CLEAR: Concise List Enrichment Analysis Reducing Redundancy

Xinglin Jia[1,2], An Phan[1,2], Karin Dorman[3], and Claus Kadelka[*2]

[1]Program in Bioinformatics and Computational Biology, Iowa State University, Iowa 50011, United States

[2]Department of Mathematics, Iowa State University, Iowa 50011, United States

[3]Department of Statistics, Iowa State University, Iowa 50011, United States

Xinglin Jia and An Phan contributed equally to this work.

Abstract

Motivation: High-throughput experiments generate genome-wide measurements for thousands of genes, which are often tested marginally. Biological processes are driven by coordinated groups of genes rather than individual genes, making gene set enrichment analysis an essential *post hoc* interpretation tool. Traditional approaches such as Over-Representation Analysis and Gene Set Enrichment Analysis test gene sets independently, which ignores the hierarchical and overlapping structure of gene set collections such as the Gene Ontology, and often leads to redundant enrichment results. Set-based approaches such as MGSA address this issue by modeling multiple gene sets simultaneously, but they rely on binary gene activation states derived from arbitrary thresholds on gene-level statistics.

Results: We introduce Concise List Enrichment Analysis Reducing Redundancy (CLEAR), a Bayesian gene set enrichment framework that jointly models gene sets while incorporating continuous gene-level statistics such as test statistics or p-values. CLEAR extends model-based gene set analysis by replacing threshold-based gene activation with a probabilistic model for continuous gene-level statistics. This approach preserves the redundancy-reduction advantages of set-based enrichment methods while avoiding the information loss introduced by binarization. Using both simulated datasets and human gene expression data, we show that CLEAR improves sensitivity compared with existing enrichment approaches while producing a more concise and interpretable set of enriched gene sets.

Availability and implementation: The source code, data, and a brief tutorial are freely available at https://github.com/jiatuya/CLEAR

Keywords: Bayesian statistics; functional enrichment analysis; RNA-seq; gene set enrichment analysis

[*]Corresponding author: ckadelka@iastate.edu

1 Introduction

Many modern high-throughput experiments generate genome-, transcriptome-, or proteome-wide measurements for thousands of genes. Hypotheses are often assessed marginally, but biological functions are typically coordinated by groups of genes rather than individual genes. Consequently, gene set analysis, also known as functional enrichment analysis, has become a fundamental tool for interpreting system-wide molecular data.

Early enrichment methods such as Over-Representation Analysis (ORA) test whether individual gene sets contain more genes of interest (e.g., differentially expressed genes) than expected by chance using statistical tests such as the hypergeometric or Fisher's exact test [Khatri et al., 2012, Maleki et al., 2020]. ORA remains widely used and is often the primary option in several popular bioinformatics tools, including DAVID [Huang et al., 2009], WebGestalt [Zhang et al., 2005], and Enrichr [Chen et al., 2013, Kuleshov et al., 2016]. Gene Set Enrichment Analysis (GSEA) introduced a ranking-based framework that avoids the need to define a threshold for genes of interest by testing whether genes in a set tend to cluster at the extreme of a genome-wide ranked list [Subramanian et al., 2005]. Because GSEA utilizes continuous gene-level statistics such as fold changes or test statistics, it has become one of the most widely adopted enrichment methods for analyzing genomics, transcriptomics, and proteomics data.

Despite their widespread use, ORA and GSEA analyze each gene set independently. This assumption is problematic for gene set collections with hierarchical and overlapping relationships, such as the Gene Ontology (GO) [Tarca et al., 2013, Ashburner et al., 2000, Consortium, 2025]. In such cases, closely related parent and child gene sets often appear simultaneously in enrichment results, producing highly redundant lists that are difficult to interpret. To address this limitation, set-based approaches such as GenGO [Lu et al., 2008] and Model-based Gene Set Analysis (MGSA) [Bauer et al., 2010] jointly model the activation of multiple gene sets using probabilistic frameworks. By accounting for gene set overlap, these methods aim to identify a concise set of active biological processes rather than long lists of redundant gene sets.

However, existing set-based methods rely on binary gene activation states derived from arbitrary threshold on gene-level statistics. This binarization discards substantial information contained in the continuous statistics produced by upstream differential expression analyses, such as effect sizes or p-values. Some extensions of MGSA incorporate the topological structure of the Gene Ontology graph [Frost and McCray, 2012, Sun et al., 2017] or additional constraints [Wang et al., 2015], yet they still disregard the relevant continuous statistical information associated with the genes.

Here we introduce Concise List Enrichment Analysis Reducing Redundancy (CLEAR), a Bayesian framework that directly models the continuous gene-level statistics rather than binary activation states. CLEAR assumes that gene-level statistics follow distinct distributions under the null and alternative hypotheses and jointly infers the activation of gene sets while accounting for overlapping annotations. By leveraging the full statistical information available for each gene, CLEAR provides a more sensitive and nuanced characterization of gene set activity while maintaining the ability to reduce redundancy among related gene sets.

We evaluate CLEAR using both simulated data and real human gene expression datasets, and compare its performance to established enrichment methods, including ORA, GSEA, and MGSA. Across a range of scenarios, CLEAR improves sensitivity while preserving the interpretability advantages of set-based models by returning concise collections of enriched gene sets. These results demonstrate that modeling continuous gene-level statistics within a joint probabilistic framework provides a powerful and flexible approach to gene set enrichment analysis.

2 Materials and Methods

2.1 Generative Model

Following the generative model described in MGSA and multi-functional analyzer (MFA) [Wang et al., 2015], we consider n genes and m predefined gene sets. The annotation structure, which determines the genes associated with each gene set, is represented by an incidence matrix $I = (I_{ij})$, where $I_{ij} = 1$ if gene i belongs to gene set j, and $I_{ij} = 0$ otherwise. Each gene set j has an unobserved activation indicator $T_j \in \{0, 1\}$, where $T_j = 1$ if gene set j is active, for example if its master regulatory gene is active, $T_j = 0$ otherwise. We assume T_j follows a Bernoulli(π) distribution, where π controls the expected proportion of active gene sets.

As in MGSA and MFA, CLEAR assumes that the hidden state of a gene i, denoted as H_i, is fully determined by the hidden state of gene sets $\boldsymbol{T} = (T_1, T_2, \ldots, T_m)$ and incidence matrix $I = (I_{ij})$. That is, a gene is active if it belongs to at least one active gene set:

$$H_i(\boldsymbol{T}) = \max_{j:I_{ij}=1} T_j. \tag{1}$$

For each gene i, we observe a statistic s_i (e.g., a test statistic or p-value) that depends on its hidden state H_i. Unlike MGSA and MFA, which model the binary state of s_i, CLEAR models these observations through continuous null $f_0(s \mid \theta_0)$ and alternative distributions $f_1(s \mid \theta_1)$. That is,

$$(s_i \mid H_i(\boldsymbol{T})) \sim \begin{cases} f_0(s \mid \theta_0), & H_i(\boldsymbol{T}) = 0, \\ f_1(s \mid \theta_1), & H_i(\boldsymbol{T}) = 1. \end{cases} \tag{2}$$

where f_0 and f_1 denote distribution families (e.g., Beta, Gamma, or Uniform) appropriate for the observed statistics (p-value or Wald test statistic), and θ_0 and θ_1 are the parameters of the distributions with priors

$$\theta_0 \sim p_1(\theta_0) \qquad \text{and} \qquad \theta_1 \sim p_2(\theta_1). \tag{3}$$

Therefore the likelihood of the observed statistics $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$ for all genes is:

$$\Pr\big(\boldsymbol{s} \mid \boldsymbol{T}, \theta_0, \theta_1\big) = \prod_{i=1}^{n} f_{H_i(\boldsymbol{T})}(s_i \mid \theta_{H_i(\boldsymbol{T})}). \tag{4}$$

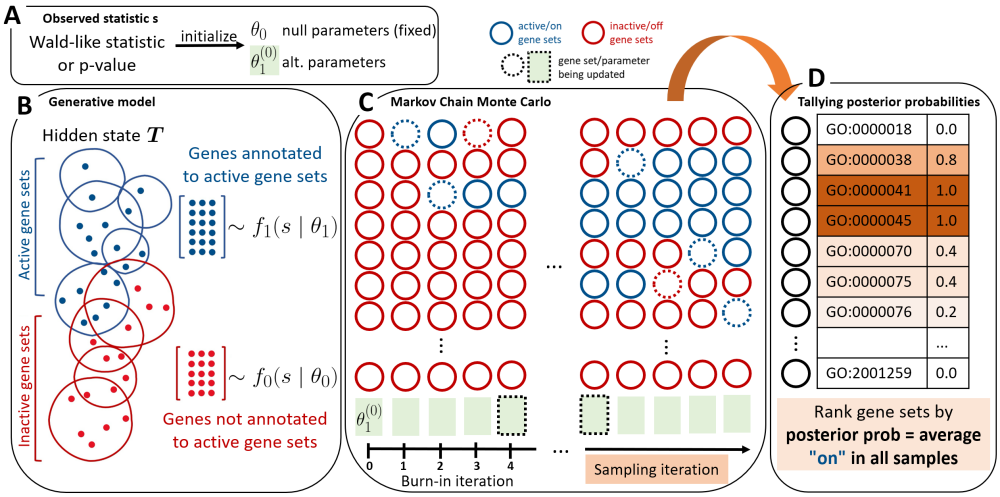


Figure 1: CLEAR method overview. CLEAR integrates a generative model for gene-level statistics with MCMC inference to identify active gene sets. (A) Observed gene-level statistics (Wald-like statistics or p-values) are used to initialize the parameters θ_0 and $\theta_1^{(0)}$. (B) Generative model of CLEAR. Each gene set j has an unobserved activation state T_j; genes belonging to active gene sets follow the alternative distribution f_1, while all other genes follow the null distribution f_0. (C) MCMC sampling procedure. Starting from an initial configuration in which all gene sets are inactive, the algorithm iteratively updates the gene set states $\boldsymbol{T}$ and the distribution parameters θ_1 (with probabilities 0.8 and 0.2, respectively) during burn-in and sampling. (D) Posterior probability of each gene set being active is computed as the proportion of sampling iterations in which the gene set is active/"on".

