and q (in the RNA). The relationship between p and q is determined by the effect size θ, which is defined as an odds ratio between q and p:

$$\theta = \frac{transcriptional\ rate\ of\ alt\ allele}{transcriptional\ rate\ of\ ref\ allele} = \frac{\frac{q}{p}}{\frac{1-q}{1-p}} = \frac{\frac{q}{1-q}}{\frac{p}{1-p}}$$

This effect size can be thought of as a fold change in transcriptional rate between alleles. A value of $\theta = 1$ indicates no allelic effect. Solving for q gives:

$$q = \frac{\theta p}{1 - p + \theta p}$$

Variates p and q are not directly observable, as they pertain to the true allele frequencies in the underlying pool of DNA and RNA molecules in the cell and cannot be exactly ascertained. However, the DNA read counts a_j and b_j provide evidence regarding the probable values of p. Similarly, the RNA read counts k_i and m_i provide evidence regarding the probable values of q. BIRD also incorporates a replicate-specific allele frequency q_i for the i^{th} RNA replicate. Replicates are explicitly modeled in the RNA but not in the DNA, as RNA replicates are typically biological replicates, whereas in typical protocols DNA replicates are technical replicates and will have low variability in allele frequencies. As described below, we use a sampling approach to perform posterior inference on all of the latent variables, most importantly for the effect size, θ.

The generative process of the BIRD model can be formally specified via its priors and likelihoods:

$$p \sim \text{uniform}(0, 1)$$

$$\forall_j\ a_j \mid a_j + b_j, p \sim \text{binomial}(a_j + b_j, p)$$

$$s \sim \text{gamma}(1.1, 3)$$

$$\theta \mid s \sim \text{lognormal}(0, s)$$

$$q = \theta p/(1-p+\theta p)$$

$$c \sim \text{gamma}(1.1, 0.0005)$$

$$q_i \mid q, c \sim \text{beta}(\text{mode} = q, \text{concentration} = c)$$

$$\forall_i\, k_i \mid k_i + m_i, q_i \sim \text{binomial}(k_i + m_i, q_i)$$

A key feature of this dependency structure is that it places a mean-1 prior on the effect size θ, thus shrinking estimates toward 1 (no effect) absent strong evidence from data likelihood. This is a desirable feature, as we expect most tested variants genome-wide to have no effect. However, when sequencing coverage is sufficiently large, the model is capable of predicting any number of causal variants as dictated by the data (i.e., the data can "overwhelm the prior"; Murphy, 2012). The resulting prior on θ is shown in Suppl. Fig. S6A. The parameters of this prior are fixed, eliminating any burden on end users to specify the prior. We specifically chose the parameters of this fixed prior so as to favor effects ranging from a halving of transcription to a doubling of transcription, as we and others have seen in previous work that a majority of variants tend to fall in that range (Vockley et al., 2015; Patwardhan et al., 2012).

All beta priors in the full BIRD model were parameterized by their mode and concentration. Shape parameters for the beta distribution can be computed from a given mode m and concentration c as follows:

$$\alpha = m(c-2) + 1$$
$$\beta = (1-m)(c-2) + 1$$

S2. Inference

We perform posterior inference using Markov chain Monte Carlo (MCMC). Using the Metropolis-Hastings algorithm (Hastings, 1970), we obtain a chain of samples from the joint posterior distribution of the latent variables p, q, q_i, and θ, conditional on the observed data. We ignore the sampled values of p, q, and q_i, as these serve only to facilitate inference on θ. We take 2000 MCMC samples and discard the first 1000 as burn-in samples. We perform no thinning.

The posterior sample values of θ are used for both estimating effect sizes and for classifying variants as regulatory or neutral. We consider two mutually exclusive alternatives:

$$H_1\colon \theta > \lambda,$$
$$H_2\colon \theta < 1/\lambda.$$

for $\lambda \geq 1$. Under H_1, the alternate allele is associated with higher expression than the reference allele, and vice-versa for H_2. For all of the results in this paper, we set λ=1. We summarize the evidence for these alternate possibilities via the posterior probabilities $\mathrm{P}(\theta>\lambda|\{a_j\},\{b_j\},\{k_i\},\{m_i\},c)$ and $\mathrm{P}(\theta<1/\lambda|\ \{a_j\},\{b_j\},\{k_i\},\{m_i\},c)$, respectively. These posterior probabilities are estimated via:

$$\mathrm{P}(H_1) = \mathrm{P}(\theta > \lambda \mid \{a_j\},\{b_j\},\{k_i\},\{m_i\},c) \approx N_{\theta>\lambda}/N,$$

