An interpretable, finite sample valid alternative to Pearson's X^2 for scientific discovery

Tavor Z. Baharav[a], **David Tse**[a], **and Julia Salzman**[b,c,d,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry, Stanford University, Stanford, 94305; [d]Department of Statistics (by courtesy), Stanford University, Stanford, 94305

Contingency tables, data represented as counts matrices, are ubiquitous across quantitative research and data-science applications. Existing statistical tests are insufficient however, as none are simultaneously computationally efficient and statistically valid for a finite number of observations. In this work, motivated by a recent application in reference-genome-free inference (1), we develop OASIS (Optimized Adaptive Statistic for Inferring Structure), a family of statistical tests for contingency tables. OASIS constructs a test-statistic which is linear in the normalized data matrix, providing closed form p-value bounds through classical concentration inequalities. In the process, OASIS provides a decomposition of the table, lending interpretability to its rejection of the null. We derive the asymptotic distribution of the OASIS test statistic, showing that these finite-sample bounds correctly characterize the p-value bound derived up to a variance term. Experiments on genomic sequencing data highlight the power and interpretability of OASIS. The same method based on OASIS significance calls detects SARS-CoV-2 and Mycobacterium Tuberculosis strains de novo, which cannot be achieved with current approaches. We demonstrate in simulations that OASIS is robust to overdispersion, a common feature in genomic data like single cell RNA-sequencing, where under accepted noise models OASIS still provides good control of the false discovery rate, while Pearson's X^2 test consistently rejects the null. Additionally, we show on synthetic data that OASIS is more powerful than Pearson's X^2 test in certain regimes, including for some important two group alternatives, which we corroborate with approximate power calculations.

Discrete data on contingency tables is ubiquitous in classical and modern data science, and is central in the social sciences and quantitative research disciplines, including biology. In modern applications these tables are frequently large and sparse, leading to a continued interest in new statistical tests for contingency tables (2). One recent motivating work is NOMAD, a new approach to genomics which maps myriad problems in genomic sequence analysis to the analysis of contingency tables. These disparate applications include discriminating phylogenetically distinct strains or alternative splicing in single cell RNA-sequencing, among others (1).

There is a rich literature that addresses testing for row and column independence in contingency tables beginning with the work of Pearson, who designed the widely used X^2 test in the early 1900s (3, 4). Other approaches include the likelihood ratio test which is asymptotically equivalent to Pearson's X^2 test under the null (4)[Section 16.3.4], permutation-based tests (e.g. MCMC-based) (5), and modeling parametric deviations from the null with log-linear models (4).

Despite the prominence of Pearson's X^2 test, it suffers from multiple statistical drawbacks which limit its utility for scientific inference. First, the X^2 test lacks robustness: it has high power against many scientifically uninteresting alternatives, for example against models where technical or biological noise cause a table to formally deviate from the specified null. We expand on this point later, and provide simulation evidence for noise stemming from biological overdispersion and contamination.

Second, the X^2 test does not provide statistically valid p-values for a finite number of observations. There is a long literature on estimating significance thresholds, centered on an asymptotic theory that assumes a fixed table size with the number of observations tending to infinity. For example, common guidelines (4) indicate that the χ^2 distribution is a bad approximation when more than 20% of the entries take value less than 5. However, in modern tables of interest, this is often the case; the biological tables which motivated this test's design have many (tens or hundreds) rows relative to the total number of observations per sample (similarly in the tens or hundreds), violating the guideline for the use of X^2.

Other methods such as log-linear models suffer from similar limitations: namely, lack of robustness, and of calibration for finite observations (4). Methods involving Monte Carlo simulations have also been studied, which provide statistically valid p-values, but despite significant recent works the lack of a closed form expression renders them too computationally intensive for large tables in practice (2, 5). Additionally, for permutation tests, B samples from the null can at best yield a p-value of 1/(B+1). Thus, many permutations are required to obtain sufficiently small p-values to reject the null, especially under the burden of multiple hypothesis correction. To our knowledge, no nontrivial p-values exist for general contingency tables that are computable in closed form and are valid for a finite number of observations.

In this work, motivated by recent applications in genomic inference (1), we introduce the OASIS framework, developing a powerful and general family of interpretable tests. Building on preliminary work (1), OASIS provides p-values that are 1) valid for a finite number of observations, 2) have a closed form expression, 3) are empirically robust to small deviations from the null distribution, and 4) empirically enable scientific inference that cannot be achieved with X^2.

To build intuition for a task we seek OASIS to perform, consider the following data. 103 patients are infected with potentially different variants of SARS-CoV-2 (6). For each, a black box produces counts of the nucleotide composition of a segment of the SARS-CoV-2 variant's genome, which can be

T.Z.B. designed, developed, analyzed, and tested the statistic. D.T. developed the statistic. J.S. designed and developed the statistic, and supervised the project. All authors wrote the manuscript. Acknowledgements: Persi Diaconis, Sourav Chatterjee, and Allison Carey for discussion. Punit Sundar ran NOMAD to generate calls for M. Tuberculosis.

J.S. and T.Z.B. have submitted a provisional patent No. 63/366,444 relating to this work. The authors declare no other competing interest.

[1]To whom correspondence should be addressed. E-mail: julia.salzman@stanford.edu

considered a categorical variable. This problem formulation and motivation stems from (1), and is discussed further in Section 6. Under the null hypothesis, each patient is infected with the same variant of the virus, and so rows in the table will be drawn from the the same distribution. If a new strain has infected a group of patients, then the row distribution for these patients will be different. The desired test should detect this, and provide post-inferential guidance on why the table was rejected, ideally discriminating populations of patients on the basis of which strains infected them.

To understand why a new test is needed, we consider Pearson's X^2 in more detail. The X^2 test statistic sums squared residuals after appropriate normalization (Figure 1B, Eq. (3)) The resulting sum is asymptotically χ^2 distributed under the null, but can significantly deviate from this for a finite number of observations. When it rejects the null, Pearson's X^2 test provides no framework for interpreting why, leading practitioners to develop and use exploratory data-analysis tools such as correspondence analysis (7–10).

OASIS seeks to improve these shortcomings by building in interpretability and analytic tractability in its construction of a linear test statistic (Figure 1, Eq. (1)). This construction enables the use of classical concentration inequalities for sums of bounded random variables, like Hoeffding's inequality, yielding p-value bounds that are valid for finite numbers of observations. As opposed to X^2 which sums squared residuals, OASIS computes a bilinear form of residuals, similar in spirit but conceptually and statistically distinct from a Lancaster decomposition of X^2 (11) and related polynomial decomposition methods (12). Fluctuations or single large residuals will be averaged, and are unlikely to generate a large test statistic. We make this observation precise later in our discussion of the linear algebraic characterizations of these approaches. The closest approaches to OASIS regarding interpretable decompositions of contingency tables involve correspondence analysis, (7–10), an exploratory method for post-facto interpretation that provides no statistical guarantees.

In the rest of this paper, we formalize OASIS, state several of its theoretical properties, contrast it with X^2, and present several variants and extensions of the OASIS test. Here, we do not seek to expound on the full theoretical generality of OASIS, but instead provide an applied exposition and disciplined framework for computing some optimization-based instantiations of the statistic, and illustrate the applied performance of OASIS in simulations and in real biological data. Simulations show that OASIS is a robust test that controls the level in a variety of models where the null is formally violated, but without a biological or scientific meaning. Simulated alternatives show that OASIS has power in many settings of interest, and in fact has more power than X^2 in a variety of settings including for two group alternatives. Biological examples show that OASIS provides scientific inference currently impossible with Pearson's X^2 test across applied genomics problems. For example, OASIS has 100% accuracy in distinguishing discrete patient populations infected with Omicron-BA.1 and BA.2 from those infected with the Delta variant without knowledge of a reference genome or any sample metadata, a task that to our knowledge was previously impossible. In a different biological domain, analyzing M. Tuberculosis sequencing data, OASIS precisely partitions samples from two sub-sublineages, again with no metadata or reference-genome. Both datasets were chosen arbitrarily from the sequence read archive (13) and were analyzed with the same method with the same parameters. Finally, the ideas in OASIS enable more general tests and analyses which we discuss in the conclusion. A glimpse at such examples include a disciplined alternative statistical framework to the singular value decomposition to decompose matrices, and inference on multiple tables defined on the same set of columns.

1. Problem Formulation

As is standard in contingency table analysis, the observed matrix of counts is taken as $X \in \mathbb{N}^{I \times J}$. Defining $[m] = \{1, 2, \ldots, m\}$, a contingency table is then defined by pairs of observations of a row ($[I]$-valued) categorical random variable, and a column ($[J]$-valued) categorical random variable. In this setting, we focus on the case where the columns correspond to samples (explanatory random variable) and the rows correspond to the response variable (4). Thus, we are interested in whether the conditional row distribution is the same for each column. Define $X^{(j)}$ as the j-th column of X, and $n_j = \sum_{i=1}^{I} X_{ij}$ as the total number of counts in column j. The null model studied is:

Definition 1 (Null Model)**.** *Conditional on the column totals $\{n_j\}_{j=1}^{J}$, each column of the contingency table X is $X^{(j)} \sim$ multinomial$(n_j, \boldsymbol{p})$, drawn independently for all j, for some common vector of (unknown) row probabilities $\boldsymbol{p}$.*

Contingency tables are well studied, and we refer the reader to (4) for further background. A classical example table studies the relationship between aspirin use and heart attacks, where the columns correspond to aspirin or placebo use, and the rows correspond to Fatal Attack, Nonfatal Attack, or No Attack. It is found that conditioning on whether the subject takes aspirin or not yields a statistically significant difference in outcome.

For our analysis, we define $M = \sum_j n_j$ as the total number of observations in the table. $\|\cdot\|$ denotes the vector ℓ_2 norm or spectral norm for matrices, unless otherwise specified. When applied to a vector, $\sqrt{\cdot}$ denotes the entrywise square root. Finally, for a random variable we use $\xrightarrow{D}$ to denote convergence in distribution.

2. OASIS test statistic

The OASIS test statistic has a natural linear algebraic construction and interpretation. Taking as input $\boldsymbol{f} \in \mathbb{R}^I$ and $\boldsymbol{c} \in \mathbb{R}^J$, and defining the expected matrix E and the centered and normalized table $\tilde{X}$, the test statistic S is computed as:

$$
\begin{aligned}
E &= \frac{1}{M} X \mathbf{1}\mathbf{1}^\top X \\
\tilde{X} &= (X - E)\,\mathrm{diag}\left(1/\sqrt{X^\top \mathbf{1}}\right) \\
S &= S(\boldsymbol{f}, \boldsymbol{c}) = \boldsymbol{f}^\top \tilde{X} \boldsymbol{c}
\end{aligned}
\qquad [1]
$$

$\tilde{X}$ is normalized so that under the null, the variance of $\boldsymbol{f}^\top \tilde{X}^{(j)}$ is constant across all j up to a small term involving the correlation between $X^{(j)}$ and E. This linearity enables the construction of a finite-sample valid p-value bound for fixed $\boldsymbol{f}, \boldsymbol{c}$, which operate on the rows and columns of the table respectively. In preliminary work these vectors were chosen randomly (1), leading to a specific case of the OASIS test we call OASIS-rand, albeit with a looser analysis and worse p-value bound. Other

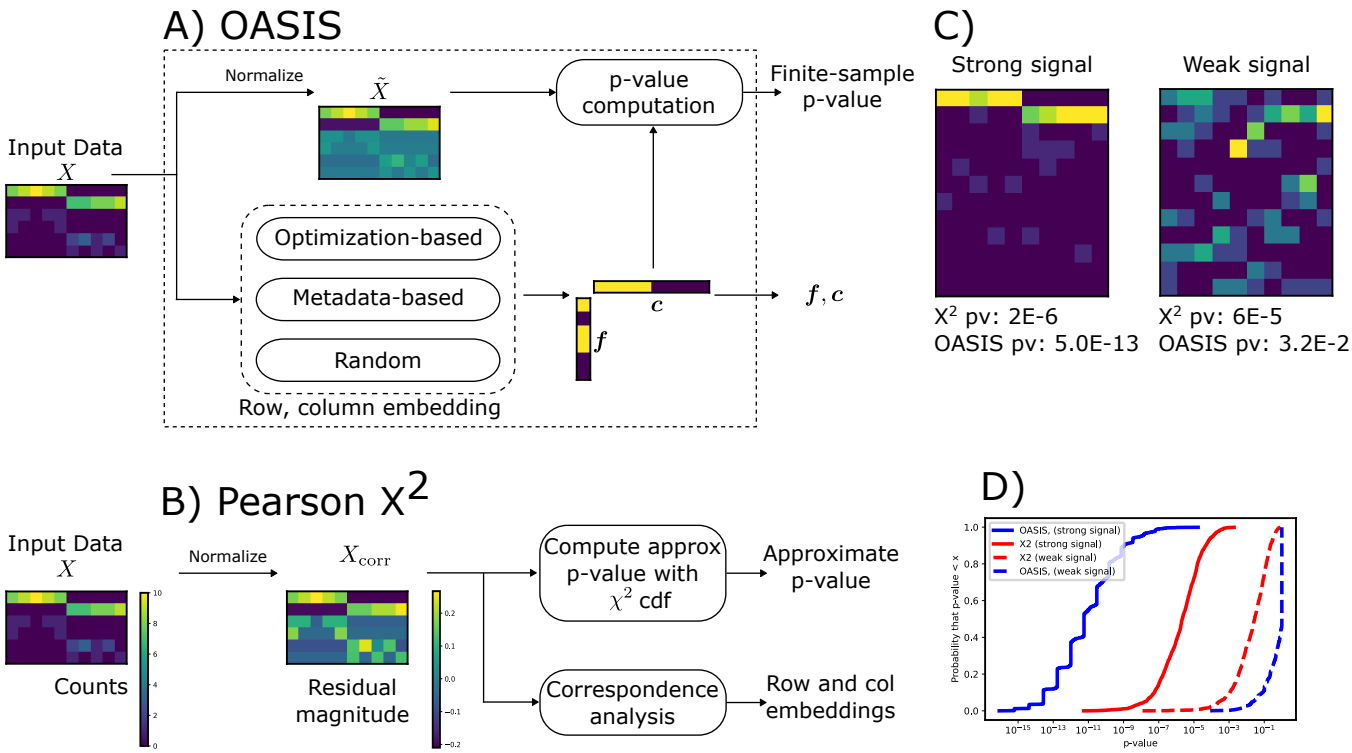

Fig. 1. Comparison of OASIS and Pearson's X^2 test for input matrix $X \in \mathbb{N}^{I \times J}$. **A)** OASIS computes a matrix of residuals $\tilde{X}$ as in Eq. (1). Row (column) score vectors $\boldsymbol{f} \in \mathbb{R}^I$ ($\boldsymbol{c} \in \mathbb{R}^J$) are generated by one of several options (metadata can be incorporated if desired, e.g. an ordering of the rows, or similarity information regarding the columns), which map each row (column) to a scalar value. These vectors are used to compute the test statistic S in Eq. (1), and finite-sample p-value bound using concentration inequalities. **B)** X^2 computes a matrix of residuals X_{corr} as in Eq. (2), which is more sensitive to deviations in low count rows (assigning a larger magnitude residual to bottom 4 rows in example X). It then computes the X^2 statistic as in Eq. (3), and provides an asymptotically valid p-value via a distributional approximation. For interpretability, practitioners often use correspondence analysis (4) to interpret rejection of the null, a procedure with no statistical guarantees, which can fail to detect desired structure. **C)** depicts two example counts tables the one on the left corresponding to concentrated signal, with the one on the left corresponding to diffuse signal. Both tables have 100 counts distributed evenly over 10 columns, with 12 available rows. The generative model for these is detailed in Section 14.A, with additional plots provided. X^2 assigns both of them similar significance, but OASIS (`OASIS-opt` detailed in Section 4) assigns a much smaller p-value to the left than the right, as is visually clear. **D)** plots the empirical CDF of the p-values of this variant of OASIS and Pearson's X^2. This is shown for the two classes of tables; one with a strong concentrated two-group signal, and one with a diffuse signal. OASIS yields significantly improved p-values for the case with strong signal, and substantially worse power than X^2 in the weak signal case, which visually looks like noise. X^2 on the other hand yields much more similar performance in the two settings. Here, `OASIS-opt` is shown, which is run over 5 independent splits of the dataset (additional details in Section 14.A).

approaches to selecting $\boldsymbol{f}, \boldsymbol{c}$ based on experimental design or data-adaptive optimization are also possible; we discuss and develop a class of optimization-based approaches in Section 4.

A. Comparison with Pearson's X^2 test statistic. The linear algebraic formulations of Pearson's X^2 statistic and correspondence analysis highlight their fundamental differences from OASIS. X^2 is computed as:

$$X_{\text{corr}} = \text{diag}\left(1/\sqrt{X\mathbf{1}}\right) \underbrace{(X - E)\,\text{diag}\left(1/\sqrt{X^\top \mathbf{1}}\right)}_{\tilde{X}} \qquad [2]$$

$$X^2 = \sum_{i,j} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} = M\left\|X_{\text{corr}}\right\|_F^2. \qquad [3]$$

This can be compared with OASIS, where the p-value bound depends on S^2 as defined in Eq. (1). This highlights two main differences between X^2 and OASIS. First, the X^2 statistic left normalizes using row frequencies, making finite-sample analysis of the test-statistic difficult. OASIS treats rows and columns asymmetrically and only normalizes by column frequencies, as these are given by the model. Second, the X^2 statistic squares each residual, and then sums these quantities. OASIS computes a bilinear form of the residual matrix with $\boldsymbol{f}$ and $\boldsymbol{c}$, squaring the resulting sum (as S^2 is used in the bounds). This intuitively allows residuals resulting from unstructured deviations to average out, focusing the power of the test on structured deviations from the null.

Since Pearson's X^2 test does not provide guidance for the the statistical reason the null is rejected, practitioners commonly employ correspondence analysis for this task (8). Correspondence analysis studies the matrix of standardized residuals X_{corr} defined in Eq. (2). This method computes the singular value decomposition of X_{corr}, and projects rows and columns along the first few principal vectors to obtain low dimensional embeddings for both the rows and the columns. As we show, OASIS provides a statistically grounded and similarly interpretable alternative to this approach, by analyzing $\tilde{X}$.

Since Pearson's X^2 test approximates X^2 as being χ^2-distributed with $(I-1)(J-1)$ degrees of freedom under the null of row and column independence, its power decays with the number of rows. This yields several important alternative hypothesis regimes where OASIS is predicted to have higher power than X^2, such as time series or 2-group alternatives when the total number of counts M is small relative to $I \times J$, i.e. the table is sparse (details in Section 6.A.1).

3. Analysis of OASIS statistic

Due to the bilinear form of the OASIS test statistic we are able to provide theoretical guarantees regarding its concentration properties, both for finite sample p-value bounds and asymptotically. OASIS's design provides interpretation for rejection of the null through $\boldsymbol{f}, \boldsymbol{c}$, and allows for the construction of a measure of effect size, which quantifies the magnitude of deviation from the null, deconfounding the total number of observations M.

A. p-value bound. A preliminary version of OASIS was designed so that p-value bounds could easily be obtained by applying classical concentration inequalities to the test statistic (1); here we improve these bounds, derive asymptotic ones, and show our newly derived finite-sample bounds have a matching form with those obtained from the asymptotic characterization. For notational simplicity we define the quantity γ which measures the similarity between $\boldsymbol{c}$ and $\boldsymbol{n} = X^\top \mathbf{1}$ the column counts

$$\gamma = \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|}, \sqrt{\frac{\boldsymbol{n}}{M}} \right\rangle^2, \qquad [4]$$

noting that $\boldsymbol{n}$ is a function of the observed table X, and so when necessary we write this quantity as γ_X. Observe that $\gamma \in [0,1]$ by Cauchy-Schwarz. While the p-value bound can be computed for any $\boldsymbol{f}, \boldsymbol{c}$, we provide a constrained variant below for simplicity.

Proposition 1. *Under the null hypothesis, for any fixed $\boldsymbol{f} \in [0,1]^I$ and $\boldsymbol{c} \in \mathbb{R}^J$ with $\|\boldsymbol{c}\|_2 \leq 1$, if $\gamma < 1$, the OASIS test statistic $S = S(\boldsymbol{f}, \boldsymbol{c})$ satisfies*

$$\mathbb{P}\left(|S| \geq s\right) \leq 2 \exp\left(-\frac{2s^2}{1-\gamma}\right).$$

We provide an unconstrained version of this bound along with the proof details in Section 10.B, where we additionally note that for $\gamma = 1$, $S = 0$ with probability 1. As intuition predicts, since the OASIS test statistic is the sum of independent increments, it converges in distribution to a Gaussian as the number of observations goes to infinity, as long as $\gamma \neq 1$.

Proposition 2. *Define $\sigma_{\boldsymbol{f}}^2 = \mathrm{Var}_{Z\sim\boldsymbol{p}}(f_Z) = \sum_i p_i f_i^2 - \left(\sum_i p_i f_i\right)^2$, where $\boldsymbol{p}$ is the common row distribution under the null, and consider fixed $\boldsymbol{f}, \boldsymbol{c}$. For a sequence of tables $\{X_t\}_{t=1}^\infty$ where each column is an independent draw from a multinomial distribution with probability vector $\boldsymbol{p}$, with $\min X_t^\top \mathbf{1} \to \infty$ and $\gamma_{X_t} < 1$ for all t, the OASIS test statistic $S_t = S(X_t, \boldsymbol{f}, \boldsymbol{c})$ satisfies*

$$\frac{1}{\sqrt{1-\gamma_{X_t}}} S_t \xrightarrow{D} \mathcal{N}\left(0, \sigma_{\boldsymbol{f}}^2 \|\boldsymbol{c}\|^2\right).$$

The intuition for this result is that each entry of $\tilde{X}^\top \boldsymbol{f}$ is asymptotically distributed as $\mathcal{N}(0, \sigma_{\boldsymbol{f}}^2)$, up to the negative correlation stemming from the unknown $\mu = \mathbb{E}_{Z\sim\boldsymbol{p}}[f_Z]$. Then, S is a linear combination of $\boldsymbol{f}^\top \tilde{X}$ with weights $\boldsymbol{c}$, and so S has variance essentially $\sigma_{\boldsymbol{f}}^2 \|\boldsymbol{c}\|^2$, up to the $(1-\gamma)$ scaling. As a corollary, an asymptotically valid p-value can be constructed using the sample variance $\hat{\sigma}_{\boldsymbol{f}}^2$. Note that the finite sample bound in Proposition 1 takes the form of a Gaussian tail bound, except that it upper bounds the variance of f_Z, a $[0,1]$-valued random variable, by $\frac{1}{4}$.

B. Effect Size. Through $\boldsymbol{f}, \boldsymbol{c}$, OASIS assigns scalar values to each row and column. For intuition, the original version of OASIS (1) assigned each sample to one of two groups by utilizing $c_j = \pm 1$, which is simply a partition of the samples into two groups. Building on this intuition, we formalize an effect size measure for OASIS, which is computed as the difference in mean as measured by $\boldsymbol{f}$ between the two sample groups induced by the sign of c_j. Defining $A_+ = \{j : c_j > 0\}$ and $A_- = \{j : c_j < 0\}$, the effect size is computed as

$$\hat{\Delta} \triangleq \left| \frac{1}{\sum_{j\in A_+} n_j} \sum_{j\in A_+} n_j \hat{\mu}_j - \frac{1}{\sum_{j\in A_-} n_j} \sum_{j\in A_-} n_j \hat{\mu}_j \right|, \qquad [5]$$

where $\hat{\mu}_j = \left[X^\top \boldsymbol{f}\right]_j / n_j$. Defining $\delta_{\mathrm{TV}}(\cdot,\cdot)$ as the total variation distance between two probability distributions, and $\hat{\boldsymbol{p}}_+$ as the empirical row distribution over samples in A_+ (similarly for $\hat{\boldsymbol{p}}_-$), we note that the effect size measure $\hat{\Delta}$ in Eq. (5) satisfies for all $\boldsymbol{f} \in [0,1]^I$

$$0 \leq \hat{\Delta} \leq \delta_{\mathrm{TV}}(\hat{\boldsymbol{p}}_+, \hat{\boldsymbol{p}}_-). \qquad [6]$$

The effect size is bounded between 0 and 1, and cannot be too large if the empirical distributions on the rows is similar between the two groups. The proof of the relationship between effect size and total variation distance, and the motivation for this effect size measure stemming from a simple two group alternative, are detailed in Section 10.D. Empirically, as we later show, this effect size measure allows OASIS to prioritize scientifically interesting tables.

4. Optimization-based approach to selecting $\boldsymbol{f}, \boldsymbol{c}$

We have discussed OASIS's test statistic for a fixed $\boldsymbol{f}$ and $\boldsymbol{c}$. The natural question is then, how to choose $\boldsymbol{f}$ and $\boldsymbol{c}$? Many options are possible. Here, we focus on an intuitive method, `OASIS-opt`, that partitions the observed counts, using part of the table to construct $\boldsymbol{c}, \boldsymbol{f}$ that optimize a p-value bound, and the held-out "test" data to compute a statistically valid p-value (Figure 2A).

A. Maximization-based $\boldsymbol{c}, \boldsymbol{f}$. For simplicity of interpretation and ease of optimization, we examine the p-value bound without the $1-\gamma$ factor. In this case, the p-value bound in Proposition 1 is minimized (ignoring issues of uniqueness) by $\boldsymbol{f}, \boldsymbol{c}$ as:

$$\boldsymbol{f}^\star, \boldsymbol{c}^\star = \underset{0\leq \boldsymbol{f}\leq 1, \|c\|_2 \leq 1}{\mathrm{argmax}} \left| \boldsymbol{f}^\top \tilde{X} \boldsymbol{c} \right|. \qquad [7]$$

Due to the structure of $\tilde{X}$, any $\boldsymbol{c}$ that is a local maxima of Eq. (7) yields $\gamma = 0$. Additionally, any $\boldsymbol{f}$ that is a local maxima of Eq. (7) will take only the discrete values $\{0,1\}$. This is due to the use of Hoeffding's inequality in the construction of the p-value bound, and alternative bounds based on the asymptotic p-value bound in Proposition 2, or finite sample ones based on Bernstein's inequality, for example, would yield non-binary optimized $\boldsymbol{f}$ (Section 11).

