An Information Theoretic Framework for Protein Activity Measurement

Aaron T. Griffin[1,2], Lukas J. Vlahos[2], Codruta Chiuzan[3], Andrea Califano[2,4,5,6,7,8]

[1] Medical Scientist Training Program, Columbia University Irving Medical Center, New York, NY, USA

[2] Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA

[3] Department of Biostatistics, Columbia University Irving Medical Center, New York, NY, USA

[4] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

[5] Department of Medicine, Columbia University, New York, NY, USA

[6] JP Sulzberger Columbia Genome Center, Columbia University Irving Medical Center, New York, NY, USA

[7] Department of Biomedical Informatics, Columbia University, New York, NY, USA

[8] Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA

Abstract

Nonparametric analytical Rank-based Enrichment Analysis (NaRnEA) is a novel gene set analysis method which leverages an analytical null model derived under the Principle of Maximum Entropy. NaRnEA critically improves over two widely used methods – Gene Set Enrichment Analysis (GSEA) and analytical Rank-based Enrichment Analysis (aREA) – as shown by differential activity measurement of ~2,500 transcriptional regulatory proteins across three cohorts in The Cancer Genome Atlas (TCGA) based on the enrichment of their transcriptional targets in differentially expressed genes. Phenotype-matched proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) was used to evaluate measurement accuracy. We show that the sample-shuffling empirical null models leveraged by GSEA and aREA are overly conservative, a shortcoming that is critically addressed by NaRnEA's optimal analytical null model.

Introduction

The ever-increasing availability of next-generation sequencing technologies and highly accurate annotations for prokaryotic and eukaryotic genomes, especially the human genome, have transformed biology into a data-rich scientific discipline[1]. Thus, leveraging large-scale, gene-level biochemical measurements to support mechanistic inferences involving biological and cellular processes[2,3], as well as to measure protein and pathway activity[4], has become a critical priority in computational biology[5].

Not surprisingly, gene set analysis methods, which were developed precisely to address some of these challenges, have rapidly emerged as some of the most utilized tools in biomedical research. These methods provide a statistical assessment of the biological relevance of a specific set of genes – ranging from signal transduction pathways, gene ontology categories, and transcriptional targets of regulatory proteins, among others – based on their ranking according to a quantitative metric of interest. Most frequently, differential expression between two cellular states is used as the ranking criterion, but virtually any criteria to rank genes in a list can be used; see (Maleki et al. 2020 Front Genet)[6] and (Das et al. 2020 Entropy)[7] for recent reviews comparing the wide variety of published gene set analysis methods, as well as the statistical assumptions implicit to each method.

In this manuscript, we leverage the Principle of Maximum Entropy from Information Theory[8] to derive a provably optimal solution for determining the statistical significance of gene set enrichment which is implemented in the novel algorithm Nonparametric analytical Rank-based Enrichment Analysis (NaRnEA). We show that NaRnEA overcomes key limitations of two widely used gene set analysis methods – Gene Set Enrichment Analysis (GSEA)[9] and analytical Rank-based Enrichment Analysis (aREA)[4]. In addition, while the magnitude of a gene set's enrichment test statistic is strongly dependent on the size of the gene set, NaRnEA provides a Proportional Enrichment Score (PES) as an effect size that is virtually independent of gene set size. For simplicity, we focus on enrichment analysis in the context of differential gene expression; however, NaRnEA can trivially generalize to other quantitative ranking metrics and datasets.

Accurate benchmarking of gene set analysis methods using experimental data is challenging because the gene sets commonly used by these methodologies are often literature-based lists (e.g. KRAS pathway genes[10]) that lack any kind of independent, systematic validation[5-7]. However, recent work in cancer systems biology has shown that these methods can also be used to accurately measure the differential activity of transcriptional regulatory proteins, which provides a more objective benchmark that can be experimentally assessed. We define differential protein activity as the contribution of a transcriptional regulatory protein to the implementation of a specific differential gene expression signature. Consistent with this definition and akin to using a highly multiplexed gene reporter assay, we introduced the Virtual Inference of Protein Activity by Enriched Regulon Analysis (VIPER) algorithm[4] to measure protein activity based on the enrichment of transcriptional targets (i.e. regulon gene set) in a signature of differentially expressed genes.

The tissue-specific regulons necessary for these analyses can be effectively reverse-engineered using a variety of methods[11]. In the context of this study, we use the ARACNe3 algorithm, which is the most recent implementation of the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe)[12-14]; the previous versions of this algorithm have been experimentally validated as ARACNE-inferred regulons have been used extensively to measure differential protein activity in combination with VIPER, effectively elucidating Master Regulator (MR) proteins representing mechanistic determinants of tumor transcriptional state[15-17]. This paradigm has also been extended to measure differential protein activity in single cells, even when no transcripts can be detected for the encoding gene, with accuracy comparable to antibody-based experimental methods[18]. Thus, large-scale, mass spectrometry-based protein abundance protein abundance measurements and phenotype-matched gene expression profiles provide a unique, objective benchmark to evaluate improvements to the measurement of differential protein activity using gene set analysis methodologies.

Our study shows that, when assessed across three primary tumor cohorts from TCGA – lung adenocarcinoma (LUAD), colon adenocarcinoma (COAD), and head-neck squamous cell carcinoma (HNSC) – NaRnEA measures statistically significant differential activity for hundreds of transcriptional regulatory proteins that is otherwise undetectable by GSEA and aREA. Crucially, CPTAC proteomic data analysis showed significant correlation between differentially active proteins, as measured by NaRnEA, and differential protein abundance, as measured from CPTAC data.

NaRnEA and ARACNe3 are freely available for research use on GitHub (https://github.com/califano-lab/NaRnEA).

Results

Nonparametric analytical Rank-based Enrichment Analysis

NaRnEA is a nonparametric, frequentist statistical method designed to infer gene set enrichment according to the paradigm of null hypothesis significance testing. The nonparametric formulation of NaRnEA enables inference of gene set enrichment in gene expression signatures estimated with a variety of differential gene expression methods (e.g. DESeq2[19], Mann-Whitney U test). NaRnEA requires each member of a gene set to be parameterized with two values: (i) the Association Weight and (ii) the Association Mode. The Association Weight represents the likelihood that the specific gene truly belongs to the gene set. For example, in gene sets representing the regulon of a transcriptional regulatory protein, the Association Weight would indicate the confidence in the gene as a *bona fide* transcriptional target of the protein. The Association Mode represents whether the expression of the gene is positively or negatively associated with the enrichment of the gene set. This would, for instance, differentiate positive and negative regulators of apoptosis in a literature-curated apoptosis gene set. The Association Weight is a positive value, while the Association Mode should be a value in the [-1,1] range. As a result, NaRnEA can assess enrichment of simple gene sets, such as those constructed based on biological processes or pathways, as well as enrichment of more complex gene sets that incorporate substantial prior knowledge, such as regulons whose genes are associated with a specific likelihood and regulation directionality.

NaRnEA derives a fully analytical null model for gene set enrichment using the information theoretic Principle of Maximum Entropy (see Methods). Leveraging this provably optimal null model, NaRnEA calculates a gene set enrichment test statistic – referred to as the Normalized Enrichment Score (NES) – which is an asymptotically normal random variable when the gene set is not enriched in the gene expression signature (i.e. the null hypothesis for gene set analysis). A statistically significant positive or negative value of the NaRnEA NES provides evidence for positive or negative gene set enrichment, respectively. In addition to the NES, NaRnEA calculates a gene set enrichment effect size – referred as the Proportional Enrichment Score (PES) (see Methods). The NaRnEA PES is calculated by adjusting the NaRnEA NES by its maximum (or minimum) possible value in the case of positive (or negative) gene set enrichment, thus accounting for features specific to the gene set's size and parameterization as well as the gene expression signature itself. The NaRnEA PES is positive when the gene set is positively enriched, negative when the gene set is negatively enriched, and has a maximum magnitude of 1, thus providing a metric that is universally comparable between different gene set sizes (i.e. gene set size-independent) and across different gene expression signatures.

Gene Set Analysis Method Comparison

GSEA requires properly normalized gene expression profiles from samples representing a test phenotype and a reference phenotype; the differential expression of each gene is then estimated using an appropriate statistical test (e.g. Mann-Whitney U test). The enrichment of a gene set in this gene expression signature is calculated with a weighted Kolmogorov-Smirnov-like statistic (i.e. the GSEA enrichment score). It is recommended that the null distribution of the GSEA enrichment score for a particular gene set should be approximated using an empirical phenotype-based permutation procedure[9,20]; two-sided empirical p-values are calculated accordingly.

aREA accepts the same input data with differential gene expression estimated using Welch's unpaired t-test. The enrichment of a gene set in this gene expression signature is calculated using a three-tailed approach, returning an enrichment score test statistic. It is recommended that the null distribution for the aREA enrichment score for a particular gene set should also be approximated using an empirical phenotype-based permutation procedure or an approximate analytical approach if there are not enough samples for the permutation-based strategy; two-sided empirical p-values are calculated accordingly[4].

Benchmarking Gene Set Analysis Methods based on Differential Protein Activity Measurement

To compare the performance of different gene set analysis methods, we used NaRnEA, GSEA, and aREA to measure the differential activity of 2,491 transcription factor (TF) and co-transcription factor (coTF) proteins based on the enrichment of their ARACNe3-inferred regulons (i.e. transcriptional target gene sets) in genes

differentially expressed between primary tumor samples and adjacent normal tissue samples in three cohorts from TCGA as discussed in (*Alvarez et al. Nat Genet 2016*)[4]. For this, we reverse-engineered a tumor-specific regulons (TS-regulons) for each TF and coTF using LUAD, COAD, or HNSC primary tumor gene expression profiles from the respective TCGA cohorts. These cancer types were selected because of (a) the availability of phenotype-matched primary tumor and adjacent normal tissue RNA-Seq gene expression profiles in TCGA and (b) the availability of phenotype-matched primary tumor and adjacent normal tissue mass spectrometry protein abundance profiles in CPTAC, supporting direct assessment of transcriptional regulatory proteins that have differential activity and differential abundance between malignant and normal tissues. To avoid potential overfitting, primary tumor gene expression profiles which were used in the calculate of differential gene expression signatures were excluded from regulon reverse-engineering with ARACNe3. The remaining gene expression profiles from each cohort were used to reverse-engineer a TS-regulon for each TF and coTF protein with ARACNe3, the latest implementation of the ARACNe algorithm (see Methods). A null model regulon (NM-regulon) was constructed from each TS-regulon by replacing ARACNe3-inferred transcriptional targets with an equal number of genes randomly selected from genes not originally in the TS-regulon (i.e. the regulon gene set complement).