$$\mathrm{P}(H_2) = \mathrm{P}(\theta < 1/\lambda \mid \{a_j\},\{b_j\},\{k_i\},\{m_i\},c) \approx N_{\theta<1/\lambda}/N,$$

for $N_{\theta<1/\lambda}$ the number of MCMC samples for which $\theta<1/\lambda$, $N_{\theta>\lambda}$ the number of MCMC samples for which $\theta>\lambda$, and N the total number of accepted MCMC samples. Note that $\mathrm{P}(\theta<1/\lambda|\{a_j\},\{b_j\},\{k_i\},\{m_i\},c) + \mathrm{P}(\theta>\lambda|\{a_j\},\{b_j\},\{k_i\},\{m_i\},c) \leq 1$, and that H_1 and H_2 are mutually exclusive. We take the maximum of these two posterior probabilities as a summary of the evidence that the variant is regulatory (non-neutral):

$$P_{reg} = \max(\ \mathrm{P}(\theta<1/\lambda|\{a_j\},\{b_j\},\{k_i\},\{m_i\},c)\ ,\ \mathrm{P}(\theta>\lambda|\{a_j\},\{b_j\},\{k_i\},\{m_i\},c)\ )$$

Given a threshold t, we can perform classification by predicting variants for which $P_{reg} > t$ to be regulatory variants and all others to be neutral. By considering all possible thresholds, we can compute receiver operating characteristic (ROC) curves, which we summarize via the area under the curve (AUC).

Estimation of effect sizes is performed by computing the median of the chain of MCMC samples for θ. In addition to this point estimate, we report a 95% symmetric credible interval for θ, based on the MCMC samples, to indicate the degree of uncertainty in the point estimate.

S3. Implementation and Availability

BIRD is implemented in the probabilistic programming language STAN (Carpenter et al., 2017), and is distributed as STAN model files that can be run in R or python, or compiled to C++. For the experiments performed here, we used cmdstan version 2.17.0 and GCC version 7.3.0.

Sampling is performed using Metropolis-Hastings with the Hamiltonian Monte Carlo proposal (Duane et al., 1987). Using a burn-in of 1000 samples followed by an additional 1000 samples for inference, a single variant can be analyzed in approximate 2.5 seconds on a single CPU core.

BIRD is available for free download at:

http://www.geneprediction.org/bird/

A web tool is also provided that allows users to estimate the required sequencing depth and number of replicates needed to detect regulatory variants at a given sensitivity and false discovery rate, over a range of different allele frequencies and effect sizes, using our model (Suppl. Fig. S7). This tool is intended to aid in the design of experiments and the selection of optimal sequencing depths. The web tool is available at:

http://67.159.92.22:8080/

S4. Evaluation Methods

We assessed the predictive accuracy of BIRD on both real and simulated genetic variants. For both real and simulated data, we assessed the estimation error for θ by computing the root mean squared error (RMSE). RMSE reflects the error in an estimate of a parameter as compared to the true value of that parameter. For simulated data, the true value of θ is known from the simulation, but for real data the true value of θ is unknown. In order to compute RMSE for real data, estimates of θ from downsampled read counts are compared to θ_{true}, where θ_{true} is estimated from large read counts (prior to downsampling) via:

$$\theta_{true} = \frac{\frac{RNA_{alt}}{DNA_{alt}}}{\frac{RNA_{ref}}{DNA_{ref}}} = \frac{\frac{RNA_{alt}}{RNA_{ref}}}{\frac{DNA_{alt}}{DNA_{ref}}}$$

where RNA_{alt}, RNA_{ref}, DNA_{alt}, and DNA_{ref} are raw (non-downsampled) read counts; *alt* denotes the alternate allele and *ref* denotes the reference allele. No pseudocounts are applied.

For real data, only variants with high coverage (Suppl. Text S5) and with minor allele frequency ≥ 1% were considered. Because only variants with high coverage were considered, and because θ_{true} was computed from large read counts prior to downsampling, we expect θ_{true} to be close to the true value of θ.