The optimization problem in Eq. (7) is non-convex, but is bi-convex in $\boldsymbol{f}$ and $\boldsymbol{c}$. This enables simple (closed form) iterative optimization via alternating maximization to a local maxima, which we encapsulate as an algorithm called `OASIS-opt`. Analysis details and alternating maximization updates are given in Section 11.

Due to the non-convexity of the overall objective, the starting point from which alternating maximization is run matters, and so several initializations (chosen e.g. randomly) could be utilized. We show in Figure 1 how using additional random splits and applying Bonferroni correction to the generated p-values can yield an improved test, ensuring rejection of highly significant tables. However, there is no free lunch; while the resulting method is more likely to detect highly significant tables, the burden of multiple hypothesis correction over additional tables increases. The number of random initializations needed, and potential for cross-validated approaches for computing $\boldsymbol{f}, \boldsymbol{c}$ to improve robustness at the expense of multiple testing correction, are left for future work. These two sources of randomness; statistical randomness in data-splitting, and and randomness for computational efficiency of solving the inner maximization problem, are distinct, with the former being necessary and the latter being a computational tool. This is discussed further in Section 14.A.

B. Relation to the Singular Value Decomposition. Note that if OASIS optimized the raw magnitude of the test-statistic $\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}$, for ℓ_2 constrained $\boldsymbol{c}$ and $\boldsymbol{f}$, this maximum would be attained by $\boldsymbol{f}$ taking the value of the principle left singular vector of $\tilde{X}$, $\boldsymbol{c}$ as the principle right singular vector, and the optimum value would be the maximum singular value, by the variational characterization of the SVD. `OASIS-opt` provides an alternative decomposition of a contingency table to the ubiquitous SVD, within a disciplined statistical framework. While the SVD has a statistical interpretation in some parametric settings, `OASIS-opt`'s alternative decomposition provides a fresh and potentially more statistically tractable and interpretable decomposition.

As in Eq. (3), the X^2 test statistic can be expressed as $X^2 = \|X_{\text{corr}}\|_F^2$. This can be approximated as $\frac{1}{I}\|\tilde{X}\|_F^2$ in the regime where all rows have similar counts. Comparatively, the OASIS test statistic attains a maximum value (up to the ℓ_2 as opposed to ℓ_∞ constraint on $\boldsymbol{f}$) of $\|\tilde{X}\|_2$. Note that the ℓ_2 ball is contained within the ℓ_∞ ball, and so the optimal value of the OASIS test-statistic is in fact lower bounded by $\|\tilde{X}\|_2$. The Frobenius norm is the sum of squared singular values, and so X^2 is essentially summing over all possible directions of deviation. This makes it powerful, but overpowered against uninteresting and unexplainable alternatives such as those with a flat spectrum. OASIS on the other hand computes significance by projecting deviations from the expectation in a single direction. Intuitively, this denoises all the lower signal components, and ensures that OASIS remains robust to biological noise, and yields interpretable results.

An SVD of $\tilde{X}$ or X_{corr} offers one approach to generating $\boldsymbol{c}, \boldsymbol{f}$ for inference with OASIS, however this choice is in general unlikely to yield desirable $\boldsymbol{c}$ and $\boldsymbol{f}$. Empirically, this choice gives worse p-values with less meaningful $\boldsymbol{f}, \boldsymbol{c}$ than `OASIS-opt`, as shown for a toy example in Figure 6. One reason for this is that the OASIS p-value bound penalizes the ℓ_∞ norm of $\boldsymbol{f}$ rather than the ℓ_2 of the SVD. Directly optimizing the p-value objective (with an ℓ_∞ constraint) naturally yields improved p-values. Thus, `OASIS-opt` can for some problems be used as an improved and statistically valid variant of correspondence analysis.

5. A disciplined statistical framework for subgroup classification with `OASIS-opt`

OASIS in its basic form provides a statistical test for a contingency table, against the null of row and column independence. However, after the null has been rejected, many followup questions often remain. One natural question, after the samples have been found to violate the null when combined using $\boldsymbol{c}$, is whether there exist different ways of grouping samples to still yield statistically significant deviations. A broader question is whether, studying many tables defined on the same sets of columns, there is some global inference that can be drawn regarding the underlying clustering and similarity of the samples.

A. Iterative analysis of contingency tables with `OASIS-iter`. Addressing the first question, regarding iterative testing of a single contingency table, we propose a variant of OASIS we call `OASIS-iter`, which is a simple wrapper on top of `OASIS-opt`. First, we define a method, `OASIS-perp`, which takes in a matrix X and a set of vectors $\{v^{(k)}\}$, and optimizes the objective in Eq. (8). This objective is identical to that of `OASIS-opt`, except that the optimizing $\boldsymbol{c}$ is constrained to be orthogonal to all vectors in the input set $\{v^{(k)}\}$. To optimize this objective we again use alternating maximization, with updates derived in Section 11 (Equations 34, 39).

$$\boldsymbol{f}^\star, \boldsymbol{c}^\star = \underset{\substack{0 \le \boldsymbol{f} \le 1, \|c\|_2 \le 1 \\ \boldsymbol{c}^\top \boldsymbol{v}^{(k)} = 0 \ \forall k \in [K]}}{\operatorname{argmax}} \left| \boldsymbol{f}^\top \tilde{X} \boldsymbol{c} \right|. \qquad [8]$$

Algorithm 1 `OASIS-iter`

Input: Contingency table X
$X_{\text{train}}, X_{\text{test}} \leftarrow$ `splitDataset`(X) ▷ Split data randomly into train and test portions
for $i = 1, 2, \ldots$ **do**
  # Run `OASIS-opt` with constraint: $\boldsymbol{c}^{(i)} \perp \boldsymbol{c}^{(j)} \ \forall j < i$
  $\boldsymbol{c}^{(i)}, \boldsymbol{f}^{(i)} \leftarrow$ `OASIS-perp`$(X_{\text{train}}, \{\boldsymbol{c}^{(j)}\}_{j=1}^{i-1})$
  # Compute OASIS p-value bound
  $p^{(i)} \leftarrow$ `pvBound`$(X_{\text{test}}, \boldsymbol{f}^{(i)}, \boldsymbol{c}^{(i)})$
  if $p^{(i)} > 0.05$ **then**
    Break
return $[\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}, \ldots], \ [\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}, \ldots], \ [p^{(1)}, p^{(2)}, \ldots]$

Each $\boldsymbol{c}^{(i)}$ outputted by this procedure represents a different, orthogonal direction along which this table can be partitioned so as to yield a significant p-value. This approach provides a statistical stopping criterion for cluster identification; once the obtained $p^{(k+1)}$ exceeds the desired significance level (e.g. 0.05), the number of clusters can be estimated as k. From an algorithmic perspective, this method is a simple extension of `OASIS-opt`, as the restriction that each $\boldsymbol{c}^{(i)}$ is orthogonal to all previous $\boldsymbol{c}$ is a convex constraint. The derivation for this orthogonally-constrained optimization is detailed in Section 11.D, and simulations on synthetic data are in Section 14.F.

B. Data-aggregation. With this new primitive of iterative analysis of a single table, one natural question is whether we can aggregate the data across all tables, and run `OASIS-iter` to iteratively identify all underlying clusters. We design a method

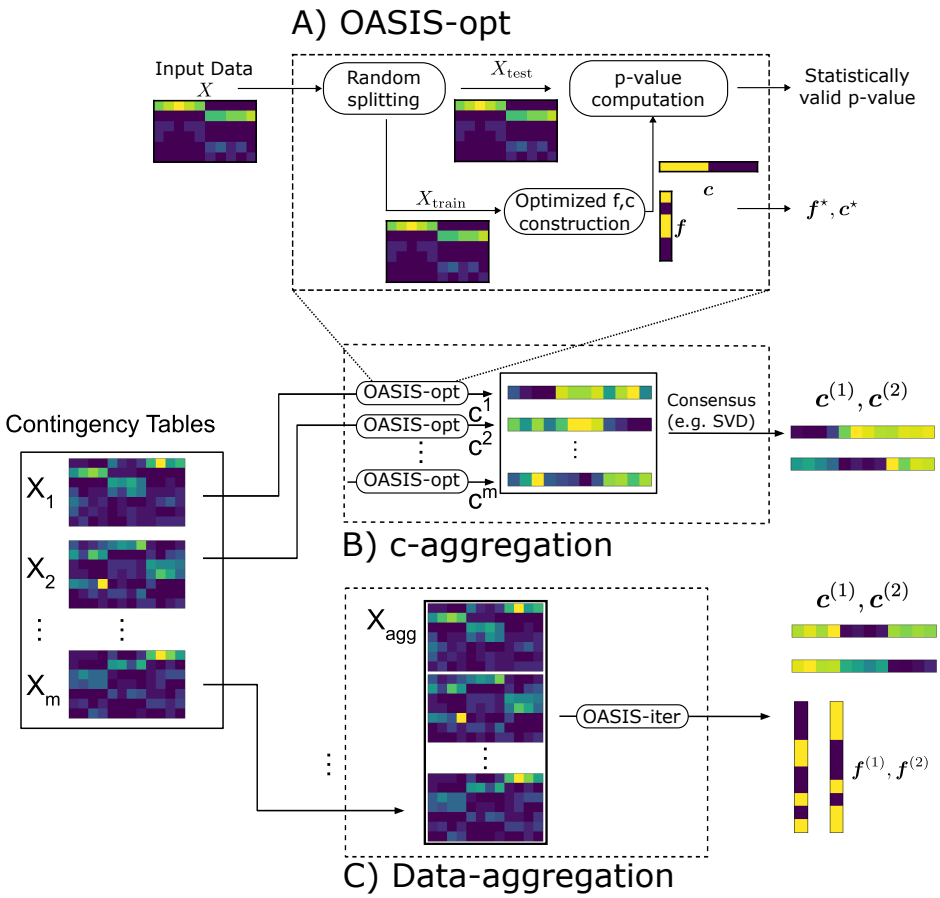

Fig. 2. Figure showing the algorithms and methods we build on top of the base OASIS framework. **A)** `OASIS-opt`, the primary instantiation of OASIS used in this work. `OASIS-opt` employs sample splitting to find optimized $\boldsymbol{f}, \boldsymbol{c}$ on the training data, before generating a statistically valid p-value on the held out test data. B,C depict two algorithms we propose for inferring latent structure from a collection of tables defined on the same set of columns. As building blocks, we use `OASIS-opt` for $\boldsymbol{c}$-aggregation, and `OASIS-iter` for data-aggregation, as depicted. **B)** $\boldsymbol{c}$-aggregation (Algorithm 3), which performs inference on each table marginally using `OASIS-opt`, and aggregates the resulting $\boldsymbol{c}$ vectors, can be used to identify latent structure across the samples. **C)** Data-aggregation (Algorithm 2), which stacks the contingency tables into one large data matrix X_{agg}, and performs iterative analysis on this table using `OASIS-iter` (Algorithm 1).

Algorithm 2 `Data-aggregation`

Input: List of contingency tables $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$
For each $i \in [m]$, discard all rows of $X^{(i)}$ with fewer than 10 counts, then compute M_i as the total remaining counts
Construct $X_{\text{agg}} = \begin{bmatrix} X_1/\sqrt{M_1} \\ X_2/\sqrt{M_2} \\ \vdots \\ X_m/\sqrt{M_m} \end{bmatrix}$
$[\boldsymbol{c}^{(1)}, \ldots], [\boldsymbol{f}^{(1)}, \ldots], [p^{(1)}, \ldots] \leftarrow$ `OASIS-iter`(X_{agg})
 # Can utilize $\{p^{(i)}\}$ to determine number of components
return $\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}, \boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}$

for this task called data-aggregation Algorithm 2, which takes as input a set of tables. These tables should have shared latent structure, and can be obtained, as in our setting, by filtering `OASIS-opt`'s calls for tables with a large effect size, many observations, and a significant p-value after performing Benjamini-Yekutieli (14) correction across a much larger list of tables outputted from NOMAD (1). These tables are then vertically concatenated into one larger table, X_{agg}, upon which `OASIS-iter` is run. This method is shown graphically in Figure 2B, and is formalized in Algorithm 2. For ease of visualization, in this work we terminate this method after 2 iterations, and utilize the outputted $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ as the 2D sample embeddings.

Note that this aggregated table does not generally satisfy the contingency table null, due to the correlation in entries from the different sub-tables. Algorithmically, we discard rows in this table with fewer than 10 observations, to minimize the computational burden. Additionally, to avoid having a single high-count sub-table dominate all the others in this computation, we normalize the counts in each sub-table by $1/\sqrt{M}$, the total number of counts in that sub-table.

C. $\boldsymbol{c}$-aggregation. Due to the statistical and practical issues with normalizing and aggregating tables before performing statistical testing, another possible method is to independently identify 1D embeddings for each table, and then aggregate these results. We call this method $\boldsymbol{c}$-aggregation, Algorithm 3, which utilizes `OASIS-opt` to generate $\boldsymbol{c}$ for each table, stacks these vectors vertically into a matrix C, and then computes a low dimensional embedding for the samples, e.g. with an SVD. This is formally described in Algorithm 3, and shown in Figure 2C. As before, the tables used as input for this

Algorithm 3 $\boldsymbol{c}$ `aggregation`

Input: List of contingency tables $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$
$\boldsymbol{f}^{(i)}, \boldsymbol{c}^{(i)} \leftarrow$ `OASIS-opt`$(X^{(i)})$ for all $i \in [m]$
Construct $C = \begin{bmatrix} - & \boldsymbol{c}_1 & - \\ - & \boldsymbol{c}_2 & - \\ & \vdots & \\ - & \boldsymbol{c}_m & - \end{bmatrix}$
Fill missing entries in C with 0
Compute SVD of $C = U\Sigma V^\top$
 # Can use Σ to decide on number of components to use
return $\boldsymbol{v}_1, \boldsymbol{v}_2$, the first two right singular vectors

method could be filtered for those with a large effect size, many observations, and a significant post-correction p-value.

In the construction of C, a value of 0 is assigned to missing entries (samples which had no observations for a given table). This highlights an interesting feature in this optimization procedure; not all samples must be present in all tables in order for this joint inference to be performed. Defining the i-th right singular vector of C as $\boldsymbol{c}^{(i)}$, $\boldsymbol{c}^{(1)}$ is mathematically equivalent to finding a vector which maximizes the average squared inner product with each $\boldsymbol{c}$.

6. Numerical results

Many problems in modern data science, including in genomics, map to contingency tables. Analysis of contingency tables in genomics applications, such as those that treat sequences or genes as counts, is core to biological inference. In these applications, noise is often introduced, sometimes without a good parametric model. We now show that OASIS is robust to some important classes of noise models, and that OASIS has substantial power against several classes of alternative hypotheses of interest. We also show that OASIS can be used to perform classification tasks not possible with Pearson's X^2, enabling reference-genome-free strain classification in SARS-CoV-2 and M. Tuberculosis. For OASIS, we show the performance of `OASIS-opt` as well as `OASIS-rand`, utilizing the improved p-value bounds presented in this work.

A. Empirical robustness against uninteresting alternatives.

Next generation sequencing data is commonly treated as count data on contingency tables, for example with single-cell RNA sequencing often expressed as a gene counts matrix. While the statistical null typically tests whether observations in each sample are identically distributed, biochemical noise introduced during sampling generates overdispersed, non-identically distributed observations, violating this null. The field has converged on modeling such data with negative binomial distributions (15, 16). Probabilistically, a negative binomial random variable is equivalent to a Poisson random variable with a random, gamma-distributed, mean. In terms of the observed table, this would yield an alternative probabilistic model where $X_{ij} \sim \text{Pois}(\Lambda)$, where $\Lambda \sim \Gamma(\lambda/\alpha, \alpha)$ where λ is the expected number of observations, and under the true null $X_{ij} \sim \text{Pois}(\lambda)$ counts would be observed ($\alpha = 0$). This does not satisfy the contingency table null for $\alpha > 0$, but also does not represent a biologically meaningful deviation. Therefore, a robust test would be able to operate in noise regimes practical for biological sequencing data. OASIS is able to accomplish this, as shown in Table 1. Taking a table with a mean of $\lambda \approx 129$ observations per sample with a uniform row distribution, for 20 rows and 10 columns `OASIS-opt` rejects 7.5% of tables, while Pearson's X^2 test rejects 95.7% of tables. Additional simulation details in Section 14.B.1.

num rows	num cols	X^2	OASIS-rand	`OASIS-opt`
5	10	0.143	0.003	0.010
	50	0.318	0.006	0.019
	400	0.318	0.006	0.019
20	10	0.957	0.054	0.075
	50	1.000	0.065	0.193
	400	1.000	0.078	0.874
100	10	1.000	0.947	0.996
	50	1.000	0.996	1.000
	400	1.000	0.998	1.000

Table 1. Power against negative binomial-based noise model as in (15). Uniform row distribution, full plots and details in Section 14.B.1.

As expected, Pearson's X^2 rejects nearly all tables generated under this negative binomial sampling model (Table 1). Intuitively, the OASIS statistic should be robust to the negative binomial alternatives, as the deviation is unstructured, and in the parameter regimes explored here it is. For all tests, the rejection fraction is monotonic in the number of rows and columns, but OASIS retains significant robustness compared to X^2. These tabulated simulation results show the fraction of rejected tables (for $\alpha = 0.05$) at the level of negative binomial overdispersion predicted by the sampling depth (15).

More generally, an applied task in genomics and other applications is to reject the null only when it is "interesting". Below, we refer to an alternative as uninteresting if it can be explained by a small number of outlying matrix counts, or from sampling from a set of distributions $\{\boldsymbol{p}^{(j)}\}$ where $\boldsymbol{p}^{(j)}$ is the row distribution of the j-th column, and for each j, k $\|\boldsymbol{p}^{(j)} - \boldsymbol{p}^{(k)}\|_1 < \epsilon$. Such distributions, while statistically different from the null, represent small effects that may be due to contamination or equipment error, and are not a priority to detect. OASIS provides robustness against these sources of unstructured noise that represent biologically uninteresting alternatives. We characterize this with simulations against ℓ_1 corruption of each individual column's probability distributions in Figure 11.

A.1. Power against simulated alternatives. OASIS has substantial power (in some regimes more than X^2) against large, structured deviations from independence, such as when the samples can be partitioned into two groups with nearly disjoint supports. These examples arise in important applications such as detecting recombination in B cells that generate antibodies–V(D)J recombination–and differentially regulated alternative splicing, among others (simulated in Section 14.C).

Approximate power calculations corroborate these numerical results. For a simulated alternative, consider the toy setting where we observe the expected underlying alternative matrix; for sample j with row probabilities $\boldsymbol{p}^{(j)}$ we observe $X^{(j)} = n_j \boldsymbol{p}^{(j)}$ instead of the random draw $X^{(j)} \sim \text{multinomial}(n_j, \boldsymbol{p}^{(j)})$. Analyzing OASIS and X^2 in this deterministic setting, we show that OASIS has power comparable to and in some regimes exceeding X^2, in particular when the number of rows is large, as tabulated in Table 2 (analysis details in Section 15). We conjecture that while OASIS may not

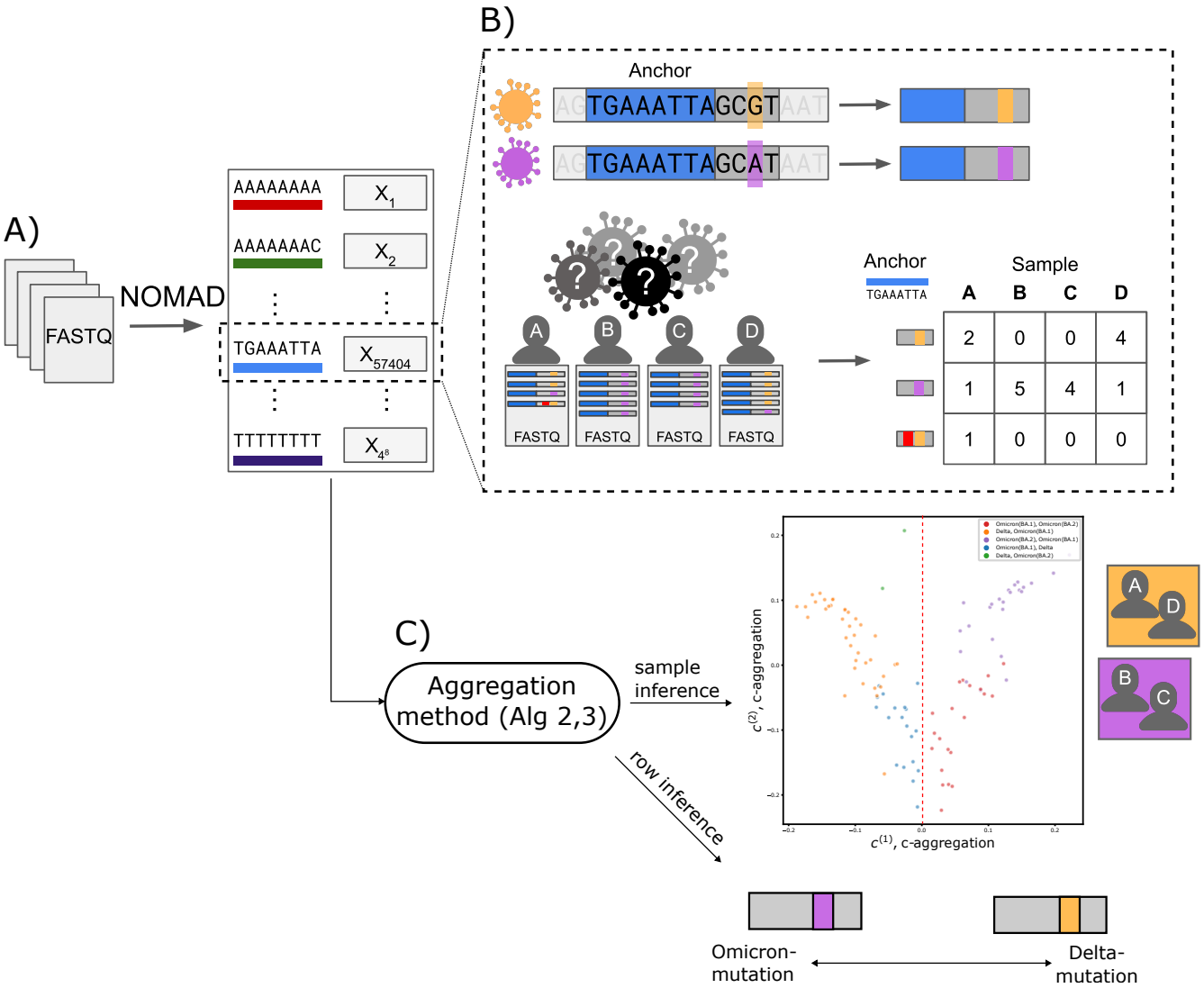

Fig. 3. A: NOMAD generates contingency tables from genomic sequencing data, here FASTQ files, for all possible anchors (length k genomic sequences) (1). **B:** Shown in greater detail is the process for 1 specific anchor, TGAAATTA. This anchor highlights a mutation between two strains of SARS CoV-2, between Omicron (purple) and Delta (orange). Below, viral sequencing data from the 4 individuals (samples) infected with COVID is shown. However, it is a priori unknown which strain each individual was infected with, and only an incomplete reference genome is available that does not reflect all strain level variation. For the fixed anchor sequence (shown in blue), NOMAD counts for each sample the frequency of sequences that occur immediately after, and generates a contingency table, where the columns are indexed by the samples and the rows are indexed by the sequences. Shown in the plot is one read in sample A which underwent sequencing error, shown in red, and thus yielded an additional discrete observation– a sequence– resulting in an extra row. This sequencing error leads to tables with many rows with low counts. Note that we cannot know a priori which rows of this table are due to sequencing error, as we are simply observing raw sequencing data; there is no available ground truth, unless we are given a reference genome and are able to successfully align the reads to the correct location, the exact problem NOMAD tries to avoid. **C:** The contingency tables generated by NOMAD are defined over the same set of samples (patients), so we can use these to infer sample origins. Plot shown depicts results of running c-aggregation run on SARS-CoV-2 data (6), perfectly predicting whether a patient has Delta or not, and yielding highly predictive performance for subvariant classification (Omicron BA.1 versus BA.2). data-aggregation yields high predictive power over the strain of targets, with a 93% correct classification accuracy of whether a target was Delta or not. In the depicted toy example, this would correspond to grouping targets and individuals by strain as shown.

exhibit the optimal asymptotic rate against certain classes of alternatives (as shown by the unique per-sample expression setting, motivated by V(D)J recombination), it does have power going to 1 as the number of observations goes to infinity across a broad class of alternatives. In additional simulations, we show that OASIS is properly characterized by its asymptotic p-value (Figure 27) and so is asymptotically calibrated.

B. SARS-CoV-2 variant detection. To illustrate OASIS's performance we arbitrarily selected a public dataset of SARS-CoV-2 infections from the short read archive, here focusing on one generated in France tracking patients where Omicron and Delta were co-circulating (6). We show that OASIS detects variants in SARS-CoV-2 by mapping sequencing data from each of 103 patients' nasal swabs (6) to a contingency table via NOMAD, as described in (1). NOMAD generates a contingency table for each length k-subsequence present (called an anchor k-mer) from genomic sequencing (Figure 3). A statistical test with good scientific performance will identify anchors near known mutations in the SARS-CoV-2 genome that distinguish Omicron and Delta. Data processing details are deferred to Section 13.

	OASIS p-value	Pearson's X^2 p-value	Notes
Two-group	$\exp\left(-M\delta_{\mathsf{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})\right)$	$\exp\left(-\frac{M^2}{I\times J}\right)$	OASIS is more powerful when $\frac{M}{I\times J}$ is small
Time series	$\exp\left(-M\delta^2_{\mathsf{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})\right)$	$\exp\left(-M\delta^2_{\mathsf{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}) \times \frac{M}{I\times J}\right)$	Same behavior as two group setting.
Unique per-sample expression	$\exp(-M)$	$\exp(-M^2)$	X^2 more powerful, but OASIS has power going to 1
One deviant sample	$\exp(-M/J)$	$\exp(-M^2/J^2)$	X^2 has too much power

Table 2. Table highlighting approximate power calculation results from Section 15. Multiplicative constants (leading, and in the exponent) are not included in this table for clarity.

