

1 Bayesian modelling of high-throughput sequencing assays with malacoda

2

3 Andrew R. Ghazi¹, Xianguo Kong², Ed S. Chen³, Leonard C. Edelstein², Chad A. Shaw³

4 1. Quantitative and Computational Biosciences, Baylor College of Medicine, Houston,

5 Texas

6 2. Cardeza Foundation for Hematologic Research, Thomas Jefferson University,

7 Philadelphia, Pennsylvania

8 3. Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas

9

10 **Abstract**

11 NGS studies have uncovered an ever-growing catalog of human variation while leaving
12 an enormous gap between observed variation and experimental characterization of variant
13 function. High-throughput screens powered by NGS have greatly increased the rate of variant
14 functionalization, but the development of comprehensive statistical methods to analyze screen
15 data has lagged behind. In the massively parallel reporter assay (MPRA), short barcodes are
16 counted by sequencing DNA libraries transfected into cells and output RNA in order to
17 simultaneously measure the shifts in transcription induced by thousands of genetic variants.
18 These counts present many statistical challenges, including over-dispersion, depth dependence,
19 and uncertain DNA concentrations. So far, the statistical methods used have been rudimentary,
20 employing transformations on count level data and disregarding experimental and technical
21 structure while failing to quantify uncertainty in the statistical model.

22

23 We have developed an extensive framework for the analysis of NGS functionalization
24 screens available as an R package called malacoda (available from
25 github.com/andrewGhazi/malacoda). Our software implements a probabilistic, fully Bayesian
26 model of screen data. The model uses the negative binomial distribution with gamma priors to
27 model sequencing counts while accounting for effects from input library preparation and
28 sequencing depth. The method leverages the high-throughput nature of the assay to estimate the
29 priors empirically. External annotations such as ENCODE data or DeepSea predictions can also
30 be incorporated to obtain more informative priors – a transformative capability for data
31 integration. The package also includes quality control and utility functions, including automated
32 barcode counting and visualization methods.

33 To validate our method, we analyzed several datasets datasets using malacoda and
34 alternative MPRA analysis methods. These data include experiments from the literature,
35 simulated assays, and primary MPRA data. We also used luciferase assays to experimentally
36 validate the strongest hits from our primary data, as well as variants for which the various
37 methods disagree and variants detectable only with the aid of external annotations.

38 **Author Summary**

39 Genetic sequencing technology has progressed rapidly in the past two decades. Huge genomic
40 characterization studies have resulted in a massive quantity of background information across the
41 entire genome, including catalogs of observed human variation, gene regulation features, and
42 computational predictions of genomic function. Meanwhile, new types of experiments use the
43 same sequencing technology to simultaneously test the impact of thousands of mutations on gene
44 regulation. While the design of experiments has become increasingly complex, the data analysis
45 methods deployed have remained overly simplistic, often relying on summary measures that

46 discard information. Here we present a statistical framework called for the analysis of massively
47 parallel genomic experiments designed to incorporate prior information in an unbiased way. We
48 validate our method by comparing our method to alternatives on simulated and real datasets, by
49 using different types of assays that provide a similar type of information, and by closely
50 inspecting an example experimental result that only our method detected. We also present the
51 method's accompanying software package which provides an end-to-end pipeline that provides
52 a simple interface for data preparation, analysis, and visualization.

53 **Introduction**

54 The advent of next generation sequencing (NGS) has generated an explosion of observed
55 genetic variation in humans. Variants with unclear effects greatly outnumber those with obvious,
56 severe impact; the 1000 Genomes Project [1] has estimated that a typical human genome has
57 roughly 150 protein-truncating variants, 11,000 peptide-sequence altering variants, and 500,000
58 variants falling into known regulatory regions. Simultaneously, genome-wide association studies
59 (GWAS) have found strong statistical associations between thousands of noncoding variants and
60 hundreds of human phenotypes [2,3]. Traditional methods of assessing the regulatory impact of
61 variants are slow and low-throughput: luciferase reporter assays require multiple replications of
62 cloning individual genomic regions, transfection into cells, and measurement of output intensity.

63 Massively Parallel Reporter Assays (MPRA), overviewed in Figure 1, were developed to
64 assess simultaneously the transcriptional impact of thousands of genetic variants [4]. The
65 simplest form of MPRA uses a carefully designed set of barcoded oligonucleotides containing
66 roughly 150 base pairs of genomic context surrounding variants of interest. There are typically
67 thousands of variants selected by preliminary evidence from GWAS, and there are usually ten to

68 thirty replicates of each allele with different barcodes. The oligonucleotides are cloned into
69 plasmids, making a complex library that is then transfected into cells. The cells use the library as
70 genetic material and actively transcribe the inserts. Because the barcodes are preserved by
71 transcription, counting the RNA products of each variant construct by re-identifying each
72 barcode in the NGS product provides a direct measure of the transcriptional output of a given
73 genetic variant. By designing the oligonucleotide library to contain multiple barcodes of both the
74 reference and alternate alleles for each variant, one can statistically assess the transcription shift
75 (TS) for each variant.

76 **Fig1. Diagram of MPRA** MPRA simultaneously assess the transcription shift of thousands of
77 variants. The diagram shows six constructs with two variants, but in practice the size of the
78 oligonucleotide library is only limited by cost. A typical MPRA has tens to hundreds of
79 thousands of oligonucleotides to assay thousands of variants.

80 MPRA have successfully identified many transcriptionally functional variants [5, 6, 7],
81 but the accompanying statistical analyses have been rudimentary. Initial studies focused on the
82 computation of the “activity” for each barcode in each RNA sample. This involves averaging
83 across depth-adjusted counts to compute a normalizing DNA factor for each barcode, then
84 dividing RNA counts by the DNA factor and taking the log of this ratio. Then a t-test is used to
85 compare the activity measurements for each allele, followed by assay-wide multiple-testing
86 corrections. The key limitations include ignoring systematic variation due to unknown DNA
87 concentrations, the application of heavy transformation and summarization to the data prior to
88 modelling, and the failure to include the reservoir of prior data and biological knowledge
89 concerning genes and genomic regions. The methods mpralm [8] and MPRAscore [9] are more
90 recent methods, but they suffer from a number of limitations: failure to model variation in input

91 DNA concentrations, aggregation of data across barcodes and sequencing samples without
92 modeling systematic sources of variation, and over-reliance on point estimates of dispersion that
93 cause systematic errors in transcription shift estimates.

94 Other areas of genomic analysis have generated a wealth of information on genomic
95 structure and function, frequently specific to particular genomic contexts and variants. For
96 example, the ENCODE project [10] provides genome-wide ChIP-seq data on transcription
97 binding profiles, histone marks, and DNA accessibility. Computational methods such as
98 DeepSea [11] use machine learning to provide variant-specific predictions on chromatin effects.
99 Genome-wide databases like ENCODE and computational predictors like DeepSea contain real
100 information about variant effects, but the method for incorporating this information into a
101 statistical framework for experimental analysis of variants has been unclear.

102 We hypothesized that a Bayesian approach to high throughput NGS screens such as
103 MPRA would improve statistical sensitivity and specificity and yield more accurate estimates of
104 variant function, particularly when incorporating prior information. The Bayesian approach
105 offers a flexible modeling system that can flexibly fit hierarchical model structures of count data
106 while also directly accounting for experimental sources of variation. The Bayesian approach also
107 enables the integration of prior information and probabilistic modelling of dispersion parameters.
108 These advantages offer significant improvements in statistical efficiency and provide advantages
109 for formulating systems-level hypotheses -- for example, the impact of specific transcription
110 factors -- that are absent from other approaches. Here we present *malacoda*, an end-to-end
111 Bayesian statistical framework that addresses the gaps in the prior approaches while providing
112 novel methods for incorporating prior information. The malacoda method centers on MPRA but
113 also has potential extension to a broad array of NGS-based high-throughput screens. We

114 establish the superior performance of malacoda on MPRA compared to alternatives using
115 simulation studies. Then, we apply the method to previously published findings to make new
116 biological discoveries that we explore in the paper. We also apply malacoda to primary MPRA
117 studies that we performed. The results demonstrate that using malacoda we can discover
118 biologically important findings that were missed by prior approaches. We have made the
119 software available as an open source R package on GitHub.