This generative formulation links gene-level statistics to latent gene set activity, allowing CLEAR to infer active biological processes directly from continuous gene-level measurements.

2.1.1 Generative Model for Gene-level Statistics

Marginal differential expression (DE) analysis is a common upstream step in RNA-seq and other genomics studies, producing gene-level statistics such as test statistics or p-values. Methods such as DESeq2 assess differential expression using Wald tests on estimated log fold changes [Love et al., 2014]. Under the null hypothesis of no differential expression, the resulting Wald statistics approximately follow a standard normal distribution. Under the alternative hypothesis, the up- or down-regulated genes can be modeled as normally distributed with a shifted mean ($\mu_1 \neq 0$) and increased variance ($\sigma_1^2 > 1$) [Wolf et al., 2023]. Therefore, the combined distribution of up- and down-regulated genes is a mixture distribution with at least two components.

CLEAR models these gene-level statistics using distinct distributions under the null and alternative hypotheses. To limit the number of parameters it must estimate, CLEAR models the absolute value of Wald statistics using truncated Normal distributions, under both the null and the alternative (Table 1). This approach captures the increased magnitude of differential expression signals while avoiding separate modeling of up- and down-regulation. The same modeling assumption can also be applied to other DE analysis statistics, such as the moderated t-statistic from

Table 1: Distributions of s_i under null and alternative hypotheses (Wald statistics).

Component	**Description**
s_i	Absolute observed test statistic (Wald-like)
$f_0(s)$	Truncated standard normal $\mathcal{N}_{[0,\infty)}(0,1)$
$f_1(s)$	Truncated normal $\mathcal{N}_{[0,\infty)}(\mu_1, \sigma_1^2)$

Table 2: Initialization and update rules for alternative distributions.

Type	Truncated Normal (|Wald statistics| or $-\log_{10}(p)$)	Beta (p)	Gamma ($-\log p$)
Deterministic	$\mu^{(0)} = \hat{\mu},\ \sigma^{(0)} = \hat{\sigma}$	$a^{(0)} = 0.5,\ b^{(0)} = 1$	$\alpha^{(0)} = \frac{\hat{\mu}^2}{\hat{\sigma}^2},\ \kappa^{(0)} = \frac{\hat{\sigma}^2}{\hat{\mu}}$
Random	$\theta_1^{(0)} = \theta_1\, u, \quad u \sim \text{Uniform}(0.5, 1.5)$		
Update rule	Propose one parameter at a time: $\theta'_{1j} \sim N(\theta_{1j}^{(k)}, \tau_j^2)$; re-propose if boundaries are violated		

Note. The quantities $\hat{\mu}$ and $\hat{\sigma}$ are the empirical mean and standard deviation of the transformed statistics appropriate to each model (e.g., |Wald statistics|, $-\log_{10}(p)$, or $-\log(p)$). For the Gamma alternative, (α, κ) denote the shape and scale, and their initial values are obtained using method-of-moments estimates. The scalar u is sampled from Uniform$(0.5, 1.5)$ and applied to θ_1 for random initialization.

limma [Smyth, 2005].

When test statistics are unavailable or poorly approximated by Normal distributions, CLEAR can instead model p-values or transformed p-values (e.g., $-\log_{10}(p)$). Under the null hypothesis, p-values follow a uniform distribution on $[0, 1]$, by definition. Under the alternative hypothesis, p-values concentrate near zero in a right-skewed distribution, which can be modeled by Beta distributions [Allison et al., 2002, Pounds and Morris, 2003].

In practice, transformed p-values may be easier to model. For instance, when sample sizes and effect sizes are sufficiently large, $-\log_{10}(p)$ approximately follows a normal distribution under the alternative hypothesis [Boos and Stefanski, 2011], while under the null hypothesis it follows an exponential distribution with rate $\ln 10$. Because $-\log_{10}(p)$ is non-negative, the alternative distribution is modeled using a truncated Normal distribution defined on $s \geq 0$.

In addition, CLEAR also offers an option to model $-\log(p)$-values as Gamma-distributed under the alternative hypothesis, paired with Exp(1) under the null hypothesis. Compared to the truncated Normal, the Gamma distribution offers a natural generalization of the alternative model with flexible tail behavior, which is particularly useful when modeling asymmetric, right-skewed distributions.

These alternative likelihood models allow CLEAR to flexibly accommodate different forms of gene-level statistics while maintaining the same underlying probabilistic framework.

2.1.2 Priors

Similar to MGSA, we placed a prior on the gene set activation probability π [Bauer and Gagneur, 2023]:

$$\pi \sim \text{Uniform}(0, 20/m), \tag{5}$$

Table 3: Prior distributions for alternative model parameters.

	Truncated Normal ($\lvert t\rvert$ or $-\log_{10}(p)$)	Beta (p)	Gamma ($-\log p$)
Parameterization	$\mu,\ \sigma$	a in Beta$(a, 1)$	$\alpha,\ \kappa$
Prior for Location/Shape	$\mu \sim$ Cauchy$(\hat{\mu},\ 2.5)$	$a \sim$ Uniform$(0, 1)$	$\alpha \sim$ Exp(1)
Prior for Scale/Rate	$\sigma \sim$ half-Cauchy$(0,\ 2.5)$	$b \equiv 1$ (fixed)	$\kappa \sim$ half-Cauchy$(0, 2.5)$

which reflects the prior belief that only a small fraction of gene sets are active a priori and encourages sparse gene set configurations. In addition, an empirical Bayes estimate of π can be obtained by maximizing the marginal likelihood; however, this quantity is difficult to compute. We show that the resulting estimate $\hat{\pi}$ is close to the posterior mean proportion of active gene sets (Supplement A).

For θ_1 we used weakly informative priors (Table 3) allowing wide parameter exploration while penalizing implausibly large values, following established practice in Bayesian modeling [Gelman, 2006, Gelman et al., 2008, Lemoine, 2019]. These priors accommodate the behaviors of the statistics under the alternative (e.g., p-values concentrated near zero).

For the $[0, \infty)$-truncated Normal alternative models for $|s|$ or $-\log_{10}(p)$, we assumed a Cauchy prior with empirically estimated location $\hat{\mu}$ and scale of 2.5 for the location parameter μ,

$$\mu \sim \text{Cauchy}(\hat{\mu}, 2.5), \tag{6}$$

following Gelman and colleagues' recommendation [Gelman et al., 2008, Bruch and Felderer, 2022]. And for σ, we used a half-Cauchy(0, 2.5) prior

$$\sigma \sim \text{half-Cauchy}(0, 2.5), \tag{7}$$

which is a widely recommended weakly informative prior for scale parameters to balance concentration near zero with heavy tails [Gelman, 2006, Polson and Scott, 2012, Bruch and Felderer, 2022].

We expect untransformed p-values under the alternative to be smaller than uniform and have a decreasing density on $[0, 1]$ [Allison et al., 2002, Pounds and Morris, 2003, Gadbury et al., 2004, Gadbury and Allison, 2012]. Therefore, we choose a Beta$(a, 1)$ distribution with $0 < a < 1$, which is strictly decreasing on $[0, 1]$. We applied a noninformative Uniform prior on a:

$$a \sim \text{Uniform}(0, 1). \tag{8}$$

For the Gamma alternative for $-\log p$-values, with shape α and scale κ, we use a weakly informative Exponential prior on α:

$$\alpha \sim \text{Exp}(1), \tag{9}$$

which creates a preference for shapes near the null ($\alpha = 1$) while allowing flexibility. For the scale

parameter κ, we again apply a half-Cauchy prior:

$$\kappa \sim \text{half-Cauchy}(0, 2.5). \tag{10}$$

Under the null hypothesis, there are no free parameters θ_0, and depending on user choice, θ_1 is (μ, σ) of the truncated Normal, a of the Beta, or (α, κ) of the Gamma.

2.2 Markov Chain Monte Carlo (MCMC) estimation

We sample from the posterior distribution of gene set states $\boldsymbol{T}$ given the observed statistics $\boldsymbol{s}$,

$$p\big(\boldsymbol{T}, \theta_0, \theta_1 \mid \boldsymbol{S}\big) \propto p(\boldsymbol{T})p(\theta_0)p(\theta_1)\prod_{i=1}^{n} f_{H_i(\boldsymbol{T})}(s_i \mid \theta_{H_i(\boldsymbol{T})}), \tag{11}$$

using a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm [Robert and Casella, 2010]. As in MGSA, at each MCMC iteration, we update the gene set state with probability 0.8 and update the parameters of the alternative distributions with probability 0.2. Our MCMC algorithm is detailed here and illustrated in Fig. 1.

1. **Initialize.** An initial state $\boldsymbol{T}^{(0)}$ is selected. In practice, we follow MGSA's default and set $\boldsymbol{T}^{(0)} \equiv 0$, meaning all gene sets are initially inactive. We initialize the distribution parameters θ_1 using simple summary statistics computed from the full set of observed statistics, which provides a neutral starting point that lies between the null and alternatives and avoids implausibly extreme values (Fig. 1A, Table 2).

2. **Updates of $\boldsymbol{T}$.** At each iteration k, the current gene set state is denoted as $\boldsymbol{T}^{(k)}$. If iteration k is a gene set update, we propose a new state $\boldsymbol{T}'$ with proposal probability

$$\Pr(\boldsymbol{T}' \mid \boldsymbol{T}^{(k)}) = \begin{cases} 1/|\boldsymbol{T}| & \text{if } \|\boldsymbol{T}' - \boldsymbol{T}^{(k)}\|_1 = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

   where $|\boldsymbol{T}|$ is the number of gene sets. That is, we propose to flip the state of a single gene set by either activating or deactivating it.