The LUAD transcriptional regulatory network was reverse-engineered from 476 primary tumor gene expression profiles and contains 2,491 regulators, 19,351 targets, and 947,776 interactions. The COAD transcriptional regulatory network was reverse-engineered from 437 primary tumor gene expression profiles and contains 2,491 regulators, 19,351 targets, and 646,341 interactions. The HNSC transcriptional regulatory network was reverse-engineered from 457 primary tumor gene expression profiles and contains 2,491 regulators, 19,351 targets, and 810,683 interactions.

Based on previous work, we assume that enrichment of TS-regulons in a differential gene expression signature will effectively measure the differential activity of the corresponding transcriptional regulatory protein[4,18], while NM-regulons should not produce any statistically significant enrichment, thus providing an optimal comparative null model. It is thus reasonable to expect that a subset of TS-regulons, yet no NM-regulons, will demonstrate statistically significant enrichment in the corresponding gene expression signature, as assessed by comparing gene expression profiles from primary tumor and adjacent normal tissue. In this benchmarking, TS-regulons which are not correctly inferred will behave like NM-regulons and will therefore not contribute to the difference in performance between the gene set analysis methods. We thus propose that the performance of a gene set analysis method can be evaluated based on (1) the ability of each method to differentiate between TS-regulons and NM-regulons, (2) the number of TS-regulons with statistically significant differential protein activity, (3) the number of NM-regulons with statistically significant differential protein activity, and (4) the agreement between the differential protein activity measured by each method and the differential protein abundance measured by mass spectrometry.

Gene Set Analysis Sensitivity and Specificity

We used NaRnEA, GSEA, and aREA to compute the enrichment of both TS-regulons and NM-regulons in differential gene expression signatures as assessed by comparing gene expression profiles of 57 LUAD, 41 COAD, and 43 HNSC patient-matched primary tumor and adjacent normal tissue samples from TCGA.

We generated receiver operating characteristic (ROC) curves for NaRnEA, GSEA, and aREA within each TCGA cohort by designating TS-regulons as cases (i.e. positives) and NM-regulons as controls (i.e. negatives) under the assumption that none of the NM-regulons should produce statistically significant enrichment, while a substantial fraction of the TS-regulons should be statistically significantly enriched in the corresponding tumor vs. tissue gene expression signature. The statistical significance of the Area Under the ROC curve (AUROC) was calculated using a one-tailed Mann-Whitney U test and the 95% confidence interval for each AUROC was estimated using DeLong's method[21]; DeLong's test was used to compare AUROC values for each pair of methods within each cancer subtype.

Based on this analysis, NaRnEA statistically significantly outperformed GSEA and aREA in its ability to differentiate between TS-regulons and NM-regulons based on the enrichment two-sided p-values, achieving a highly statistically significant AUROC in LUAD (AUROC = 0.741, p = 4.13e-191), COAD (AUROC = 0.698, p = 2.05e-129), and HNSC (AUROC = 0.716, p = 4.42e-154) (Figure 1). The AUROC for GSEA was statistically significant only in HNSC (AUROC = 0.523, p = 2.75e-3), while the AUROC for aREA was statistically significant

only in LUAD (AUROC = 0.555, p = 1.07e-11) and HNSC (AUROC = 0.521, p = 5.71e-3). DeLong's test shows that NaRnEA statistically significantly outperformed both GSEA and aREA in all three cohorts (FWER < 0.05).

Direct analysis of the two-sided p-values provides an estimate of the fraction of TS-regulons and NM-regulons identified as statistically significantly enriched by the different gene set analysis methods based on control of the False Positive Rate (FPR < 0.05), the False Discovery Rate (FDR < 0.05), or the Family Wise Error Rate (FWER < 0.05). In agreement with the ROC curve analysis, NaRnEA was the only gene set analysis method capable of identifying statistically significantly enriched TS-regulons after correcting for multiple hypothesis testing within each cancer cohort (Table 1). Furthermore, after correcting for multiple hypothesis testing, NaRnEA did not identify any statistically significantly enriched NM-regulons, demonstrating that NaRnEA is substantially more sensitive than GSEA or aREA, without loss of specificity.

Like a Volcano Plot[22], NaRnEA-based enrichment of TS-regulons for each tumor cohort can be effectively visualized by plotting the absolute value of the NaRnEA NES vs. the NaRnEA PES (Figure 2).

Comparison of Gene Set Enrichment and Differential Protein Abundance

We then proceeded to assess whether the differential protein activity measured by NaRnEA from TCGA transcriptomic data tracked with differential protein abundance, as experimentally assessed by mass spectrometry data from CPTAC. To this end, we used the Mann-Whitney U test to assess whether TF and coTF proteins were statistically significantly differentially abundant between primary tumor and phenotype-matched adjacent normal tissue samples, based on CPTAC proteomic data (FDR < 0.05). Specifically, we analyzed (n_{tumor} = 110, 97, 109) vs. (n_{tissue} = 101, 100, 64) from LUAD, COAD, and HNSC cohorts in CPTAC, respectively. While regulatory protein abundance is not a perfect proxy for its transcriptional activity – since protein activity is further determined by post-translational modifications, complex formation, and subcellular localization – we assumed that the two should be significantly correlated. Furthermore, since the statistical significance of differential protein activity and differential protein abundance were measured from gene expression data in TCGA and mass spectrometry data in CPTAC, respectively, the data are completely independent under the null hypothesis.

For each cancer type, we compared only the proteins for which a TS-regulon was available. Since mass spectrometry-based proteomic data do not provide a precise quantitative metric, we corrected the Mann-Whitney U test two-sided p-values from CPTAC for multiple hypothesis testing and classified each TF or coTF as upregulated (UP, Mann-Whitney U test rank-biserial correlation > 0, FDR < 0.05), downregulated (DOWN, Mann-Whitney U test rank-biserial correlation < 0, FDR < 0.05), or unchanged (NS, FDR ≥ 0.05). Similarly, we also corrected the NaRnEA two-sided p-values from TCGA for multiple hypothesis testing and classified each TF or CTF as activated (UP, NaRnEA PES > 0, FDR < 0.05), deactivated (DOWN, NaRnEA PES < 0, FDR < 0.05), or unchanged (NS, FDR ≥ 0.05).

We then tested for an association between NaRnEA-based differential protein activity from TCGA and Mann-Whitney U test-based differential protein abundance from CPTAC using a three-by-three contingency table for each cancer type (Table 2). We computed the agreement between the rows and columns using Kendall's Tau B correlation coefficient, which adjusts for tied values within each of the three marginal categories; the statistical significance of this dependence was calculated with a Chi-Squared test. Based on this analysis, we find that the differential protein activity measured by NaRnEA in TCGA was statistically significantly positively correlated with the differential protein abundance from CPTAC for LUAD (Kendall's Tau-B = 0.382, p = 3.14e-51), COAD (Kendall's Tau-B = 0.281, p = 1.12e-17), and HNSC (Kendall's Tau-B = 0.342, p = 3.63e-37) cancer types.

After relaxing the threshold for statistically significant differential protein activity (FDR < 0.25), the same analysis demonstrated that differential protein activity measured by aREA in the TCGA LUAD cohort was statistically significantly positively correlated with the differential protein abundance from the CPTAC LUAD cohort yet with much lower statistical significance and a smaller effect size (Kendall's Tau-B = 0.133, p = 2.75e-6). Neither the original analysis (FDR < 0.05) nor the modified analysis (FDR < 0.25) could be performed for GSEA or aREA in any of the other cohorts as no statistically significant differential protein activity could be measured by these gene set analysis methods even at the more relaxed levels. This reflects the fact that most approaches use GSEA or aREA without correcting for multiple hypothesis testing.

The low sensitivity of both GSEA and aREA can be traced directly to their reliance on null models for gene set enrichment that are estimated by an empirical phenotype-based permutation procedure this biased and underpowered. Indeed, we can show that >90% of null gene expression signatures produced by the empirical

phenotype-based permutation procedure are statistically significantly correlated with the original gene expression signature. Thus, the resulting empirical null model for gene set enrichment is contaminated with enrichment scores that follow the alternative distribution rather than the null distribution, since the gene set in question will be truly enriched in the supposed null gene expression signatures which are correlated with the true gene expression signature. This results in overestimation of empirical two-sided p-values and a substantial loss of statistical power for gene set analysis.

While aREA claims to provide an alternative method for assessing the statistical significance of gene set enrichment using an analytical null model, no formal proof for the validity of this analytical null model has been published. We can show that the aREA analytical null model is also flawed by testing for the enrichment of each of the TS-regulons from the TCGA LUAD cohort in null signatures produced by shuffling gene identifiers for the LUAD tumor vs. tissue gene expression signature; these null signatures contain realistic differential gene expression values but are no longer correlated with the original gene expression signature. The resulting aREA normalized enrichment scores should be normally distributed with a mean of zero and variance of one. Using a one-sample Kolmogorov-Smirnov test we can evaluate the standard normality of the aREA normalized enrichment scores for each regulon across 1,000 shuffled gene expression signatures; after correcting for multiple hypothesis testing (FWER < 0.05), we reject the null hypothesis of standard normality for 100% of the TS-regulons. An identical analysis with NaRnEA NES values does not lead us to reject the null hypothesis of standard normality for any of the TS-regulons. This numerically confirms the mathematical accuracy of the optimal NaRnEA analytical null model while simultaneously demonstrating that the aREA analytical null model is mathematically invalid.

GSEA offers an alternative method for assessing the statistical significance of gene set enrichment using a null model estimated by an empirical gene-based permutation procedure; this approach may be regarded as an approximation of the NaRnEA analytical null model which still does not overcome the other limitations of GSEA (e.g. inability to incorporate Association Weight and/or Association Mode parameters, empirical minimum p-values determined by the number of permutations for the empirical null model, high computational complexity). Most importantly, assessing statistical significance below the 10^{-5} threshold would require an extraordinarily large number of permutations for the empirical null model, thus making the procedure computationally intractable.

Taken together, these results show that NaRnEA is dramatically more effective than both GSEA and aREA in identifying biologically relevant enrichment of gene sets.