120 **Methods**

121 **Overview**

122 In malacoda we utilize a negative binomial model for NGS to consider barcode counts
123 with empirically estimated gamma priors, and we explicitly model variation in the input DNA
124 concentrations for each barcode. By default the method marginally estimates the priors from the
125 maximum likelihood estimates of each variant in the assay; the method also supports informative
126 prior estimation by using external genomic annotations for each variant as weights. This
127 approach enables disparate knowledge sources to inform the results in a principled, systematic,
128 and calculation. The probabilistic model underlying malacoda uses the NGS data directly
129 without transformation, and it accounts for all known sources of experimental variation and
130 uncertainty in model parameters. Finally, the method provides estimate shrinkage as a method
131 for avoiding false positives.

132 **Description of the statistical model**

133 MPRA data are the counts of the barcoded DNA input from sequencing the plasmid
134 library and counts of the barcoded RNA outputs from sequencing the RNA content extracted

135 from passaged cells. The DNA counts vary according to the sequencing depth of the sample as
136 well as due to the inherent noise in library preparation. The RNA measurements also vary
137 according to sequencing depth, but they are also affected by the DNA input concentration and
138 the inherent transcription rate of their associated region of genomic context. Figure 2A shows a
139 subset of a typical MPRA dataset, with two barcodes of each allele for two variants and several
140 columns of counts. We find that typically MPRA are performed with four to six RNA
141 sequencing replicates and a smaller number of DNA replicate samples. Figure 2B shows a
142 simplified Kruschke diagram of the model underlying malacoda, using the mean-dispersion
143 parameterization of the negative binomial. More explicitly,

$$144 \quad \text{Counts}_{DNA} \sim \text{NegBin}(\text{depth}_s \cdot \mu_{bc} \cdot \varphi_{DNA})$$

$$145 \quad \text{Counts}_{RNA} \sim \text{NegBin}(\text{depth}_s \cdot \mu_{bc} \cdot \mu_{allele}, \varphi_{RNA})$$

146 Where depth_s indicates the depth of a particular sequencing sample, μ_{bc} indicates the
147 unknown concentration of a particular barcode in the plasmid library, and μ_{allele} indicates the
148 effect of the genomic context of a given allele of a given variant. There are separate dispersions
149 parameters φ for both DNA and RNA. The means μ and dispersions φ come from their own
150 gamma priors.

151 The negative binomial distribution is a natural choice for modelling NGS count data
152 given its ability to accurately fit overdispersed observations frequently seen in sequencing data
153 [12]. Briefly, the observed dispersion in NGS count data usually exceeds that expected from
154 simpler binomial or Poisson models. We chose gamma distributions as priors for several reasons.
155 They have the appropriate $[0, \infty)$ support, and for a non-negative random variable whose
156 expectation and expected log exist, they are the maximum entropy distribution. Additionally,

157 they are characterized by two parameters, allowing the prior estimation process to accurately fit
158 the observed population of negative binomial estimates. Probabilistic modelling of the dispersion
159 parameters is key -- as demonstrated by simulation in S1 Appendix. This practice helps avoid
160 pitfalls common to methods based on point estimates of dispersion parameters. The barcode-
161 level count data model is a central contribution of the malacoda method.

162 **Fig 2. MPRA data and malacoda priors** A) The table shows a subset of our primary MPRA
163 data. Highlighted cell containing 759 is influenced both by the sequencing depth of its sample
164 (column) and the unknown input DNA concentration of its barcode (row). B) A simplified
165 Kruschke diagram of the generative model underlying malacoda C) A conceptual diagram
166 demonstrating three prior types available from malacoda. The marginal prior (left) weights all
167 variants in the assay equally, while the grouped and conditional priors utilize informative
168 annotations as weights in the prior estimation process.

169 After computing the joint posterior on all model parameters, the posterior on transcription
170 shift is computed as a generated quantity by taking the difference between log means of the
171 alternate and reference alleles. 95% highest density interval on TS is used to make binary calls
172 on whether a variant is functional or non-functional. If the interval excludes zero as a credible
173 value, the variant is labelled as functional. An optional “region of practical equivalence” can be
174 defined on a per-assay basis when there is particular interest in rejecting transcription shift values
175 around zero [13].

176 **Empirical priors**

177 The gamma priors are fit empirically by maximum likelihood estimation. Specifically,
178 each variant-level model is first fit by maximizing the likelihood component of the malacoda

179 model, then gamma distributions are fit to those estimates for the means and dispersions of the
180 DNA, reference RNA, and alternate RNA. This approach offers several benefits. First, it
181 leverages the high-throughput nature of the assay. The full dataset determines the prior; in
182 situations with thousands of variants the individual contribution of each variant to the prior is
183 negligible. Secondly, it constrains the prior to be reasonable in the context of a given assay.
184 Specific circumstances regarding library preparation, sequencer properties, cell culture
185 conditions, and other unknown factors will cause the underlying statistical properties of each
186 MPRA to be unique. A less informed, general-purpose prior, such as $\text{gamma}(\alpha = .001, \beta = .001)$,
187 would assign a considerable amount of probability density to unreasonable regions of parameter
188 space. Empirical estimation ensures that the priors capture the reasonable range of values for
189 each parameter while avoiding putting unwarranted density on extreme values [14]. Finally, by
190 sharing information between variants, empirical priors provide estimate shrinkage. The prior
191 effectively regularizes all parameter estimates, a behavior which is important in multi-parameter
192 models with relatively little data per parameter. This in turn acts as a natural safeguard against
193 false positives, thus removing the need for *post hoc* multiple testing correction.

194 In order to incorporate external knowledge, the malacoda method also allows users to
195 provide arbitrary annotations to supplement the analysis. Figure 2C contrasts the marginal prior
196 estimation (left) with two prior types that make use of external annotations. These priors make
197 use of the information in the annotations by employing the principle that similarly annotated
198 variants should perform similarly in the assay. When the annotations are simply a set of
199 descriptive categories (for example predictions of likely benign, uncertain, or likely functional),
200 the grouped prior (2C, center) simply fits a prior distribution within each subset. When the
201 annotations are continuous values, the conditionally weighted (2C, right) prior employs a kernel

202 smoothing process to estimate the prior. To estimate the prior for a single variant, it initializes a
203 t-distribution kernel centered at the annotation of the variant in question, then gradually widens
204 this kernel until the n -th most highly weighted variant (where n is a configurable tuning
205 parameter defaulting to 100) has a weight of at least one percent of that of the most influential
206 variant. While the diagram in figure 2C shows this for only a single informative annotation on
207 the horizontal axis, the code allows for an arbitrary number of continuous predictors to be used.

208 **Simulation and Validation Studies**

209 We took several approaches to validate and compare the malacoda method with
210 alternatives. First, we simulated MPRA data using across a realistic grid of parameters governing
211 the fraction of truly functional variants, the number of variants in the assay, and the number of
212 barcodes per allele. These simulations also modelled distinct sequencing samples, varying
213 sequencing depth, and barcode failure during library preparation. We then compared malacoda to
214 alternative methods including the t-test, mpralm, and MPRAscore. Across these simulations we
215 compared performance metrics such as area-under-curve (AUC) and estimate accuracy.
216 Secondly, we applied malacoda and alternative methods to real MPRA data from the Ulirsch
217 dataset [5], using inter-method consensus as a performance metric. We repeated this with our
218 own primary MPRA data on variants related to platelet function. Finally, we tested a subset of
219 variants where the various methods disagreed with luciferase reporter assays to assess
220 consistency with MPRA estimates of variant function.