3. **Updates of θ_1.** If iteration k is a distribution parameter update, we use a Gaussian random-walk proposals centered at the current values:

$$\theta'_{1j} \sim N(\theta_{1j}^{(k)}, \tau_j^2), \tag{13}$$

   where $j = 1, 2$ indexes the elements of θ_1 and τ_j is a proposal standard deviation tuned adaptively during burn-in based on proposal acceptance ratio to improve mixing [Andrieu and Thoms, 2008, Maire et al., 2019, Roberts and Rosenthal, 2009]. Every 1000 iterations during the burn-in period, if the acceptance ratio exceeds 0.45, τ is multiplied by 1.1; if it falls below 0.25, it is multiplied by 0.9. The specific thresholds and multipliers are arbitrarily

chosen, but fall in the general recommended ranges [Roberts and Rosenthal, 2001]. Parameter proposals that violate the parameter constraints are rejected (e.g., $\sigma \leq 0, \alpha \leq 0$, $\kappa \leq 0$, or $a \notin (0, 1)$).

This sampling scheme jointly infers gene set activity and the distribution of gene-level statistics by alternating updates of the active gene sets and the parameters of the alternative distributions.

2.3 In silico data generation

We generated in silico data based on real Gene Ontology (GO) gene set annotation data. For each simulated dataset, we randomly selected $m_{\text{on}} \in \{5, 10, 20, 50\}$ from 883 *E. coli* GO-annotated gene sets [Gene Ontology Consortium, 2024]. The selected gene sets were required to be non-hierarchically related so that no chosen gene set was a subset of another. All genes annotated to at least one selected gene set were labeled as "active", while the remaining genes were labeled "inactive".

Gene-level test statistics were generated by simulating t-statistics for the active and inactive genes. Active genes were sampled from a non-central t distribution $t_\nu(\mu_t)$ with $\mu_t > 0$, while inactive genes were sampled from $t_\nu(0)$. The degrees of freedom ν were set to 50 and 3 to represent experiments with sufficient and limited sample sizes, respectively. The non-centrality parameter (NCP) μ_t, which controls the effect size, was varied across the range $\mu_t \in \{1, 2, 3, 5, 10\}$ based on empirical distributions observed in human RNA-seq datasets (Fig. S6).

To mimic both up- and down-regulation, the sign of each simulated t-statistic was flipped with probability 0.5. Two-sided p-values were then computed from the simulated t-statistics. This simulation scheme mimics a typical differential expression analysis pipeline used in gene expression studies, where gene-wise t-statistics are computed prior to enrichment analysis. Because the effective sample size $\nu + 1$ is typically constant across genes in differential expression tests, varying μ_t allows us to evaluate enrichment methods under different signal strengths.

2.4 Real data

2.4.1 Datasets

We collected several datasets published in popular benchmarking papers, including 15 human RNA-seq datasets from The Cancer Genome Atlas [Geistlinger et al., 2020], and 24 human microarray datasets in Gene Expression Omnibus (GEO) [Tarca et al., 2013]. The microarray datasets were restricted to only those with phenotypically relevant gene sets determined in [Geistlinger et al., 2020]. A detailed list of all datasets used can be found in Table S1.

For the TCGA datasets, raw count data for each cancer dataset were downloaded using the `GSEABenchmarkeR` R package, version 1.24.0 [Geistlinger et al., 2023]. We retained only genes with a minimum of 10 counts in at least three samples, consistent with DESeq2 best practices [Love et al., 2014, 2025]. Differential expression analysis on 15 TCGA datasets was performed using `DESeq2` without p-value adjustment [Love et al., 2014].

For the GEO microarray datasets, we obtained already preprocessed and normalized expression values from the `KEGGdzPathwaysGEO` R package, version 1.42.0 [Bhatti and Tarca, 2024]. We performed differential expression analysis using `limma` following [Tarca et al., 2012, Smyth, 2005] to get moderated t-statistics and unadjusted p-values. The `limma` package computes moderated t-statistics with a shrunken standard error after estimating the variance with empirical Bayes [Ritchie et al., 2015]. After that, probes were mapped to gene symbols using either Affymetrix Affymetrix HG-U133A or HG-U133_Plus_2 Array annotation data in R, as recorded in the metadata of each microarray dataset [Carlson, 2021a,b].

We used the GO biological processes (obtained from GO Data Archive, 2024-11-03 release) [Carbon and Mungall, 2024]. GO annotations for all human genes were retrieved from the same time point. All annotations were propagated to the root following the true path rule [Ashburner et al., 2000, Consortium, 2025]. For each dataset, GO gene sets were intersected with measured genes (i.e., genes not present in the dataset were excluded from GO gene sets for that dataset), and equivalent GO gene sets were combined into a single GO gene set (e.g., `mitotic spindle elongation` and `mitotic spindle midzone assembly` were combined into `mitotic spindle elongation & mitotic spindle midzone assembly` since they annotate the same genes). After all preprocessing steps, we retained approximately 3,000 GO gene sets of sizes between 20 and 500 genes. The upper gene set size cutoff removes overly broad and generic processes, such as `reproduction`, that do not pinpoint meaningful biological signals in the data. Lower gene set size cutoff eliminates very specific processes that are associated only with a few genes because they may represent noise rather than biological signals. Note that most processes have only very few gene annotations so the use of a lower cutoff dramatically decreases the number of considered gene sets and improves run time.

MGSA and GSEA were performed in R with Bioconductor packages (`mgsa` version 1.50 and `clusterProfiler::GSEA` version 4.10) [Bauer and Gagneur, 2023, Wu et al., 2021, 2023]. Scripts to run hypergeometric method and CLEAR are available on our GitHub repository. For MGSA and CLEAR, all datasets were run with 1,000,000 iterations (the first 500,000 are burn-in iterations), gene set states and hyperparameters are updated with probabilities 0.8 and 0.2, respectively. For all datasets, unadjusted p-values were used with the following methods: CLEAR tnormal (truncated normal), CLEAR beta, CLEAR gamma, MGSA, hypergeom (ORA). Wald statistics (from RNA-seq datasets) or moderated t-statistics (from microarray datasets) were used in GSEA and CLEAR tnormal.

All seven methods were provided with the same list of genes and GO gene sets to enable a fair comparison of the enrichment analysis output. For all variants of CLEAR, we repeated the analysis ten times for each dataset (to reduce variability from MCMC initialization), and ranked the final list of gene sets by highest sum of posterior probability over all repeats.

2.4.2 Metrics

To assess biological relevance in real datasets, we evaluated how well each method enriched known disease-relevant biological processes. We obtained a biological "ground truth" based on known

disease driver genes and oncogenic processes [Geistlinger et al., 2020]. Disease-relevant gene sets were extracted from the MalaCards disease database, which measures disease relevance of genes from experimental evidence and the literature [Rappaport et al., 2014]. In our study, cancer- and disease-relevant gene sets with 20 to 500 associated genes were considered positives, while other gene sets of the same size range were treated as negatives. The number of disease-relevant gene sets for each dataset is reported in Table S1.

Gene set–wise precision-recall curves were constructed using the positive and negative gene sets, and performance was summarized by the area under the precision–recall curve (PR-AUC). Following standard practice, the curve was anchored by projecting the precision at the first non-zero recall to recall = 0 and extending to recall = 1 with precision equal to the prevalence (i.e., the fraction of positive gene sets). Gene sets returned by each method were ranked either by highest posterior probability (for MGSA and CLEAR), or by lowest p-value (for GSEA and ORA). Additionally, since each dataset has a different number of phenotypically relevant gene sets (ranging from 6 to 282 gene sets, see Table S1), we normalized the PR-AUC by the prevalence of each dataset. Under this normalization, a random ranking of gene sets yields a normalized PR-AUC of 1.0.

Since we compare ORA and GSEA with set-based methods (MGSA and CLEAR), we also assessed the overlap of the top-ranking gene sets to emphasize the advantage of analyzing all gene sets simultaneously. Set-based methods are expected to select gene sets that share fewer genes, avoiding the long lists of overlapping gene sets commonly produced by independent testing approaches [Jantzen et al., 2011, Simillion et al., 2017]. To quantify this property, we selected the top 20 highest-ranked gene sets from each method (breaking ties randomly when necessary), and computed the pairwise overlap coefficient between gene sets based on their annotated genes [Candia and Ferrucci, 2024]. The average of all pairwise overlap coefficients defines the *gene set overlap* metric used in this study. In addition, we also calculated the average size of the top 20 gene sets returned by each method, which serves as a proxy for the specificity of the enriched biological processes.

Performance differences across enrichment methods were assessed using the Friedman test, treating datasets as blocks and methods as repeated/grouped measures. This non-parametric test accounts for grouped measurements across datasets and does not assume normality of the performance metric [Friedman, 1937]. When the global test was significant, we applied the Nemenyi test for post-hoc pairwise comparisons [Nemenyi, 1963].

3 Result

We evaluated the performance of CLEAR against MGSA, ORA, and GSEA on both in silico and real data [Bauer and Gagneur, 2023, Subramanian et al., 2005, Zhang et al., 2005, Huang et al., 2009]. The data contain gene sets ranked either by descending posterior probability (CLEAR and MGSA) or increasing p-value (ORA and GSEA). For in silico datasets, precision-recall (PR) curves and redundancy were used to capture the power and accuracy of the methods.