OASIS and X^2 yield substantially different results on the 100,914 tables generated by NOMAD. We demonstrate the improvement in biological inference enabled by OASIS, utilizing as a measure of biological relevance for each method's called tables how well these calls can predict sample metadata. For each sample, we have associated metadata indicating the clinical ground truth of whether the patient (sample) was infected with Delta. For each table called by `OASIS-opt`, we use as its 1D sample embedding the vector $\boldsymbol{c}$, taking the sign of each entry to generate a two-group partitioning. The measure of biological relevance used is then computed as the absolute cosine similarity, $s(x, y) = \frac{1}{\|x\|\|y\|}|\langle x, y\rangle|$, between the sample metadata and the sign of $\boldsymbol{c}$. This directly corresponds to the fraction of agreed upon coordinates in x, y. For X^2 this process is mirrored, where instead of $\boldsymbol{c}$, the principle right singular vector for correspondence analysis is used.

Out of the 100,914 tables, OASIS rejects 28,430, and Pearson's X^2 test rejects 71,543 tables. However, when the tables that X^2 rejects are decomposed, they do not yield signal that correlates well with the ground truth; examining quantiles of the two empirical distributions, the 0.5 and 0.9 quantiles of these distributions are 0.22 and 0.66 for OASIS, as opposed to 0.10, 0.52 for X^2. We show in Figure 20, for these significant tables, correspondence analysis yields similarly high correlation, indicating that it is the improved detection of OASIS, rather than the handicap of correspondence analysis versus OASIS's $\boldsymbol{c}$, that is yielding the larger similarities. In addition, of the 389 anchors that OASIS calls that X^2 does not, the 0.5 and 0.9 quantiles of the absolute cosine similarity between these binarized $\boldsymbol{c}$ and the underlying ground truth of whether a patient has Delta are 0.15, 0.76, indicating many biologically significant tables that were missed by X^2.

Pearson's X^2 test rejects 71,543 out of the total 100,914 tables tested. Analyzing all 16,611 tables rejected by X^2 with a p-value of 0, the 1D embeddings obtained from correspondence analysis had low absolute cosine similarity with the biological ground truth 0.087 and 0.40 at 0.5 and 0.9 quantiles. Looking at specific rejected tables, we select two of the most significant tables rejected by X^2 but not by `OASIS-opt`. Correspondence analysis yields embeddings with minimal correlation with the ground truth, 0.15 and 0.02 for the two tables selected, one of which visually appears to have just detected one deviating sample (Figure 18).

In comparison, `OASIS-opt` yields 28,430 total calls. $\boldsymbol{c}$ found by `OASIS-opt` have a significantly higher absolute cosine similarity with the clinical ground truth of whether the patient had a Delta infection, as discussed. For the 2114 tables that `OASIS-opt` assigns a p-value of 0, the cosine similarity with ground truth has 0.5 and 0.9 quantiles of 0.69 and 0.79. Systematic investigation of `OASIS-opt`'s top significance calls show a strong enrichment for being concordant with clinical metadata describing whether a patient had a Delta infection. We consider for this analysis tables with effect size in the top 10% and total counts $M > 1000$, as these are predicted to have high correspondence with this clinical metadata, yielding 2495 remaining tables. Of these, the absolute cosine similarity between the identified $\boldsymbol{c}$ with the indicator of whether a patient has Delta has 0.5 and 0.9 quantiles of 0.72 and 0.82. In Figure 20, we show the ECDF of all 28,430 of `OASIS-opt`'s calls absolute cosine similarity, and all 71,543 calls from X^2, showing an order of magnitude difference in fraction of identified tables with e.g. $> .6$ cosine similarity.

Comparing with the original statistic used in NOMAD (1), OASIS-rand identifies 5932 significant tables. All except 48 of these are identified by `OASIS-opt`. The two most significant tables that are called by `OASIS-opt` (with an effect size in the top 10% and counts greater than 1000) and not by OASIS-rand are shown in Figure 19, both having a high absolute cosine similarity with sample metadata, 0.76, showing that the additional calls `OASIS-opt` is providing are biologically significant. We show additionally that OASIS provides calls beyond those possible with X^2, showing in Figure 17 the 5 tables in the reduced anchor list above (effect size and counts filtered), which all have large cosine similarities with the ground truth metadata.

B.1. Data-aggregation. Running data-aggregation (Algorithm 2) on the SARS-CoV-2 dataset, yields important inference on both targets and samples. We select the first two components, as the first component contains 57% of the total power as measured by $\sigma_1^2/\sum_k \sigma_k^2$, and with the second contained 76% of the cumulative power (full spectrum shown in Figure 21). Focusing first on samples, the first two $\boldsymbol{c}$ output by data-aggregation have high predictive power over strain information. A simple threshold on $\boldsymbol{c}^{(1)}$ yields perfect prediction of whether a patient has Delta or not. The second vector, $\boldsymbol{c}^{(2)}$, enables classification of Omicron subvariants. A linear predictor on $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ yields 94/103 (91%) accuracy in predicting whether the primary strain is Omicron BA.1. These results are shown in Figure 4.

Running data-aggregation on these matrices also allows us to perform joint inference on targets from different tables. The outputted $\boldsymbol{f}^{(1)}$ of this procedure has significant predictive power over whether a target corresponds to an Omicron mutation or not. We show this by using Bowtie (17) to check whether a target noiselessly aligns to a Delta assembly, an Omicron BA.1 assembly, or an Omicron BA.2 assembly (details in Section 13). Of the 51,557 anchor-target pairs (rows of X_{agg}), we filtered for those targets which constitute at least 5% of the total counts for that anchor. Then, we label a target as Delta if it maps to the Delta assembly and neither of the Omicron ones, and Omicron if it maps to at least one of the Omicron assemblies but not the Delta one. This yields 4836 sufficiently abundant anchor-target pairs which map, consti-

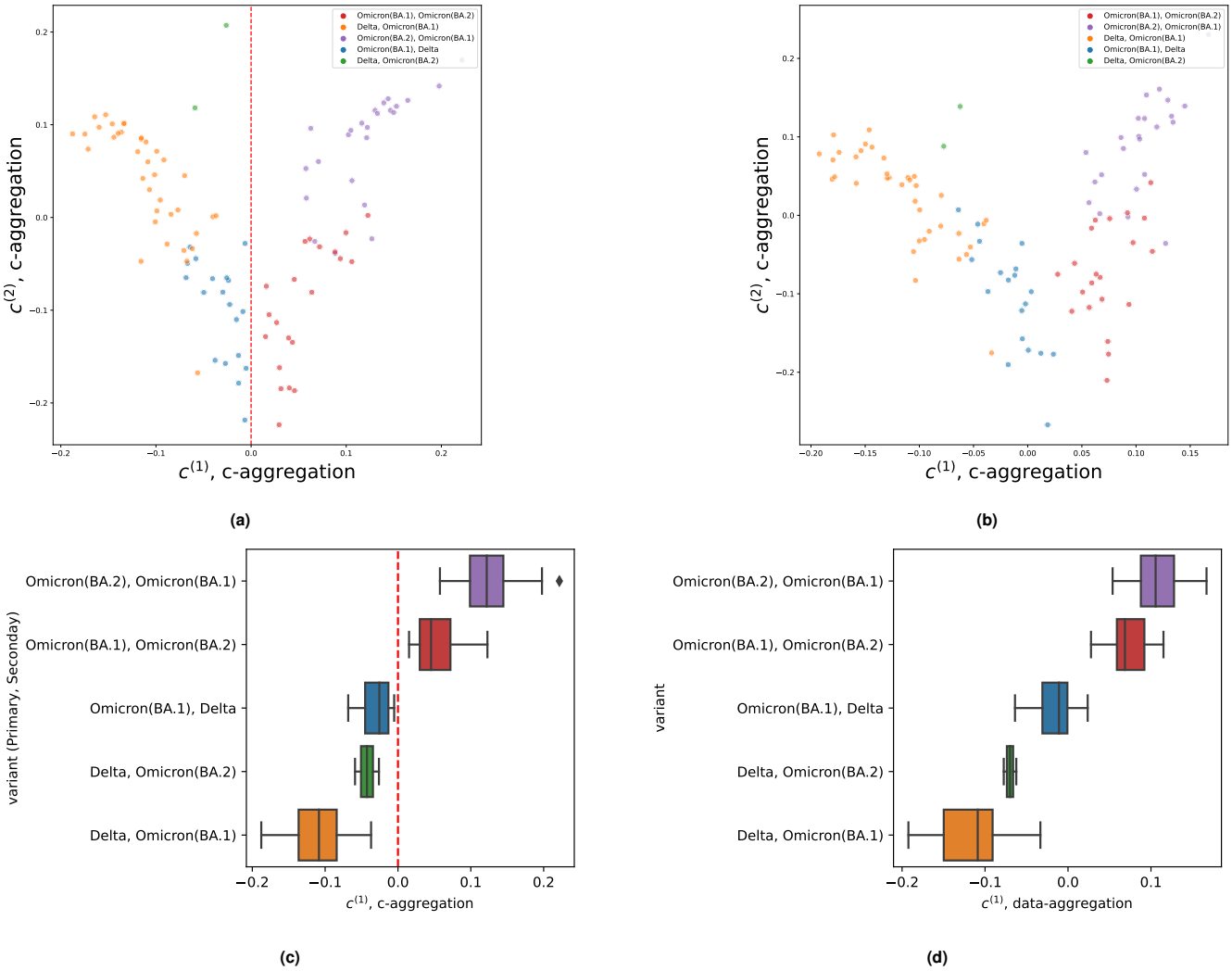

Fig. 4. Interpretation of OASIS-rejected null from SARS-CoV-2 coinfection data (6). Tables were generated by NOMAD (1) and tested with `OASIS-opt`. (a) and (c) depict the results of $\boldsymbol{c}$-aggregation (Algorithm 3). (a) shows the generated 2D embeddings, which perfectly classify whether a patient has Delta or not at $\boldsymbol{c}^{(1)} \gtrless 0$, highlighted in (c). (b) and (d) depict the results of data-aggregation (Algorithm 2). Perfect separation of Delta vs non-Delta samples, no longer at $\boldsymbol{c}^{(1)} \gtrless 0$. Boxplot shows 1D projection. Analysis details in Section 13.

tuting 2,479 of the 2,495 anchors tested. Comparing with this ground truth vector of whether a target was classified as Delta or not, $\boldsymbol{f}$ had an absolute cosine similarity of 0.86, corresponding to 4496/4836 (93%) targets correctly classified. This is with no additional parameter tuning. The targets that are incorrectly classified have the potential to uncover novel biology, which is beyond the scope of this paper. Examining the first two anchors with targets that are considered misclassified, we observe that the first strongly suggests an unannotated Omicron deletion, while the second perfectly identifies and predicts an annotated Omicron deletion, however due to the lookahead distance used (a parameter in NOMAD (1)) both targets perfectly map to the Wuhan reference. We provide alignments for these results in Figure 23.

B.2. $\boldsymbol{c}$-aggregation. We additionally apply $\boldsymbol{c}$-aggregation (Algorithm 3) to this dataset, restricting our attention to the first two right singular vectors of C. The first singular vector of the matrix C perfectly partitions individuals infected with the Delta strain from those that were not (at the threshold $\boldsymbol{c}^{(1)} \gtrless 0$). The second singular vector differentiates between the Omicron BA.1 and BA.2 subvariants as shown in Figure 4 where the x-axis is the principle right singular vector and the y-axis is the second right singular vector. When tasked with predicting whether a sample has BA.1 as its primary variant, a simple linear classifier in this two dimensional space is able to correctly classify 95/103 (92%) of the samples. Further analysis details are in Section 13.A.

Since $\boldsymbol{c}$-aggregation provides inference on the samples, it can be used to indirectly provide inference on the targets. We show this by utilizing $\hat{y} = \text{sign}(X_{\text{agg}}\boldsymbol{c}^{(1)})$ as a predictor of whether a patient is infected with Delta or not, which provides an absolute cosine similarity of 0.80, yielding correct predictions on 4357/4836 targets (90%). This highlights the power of data-aggregation in performing joint inference directly on the rows.

Together, these analyses support the idea that OASIS finds tables representing important biological differences, and can provide auxiliary information on samples ($\boldsymbol{c}$) that can guide

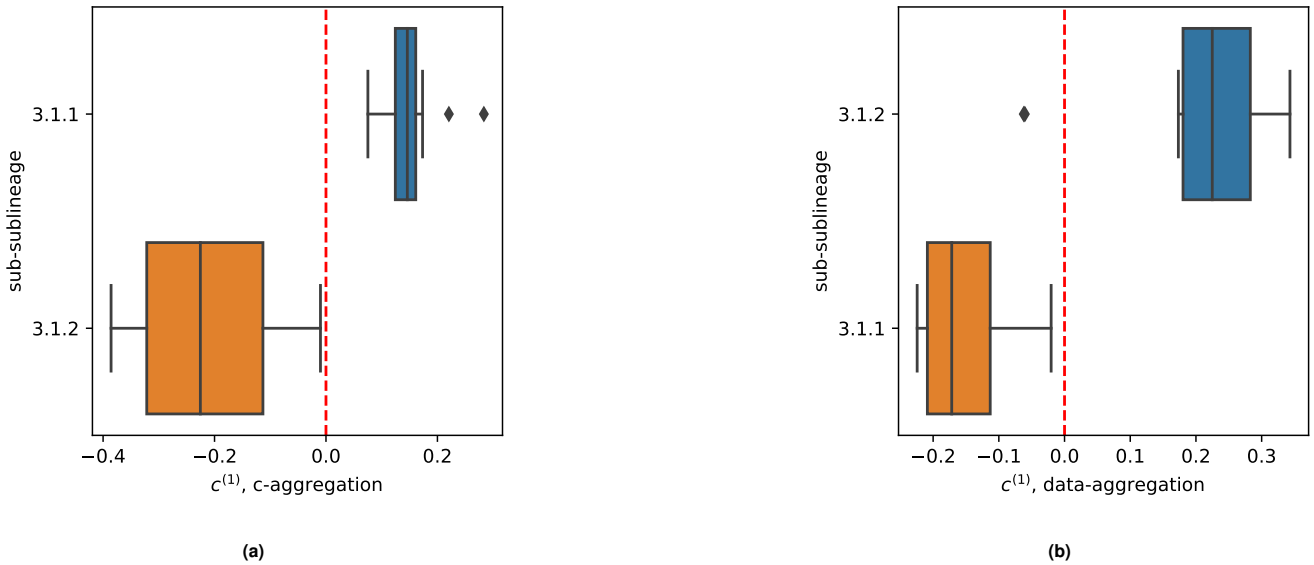

Fig. 5. Interpretation of OASIS-rejected null from M. Tuberculosis data (6). Tables were generated by NOMAD (1) and tested with OASIS. (a) and (c) depict the results of $\boldsymbol{c}$-aggregation (Algorithm 3). (a) shows the generated 2D embeddings, which perfectly classify patients based on sub-sublineage at $\boldsymbol{c}^{(1)} \gtrless 0$, highlighted in (c). (b) and (d) depict the results of data-aggregation (Algorithm 2). Two samples are misclassified (visually, one on top of the other), but with a much larger margin for the rest. 2D plot in Figure 24.

scientific inference within a disciplined framework. These features are absent from X^2 or exploratory correspondence analysis.

C. M. Tuberculosis strain identification. We tested the generality of `OASIS-opt`'s inferential power to classify microbial variants in a different microorganism, the bacterium M. Tuberculosis. We processed data from 25 isolates from two sub-sublineages known to derive from the sublineage 3.1: 3.1.1 and 3.1.2 (Accession ID PRJEB41116 (18)). Currently, bacterial typing is based on manual curation and SNPs, is highly manually intensive, and requires mapping to a reference genome.

As with SARS-CoV-2, we utilized NOMAD to generate tables which we tested with `OASIS-opt`. 80,519 tables were called by `OASIS-opt` (corrected p-value), 258 with effect size in the 90-th quantile and total count $M > 1000$. To avoid biological noise, we preemptively filtered out tables with targets with repetitive sequences ($>$ 10 repeated basepairs). `OASIS-opt` was run blind to sample metadata and the M. Tuberculosis reference genome.

C.1. Data-aggregation. We run data-aggregation, Algorithm 2, on the filtered Tuberculosis tables described above. The top two identified vectors, $\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}$, yield a well separated partitioning, in terms of sub-sublineage. Two samples are misclassified, with the rest being perfectly predicted by $\boldsymbol{c}^{(1)} \gtrless 0$, yielding an accuracy of 23/25 (92%) (Figure 5).

C.2. $\boldsymbol{c}$-aggregation. We utilized $\boldsymbol{c}$-aggregation on the same set of filtered tables. The first singular value constitutes 49% of the power in the matrix, and the top 2 singular values comprise 72% of the spectrum's power. A 1D embedding from the first singular vector leads to a perfectly separation of sub-sublineages, classifying at the threshold $\boldsymbol{c}^{(1)} \gtrless 0$, as shown in Figure 5.

7. Discussion

In this paper we have described several applications and extensions of OASIS, but there are many interesting directions of ongoing and future work. A known failing of X^2 is its inability utilize available metadata regarding the rows and columns of the matrix (e.g. ordinal structure). OASIS however can incorporate this knowledge, in its construction of $\boldsymbol{f}, \boldsymbol{c}$ (discussed in Section 12.A). Additionally, with its natural effect size measure, OASIS can be used as a coefficient of correlation between two random variables, even continuous valued ones (discussed in Section 12.B). As we have shown, OASIS empirically has power against a wide variety of alternatives, and can be shown in toy models to have more power than Pearson's X^2 test in some regimes (discussed further in Section 15). While Hoeffding's inequality for sums of bounded random variables provides one p-value, alternatives can be constructed using different methods, such as empirical variants of Bernstein's inequality (19) or ones based on Stein's method (20), leading to alternative optimization criterion for $\boldsymbol{f}, \boldsymbol{c}$ (discussed in Section 12.C). OASIS's statistic can also be extended to different nulls (e.g. volume-based (21)) by using alternative concentration inequalities (22). While OASIS currently analyzes 2 dimensional tables, its simple and theoretically tractable approach of centering, normalizing, and projecting can also be generalized to tensors (discussed in Section 12.D). Regularization could also be imposed on $\boldsymbol{f}$ to enable more interpretable inference on rows. Finally, as currently presented, in a high dimensional setting with large numbers of tables generated in a single experiment over the same set of samples, OASIS performs testing on each table independently. We have introduced two approaches for jointly analyzing tables, data-aggregation and $\boldsymbol{c}$-aggregation, to identify latent relationships between samples. Ongoing and future work investigates other approaches, including adaptive ones, to perform this statistical inference and testing on tensors.

While we would not be surprised if such tests have been previously introduced, we have not been able to locate them in the literature.

8. Conclusion

This paper provides a theoretical framework for and applied extensions of a new test, OASIS, that maps statistical problems involving discrete data to a statistic that admits a closed form finite sample p-value. Here, we have focused on practical scientific problems, applying OASIS to data in contingency tables, however this framework can be extended to other statistics and data formats.

We have focused the development of OASIS on genomics applications, a rapidly expanding field with diverse scientific applications from single cell genomic inference, to viral and microbial strain detection. The field still relies on classical statistical tests and parametric models. OASIS provides an alternative to these approaches that is both empirically robust and scientifically powerful. On real and simulated data, `OASIS-opt` prioritizes biologically significant signals: without metadata or any reference genome or specialization to the application at hand, it can classify viral variants including Omicron sub-variants and Delta in SARS-CoV-2 as well as sub-sublineages of Mycobacterium Tuberculosis. Moreover, it is robust to noise introduced during sequencing, the genomic data generation process, including in single cell genomics (23). OASIS provides a tool towards answering questions in mechanistic biology that are difficult or impossible to address with current statistical approaches for genomics. Today, these tasks are difficult and requires intensive manual labor (6). In addition to its scientifically accurate output, this lineage assignment is performed in a rigorous statistical framework that promises further theoretical extensions in clustering.

In summary, OASIS is a finite-sample valid test that has many important statistical properties not enjoyed by X^2 which will enable its use across many disciplines in modern data science. It is computationally simple, robust against deviations from the null in scientifically uninteresting directions, and provides a statistically rigorous way for the analyst to interpret rejections of the null hypothesis. In addition to finite sample p-values, the OASIS statistic admits an asymptotic distributional characterization under the null. Simulations support the theoretical guarantees provided for OASIS, showing its good power in scientifically important settings, and indeed, increased power over X^2 in some.

1. K Chaung, T Baharav, I Zheludev, J Salzman, A statistical, reference-free algorithm subsumes myriad problems in genome science and enables novel discovery. *bioRxiv* (2022).
2. Y Chen, P Diaconis, SP Holmes, JS Liu, Sequential monte carlo methods for statistical analysis of tables. *J. Am. Stat. Assoc.* **100**, 109–120 (2005).
3. K Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* **50**, 157–175 (1900).
4. A Agresti, *Categorical data analysis*. (John Wiley & Sons) Vol. 792, (2012).
5. P Diaconis, B Sturmfels, Algebraic algorithms for sampling from conditional distributions. *The Annals statistics* **26**, 363–397 (1998).
6. A Bal, et al., Detection and prevalence of sars-cov-2 co-infections during the omicron variant circulation in france. *Nat. Commun.* **13**, 1–9 (2022).
7. GR Iversen, Decomposing chi-square: A forgotten technique. *Sociol. Methods & Res.* **8**, 143–157 (1979).
8. MJ Greenacre, Correspondence analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 613–619 (2010).
9. LL Hsu, AC Culhane, Correspondence analysis for dimension reduction, batch integration, and visualization of single-cell rna-seq data. *Sci. Reports* **13**, 1–17 (2023).
10. J Salzman, *Spectral analysis with markov chains*. (Citeseer) Vol. 68, (2007).
11. H Lancaster, The derivation and partition of χ^2 in certain discrete distributions. *Biometrika* **36**, 117–129 (1949).
12. P Diaconis, RC Griffiths, Reproducing kernel orthogonal polynomials on the multinomial distribution. *J. Approx. Theory* **242**, 1–30 (2019).
13. R Leinonen, H Sugawara, M Shumway, INSD Collaboration, The sequence read archive. *Nucleic acids research* **39**, D19–D21 (2010).
14. Y Benjamini, D Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Annals statistics* pp. 1165–1188 (2001).
15. MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21 (2014).
16. R Jiang, T Sun, D Song, JJ Li, Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome biology* **23**, 1–24 (2022).
17. B Langmead, C Trapnell, M Pop, SL Salzberg, Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology* **10**, 1–10 (2009).
18. V Dreyer, et al., High fluoroquinolone resistance proportions among multidrug-resistant tuberculosis driven by dominant l2 mycobacterium tuberculosis clones in the mumbai metropolitan region. *Genome Medicine* **14**, 95 (2022).
19. A Maurer, M Pontil, Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* (2009).
20. S Chatterjee, Stein's method for concentration inequalities. *arXiv preprint math/0604352* (2006).
21. P Diaconis, B Efron, Testing for independence in a two-way table: new interpretations of the chi-square statistic. *The Annals Stat.* **13**, 845–874 (1985).
22. W Hoeffding, The large-sample power of tests based on permutations of observations. *The Annals Math. Stat.* pp. 169–192 (1952).
23. R Dehghannasiri, et al., Unsupervised reference-free inference reveals unrecognized regulated transcriptomic complexity in human single cells. *bioRxiv* (2022).
24. MJ Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. (Cambridge University Press) Vol. 48, (2019).
25. P Billingsley, *Probability and measure*. (John Wiley & Sons), (2008).
26. S Chatterjee, A new coefficient of correlation. *J. Am. Stat. Assoc.* **116**, 2009–2022 (2021).
27. M Kokot, M Długosz, S Deorowicz, Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).

9. Data Availability

M. Tuberculosis data is publicly available under Accession ID PRJEB41116 (18). SARS-CoV-2 data is publicly available under Accession ID PRJNA817806 (6). Code for NOMAD is publicly available at https://github.com/salzman-lab/nomad, along with code for OASIS inference.

10. Proofs

As discussed, OASIS's test statistic S has a simple linear algebraic characterization. For $X \in \mathbb{N}^{I \times J}$ the expected matrix E can be written as the rank 1 outer product between the row and column sums (dividing by M). The centered matrix is then right normalized to account for the unequal sampling depth, dividing by the square root of the column sums to appropriately normalize, $W_n^{1/2}$, yielding the centered and right normalized matrix $\tilde{X}$. Multiplying from the left by $\boldsymbol{f}$ yields a deviation statistic for each column, which is aggregated in a $\boldsymbol{c}$-weighted sum to yield the final test statistic S.

To analyze OASIS's test statistic, we define the random variables $\{Z_{j,k}\}$ for $j = 1, \ldots, J$ and $k = 1, \ldots n_j$. $Z_{j,k}$ denotes the row identity of the k-th observation from the j-th column, meaning that $Z_{j,k} \in [I] = \{1, 2, \ldots, I\}$. The table can then be equivalently constructed from the $\{Z_{j,k}\}$ by taking $X_{i,j} = \sum_{k=1}^{n_j} \mathbb{1}\{Z_{j,k} = i\}$. Under the null hypothesis, $\{Z_{j,k}\}$ are all independently drawn from the same probability distribution $\boldsymbol{p}$ over $[I]$. This ordering $k = 1, \ldots, n_j$ represents a random ordering of the counts observed. This reformulation allows OASIS's test statistic to be expressed as a weighted sum of independent random variables, enabling the use of classical concentration inequalities.