221 **Software**

222 Our method is available as an R package from github.com/andrewGhazi/malacoda. The
223 package includes detailed installation instructions, extensive help documentation, an analysis

224 walkthrough vignette, and implementations of traditional activity-based analysis methods. The
225 package also includes functionality to extract, quality-filter, and count barcodes from a set of
226 FASTQ files through an application of the FASTX-Toolkit [15]. Through an interface with the
227 FreeBarcodes package [16], the package can also decode sequencer errors in the barcodes of an
228 assay that has been designed using our previous work, mpradesigntools [17]. In our experience
229 this typically recaptures about 5% additional data with no additional cost beyond a line of code
230 during the assay design. The package also contains plotting functionality to help visualize the
231 results of analyses.

232 **Experimental Methods**

233 In order to collect experimental measurements of the transcriptional impact of variants
234 through means other than MPRA, we performed luciferase reporter assays on sixteen variants.
235 Four were among the strongest signals detected in our MPRA, six were variants from our MPRA
236 where the statistical methods disagreed, and six were variants from the Ulirsch dataset [5] where
237 the malacoda marginal and DeepSea-based [11] conditional prior model fits disagreed.

238 150-200bp genomic DNA sequences flanking the variants were amplified by PCR using
239 K562 lymphoblast (ATCC) genomic DNA as template, then cloned into PGL4.28 minimum
240 promoter luciferase reporter vector (Promega) at NheI and HindIII sites. Counterpart SNP
241 variants were generated by site-directed mutagenesis. All the constructs were validated by DNA
242 sequencing. 3 μ g plasmid preparations were co-transfected with 0.5 μ g β -gal plasmid into 1x10⁶
243 of K562 cells with Lipofectamine 2000 based on manufacturer's instructions. Each assay was
244 repeated with 3 independent plasmid preparations. 24 hours post transfection, luciferase and β -
245 gal were measured. Luciferase units were then normalized to β -gal values.

246 **Results**

247 **Simulation Studies**

248 We evaluated our simulation results in three ways. First, we focused on the accuracy of
249 transcription shift estimates. Figure 3A shows the results of analyzing one simulated dataset,
250 with the true value of the simulation's transcription shift plotted on the x-axis, with the model
251 estimates on the y-axis. For each fit of each simulation using each analysis method, we analyzed
252 performance using two metrics: standard deviation of estimates for truly non-functional variants
253 at zero (center dots, lower is better) and correlation with the truth for truly functional variants
254 with nonzero effects (off-center dots, higher is better).

255

256 **Fig 3. Simulation results** A) The figure compares TS values used to generate simulated data to
257 malacoda TS estimates. Simulated MPRA assays use a varying fraction of variants that are truly
258 non-functional (center). B) ROC curves assess the performance of each method on a randomly
259 selected assay with 3000 variants, 5% truly functional variants, and 10 barcodes per allele. C)
260 Performance metrics averaged across multiple simulations under the same conditions as B. D) A
261 scatterplot demonstrates the relationship between luciferase-based estimates of TS against
262 MPRA-based estimates.

263

264 Second, we also computed area under the curve (AUC) for each method. Bayesian
265 methods such as malacoda explicitly do not consider a null hypothesis and therefore do not
266 output p-values; in order to create an analogous output quantity to derive an ROC curve we
267 instead computed one minus the minimum HDI width necessary to include zero as a credible
268 transcription shift value to distinguish true and false positives. Figure 3B shows the ROC curves

269 by method for a randomly chosen simulation with ten barcodes per allele, 5% truly functional
270 variants, and 3000 variants. Figure 3C shows that across all simulations with these
271 characteristics, malacoda consistently showed the highest median AUC, highest correlation with
272 the truth for functional variants, and the lowest spread among estimates of truly nonfunctional
273 variants. Other simulation grid points are shown in S2 Appendix, and these display similar
274 patterns.

275 In order to examine the performance of malacoda on real data, we applied the various
276 methods to both the Ulirsch data [5] and to our own primary dataset. Unlike the case with
277 simulations, the underlying true values are not known. However, inter-method consensus can
278 serve as a performance metric -- alternative methods presumably fail in different ways, so if they
279 tend to disagree with one another but agree with malacoda, that would imply that malacoda is
280 working well across the cases where others fail. Indeed, Figure 4 shows that the other methods
281 tend to correlate with malacoda better than the other alternatives. The one exception is when
282 applied to our dataset, mpralm tends to agree best with the t-test method. Given that linear
283 models underlie both mpralm and the t-test method, it seems plausible that they would
284 sometimes show similar results.

285

286 **Fig 4. Inter-method consensus** A) A pairwise plot of TS estimates in our MPRA, showing that
287 other methods generally agree with malacoda more than each other. Color indicates local density
288 of points. B) A pairwise plot of TS estimates using both the marginal and DeepSea-based
289 malacoda priors in the Ulirsch dataset, showing a similar outcome.

290 **Biological results**

291 The number of luciferase reporter assays we performed was not enough to overcome the
292 amount of noise inherent to light intensity-based measurements, thus we did not have enough
293 data to clearly demonstrate that any of the MPRA analysis methods outperform the others in
294 terms of correlation with luciferase results. However, the results show that the various methods
295 are consistent with MPRA-based estimates Figure 3D, providing further evidence that MPRA
296 results are biologically realistic.

297 We closely inspected a particular biological discovery to demonstrate malacoda's ability
298 to identify low-signal variants. One of the functional variants we identified with malacoda using
299 the DeepSea-based conditional prior in the Ulirsch dataset [5] is rs11865131; this variant is
300 identified by malacoda but not by any of the other methods. The variant rs11865131 is in an
301 intron within the *NPRL3* gene which encodes the Natriuretic Peptide Receptor Like 3 protein.
302 *NPRL3* is part of the GTP-ase activating protein activity toward Rags [18] (GATOR1) complex.
303 The GATOR1 complex inhibits mammalian target of rapamycin (*MTOR*) by inhibiting *RRAGA*
304 function (reviewed in [18] *MTOR* signaling has been implicated in platelet aggregation and
305 spreading in addition to aging associated venous thrombosis [19, 20]. Analysis of the
306 rs11865131 locus indicates that it colocalizes with ENCODE ChIP-Seq peaks for 36
307 transcription factors in K562 erythroleukemia cells as well as containing enhancer histone
308 epigenetic marks. Together, these data indicate that this is likely an important regulatory region.
309 In addition to the heterologous K562 cell line, data from cultured megakaryocytes indicates that
310 rs11865131 lies within *RUNXI* and *SCL* ChIP-Seq peaks, two well-studied megakaryopoietic
311 transcription factors [21]. This agrees with our data that platelet *NPRL3* mRNA is positively
312 associated with platelet count in healthy humans [22, 23]. These data indicate that malacoda has

313 identified a likely important regulatory region for megakaryocytes and platelets that was missed
314 by other MPRA analysis methods.

315 **Discussion**

316 We developed a fully Bayesian framework for the analysis of NGS high throughput
317 screens with specific application to MPRA studies. The method is an advance in statistical and
318 computational science for these data - a fully Bayesian model that probabilistically incorporates
319 all known sources of variation. The method does a better job of identifying true positives in
320 simulated data and performs well in empirical studies. The method identified a previously missed
321 functional variant in the *NPRL3* gene that has confirmatory evidence from a variety of other
322 studies. Particular advantages of the method are accurate estimation of variant effects, the
323 treatment of the dispersion parameter in both estimation and inference, and the potential to
324 incorporate informative prior information.