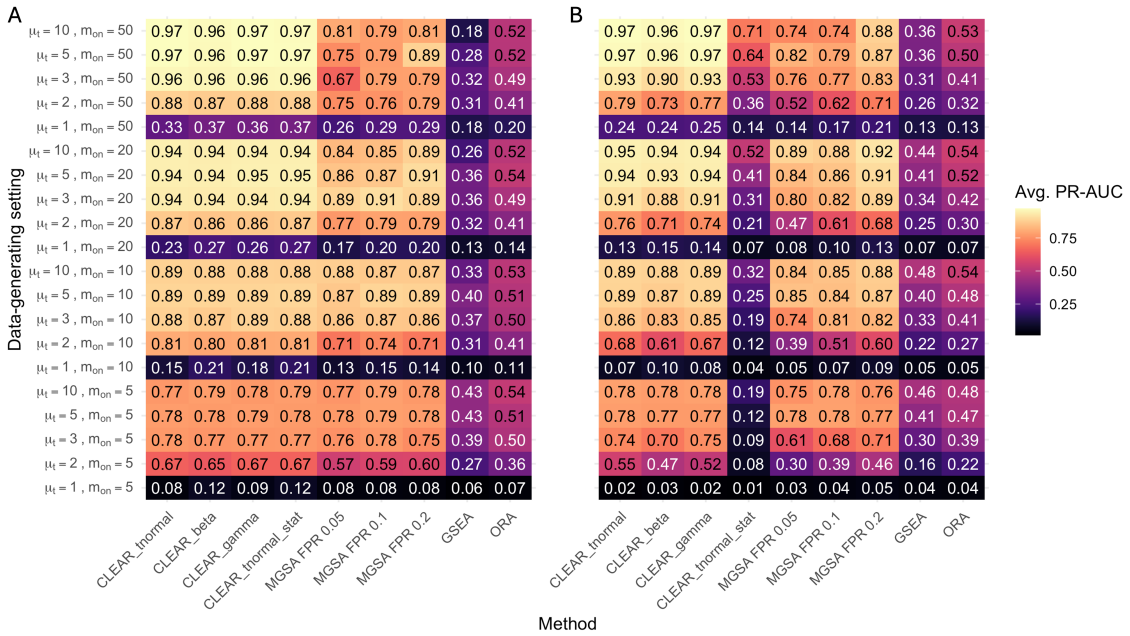


Figure 2: Average precision-recall AUC (PR-AUC) of enrichment methods in simulated data. Rows describe different data-generation configurations, characterized by the signal strength μ_t and the number of activated gene sets m_{on}. Columns describe enrichment methods. Panel A shows results for a sufficient sample size ($\nu = 50$). Panel B shows results for a limited sample size ($\nu = 3$).

3.1 In silico data analysis

PR curves measure how effectively a method recovers true positive gene sets. Precision is the proportion of true positives among all returned positives, while recall (also known as sensitivity) is the proportion of true positives identified out of all positives. For each simulated dataset, we computed the area under the PR curve (PR-AUC) to summarize overall performance.

Across simulated datasets, CLEAR consistently achieved higher PR-AUC than existing enrichment methods, particularly under moderate to strong signal conditions. The improvement was most pronounced when gene-level statistics closely followed the assumed likelihood distributions (Fig. 2).

When the sample size is sufficiently large ($\nu = 50$), all variants of CLEAR (truncated Normal (p-values), Beta, Gamma, and truncated Normal (test statistics)) performed equally well with highest PR-AUC among all methods being tested, followed by p-value-based models, MGSA, ORA, and then GSEA (Fig. 2A). Compared to MGSA and the traditional methods, CLEAR showed a clear advantage when the signal-to-noise ratio was high (e.g., $\mu_t = 10$). When the signal is very weak (e.g., $\mu_t = 1$), all methods perform poorly and the performance gain of p-value-based CLEAR models over MGSA was smaller.

When the sample size is small ($\nu = 3$), performance declines for all methods due to increased statistical noise. The test statistics model was particularly affected because the Normality assumption for the alternative distribution no longer holds well in small-sample settings. In contrast, the p-value-based CLEAR models maintained strong performance and consistently outperformed other methods in most scenarios.

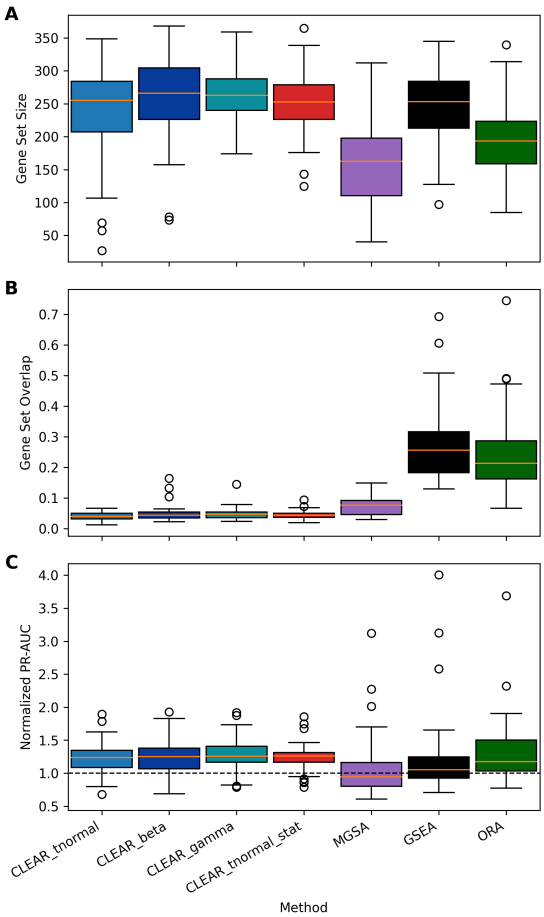


Figure 3: Method performance across 15 TCGA RNA-seq datasets and 24 GEO microarray datasets. Panels A and B show the average gene set size and gene set overlap for the top 20 gene sets enriched by each method. Panel C shows the normalized PR-AUC values where phenotypically relevant gene sets are considered positives. Orange lines display the median value for each method. Dotted black line in panel C represents the expected value of normalized PR-AUC.

Overall, CLEAR variants substantially outperformed MGSA, ORA, and GSEA when the signal strength was moderate to high. Under very weak signals (e.g., NCP $= \mu_t = 1$), all methods performed poorly, and the differences between methods were smaller.

3.2 Real data analysis

We first examined the characteristics of gene sets returned by each enrichment method. CLEAR and GSEA tended to return larger gene sets than MGSA and ORA (Fig. 3A). For CLEAR, this behavior likely reflects the hierarchical structure of the GO graph. When both a parent gene set and its child gene sets carry detectable signal, CLEAR often selects the parent gene set while excluding the children.

Next, we evaluated redundancy among the top enriched gene sets (Fig. 3B). CLEAR produced the lowest gene set overlap, followed by the other set-based method MGSA. In contrast, the tra-

ditional enrichment methods GSEA and ORA returned substantially more overlapping gene sets. Notably, CLEAR achieved the lowest overlap despite tending to return relatively large gene sets (Fig. 3A), indicating that CLEAR effectively selects representative gene sets rather than multiple closely related gene sets.

Finally, we assessed biological relevance using normalized PR-AUC scores, which measure how effectively each method recovers known disease-relevant biological processes (Fig. 3C). Normalization allows comparison across datasets with different proportions of positive gene sets. Raw PR-AUC values are reported in the Supplement (Fig. S5). For most datasets, all methods performed only slightly better than a random prediction, hinting at very low signal-to-noise ratios. The Friedman global test [Friedman, 1937], using datasets as the grouping variable, rejected the null hypothesis that all methods performed similarly ($p = 1.6 \times 10^{-10}$). CLEAR achieved performance comparable to ORA ($p \approx 1.0$, Nemenyi post-hoc test), implying that CLEAR performed no worse than ORA [Woolson, 2007, Benavoli et al., 2016]. All versions of CLEAR significantly outperformed MGSA and GSEA ($p < 0.03$, Nemenyi post-hoc test), except for CLEAR tnormal versus GSEA ($p = 0.26$). All post-hoc test results are reported in Fig. S7.

It is important to note that the disease-relevant biological processes used as positives in the normalized PR-AUC metric are often hierarchically related. For example, in the BRCA dataset the gene set `cell surface receptor protein tyrosine kinase signaling pathway (GO:0007169)` and three of its child gene sets `GO:0008543, GO:0038084, GO:0048009` are all considered positives. CLEAR selected only the parent gene set, whereas GSEA and ORA selected all four gene sets, thereby receiving four true positives in the precision-recall metric. This evaluation framework therefore favors methods that return redundant gene sets. Despite this disadvantage, CLEAR achieved comparable PR-AUC to ORA and better PR-AUC than MGSA, and GSEA, and a hypothetical random prediction. Overall, CLEAR consistently returned less redundant gene sets while achieving comparable or better biological relevance than traditional enrichment methods.

We evaluated elapsed runtime and memory usage for all methods on two human RNA-seq datasets, BLCA (3,304 GO gene sets) and BRCA (3,334 GO gene sets), using the GNU time utility running on a single CPU. Among the evaluated methods, CLEAR had the longest runtime. For one run of 1,000,000 iterations, CLEAR beta and CLEAR gamma required approximately 20 minutes, whereas CLEAR tnormal (using either p-value or Wald statistic input) required approximately 10 minutes. Under the same MCMC iteration setting, MGSA completed in a few seconds with an efficient implementation in C. GSEA with the fgsea preranked implementation required approximately 2 minutes, and ORA completed in a few seconds per dataset [Korotkevich et al., 2021]. Although CLEAR is computationally slower than methods such as ORA, GSEA, and MGSA, this is not unexpected for a Bayesian-based approach that models all gene sets simultaneously and updates a high-dimensional joint state space through MCMC.

4 Discussion

We developed a novel enrichment analysis method, CLEAR, that extends set-based gene set analysis to directly model continuous gene-level statistics. By replacing binary gene activation with a probabilistic model for gene-level test statistics or p-values, CLEAR retains the redundancy-reduction advantages of set-based enrichment approaches while avoiding the information loss introduced by thresholding.

Across simulated datasets, CLEAR substantially outperformed existing enrichment methods under moderate to strong signal conditions. In real biological datasets (with seemingly much lower signal-to-noise ratios), CLEAR achieved equal and better biological relevance than traditional enrichment methods (GSEA, ORA, MGSA) while returning substantially fewer redundant sets of enriched GO gene sets. Together, these results demonstrate that modeling gene-level statistics directly can improve sensitivity while maintaining interpretable enrichment results.