Real variants satisfying the required coverage and allele frequency filters (Suppl. Text S5) were used to stochastically generate sets of 50,000 downsampled variants for testing. Read counts were downsampled uniformly at random (i.e., no effort was made to retain specific allele ratios or effect sizes) to 30, 50, 100, 500, and 1000 reads per variant in DNA and similarly in RNA, resulting in 5 test sets of 50,000 variants each. BIRD was then applied to the downsampled data to produce a posterior median estimate of θ. RMSE was computed both for BIRD and for a commonly used *ad hoc* estimator based on a simple ratio of read counts:

$$\theta_{ad\ hoc} = \frac{\frac{RNA_{alt}}{DNA_{alt}}}{\frac{RNA_{ref}}{DNA_{ref}}} = \frac{\frac{RNA_{alt}}{RNA_{ref}}}{\frac{DNA_{alt}}{DNA_{ref}}}$$

where RNA_x and DNA_x are downsampled read counts for allele x, and a pseudocount of 1 was added to all downsampled counts to avoid zero counts. The pseudocount was applied only for the *ad hoc* estimator and was not applied to the inputs to other models. The *ad hoc* estimator is based on the maximum likelihood estimates of p (alternate allele frequency in DNA) and q (alternate allele frequency in RNA). These maximum likelihood estimates (computed as ratios of read counts) are commonly used in allele-specific expression analyses and for allelic analysis in high-throughput reporter assays (for example, $\log_2(\theta_{ad\ hoc})$ is identical to the BAB score in Zhang et al., 2018).

Note that we expect the *ad hoc* estimate to be increasingly unstable as variant coverage decreases. In contrast, we expect the posterior estimate under the BIRD model to be more robust, due to increased reliance on priors at low read counts, as well as the explicit modeling of between-replicate variance.

Note that while the formulas given above for θ_{true} and $\theta_{ad\ hoc}$ are identical, their values will not be identical in general, because θ_{true} is computed from high-coverage read counts prior to downsampling, whereas $\theta_{ad\ hoc}$ is computed from randomly downsampled read counts and incorporates pseudocounts. As the difference between downsampled counts and raw (non-downsampled) counts diminishes, the difference between $\theta_{ad\ hoc}$ and θ_{true} should also diminish. This may bias RMSE estimates in favor of $\theta_{ad\ hoc}$ as compared to other estimators such as our BIRD model, particularly when downsampled read counts are large. As such, the advantage of the BIRD model over the *ad hoc* method may be underestimated in our comparisons.

In order to assess which components of BIRD contribute most to its predictive accuracy, we tested several handicapped versions of the model. In particular, we considered (1) ignoring replicate structure by pooling read counts across replicates (Suppl. Fig. S4); and (2) removing θ and its prior from the model and computing θ directly from posterior estimates of p and q (Suppl. Fig. S3). These modifications were performed separately.

We compared the classification accuracy of the full BIRD model to: (1) a simpler version called Swift (Suppl. Text S8); (2) simpler versions of BIRD lacking the prior on effect size (Suppl. Fig. S3) or lacking modeling of replicates (Suppl. Fig. 4A,B); (3) the standard beta-binomial test (see below); (4) the standard Fisher's exact test; and (5) QuASAR-MRPA (Kalita et al., 2018B), which utilizes a likelihood ratio test based on beta-binomial likelihoods.

The beta-binomial test is based on the beta-binomial distribution:

$$P(k|k+m,a,b) = \int_0^1 P(p|a,b)P(k|k+m,p)dp$$

where k and m are the alternate and reference allele read counts in RNA, respectively, and a and b are the alternate and reference allele read counts in DNA, respectively. The beta-binomial utilizes a beta prior $P(p|a,b)$ on the alternate allele frequency, p, and a binomial likelihood $P(k|k+m,p)$ on the RNA read counts. The alternate allele frequency, p, is integrated out to produce a posterior predictive distribution of new data (k, m) given previous data (a, b), under the assumption that the new and old data come from the same distribution. That is, we use the beta-binomial to model the null hypothesis that the alternate allele frequency in the underlying population is the same in the RNA and the DNA; violation of this assumption will be indicative of a regulatory effect. In order to use the beta-binomial as a null hypothesis test, we compute the tail probability of the distribution to arrive at a p-value.

S5. Generation of Human Data

Two sets of human variants were tested. The first set consisted of variants in a GWAS locus for fetal adiposity. We assayed these variants using Population STARR-seq ("Pop-STARR": Vockley et

al., 2015) in human HepG2 cells, human adipocytes, and human pre-adipocytes. The second set were previously synthesized and assayed using a massively parallel reporter assay by Tewhey et al. (2016). For the Tewhey data, we used only data from LCLs from individuals NA12878 and NA19239.

For the fetal adiposity data set, only variants with high coverage (at least 1000 DNA reads and 1000 RNA reads per variant, and having at least 10 reads for both the reference and alternate alleles in both DNA and RNA) and with minor allele frequency ≥ 1% were considered, to ensure accurate estimation of θ_{true}. For the Tewhey et al. (2016) data, the same thresholds were applied, except that at least 10,000 RNA reads and 10,000 DNA reads per variant were required. Note that these thresholds were not applied to the model inputs; they were only used to choose the variants for which a "true" effect size (θ_{true}) could be reliably estimated (based on high read counts), for testing the models after downsampling. After downsampling, read counts can be arbitrarily small (including zero), so that models were not tested only at high coverage.