325 The functional discovery of the variant rs11865131 represents a demonstration of the
326 power of the malacoda method to identify biologically important results missed by alternative
327 methods. This variant lies in an intronic region of the gene *NPRL3*, and protein coding
328 approaches to variant analysis would overlook this regulatory variant. Multiple lines of evidence
329 point to the biological relevance of this variant, including epigenetic and transcription factor
330 binding data as well as evidence of association with platelet count in healthy humans.

331 There are downsides to our method. First, Bayesian methods that estimate a joint
332 posterior on many parameters by MCMC are significantly slower than optimization approaches.
333 To address this, we fit our models with Stan [24], which allows us to perform a first pass fit with
334 Automatic Differentiation Variational Inference [25] and, if seemingly worthwhile, to perform a
335 final fit with Stan's state-of-the-art No-U-Turn Sampler. Despite this measure, our marginal prior

336 analysis of 8251 variants from the Ulirsch dataset with 50,000 MCMC samples using no
337 variational first pass took over fifteen hours when parallelized across eighteen threads on two
338 Intel Xeon X5675 3.07GHz processors. Nevertheless, an analysis that runs in hours is reasonable
339 for an assay that takes weeks to perform.

340 Secondly, the efficacy of our method does not account for uncertainty in our empirical
341 prior estimation functionality [14]. The R package includes a fully hierarchical model that adds
342 an additional layer of hyperparameters in order to probabilistically model the gamma priors and
343 all other parameters for an entire MPRA dataset at once, but this approach falls outside the
344 intended scope of the malacoda framework. This model, featuring hundreds of thousands of
345 parameters, is presently too complex to fit in practice.

346 The statistical method and validation work presented in this article has focused primarily
347 on the analysis of “typical” MPRA: two alleles per variant, in a single tissue type, with no other
348 experimental perturbations. However, we have expanded the modelling capabilities of the
349 package beyond these limitations. Models tailored to more exotic experimental structures, such
350 as arbitrary numbers of alleles per variant, multiple tissue types, or cell-culture perturbations, are
351 also included with the package. We also have expanded the model framework included in the
352 package beyond MPRA into CRISPR screen modelling: the counts of gRNAs targeting specific
353 genes in survival/dropout screens can make use of an analogous negative binomial structure with
354 similar empirical gamma priors. This opens the path to incorporating gene-level annotations into
355 Bayesian CRISPR screen analysis.

356 Sophisticated high-throughput assays are a central component to the future of genomics.
357 Therefore, the statistical methods used for these data should be as efficient as possible,
358 accounting for all sources of variation and quantifying the resulting uncertainty. Our software,

359 malacoda, provides an end-to-end framework for the probabilistic analysis of MPRA data.
360 Through our well-documented, easy-to-use R package, users can perform sequencing error
361 correction and data pre-processing before executing a fully Bayesian analysis in as little as two
362 lines of code. When informative annotations on variant function are available, malacoda is
363 capable of taking full advantage through a conditional prior estimation process. We hope that
364 this work may act as a stepping stone towards further integrative, probabilistic analysis in the
365 field of high-throughput genomics.

366

367 **References**

368

- 369 1. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A
370 global reference for human genetic variation. *Nature* [Internet]. 2015;526(7571):68–74.
371 doi: 10.1038/nature15393
- 372 2. Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al.
373 Potential etiologic and functional implications of genome-wide association loci for
374 human diseases and traits. *Proc Natl Acad Sci USA*. 2009;106(23):9362–7. doi:
375 10.1073/pnas.0903103106
- 376 3. Nishizaki SS, Boyle AP. Mining the Unknown: Assigning Function to Noncoding Single
377 Nucleotide Polymorphisms. *Trends Genet* [Internet]. 2017;33(1):34–45. doi:
378 10.1016/j.tig.2016.10.008
- 379 4. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic
380 dissection and optimization of inducible enhancers in human cells using a massively

- 381 parallel reporter assay. *Nat Biotechnol* [Internet]. 2012;30(3):271–7. doi:
382 10.1038/nbt.2137
- 383 5. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic
384 functional dissection of common genetic variation affecting red blood cell traits. *Cell*
385 [Internet]. 2016;165(6):1530–45. doi: 10.1016/j.cell.2016.04.048
- 386 6. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification
387 of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*
388 [Internet]. 2016;165(6):1519–29. doi: 10.1016/j.cell.2016.04.027
- 389 7. Shen SQ, Myers CA, Hughes AEO, Byrne LC, Flannery JG, Corbo JC. Massively
390 parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.*
391 2016;26(2):238–55. doi: 10.1101/gr.193789.115
- 392 8. Myint L, Avramopoulos DG, Goff LA, Hansen KD. Linear models enable powerful
393 differential activity analysis in massively parallel reporter assays. *BMC Genomics.*
394 2019;20(1):1–19. doi: 10.1186/s12864-019-5556-x
- 395 9. Niroula A, Ajore R, Nilsson B. MPRAscore: robust and non-parametric analysis of
396 massively parallel reporter assays. *Bioinformatics.* 2019;(July):1–3. doi:
397 10.1093/bioinformatics/btz591
- 398 10. Consortium EP. An integrated encyclopedia of DNA elements in the human genome.
399 *Nature.* 2013;489(7414):57–74. doi: 10.1038/nature11247.An
- 400 11. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-
401 based sequence model. (DeepSea). *Nat Methods* [Internet]. 2015;12(10):931–4. doi:
402 10.1038/nmeth.3547

- 403 12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
404 RNA-seq data with DESeq2. *Genome Biol* [Internet]. 2014;15(12):550. doi:
405 10.1186/s13059-014-0550-8
- 406 13. Kruschke J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed.
407 London: Academic Press; c2015. P.336-40.
- 408 14. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data*
409 *Analysis*. Third Edition. Boca Raton, FL: CRC Press; 2013. p. 51-6, p. 102-4.
- 410 15. Assaf G, Hannon GJ. FASTX-Toolkit [Internet]. 2010. Available from:
411 http://hannonlab.cshl.edu/fastx_toolkit/index.html
- 412 16. Hawkins JA, Jones SK, Finkelstein IJ, Press WH. Indel-correcting DNA barcodes for
413 high-throughput sequencing. *Proc Natl Acad Sci* [Internet]. 2018;115(27):E6217–26. doi:
414 10.1073/pnas.1802640115
- 415 17. Ghazi AR, Chen ES, Henke DM, Madan N, Edelstein LC, Shaw CA. Design tools for
416 MPRA experiments. *Bioinformatics*. 2018;34(15):2682–3. doi:
417 10.1093/bioinformatics/bty150
- 418 18. Shaw RJ. GATORs take a bite out of mTOR. *Science*. 2013;340(6136):1056–7. doi:
419 10.1126/science.1240315
- 420 19. Aslan JE, Tormoen GW, Loren CP, Pang J, McCarty OJT. S6K1 and mTOR regulate
421 Rac1-driven platelet activation and aggregation. *Blood*. 2011;118(11):3129–36. doi:
422 10.1182/blood-2011-02-331579
- 423 20. Yang J, Zhou X, Fan X, Xiao M, Yang D, Liang B, et al. MTORC1 promotes aging-
424 related venous thrombosis in mice via elevation of platelet volume and activation. *Blood*.
425 2016;128(5):615–24. doi: 10.1182/blood-2015-10-672964