In the simulation study, gene-level test statistics were generated from a t-distribution with different degrees of freedom ν to represent experiments with sufficient or limited sample sizes. This design mimics realistic differential expression analysis pipelines while avoiding biases that could arise from sampling statistics directly from assumed model distributions (e.g., mixtures of Beta and Uniform distributions for p-values). As a result, CLEAR does not gain an artificial advantage in the benchmarking experiments.

When the sample size is small ($\nu = 3$), the simulated test statistics deviate from the truncated Normal assumption used in the statistic-based CLEAR model, leading to reduced performance (Fig. 2). In contrast, the p-value-based CLEAR models were more robust because the Uniform$(0, 1)$ distribution assumed under the null model remained true regardless of the underlying test statistic distribution. Consequently, the p-value-based CLEAR models maintained strong performance even under small-sample conditions.

CLEAR addresses MGSA's binarization threshold issue

CLEAR extends MGSA's binary generative framework (that relies heavily on the correct choice of arbitrary thresholds) by replacing threshold-based gene activation with a probabilistic model for threshold-free, continuous gene-level statistics. In MGSA, genes must be classified as active or inactive based on an arbitrary threshold, which can lead to information loss and sensitivity to threshold choice. CLEAR instead models gene-level statistics or p-values directly using parametric distributions under the null and alternative hypotheses. This approach avoids binarization and allows the model to capture more subtle gene-level signals.

CLEAR and MGSA belong to a broader class of set-based Bayesian enrichment approaches that jointly model gene sets rather than analyzing them independently, as done by traditional enrichment methods such as ORA and GSEA. By modeling gene sets jointly, set-based approaches address the redundancy and dependency inherent in gene set collections such as the Gene Ontology, reducing the tendency to return long lists of highly overlapping enriched gene sets.

Benchmarking on real datasets remains challenging

CLEAR relies on differences between the gene-level statistics of genes under the alternative distribution (genes associated with the condition of interest) and those under the null distribution. In an ideal setting, genes belonging to phenotypically relevant gene sets would exhibit gene-level statistics clearly distinct from background genes. However, in the real datasets we observed substantial overlap between the p-value and Wald statistic distributions of genes belonging to phenotypically relevant gene sets and those outside them (Fig. S6). This pattern was observed across nearly all datasets, indicating that gene-level statistics alone often provide limited separation between benchmark-positive and benchmark-negative genes.

In addition, Fig. S6 also shows deviations from the model assumptions underlying CLEAR. In particular, the empirical distributions of benchmark-negative genes often differ from the assumed null distributions (Uniform for p-values in CLEAR beta, $\exp(1)$ for $-\log p$ in CLEAR gamma, and truncated Normal for |Wald statistics| in CLEAR tnormal). A likely explanation is that benchmark-negative genes in real data rarely represent a pure null set; instead, they may include genes with weak or indirect associations with the studied condition. In fact, many benchmark-negative genes exhibit unexpectedly small p-values, suggesting that curated phenotypically relevant gene sets are at best an imperfect proxy for true biological ground truth.

More broadly, benchmarking enrichment methods on real data is inherently challenging because definitive ground truth gene sets are rarely available. Instead, benchmarks rely on prior biological knowledge, which can be incomplete or inconsistent [Hung et al., 2012]. Even with carefully curated benchmarking datasets, it remains difficult to assess method performance in real biological systems such as cancer, where consensus pathway definitions are still evolving [Chen et al., 2021, Menyhart et al., 2024]. Similar challenges arise in simpler model organism experiments, where multiple biological processes may plausibly be considered relevant to the experimental condition [Siso et al., 2012, Bendjilali et al., 2017]. Our benchmarking therefore follows established practice in the literature while acknowledging that evaluations based on curated gene sets remain approximate in the absence of definitive biological ground truth.

Limitations

Despite the advantages, CLEAR has a few limitations. First, MCMC convergence can be elusive given the high dimension of the parameter space. CLEAR jointly samples discrete and continuous parameters from a posterior distribution with potentially many local optima. MGSA also has the same issue, and its documentation recommends multiple restarts because convergence is difficult to access. We provide a detailed convergence analysis in Appendix B. Given this limitation, we recommend running multiple independent chains for each dataset to help confirm that the model has properly converged.

Second, runtime is a practical limitation of CLEAR. While ORA and GSEA evaluate each gene set individually, MGSA and CLEAR use a computationally intensive Bayesian framework that jointly models all gene sets. However, the current implementation of CLEAR takes longer

than MGSA because it uses a more complex statistical framework. In MGSA, the parameters false positive rate and false negative rate lie in a bounded interval $(0, 1)$, and MGSA defines a discrete grid over these parameters and pre-calculates the likelihood contributions for true positives, true negatives, false positives, and false negatives to be reused during MCMC and reduce computational cost. In contrast, CLEAR does not discretize parameters on a predefined grid because its continuous parameters (θ_1 for f_1) do not have similarly bounded ranges. As a result, the probability density of genes must be recomputed whenever the model parameters are updated during MCMC (20% of iterations), which becomes the main computational bottleneck. We also note that the current implementation of CLEAR is written in R, whereas MGSA is implemented in C, so part of the observed runtime difference likely arises from implementation efficiency rather than from the statistical framework alone. Future work should focus on improving computational efficiency through code profiling, lower-level implementation, and algorithmic optimization.

Third, CLEAR's performance may be affected when empirical data deviate substantially from the assumed statistical distribution. CLEAR relies on predefined distributions to model the null and alternative hypotheses (e.g., Uniform and Beta). If the true distribution of the data differs from these assumptions, model fit accuracy may be reduced. In practice, different datasets may exhibit different mixture patterns in their gene-level statistics. Therefore, we recommend visually inspecting the distribution of gene-level statistics (e.g., p-values or test statistics) before fitting the model to achieve the best performance.

5 Conclusion

Existing functional enrichment methods often rely on independent testing of gene sets or binarized gene activation states, which can discard valuable information contained in continuous gene-level statistics. CLEAR addresses these limitations by modeling continuous gene-level statistics within a probabilistic framework that distinguishes between null and alternative gene distributions. Across simulated datasets, CLEAR demonstrated improved sensitivity compared with existing enrichment approaches. In real biological datasets, CLEAR achieved biological relevance comparable to traditional enrichment methods while producing substantially less redundant enrichment results. By combining the redundancy control of set-based enrichment approaches with a threshold-free probabilistic model for gene-level signals, CLEAR provides a practical, robust and interpretable framework for gene set enrichment analysis.

Competing interests

No competing interest is declared.

Author contributions statement

Conceptualization: X.J., A.P., C.K. Methodology: X.J., A.P., K.D., C.K. Software: X.J., A.P. Formal analysis: X.J., A.P. Supervision: K.D., C.K. Writing – original draft: X.J., A.P. Writing – review & editing: X.J., A.P., K.D., C.K.

Data availability

The source code and data are freely available at https://github.com/jiatuya/CLEAR. Publicly available datasets were used in this manuscript, as summarized in Section 2.4.1. Preprocessed datasets (TCGA RNA-seq and GEO microarray) are available in the Figshare folder at https://doi.org/10.6084/m9.figshare.31891432.

References

D. B. Allison, G. L. Gadbury, M. Heo, J. R. Fernández, C.-K. Lee, T. A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, 39(1):1–20, Mar. 2002.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, Dec. 2008.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

S. Bauer and J. Gagneur. mgsa: Model-based gene set analysis, 2023. URL https://bioconductor.org/packages/mgsa. R package version 1.50.0.

S. Bauer, J. Gagneur, and P. N. Robinson. GOing bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.*, 38(11):3523–3532, June 2010.

A. Benavoli, G. Corani, and F. Mangili. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, 2016.

N. Bendjilali, S. MacLeon, G. Kalra, S. D. Willis, A. N. Hossian, E. Avery, O. Wojtowicz, and M. J. Hickman. Time-course analysis of gene expression during the Saccharomyces cerevisiae hypoxic response. *G3: Genes, Genomes, Genetics*, 7(1):221–231, 2017.

G. Bhatti and A. L. Tarca. KEGGdzPathwaysGEO: KEGG disease datasets from GEO, 2024. URL https://bioconductor.org/packages//release/data/experiment/html/KEGGdzPathwaysGEO.html. R package version 1.40.0.

D. D. Boos and L. A. Stefanski. P-value precision and reproducibility. *Am. Stat.*, 65(4):213–221, 2011.

C. Bruch and B. Felderer. Prior choice for the variance parameter in the multilevel regression and poststratification approach for highly selective data. a monte carlo simulation study. *Austrian J. Stat.*, 51(4):76–95, Aug. 2022.

J. Candia and L. Ferrucci. Assessment of Gene Set Enrichment Analysis using curated RNA-seq-based benchmarks. *PLOS ONE*, 19(5):e0302696, May 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0302696. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0302696.

S. Carbon and C. Mungall. Gene Ontology data archive, Nov. 2024. URL https://doi.org/10.5281/zenodo.14083199.

M. Carlson. hgu133a.db: Affymetrix Affymetrix HG-U133A array annotation data (chip hgu133a), 2021a. R package version 3.13.0.

M. Carlson. hgu133plus2.db: Affymetrix affymetrix hg-u133_plus_2 array annotation data (chip hgu133plus2), 2021b. R package version 3.13.0.

E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

Y. Chen, F. J. Verbeek, and K. Wolstencroft. Establishing a consensus for the hallmarks of cancer based on Gene Ontology and pathway annotations. *BMC bioinformatics*, 22(1):178, 2021.

T. G. Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, page gkaf1292, 12 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf1292. URL https://doi.org/10.1093/nar/gkaf1292.

B. E. Deonovic and B. J. Smith. Convergence diagnostics for MCMC draws of a categorical variable. *arXiv [stat.ME]*, June 2017.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522.

H. R. Frost and A. T. McCray. Markov chain ontology analysis (MCOA). *BMC Bioinformatics*, 13(1):23, Feb. 2012.

G. L. Gadbury and D. B. Allison. Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. *PLoS One*, 7(10):e46363, Oct. 2012.