Discussion

It is widely appreciated that the rigor and reproducibility of scientific research depends on the use of computational and experimental methodologies which are sufficiently sensitive to make meaningful inferences while maintaining strong control over the Type I error rate to reduce spurious findings[23]. Gene set analysis methods are finding increasing application for hypothesis generation[24], precision oncology[15], systems pharmacology[25], analysis of single-cell transcriptomic data[18,26], and biomarker development[27]. Here we demonstrate that NaRnEA significantly outperforms GSEA and aREA for the purpose of gene set analysis in three independent cancer types; despite the widespread use of both competing gene set analysis methods, NaRnEA is the only method capable of consistently distinguishing between biologically coherent gene sets and gene sets constructed at random in these cohorts. Furthermore, the substantial agreement between NaRnEA-based differential protein activity in TCGA cohorts and mass spectrometry-based differential protein abundance in phenotype-matched CPTAC cohorts confirms that gene set enrichment inferred by NaRnEA cannot be explained away as erroneous false positives. Indeed, the specificity of NaRnEA is established by the fact that NaRnEA does not measure statistically significant differential protein activity for any NM-regulons in TCGA cohorts. To encourage immediate use by members of the scientific community, both NaRnEA and ARACNe3 are freely available for research use on GitHub (https://github.com/califano-lab/NaRnEA).

Methods

NaRnEA Mathematical Framework

The following section serves as a comprehensive derivation of the mathematical and conceptual framework underlying NaRnEA, which is designed to test for the enrichment of a gene set in a gene expression signature using a null model for gene set enrichment derived under the Principle of Maximum Entropy from information theory[8,28]; this formulation renders the NaRnEA analytical null model unique and optimal from an information theoretic perspective in the sense that any other null model must make operate under assumptions that cannot be guaranteed when the null hypothesis is true (i.e. that the gene set being analyzed is not enriched in the gene expression signature in question).

Let $\{Z_g\}_{AB}$ be a vector of length (G), where $g \in \{1, \dots, G\}$, that represents the differential gene expression signature between biological phenotype (A) and biological phenotype (B). Let (X_{gA}) be a random variable that represents the relative molar concentration of the transcripts originating from the gth gene in biological phenotype (A). Similarly, let (X_{gB}) be a random variable which represents the relative molar concentration of the transcripts originating from the gth gene in biological phenotype (B). NaRnEA assumes that (Z_{gAB}) has the following properties:

i. $Z_{gAB} \in \mathbb{R}$

ii. $Z_{gAB} = -Z_{gBA}$

iii. $Z_{gAB} \geq 0 \; iff \; \Pr(X_{gA} \geq X_{gB}) \geq \frac{1}{2}$

iv. $Z_{gAB} < 0 \; iff \; \Pr(X_{gA} \geq X_{gB}) < \frac{1}{2}$

v. $|Z_{gAB}| \geq |Z_{kAB}| \; iff \; \left|\Pr(X_{gA} \geq X_{gB}) - \frac{1}{2}\right| \geq \left|\Pr(X_{kA} \geq X_{kB}) - \frac{1}{2}\right| \; for \; g \neq k$

where $\mathbb{R}$ is the set of all real numbers and $\Pr(\cdot)$ is the probability of an event.

NaRnEA tests for gene set enrichment after nonparametrically transforming the original gene expression signature; the properties of (Z_{gAB}) are necessary and sufficient to guarantee that this nonparametric transformation preserves the relative magnitude and directionality of differential expression for each gene between phenotype (A) and phenotype (B).

NaRnEA leverages the Association Weight and Association Mode parameters so that users may construct gene sets whose members contribute unequally to gene set enrichment and which exhibit complex and/or multimodal enrichment in the gene expression signature. ARACNe3-inferred regulon gene sets for transcriptional regulatory proteins are an excellent example of such complex gene sets since the gene expression of transcriptional targets may be positively or negatively correlated with the transcriptional regulator's activity and the expression of transcriptional targets can demonstrate varying degrees of dependency with respect to the transcriptional regulatory protein activity.

We describe the statistical framework of NaRnEA from the biological perspective of measuring differential activity of transcriptional regulatory proteins using regulon gene sets, but we note that NaRnEA can be immediately generalized to gene sets whose members share a common gene ontology or belong to the same biochemical pathway; for these gene sets, Association Weight and Association Mode parameters must be estimated based on domain-specific prior knowledge.

Let (AW_{gr}) be the Association Weight between the gth gene and the rth transcriptional regulator. Let (X_g) be the random variable representing the relative molar concentration of the transcripts originating from the gth gene. Let (Y_r) be the random variable representing the biochemical activity of the rth transcriptional regulator. Let $I[X;Y]$ be the mutual information between the random variables (X) and (Y). NaRnEA assumes that (AW_{gr}) has the following properties:

i. $AW_{gr} \geq 0$

ii. $AW_{gr} > 0 \; iff \; I[X_g; Y_r] > 0$

iii. $AW_{gr} \geq AW_{kr} \; iff \; I[X_g; Y_r] \geq I[X_k; Y_r] \; for \; g \neq k$

Let (AM_{gr}) be the Association Mode between the gth gene and rth transcriptional regulator. Let $SCC[X,Y]$ be the Spearman correlation coefficient between the random variables (X) and (Y). NaRnEA assumes that (AM_{gr}) has the following properties:

i. $AM_{gr} \in [-1, 1]$

ii. $AM_{gr} \geq 0 \; iff \; SCC[X_g, Y_r] \geq 0$

iii. $AM_{gr} < 0 \; iff \; SCC[X_g, Y_r] < 0$

iv. $|AM_{gr}| \geq |AM_{kr}| \; iff \; |SCC[X_g, Y_r]| \geq |SCC[X_k, Y_r]| \; for \; g \neq k$

To test for the enrichment of the regulon gene set belonging to the rth transcriptional regulator in the gene expression signature $\{Z_g\}_{AB}$, NaRnEA calculates the Directed Enrichment Score (D_{rAB}), which is sensitive to direction and magnitude of gene set member differential gene expression, as well as the Undirected Enrichment Score (U_{rAB}), which is sensitive only to the magnitude of gene set member differential gene expression, as follows:

$$D_{rAB} = \sum_{g=1}^{G} D_{grAB}$$

$$D_{grAB} = (AW_{gr})(AM_{gr})(r_{gAB} s_{gAB})$$

$$U_{rAB} = \sum_{g=1}^{G} U_{grAB}$$

$$U_{grAB} = (AW_{gr})(1 - |AM_{gr}|)(r_{gAB})$$

$$r_{gAB} = rank(|Z_{gAB}|) \qquad r_{grAB} \in \{1, \dots, G\}$$

$$s_{gAB} = sign(Z_{gAB}) \qquad s_{grAB} \in \{1, -1\}$$

The nonparametric transformation of the original gene expression signature $\{Z_g\}_{AB} \longmapsto \{r_g s_g\}_{AB}$ preserves the relative magnitude and directionality of differential expression for each gene between the two phenotypes. Additionally, it simplifies subsequent calculations and allows NaRnEA to analyze gene expression signatures estimated with a variety of differential gene expression methods.

From the previously stated assumptions made by NaRnEA about the biological interpretability of (Z_{gAB}), (AW_{gr}), and (AM_{gr}), we can clarify the inference of gene set enrichment using the language of frequentist null hypothesis significance testing as follows:

H_o: the rth gene set is not enriched in $\{Z_g\}_{AB}$.

H_a^+: the rth gene set is positively enriched in $\{Z_g\}_{AB}$.

H_a^-: the rth gene set is negatively enriched in $\{Z_g\}_{AB}$.

Here the alternative hypothesis of gene set enrichment has been subdivided into two mutually exclusive versions; each of these versions implies that the members of the rth gene set are enriched at one or both extremes of the gene expression signature $\{Z_g\}_{AB}$. Under (H_a^+) we would expect $(Z_{gAB} \geq 0)$ for those members of the rth gene set for which $(AM_{gr} \geq 0)$ and $(Z_{gAB} < 0)$ for those members of the rth gene set for which $(AM_{gr} < 0)$. In contrast, under (H_a^-) we would expect $(Z_{gAB} < 0)$ for those members of the rth gene set for which $(AM_{gr} \geq 0)$ and $(Z_{gAB} \geq 0)$ for those members of the rth gene set for which $(AM_{gr} < 0)$.

We can derive the marginal sampling distributions for (D_{rAB}) and (U_{rAB}) under (H_o) using the information theoretic Principle of Maximum Entropy, which states that we should select, as our null model, the probability distribution with the greatest entropy constrained on all information available under (H_o). In terms of available information under (H_o), we know the following:

i. the genes for which $(AW_{gr} > 0)$ are members of the rth gene set.

ii. the members of the rth gene set are present in $\{Z_g\}_{AB}$.

Following the nonparametric transformation of the gene expression signature, we know that each member of the gene set may occupy one of (G) possible values $\{r_1 s_1, \dots, r_G s_G\}_{AB}$. Thus, for all members of the rth gene set, (D_{grAB}) and (U_{grAB}) are discrete random variables under (H_o).

Let $H[X]$ be the Shannon entropy for the random variable (X). Without any additional information and for (G) much larger than the number of members in the rth gene set, $H[D_{grAB}|H_o]$ and $H[U_{grAB}|H_o]$ are maximized when $p(r_{gAB}s_{gAB}|H_o)$ takes the form of a discrete uniform distribution over the entire set of $\{r_1 s_1, \dots, r_G s_G\}_{AB}$ (i.e. it is equiprobable that a gene set member is located in any position in the nonparametric gene expression signature). Furthermore, without any constraints regarding higher-order dependencies between the members of the rth gene set under (H_o), $H[D_{rAB}|H_o]$ and $H[U_{rAB}|H_o]$ are maximized when all members of the rth gene set are independent under (H_o).