- 426 21. Chacon D, Beck D, Perera D, Wong JWH, Pimanda JE. BloodChIP: A database of
427 comparative genome-wide transcription factor binding profiles in human blood cells.
428 Nucleic Acids Res. 2014;42(D1):172–7. doi: 10.1093/nar/gkt1036
- 429 22. Simon LM, Edelstein LC, Nagalla S, Woodley AB, Chen ES, Kong X, et al. Human
430 platelet microRNA-mRNA networks associated with age and gender revealed by
431 integrated plateletomics. Blood. 2014;123(16):37–45. doi: 10.1182/blood-2013-12-
432 544692
- 433 23. Edelstein LC, Simon LM, Montoya RT, Holinstat M, Chen ES, Bergeron A, et al. Racial
434 differences in human platelet PAR4 reactivity reflect expression of PCTP and miR-376c.
435 Nat Med [Internet]. 2013;19(12):1609–16. doi: 10.1038/nm.3385
- 436 24. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M,
437 Guo J, Li P, Riddell A. Stan: A probabilistic programming language. J Stat Softw.
438 2017;76(1). doi: 10.18637/jss.v076.i01
- 439 25. Kucukelbir A, Blei DM, Gelman A, Ranganath R, Tran D. Automatic Differentiation
440 Variational Inference. J Mach Learn Res. 2017;18:1–45. Available from:
441 <https://arxiv.org/abs/1603.00788>

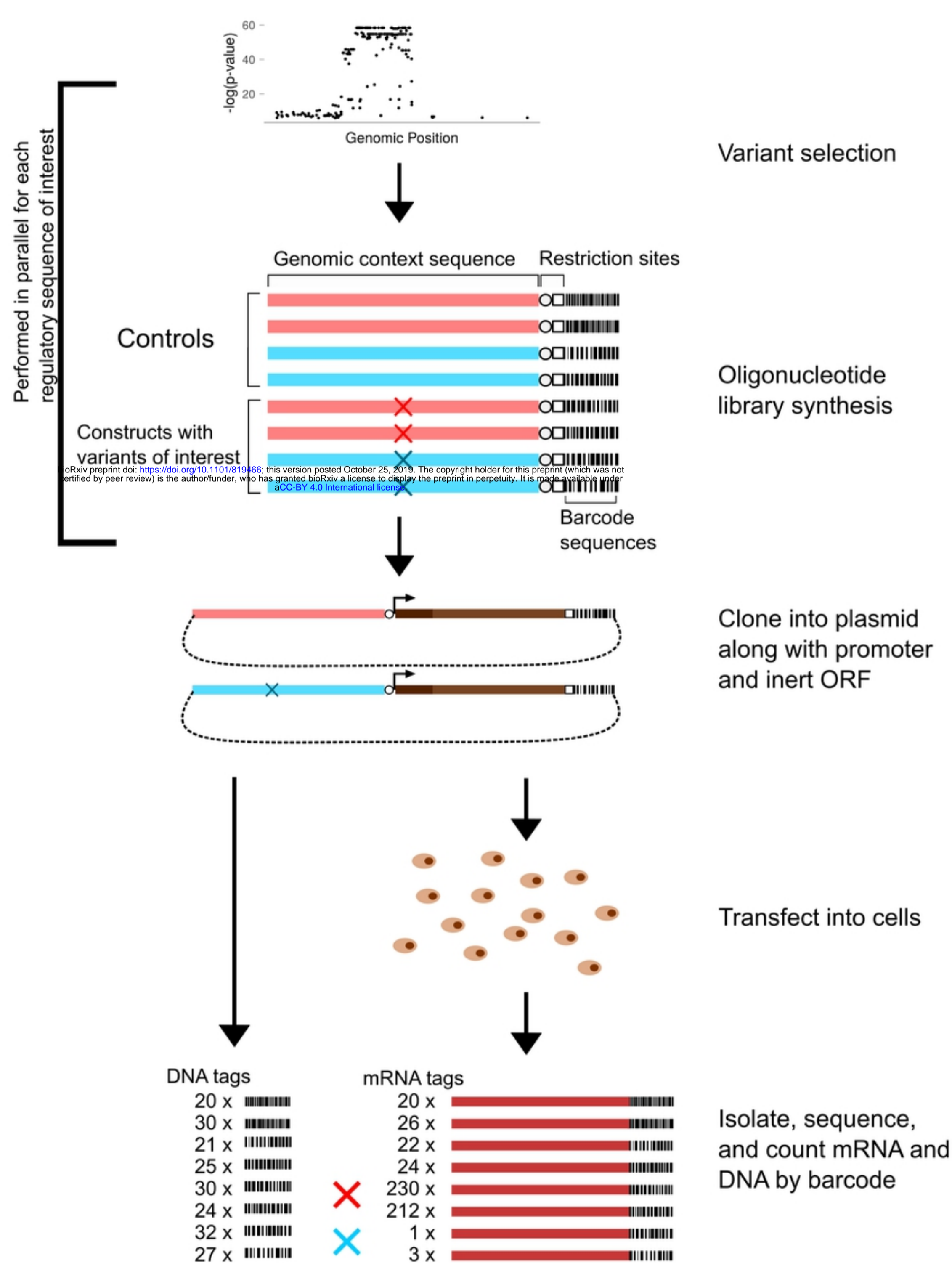
442 **Supporting Information**

443 **S1 Appendix. Negative Binomial variance estimation.**

444 **S2 Appendix. Simulation details and extended results.**

445 **S3 Dataset. RData file of luciferase and MPRA results.** An RData file that loads two objects:
446 `luc_results`, a table of the luciferase results, and `mpra_results`, giving the primary data on MPRA
447 counts for the variants tested with luciferase

448 **S4 Dataset. RData file of estimate comparisons.** The data necessary to produce Figure 4. An
449 RData file that contains two data frames: `ulirsch_comparisons` and `primary_comparisons`. Each
450 row corresponds to one variant, and each column corresponds to a given analysis method. The
451 values in the table give the transcription shift estimates.



A

variant_id	allele	barcode	MPRA_DNA1	MPRA_DNA2	MPRA_RNA1	MPRA_RNA2	MPRA_RNA3
7_79758455_C_T_CD36	ref	GCCATAAGCAGTCT	473	788	3329	8337	5106
		TTACGAATAGTGCG	362	549	3571	7342	4259
	alt	TAGCTGTTCTGAC	1807	2887	1788	4422	3166
		ATGCCGTTGCGATT	48	48	40	0	48
rs11749731	ref	AACCGTCGCGTAGT	543	868	248	759	489
		CACGCAATGTCTTA	173	246	93	89	75
	alt	CTTCGTACTIONTCC	412	638	370	685	707
		AGGACGCAATACAA	284	569	520	1107	1090
rs2236053	ref	CTACCGCGTCACTA	457	660	1616	3875	3164
		CTACCGCGTGTAGG	69	123	314	366	365
	alt	ATCTGTGCGCGCTAT	165	248	540	998	593
		TAGCGTGTACTTCA	1122	1708	3819	8397	6181

bioRxiv preprint doi: <https://doi.org/10.1101/819466>; this version posted October 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