The following table gives the numbers of variants retained after filtering:

Data	Total	Retained
Tewhey et al.	30667	9014
Adipocytes	372	372
Preadipocytes	407	407
HepG2	861	173

Suppl. Table T5: Numbers of variants before and after filtering.

Experimental methods for the fetal adiposity locus in HepG2 cells were described previously (Guo et al., 2017). Experimental methods for the fetal adiposity locus in pre-adipocytes and adipocytes were as follows. 36 million subcutaneous human white pre-adipocytes were obtained from Promocell (C-12731) and electroporated with the STARR-seq library using a Biorad Electroporator (170 volts, 950 μF, 2 μM, infinite capacitance) using 8 μg of plasmid split across 8 electroporations. The electroporations were then pooled and split across eight T-75 plates. The cells were then grown in pre-adipocyte growth medium (C-27410) for 48 hours as per manufacturer's protocol. After 48 hours, 5 plates were washed twice with PBS, incubated in 12 ml of PBS + 400 μL of DNAse each at 37°C for 4 minutes, washed with PBS once again, and harvested for RNA using a Qiagen RNeasy Midi kit as per manufacturer's protocol. The media on the remaining 3 plates was changed to pre-adipocyte differentiation media (C-27436) according to the manufacturer's protocol. After 48 hours the RNA from the 3 plates was harvested using the same method used on the 5 plates 48 hr prior. Libraries for the eight experimental samples and pooled plasmid controls (8 separate PCRs) were then prepared per the methods outlined by Vockley et al. (2015).

S6. Simulations

In addition to the human variants, we simulated a large number of data sets consisting of simulated regulatory variants ($\theta < 1$) and neutral variants ($\theta = 1$), at a variety of simulated effect

sizes, allele frequencies, and sequencing depths, and with a variety of replicate structures. Simulating different combinations of these parameters allowed us to both assess the robustness of our model across a range of scenarios, and to identify tradeoffs between different parameters such as variant coverage and number of replicates. For each simulated data set, we assessed classification accuracy (the ability to correctly classify variants as regulatory or neutral) via receiver operating characteristics (ROC) curves, which we summarize via the area under the ROC curve (AUC).

Our simulator closely emulates the data generation process for real genetic variants, so that we expect simulated data sets to be similar in all numerical aspects to real genomic variant data from a reporter assay. For each variant, we first simulate the generation of plasmid DNA by drawing a plasmid allele frequency from a beta distribution:

$$p \sim \text{beta}(\text{mode}=v, \text{concentration}=c_2),$$

where v is set to a fixed, chosen value for each simulation, and c_2 was estimated from data published by Tewhey et al. (2016) (Suppl. Text S7). Values of p=0 or p=1 are rejected and re-sampled. The effect size, θ, is set to a fixed, chosen value for each simulation. We then compute q deterministically from p and θ, via:

$$q = \frac{\theta p}{1 - p + \theta p}$$

The total per-variant sequencing depths N_{DNA} for DNA and N_{RNA} for RNA are fixed for each simulation to a chosen value. Alternate (D_{alt}) and reference (D_{ref}) allele counts for plasmid DNA are sampled from a binomial distribution parameterized by the sampled plasmid allele frequency p and the DNA sequencing depth N_{DNA}:

$$D_{alt} \mid N_{DNA}, p \sim \text{binomial}(N_{DNA}, p)$$
$$D_{ref} = N_{DNA} - D_{alt}$$

For RNA reads, we first simulate stochasticity in allele frequencies between replicates, to reflect randomness introduced during library preparation as well as natural biological variability in gene expression:

$$F_i \mid q, c_1 \sim \text{beta}(\text{mode} = q, \text{concentration} = c_1).$$

where q was computed previously from θ and p, and c_1 was estimated from data published by Tewhey et al. (2016) (Suppl. Text S7). F_i is taken to be the alternate allele frequency in the i^{th} RNA replicate.

We then generate unevenness in coverage between replicates using a linear model. We observed in the Tewhey et al. (2016) data that the mean read count in the smallest replicate is typically not less than ~62% of the mean read count in the largest replicate (Suppl. Text S9.1). Letting R denote the number of RNA replicates, we solve for β and M_i (for $1 \leq i \leq R$) such that:

$$M_1 = 0.62 M_R$$
$$M_i = M_1 + i\beta$$

subject to $\sum_i M_i = N_{RNA}$. M_i is taken to be the total read count per variant for the i^{th} replicate.