A. Proof of original p-value. The original p-value bound proposed in (1) is:

$$\mathbb{P}\left(|S(\boldsymbol{f}, \boldsymbol{c})| \geq s\right) = 2\exp\left(-\frac{2(1-\xi)^2 s^2}{\sum_j c_j^2}\right) + 2\exp\left(-\frac{2\xi^2 M s^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}\right)$$
$$\text{where } \xi = \left(1 + \sqrt{\frac{M \sum_j c_j^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}}\right)^{-1}. \quad [9]$$

Using the quantity γ defiend in this paper, this can be expressed as:

$$\mathbb{P}\left(|S(\boldsymbol{f}, \boldsymbol{c})| \geq s\right) = 2\exp\left(-\frac{2(1-\xi)^2 s^2}{\|\boldsymbol{c}\|^2}\right) + 2\exp\left(-\frac{2\xi^2 s^2}{\|\boldsymbol{c}\|^2 \gamma}\right)$$
$$\text{where } \xi = \left(1 + \frac{1}{\sqrt{\gamma}}\right)^{-1}. \quad [10]$$

The p-value bound in (1) can be proved in the following manner. First, we estimate the expectation (unconditional on sample identity) of $\boldsymbol{f}$ on the observations as $\bar{\mu}$. Then, we estimate the expectation of $\boldsymbol{f}$ for each column j separately, as $\hat{\mu}_j$. Then, we construct an estimate for the deviation of a column j from the table average by constructing $S_j = \sqrt{n_j}(\hat{\mu}_j - \bar{\mu})$, normalizing by $\sqrt{n_j}$ to ensure that each S_j will have essentially constant variance (up to the correlation between $\bar{\mu}$ and $\hat{\mu}_j$). Finally, we construct our overall test statistic S as the $\boldsymbol{c}$-weighted sum of the S_j, i.e. $S = \sum_j c_j S_j$. In summary:

$$\bar{\mu} = \frac{1}{M} \sum_{i,j} f_i X_{i,j} = \frac{1}{M} \sum_{j,k} f_{Z_{j,k}} \quad [11]$$

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i,j} f_i X_{i,j} = \frac{1}{n_j} \sum_{k=1}^{n_j} f_{Z_{j,k}} \quad [12]$$

$$S_j = \sqrt{n_j}(\hat{\mu}_j - \hat{\mu}) \quad [13]$$

$$S = \sum_j c_j S_j \quad [14]$$

To provide a p-value, we use Hoeffding's inequality:

Lemma 1 (Hoeffding's inequality, Prop. 2.1 (24))**.** *Suppose that the random variables X_i, $i = 1, \ldots, n$ are independent, and X_i has mean μ_i and $X_i \in [a_i, b_i]$. Then for all $t \geq 0$ we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2}\right).$$

Taking the true (unknown) distribution of the $\{Z_{j,k}\}$ under the null to be $Z_{j,k} \sim \boldsymbol{p}$, we define $\mu \triangleq \mathbb{E}[f_{D_{1,1}}]$. Then, as in (1),

$$\begin{aligned}
\mathbb{P}(|S(\boldsymbol{f},\boldsymbol{c})| \geq \epsilon) &\leq \mathbb{P}\left(\left|\sum_{j,k} c_j \frac{f_{Z_{j,k}} - \bar{\mu}}{\sqrt{n_j}}\right| \geq \epsilon\right) \\
&= \mathbb{P}\left(\left|\sum_{j,k} c_j \frac{f_{Z_{j,k}} - \mu}{\sqrt{n_j}} + (\mu - \bar{\mu}) \sum_j c_j \sqrt{n_j}\right| \geq \epsilon\right) \\
&\leq \min_{\xi \in (0,1)} \left[\mathbb{P}\left(\left|\sum_{j,k} c_j \frac{f_{Z_{j,k}} - \mu}{\sqrt{n_j}}\right| \geq (1-\xi)\epsilon\right) + \mathbb{P}\left(\left|(\mu - \bar{\mu}) \sum_j c_j \sqrt{n_j}\right| \geq \xi\epsilon\right)\right] \\
&\overset{(a)}{=} \min_{\xi \in (0,1)} \left[\mathbb{P}\left(\left|\sum_{j,k} \frac{c_j}{\sqrt{n_j}} \left(f_{Z_{j,k}} - \mu\right)\right| \geq (1-\xi)\epsilon\right) + \mathbb{P}\left(\left|\frac{1}{M} \sum_{j,k} \left(f_{Z_{j,k}} - \mu\right)\right| \geq \frac{\xi\epsilon}{\sum_j c_j \sqrt{n_j}}\right)\right] \\
&\overset{(b)}{\leq} \min_{\xi \in (0,1)} \left[2\exp\left(-\frac{2(1-\xi)^2\epsilon^2}{\sum_{j,k} \frac{c_j^2}{n_j}}\right) + 2\exp\left(-\frac{\frac{2\xi^2 M^2 \epsilon^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}}{M}\right)\right] \\
&= \min_{\xi \in (0,1)} \left[2\exp\left(-\frac{2(1-\xi)^2\epsilon^2}{\sum_{j:n_j>0} c_j^2}\right) + 2\exp\left(-\frac{2\xi^2 M \epsilon^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}\right)\right]
\end{aligned}$$

The first inequality comes from a union bound (if neither of the two conditions are met, then $|S| < \epsilon$. (a) assumes that the denominator $\sum_j c_j \sqrt{n_j} \neq 0$, as otherwise ξ can simply be taken to be 0. In (b) we utilize Hoeffding's inequality on the two terms; in the first we note that the j,k-th term, $\frac{c_j}{\sqrt{n_j}}(f_{Z_{j,k}} - \mu)$ is indeed 0 mean, and is bounded as $\mu \in [0,1]$ is simply a constant offset that shifts both the min and max by the same amount (retaining the dynamic range of $f_{Z_{j,k}} \in [0,1]$). Concretely,

$$Y_{j,k} \triangleq \frac{c_j}{\sqrt{n_j}}(f_{Z_{j,k}} - \mu) \in \left[\frac{-|c_j \mu|}{\sqrt{n_j}}, \frac{|c_j(1-\mu)|}{\sqrt{n_j}}\right] \implies \max Y_{j,k} - \min Y_{j,k} = \frac{|c_j|}{\sqrt{n_j}}. \qquad [15]$$

Similarly, for the second term, each summand has a range of 1.

This bound can easily be optimized over $\xi \in (0,1)$ to within a factor of 2 of optimum by equating the two terms, which is achieved when:

$$\xi = \left(1 + \sqrt{\frac{M \sum_{j:n_j>0} c_j^2}{\sum_j c_j \sqrt{n_j}}}\right)^{-1}. \qquad [16]$$

This leads to the stated p-value bound in (1) of

$$\mathbb{P}\left(|S(\boldsymbol{f},\boldsymbol{c})| \geq s\right) = 2\exp\left(-\frac{2(1-a)^2 s^2}{\sum_{j:n_j>0} c_j^2}\right) + 2\exp\left(-\frac{2a^2 M s^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}\right) \quad \text{with} \quad a = \left(1 + \sqrt{\frac{M \sum_j c_j^2}{\left(\sum_j c_j \sqrt{n_j}\right)^2}}\right)^{-1}. \qquad [17]$$

B. Proof of improved p-values. Again, we utilize define $\bar{\mu} = \frac{1}{M}\boldsymbol{f}^\top X \mathbf{1}$ as the estimate of $\mu = \mathbb{E}_{D \sim \boldsymbol{p}}[f_D]$, where $\boldsymbol{p}$ is the unknown common row distribution under the null. Then, for fixed $\boldsymbol{f}, \boldsymbol{c}$, our test statistic can be simplified as

$$\begin{aligned}
S(\boldsymbol{f},\boldsymbol{c}) &= \sum_j c_j \sqrt{n_j}\left(\hat{\mu}_j - \bar{\mu}\right) \\
&= \left[\sum_j c_j \sqrt{n_j}\left(\frac{1}{n_j}\sum_{k=1}^{n_j} f_{Z_{j,k}}\right)\right] - \left[\sum_j c_j \sqrt{n_j}\left(\frac{1}{M}\sum_{\ell k} f_{D_{\ell k}}\right)\right] \\
&= \sum_{jk} \frac{c_j}{\sqrt{n_j}} f_{Z_{j,k}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M} \sum_{jk} f_{Z_{j,k}} \\
&= \sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right) f_{Z_{j,k}} \\
&= \sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right) \left(f_{Z_{j,k}} - \mu\right).
\end{aligned} \qquad [18]$$

Note that, ignoring $f_{Z_{j,k}}$, this summation sums to 0 (as constructed, to allow for the test statistic to have mean 0 independent of $\boldsymbol{f}$ and D). This allows us to subtract the mean $\mu \triangleq \mathbb{E}_{D\sim p}[f_D]$ inside the summation in the last line, yielding a sum of mean 0 terms.

Note that if $\boldsymbol{c} = \sqrt{\boldsymbol{n}/M}$ or $\boldsymbol{c} = -\sqrt{\boldsymbol{n}/M}$ (equivalently, if $\gamma = 1$), then each coefficient will be equal to 0, i.e. $\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right) = 0$ for all j, k. In this case, $S(\boldsymbol{f}, \boldsymbol{c}) = 0$ with probability 1.

If $\gamma < 1$, then examining each term in the summation in Eq. (18), we can see that

$$Y_{j,k} \triangleq \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right) \implies \max Y_{j,k} - \min Y_{j,k} \leq \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right| \times \left(\max_i f_i - \min_i f_i\right) \qquad [19]$$

Thus, observing that each $Y_{j,k}$ has mean 0, and is bounded as above, we can apply Hoeffding's inequality to the sum of these $Y_{j,k}$. For the simplification below we consider the case where $\max_i f_i - \min_i f_i = 1$, but this is without loss of generality as dividing by $\max_i f_i - \min_i f_i$ simply rescales ϵ. Note that this condition of $\gamma < 1$ ensures that the denominator is nonzero.

$$\begin{aligned}
\mathbb{P}\left(|S(\boldsymbol{f}, \boldsymbol{c})| \geq \epsilon\right) &= \mathbb{P}\left(\left|\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)\right| \geq \epsilon\right) \\
&\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\sum_{j:n_j>0}\left(c_j - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\sqrt{n_j}\right)^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\sum_{j:n_j>0}\left[c_j^2 - 2c_j\sqrt{n_j}\frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M} + \frac{\left(\sum_\ell c_\ell \sqrt{n_\ell}\right)^2}{M^2} n_j\right]}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\sum_j c_j \sqrt{n_j}\right)^2}\right). \qquad [20]
\end{aligned}$$

The only inequality used is Hoeffding's; the rest is simplification and manipulation.

This improves upon the previous result, due to the smaller denominator. It also follows much more similarly to the asymptotic normality argument, where the variance of S appears in the denominator, up to the variance of $f_{Z_{j,k}}$. Essentially, this finite-sample valid bound takes the same form as the asymptotic bound, but upper bounds the variance of $f_{Z_{j,k}}$ to be at most $\frac{1}{4}$.

Thus, our final p-value bound is

$$\begin{aligned}
\mathbb{P}\left(|S(\boldsymbol{f}, \boldsymbol{c})| \geq s\right) &= 2\exp\left(-\frac{2s^2}{\left(\max_i f_i - \min_i f_i\right)^2\left(\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\sum_j c_j \sqrt{n_j}\right)^2\right)}\right) \\
&= 2\exp\left(-\frac{2s^2}{\left(\max_i f_i - \min_i f_i\right)^2 \|\boldsymbol{c}\|^2 (1-\gamma)}\right)
\end{aligned}$$

We state this general proposition below, noting that Proposition 1 follows as an immediate corollary.

Proposition 3. *For any fixed $\boldsymbol{f} \in \mathbb{R}^I$ and $\boldsymbol{c} \in \mathbb{R}^J$, if $\gamma = 0$ then $S = 0$ with probability 1. If $\gamma < 1$, then*

$$\mathbb{P}\left(|S(\boldsymbol{f}, \boldsymbol{c})| \geq s\right) \leq 2\exp\left(-\frac{s^2}{2\|\boldsymbol{f}\|_\infty^2 \|\boldsymbol{c}\|_2^2 (1-\gamma)}\right),$$

where $\|\boldsymbol{f}\|_\infty$ can be tightened to $(\max_i f_i - \min_i f_i)/2$.

C. Asymptotic distribution of OASIS test-statistic. Here, we utilize the Lyapnuov Central Limit Theorem to show that OASIS's test statistic is asymptotically normally distributed.

Theorem 1 (Theorem 27.3 (25))**.** *Suppose that for each M the sequence $X_{M1}, \ldots, X_{MM}$ is independent and satisfies*

$$\mathbb{E}[X_{Mk}] = 0, \quad \sigma_{Mk}^2 = \mathbb{E}[X_{Mk}^2], \quad s_M^2 = \sum_{k=1}^{M} \sigma_{Mk}^2$$

for all M, k, where the means and variances are assumed to be finite, and $s_M^2 > 0$ for large M. Defining $S_M = \sum_{k=1}^{M} X_{Mk}$, if

$$\lim_{M} \sum_{k=1}^{M} \frac{1}{s_M^{2+\delta}} \mathbb{E}\left[|X_{Mk}|^{2+\delta}\right] = 0$$

holds for some positive δ (Lyapunov's condition), then $S_M/s_M \xrightarrow{D} Z$ where $Z \sim \mathcal{N}(0,1)$.

Proof. Continuing from Eq. (18), the test-statistic S can be expressed as

$$X_{M\tau(j,k)} = \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)$$

$$S(\boldsymbol{f}, \boldsymbol{c}) = \sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right) = \sum_{\tau=1}^{M} X_{M\tau},$$

for an appropriate reindexing τ which maps jk indices to $[M]$. Since $f_i \in [0,1]$ for all i, each $X_{M\tau}$ has finite variance, and has mean 0 due to the centering.

Computing the variance of S_M, which is s_M^2 as the terms in S_M have mean 0 and are independent, yields

$$\begin{aligned} s_M^2 &= \sum_{jk} \mathbb{E}\left[\left(\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)\right)^2\right] \\ &= \sigma_{\boldsymbol{f}}^2 \sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)^2 \\ &= \sigma_{\boldsymbol{f}}^2 \left(\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\sum_\ell c_\ell \sqrt{n_\ell}\right)^2\right). \end{aligned} \tag{21}$$

Note that this is greater than 0 by assumption in the Proposition statement (γ is bounded away from 1). Recall that γ is defined as below

$$\gamma = \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|}, \sqrt{\frac{\boldsymbol{n}}{M}} \right\rangle^2,$$

implying that our desired result is

$$\frac{1}{\sqrt{1-\gamma}} S(\boldsymbol{f}, \boldsymbol{c}) \xrightarrow{D} \mathcal{N}\left(0, \|\boldsymbol{c}\|^2 \sigma_{\boldsymbol{f}}^2\right). \tag{22}$$

Satisfying Lyapunov's condition: we show that this condition holds for $\delta = 1$, which entails bounding third moments. Note that $|f_{Z_{j,k}} - \mu| \leq 1$ a.s., and so $\mathbb{E}[|f_{Z_{j,k}} - \mu|^3] \leq \mathbb{E}[|f_{Z_{j,k}} - \mu|^2] = \sigma_{\boldsymbol{f}}^2$. By Cauchy-Schwarz, even the largest coefficient is vanishing:

$$\left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right| \leq \left|\frac{c_j}{\sqrt{n_j}}\right| + \frac{\|\boldsymbol{c}\|}{\sqrt{M}} \leq \frac{2\|\boldsymbol{c}\|}{\min_j \sqrt{n_j}}. \tag{23}$$

Using this, we can upper bound the sum of the third moments:

$$\begin{aligned} \sum_{jk} \mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)\right|^3\right] &= \sum_{jk} \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right|^3 \mathbb{E}\left[\left|f_{Z_{j,k}} - \mu\right|^3\right] \\ &\leq \sum_{jk} \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right|^3 \sigma_{\boldsymbol{f}}^2 \\ &\leq \left(\frac{2\|\boldsymbol{c}\|}{\min_j \sqrt{n_j}}\right) \sigma_{\boldsymbol{f}}^2 \left(\sum_{jk} \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right|^2\right). \end{aligned} \tag{24}$$

Verifying Lyapunov's condition:

$$\begin{aligned}
\frac{1}{s_n^3}\sum_{jk}\mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}}-\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)f_{Z_{j,k}}\right|^3\right] &= \frac{\sum_{jk}\mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}}-\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)f_{Z_{j,k}}\right|^3\right]}{\left(\sum_{jk}\mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}}-\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)f_{Z_{j,k}}\right|^2\right]\right)^{3/2}} \\
&\leq \frac{\left(\frac{2\|\boldsymbol{c}\|}{\min_j\sqrt{n_j}}\right)\sigma_{\boldsymbol{f}}^2\left(\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}}-\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)^2\right)}{\left(\sigma_{\boldsymbol{f}}^2\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}}-\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)^2\right)^{3/2}} \\
&= \frac{2\|\boldsymbol{c}\|}{\min_j\sqrt{n_j}}\left(\sigma_{\boldsymbol{f}}^2\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}}-\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)^2\right)^{-1/2} \\
&= \frac{2\|\boldsymbol{c}\|}{\sigma_{\boldsymbol{f}}\min_j\sqrt{n_j}}\left(\|\boldsymbol{c}\|^2(1-\gamma)\right)^{-1/2} \\
&= \frac{2}{\sigma_{\boldsymbol{f}}\sqrt{(1-\gamma)\min_j n_j}} \qquad [25]
\end{aligned}$$

We can see that the last line goes to 0 as all $n_j \to \infty$ simultaneously as long as $\sigma_{\boldsymbol{f}} > 0$ and $\gamma < 1$, which were assumed in the Theorem statement, and thus by the Lyapunov CLT $S(\boldsymbol{f},\boldsymbol{c})$ we have the desired result. The variance of $S(\boldsymbol{f},\boldsymbol{c})$ has already been computed as s_M^2, and so

$$\frac{1}{\sqrt{1-\gamma}}S(\boldsymbol{f},\boldsymbol{c}) \xrightarrow{D} \mathcal{N}\left(0, \|\boldsymbol{c}\|^2\sigma_{\boldsymbol{f}}^2\right). \qquad [26]$$

Note that if $\sigma_{\boldsymbol{f}} = 0$ then $S = 0$ w.p. 1, which matches this result under the convention that a Gaussian with variance 0 is a Dirac point mass at 0. □

With this proposition, we can construct an asymptotically valid p-value bound as

$$\mathbb{P}(|S(\boldsymbol{f},\boldsymbol{c})| \geq s) = 2\Phi\left(-\frac{s}{\hat{\sigma}_{\boldsymbol{f}}\|c\|\sqrt{1-\gamma}}\right) \leq 2\exp\left(-\frac{s^2}{2\hat{\sigma}_{\boldsymbol{f}}^2\|\boldsymbol{c}\|^2(1-\gamma)}\right), \qquad [27]$$

where Φ is the CDF of a standard Gaussian random variable. This holds by the continuous mapping theorem, as $\hat{\sigma}_{\boldsymbol{f}}^2$ is an asymptotically consistent estimator of $\sigma_{\boldsymbol{f}}$. Note that this form precisely matches that of the finite-sample p-value bound in Proposition 1, up to bounding $\hat{\sigma}_{\boldsymbol{f}} \leq \frac{1}{4}$.

D. Effect Size. This effect size is motivated by a simple two group alternative. Consider the case where our samples originate from two groups, A, and B. Each group has a characteristic row distribution, $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$ respectively, where observations from a sample in group A are drawn i.i.d. (independent and identically distributed) from $\boldsymbol{p}_A$ (similarly with B and $\boldsymbol{p}_B$). Then, denoting D as the identity of the random row drawn from $\boldsymbol{p}_A$ or $\boldsymbol{p}_B$, the effect size estimate $\hat{\Delta}$ is nothing but the plug-in estimate of $\Delta = \left|\mathbb{E}_{D\sim\boldsymbol{p}_A}\left[f_D\right] - \mathbb{E}_{D\sim\boldsymbol{p}_B}\left[f_D\right]\right|$ the difference between $\mathbb{E}\left[f_D\right]$ where $D \sim \boldsymbol{p}_A$ or $D \sim \boldsymbol{p}_B$. This is an intuitive measure of discrepancy, as OASIS declaring a discovery based on a specific $\boldsymbol{f}$ and $\boldsymbol{c}$ indicates that when partitioned into groups by $\boldsymbol{c}$, the two groups of samples have significantly different row distributions as measured by $\boldsymbol{f}$. This measure Δ is nonnegative, and can be naturally bounded above by the total variation distance between the empirical row distributions of columns where $c > 0$ and where $c < 0$. Taking the supremum over all $\boldsymbol{f} \in [0,1]^I$, this effect size measure takes a value of 0 if and only if the empirical row distributions are the same between the two clusters, and 1 if they're disjoint.

Proof of Eq. (6). The lower bound on $\hat{\Delta}$ follows trivially from the absolute value. The upper bound follows from the reformulation of $\hat{\Delta}$ as:

$$\hat{\Delta} = \left|\sum_i(\hat{p}_{+,i} - \hat{p}_{-,i})f_i\right| \leq \max_{g\in[0,1]^I}\left|(\hat{\boldsymbol{p}}_+ - \hat{\boldsymbol{p}}_-)^\top g\right| = \frac{1}{2}\|\hat{\boldsymbol{p}}_+ - \hat{\boldsymbol{p}}_-\|_1 = \delta_{\mathrm{TV}}(\hat{\boldsymbol{p}}_+, \hat{\boldsymbol{p}}_-)$$

The evaluation into the ℓ_1 norm can be seen by the fact that since $\hat{\boldsymbol{p}}_+$ and $\hat{\boldsymbol{p}}_-$ are probability distributions, shifting g by a constant does not change the objective value. Thus, we can equivalently maximize over $g \in [-1/2, 1/2]^I$. Then, for any vector x,

$$\max_{g:\|g\|_\infty\leq 1} g^\top x = \|x\|_1, \qquad [28]$$

by the dual norm characterization of the ℓ_1 norm. □

This effect size is a plug-in estimate for the two-group setting quantity of

$$\Delta \triangleq \left|\mathbb{E}_{D\sim \boldsymbol{p}_A}\left[f_D\right] - \mathbb{E}_{D\sim \boldsymbol{p}_B}\left[f_D\right]\right|. \tag{29}$$

D.1. Beyond binary c. There is no clear extension of this effect size measure beyond binary $\boldsymbol{c}$. We discuss some candidate measures and their drawbacks below. For simplicity, in this section we consider $\boldsymbol{f} \in [-1,1]^I$ instead of $\boldsymbol{f} \in [0,1]^I$, to avoid recentering. One candidate measure which encourages binary $\boldsymbol{f}$ is:

$$\hat{\Delta} = \left|\frac{\boldsymbol{f}^\top X \boldsymbol{c}}{\boldsymbol{1} X |\boldsymbol{c}|}\right|. \tag{30}$$

A more gradual and natural alternative which allows for non-extremized $\boldsymbol{f}$ is:

$$\hat{\Delta} = \left|\frac{\boldsymbol{f}^\top X \boldsymbol{c}}{|\boldsymbol{f}|^\top X |\boldsymbol{c}|}\right|. \tag{31}$$

These quantities are both bounded between 0 and 1. The first attains a value of 1 only if $f_i X_{i,j} c_j = X_{i,j}|c_j|$ (extending the range of $\boldsymbol{f}$ to $[-1,1]$). The second requires that $f_i X_{i,j} c_j = |f_i| X_{i,j} |c_j|$. However, note that this can be trivially achieved by taking $c_j = 0$ for all except one sample, and taking $\boldsymbol{f}$ to be 1 for all rows, or even more easily by taking $\boldsymbol{f}$ and $\boldsymbol{c}$ to both be all ones vectors. This will not yield a significant p-value, however, due to the structure of $\tilde{X}$. Note that a similar approach can be taken for binary c_j, but then it runs into the above problems.

11. Optimization primitive

Recall the p-value bound derived in Proposition 1 for the test statistic $S = \boldsymbol{f}^\top \tilde{X} \boldsymbol{c}$, restated below for convenience:

$$2\exp\left(-\frac{2S^2}{\|\boldsymbol{c}\|^2(1-\gamma)}\right). \tag{32}$$

A. Maximizing over $\boldsymbol{f}$. OASIS's p-value is monotonically decreasing in $|S|$, the test statistic, so the vector $\boldsymbol{f}$ that minimizes this finite-sample valid p-value is also the one that maximizes $|S|$, as Eq. (32) only depends on $\boldsymbol{f}$ through S. For fixed $\boldsymbol{c}$, the optimal $\boldsymbol{f}^*$ can be computed as

$$\boldsymbol{f}^* = \operatorname*{argmax}_{0\le \boldsymbol{f} \le 1} \left|\boldsymbol{f}^\top \tilde{X} \boldsymbol{c}\right|. \tag{33}$$

Defining $\boldsymbol{v} = \tilde{X}\boldsymbol{c}$, Eq. (33) is optimized by either

$$f_i^\star = \begin{cases} 1 & \text{if } v_i > 0, \\ 0 & \text{if } v_i \le 0. \end{cases} \quad \text{or} \quad \boldsymbol{f}_t^\star = \begin{cases} 1 & \text{if } v_i < 0, \\ 0 & \text{if } v_i \ge 0. \end{cases} \tag{34}$$

Comparing the objective values of these two solutions yields an optimal solution which is unique up to entries in $\boldsymbol{v}$ being 0, and the objective values of these two solutions being the same.

Note that due to the rescaling of each point's contribution by $\sqrt{n_j}$ (downweighting), this will not necessarily yield an effect size equal to the total variation distance between the two empirical distributions. In the two group setting this can be seen as constructing the empirical distribution for each cluster after downweighting the counts from each sample by $1/\sqrt{n_j}$, and the maximizing $\boldsymbol{f}$ is 1 when the $c_j = +1$ cluster's frequency is larger and -1 when the $c_j = -1$ cluster's frequency is larger.

B. Maximizing over $\boldsymbol{c}$. Maximizing over $\boldsymbol{c}$ requires an extra step as $\boldsymbol{c}$ appears in both the numerator and denominator of the p-value bound in Eq. (32). If we consider the case when γ is negligible then this term of $(1-\gamma)$ in the denominator can be ignored. This is practically the case when we are performing inference, as $\boldsymbol{c}$ will be balanced. Then, this reduces the problem to minimizing the exponent, i.e. maximizing $S^2/\|\boldsymbol{c}\|^2$. In this setting we obtain, defining $S_j = \left[\tilde{X}^\top \boldsymbol{f}\right]_j$:

$$\boldsymbol{c}^* = \operatorname*{argmax}_{\|\boldsymbol{c}\|_2 \le 1} \left|\sum_j c_j S_j\right| \implies c_j^* \propto S_j. \tag{35}$$

This means that it is not necessarily optimal to select $c_j = \pm 1$. The intuition behind this is that if certain samples have extremely large S_j, then we want to assign them c_j close to 1. If other samples have only moderately large S_j, we still wish to assign them positive c_j, but we are not as confident in their binary cluster assignment and so assign them only a moderate c_j. And, if a sample has S_j close to 0, then it is not identifying well with the $c_j = +1$ or $c_j = -1$ clusters, and thus should be assigned a value of c_j close to 0. Thus, a set of optimal c_j can be constructed as $c_j^* = \frac{S_j}{\|\vec{S}\|_2}$.