From this maximum entropy null model for gene set enrichment, we may derive the expected value and variance of (D_{rAB}) under (H_o) as follows:

$$E[D_{rAB}|H_o] = \sum_{g=1}^{G} E[D_{grAB}|H_o]$$

$$E[D_{grAB}|H_o] = (AW_{gr})(AM_{gr})E[r_{gAB}s_{gAB}|H_o]$$

$$E[r_{gAB}s_{gAB}|H_o] = \sum_{k=1}^{G}(r_{kAB}s_{kAB})Pr(r_{gAB}s_{gAB} = r_{kAB}s_{kAB}|H_o) = \frac{1}{G}\sum_{k=1}^{G} r_{kAB}s_{kAB}$$

$$Var[D_{rAB}|H_o] = \sum_{g=1}^{G} Var[D_{grAB}|H_o]$$

$$Var[D_{grAB}|H_o] = E[D_{grAB}^2|H_o] - E[D_{grAB}|H_o]^2$$

$$E[D_{grAB}^2|H_o] = (AW_{gr})^2(AM_{gr})^2 E[r_{gAB}^2|H_o]$$

$$E[r_{gAB}^2|H_o] = \sum_{k=1}^{G}(r_{kAB}^2)\,\mathrm{Pr}(r_{gAB} = r_{kAB}|H_o) = \frac{1}{G}\sum_{k=1}^{G} r_{kAB}^2 = \frac{(2G^2 + 3G + 1)}{6}$$

From these first two moments we can compute $NES[D_{rAB}]$, the Normalized Directed Enrichment Score for the rth gene set in the gene expression signature $\{Z_g\}_{AB}$, as follows:

$$NES[D_{rAB}] = \frac{D_{rAB} - E[D_{rAB}|H_o]}{\sqrt{Var[D_{rAB}|H_o]}}$$

Similarly, from the maximum entropy null model for gene set enrichment, we may derive the expected value and variance of (U_{rAB}) under (H_o) as follows:

$$E[U_{rAB}|H_o] = \sum_{g=1}^{G} E[U_{grAB}|H_o]$$

$$E\left[U_{grAB}\middle|H_o\right] = \left(AW_{gr}\right)\left(1 - \left|AM_{gr}\right|\right)E\left[r_{gAB}\middle|H_o\right]$$

$$E\left[r_{gAB}\middle|H_o\right] = \sum_{k=1}^{G}(r_{kAB})\Pr\left(r_{gAB} = r_{kAB}\middle|H_o\right) = \frac{1}{G}\sum_{k=1}^{G} r_{kAB} = \frac{(G+1)}{2}$$

$$Var[U_{rAB}|H_o] = \sum_{g=1}^{G} Var\left[U_{grAB}\middle|H_o\right]$$

$$Var\left[U_{grAB}\middle|H_o\right] = E\left[U_{grAB}^2\middle|H_o\right] - E\left[U_{grAB}\middle|H_o\right]^2$$

$$E\left[U_{grAB}^2\middle|H_o\right] = \left(AW_{gr}\right)^2\left(1 - \left|AM_{gr}\right|\right)^2 E\left[r_{gAB}^2\middle|H_o\right]$$

$$E\left[r_{gAB}^2\middle|H_o\right] = \sum_{k=1}^{G}\left(r_{kAB}^2\right)\Pr\left(r_{gAB} = r_{kAB}\middle|H_o\right) = \frac{1}{G}\sum_{k=1}^{G} r_{kAB}^2 = \frac{(2G^2 + 3G + 1)}{6}$$

From these first two moments we can compute $NES[U_{rAB}]$, the Normalized Undirected Enrichment Score for the rth gene set in the gene expression signature $\{Z_g\}_{AB}$, as follows:

$$NES[U_{rAB}] = \frac{U_{rAB} - E[U_{rAB}|H_o]}{\sqrt{Var[U_{rAB}|H_o]}}$$

Since $NES[D_{rAB}]$ is a random variable that is equal to the sum of independent random variables under (H_o), according to the Lindeberg Central Limit Theorem[29], if

$$\lim_{G\to\infty} \frac{\sum_{g=1}^{G} E\left[\left(D_{grAB} - E\left[D_{grAB}\middle|H_o\right]\right)^2 \times \mathbb{I}\left\{\left|D_{grAB} - E\left[D_{grAB}\middle|H_o\right]\right| > \varepsilon\sqrt{Var[D_{rAB}|H_o]}\right\}\middle|H_o\right]}{Var[D_{rAB}|H_o]} = 0$$

for all $(\varepsilon > 0)$ where $\mathbb{I}\{\cdot\}$ is the identity operator, then

$$p(NES[D_{rAB}]|H_o) \xrightarrow{D} \mathcal{N}(0,1)$$

where $\xrightarrow{D}$ is convergence in distribution. Thus $NES[D_{rAB}]$ is asymptotically a standard normal random variable under (H_o) according to the maximum entropy null model for gene set enrichment if Lindeberg's condition is met.

Similarly, since $NES[U_{rAB}]$ is a random variable that is equal to the sum of independent random variables under (H_o), according to the Lindeberg Central Limit Theorem, if

$$\lim_{G\to\infty} \frac{\sum_{g=1}^{G} E\left[\left(U_{grAB} - E\left[U_{grAB}\middle|H_o\right]\right)^2 \times \mathbb{I}\left\{\left|U_{grAB} - E\left[U_{grAB}\middle|H_o\right]\right| > \varepsilon\sqrt{Var[U_{rAB}|H_o]}\right\}\middle|H_o\right]}{Var[U_{rAB}|H_o]} = 0$$

for all $(\varepsilon > 0)$ where $\mathbb{I}\{\cdot\}$ is the identity operator, then

$$p(NES[U_{rAB}]|H_o) \xrightarrow{D} \mathcal{N}(0,1)$$

where $\xrightarrow{D}$ is convergence in distribution. Thus $NES[U_{rAB}]$ is asymptotically a standard normal random variable under (H_o) according to the maximum entropy null model for gene set enrichment if Lindeberg's condition is met.

Due to the nonparametric transformation of the gene expression signature, we find that this convergence to normality occurs sufficiently well to achieve strong control of the Type I error rate for gene sets with at least 30 members that are well-balanced (i.e. the Association Weight and Association Mode values for the gene set members do not vary over a large range).

Furthermore, we may generalize these statements using the multivariate central limit theorem and make the following claim about the joint distribution of $NES[D_{rAB}]$ and $NES[U_{rAB}]$ under (H_o):

$$p\left(\begin{pmatrix} NES[D_{rAB}] \\ NES[U_{rAB}] \end{pmatrix} \middle| H_o\right) \xrightarrow{D} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & Cov[NES[D_{rAB}], NES[U_{rAB}]|H_o] \\ Cov[NES[D_{rAB}], NES[U_{rAB}]|H_o] & 1 \end{pmatrix}\right)$$

$$Cov[NES[D_{rAB}], NES[U_{rAB}]|H_o] = \rho[NES[D_{rAB}], NES[U_{rAB}]|H_o]$$

$$\rho[NES[D_{rAB}], NES[U_{rAB}]|H_o] = \rho[D_{rAB}, U_{rAB}|H_o]$$

$$\rho[D_{rAB}, U_{rAB}|H_o] = \frac{Cov[D_{rAB}, U_{rAB}|H_o]}{\sqrt{Var[D_{rAB}|H_o]Var[U_{rAB}|H_o]}}$$

$$Cov[D_{rAB}, U_{rAB}|H_o] = E[D_{rAB}U_{rAB}|H_o] - E[D_{rAB}|H_o]E[U_{rAB}|H_o]$$

$$E[D_{rAB}U_{rAB}|H_o] = E\left[\left(\sum_{g=1}^{G} D_{grAB}\right)\left(\sum_{g=1}^{G} U_{grAB}\right)\middle|H_o\right]$$

$$\ldots = E\left[\sum_{g=1}^{G} D_{grAB}U_{grAB} + \sum_{g\neq k} D_{grAB}U_{krAB}\,\middle|H_o\right]$$

$$\ldots = \sum_{g=1}^{G} E\left[D_{grAB}U_{grAB}\middle|H_o\right] + \sum_{g\neq k} E\left[D_{grAB}\middle|H_o\right]E[U_{krAB}|H_o]$$

$$E\left[D_{grAB}U_{grAB}\middle|H_o\right] = \left(AW_{gr}\right)^2\left(AM_{gr}\right)\left(1 - \left|AM_{gr}\right|\right)E\left[r_{gAB}^2 s_{gAB}\middle|H_o\right]$$

$$E\left[r_{gAB}^2 s_{gAB}\middle|H_o\right] = \sum_{k=1}^{G}\left(r_{kAB}^2 s_{kAB}\right)\Pr\left(r_{gAB}^2 s_{gAB} = r_{kAB}^2 s_{kAB}\middle|H_o\right) = \frac{1}{G}\sum_{k=1}^{G} r_{kAB}^2 s_{kAB}$$

where $Cov[X, Y]$ is the covariance between the random variables (X) and (Y) and $\rho[X, Y]$ is the Pearson product-moment correlation between the random variables (X) and (Y).

To summarize, according to the maximum entropy null model for gene set enrichment and the multivariate generalization of the Lindeberg Central Limit Theorem, $NES[D_{rAB}]$ and $NES[U_{rAB}]$ form an asymptotically bivariate normal random vector under (H_o) if Lindeberg's condition is met for both $NES[D_{rAB}]$ and $NES[U_{rAB}]$ under (H_o).

From the assumptions about the distribution of the gene set members in the gene expression signature under the alternative hypothesis, we expect $(NES[D_{rAB}] > 0)$ under (H_a^+) and $(NES[D_{rAB}] < 0)$ under (H_a^-). In contrast, we expect $(NES[U_{rAB}] > 0)$ under (H_a^+) and $(NES[U_{rAB}] > 0)$ under (H_a^-). To integrate the statistical significance of $NES[D_{rAB}]$ and $NES[U_{rAB}]$ we perform the following change of variables:

$$NES_{rAB}^{+} = \frac{NES[D_{rAB}] + NES[U_{rAB}]}{\sqrt{2 + 2Cov[NES[D_{rAB}], NES[U_{rAB}]|H_o]}}$$

$$NES_{rAB}^{-} = \frac{NES[D_{rAB}] - NES[U_{rAB}]}{\sqrt{2 - 2Cov[NES[D_{rAB}], NES[U_{rAB}]|H_o]}}$$

This change of variables can be represented with an affine transformation of the original bivariate normal random vector as follows:

$$\begin{pmatrix} NES_{rAB}^{+} \\ NES_{rAB}^{-} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2+2\theta}} \\ \sqrt{\frac{1}{2-2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix}\begin{pmatrix} NES[D_{rAB}] \\ NES[U_{rAB}] \end{pmatrix}$$

where $\theta = Cov[NES[D_{rAB}], NES[U_{rAB}]|H_o]$.