G. L. Gadbury, G. P. Page, J. Edwards, T. Kayo, T. A. Prolla, R. Weindruch, P. A. Permana, J. D. Mountz, and D. B. Allison. Power and sample size estimation in high dimensional biology. *Stat. Methods Med. Res.*, 13(4):325–338, Aug. 2004.

L. Geistlinger, G. Csaba, M. Santarelli, M. Ramos, L. Schiffer, N. Turaga, C. Law, S. Davis, V. Carey, M. Morgan, R. Zimmer, and L. Waldron. Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*, 22(1):545–556, Mar. 2020. ISSN 1467-5463. doi: 10.1093/bib/bbz158. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7820859/.

L. Geistlinger, G. Csaba, M. Santarelli, M. Ramos, L. Schiffer, C. Law, N. Turaga, S. Davis, V. Carey, M. Morgan, R. Zimmer, and L. Waldron, 2023. URL https://bioconductor.org/packages/release/bioc/html/GSEABenchmarkeR.html. R package version 1.22.0.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, Sept. 2006.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A WEAKLY INFORMATIVE DEFAULT PRIOR DISTRIBUTION FOR LOGISTIC AND OTHER REGRESSION MODELS. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

Gene Ontology Consortium. Gene ontology annotations (gaf 2.2): release 2024-10-11. http://geneontology.org, 2024.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

J.-H. Hung, T.-H. Yang, Z. Hu, Z. Weng, and C. DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.

S. G. Jantzen, B. J. Sutherland, D. R. Minkley, and B. F. Koop. Go trimming: Systematically reducing redundancy in large Gene Ontology datasets. *BMC research notes*, 4(1):267, 2011.

P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):e1002375, 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002375. Publisher: Public Library of Science.

G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, and A. Sergushichev. Fast gene set enrichment analysis, Feb. 2021. URL https://www.biorxiv.org/content/10.1101/060012v3. Pages: 060012 Section: New Results.

M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

N. P. Lemoine. Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *Oikos*, 128(7):912–928, July 2019.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

M. I. Love, S. Anders, and W. Huber. Analyzing RNA-seq data with DESeq2, 2025. URL https://bioconductor.org/packages//release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html.

Y. Lu, R. Rosenfeld, I. Simon, G. J. Nau, and Z. Bar-Joseph. A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.*, 36(17):e109, Oct. 2008.

F. Maire, N. Friel, A. Mira, and A. E. Raftery. Adaptive incremental mixture Markov chain Monte Carlo. *J. Comput. Graph. Stat.*, 28(4):790–805, June 2019.

F. Maleki, K. Ovens, D. J. Hogan, and A. J. Kusalik. Gene set analysis: Challenges, opportunities, and future research. *Front. Genet.*, 11:654, 2020.

O. Menyhart, W. J. Kothalawala, and B. Győrffy. A gene set enrichment analysis for the cancer hallmarks. *Journal of Pharmaceutical Analysis*, page 101065, 2024.

P. B. Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963.

N. G. Polson and J. G. Scott. On the half-cauchy prior for a global scale parameter. *Bayesian Anal.*, 7(4):887–902, Dec. 2012.

S. Pounds and S. W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, July 2003.

N. Rappaport, M. Twik, N. Nativ, G. Stelzer, I. Bahir, T. I. Stein, M. Safran, and D. Lancet. Malacards: A comprehensive automatically-mined database of human diseases. *Current Protocols in Bioinformatics*, 47(1):1–24, 2014.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, Apr. 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/.

C. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer Texts in Statistics. Springer, New York, NY, Nov. 2010.

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.*, 16(4):351–367, Nov. 2001.

G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *J. Comput. Graph. Stat.*, 18(2): 349–367, Jan. 2009.

C. Simillion, R. Liechti, H. E. Lischer, V. Ioannidis, and R. Bruggmann. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC bioinformatics*, 18(1):151, 2017.

M. I. G. Siso, M. Becerra, M. L. Maceiras, Á. V. Vázquez, and M. E. Cerdán. The yeast hypoxic responses, resources for new biotechnological opportunities. *Biotechnology letters*, 34(12):2161–2173, 2012.

G. K. Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, Oct. 2005.

D. Sun, Y. Liu, X.-S. Zhang, and L.-Y. Wu. NetGen: a novel network-based probabilistic generative model for gene set functional enrichment analysis. *BMC Syst. Biol.*, 11(Suppl 4):75, Sept. 2017.

A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, June 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-136. URL https://doi.org/10.1186/1471-2105-13-136.

A. L. Tarca, G. Bhatti, and R. Romero. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PLoS ONE*, 8(11):e79217, Nov. 2013. doi: 10.1371/journal.pone.0079217. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3829842/.

Z. Wang, Q. He, B. Larget, and M. A. Newton. A multi-functional analyzer uses parameter constraints to improve the efficiency of model-based gene-set analysis. *aoas*, 9(1):225–246, Mar. 2015.

S. Wolf, D. Melo, K. M. Garske, L. F. Pallares, A. J. Lea, and J. F. Ayroles. Characterizing the landscape of gene expression variance in humans. *PLoS genetics*, 19(7):e1010833, 2023.

R. F. Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.

T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, xiaochong Fu, S. Liu, X. Bo, and G. Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021. doi: 10.1016/j.xinn.2021.100141.

T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, xiaochong Fu, S. Liu, X. Bo, and G. Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data, 2023. URL https://bioconductor.org/packages/clusterProfiler. R package version 4.10.0.

B. Zhang, S. Kirov, and J. Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research*, 33(suppl_2):W741–W748, 2005.

Supplementary methods and figures - Jia et al.

A Empirical Bayes Interpretation of Gene set Activation Probability π

MGSA assumes that gene set activation indicators $T_j \in \{0, 1\}$ follow

$$T_j \mid \pi \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi), \quad j = 1, \ldots, m, \tag{14}$$

with a hyperprior distribution $\pi \sim \text{Uniform}(0, 20/m)$, where m is the number of gene sets. Let s denote the observed statistics and let θ_0, θ_1 denote the null and alternative parameters, respectively. In empirical Bayes, we can estimate π using marginal likelihood,

$$m_s(\pi) = \sum_T \int p(s \mid T, \theta_0, \theta_1)\, p(T \mid \pi)\, p(\theta_0)\, p(\theta_1)\, d\theta_0\, d\theta_1. \tag{15}$$

However, the marginal likelihood is generally difficult to compute. Instead, we consider the posterior distribution under a fixed value π,

$$p(T, \theta_0, \theta_1 \mid s; \pi) = \frac{p(s \mid T, \theta_0, \theta_1)\, p(T; \pi)\, p(\theta_0)\, p(\theta_1)}{m_s(\pi)}. \tag{16}$$

Suppose an MCMC sample $\{T^{(i)}, \theta_0^{(i)}, \theta_1^{(i)}\}_{i=1}^{\ell}$ is obtained under a particular π_1. Then the marginal likelihood ratio can be approximated by

$$B_\ell(\pi) = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{p(T^{(i)}; \pi)}{p(T^{(i)}; \pi_1)} \xrightarrow{\text{a.s.}} \frac{m_s(\pi)}{m_s(\pi_1)}. \tag{17}$$

Therefore,

$$\arg\max_\pi m_s(\pi) = \arg\max_\pi B_\ell(\pi). \tag{18}$$

In the current setting, the maximum is

$$\hat{\pi} = \frac{1}{\ell m} \sum_{i=1}^{\ell} \sum_{j=1}^{m} T_j^{(i)} \approx \frac{1}{m} \sum_{j=1}^{m} \Pr(T_j = 1 \mid s), \tag{19}$$

which shows that the empirical Bayes estimate of π approximates the posterior mean proportion of active gene sets.

B Convergence Analysis

Convergence is an important indication that the posterior has been correctly estimated for models that rely on posterior sampling. For MGSA and CLEAR, the parameters include gene set activation,

treated as a binary parameter, along with continuous parameters, such as false positive rate and false negative rate for MGSA, and the parameters θ_1 of the alternative distributions for CLEAR. We employed separate strategies to evaluate the convergence of the continuous vs. discrete parameters. In addition, we used trace plots to monitor posterior changes over iterations.

B.1 Discrete Parameter Convergence Assessment

We assessed the convergence of binary gene set indicators (0 for "off" and 1 for "on") following Deonovic and Smith's framework for categorical MCMC who offer two χ^2 statistics for testing convergence of discrete parameters [Deonovic and Smith, 2017].

B.1.1 Homogeneity of the Marginal Probability using Weiss Test

We applied a Weiss test for convergence of posterior probabilities across all MCMC chains. The standard Pearson χ^2 statistic was adjusted by a variance inflation factor, $\hat{c}$, to account for the dependence of MCMC sampling following the recommendations of Deonovic and Smith. Factor $\hat{c}$ is estimated by modeling the binary sequence as a Discrete Autoregressive process of order 1 (DAR(1)).

B.1.2 Homogeneity of the Transition Probability using Billingsley Test

In addition to the Weiss test, we applied the Billingsley test to transition probabilities, i.e., $0 \to 1$, $1 \to 1$, etc., to assess convergence of chain dynamics. We constructed a contingency table of transition counts for each gene set and estimated chain-specific transition probabilities. Then we computed a χ^2 statistic to compare these transition probabilities with the pooled transition probabilities across all chains.

The Weiss test and Bilingsley tests were performed for each gene set independently. We specifically focused the convergence analysis on "high-confidence" gene sets (gene sets with a posterior probability > 0.5) to ensure that the gene sets identified with highest support were robustly identified across independent runs.

B.2 Continuous Parameter Convergence Assessment

For continuous parameters (e.g., mean and standard deviation from a truncated Normal), we assessed convergence using the standard $\hat{R}$ [Gelman and Rubin, 1992]. An $\hat{R}$ value below 1.1 was considered indicative of convergence.