Note that due to the non-convexity of the objective, the starting initialization matters. One candidate approach is randomly initializing $\boldsymbol{c}^{(0)}$ and then performing alternating maximization from this random initialization. Note that this can naturally be extended to start from several random initializations, and use the best performing initialization.

C. Factor of $1-\gamma$. The p-value bound optimizing $\boldsymbol{f}, \boldsymbol{c}$ are given as:

$$\underset{\boldsymbol{f}\in[0,1]^I, \|\boldsymbol{c}\|\leq 1}{\operatorname{argmin}} 2\exp\left(-\frac{2S^2}{1-\gamma}\right) = \underset{\boldsymbol{f}\in[0,1]^I, \|\boldsymbol{c}\|\leq 1}{\operatorname{argmax}} \frac{S^2}{1-\gamma} \tag{36}$$

Since $\gamma \in [0,1]$, the optimal solution in Eq. (36) has value upper bounded as below:

$$\max_{\boldsymbol{f}\in[0,1]^I, \|\boldsymbol{c}\|\leq 1} \frac{S^2}{1-\gamma} \leq \max_{\boldsymbol{f}\in[0,1]^I, \|\boldsymbol{c}\|\leq 1} S^2. \tag{37}$$

We show that the optimizing $\boldsymbol{c}$ we obtain yields $\gamma = 0$, as at a local maxima (for fixed $\boldsymbol{f}$, $\boldsymbol{c}$ is optimized as being proportional to $\tilde{X}^\top \boldsymbol{f}$:

$$\boldsymbol{c} \propto \operatorname{diag}(1/\sqrt{\boldsymbol{n}}) \left(X^\top - \frac{1}{M} X^\top \mathbf{1}\mathbf{1}^\top X^\top\right) \boldsymbol{f}, \tag{38}$$

and so

$$\begin{aligned}
\gamma &= \left\langle \frac{\boldsymbol{c}}{\|\boldsymbol{c}\|}, \sqrt{\frac{\boldsymbol{n}}{M}} \right\rangle^2 \\
&= \frac{1}{M\|\boldsymbol{c}\|^2} \left\langle \operatorname{diag}(1/\sqrt{\boldsymbol{n}}) \left(X^\top - \frac{1}{M} X^\top \mathbf{1}\mathbf{1}^\top X^\top\right) \boldsymbol{f}, \sqrt{\boldsymbol{n}} \right\rangle^2 \\
&= \frac{1}{M\|\boldsymbol{c}\|^2} \left(\boldsymbol{f}^\top \left(X - \frac{1}{M} X \mathbf{1}\mathbf{1}^\top X\right) \mathbf{1}\right)^2 \\
&= 0.
\end{aligned}$$

D. Iterative testing of contingency tables. Recalling the initial intuition for the OASIS test, where there are two groups of samples each with a characteristic probability distribution over the rows, the optimal $\boldsymbol{c}$ vector in this case partitions the samples by type (positive on one group of samples and negative on the other). However, these scalar-valued vectors $\boldsymbol{c}$ are limited by their one dimensionality, and so a natural question is whether OASIS can identify more than 2 clusters within a table. In genomics contexts this translates to detecting subclusters or multiple cell-types.

Considering the matrix viewpoint of OASIS, the test statistic can be thought of as determining whether the centered and right normalized contingency table has a sufficiently large principle singular value. We show that a simple extension of OASIS allows for detection of how many ways the contingency table significantly deviates from the null, with a statistically-valid stopping condition.

Consider that one initial partitioning vector $\boldsymbol{c}^{(0)}$ is identified, for example distinguishing Delta vs Omicron samples in the running SARS-CoV-2 example, and we want to determine if there are additional statistically significant ways of partitioning this data not along this same direction $\boldsymbol{c}^{(0)}$. One way of accomplishing this is by attempting to identify new $\boldsymbol{c}, \boldsymbol{f}$ which yield a significant p-value, requiring that this new vector $\boldsymbol{c}$ be orthogonal to the previously found vector $\boldsymbol{c}^{(0)}$. In the previous example, this could be a vector separating Omicron BA.1 and BA.2. The constraint $\boldsymbol{c}^\top \boldsymbol{c}^{(0)} = 0$ maintains the convexity of the optimization problem.

This naturally generalizes to multiple orthogonality constraints where $\boldsymbol{c}$ is restricted to be orthogonal to a subspace. As discussed, OASIS's initial test statistic can be thought of as identifying one direction of substantial signal. Restricting this new $\boldsymbol{c}$ to be orthogonal to the first one is, loosely speaking, identifying the second largest singular value and its corresponding right singular vector. Thus, this iterative procedure of finding $\boldsymbol{c}$ orthogonal to all previously found $\boldsymbol{c}$ and terminating when this no longer yields a significant p-value, can be thought of as looking at the spectrum of $\tilde{X}$ until a significant dropoff occurs.

In this iterative setting, the objective of interest is

$$\boldsymbol{c}^\star = \underset{\substack{\|\boldsymbol{c}\|_2\leq 1, \\ \boldsymbol{c}^\top \boldsymbol{v}=0}}{\operatorname{argmax}} \left|\boldsymbol{f}^\top \vec{S}\right| \implies c_j^* \propto S_j - \frac{v^\top \vec{S}}{\|v\|_2^2} v_j. \tag{39}$$

E. Metadata-based $\boldsymbol{c}$. In some settings we may be given additional information regarding the samples (e.g. cell-type information) in the form of a vector $\tilde{\boldsymbol{c}} \in [-1,1]^J$. In the simplest case, we fully trust this metadata, and want to see if this sample partitioning yields a significant p-value. In this proposed method, we utilize exactly $\tilde{\boldsymbol{c}}$, find the optimizing $\boldsymbol{f}^\star$ from the split data, and return this pair for inference.

One additional approach of potential interest is if metadata $\tilde{\boldsymbol{c}}$ is given, but we are still free to reweight samples: not all samples are equally representative. If we require that each c_j retains the same sign as $\tilde{c}_j$, then we can formulate this as:

$$\boldsymbol{c}^\star = \underset{\substack{c\,:\,\|c\|\leq 1, \\ c_j\tilde{c}_j\geq 0\ \forall\ j}}{\operatorname{argmax}} \left|\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}\right|. \tag{40}$$

While the overall objective is non-convex, it is still convex in $\boldsymbol{f}$ for fixed $\boldsymbol{c}$, and convex in $\boldsymbol{c}$ for fixed $\boldsymbol{f}$. The maximization of $\boldsymbol{f}$ is the same as before, and $\boldsymbol{c}$ is now maximized as, defining $S_j = [\tilde{X}^\top \boldsymbol{f}]_j$:

$$c_j \propto \begin{cases} S_j & \text{if } \tilde{c}_j S_j \geq 0, \\ 0 & \text{if } \tilde{c}_j S_j < 0. \end{cases} \quad \text{or} \quad c_j \propto \begin{cases} -S_j & \text{if } \tilde{c}_j S_j \leq 0, \\ 0 & \text{if } \tilde{c}_j S_j > 0. \end{cases} \tag{41}$$

To optimize the overall truncated objective, we can still run alternating maximization.

F. Asymptotic p-value bound. The asymptotic p-value bound provided in Proposition 2 leads to the following maximization objective

$$\max_{\boldsymbol{f},\boldsymbol{c}} \frac{\left(\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}\right)^2}{\|\boldsymbol{c}\|^2 \hat{\sigma}_{\boldsymbol{f}}^2} \tag{42}$$

where $\hat{\sigma}_{\boldsymbol{f}}^2 = \sum_i \hat{p}_i f_i^2 - \left(\sum_i \hat{p}_i f_i\right)^2$ and $\hat{\boldsymbol{p}} \triangleq \frac{1}{M} X\mathbf{1}$, considering the case where γ is negligible.

As before, this optimization problem is non-convex, but can be optimized via alternating maximization. For fixed $\boldsymbol{f}$, $\boldsymbol{c}$ is optimized as before, with $\boldsymbol{c} \propto \tilde{X}^\top \boldsymbol{f}$. For fixed $\boldsymbol{c}$, defining $\boldsymbol{v} = \tilde{X}\boldsymbol{c}$, $\boldsymbol{f}$ is optimized as

$$\operatorname*{argmax}_{\boldsymbol{f}} \frac{\left(\boldsymbol{f}^\top \boldsymbol{v}\right)^2}{\sum_i \hat{p}_i f_i^2 - \left(\sum_i \hat{p}_i f_i\right)^2}. \tag{43}$$

Proposition 4. *An optimizing $\boldsymbol{f}$ for* Eq. (43) *can be constructed as $f_1 = v_1/\hat{p}_1$, and*

$$f_i = \frac{v_i}{\hat{p}_i} + \frac{\sum_k v_k}{\hat{p}_1} \tag{44}$$

for $i = 2, \ldots, I$, where $\boldsymbol{v} = \tilde{X}\boldsymbol{c}$ and $\hat{\boldsymbol{p}} = \frac{1}{M} X\mathbf{1}$.

This can be seen by taking the partial derivative with respect to f_i, and noting that the optimizing $\boldsymbol{f}$ is shift invariant, and that the objective value does not change with rescaling of $\boldsymbol{f}$.

Proof. Considering the objective for the asymptotic p-value:

$$\boldsymbol{f}^\star, \boldsymbol{c}^\star = \operatorname*{argmax}_{0 \le \boldsymbol{f} \le 1, \|c\|_2 \le 1} \frac{\left(\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}\right)^2}{\sigma_{\boldsymbol{f}}^2}. \tag{45}$$

Since the numerator and denominator both scale quadratically in the magnitude of $\boldsymbol{f}$, we can instead treat the denominator as a constraint. This entails constraining the denominator $\sigma(\boldsymbol{f}) \triangleq \sum \hat{p}_i f_i^2 - (\sum \hat{p}_i f_i)^2$ to be at most 1, and maximizing the numerator subject to $\sigma(\boldsymbol{f}) \le 1$. Since $\sigma(\boldsymbol{f})$ is a convex function of $\boldsymbol{f}$, we can simply find a stationary point with respect to $\boldsymbol{f}$ and this will be the optimum.

Taking the partial derivative with respect to f_i and equating to 0, we obtain (defining $\mu = \sum \hat{p}_i f_i$)

$$\begin{aligned}
0 = \frac{\partial}{\partial f_i} &= \frac{2(\boldsymbol{f}^\top \boldsymbol{v}) v_i \left(\sum \hat{p}_i f_i^2 - (\sum \hat{p}_i f_i)^2\right) - \left(\boldsymbol{f}^\top \boldsymbol{v}\right)^2 \left[2\hat{p}_i f_i - 2\left(\sum \hat{p}_i f_i\right)\hat{p}_i\right]}{\left(\sum \hat{p}_i f_i^2\right)^2} \\
2(\boldsymbol{f}^\top \boldsymbol{v}) v_i \left(\sum \hat{p}_i f_i^2 - \mu^2\right) &= \left(\boldsymbol{f}^\top \boldsymbol{v}\right)^2 \left[2\hat{p}_i f_i - 2\mu \hat{p}_i\right] \\
v_i \left(\sum p_k f_k^2 - \mu^2\right) &= \left(\boldsymbol{f}^\top \boldsymbol{v}\right) \hat{p}_i \left[f_i - \mu\right] \\
\frac{\hat{p}_i \left(f_i - \mu\right)}{v_i} &= \frac{\sum p_k f_k^2 - \mu^2}{\boldsymbol{f}^\top \boldsymbol{v}}
\end{aligned} \tag{46}$$

We can write out a system of $I+1$ equations in μ and the I unknown f_i's, and solve by setting $f_1 = v_1/\hat{p}_1$ (shift invariant, can be arbitrary), and all the RHS of the above equations to 1 (scale invariant). This yields $\mu = \frac{1}{p_1} \sum v_i$, on top of $f_1 = v_1/\hat{p}_1$, and

$$f_i = \frac{v_i}{\hat{p}_i} + \frac{\sum_k v_k}{p_1} \tag{47}$$

□

G. Comparison with SVD. SVD is frequently used for matrix decomposition and interpretation tasks, but doesn't inherently minimize the p-value bound of Proposition 1 or yield desirable partitionings. Mathematically, the SVD yields $\boldsymbol{f}, \boldsymbol{c}$ which maximize

$$\max_{\boldsymbol{c},\boldsymbol{f}} \frac{\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}}{\|\boldsymbol{f}\|_2 \|\boldsymbol{c}\|_2}, \tag{48}$$

as opposed to the p-value objective of (disregarding the $1-\gamma$ factor)

$$\max_{\boldsymbol{c},\boldsymbol{f}} \frac{\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}}{\|\boldsymbol{f}\|_\infty \|\boldsymbol{c}\|_2}. \tag{49}$$

Note that taking an SVD and setting $\boldsymbol{f}$ to be the left singular vector does not minimize the p-value bound. This is because $\boldsymbol{f}$ has only a boundedness constraint, while the left singular vector has an ℓ_2 constraint. For a fixed $\boldsymbol{c}$, taken to be the right singular vector or otherwise, maximization of the linear program

$$\max_{0 \leq \boldsymbol{f} \leq 1} \left| f^\top \tilde{X} \boldsymbol{c} \right| \qquad [50]$$

is attained at a corner point, and that if entries of $\tilde{X}\boldsymbol{c}$ are not equal to 0 then this optimizing $\boldsymbol{f}$ is unique up to the absolute value ($\boldsymbol{f}$ and $1 - \boldsymbol{f}$ yielding the same objective value). For any $\boldsymbol{f}$ with entries not in $\{0, 1\}$, at least as good of an objective value can be attained by extremizing $\boldsymbol{f}$ according to the entries of $\tilde{X}\boldsymbol{c}$.

Even if the SVD-based $\boldsymbol{f}$ is shifted and extremized to be in $[0, 1]^I$, this is still not optimizing the desired objective. One simple example where an SVD does not yield the desired partitioning is shown in Figure 6.

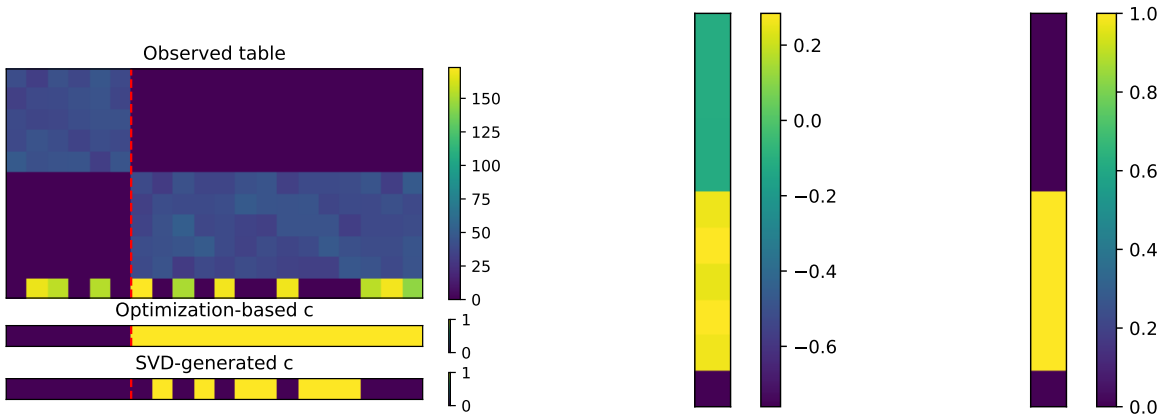

Fig. 6. Failure of SVD approach, which does not identify the visually clear planted structure. On the left the principle left singular vector of the centered and right normalized counts matrix, constituting the $\boldsymbol{f}$ that would be generated by an SVD. On the left, the SVD-based $\boldsymbol{f}$ does identify the row groupings, but utilizing only half the dynamic range for the first 10 rows. On the right, the optimization-based method more clearly identifies the row structure, forcing the bottom row to a value of 0, and utilizing the full dynamic range for the first 10 rows.

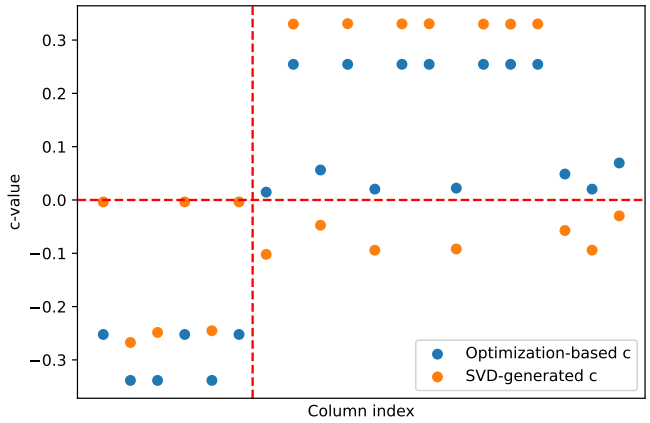

Fig. 7. Plot of $\boldsymbol{c}$ vectors identified by SVD and OASIS optimization-based approach.

Figure 6 shows a planted model, where there are two subgroups of columns, one which expresses rows 1-5, and the other which expresses rows 6-10. The values in these entries is approximately 40. However, there is a confounding final row, which randomly groups the columns into two new groups, and yields counts either 160 or 0. Multinomial sampling provides slight noise to make this more realistic. While both methods do call this table as significant (extreme example chosen for clarity of exposition), an SVD-based approach is unable to identify the latent column structure. This is because of its ℓ_2 normalization of $\boldsymbol{f}$. The latent structure is identified by a vector $\boldsymbol{f}$ which takes value 1 on the first 5 rows, 0 on the next 5, and 1/2 (i.e., disregarding), the final row. However, an SVD restricts to an ℓ_2 constraint, and since the last row has more counts, it is selected as the dominant component in the SVD-based $\boldsymbol{f}$. Taking an SVD, the magnitude of this last component in $\boldsymbol{f}$ is .76, where all other components have magnitude at most .28. These are depicted in more details in Figure 6.

12. Additional discussion

A. Metadata-based $\boldsymbol{f}, \boldsymbol{c}$. OASIS is stated generally, and treats the rows as the set $[I]$ and the columns as the set $[J]$, with no additional information regarding shared structure. However, considering the motivating biological application, rows have additional information (the associated target sequence in $\{A, T, C, G\}^{27}$ (1)). Considering this, one potential direction to construct more biologically meaningful inference is to ensure that similar sequences have similar f_i values. This can be accomplished by constructing $\boldsymbol{f}$ to be Lipschitz with respect to the Levenshtein distance between the targets of the associated rows.

Similarly, in the presence of sample metadata (e.g. cell-type information), the user may wish to incorporate this information into constructing $\boldsymbol{c}$. For example, one may wish to restrict each entry of $\boldsymbol{c}$ to be the same sign as it is in a metadata vector $\tilde{\boldsymbol{c}}$. Alternatively, we may simply wish to use cell-type information to constrain all cells of the same cell-type to have a c_j of the same sign. These, and many other metadata-aided approaches, can be framed as optimization problems that admit efficient alternating maximization solutions.

B. OASIS as a coefficient of correlation. OASIS can also be used to test the dependence between two random variables. Consider observations of two real valued random variables (X, Y), where we observe draws from the joint distribution (X_i, Y_i) and want to test against the null where X and Y are independent. One recent measure of dependence was proposed (26). OASIS can also be used to test against this, by quantizing the random variables (e.g. into deciles), and constructing a contingency table from these categorical (quantized) observations. For any quantization, OASIS can yield a finite-sample valid p-value against the null of independence. This empirically performs quite well, and can yield an effect size estimate comparable to Xicor (26). With no tuning of quantization (fixed at deciles), we show OASIS's effect size and power in Section 14.H.

C. Bernstein-based p-value bounds. The finite sample valid p-value bounds provided in Proposition 1 suffer from the fact that they do not incorporate the variance of f_D where $D \sim \boldsymbol{p}$, and simply rely on the boundedness of f_D. One natural concentration inequality that utilizes both the boundedness of a random variable as well as its variance is (19). An improved p-value bound can potentially be obtained by bounding the true variance $\sigma_{\boldsymbol{f}}^2$ with high probability by some function of the empirical variance $\hat{\sigma}_{\boldsymbol{f}}^2$ on the training set, and utilizing Bernstein's inequality instead of Hoeffding's, applied to the same test statistic.

D. Tensor analysis. In computational genomics, it is of high priority to discover effects or partitionings that are reproducible. In the setting of (1) this motivates going beyond treating each table independently, and instead analyzing the anchor x sample x target tensor. OASIS's framework enables this, with metadata-guided construction of $\boldsymbol{f}, \boldsymbol{c}$, to be performed jointly.

Note that more generally OASIS, with its simple linear structure, suggests a simple test for tensors. If we have a 3 dimensional tensor, we can construct and analyze a test statistic similar to OASIS by centering the tensor, appropriately normalizing by sampling depth, then taking inner products along each of the 3 dimensions to reduce to a scalar.

13. SARS-CoV-2 Analysis details

We used dehosted sequencing data of nasopharyngeal swabs with coinfections available under accession PRJNA817806 (6).

We use the optimization-based p-values with 25% train/test split to compute optimizing $c^\star, f^\star$ vectors, and compute p-values on this held out test set. We then utilize anchors whose BY corrected p-value is less than 0.05, whose effect size (when $c^\star$ is binarized) is in the top 10% of anchors, and have a total number of counts $M > 1000$. This yields 2495 anchors.

We then iterate over these 2495 anchors, and rerun the alternating maximization procedure to generate $c^\star, f^\star$ on the full table.

Note that in addition to the aforementioned thresholds, there are several potential avenues towards filtering that we did not finetune, including how balanced the clustering is (ensure that the table is not simply one deviating column), how similar a $\boldsymbol{c}$ vector is to others (not detect random splits, e.g. due to SNPs), and reducing similar anchors (noting that one single base-pair mutation may yield up to k significant anchors that are one base-pair offsets of each other).

The alignment was performed using bowtie (17). We used 4 publicly available COVID reference genomes for the strains circulating during this time period: the Wuhan reference genome (NCBI assembly NC_045512.2), Delta (OK091006.1), Omicron BA.1 (OX315743.1), and Omicron BA.2 (OX315675.1). A bowtie index comprised of the three COVID genomes was constructed (bowtie-build on a fasta file containing three reads, one for each assembly).

We then construct a fasta file from the anchors passing the above filtering steps, and their targets which comprise $> 5\%$ of the total counts for that anchor. This fasta is aligned to the bowtie index using (-v 0 -a) options to ensure that we only get exact matches, and record all possible matches.

A. $\boldsymbol{c}$ aggregation. To aggregate the $\boldsymbol{c}^{(a)}$ vectors obtained from experiments on different anchors a (different tables), one natural objective is

$$\boldsymbol{c}^\star = \underset{\|\boldsymbol{v}\| \leq 1}{\operatorname{argmax}} \sum_a \langle \boldsymbol{c}^{(a)}, \boldsymbol{v} \rangle^2 = \underset{\|\boldsymbol{v}\| \leq 1}{\operatorname{argmax}} \|\boldsymbol{C}\boldsymbol{v}\|_2 . \quad [51]$$

Here, we stacked the $\boldsymbol{c}^{(a)}$ vectors to obtain a matrix $\boldsymbol{C}$. Analyzing this objective, an optimizing $\boldsymbol{v}$ is attained by the principle right singular vector of $\boldsymbol{C}$. To analyze this further, we also plot the second singular vector, and see that this separates Omicron BA.1 from BA.2.

B. Orthogonal optimization. We can also optimize directly on the data matrices, as discussed, and identify an optimizing $\boldsymbol{c}$ on the combined data matrices, instead of first computing the optimizing $\boldsymbol{c}$ on each table and then aggregating them together. We can then iteratively find orthogonal $\boldsymbol{c}$ vectors, to further partition the samples. To merge the tables, we discard all anchor target pairs that have fewer than 10 counts, and normalize each count by $1/\sqrt{M}$, to ensure that high count tables do not completely outweigh low count ones. This yields a visually similar clustering to $\boldsymbol{c}$-aggregation, but does not perfectly predict with $\boldsymbol{c}^{(1)} \gtrless 0$ whether the sample has Delta or not. There is still perfect classification accuracy on $\boldsymbol{c}^{(1)}$ for predicting whether a sample has Delta or not, and 91% accuracy for predicting from $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ whether a sample has Omicron BA.1 as its primary strain or not (correctly classified 94/103).

14. Additional simulations

For brevity and flow we did not include all numerical results in the main text. We show below additional plots.

A. Comparison of OASIS and X^2. We define a class of alternatives parameterized by their total number of counts M and a corruption parameter ϵ as follows (the number of rows and columns are fixed as 12 and 10 respectively).

We study the pure signal model, where the probability vector for the first half (5) of the samples is the first standard basis vector $\boldsymbol{p}^{(j)} = e_1$ for $j = 1, \ldots, J/2$, and the latter half of the samples have $\boldsymbol{p}^{(j)} = e_2$ for $j = J/2+1, \ldots, J$. To vary between structured and unstructured signal, we mix each columns probability distribution with another probability distribution $\boldsymbol{q}^{(j)}$, which is generated independently for each column. Concretely,

$$\boldsymbol{p}^{(j)} = (1-\epsilon)e_{1+\mathbb{1}\{j>J/2\}} + \epsilon \boldsymbol{q}^{(j)}. \qquad [52]$$

The observed matrix is generated by drawing $X^{(j)} \sim \text{multinomial}(M/J, \boldsymbol{p}^{(j)})$, independently for each of the J columns. Each $\boldsymbol{q}^{(j)}$ was generated by taking each entry to be independently uniformly distributed between $[0,1]$, and then normalized to a valid probability distribution.