We then simulate read counts for alternate and reference alleles in the RNA:

$$R_{alt,i} \mid M_i, F_i \sim \text{binomial}(M_i, F_i)$$
$$R_{ref,i} = M_i - R_{alt,i}$$

We simulated a range of effect sizes (0.50, 0.75, 0.90, 1), a range of minor allele frequencies (0.0001 to 0.5), a range of variant read coverages per variant (30, 50, 100, 500, 1000, 5000, 10000, and 1000000 reads per variant), and different numbers of replicates (1, 2, 5, 10, 25, 100).

We used these simulated data sets to compare the classification accuracy and estimation error of the full BIRD model, handicapped versions of the BIRD model, the Fisher's exact test, the beta-binomial test, and a recently published method called QuASAR-MPRA (Kalita et al., 2018B).

S7. Estimation of Concentration Parameters

To ensure that our simulated data sets were statistically similar to real data, we estimated dispersion parameters from real data sets and then used those dispersion parameters in our simulator (Suppl. Text S6). To estimate the dispersion parameters, we implemented the multi-variant (multi-site) model depicted in Suppl. Fig. S2. This model is similar to the BIRD model, except that it generates multiple variants (sites) simultaneously. The different variants generated by the model share common dispersion parameters s, c_1, and c_2.

We separately applied the multi-variant model to 1000 variants from the Tewhey et al. (2016) LCL data, 372 variants from the adipocyte data, 407 variants from the pre-adipocyte data, and 173 variants from the Guo et al. (2017) HepG2 data. MCMC was used to produce dispersion estimates via the posterior median; separate estimates were produced for each of the four data sets.

Parameter estimates for the four data sets are provided in the following table:

Data	**c1**	**c2**	**s**
Tewhey et al. (2016)	124.6	71.9	0.12
adipocytes	20.7	23.2	1.41
preadipocytes	28.1	24.1	1.19
HepG2 (Guo et al., 2017)	22.9	21.5	1.38

Suppl. Table T7: Parameter estimates for different data sets.

Estimates from the Tewhey et al. (2016) LCL data were used in the simulator as described in Suppl. Text S6, for all simulations except those supporting Suppl. Fig. S20. Simulations based on

parameters estimated from the Guo et al. (2017) HepG2 data were used For Suppl. Fig. S20, for comparison.

S8. Optimized Version of the Model ("Swift")

Because the full BIRD model requires MCMC and is therefore costly to apply to data sets consisting of millions of variants, we also implemented a simplified model that facilitates much faster inference. The model, called Swift (Suppl. Fig. S5), has a simplified structure that enables sampling directly from the joint posterior distribution, without need for MCMC. As a result, inference with the Swift model is approximately 5000 times faster than MCMC-based inference with the full BIRD model (Suppl. Text S9.5).

In order to allow direct sampling from the posterior, we decompose the joint distribution of p and q as:

$$\mathrm{P}(p,q|a,b,k,m,c) = P(p|a,b) \times P(q|p,k,m,c)$$

where read counts are pooled across replicates, so that $a = \sum_i a_i$, $b = \sum_i b_i$, $k = \sum_i k_i$, and $m = \sum_i m_i$. The above decomposition makes the conditional independence assumptions that (1) p is independent of the RNA (and of c) given the DNA, and (2) that q is independent of the DNA given the RNA (and given p). Assuming a uniform prior on p, the two terms on the right in the equation above can be written:

$$P(p|a,b) \propto \mathrm{binomial}(a|a+b,p) \times \mathrm{beta}(1,1) \propto \mathrm{beta}(a+1,b+1)$$

$$P(q|p,k,m,c) \propto \mathrm{binomial}(k|k+m,q) \times \mathrm{beta}(\mathrm{mean}=p,\mathrm{conc}=c) \propto \mathrm{beta}(k+\alpha,m+\beta)$$

for $\alpha = p(c-2)$ and $\beta = (1-p)(c-2)$, for concentration parameter c; this parameterization results in a beta prior on q having mean p, resulting in shrinkage of q toward p. That shrinkage of q toward P results in shrinkage of effect sizes toward 1 as in the full BIRD model.

We draw each sample (p_i,q_i) by first drawing p_i from $P(p|a,b)$ and then drawing q_i from $P(q|p_i,k,m,c)$, rejecting any sample of 0 or 1 for either p_i or q_i. Note that the index i here denotes the MCMC sample number (not replicate). We selected a fixed concentration of $c=100$ based on visual inspection of the resulting prior on q (see below). We assessed classification accuracy and estimation error for several values of c (Suppl. Text. S9.5).