B.3 Convergence Result and Discussion

We performed convergence analysis on simulated datasets with $\nu = 50$. We sampled 5 out of 100 replicates across all combinations of NCP and m_{on}, totaling 100 datasets. For each dataset, we ran 5 independent MCMC chains and recorded the posterior probabilities of gene sets, transition probabilities, continuous parameter values, and the posterior value. The chi-square tests showed

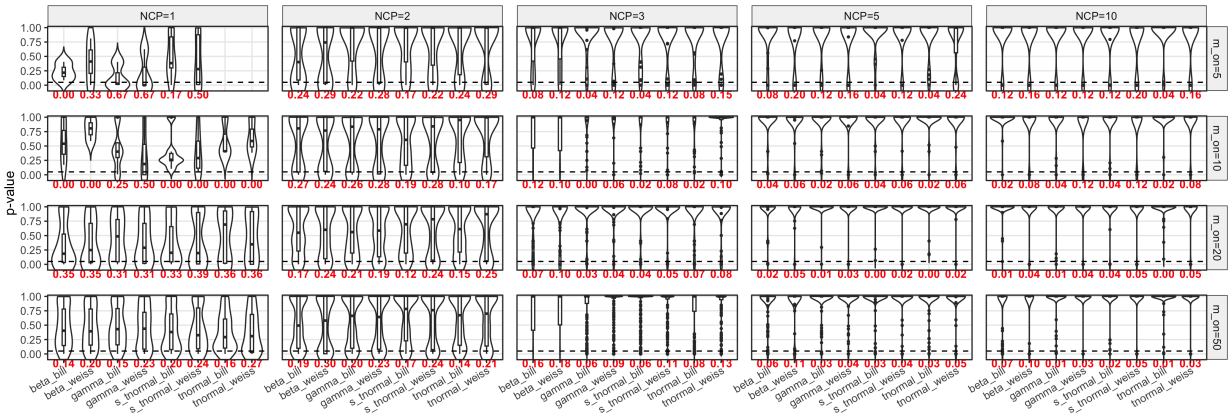


Figure S1: Weiss test and Billingsley test p-values for gene sets with average posterior probability $>$ 0.5. Results are computed for simulated data sets under different data-generation configurations by the signal strength NCP (μ_t) and the number of activated gene sets (m_{on}). The dashed line marks the significance threshold (0.05). Red numbers at the bottom of each violin are the proportion of p-values below 0.05

most high-confidence gene sets had non-significant p-values (> 0.05), indicating convergence of the binary sampling for those gene sets. However, in low-signal datasets, the proportion of converged gene sets tends to be lower than in high-signal datasets (Fig. S1). The $\hat{R}$ diagnostic showed good convergence of the continuous parameters for most datasets. However, lack of convergence was observed in the combinations of $(\text{NCP} = 5, m_{on} = 5)$, $(10, 5)$ and $(10, 10)$ (Fig. S2).

CLEAR is a high-dimensional Bayesian model that samples from both discrete and continuous spaces. When the number of gene sets is high, it is challenging for CLEAR to fully explore the posterior distribution and converge to the global optimal, despite the large number of iterations. If multiple optimal or near-optimal solutions exist, different chains may converge to different regions of the posterior space. In the convergence analysis of the simulated datasets, we observed that some datasets having poor convergence metrics (i.e., $\hat{R} > 1.1$) exhibited multimodal behavior, that different chains converged to different posterior configurations with one or more gene sets having noticeably different posterior probabilities across chains (Fig. S3, Fig. S4).

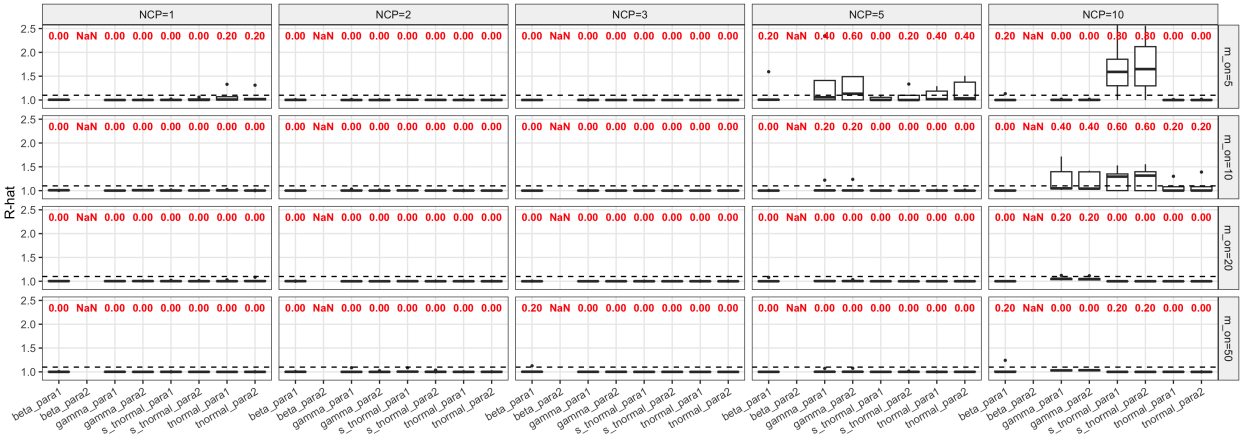


Figure S2: Gelman-Rubin convergence ($\hat{R}$) for each model parameter. Results are computed for simulated data sets under different data-generation configurations by the signal strength NCP (μ_t) and the number of activated gene sets (m_{on}). The dashed line marks the convergence threshold (1.1). Red numbers at the top of each violin are the proportion of $\hat{R}$ above 1.1

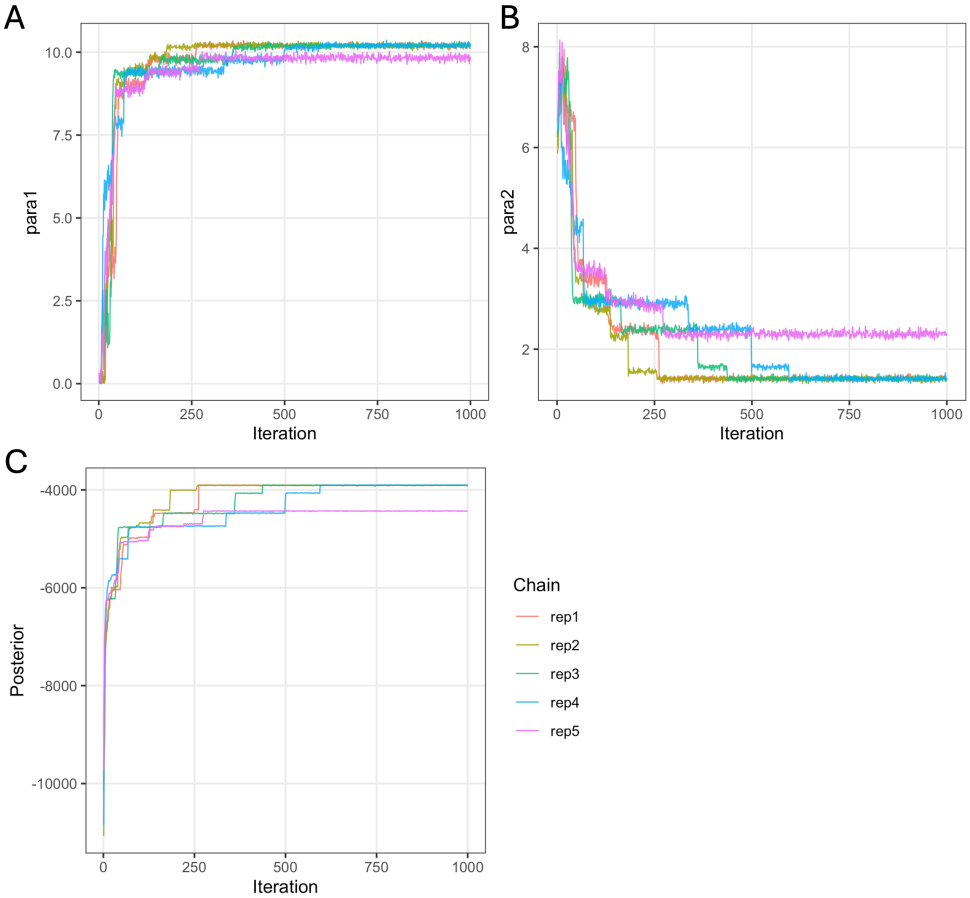
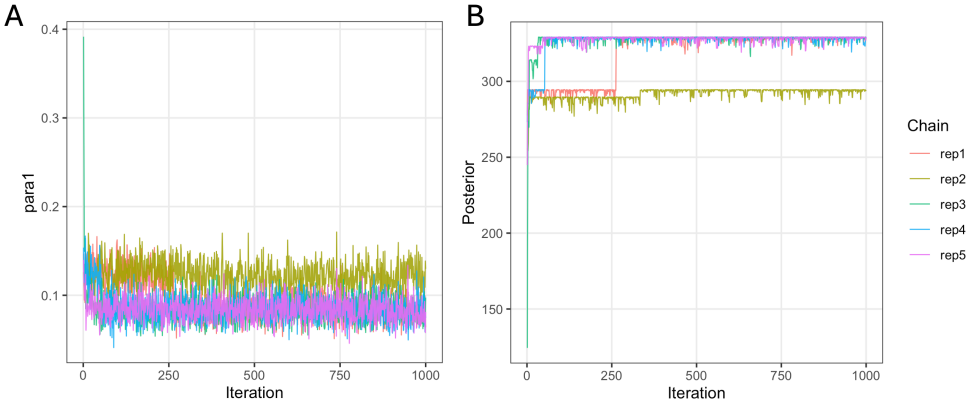


Figure S3: Convergence of CLEAR under the statistics model on a simulated dataset ($\mu_t = 10, m_{\text{on}} = 50$). The trace plots record every 1000 MCMC steps from the beginning to the end. (A) Trace plot of μ across five MCMC chains. (B) Trace plot of σ across five MCMC chains. (C) Trace plot of posterior across five MCMC chains.