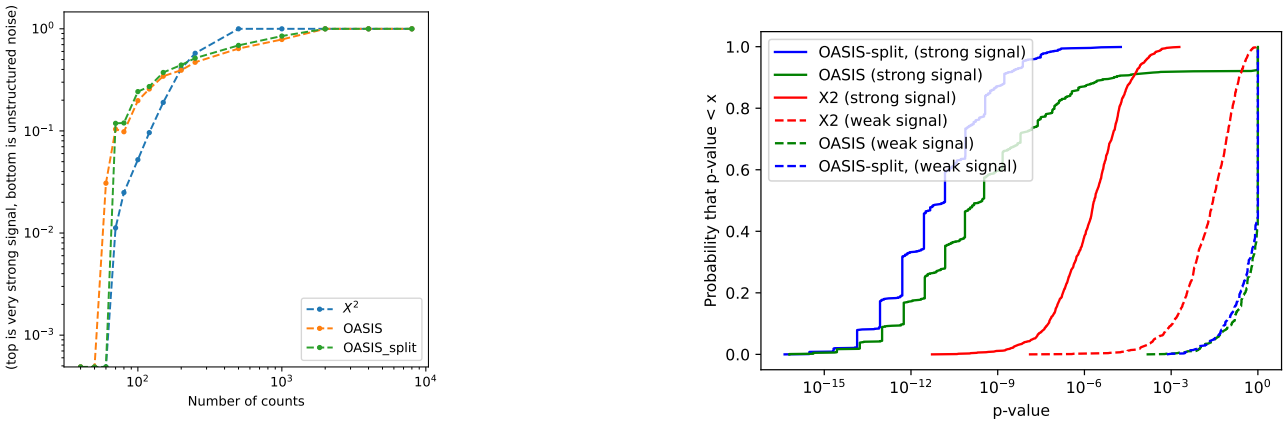

Fig. 8. Following Figure 1, this plot shows the performance of `OASIS-opt` and X^2 over the class of alternatives defined in Eq. (52). The plot shows an empirical achievability curve, where the y-axis indicates how structured versus unstructured the alternative is, and the x-axis indicates the total number of counts in the table. The more up and to the left a curve is, the better. For any alternative (any $\epsilon \in [0,1]$ for our model), a statistical test should reject the null with enough observations. Here, we show that empirically, OASIS is more likely to reject highly structured tables with lower counts, whereas Pearson's X^2 test is more likely to reject unstructured tables with moderate counts. `OASIS-opt` with 5 random data splits is shown. The flat behavior between adjacent data points for `OASIS-opt` is due to the sample splitting: an increase in number of observations does not necessarily translate to an increase in number of observations in the test data, e.g. going from 7 to 8 observations per sample. The right plot shows the subplot in Figure 1A, additionally showing curves when 5 random splits are used for `OASIS-opt`. As can be seen, this removes the tail of the `OASIS-opt` p-values, improving the algorithmic performance in this setting. The jagged behavior in the plot is due to the finite number of observations ($M = 100$ in this plot), leading to a finite number of values the optimized test-statistic will take, and so even aggregating over 1000 trials, the curve does not smooth out as it does for X^2, which can take many more possible values.

In the simulation of this method, we note that as initially stated, `OASIS-opt` has a large potential drawback. There are two sources of randomness used in the algorithm; the first time in splitting the data into train and test, and then in solving the non-convex optimization problem. The first source (splitting into train and test) is fundamental, whereas the second is simply for computational efficiency; if computational complexity is not an issue, we could enumerate over all possible $\boldsymbol{f} \in \{0,1\}^I$ and identify the optimal $\boldsymbol{c}$ for each, yielding the optimal solution to the inner optimization problem deterministically. However, the randomness in sample splitting is necessary, and can yield a bad bottleneck. If the training data is not representative of the test data, then no matter how well the optimization problem is solved on the training data, it need not yield a significant p-value on the test data. Thus, a natural consideration is to generate multiple random splits of the data, and multiple hypothesis correct over these. In Figure 1D, results are shown using 5 random splits. We additionally show results using just one random split in Figure 8.

B. Robustness against simulated alternatives. OASIS provides robustness against many undesirable alternatives, beyond just negative binomial overdispersion. We provide additional simulation evidence, showing OASIS's robustness against corruption of each individual columns probability distributions Figure 11.

B.1. Negative binomial overdispersion. As discussed, OASIS is robust to negative binomial overdispersion, as modelled in (15). Under the statistical null, for a common row distribution $\boldsymbol{p}$, if the number of observations in each column is Poisson distributed with rate λ, then the observed counts satisfy $X_{i,j} \sim \text{Pois}(\lambda p_i)$, independently drawn for each entry. However, as shown by (15), sequencing data is overdispersed in practice, and should be modelled as negative binomial. For an overdispersion parameter α (where $\alpha = 0$ corresponds to the true null), we thus generate data as $X_{ij} \sim \text{Pois}(\Gamma(\lambda p_i/\alpha, \alpha))$. Following the trends depicted in (15), we model our overdispersion parameter as a function of the expected number of counts to be observed x, as as $\alpha(x) = 3/x$ if $x < 150$, and $\alpha(x) = 0.02$ otherwise. In these plots, the number of observations per column was varied from 10 to 1000 in log-scale through 10 equally spaced points. Then one which shows most clearly the full dynamic range is one of the midpoints, index 6/10, with a value of $\exp(\ln(10) + \frac{5}{9}(\ln(1000) - \ln(10))) \approx 129$.

num rows	num cols	X^2	OASIS-rand	OASIS-opt
5	10	0.370	0.005	0.007
	50	0.765	0.009	0.006
	400	0.995	0.001	0.057
10	10	0.792	0.011	0.015
	50	1.000	0.007	0.046
	400	1.000	0.010	0.309
20	10	0.999	0.065	0.080
	50	1.000	0.058	0.213
	400	1.000	0.100	0.866
100	10	1.000	0.871	0.978
	50	1.000	0.983	1.000
	400	1.000	0.994	1.000

Table 3. Power at $\lambda \approx 129$, noise model as in (15). Exponentially distributed target distribution, described in text. Full plots in Figure 9.

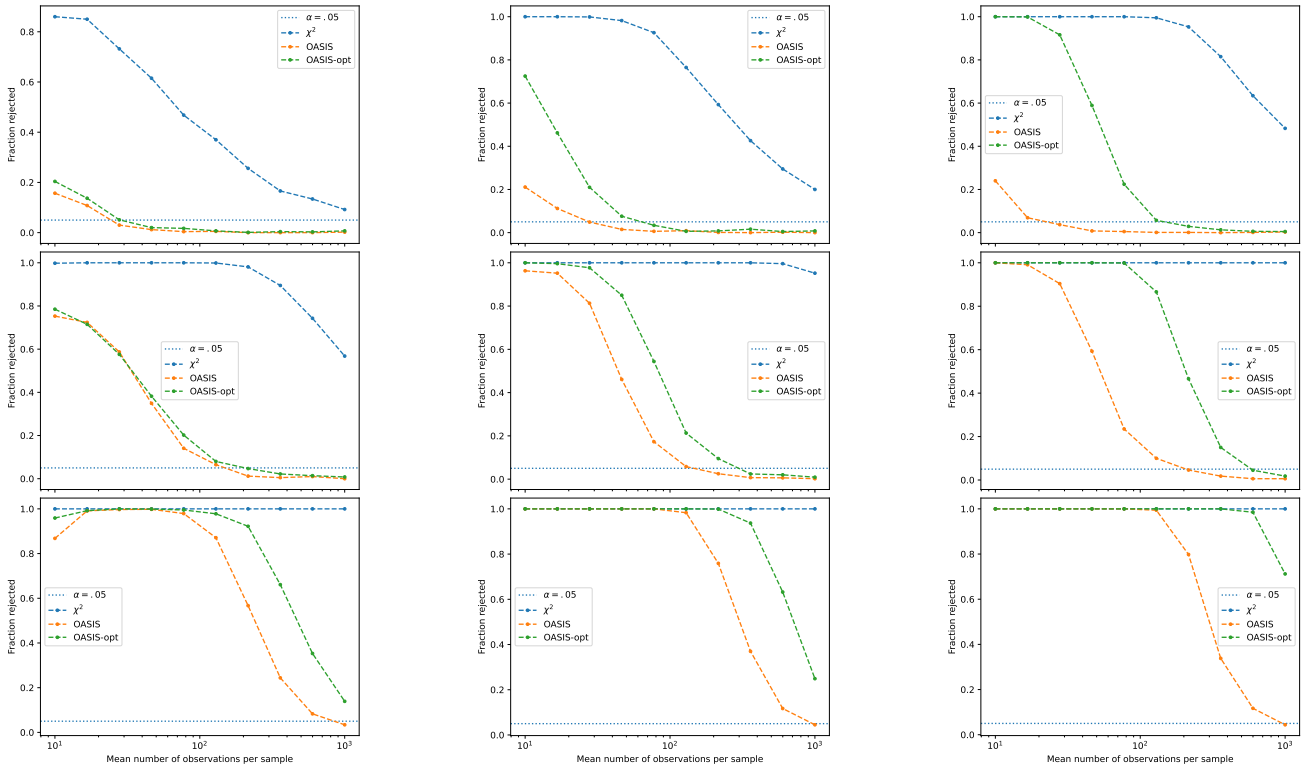

Fig. 9. Simulated data for negative binomial overdispersion. Null target distribution generated from exponential distributions, described in text. OASIS has significantly better control of the false positive rate than the X^2-test, and requires substantially fewer samples per column to control the FDR. Left to right, the columns correspond to tables having 10, 50, and 400 columns. Top to bottom, the rows indicate 5 rows, 20 rows, and 100 rows.

To generate a random target distribution from the null, we simulated two settings. One parameter free method is to take a uniform target distribution. Another way, modelling the more realistic setting where some targets are more likely than

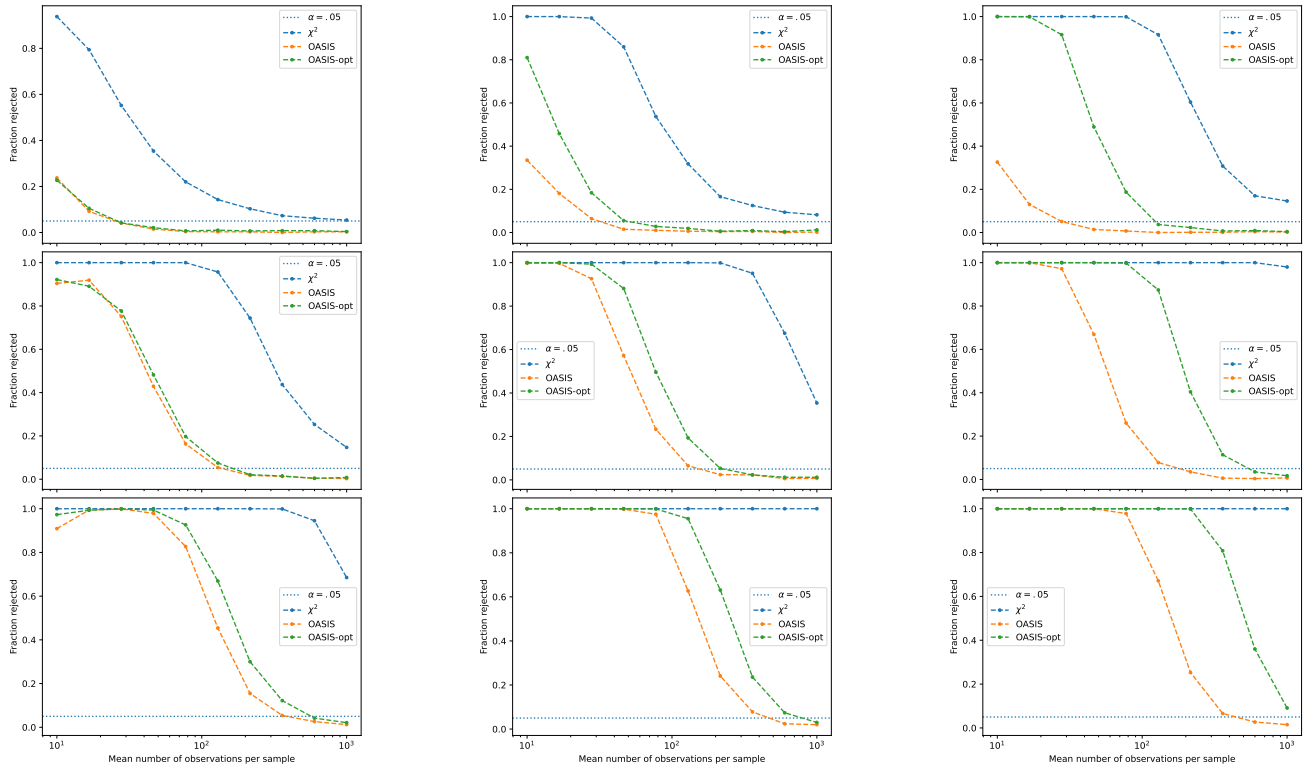

Fig. 10. Simulated data for negative binomial overdispersion. Null target distribution is uniform over rows. OASIS has significantly better control of the false positive rate than the X^2-test, and requires substantially fewer samples per column to control the FDR. Left to right, the columns correspond to tables having 10, 50, and 400 columns. Top to bottom, the rows indicate 5 rows, 20 rows, and 50 rows.

others, is to generate a vector $\tilde{\boldsymbol{p}}$ where each entry is drawn independently from some distribution that has nontrivial tails, and renormalize this to a probability distribution, $\boldsymbol{p} = \tilde{\boldsymbol{p}}/\|\tilde{\boldsymbol{p}}\|_1$. In this work, we utilized an exponential distribution with rate parameter 5.

In either setting, uniform or exponential target distribution (Figures 9 and 10), OASIS provides significantly better control of the discovery rate against this uninteresting alternative stemming from overdispersion.

B.2. Robustness again ℓ_1 corruption. In this setting, we study an alternative where each column is uniformly distributed, but is then mixed (with weight ϵ) with a probability distribution $\boldsymbol{q}^{(j)}$, which is generated independently for each column. Concretely,

$$\boldsymbol{p}^{(j)} = (1-\epsilon)\frac{1}{r}\mathbf{1} + \epsilon \boldsymbol{q}^{(j)}. \qquad [53]$$

Where the alternative matrix is generated by taking a Poisson number of observations from each column, multinomial with that columns probability vector $\boldsymbol{p}^{(j)}$, independently for each column. To avoid additional parameters, a uniform target distribution was used as the true null. Each $\boldsymbol{q}^{(j)}$ was generated by taking each entry to be uniformly distributed between $[0, 1]$. This is shown pictorially in Figure 11.

In simulations, we observe that OASIS is extremely robust to these types of perturbations; whereas X^2 immediately identifies that these tables deviate from the null, OASIS has a much more difficult time identifying structured deviations, and maintains control of the discovery rate for much larger corruption magnitudes, especially for larger tables.

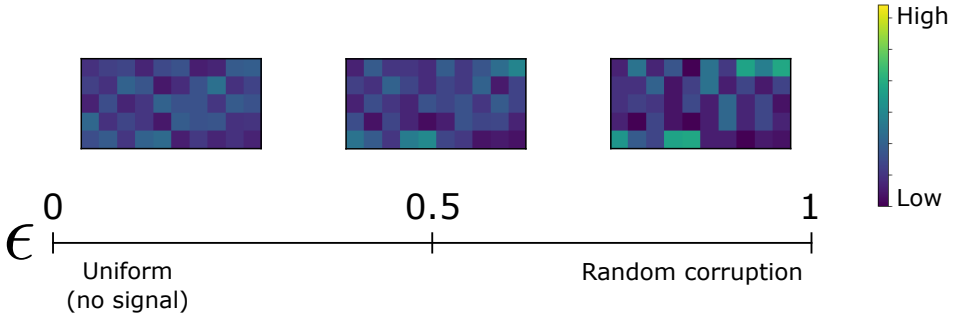

Fig. 11. Simulated data under independent ℓ_1 corruption to each columns probability vector. X^2 has significant power against this alternative, even though it represents contamination and a "biological null".

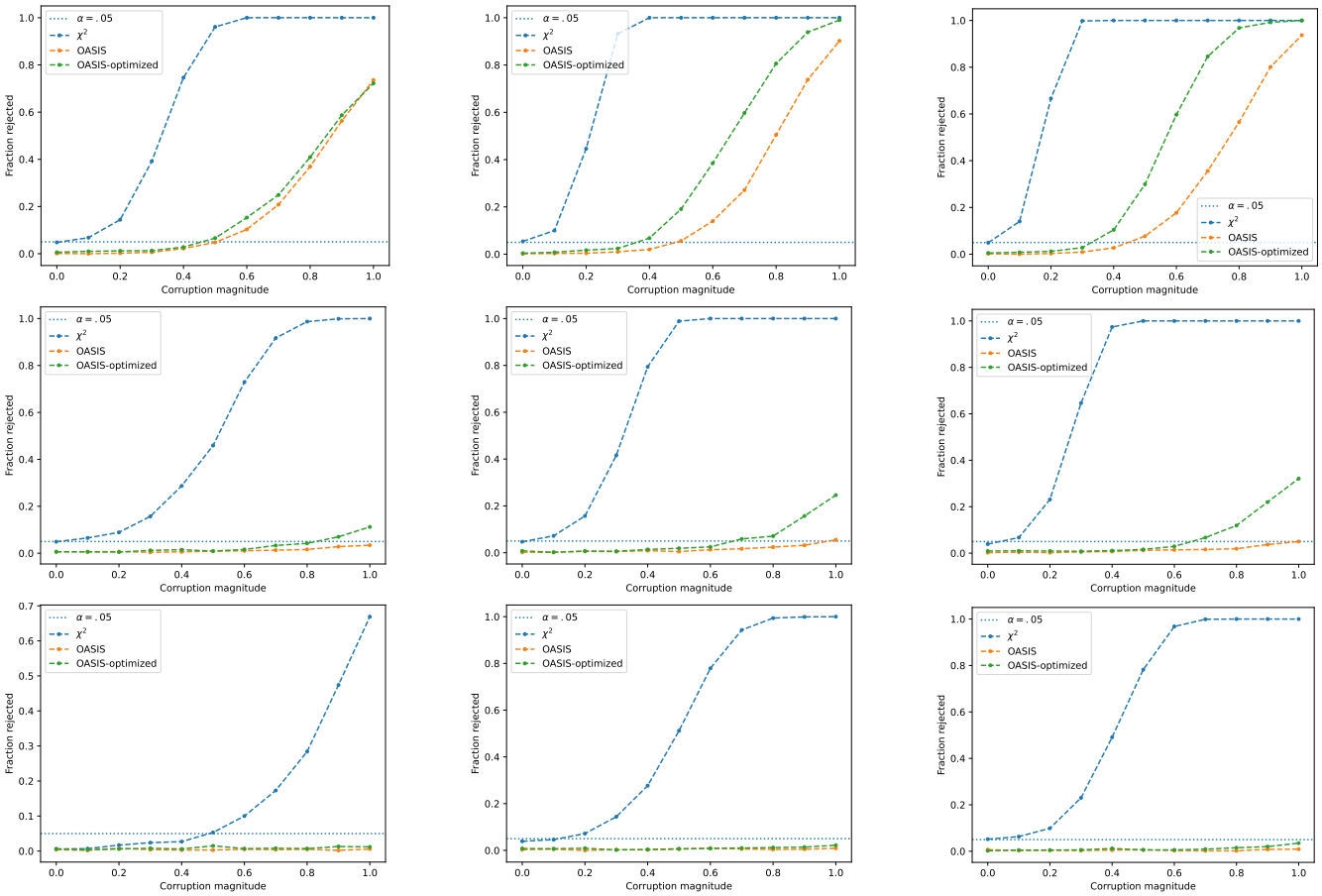

Fig. 12. Simulated data for independent ℓ_1 corruption of each column. First row corresponds to 20 rows, second to 100, and third to 500. First column corresponds to 10 columns, second to 50 columns, third to 100 columns. Number of observations in each column is Poisson distributed with mean 100, independent across columns. Observations are drawn independently, and are multinomially distributed for each column with probability vector as defined in Eq. (53).

C. Power against simulated alternatives. OASIS has power against alternatives with unique per-sample expression (motivated by V(D)J recombination) Figure 13, and can be shown to have more power against splicing-type alternatives Figure 15 than the X^2 test in some settings.

C.1. Unique per-sample expression. In this setting, the probability distribution of column $j \in [J]$ is $\boldsymbol{p}^{(j)} \in [0,1]^I$, with $I = J$:

$$p_i^{(j)} = \begin{cases} (1-\epsilon) + \epsilon/r & \text{if } i = j, \\ \epsilon/r & \text{otherwise.} \end{cases} \qquad [54]$$

This setting is shown pictorially in Figure 13, where a corruption $\epsilon = 0$ corresponds to pure signal, and $\epsilon = 1$ corresponds to a uniform target distribution for all columns (no signal). We provide simulation results for this setting in Figure 14, where in most settings Pearson's X^2 has more power than OASIS, but not by a significant margin.

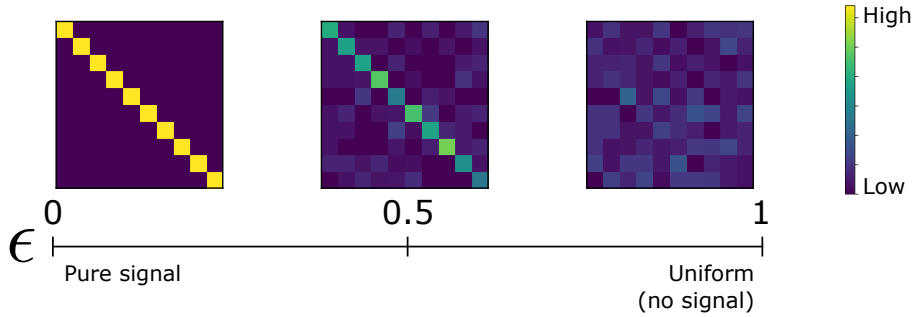

Fig. 13. Simulated data for an alternative with unique per-sample expression, motivated by V(D)J recombination. All methods have power.

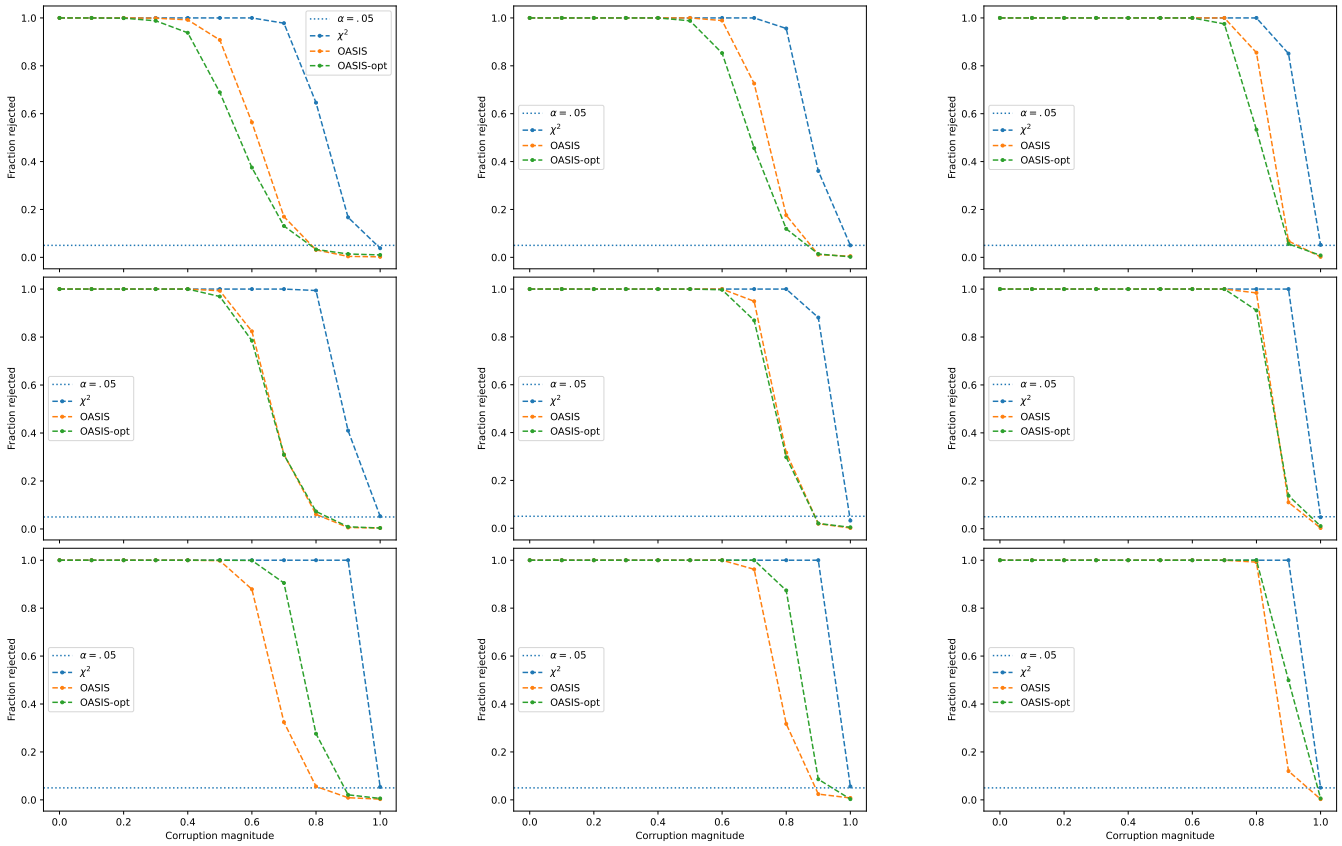

Fig. 14. Simulated data for unique per-sample expression alternative. First row corresponds to 10 rows and columns, second to 25, and third to 100. First column corresponds to a mean of 10 observations per column, second to 20, and third to 50. Number of observations in each column is Poisson distributed with the given mean, independent across columns. Probabilistic model in Eq. (54).

C.2. Two-group setting. Here we show OASIS's power against alternative splicing type alternatives across a broader range of parameters. Alternative tables were generated as taking two groups of samples of equal size, and defining the two group's probability distributions as, for $k = 1, 2$

$$p_i^{(k)} = \begin{cases} 1 - \epsilon + \epsilon/I & \text{if } i = k, \\ \epsilon/I & \text{otherwise.} \end{cases} \quad [55]$$

This indicates that $\boldsymbol{p}^{(1)}$ has most of its counts in the first row, and $\boldsymbol{p}^{(2)}$ in the second. This is shown pictorially in Figure 15.