Given a sample (p_i,q_i), we compute θ_i via:

$$\theta_i = \frac{\frac{q_i}{p_i}}{\frac{1-q_i}{1-p_i}}$$

Parameterizing the beta prior on q with $c=100$ as described above induces an implicit prior on θ that shrinks $\log_2\theta$ toward 0 and contains a majority of its mass between a halving and a doubling

of transcription (Suppl. Fig. S6B), similarly to the gamma-lognormal prior used in the full BIRD model (Suppl. Fig. S6A).

We draw 1000 samples for each variant. Summaries based on the posterior median and 95% credible interval are reported as with the full BIRD model.

S9. Supplementary Results

S9.1 Results on Human Data

S9.1.1 Fetal Adiposity Data

174 regions from chromosome 3 were captured from 760 human donors as described previously (Vockley et al., 2016), totaling 70940 bases. A total of 173, 407, and 372 variants having at least 1000 reads were tested in HepG2 cells, pre-adipocytes, and adipocytes, respectively. One replicate of DNA and three, five, and three replicates of RNA were sequenced in HepG2 cells, pre-adipocytes, and adipocytes, respectively. Distributions of read coverages in HepG2 cells for DNA (median = 15731, std. dev. = 26482.1) and RNA (median = 35235; std. dev. = 95318.0) are shown in Suppl. Fig. S13A,B.

To quantify the unevenness in coverage between RNA replicates, for use in generating uneven coverage in our simulator, we define the statistic U_{mean} as follows. For a single site, let N_{min} denote the smallest read count for that site across all replicates, and let N_{max} denote the largest read count for that site across replicates. Define U_{mean} to be the mean ratio N_{min}/N_{max} across all sites. For the HepG2 data set, U_{mean} = 0.59.

When comparing variants with weak effects ($|\log_2(\theta)|<0.5$) to those with stronger effects ($|\log_2(\theta)|>0.5$), and also comparing uncommon (MAF<0.1) to common (MAF>0.1) variants, those with stronger effects were enriched for being uncommon in the general population (Suppl. Fig. S21C; Fisher's exact: p = 0.0003 in preadipocytes, p = 0.04 in adipocytes, p = 1 in HepG2), which would be consistent with natural selection reducing allele frequencies of deleterious variants in open chromatin regions in the GWAS locus. Of 523 tested variants in the GWAS locus (the union across the three cell types), 96 were identified by BIRD as regulatory variants with high posterior probability ($\mathrm{P(regulatory)}>0.99$) in at least one cell type, a proportion of 18.3%, which is a 1.45-fold enrichment over the proportion (12.6% = 3878/30667) in the genome-wide Tewhey et al. data (Suppl. Text S9.1.2).

The GWAS variants were also enriched for variants having a negative effect on transcription (Suppl. Fig. S21A), which would be consistent with disruption of binding sites for transcriptional activators; a similar enrichment of negative effects has been documented by previous studies (e.g., Kwasnieski et al., 2012; Patwardhan et al., 2012). QuASAR-MPRA (Suppl. Text S9.6) estimates were also shifted toward negative effects in the GWAS locus as compared to the Tewhey et al. data (Suppl. Fig. S21B). Comparing medians of BIRD's estimates between the GWAS locus and the Tewhey et al. data, the difference was highly significant (median for GWAS variants = 0.866; median for Tewhey et al. data = 1.00; Wilcoxon W=4965800, p=7.182771e-51). The medians of QuASAR-MPRA's estimates were very similar to BIRD's, and

were also significantly different between the two data sets (median for GWAS variants = 0.848; median for Tewhey et al. data = 0.989; Wilcoxon W=5986000, p=1.150879e-23).

S9.1.2 Tewhey et al. (2016) Data

A total of 30673 variants were synthesized and assayed by Tewhey et al. (2016) using a massively parallel reporter assay (MPRA). Five DNA replicates and thirteen RNA replicates were sequenced; we used only data from LCLs from individuals NA12878 and NA19239. For this data U_{mean} = 0.62. Distributions of read coverages for DNA (median = 7750.3; std. dev. = 4922.2) and RNA (median = 17712.8; std. dev. = 30276.9) are shown in Suppl. Fig. S13C,D.

S9.2 Impacts of Removing Effect Prior from the Model

A modified version of BIRD was created that lacks θ and its prior (Suppl. Fig. S3). Because the model lacks θ, paired samples were drawn for p and q from their joint posterior distribution under the model, using MCMC, and a value of θ was computed for each of these samples via:

$$\theta = \frac{\frac{q}{1-q}}{\frac{p}{1-p}}$$

These values of θ were then used to compute posterior median estimates as in the full BIRD model.