It is a well-established result that an affine transformation of a multivariate normal random vector produces a multivariate normal random vector[30]. More formally, if $(Y = c + BX)$ is an affine transformation of $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ where (c) is a (2×1) vector of constants and (B) is a (2×2) matrix, then $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, where

$$\mu_Y = c + B\mu_X$$

$$\Sigma_Y = B\Sigma_X B^T$$

It follows from this statement that (NES_{rAB}^{+}) and (NES_{rAB}^{-}) form a bivariate normal random vector under (H_o) whose mean vector and variance-covariance matrix may be calculated as follows:

$$\begin{pmatrix} E[NES_{rAB}^{+}|H_o] \\ E[NES_{rAB}^{-}|H_o] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2+2\theta}} \\ \sqrt{\frac{1}{2-2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix} \begin{pmatrix} E[NES[D_{rAB}]|H_o] \\ E[NES[U_{rAB}]|H_o] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} Var[NES_{rAB}^{+}|H_o] & Cov[NES_{rAB}^{+}, NES_{rAB}^{-}|H_o] \\ Cov[NES_{rAB}^{+}, NES_{rAB}^{-}|H_o] & Var[NES_{rAB}^{-}|H_o] \end{pmatrix} = \cdots$$

$$\cdots = \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2+2\theta}} \\ \sqrt{\frac{1}{2-2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix} \begin{pmatrix} Var[NES[D_{rAB}]|H_o] & \theta \\ \theta & Var[NES[U_{rAB}]|H_o] \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2+2\theta}} \\ \sqrt{\frac{1}{2-2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix}^T$$

$$\cdots = \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2+2\theta}} \\ \sqrt{\frac{1}{2-2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix} \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2-2\theta}} \\ \sqrt{\frac{1}{2+2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix}$$

$$\cdots = \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2+2\theta}} \\ \sqrt{\frac{1}{2-2\theta}} & -\sqrt{\frac{1}{2-2\theta}} \end{pmatrix} \begin{pmatrix} \sqrt{\frac{1}{2+2\theta}} + \theta\sqrt{\frac{1}{2+2\theta}} & \sqrt{\frac{1}{2-2\theta}} - \theta\sqrt{\frac{1}{2-2\theta}} \\ \theta\sqrt{\frac{1}{2+2\theta}} + \sqrt{\frac{1}{2+2\theta}} & \theta\sqrt{\frac{1}{2-2\theta}} - \sqrt{\frac{1}{2-2\theta}} \end{pmatrix}$$

$$Var[NES_{rAB}^{+}|H_o] = \frac{1}{2+2\theta} + \frac{\theta}{2+2\theta} + \frac{\theta}{2+2\theta} + \frac{1}{2+2\theta} = \frac{2+2\theta}{2+2\theta} = 1$$

$$Var[NES_{rAB}^{-}|H_o] = \frac{1}{2-2\theta} - \frac{\theta}{2-2\theta} - \frac{\theta}{2-2\theta} + \frac{1}{2-2\theta} = \frac{2-2\theta}{2-2\theta} = 1$$

$$Cov[NES_{rAB}^{+}, NES_{rAB}^{-}|H_o] = \sqrt{\frac{1}{2+2\theta}}\sqrt{\frac{1}{2-2\theta}} - \theta\sqrt{\frac{1}{2+2\theta}}\sqrt{\frac{1}{2-2\theta}} + \theta\sqrt{\frac{1}{2+2\theta}}\sqrt{\frac{1}{2-2\theta}} - \sqrt{\frac{1}{2+2\theta}}\sqrt{\frac{1}{2-2\theta}} = 0$$

$$\begin{pmatrix} Var[NES_{rAB}^{+}|H_o] & Cov[NES_{rAB}^{+}, NES_{rAB}^{-}|H_o] \\ Cov[NES_{rAB}^{+}, NES_{rAB}^{-}|H_o] & Var[NES_{rAB}^{-}|H_o] \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Thus, (NES_{rAB}^{+}) and (NES_{rAB}^{-}) are independent standard normal random variables under (H_o). This is statistically advantageous since we expect $(NES_{rAB}^{+} > 0)$ under (H_o) and $(NES_{rAB}^{-} < 0)$ under (H_o). We may leverage the asymptotic standard normality of (NES_{rAB}^{+}) and (NES_{rAB}^{-}) to calculate the one-tailed p-values for these test statistics as follows:

$$p_{rAB}^{+} = 1 - \Phi(NES_{rAB}^{+})$$
$$p_{rAB}^{-} = \Phi(NES_{rAB}^{-})$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Since (H_a^{+}) and (H_a^{-}) are mutually exclusive, we may use $\min\{p_{rAB}^{+}, p_{rAB}^{-}\}$ to determine which version of the alternative hypothesis is more probable. This decision constitutes a form of multiple hypothesis testing; as a result, we must adjust the final two-sided p-value to control the NaRnEA Type I error rate. Given that (p_{rAB}^{+}) and (p_{rAB}^{-}) are independent under (H_o), the final NaRnEA two-sided p-value for the rth gene set may be calculated as follows:

$$p_{rAB} = 1 - (1 - \min\{p_{rAB}^{+}, p_{rAB}^{-}\})^2$$

The final NaRnEA two-sided p-value is equal to the probability of observing at least one value less than or equal to $\min\{p_{rAB}^{+}, p_{rAB}^{-}\}$ when drawing two independently and identically distributed observations from $U[0,1]$, (i.e. the sampling distribution of the independent one-tailed p-values under the null hypothesis). We combine the directional information implied by $\min\{p_{rAB}^{+}, p_{rAB}^{-}\}$ with the magnitude of the final NaRnEA two-sided p-value to calculate (NES_{rAB}), the final NaRnEA Normalized Enrichment Score for the enrichment of the rth gene set in the gene expression signature $\{Z_g\}_{AB}$, as follows:

$$NES_{rAB} = \begin{cases} \Phi^{-1}\left(1 - \frac{p_{rAB}}{2}\right) \; iff \; p_{rAB}^{+} < p_{rAB}^{-} \\ \Phi^{-1}\left(\frac{p_{rAB}}{2}\right) \; iff \; p_{rAB}^{+} > p_{rAB}^{-} \end{cases}$$

As a result, the final NaRnEA Normalized Enrichment Score for the enrichment of the rth gene set in the gene expression signature $\{Z_g\}_{AB}$ is endowed with the following properties:

i. $p(NES_{rAB}|H_o) \xrightarrow{D} \mathcal{N}(0,1)$
ii. $E[NES_{rAB}|H_a^{+}] > 0$
iii. $E[NES_{rAB}|H_a^{-}] < 0$

Thus, (NES_{rAB}) serves as a single test statistic for gene set enrichment and the statistical significance of (NES_{rAB}) may be computed analytically using the standard normal cumulative distribution function.

As a test statistic, we expect the sampling distribution of (NES_{rAB}) under (H_a) to vary depending on gene set size, parameterization, and features specific to the gene expression signature under investigation; therefore, we also require an effect size for gene set enrichment.

The NaRnEA Proportional Enrichment Score (PES_{rAB}) for the enrichment of the rth gene set in the gene expression signature $\{Z_g\}_{AB}$ is calculated from (NES_{rAB}) as follows:

$$PES_{rAB} = \begin{cases} \frac{NES_{rAB}}{|\max\{NES_{rAB}\}|} \; if \; NES_{rAB} > 0 \\ \frac{NES_{rAB}}{|\min\{NES_{rAB}\}|} \; if \; NES_{rAB} < 0 \end{cases}$$

This formulation endows (PES_{rAB}) with the following properties:

i. $PES_{rAB} \in [-1, 1]$

ii. $E[PES_{rAB}|H_a^+] > 0$

iii. $E[PES_{rAB}|H_a^-] < 0$

The NaRnEA Proportional Enrichment Score (PES_{rAB}) for the enrichment of the rth gene set in the gene expression signature $\{Z_g\}_{AB}$ functions like a correlation coefficient in that it is bounded on the interval [-1,1], has a positive expected value when the gene set is positively enriched, has a negative expected value when the gene set is negatively enriched, and may be used to quantitatively compare the enrichment of gene sets with different parameterizations in the same gene expression signature or across different gene expression signatures. Confidence intervals for the NaRnEA Proportional Enrichment Score (PES_{rAB}) may be approximated by bootstrapping the members of the rth gene set, recomputing (PES_{rAB}) for the bootstrapped versions of the gene set, and applying an appropriate estimation procedure[31].

Finally, NaRnEA offers post-hoc leading-edge (ledge) analysis to quantitatively evaluate the extent to which each gene set member contributes to the enrichment of the gene set in the gene expression signature. This ledge analysis is conceptually similar to the leading-edge analysis proposed for GSEA by (*Subramanian et al. 2005 Proc Natl Acad Sci*)[9], but rather than returning a dichotomous leading edge the NaRnEA ledge analysis returns a post-hoc ledge p-value for each member of the gene set.

The NaRnEA Ledge Score for the gth gene with respect to the rth gene set in the gene expression signature $\{Z_g\}_{AB}$ is computed as follows:

$$LeS_{grAB} = \begin{cases} (1 - |AM_{gr}|)(r_{gAB}) + (AM_{gr})(r_{gAB}s_{gAB}) \; if \; PES_{rAB} \geq 0 \\ (1 - |AM_{gr}|)(r_{gAB}) - (AM_{gr})(r_{gAB}s_{gAB}) \; if \; PES_{rAB} < 0 \end{cases}$$

The statistical significance of (LeS_{grAB}) may be calculated by extending the maximum entropy null model for gene set enrichment to the level of the gth gene as follows:

$$p_{grAB} = \begin{cases} \frac{1}{G}\sum_{k=1}^{G} \mathbb{I}\{LeS_{grAB} \leq (1 - |AM_{gr}|)(r_{kAB}) + (AM_{gr})(r_{kAB}s_{kAB})\} \; if \; PES_{rAB} \geq 0 \\ \frac{1}{G}\sum_{k=1}^{G} \mathbb{I}\{LeS_{grAB} \leq (1 - |AM_{gr}|)(r_{kAB}) - (AM_{gr})(r_{kAB}s_{kAB})\} \; if \; PES_{rAB} < 0 \end{cases}$$

where $\mathbb{I}\{\cdot\}$ is the identity operator. The NaRnEA post-hoc ledge p-values may be corrected for multiple hypothesis testing to control the post-hoc ledge False Discovery Rate (FDR) according to the methodology of Benjamini and Hochberg[32]. The NaRnEA post-hoc ledge precision value for a gene set member is equal to one minus the minimum post-hoc ledge FDR at which the gene set member would be declared part of the gene set ledge.

Inferring Context-Specific Transcriptional Regulatory Networks with ARACNe3

ARACNe3 is an updated implementation of the Algorithm for the Reconstruction of Accurate Cellular Networks[13,14] which builds on the ARACNe-AP codebase[12]. The goal of ARACNe3 is to reverse-engineer a context-specific transcriptional regulatory network consisting of bivariate transcriptional regulatory interactions between a set of predefined putative transcriptional regulators and putative transcriptional targets.