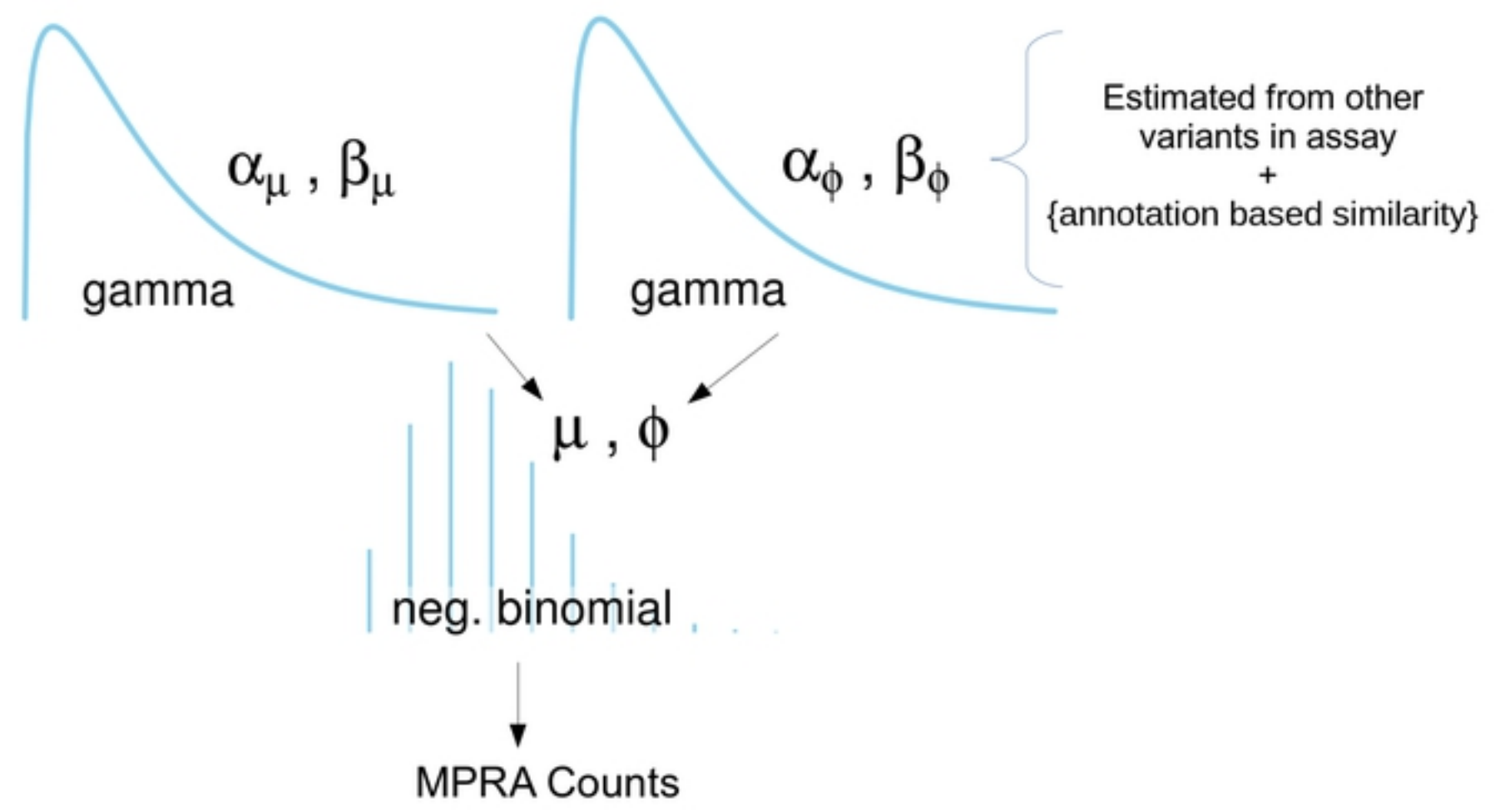
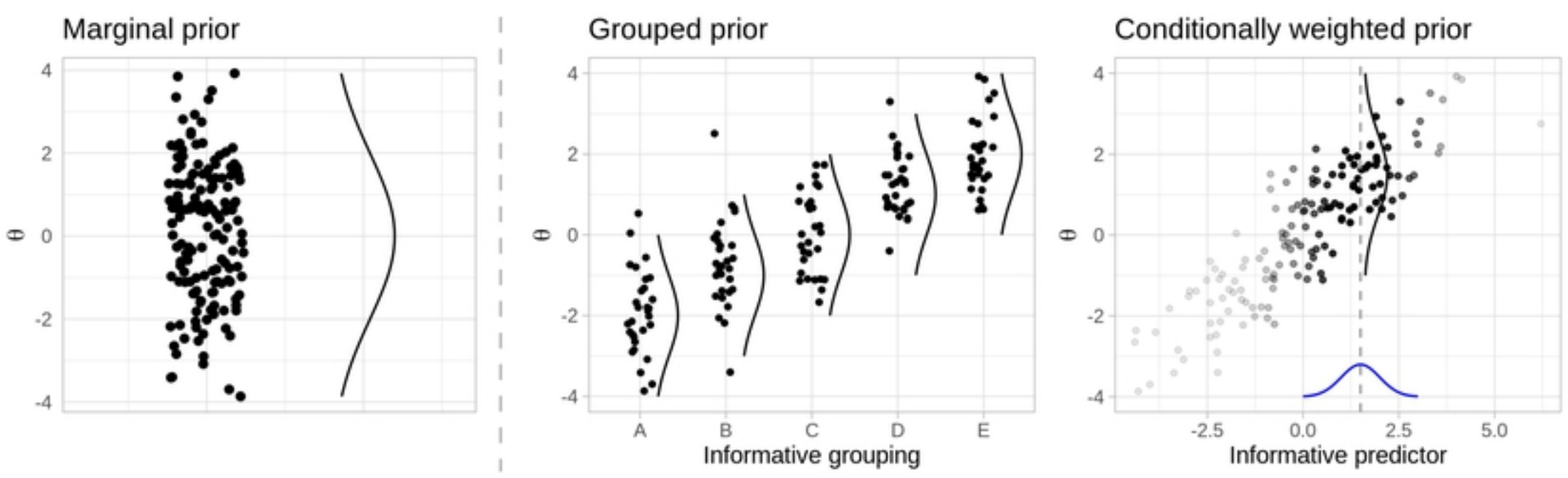
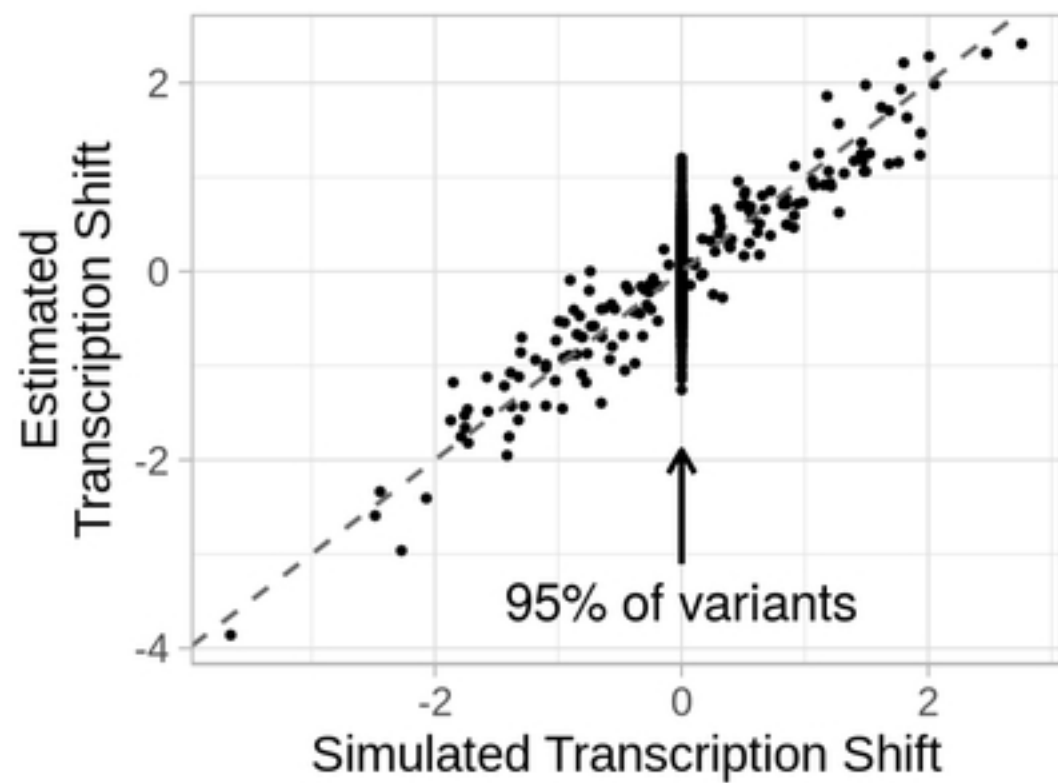
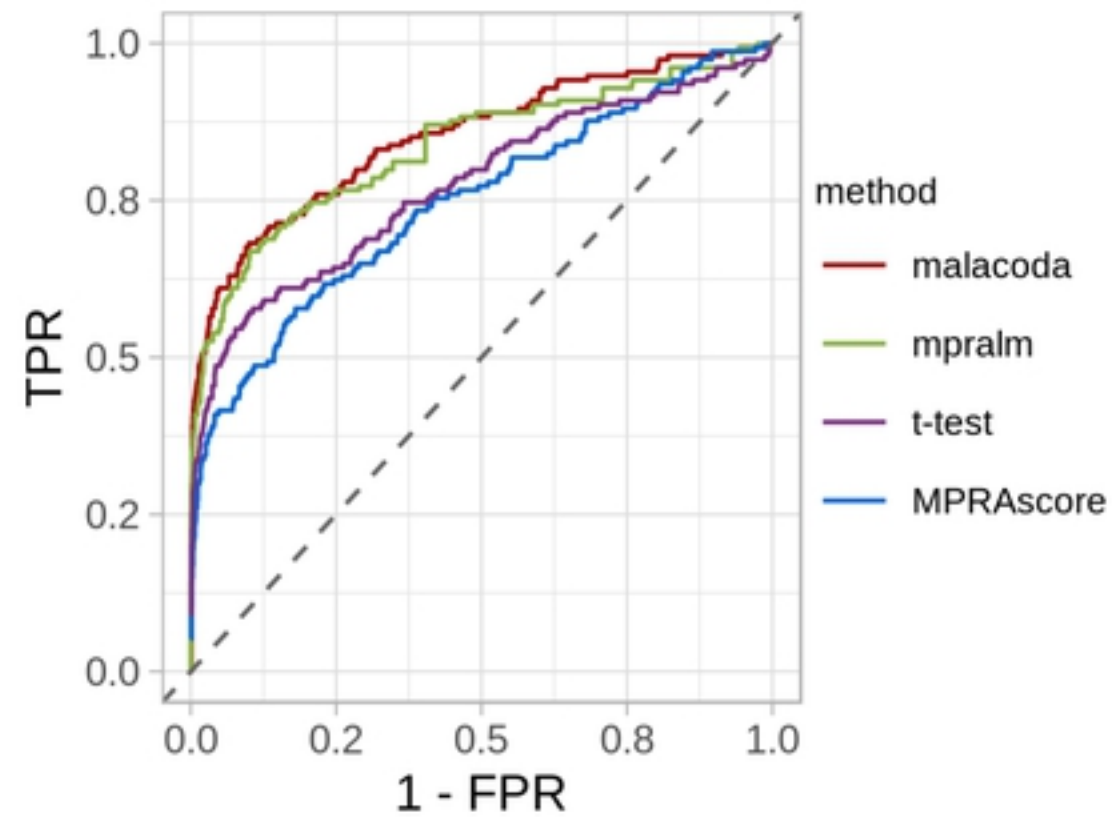
B**C**