S9.3 Impacts of Pooling Replicates

Two modified versions of BIRD were created that lacked explicit replicates (Suppl. Fig. S4A,B). In these models, counts are pooled (summed) across replicates, and a single binomial is used for the summed DNA read counts:

$$\sum_j a_j \mid \sum_j a_j + b_j,\, p \sim \text{binomial}(\sum_j a_j + b_j,\, p)$$

In model NR1 (Suppl. Fig. S4A), a single binomial is also used for the summed RNA read counts:

$$\sum_j k_j \mid \sum_j k_j + m_j,\, q \sim \text{binomial}(\sum_j k_j + m_j,\, q)$$

In comparison, model NR2 (Suppl. Fig. S4B) incorporates an additional latent variable, d, to allow for greater variance in estimates of q (effectively capturing some of the variability between replicates):

$$\sum_j k_j \mid \sum_j k_j + m_j,\, q \sim \text{binomial}(\sum_j k_j + m_j,\, d)$$
$$d \sim \text{beta}(\text{mode} = q,\, \text{concentration} = c)$$

Despite not modeling replicates, the accuracy of these models increased as larger numbers of replicates were used in simulated data sets with fixed read coverage (Fig. 4A,B, dashed lines). This is due to the fact that sums of binomial variables with different probabilities of success are distributed according to a Poisson-binomial distribution with a variance that decreases with increasing heterogeneity between probabilities of success (i.e., N and $\bar{q}$ are fixed, and σ_{rep}^2 increases):

$$Var(X) = \sum_{i=1}^{N} q_i(1-q_i) = N\bar{q}(1-\bar{q}) - N\sigma_{rep}^2$$

(Poisson, 1837), where X is the alternate allele read count, $\bar{q}$ is the mean of the parameters q_i, and σ_{rep}^2 is the between-replicate variance in the binomial parameters q_i. Here, q_i represents the alternate allele frequency in the i^{th} RNA replicate, and for this analysis we fix N at the total number of reads for the variant. Note that σ_{rep}^2 is not to be confused with the parameters of a model (such as a concentration parameter); rather, it is a property of the data, not of any particular model.

For a series of simulated data sets in which the total reads per variant was fixed (e.g., fix N=1000) and the number of replicates was increased from 1 to 10, the sample variance in read counts for the alternate allele decreased almost monotonically with increasing numbers of replicates, as predicted by the above formula (Suppl. Fig. S11A). Nearly monotonic decreases were also seen in the sample variance of the maximum likelihood estimate of q (Suppl. Fig. S11B), the sample variance of the *ad hoc* estimate of θ (Suppl. Fig S11C), and the RMSE of the *ad hoc* estimate of θ (Suppl. Fig. S11D).

These observations indicate that, for allelic assays, merely generating more experimental replicates can potentially produce more accurate inferences, for any model, even if that model does not distinguish counts from different replicates. That was the case for classification of simulated variants (Figs. 4A,B, dashed lines). However, distinguishing counts between replicates produced additional increases in classification accuracy, as demonstrated by the comparison between the full BIRD model and the models in which reads are pooled across replicates (Figs. 4A,B solid lines versus dashed lines). The full model with replicates had lower Type I error on simulated data than model NR1 that pools read counts across replicates, and lower Type II error than model NR2 that also pools read counts but includes an additional latent variable to accommodate greater variance in estimates of q (Suppl. Fig. S9). The full model also produced credible intervals and point estimates that for many simulation parameters had lower RMSE and/or higher coverage of 95% credible intervals (Suppl. Fig. S10).

S9.4 Coverage of Credible Intervals

BIRD computes a 95% credible interval (a,b) for θ by drawing N samples from the posterior distribution via MCMC and then identifying the largest value a such that $N_{\theta<a}/N \leq 0.025$ and the smallest value b such that $N_{\theta>b}/N \leq 0.025$, where $N_{\theta<a}$ is the number of samples for which $\theta < a$, and $N_{\theta>b}$ is the number of samples for which $\theta > b$.

On variants simulated to have an effect size of 0.5, a version of BIRD in which the simulated dispersion parameters are substituted into the model produced 95% credible intervals that contained the true θ in 94.1% of cases at high read depth (Suppl. Fig. S8, red bars). In practice, true dispersions are unknown, and the BIRD model places a hyperprior on the dispersions, resulting in credible intervals that have somewhat lower coverage (Suppl. Fig. S8, blue bars).