ARACNe3 accepts properly normalized gene expression profiles corresponding to independent samples from a single biological phenotype. Like previous versions of the algorithm, ARACNe3 recommends that users reverse-engineer multiple estimates of the transcriptional regulatory network topology and integrate these to form an ensemble transcriptional regulatory network. However, ARACNe-AP originally recommended that the estimates

of the transcriptional regulatory network topology should be reverse-engineered in a decorrelated manner by sampling from the original set of gene expression profiles with replacement (i.e. bootstrapping). While this approach is commonly employed in the field of ensemble machine learning (e.g. random forest bagging[33]), we find that sampling gene expression profiles with replacement increases the bias and variance of the adaptive partitioning mutual information (APMI) estimator; the increase in bias occurs when sampling with replacement produces areas of locally high density due to replicated data points and the increase in variance occurs because these areas of locally high density occur stochastically across different iterations of the bootstrapping procedure. To overcome this, ARACNe3 generates decorrelated individual networks by sampling ~63.21% of the gene expression profiles without replacement each time; this is equal to the probability of a unique sample appearing in a single bootstrap and thus achieves the same level of decorrelation without unduly increasing the bias or variance of the APMI estimator.

ARACNe3 estimates the null distribution for mutual information by computing the mutual information between ~1,000,000 pairs of copula-transformed gene expression marginals which are shuffled so that they are explicitly independent. ARACNe3 then fits a piecewise null model to these values where an empirical cumulative distribution function is used for the body of the null model up to the 95th percentile of the data and the tail of the null model is fit analytically using robust linear regression applied to logarithmically transformed tail probabilities past the 95th percentile with the mblm R package[34]. The ARACNe3 piecewise null model controls the Type I error rate for the APMI estimator more accurately than the null model implemented in ARACNe-AP and allows ARACNe3 to perform the first round of individual network pruning based on control of the False Discovery Rate (FDR), resulting in a substantial gain in power over previous versions of the algorithm which recommended performing the first round of individual network pruning based on control of the Family Wise Error Rate (FWER).

The second round of ARACNe3 individual network pruning is carried out identically to previous versions of the algorithm[13]; briefly, all three-gene cliques which remain after the first round of pruning are identified and the weakest edge of the three-gene clique is removed from the network. The final set of edges which remain after both rounds of pruning constitute an individual ARACNe3-inferred network. This procedure is carried out until one of two stopping criteria are met: (1) a prespecified maximum number of individual networks have been reverse-engineered, or (2) each putative transcriptional regulator has been assigned a prespecified minimum number of unique putative transcriptional targets. The individual networks are then integrated to form an ensemble ARACNe3-inferred transcriptional regulatory network. The mutual information and Spearman correlation for each putative transcriptional regulatory interaction in the ensemble network are estimated a final time using all gene expression profiles for greater accuracy.

The ARACNe3-inferred regulon gene set for the rth transcriptional regulator is constructed by extracting all putative transcriptional regulatory interactions for the rth transcriptional regulator from the ARACNe3-inferred context-specific transcriptional regulatory network. The Association Weight values are calculated by sorting all putative target genes based on (1) the number of individual networks in which they appeared as targets of the rth transcriptional regulator and (2) the final estimated mutual information between the rth transcriptional regulator and target gene. A copula transformation is then applied so the ARACNe3-inferred regulon gene set Association Weight values are sufficiently well-balanced to meet Lindeberg's condition and guarantee the asymptotic standard normality of the NaRnEA Normalized Enrichment Score under the null hypothesis of gene set analysis. The Association Mode values are set equal to the Spearman correlation coefficient between the rth transcriptional regulator and the gene set member as estimated from all gene expression profiles.

The LUAD context-specific transcriptional regulatory network was reverse-engineered with ARACNe3 from 476 unpaired primary tumor gene expression profiles using 2,491 putative transcriptional regulators. Gene expression profiles were downloaded using the TCGAbiolinks R package from Bioconductor[35] and normalized for sequencing depth prior to network reverse-engineering (i.e. counts per million). The first round of individual network pruning was carried out with a threshold calculated to control the FDR < 0.05. Individual networks were reverse-engineered until each putative transcriptional regulator had at least 50 unique putative transcriptional targets; this was achieved after 9 iterations. The final ensemble ARACNe3-inferred transcriptional regulatory network for TCGA LUAD has 2,491 regulators, 19,351 targets, and 947,776 regulatory interactions.

The COAD context-specific transcriptional regulatory network was reverse-engineered with ARACNe3 from 437 unpaired primary tumor gene expression profiles using 2,491 putative transcriptional regulators. Gene expression profiles were downloaded using the TCGAbiolinks R package from Bioconductor and normalized for

sequencing depth prior to network reverse-engineering (i.e. counts per million). The first round of individual network pruning was carried out with a threshold calculated to control the FDR < 0.05. Individual networks were reverse-engineered until each putative transcriptional regulator had at least 50 unique putative transcriptional targets; this was achieved after 7 iterations. The final ensemble ARACNe3-inferred transcriptional regulatory network for TCGA COAD has 2,491 regulators, 19,351 targets, and 646,341 regulatory interactions.

The HNSC context-specific transcriptional regulatory network was reverse-engineered with ARACNe3 from 457 unpaired primary tumor gene expression profiles using 2,491 putative transcriptional regulators. Gene expression profiles were downloaded using the TCGAbiolinks R package from Bioconductor and normalized for sequencing depth prior to network reverse-engineering (i.e. counts per million). The first round of individual network pruning was carried out with a threshold calculated to control the FDR < 0.05. Individual networks were reverse-engineered until each putative transcriptional regulator had at least 50 unique putative transcriptional targets; this was achieved after 12 iterations. The final ensemble ARACNe3-inferred transcriptional regulatory network for TCGA HNSC has 2,491 regulators, 19,351 targets, and 810,683 regulatory interactions.

Gene Set Enrichment Analysis

Paired gene expression profiles from TCGA were normalized using a blinded DESeq2 variance stabilizing transformation prior to analysis with GSEA which was performed as described previously by (*Subramanian et al. 2005 Proc Natl Acad Sci*)[9] using the GSEA java command line tool from the Broad Institute's Molecular Signatures Database (http://www.gsea-msigdb.org/gsea/downloads.jsp). The GSEA null model was estimated using 1,000 sample shuffling permutations and gene set enrichment was computed with the following parameters:

bash GSEA_4.1.0/gsea-cli.sh GSEA -res $cur.expression.file -cls $cur.phenotype.file #Tumor_versus_Tissue -gmx $cur.gene.set.file -collapse No_Collapse -mode Max_probe -norm meandiv -nperm 1000 -permute phenotype -rnd_type no_balance -scoring_scheme weighted -rpt_label $cur.output.label -metric Signal2Noise -sort real -order descending -create_gcts false -create_svgs false -include_only_symbols false -make_sets true -median false -num 100 -plot_top_x 10 -rnd_seed $sim.seed -save_rnd_lists false -set_max 5000 -set_min 15 -zip_report false -out $cur.output.dir

analytical Rank-based Enrichment Analysis

Paired gene expression profiles from TCGA were normalized using a blinded DESeq2 variance stabilizing transformation prior to analysis with aREA. The VIPER R package was downloaded from Bioconductor at (https://www.bioconductor.org/packages/release/bioc/html/viper.html) and was used to run aREA as described previously by (*Alvarez et al. 2016 Nat Genet*)[4] with the functions *viper::rowTtest, viper::ttestNull,* and *viper::msviper* using default parameters. The aREA empirical null model was estimated using 1,000 sample shuffling permutations.

CPTAC Proteomic Data Analysis

Log-ratio normalized protein abundance data for primary tumors and adjacent normal tissues were downloaded from CPTAC for the LUAD, COAD, and HNSC cancer types from (http://linkedomics.org)[36]. Data were loaded into R and differential protein abundance between primary tumors and adjacent normal tissue was estimated for each protein with a two-tailed Mann-Whitney U test. Gene name conversion was performed using the R package biomaRt from Bioconductor[37].

Generation and Analysis of ROC Curves

All ROC curves were generated in R using the pROC R package from CRAN[38]. The *pROC::roc.test* function was used to conduct DeLong's test and compare pairs of AUROC values for correlated ROC curves. The statistical significance of the AUROC was computed using a one-tailed Mann-Whitney U test. The 95% confidence intervals for the AUROC were computed using DeLong's method. The 95% confidence intervals for ROC curves were estimated using 2,000 stratified bootstraps with the *pROC::ci.coords* function.

Plotting and Visualization

All figures created using the *ggplot2* R package from CRAN[39].

Statistical Analysis

P-values were corrected for multiple hypothesis testing to control the FDR according to the methodology of Benjamini and Hochberg or to control the FWER according to the methodology of Bonferroni[32]. 95% confidence intervals for the binomial test of proportions were computed using the procedure of Clopper and Pearson[40].

Data Availability

All gene expression profiles from TCGA are available at (https://gdc.cancer.gov/). All protein abundance profiles from CPTAC are available at (http://linkedomics.org).

Code Availability

NaRnEA is available as an R package on GitHub at (https://github.com/califano-lab/NaRnEA). Code necessary for running ARACNe3 can be downloaded along with the NaRnEA R package on GitHub at (https://github.com/califano-lab/NaRnEA). All code used to perform the analysis and generate the figures in this manuscript can be viewed as vignettes in the NaRnEA R package on GitHub.

Author Contributions

Aaron T. Griffin created NaRnEA, conceived of the study, wrote the NaRnEA R code, created ARACNe3, wrote ARACNe3 wrapper scripts in R, downloaded and analyzed TCGA transcriptomic data, downloaded and analyzed CPTAC proteomic data, performed statistical analysis, created the figures, and wrote the manuscript. Lukas J. Vlahos wrote ARACNe3 C++ code, created the NaRnEA R package, and edited the manuscript. Codruta Chiuzan verified the mathematical accuracy of the NaRnEA statistical framework, verified the computational accuracy of the NaRnEA and ARACNe3 code, and edited the manuscript. Andrea Califano supervised the research, edited the manuscript, and secured funding.