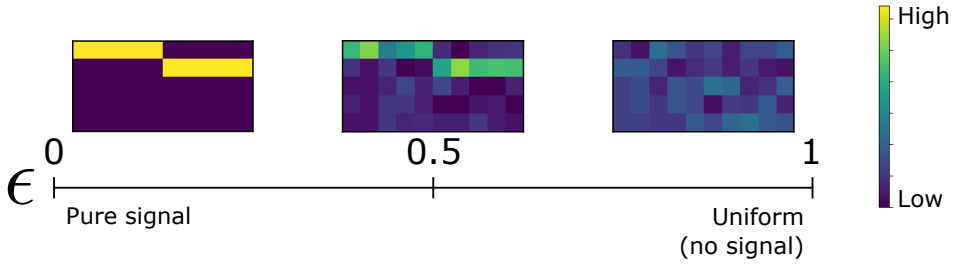

Fig. 15. Simulated data for a differentially regulated alternative splicing-type alternative. OASIS has more power than X^2.

The number of observations per sample is drawn independently from a Poisson distribution with mean 20 for each sample. Figure 16 shows how OASIS can have more power than X^2 in a variety of parameter regimes. For OASIS-rand, 50 $\boldsymbol{c}$ vectors were drawn uniformly at random from $\{-1, +1\}^J$ and 10 $\boldsymbol{f}$ vectors uniformly at random from $\{0, 1\}^I$, and all 500 pairs were tested.

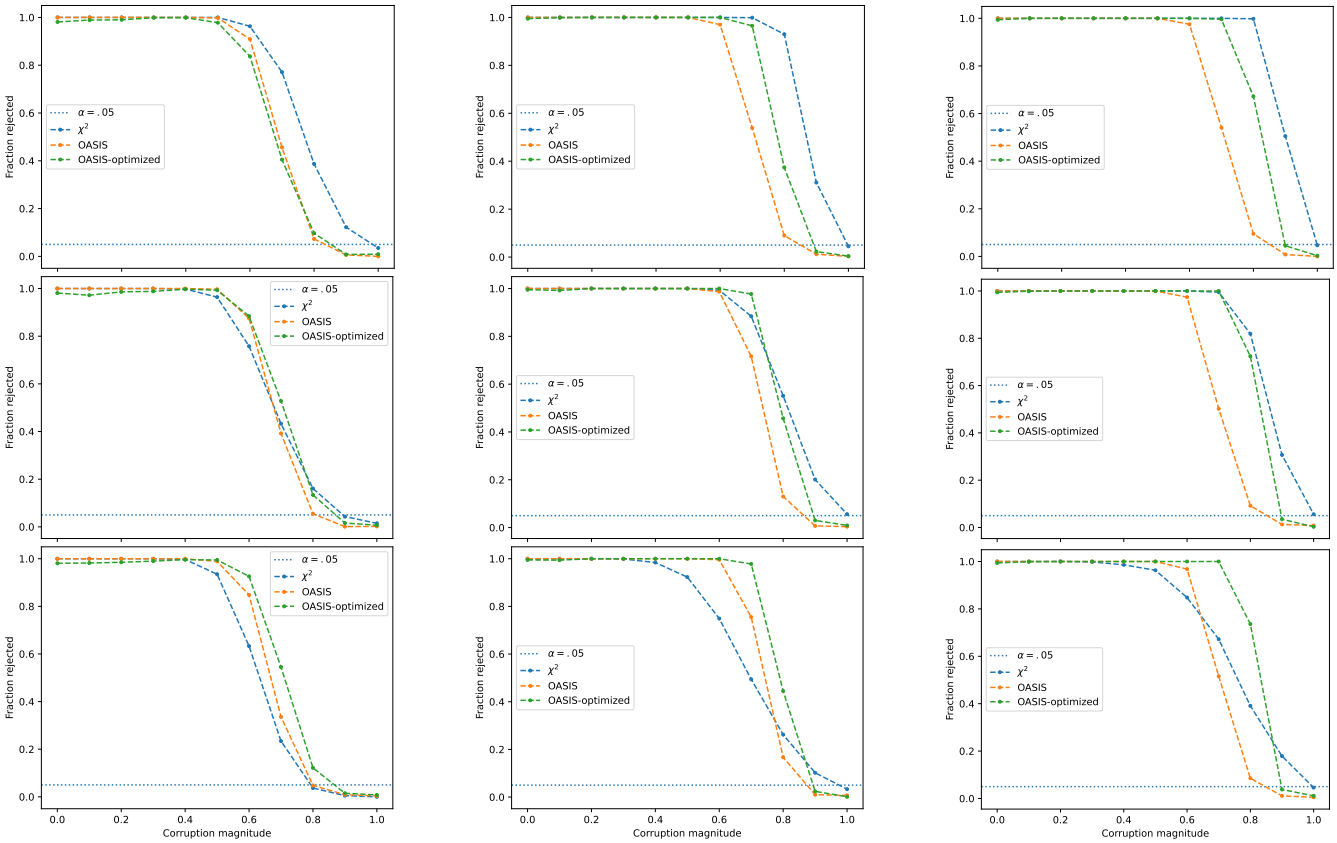

Fig. 16. Simulated data for alternative splicing-type alternative. First row corresponds to 20 rows, second to 100 rows, third to 400 rows. First column corresponds to 10 samples, second to 50 samples, third to 100 samples. Number of observations in each column is Poisson distributed with mean 20.

D. SARS-CoV-2 additional plots. In this section we provide additional results regarding the SARS-CoV-2 analysis provided in the main text.

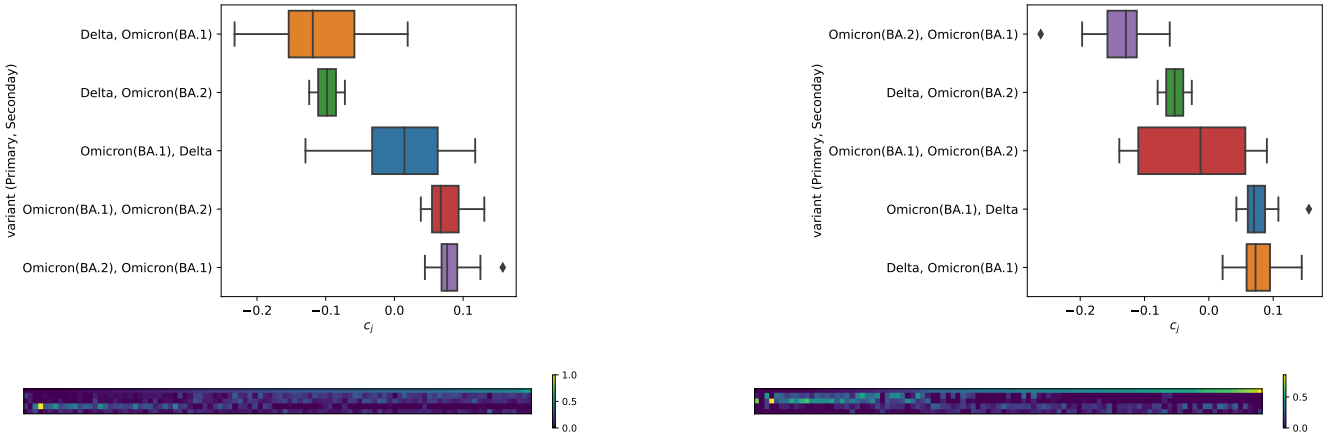

Fig. 17. Top 2 tables in terms of `OASIS-opt` p-value, not called by OASIS-rand. Both tables yield a vector $\boldsymbol{c}$ with an absolute cosine similarity of 0.76 with ground truth metadata (was the patient infected with Delta or not).

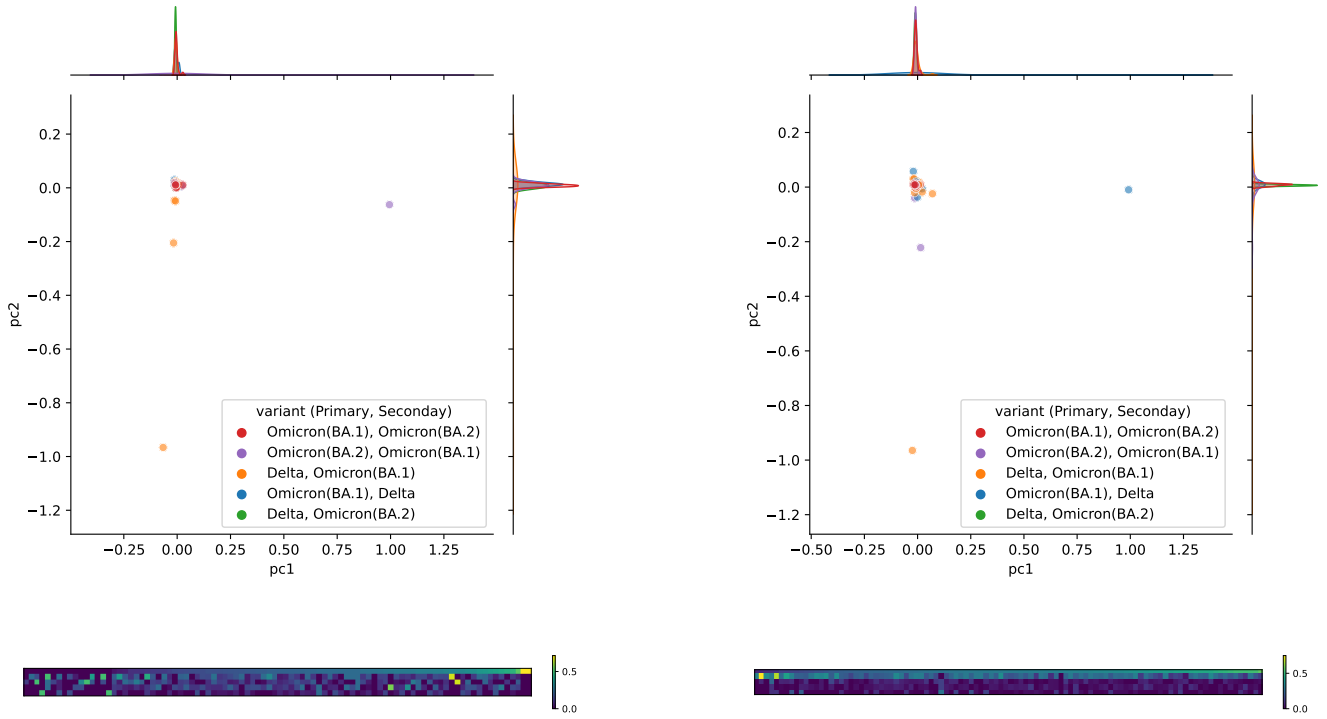

Fig. 18. Top 2 tables in terms of X^2-p-value (both with a q-value of 0), not called by `OASIS-opt` (q-value > 0.05). Since there were more than 2 tables with a X^2 p-value of 0, we chose the two with the smallest `OASIS-opt` p-value. The output 1d embedding by correspondence analysis yields an absolute cosine similarity of 0.15 and 0.02 with ground truth (whether a patient has Delta or not). Even when extended to 2 dimensions, the clusters do not separate.

We observe in Figure 20 that the improved predictive power of OASIS with respect to sample metadata is primarily driven by the tables OASIS prioritizes. When performing correspondence analysis on these tables, similar results are obtained, but when looking at all calls made by Pearson's X^2, many have worse correspondence with sample metadata.

As seen in Figure 22, iterative testing performed on the stacked data matrices yields significant p-values (of value 0, up to machine precision) through 30 iterations, due to the large number of counts, and inherent latent structure. Since there are so many counts, even vectors $\boldsymbol{c}$ that do not fully align with the latent structure will yield significant p-values. Thus, while for smaller matrices significance testing provides a good cutoff, for larger tables one may still need to find an "elbow" in the

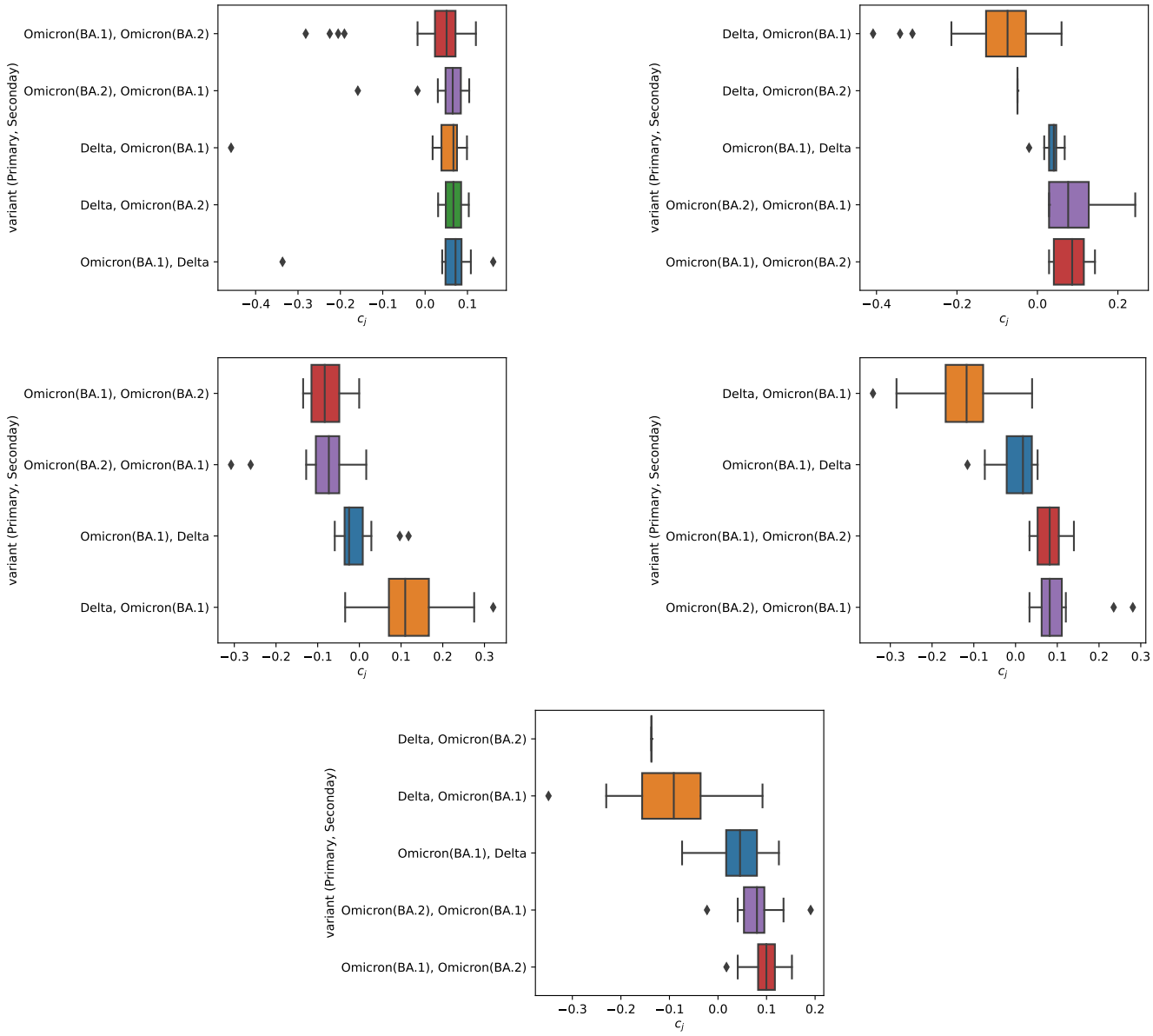

Fig. 19. 5 tables in `OASIS-opt`'s reduced list of anchors (effect size in the top 10%, $M > 1000$, significant `OASIS-opt` p-value). First is 482×100, 3688 counts, cosine similarity of 0.2. Second is 98×77, with 2540 counts, cosine similarity of 0.61. Third is 74×71, 1486 counts, cosine similarity of 0.69. Fourth is 67×69, 1299 counts, cosine similarity of 0.62. Fifth is 88×91, 1256 counts, 0.63 cosine similarity.

p-value curve. We demonstrate this in Figure 22, where we observe that the first two p-values are substantially smaller, with the rest approximately fitting a simple linearly decaying trend (when the log p-values are plotted on log scale). This threshold can be programatically identified by looking at the ratio of successive log p-values, and finding the maximum.

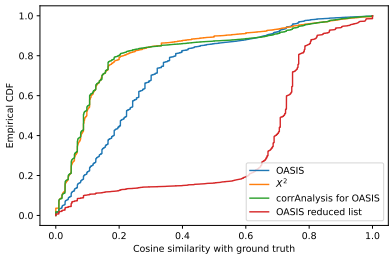

Fig. 20. Results on 100,914 tables generated by NOMAD (1) from SARS-CoV-2 data (6). Metric plotted is absolute cosine similarity between binarized sample embedding vector per table ($\boldsymbol{c}$ for OASIS, principle right singular vector from correspondence analysis for X^2) and ground truth vector indicating whether a patient (sample) has Delta or not. Empirical CCDF of absolute cosine similarity between identified vectors per table and ground truth vector is plotted for the tables that each method declares significant after multiple hypothesis correction. Out of the 100,914 tables, OASIS rejects 28,430, and Pearson's X^2 test rejects 71,543 tables. However, when decomposed, the tables that X^2 rejects do not yield signal that correlates well with the ground truth; examining quantiles of the two distributions, the 0.5 and 0.9 quantiles of these empirical distributions are 0.22 and 0.66 for OASIS, as opposed to 0.10, 0.52 for X^2.

Fig. 21. Spectrum of C matrix of stacked $\boldsymbol{c}$ vectors from $\boldsymbol{c}$-aggregation for SARS-CoV-2 data.

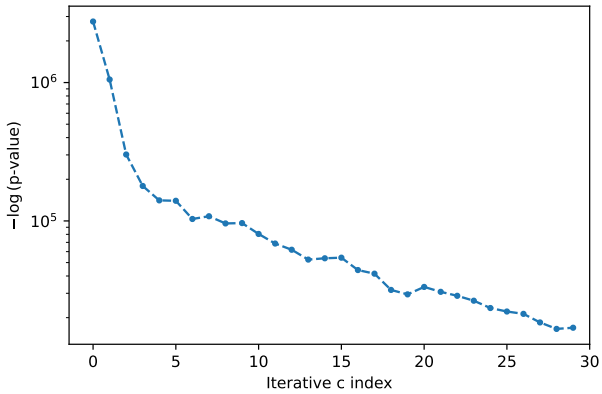

Fig. 22. Negative log of p-values for data-aggregation on stacked data matrices for SARS-CoV-2. As can be seen, after the second one, the remaining p-values follow a linear trend, indicating that only 2 components should be utilized.

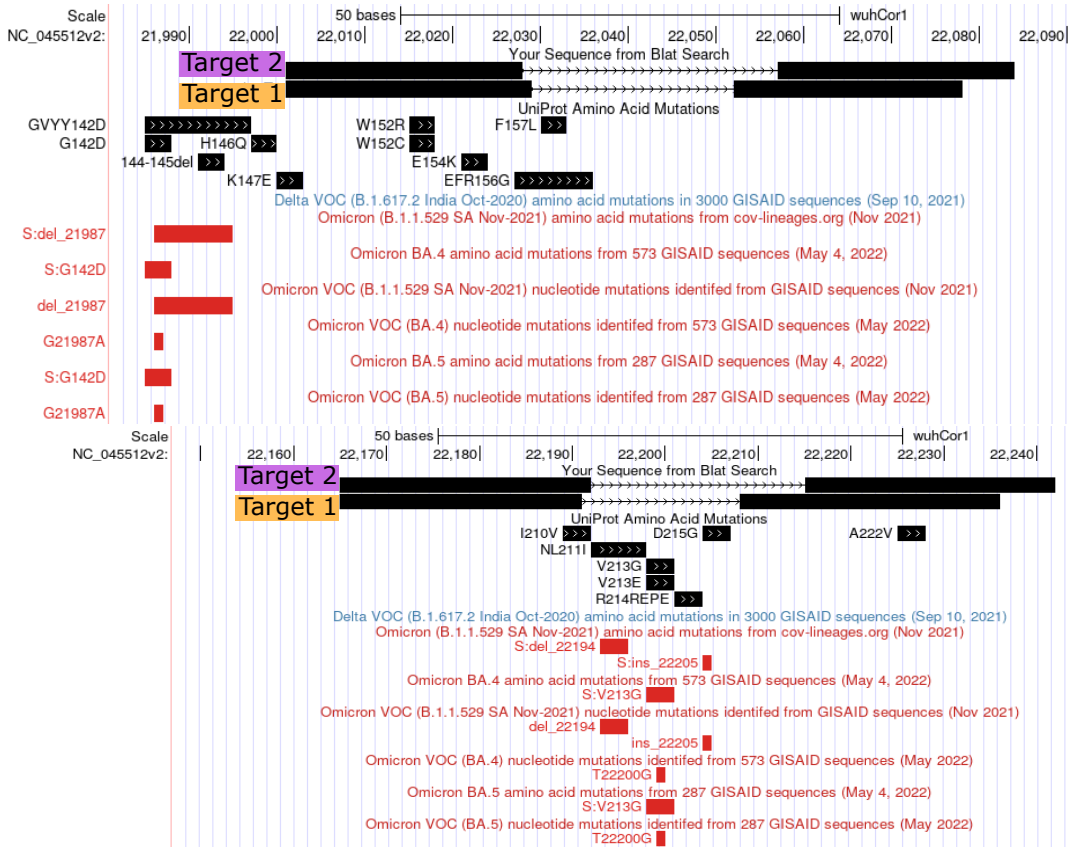

Fig. 23. BLAT to SARS-CoV-2 genome via UCSC genome browser. For each target, we align the anchor concatenated with the target, where for each of these two anchors, only 2 targets had abundance greater than 5%. Note that these targets do not come immediately after the anchor, but are taken a fixed distance ahead (called the lookahead distance (1)). In both examples, target 1 takes a value $f_i = 1$, and target 2 takes a value $f_i = 0$. The reported 93% agreement predicts that $f_i = 1$ corresponds to Delta and $f_i = 0$ corresponds to Omicron. Thus, T1 is predicted to be Delta, and T2 to be Omicron. Not shown in the first example, analyzing the raw reads from these samples shows that those from Delta samples follow the genome exactly, whereas those from Omicron samples exhibit a 6 basepair deletion in this gap, leading to the resulting contingency table. In the second example, we observe the same behavior (Omicron has a larger gap between anchor and target), this time due to an annotated Variant of Concern (VOC), a deletion.

E. Tuberculosis additional plots. For the analysis of Tuberculosis data, a newer version of the NOMAD pipeline was run to generate counts files, based on efficient k-mer counting (27). Then, the same optimization procedures were run to generate the optimized p-values

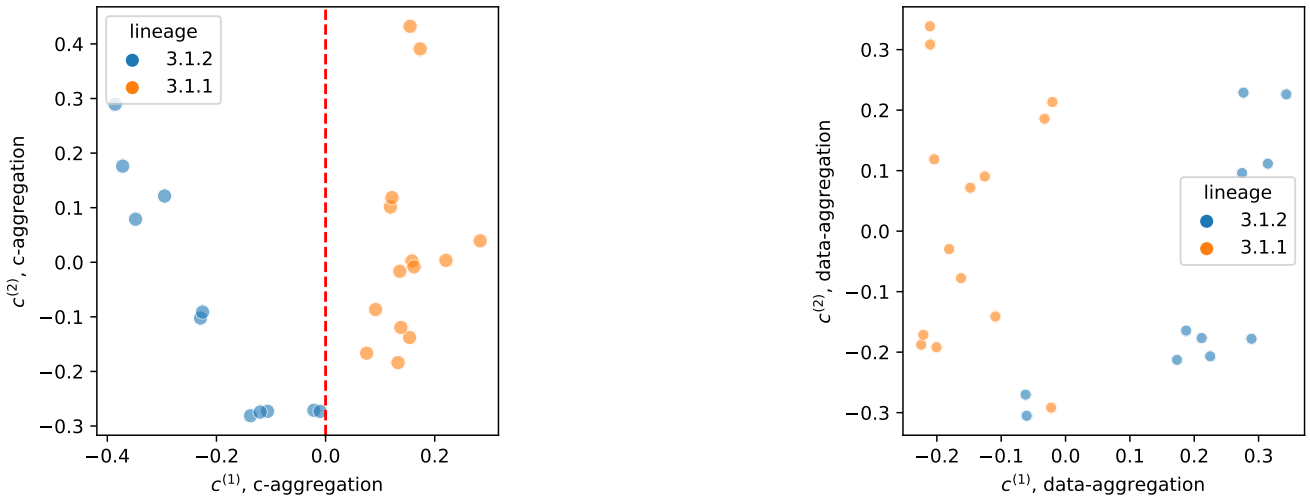

Fig. 24. 2D plots of Tuberculosis data. Left is $\boldsymbol{c}$-aggregation, right is data-aggregation.

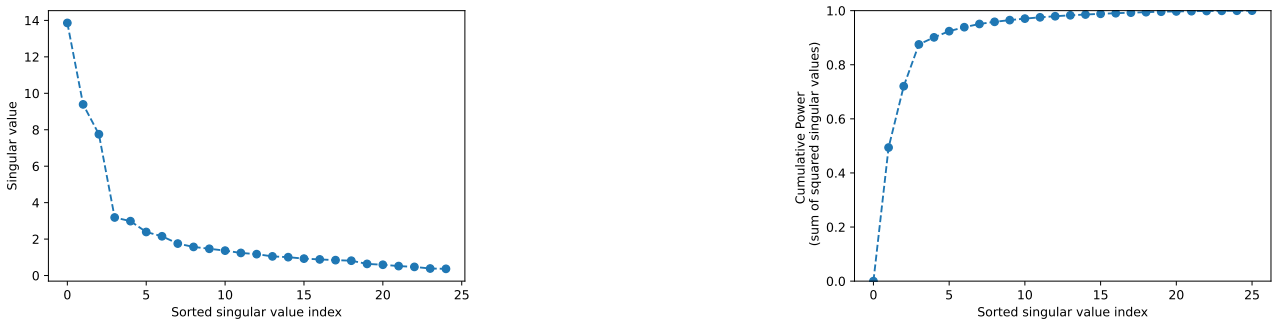

Fig. 25. Spectrum of matrix of stacked $\boldsymbol{c}$ vectors used in $\boldsymbol{c}$-aggregation, C, for Tuberculosis data.