S9.5 Performance of the Swift Model

Our optimized model, Swift, is approximately 5000 times faster than the full BIRD model, due to the fact that BIRD requires MCMC whereas Swift enables direct *i.i.d.* sampling from the posterior. Timing for a single run of BIRD and Swift on 1000 variants on a single CPU is given in Table S9.5:

	Elapsed time
BIRD	41.5 min
Swift	0.5 sec

Suppl. Table T9.5: Timing of BIRD and Swift on 1000 variants on a single CPU.

Because Swift has a simplified structure that ignores replicates, in some cases it is less accurate than BIRD, particularly at low allele frequencies (Suppl. Figs. S16, S17, S18B,C). As a result, for users with relatively small numbers of variants to test, or for variants with low allele frequency in the sample, we recommend using the full BIRD model for increased prediction accuracy.

S9.6 Results for a Competing Model

We also ran the QuASAR-MPRA model (Kalita et al., 2018B) on simulated data at a range of allele frequencies (0.005 to 0.5) and variant read coverages (10 to 5000), and found that accuracy was very high for variants with large allele frequencies, but that BIRD produced higher classification accuracy overall (median BIRD AUC=0.708 versus median QuASAR-MPRA AUC=0.549; two-sided Wilcoxon V = 627, p-value = 2.91e-10), as did Swift (median Swift AUC=0.636 versus median QuASAR-MPRA AUC=0.549; two-sided Wilcoxon V = 595, p-value = 3.821e-07), and that the differences were particularly evident for rare and uncommon variants at high read coverage (MAF$\leq$0.05, read coverage=5000) (Suppl. Fig. S16, S17, S18; Suppl. Table S1, S2). The highest minor allele frequency for which BIRD significantly outperformed QuASAR-MPRA (p-value < 0.01 and AUC difference > 5%) was MAF=0.1 (p-value=0.0008; AUC difference=5.3%). In contrast, we did not detect a difference in predictive accuracy between QuASAR-MPRA and the standard beta-binomial test (median beta-binomial AUC=0.547 versus median QuASAR-MPRA AUC=0.549; two-

sided Wilcoxon V = 349, p-value = 0.3833), finding instead that they correlated very strongly (Spearman rho=0.995; Suppl. Fig. S18A).

The same general trends were seen when the simulator was modified to use concentration parameters estimated from STARR-seq experiments in HepG2 cells (Guo et al., 2017) instead of from MPRA LCL data (Tewhey et al., 2016), as shown in Suppl. Fig. S20 and Suppl. Tables S3, S4. Differences were again significant for BIRD versus QuASAR-MPRA (median BIRD AUC=0.875 versus median QuASAR-MPRA AUC=0.522; Wilcoxon two-sided V = 150, p-value = 7.629e-05). The highest MAF for which BIRD significantly outperformed QuASAR-MPRA was MAF=0.25 (p-value=3.4765e-07; AUC difference=7.4%). Though differences were significant for Swift versus QuASAR-MPRA (median Swift AUC=0.571 versus median QuASAR-MPRA AUC=0.522; Wilcoxon two-sided V = 140, p-value = 0.001343), we consider those differences to be modest.

It should be noted that QuASAR-MPRA makes use of less information than the other three methods. BIRD, Swift, and the beta-binomial test all make use of four pieces of information: allele frequency and read coverage in DNA, and allele frequency and read coverage in RNA. The read coverage provides information regarding dispersion. While QuASAR-MPRA models allele frequency and coverage in RNA, for DNA it has access only to the allele frequency, which is provided by the end user as a fixed-point estimate; DNA read coverage is not provided to the program. Instead of estimating DNA dispersion from DNA read coverage, QuASAR-MPRA estimates its DNA concentration parameter from the RNA, under the assumption that DNA and RNA counts (meaning total counts per variant) are approximately equal. In contrast, BIRD, Swift, and the standard beta-binomial test all make explicit use of read coverage in both DNA and RNA to obtain separate estimates of dispersion for DNA and RNA.

Because QuASAR-MPRA performs substantially better on strongly imbalanced test sets than under balanced test sets (Suppl. Fig. S19), we tested QuASAR-MPRA only on imbalanced data, and omitted it from all other analyses that used balanced data. (Balanced data sets were used in all other simulations, as balanced test data produces smoother ROC curves and therefore more accurate AUC estimates.)

For variants captured from individuals and assayed directly via STARR-seq, allele frequencies are expected to be skewed toward 0 rather than 0.5, as in our fetal adiposity data (Suppl. Fig. S12A). As such, for data with skewed allele frequencies, BIRD is recommended, due to its better performance on rare variants.