Acknowledgements

This work was supported by the National Cancer Institute (NCI) Outstanding Investigator Award R35CA197745 to Andrea Califano and the NCI Cancer Target Discovery and Development Program U01CA168426 to Andrea Califano. Aaron T. Griffin was supported, in part, by the Ruth L. Kirschstein National Research Service Award (NRSA) Institutional Research Training Grant 5T32GM007367-43, and in part by the NCI Ruth L. Kirschstein National Research Service Award (NRSA) Individual Fellowship F30CA257765. The results shown here are in part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga) and in part based on data generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing Interests

Andrea Califano is an inventor on a patent describing the VIPER algorithm (US20170076035A1) and is also a founder and equity holder of DarwinHealth, Inc., a company that has licensed the VIPER algorithm from Columbia University. Columbia University is also an equity holder in DarwinHealth, Inc.

Correspondence and requests for materials should be addressed to: Andrea Califano (ac2248@cumc.columbia.edu).

References

1. Marx, V. Biology: The big challenges of big data. *Nature* **498**, 255-260 (2013).
2. Ashburner, M.*, et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
3. Croft, D.*, et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477 (2014).
4. Alvarez, M.J.*, et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**, 838-847 (2016).
5. Mangul, S.*, et al.* Systematic benchmarking of omics computational tools. *Nat Commun* **10**, 1393 (2019).
6. Maleki, F., Ovens, K., Hogan, D.J. & Kusalik, A.J. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front Genet* **11**, 654 (2020).
7. Das, S., McClain, C.J. & Rai, S.N. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy (Basel)* **22**(2020).
8. Jaynes, E.T. & Bretthorst, G.L. *Probability Theory : The Logic of Science*, (Cambridge University Press, Cambridge, UK, 2003).
9. Subramanian, A.*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
10. Broyde, J.*, et al.* Oncoprotein-specific molecular interaction maps (SigMaps) for cancer network analyses. *Nat Biotechnol* **39**, 215-224 (2021).
11. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F.M. Gene regulatory network inference resources: A practical overview. *Biochim Biophys Acta Gene Regul Mech* **1863**, 194430 (2020).
12. Lachmann, A., Giorgi, F.M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233-2235 (2016).
13. Margolin, A.A.*, et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).
14. Margolin, A.A.*, et al.* Reverse engineering cellular networks. *Nat Protoc* **1**, 662-671 (2006).
15. Alvarez, M.J.*, et al.* A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. *Nat Genet* **50**, 979-989 (2018).
16. Califano, A. & Alvarez, M.J. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat Rev Cancer* **17**, 116-130 (2017).
17. Paull, E.O.*, et al.* A modular master regulator landscape controls cancer transcriptional identity. *Cell* **184**, 334-351 e320 (2021).
18. Obradovic, A.*, et al.* Single-cell protein activity analysis identifies recurrence-associated renal tumor macrophages. *Cell* **184**, 2988-3005 e2916 (2021).
19. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
20. Tamayo, P., Steinhardt, G., Liberzon, A. & Mesirov, J.P. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat Methods Med Res* **25**, 472-487 (2016).
21. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845 (1988).
22. Cui, X. & Churchill, G.A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* **4**, 210 (2003).
23. Bishop, D. Rein in the four horsemen of irreproducibility. *Nature* **568**, 435 (2019).

24. de Leeuw, C.A., Neale, B.M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat Rev Genet* **17**, 353-364 (2016).

25. Shen, Y.*, et al.* Systematic, network-based characterization of therapeutic target inhibitors. *PLoS Comput Biol* **13**, e1005599 (2017).

26. Ding, H.*, et al.* Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm. *Nat Commun* **9**, 1471 (2018).

27. Chari, A.*, et al.* Oral Selinexor-Dexamethasone for Triple-Class Refractory Multiple Myeloma. *N Engl J Med* **381**, 727-738 (2019).

28. Shannon, C.E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379-423 (1948).

29. Lindeberg, J.W. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **15**, 211-225 (1922).

30. Gut, A. The Multivariate Normal Distribution. in *An Intermediate Course in Probability* (ed. Gut, A.) 117-145 (Springer New York, New York, NY, 2009).

31. DiCiccio, T.J. & Efron, B. Bootstrap Confidence Intervals. *Statistical Science* **11**, 189-212 (1996).

32. Benjamini, Y. & Hochberg, Y. CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **57**, 289-300 (1995).

33. Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001).

34. Lukasz, K. mblm: Median-Based Linear Models. (2019).

35. Colaprico, A.*, et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).

36. Vasaikar, S.V., Straub, P., Wang, J. & Zhang, B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res* **46**, D956-D963 (2018).

37. Durinck, S., Spellman, P.T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191 (2009).

38. Robin, X.*, et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

39. Wickham, H. ggplot2 : Elegant Graphics for Data Analysis. in *Use R!,* 1 online resource (XVI, 260 pages 232 illustrations, 140 illustrations in color (Springer International Publishing : Imprint: Springer,, Cham, 2016).

40. Clopper, C.J. & Pearson, E.S. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* **26**, 404-413 (1934).

Table Legends

Table 1. Comparative Analysis of NaRnEA, GSEA, and aREA. Comparing NaRnEA, GSEA, and aREA gene set enrichment two-sided p-values without correction (FPR < 0.05), after correcting to control the False Discovery Rate (FDR < 0.05), and after correcting to control the Family Wise Error Rate (FWER < 0.05). P-values were corrected for multiple hypothesis testing to control the FDR according to the methodology of Benjamini and Hochberg. P-values were corrected for multiple hypothesis testing to control the FWER according to the methodology of Bonferroni. Maximum likelihood estimates of the proportion of significantly enriched gene sets are shown as percentages along with 95% confidence intervals for the proportion which are calculated according to the procedure of Clopper and Pearson.

Table 2. Consistency of NaRnEA-based Differential Protein Activity from TCGA with Mass Spectrometry-based Differential Protein Abundance from CPTAC. The point estimate for Kendall's Tau-B correlation coefficient is reported along with the 95% confidence interval. Expected values for each cell are calculated under the null hypothesis of marginal independence. Z-scores are calculated by applying an inverse standard normal transformation to the quantile computed from the observed and expected values in each cell using a hypergeometric null distribution. Cells for which the z-score is significantly greater than expected under the null hypothesis (FWER < 0.05) are marked in red and cells for which the z-score is significantly less than expected under the null hypothesis (FWER < 0.05) are marked in blue.

Figure Legends

Figure 1. Comparative Benchmark of NaRnEA, GSEA, and aREA for Measurement of Differential Protein Activity. NaRnEA significantly outperforms GSEA and aREA for the purpose of distinguishing between TS-regulons and NM-regulons in TCGA transcriptomic data from (a) LUAD, (b) COAD, and (c) HNSC cancer cohorts. ROC curves are constructed from gene set enrichment two-sided p-values by setting ARACNe3-inferred regulon gene sets as cases and null gene sets as controls. 95% confidence intervals for ROC curves are estimated using 2,000 stratified bootstraps. 95% confidence intervals for the AUROC are calculated using DeLong's method. The statistical significance of the AUROC is calculated with a one-tailed Mann-Whitney U (MWU) test. Within each cancer type, AUROC values for each method are compared in a pairwise manner using DeLong's test. NaRnEA ROC curves are shown in red, GSEA ROC curves are shown in blue, and aREA ROC curves are shown in green.

Figure 2. Visualizing NaRnEA-based Measurement of Differential Protein Activity in TCGA Cohorts. Visualizing NaRnEA-based differential protein activity in TCGA transcriptomic data from the (a) LUAD, (b) COAD, and (c) HNSC cancer subtypes. The absolute value of the NaRnEA Normalized Enrichment Score (NES) is plotted vs. the NaRnEA Proportional Enrichment Score (PES) to enable simultaneous inspection of the statistical significance and effect size of the enrichment, respectively. NaRnEA two-sided p-values are corrected for multiple hypothesis testing according to the methodology of Benjamini and Hochberg to control the False Discovery Rate (FDR < 0.05). The proportion of deactivated (DOWN) and activated (UP) transcriptional regulatory proteins in each of the TCGA datasets is displayed in blue and red, respectively, along with the fraction and total number of regulons which were analyzed.

Table 1. Comparative Analysis of NaRnEA, GSEA, and aREA.

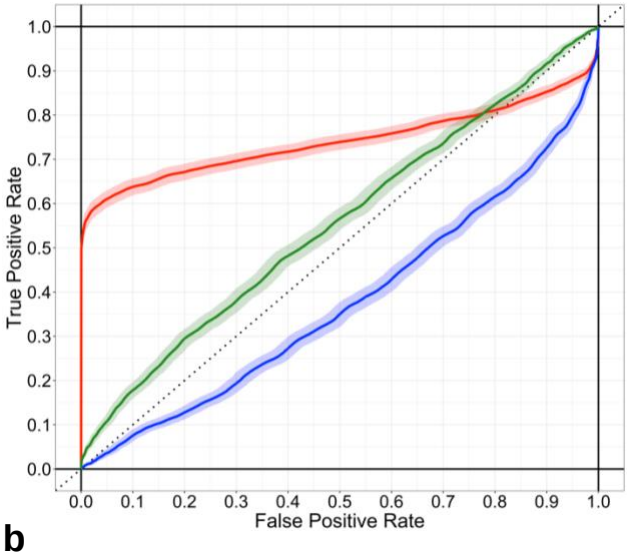
Subtype	Gene Sets	Threshold	NaRnEA [95% CI]	GSEA [95% CI]	VIPER [95% CI]
LUAD	ARACNe3	FPR < 0.05	60.98% [59.03%, 62.90%]	7.47% [6.47%, 8.57%]	9.39% [8.28%, 10.61%]
		FDR < 0.05	59.37% [57.41%, 61.31%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
		FWER < 0.05	44.92% [42.96%, 46.90%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
	Null	FPR < 0.05	4.94% [4.12%, 5.86%]	9.80% [8.66%, 11.03%]	4.05% [3.31%, 4.91%]
		FDR < 0.05	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
		FWER < 0.05	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
COAD	ARACNe3	FPR < 0.05	53.31% [51.33%, 55.29%]	6.14% [5.23%, 7.16%]	4.34% [3.57%, 5.21%]
		FDR < 0.05	50.82% [48.84%, 52.80%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
		FWER < 0.05	32.52% [30.68%, 34.40%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
	Null	FPR < 0.05	5.06% [4.23%, 5.99%]	7.91% [6.88%, 9.04%]	4.09% [3.35%, 4.95%]
		FDR < 0.05	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
		FWER < 0.05	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
HNSC	ARACNe3	FPR < 0.05	55.28% [53.30%, 57.24%]	5.78% [4.90%, 6.77%]	6.54% [5.60%, 7.59%]
		FDR < 0.05	52.47% [50.49%, 54.45%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
		FWER < 0.05	33.80% [31.94%, 35.70%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
	Null	FPR < 0.05	5.10% [4.27%, 6.04%]	2.17% [1.63%, 2.82%]	4.58% [3.79%, 5.47%]
		FDR < 0.05	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]
		FWER < 0.05	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]	0.00% [0.00%, 0.15%]

Table 2. Consistency of NaRnEA-based Differential Protein Activity from TCGA with Mass Spectrometry-based Differential Protein Abundance from CPTAC.