Figure 2

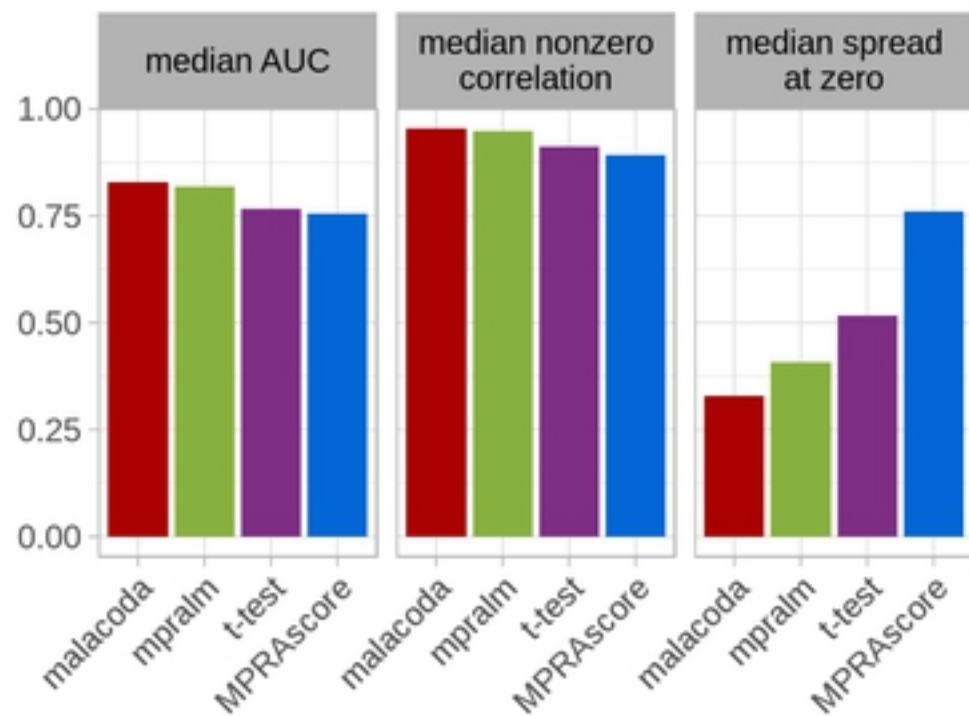
A



B



C



D

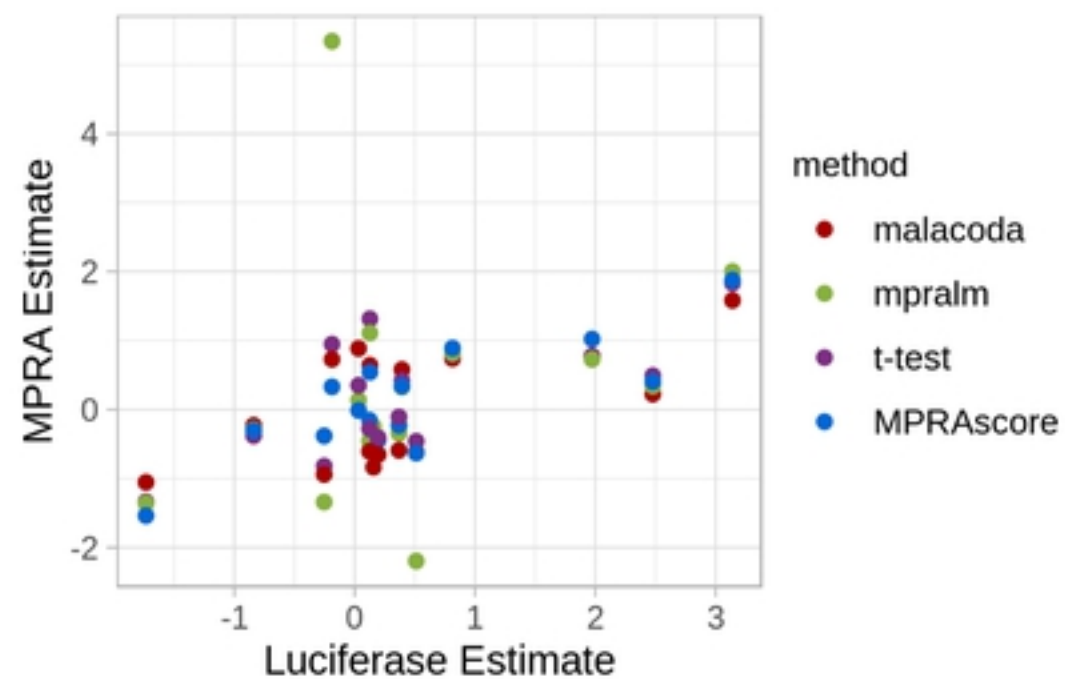