C

gene_set	avg_posterior_prob	rep1	rep2	rep3	rep4	rep5
GO:0006869	1	1	1	1	1	1
GO:0033214	1	1	1	1	1	1
GO:0034061	1	1	1	1	1	1
GO:0015891	0.9648632	0.962852	0.969938	0.973306	0.96536	0.95286
GO:0008381	0.8001696	1	8.48E-04	1	1	1
GO:0034220	0.2	0	1	0	0	0
GO:0032150	0.0085644	9.52E-04	0.022116	0.006618	0.007046	0.00609
GO:0043846	0.0064232	0	0.003408	0.008294	0.009436	0.010978
GO:0015343	0.0050916	2.80E-04	0.011468	0.010006	0.002272	0.001432
GO:0038023	0.0044148	0.003374	0.004618	0	0.002732	0.01135

Figure S4: Convergence of CLEAR under the beta model on a simulated dataset ($\mu_t = 5, m_{\text{on}} = 5$). The trace plots record every 1000 MCMC steps from the beginning to the end. (A) Trace plot of a across five MCMC chains. (B) Trace plot of posterior across five MCMC chains. (C) Posterior probabilities of the top gene sets ranked by average posterior probability across chains.

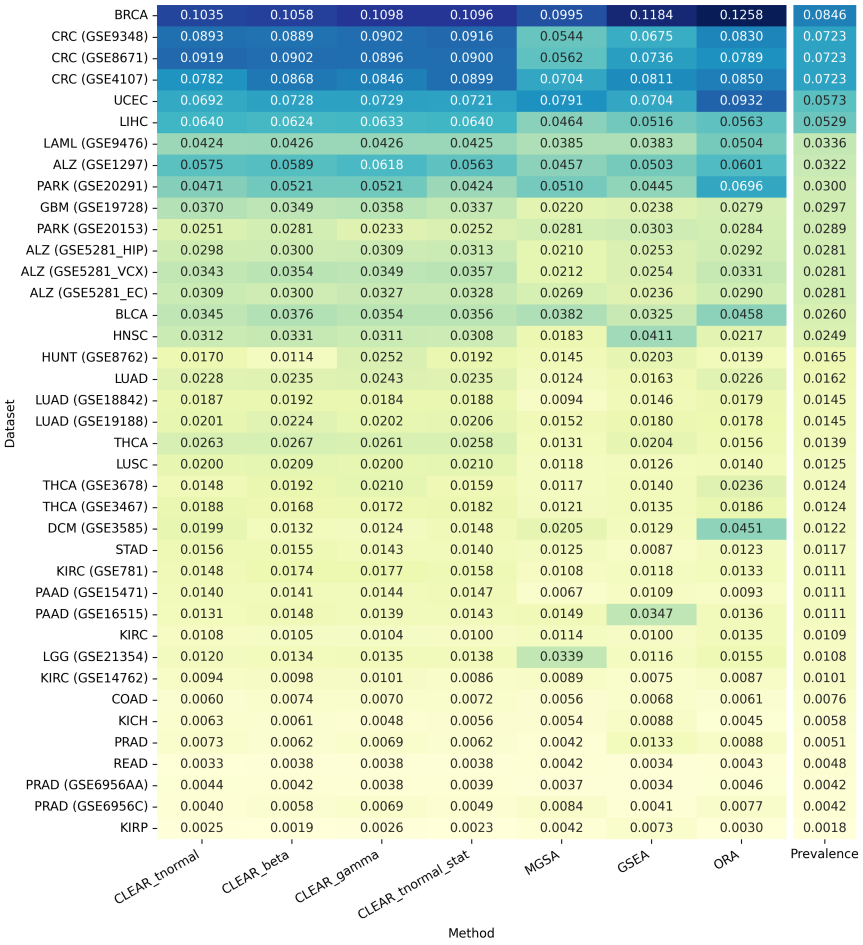


Figure S5: Heatmap showing area under precision-recall curve (PR-AUC) for all methods across 15 RNA-seq datasets (from TCGA) and 24 microarray datasets (from GEO, dataset ID is displayed), along with the prevalence of each dataset. Datasets are sorted by highest prevalence value. More than one dataset can be associated with the same type of cancer, and each cancer type has a unique set of phenotypically relevant gene sets (prevalence values could still differ due to difference in the number of valid gene sets in each dataset).

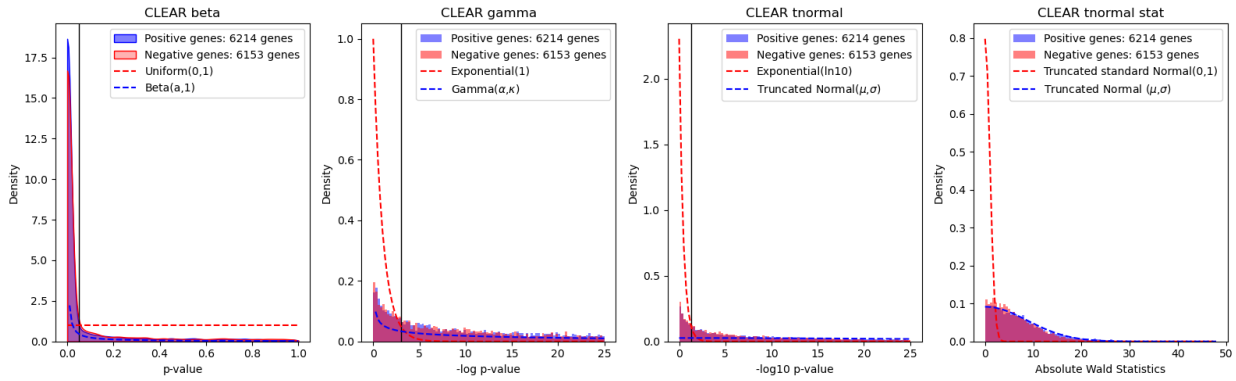


Figure S6: Gene-level statistic (p-value and Wald statistics) distributions of positive genes (genes belonging to cancer-relevant gene sets) and negative genes (remaining genes) in the BRCA dataset. Dashed red curves show the distributions under CLEAR's null hypothesis assumption. Dashed blue curves show a fitted distribution for the alternative hypothesis, where parameters were set to their posterior means.

	CLEAR_tnormal	CLEAR_beta	CLEAR_gamma	CLEAR_tnormal_stat	MGSA	GSEA	ORA	baseline
CLEAR_tnormal	1.0000	0.8834	0.9442	0.9838	0.0198	0.2626	0.9998	0.0089
CLEAR_beta	0.8834	1.0000	1.0000	0.9999	0.0001	0.0053	0.9880	0.0000
CLEAR_gamma	0.9442	1.0000	1.0000	1.0000	0.0002	0.0105	0.9972	0.0001
CLEAR_tnormal_stat	0.9838	0.9999	1.0000	1.0000	0.0006	0.0230	0.9998	0.0002
MGSA	0.0198	0.0001	0.0002	0.0006	1.0000	0.9785	0.0038	1.0000
GSEA	0.2626	0.0053	0.0105	0.0230	0.9785	1.0000	0.0901	0.9317
ORA	0.9998	0.9880	0.9972	0.9998	0.0038	0.0901	1.0000	0.0015
baseline	0.0089	0.0000	0.0001	0.0002	1.0000	0.9317	0.0015	1.0000

Figure S7: Pairwise post-hoc Nemenyi test p-values between normalized PR-AUC values of all methods. Baseline normalized PR-AUC value for each dataset is 1.0, which represents no gene set enrichment.

Table S1: List of 15 TCGA RNA-seq datasets and 24 GEO microarray datasets. Phenotypically relevant gene sets were obtained from GSEABenchmarkeR package and were used to determine positive genes and negative genes. Num_Genesets column shows the number of gene sets that have between 20 and 500 genes for each dataset.

Cancer Type	Number Positive Genes	Number Negative Genes	Number Phenotypically Relevant Genesets	Number Genesets
BLCA	3855	8223	86	3304
BRCA	6214	6153	282	3334
COAD	1744	10358	25	3301
HNSC	3620	8003	81	3259
KICH	1388	10711	19	3248
KIRC	1862	10430	36	3290
KIRP	269	11999	6	3272
LIHC	4846	5760	163	3081
LUAD	2873	9442	54	3340
LUSC	2529	9800	42	3357
PRAD	1921	10346	17	3302
READ	853	11322	16	3323
STAD	2252	9953	39	3326
THCA	3080	8942	45	3241
UCEC	5491	6644	187	3263
ALZ (GSE1297)	5363	7676	116	3600
KIRC (GSE14762)	2476	18882	39	3873
PAAD (GSE15471)	3539	17819	43	3873
PAAD (GSE16515)	3539	17819	43	3873
LUAD (GSE18842)	3473	17885	56	3873
LUAD (GSE19188)	3473	17885	56	3873
GBM (GSE19728)	5605	15753	115	3873
PARK (GSE20153)	4960	16398	112	3873
PARK (GSE20291)	4296	8743	108	3600
LGG (GSE21354)	3314	18044	42	3873
THCA (GSE3467)	3642	17716	48	3873
DCM (GSE3585)	2198	10841	44	3600
THCA (GSE3678)	3642	17716	48	3873
CRC (GSE4107)	8005	13353	280	3873
ALZ (GSE5281_EC)	5469	15889	109	3873
ALZ (GSE5281_HIP)	5469	15889	109	3873
ALZ (GSE5281_VCX)	5469	15889	109	3873
PRAD (GSE6956AA)	1620	11419	15	3600
PRAD (GSE6956C)	1620	11419	15	3600
KIRC (GSE781)	2160	10879	40	3600
CRC (GSE8671)	8005	13353	280	3873
HUNT (GSE8762)	3493	17865	64	3873
CRC (GSE9348)	8005	13353	280	3873
LAML (GSE9476)	5120	7919	121	3600