TCGA LUAD vs. CPTAC LUAD Kendall's Tau B = 0.382 [0.338, 0.427] (Chi-Squared Test p = 3.14e-51)			
	CPTAC DOWN	**CPTAC NS**	**CPTAC UP**
TCGA UP	Observed = 79 Expected = 154.70 Z-score = -9.662	Observed = 88 Expected = 107.15 Z-score = -2.616	Observed = 281 Expected = 186.15 Z-score = 11.500
TCGA NS	Observed = 136 Expected = 156.77 Z-score = -2.533	Observe = 134 Expected = 108.59 Z-score = 3.565	Observed = 184 Expected = 188.64 Z-score = -0.496
TCGA DOWN	Observed = 208 Expected = 111.53 Z-score = 13.018	Observed = 71 Expected = 77.26 Z-score = -0.874	Observed = 44 Expected = 134.21 Z-score = -12.424
TCGA COAD vs. CPTAC COAD Kendall's Tau B = 0.281 [0.223, 0.339] (Chi-Squared Test p = 1.17e-17)			
	CPTAC DOWN	**CPTAC NS**	**CPTAC UP**
TCGA UP	Observed = 104 Expected = 143.94 Z-score = -5.801	Observed = 42 Expected = 53.19 Z-score = -2.084	Observed = 163 Expected = 111.87 Z-score = 7.814
TCGA NS	Observed = 151 Expected = 149.06 Z-score = 0.354	Observe = 65 Expected = 55.09 Z-score = 1.990	Observed = 104 Expected = 115.85 Z-score = -1.714
TCGA DOWN	Observed = 113 Expected = 75.00 Z-score = 6.870	Observed = 29 Expected = 27.72 Z-score = 0.430	Observed = 19 Expected = 58.29 Z-score = -7.613
TCGA HNSC vs. CPTAC HNSC Kendall's Tau B = 0.342 [0.293, 0.391] (Chi-Squared Test p = 3.63e-37)			
	CPTAC DOWN	**CPTAC NS**	**CPTAC UP**
TCGA UP	Observed = 42 Expected = 96.38 Z-score = -8.305	Observed = 102 Expected = 126.99 Z-score = -3.335	Observed = 271 Expected = 191.63 Z-score = 10.029
TCGA NS	Observed = 95 Expected = 102.65 Z-score = -1.042	Observed = 164 Expected = 135.26 Z-score = 3.889	Observed = 183 Expected = 204.09 Z-score = -2.545
TCGA DOWN	Observed = 118 Expected = 55.97 Z-score = 10.233	Observed = 70 Expected = 73.75 Z-score = -0.510	Observed = 53 Expected = 111.28 Z-score = -8.680

Figure 1. Comparative Benchmark of NaRnEA, GSEA, and aREA for Measurement of Differential Protein Activity.

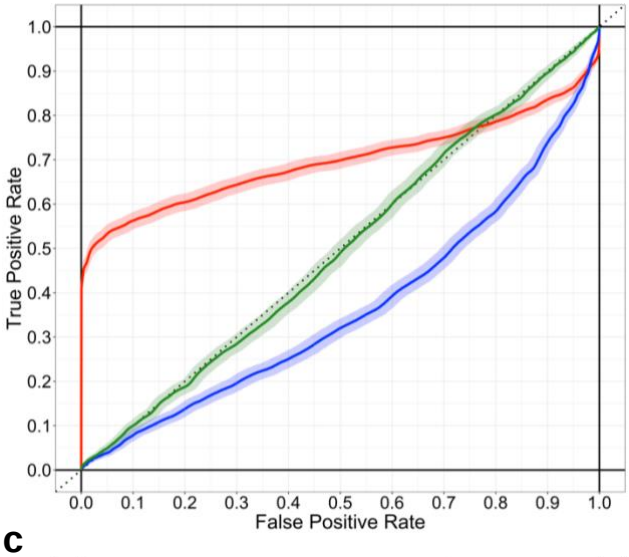
a



LUAD AUROC (DeLong and MWU)	
NaRnEA	0.741 [0.726, 0.756] (p = 4.13e-191)
GSEA	0.376 [0.361, 0.391] (p = 1.000)
VIPER	0.555 [0.539, 0.571] (p = 1.067e-11)

LUAD AUROC Comparison (DeLong)	
NaRnEA vs. GSEA	p = 1.24e-359
NaRnEA vs. VIPER	p = 8.53e-199
GSEA vs. VIPER	p = 4.13e-93

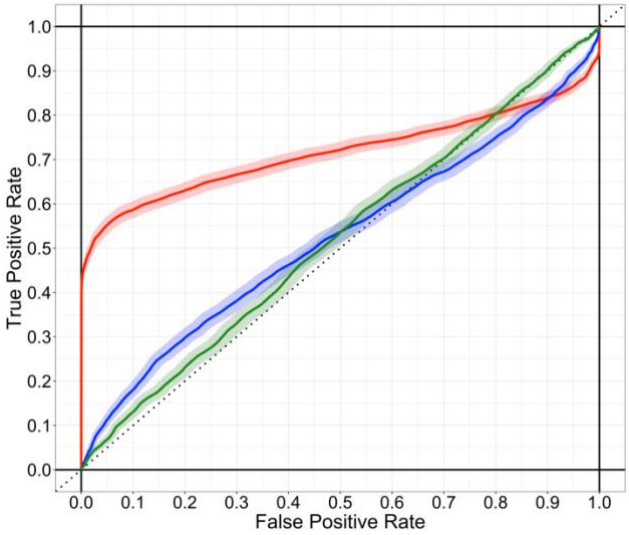
b



COAD AUROC (DeLong and MWU)	
NaRnEA	0.698 [0.682, 0.714] (p = 2.048e-129)
GSEA	0.364 [0.349, 0.379] (p = 1.000)
VIPER	0.496 [0.480, 0.512] (p = 0.704)

COAD AUROC Comparison (DeLong)	
NaRnEA vs. GSEA	p = 1.97e-262
NaRnEA vs. VIPER	p = 2.12e-207
GSEA vs. VIPER	p = 2.27e-46

c



HNSC AUROC (DeLong and MWU)	
NaRnEA	0.716 [0.701, 0.732] (p = 4.424e-154)
GSEA	0.523 [0.507, 0.539] (p = 2.753e-3)
VIPER	0.521 [0.505, 0.537] (p = 5.707e-3)

HNSC AUROC Comparison (DeLong)	
NaRnEA vs. GSEA	p = 3.85e-106
NaRnEA vs. VIPER	p = 7.60e-206
GSEA vs. VIPER	p = 0.813

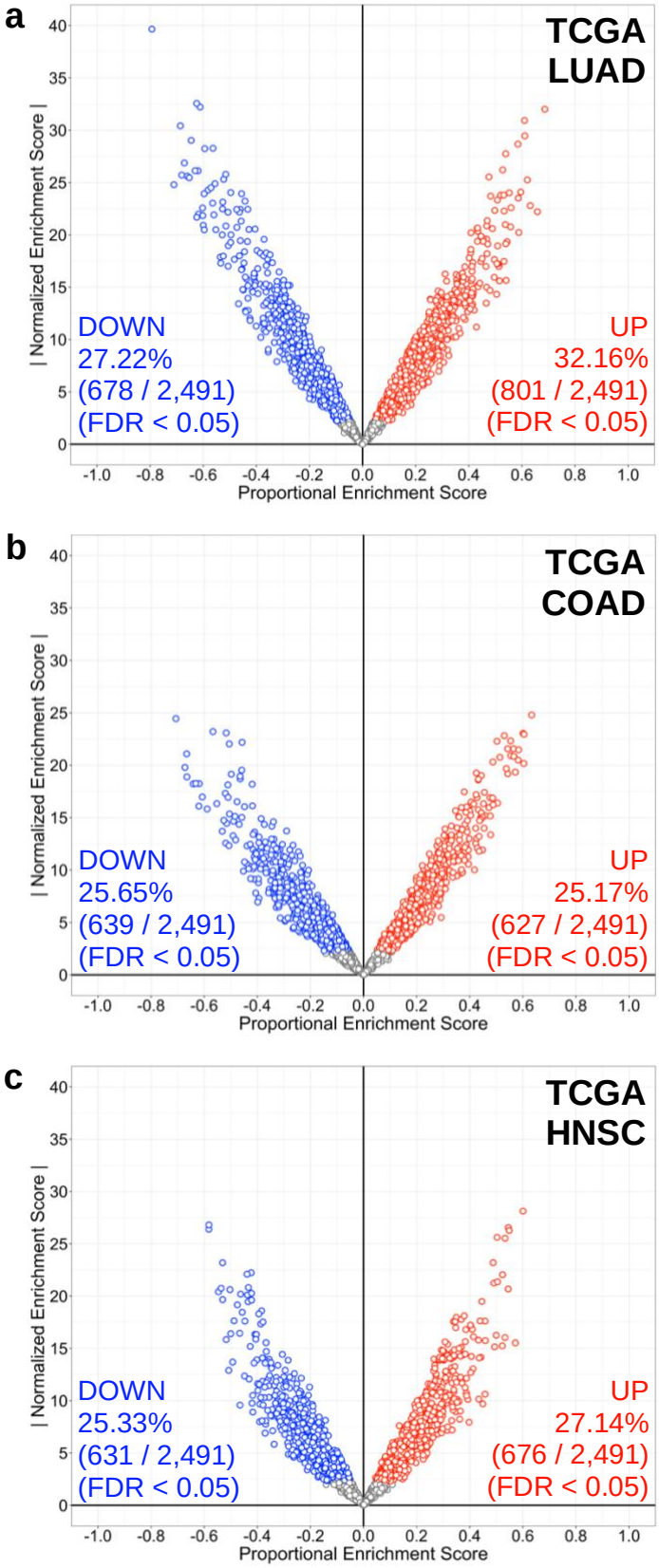



Figure 2. Visualizing NaRnEA-based Measurement of Differential Protein Activity in TCGA Cohorts.