Figure 3

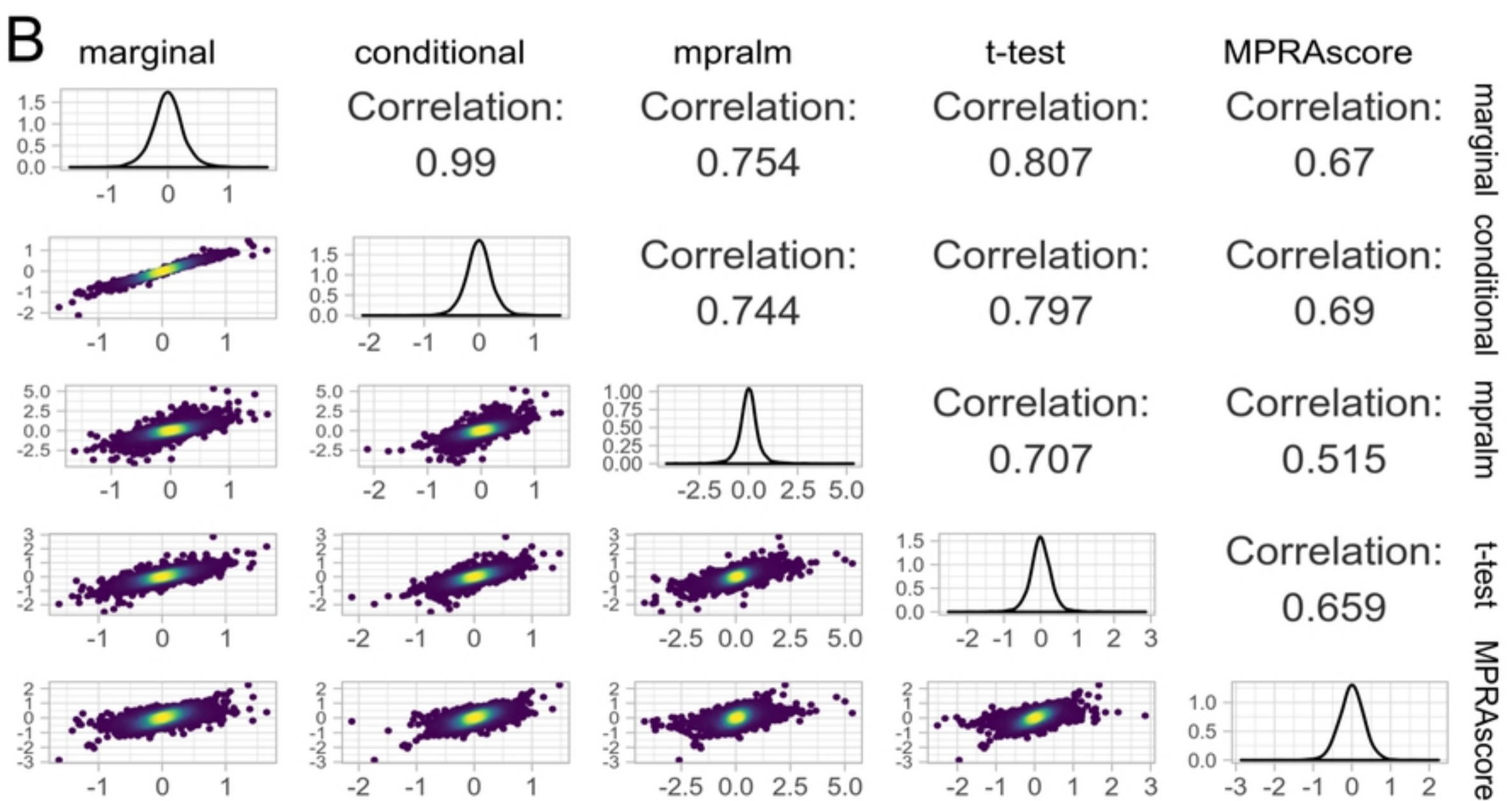
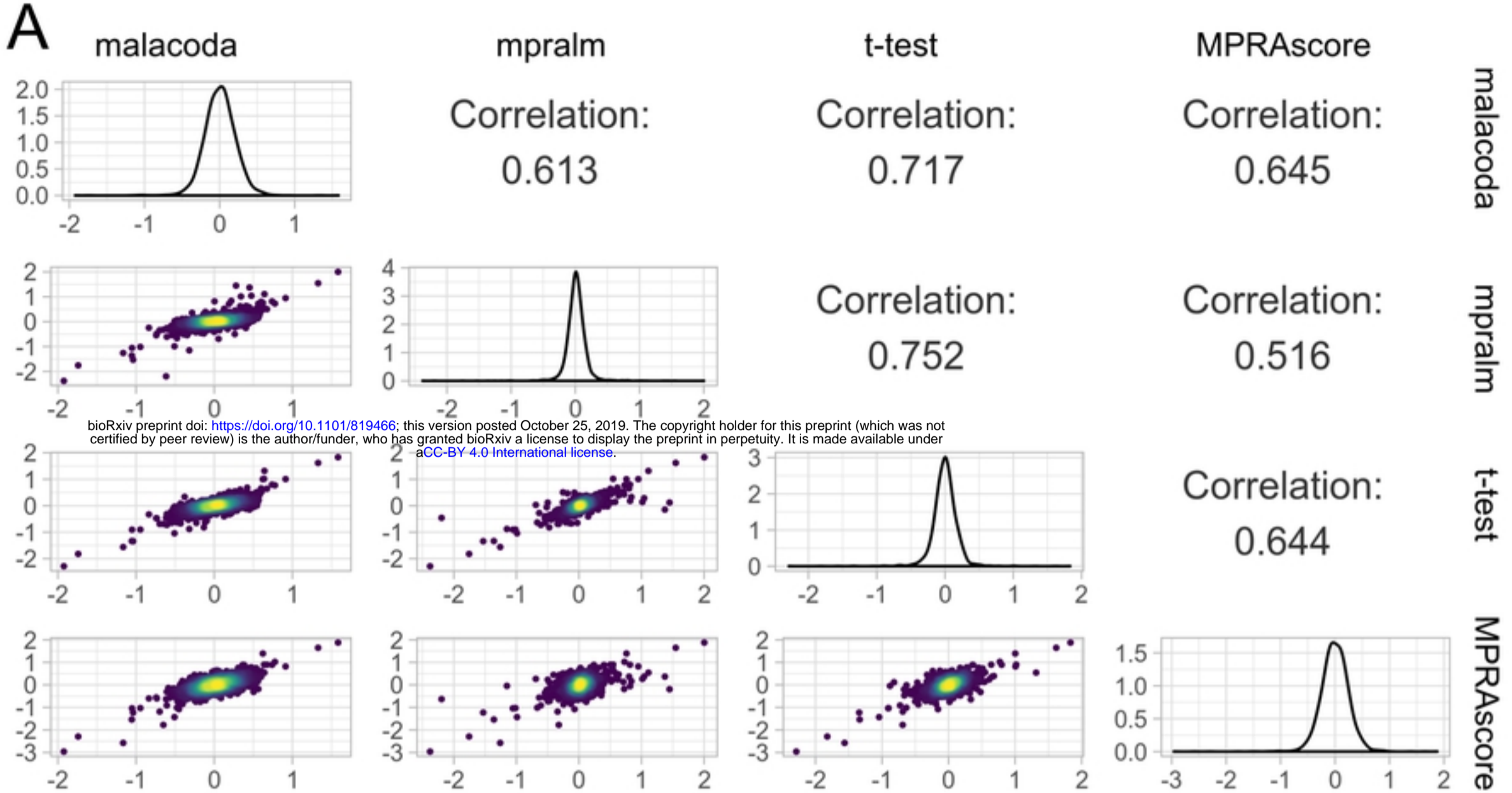


Figure 4