F. Multiple group. In Figure 26 we observe that OASIS, unaware of the underlying multi-group structure, is able to give a statistically valid stopping criterion for subclustering, correctly identifying the number of underlying clusters as 3 55% of the time when each column has Poisson with mean 20 observations per column, and up to 69% with mean 30 observations per column.

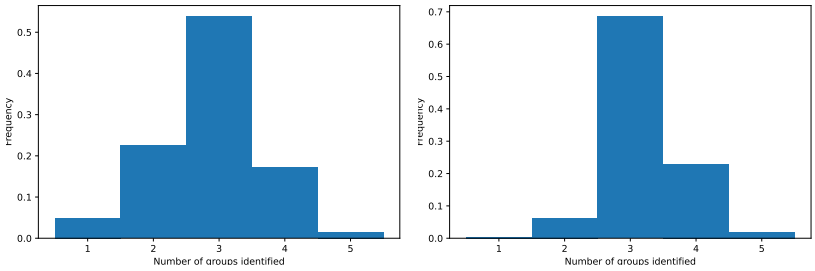

Fig. 26. Evaluation of OASIS for group discovery, in a planted setting with 3 groups, 10 rows, and 20 columns. Plotted is the number of identified $\boldsymbol{c}$ vectors using `OASIS-iter`; finding orthogonal $\boldsymbol{c}^{(k)}$ until p-value is no longer significant. Left figure shows 20 observations per column, right shows 30.

G. Asymptotic normality. As proved in Proposition 2, OASIS's test statistic is asymptotically normally distributed. However, with asymptotia, one natural question is "how large do counts need to be before the approximation is reasonable"? In Figure 27, we show that even for very small numbers of observations per column (e.g. 20), the normalized test statistic is very close to gaussian.

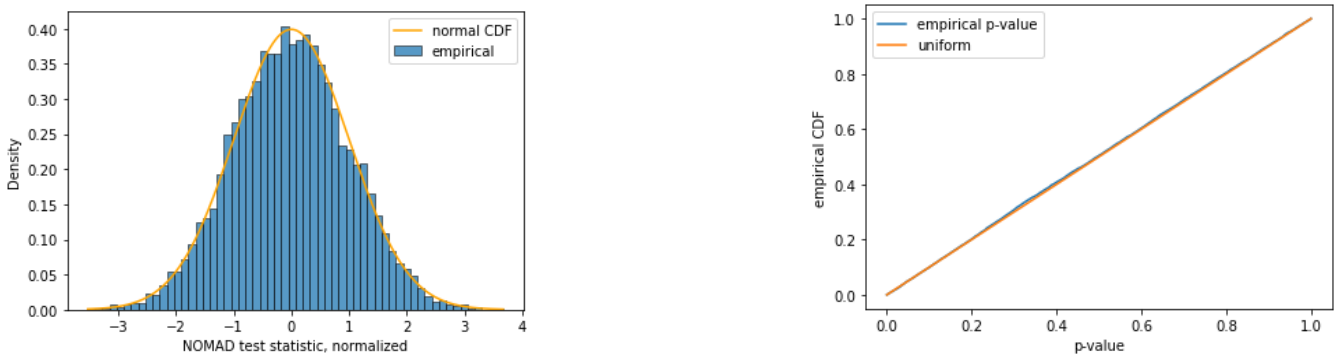

Fig. 27. Distribution of OASIS test statistic. Simulated data over 10k trials with 5 rows, 10 columns, random target distribution (normalized exponential), column counts are Poisson with mean 20. Ran 10k random tables, with random $\boldsymbol{c}$ and $\boldsymbol{f}$.

H. Correlation coefficient. As a correlation coefficient, OASIS favorably compares against xicor (26) in some settings, as shown in Figure 28. The example functions used are detailed in (26), where we show that OASIS has additional power against balanced, structured alternatives, such as circular and heteroskedastic.

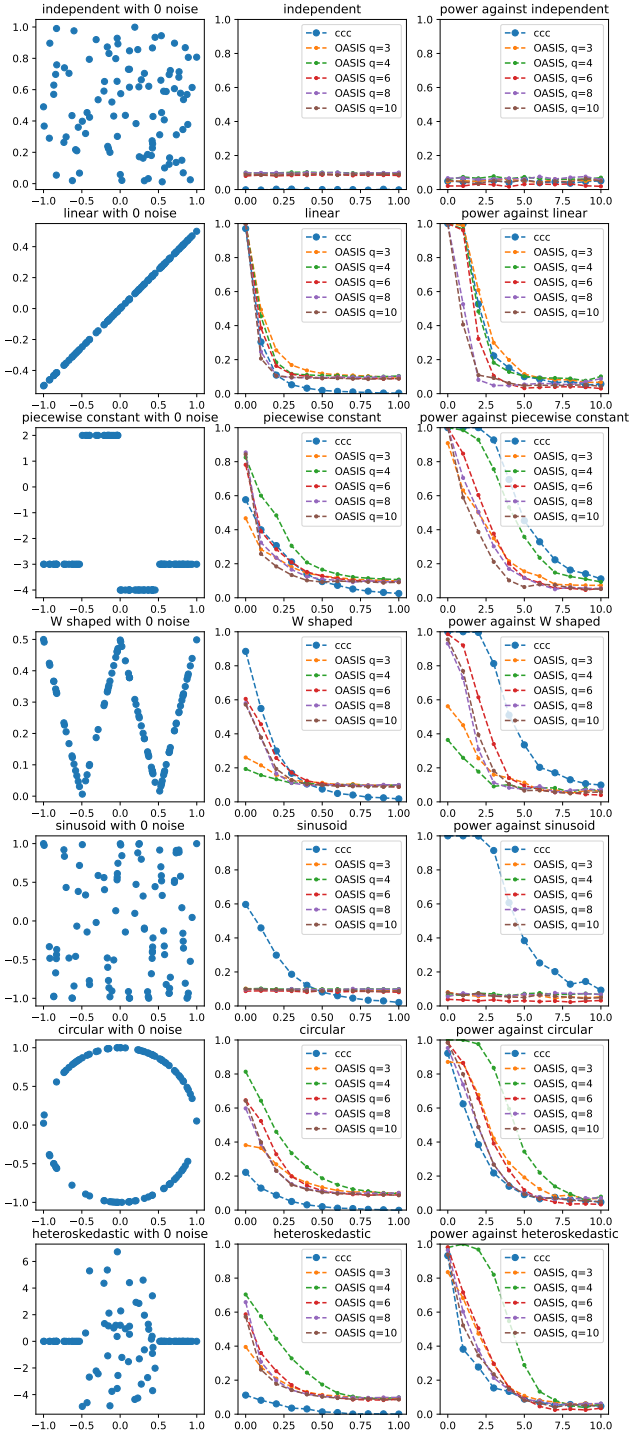

Fig. 28. Comparison of `OASIS-opt` and XICOR (26). OASIS asymptotic p-value used, 25% train test split. Each row represents a different setting, where the left column shows the noiseless problem instance $y = f(x)$, the middle column shows the magnitude of the correlation coefficient averaged over 1000 iterations per point, and the right column shows the fraction of time the null was rejected. For OASIS, the middle column shows the effect size, and the right column shows the rejection fraction utilizing the asymptotic p-value bound. The quantization level q is varied across instances; to map the input continuous-valued random variable to a discrete categorical one. This is performed independently for the row and column random variables, and binned into q bins of equal counts ($1/q$ quantiles).

15. Power analysis

Due to the simple closed form expressions for the OASIS and X^2 p-values, we can perform approximate power calculations across different alternatives. Concretely, we analyze several toy settings where we observe a scaled version of the underlying alternative matrix, and compute OASIS and X^2 p-values, showing that OASIS indeed has substantial power in these settings, and can be more powerful than X^2 in certain regimes. We conjecture that while OASIS may not exhibit the optimal asymptotic rate against certain classes of alternatives, as shown by the unique per-sample expression example it does have power going to 1 as the number of observations goes to infinity across a broad class of alternatives.

In this section we provide an approximate power analysis for the case where we observe a scaled version of the ground truth matrix. We additionally assume that OASIS identifies an optimal pair of $\boldsymbol{c}, \boldsymbol{f}$. It is reasonable to assume that this occurs in the asymptotic regime, and further since OASIS's test statistic is bilinear in $\boldsymbol{c}, \boldsymbol{f}$, its p-value is relatively smooth in these parameters, and so OASIS only needs to identify a pair of near optimal $\boldsymbol{c}, \boldsymbol{f}$ to yield a similar p-value.

The X^2 p-value for k degrees of freedom is asymptotically distributed as $\mathcal{N}(k, 2k)$. Since the X^2 test-statistic is $X^2 = \left\|\text{diag}(\sqrt{M/X\mathbf{1}})\tilde{X}\right\|_F^2$, plugging in $k = (I-1)\times(J-1)$, and examining large values of the test-statistic,

$$P_{X^2} \approx \exp\left(-\frac{X^2}{4IJ}\right) \tag{56}$$

A. Unique per-sample expression. In this setting, the alternative matrix is $c \times c$, where the j-th sample expresses target (row) j with probability $1-\alpha$, and with probability α is uniform over the rest of the $c-1$ rows. The number of observations per column are assumbed to be the same. Formally, the alternative matrix A is as below:

$$A = \left(1 - \frac{c}{c-1}\alpha\right) I_c + \frac{\alpha}{c-1}\mathbf{1}\mathbf{1}^\top \tag{57}$$

$$X = A\,\text{diag}\,(\boldsymbol{n}) \tag{58}$$

$$E = \frac{1}{c}\mathbf{1}\mathbf{1}^\top\,\text{diag}\,(\boldsymbol{n}) \tag{59}$$

In this setting, Pearson's X^2 statistic is

$$\begin{aligned} X^2 &= \sum_{ij} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \\ &= \sum_j \frac{n_j^2((1-\alpha) - \frac{1}{c})^2}{n_j/c} + \sum_{i\neq j} \frac{n_j^2(\frac{\alpha}{c-1} - \frac{1}{c})^2}{n_j/c} \\ &= c(1-\alpha-\frac{1}{c})^2 \sum_j n_j + c(\frac{\alpha}{c-1} - \frac{1}{c})^2 \sum_{i\neq j} n_j \\ &= cM(1-\alpha-1/c)^2 + \underbrace{c(c-1)M(\frac{\alpha}{c-1} - \frac{1}{c})^2}_{\text{negligible}}. \end{aligned} \tag{60}$$

Thus, for fixed c, as $M \to \infty$, the X^2 p-value can be bounded as

$$P_{X^2} = \mathbb{P}(\mathcal{N}(0,1) > M(1-\alpha)^2) \lesssim \exp(-M^2(1-\alpha)^4) \tag{61}$$

Analyzing OASIS, we assume that it identifies an optimal pair of $\boldsymbol{f}, \boldsymbol{c}$, where since n_j are balanced

$$f_i = \begin{cases} 1 & \text{if } 1 \le i \le c/2 \\ 0 & \text{if } c/2 < i \le c \end{cases} \quad \text{and} \quad c_j = \begin{cases} 1 & \text{if } 1 \le j \le c/2 \\ -1 & \text{if } c/2 < j \le c \end{cases} \tag{62}$$

constitute an optimal pair.

As $n_j = M/c$ for all j, $\hat{\mu} = 1/2$, and so

$$\hat{\mu}_j = \begin{cases} 1-\alpha/2 & \text{if } 1 \le i \le c/2 \\ \alpha/2 & \text{if } c/2 < i \le c \end{cases} \tag{63}$$

This yields

$$\begin{aligned} S &= \sum_j c_j \sqrt{n_j}\,(\hat{\mu}_j - \hat{\mu}) \\ &= \sum_j \sqrt{M/c}\frac{1-\alpha}{2} \\ &= \sqrt{cM}\frac{1-\alpha}{2} \end{aligned} \tag{64}$$

and a corresponding p-value bound

$$P_{\text{OASIS}} = \exp\left(-\frac{S^2}{\|\boldsymbol{c}\|_2^2}\right) = \exp\left(-M(1-\alpha)^2\right) \tag{65}$$

A.1. Summary. In this toy-setting of unique per-sample expression, X^2 will yield a p-value of

$$P_{X^2} \approx \exp\left(-M^2(1-\alpha)^4\right). \tag{66}$$

Comparatively, assuming OASIS identifies a pair of optimal $\boldsymbol{c}, \boldsymbol{f}$, it will obtain a p-value of

$$P_{\text{OASIS}} \leq \exp\left(-M(1-\alpha)^2\right). \tag{67}$$

While this is worse for constant α and M tending to infinity, OASIS still has power going to 1 as M increases.

B. Two-group alternative. Here we assume for simplicity that all n_j are equal, and that the clusters are of equal sizes. This can be relaxed by using a class-weighted mean estimate, which computes the mean f value over samples with $c_j > 0$, and averages this with the mean f value over samples with $c_j < 0$.

We describe the results in the context of differentially regulated alternative splicing, but the analysis holds for general 2 group alternatives with the below structure. For samples $j = 1, \ldots, J/2$, the target (row) distribution is $\boldsymbol{p}^{(+)}$, and for the second half of columns $\boldsymbol{p}^{(-)}$. X^2 naturally has power against this alternative.

In this setting we have:

$$X = n\left(\boldsymbol{p}^{(1)}\left[\boldsymbol{1}_{J/2}\ \boldsymbol{0}_{J/2}\right] + \boldsymbol{p}^{(2)}\left[\boldsymbol{0}_{J/2}\ \boldsymbol{1}_{J/2}\right]\right) \tag{68}$$

and thus

$$E = \left(\boldsymbol{p}^{(1)} + \boldsymbol{p}^{(2)}\right)\boldsymbol{1}_J^\top \operatorname{diag}(n_1, \ldots, n_J) \tag{69}$$

The Pearson X^2 statistic is then:

$$\begin{aligned}
X^2 &= \sum_{i,j} \frac{(X_{i,j} - E_{i,j})^2}{E_{i,j}} \\
&= n \sum_{i,j<J/2} \frac{\left(p_i^{(1)} - \frac{p_i^{(1)}+p_i^{(2)}}{2}\right)^2}{(p_i^{(1)} + p_i^{(2)})/2} + n \sum_{i,j\geq J/2} \frac{\left(p_i^{(2)} - \frac{p_i^{(1)}+p_i^{(2)}}{2}\right)^2}{(p_i^{(1)} + p_i^{(2)})/2} \\
&= \frac{nJ}{2} \sum_i \frac{\left(p_i^{(1)} - p_i^{(2)}\right)^2}{p_i^{(1)} + p_i^{(2)}} \\
&\begin{cases} \overset{(a)}{\geq} \frac{M}{4} \left\|p_i^{(1)} - p_i^{(2)}\right\|^2 \\ \leq M \end{cases}
\end{aligned} \tag{70}$$

in (a) we upper bounded the denominator by 2 (can potentially be tightened to 1), because this is in general not easy to compute. The sum can be upper bounded as $\|\boldsymbol{p}^{(1)}\|_1 + \|\boldsymbol{p}^{(2)}\|_1 = 2$, when the two vector's supports are disjoint.

From the $I \times J$ degrees of freedom, this yield an approximate p-value of

$$P_{X^2} \begin{cases} \exp\left(-\frac{M^2\left\|p^{(1)} - p^{(2)}\right\|^4}{8(I\times J)}\right) & \text{worst-case bound} \\ \exp\left(-\frac{M^2}{(I\times J)}\right) & \text{if supports are disjoint} \end{cases} \tag{71}$$

For OASIS, a pair of optimal $\boldsymbol{c}$ and $\boldsymbol{f}$ include where $\boldsymbol{c}$ is $+1$ for the first $J/2$ components, and -1 for the latter half, and $\boldsymbol{f}$ is $\operatorname{sign}(p^{(1)} - p^{(2)})$.

Then, the OASIS test statistic is

$$\begin{aligned}
S &= \sum_{j=1}^{J/2} \sqrt{n_j} c_j (\boldsymbol{f}^\top \boldsymbol{p}^{(1)} - \bar{\mu}) + \sum_{j=J/2+1}^{J} c_j \sqrt{n_j} (\boldsymbol{f}^\top \boldsymbol{p}^{(2)} - \bar{\mu}) \\
&\overset{(a)}{=} \sum_{j=1}^{J/2} \sqrt{n_j} \boldsymbol{f}^\top \left(\boldsymbol{p}^{(1)} - \boldsymbol{p}^{(2)}\right) \\
&= \frac{\sqrt{MJ}}{2} \boldsymbol{f}^\top \left(\boldsymbol{p}^{(1)} - \boldsymbol{p}^{(2)}\right) \\
&\overset{(b)}{=} \frac{\sqrt{MJ}}{2} \delta_{\text{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})
\end{aligned} \tag{72}$$

where in (a) we used the fact that all n_j were the same and that the clusters were equally balanced. If we use the class-weighted mean estimate sketched earlier then we can overcome the equally balanced classes problem. The n_j all being the same can be overcome by still assuming that they all go to infinity, but applying Jensen's inequality to lower bound the test statistic as above. In (b) we used the optimal $\boldsymbol{f} = \text{sign}(\boldsymbol{p}^{(1)} - \boldsymbol{p}^{(2)})$. This gives us a p-value of

$$P = 2\exp\left(-\frac{2S^2}{\|\boldsymbol{c}\|_2^2}\right) = 2\exp\left(-M\delta_{\text{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})/2\right) \quad [73]$$

B.1. Summary. The X^2 p-value can be bounded as

$$P_{X^2} \approx \exp\left(-\frac{M^2}{I \times J}\right). \quad [74]$$

OASIS has significant power against alternative splicing-type alternatives. Assuming that OASIS identifies the optimal $\boldsymbol{c}, \boldsymbol{f}$, which will be the case asymptotically, it will yield a p-value of

$$P_{\text{OASIS}} \leq \exp\left(-M\delta_{\text{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})\right). \quad [75]$$

We can see that when $I \times J \geq M$, OASIS can have more power than X^2; this is often the case we are in, where there are many rows with very low counts.

OASIS's p-value is scaling as (for constant total variation distance between the two probability vectors): $\exp(-M)$, whereas the X^2 p-value scales as $\exp(-M^2/(I \times J))$. Thus, when the matrix has low density or has many rows, i.e. $M < I \times J$, OASIS will have more power.

C. Time series. Consider the simple case where our data represents cells at progressing times during their cell-cycle. At time $t = 0$, they have target distribution $\boldsymbol{p}$, and at time $t = 1$ they have target distribution $\boldsymbol{q}$, smoothly interpolating between them. Assuming for simplicity that the samples are evenly spaced along this time course (i.e. cell j is at time $j = (j-1)/(J+1)$), and that all are observed equally (same n_j), the optimal $c_j = 2t_j - 1$, so $t = 0 \implies c_j = -1$ and $t = 1 \implies c_j = 1$.

Consider the simplest time series setting, where at time $t = 0$ cells have target distribution p, at time $t = 1$ distribution q, and vary smoothly in between, i.e. at time t the mixture distribution $p + t(q - p) = (1-t)p + tq$. For $T + 1$ evenly spaced time intervals, define the vector $\boldsymbol{t} = [0, 1/T, \ldots, (T-1)/T, 1]$. We observe that, for n observations per column, our expected data matrix X, and null mean matrix E, would be

$$\begin{aligned} X &= \left(\boldsymbol{p}(1-\boldsymbol{t})^\top + \boldsymbol{q}\boldsymbol{t}^\top\right) n \\ E &= \frac{n}{2}(\boldsymbol{p} + \boldsymbol{q})\mathbf{1}^\top \end{aligned}$$

To compute the power of Pearson's X^2 test, we have:

$$\begin{aligned} X^2 &= M \left\| \text{diag}(1/\sqrt{X\mathbf{1}})(X - E)\,\text{diag}(1/\sqrt{X^\top\mathbf{1}}) \right\|_F^2 \\ &= nM \left\| \text{diag}(1/\sqrt{X\mathbf{1}})(\boldsymbol{p} - \boldsymbol{q})(1/2 - \boldsymbol{t})^\top \right\|_F^2 \\ &\approx nM \sum_{i=1}^{I} \frac{(p_i - q_i)^2}{\frac{M}{2}(p_i + q_i)} \|1/2 - \boldsymbol{t}\|_2^2 \\ &\approx \frac{M}{6}\delta_{\text{TV}}(\boldsymbol{p}, \boldsymbol{q}) \end{aligned}$$

This leads to an approximate p-value of

$$\begin{aligned} P_{X^2} &\approx \exp\left(-\frac{M^2\delta_{\text{TV}}^2(\boldsymbol{p}, \boldsymbol{q})/36}{2I \times T}\right) \\ &= \exp\left(-nT\delta_{\text{TV}}^2(\boldsymbol{p}, \boldsymbol{q}) \times \frac{n}{72I}\right). \end{aligned}$$

To compute OASIS's power, we first compute S_j

$$\begin{aligned} S_j &= \sqrt{n}\left(\boldsymbol{f}^\top((1-t_j)\boldsymbol{p} + t_j\boldsymbol{q}) - \boldsymbol{f}^\top(\boldsymbol{p} + \boldsymbol{q})/2\right) \\ &= \sqrt{n}\boldsymbol{f}^\top((1/2 - t_j)\boldsymbol{p} + (t_j - 1/2)\boldsymbol{q}) \\ &= \sqrt{n}\boldsymbol{f}^\top(\boldsymbol{p} - \boldsymbol{q})(1/2 - t_j) \end{aligned} \quad [76]$$

We know that $c_j \propto S_j$ are optimal as

$$c_j^\star = \frac{1}{2} - t_j. \quad [77]$$

We assume that we find this exact $\boldsymbol{c}$ (noting that OASIS is robust to small estimation errors). Note that maximizing $\boldsymbol{f}$ is simple now, as the same term shows up in each column:

$$\max_{\boldsymbol{f}\in[0,1]^I} \boldsymbol{f}^\top(\boldsymbol{p}-\boldsymbol{q}) = \delta_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{q}). \tag{78}$$

Computing S,

$$\begin{aligned} S &= \sum c_j S_j \\ &= \sqrt{n}\boldsymbol{f}^\top(\boldsymbol{p}-\boldsymbol{q})\sum_j (1/2-t_j)^2 \\ &= \sqrt{n}\delta_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{q})\frac{(T+1)(T+2)}{12T} \end{aligned} \tag{79}$$

This gives OASIS a p-value of (for large enough T)

$$\begin{aligned} P(\boldsymbol{f},\boldsymbol{c}) &\le 2\exp\left(-\frac{2s^2}{\|\boldsymbol{c}\|^2}\right) \\ &\le \exp\left(-nT\delta_{\mathrm{TV}}^2(\boldsymbol{p},\boldsymbol{q})/6\right) \end{aligned} \tag{80}$$

C.1. Summary. X^2 provides a p-value bound of

$$\begin{aligned} P_{X^2} &\approx \exp\left(-\frac{M^2\delta_{\mathrm{TV}}^2(\boldsymbol{p},\boldsymbol{q})/36}{2I\times T}\right) \\ &= \exp\left(-nT\delta_{\mathrm{TV}}^2(\boldsymbol{p},\boldsymbol{q})\times\frac{n}{72I}\right). \end{aligned}$$

OASIS attains a p-value that can be bounded as

$$\begin{aligned} P_{\mathrm{OASIS}} &\le \exp\left(-M\delta_{\mathrm{TV}}^2(\boldsymbol{p},\boldsymbol{q})/6\right) \\ P_{X^2} &\approx \exp\left(-M\delta_{\mathrm{TV}}^2(\boldsymbol{p},\boldsymbol{q})\times\frac{n}{72I}\right). \end{aligned} \tag{81}$$

This displays the same behavior compared to X^2 as in alternative splicing, where OASIS can yield improved performance when $\frac{n}{I}$ is small.

D. One deviant sample. Consider the setting where one sample is maximally deviating, i.e. the observed matrix is:

$$X = \begin{bmatrix} 0 & n & \dots & n \\ n & 0 & \dots & 0 \end{bmatrix}.$$

Then, the expected matrix is

$$E = \begin{bmatrix} n(1-1/J) & n(1-1/J) & \dots & n(1-1/J) \\ n/J & n/J & \dots & n/J \end{bmatrix}.$$

Expressing the summation as grouped by $X_{00}, X_{10}, X_{0,1:}, X_{1,1:}$

$$\begin{aligned} X^2 &= \frac{(n(1-1/J))^2}{n(1-1/J)} + \frac{(n-n/J)^2}{n/J} + (J-1)\frac{(n-n(1-1/J))^2}{n(1-1/J)} + (J-1)\frac{(n/J)^2}{n/J} \\ &= n(1-1/J) + nJ(1-1/J)^2 + n/J + n(1-1/J) \\ &\approx nJ \end{aligned}$$

Thus, the Pearson X^2 test will yield an approximate p-value of

$$P_{X^2} \approx \exp\left(-\frac{(nJ)^2}{2(2J)^2}\right) \approx \exp\left(-n^2/8\right) \tag{82}$$

Comparatively, OASIS will choose $\boldsymbol{f} = [10]^\top$, which yields $S_j = \sqrt{n}/J$ for $j = 2,\dots,n$ and $S_1 = -\sqrt{n}(1-1/J)$. This yields an optimizing $\boldsymbol{c} = [-(1-1/J)1/J\dots 1/J]$, where $\|\boldsymbol{c}\|^2 = 1-1/J$ here. This yields a test-statistic value $S = \sqrt{n}\|\boldsymbol{c}\|^2$

Thus,

$$P_{\mathrm{OASIS}} = \exp(-S^2/\|\boldsymbol{c}\|^2) = \exp(-n). \tag{83}$$

D.1. Summary. X^2 provides a p-value of

$$P_{X^2} \approx \exp\left(-\frac{(nJ)^2}{2(2J)^2}\right) \approx \exp\left(-n^2/8\right), \tag{84}$$

in comparison to `OASIS-opt`'s p-value of

$$P_{\text{OASIS}} = \exp(-S^2/\|\boldsymbol{c}\|^2) = \exp(-n). \tag{85}$$

This shows how X^2 can be significantly overpowered against alternatives where only one column deviates.

E. Conclusions from power analysis. As can be seen, there are several regimes where the Pearson X^2 test has less power than OASIS, primarily when $\frac{M}{I \times J}$ is small. This highlights the difficulty with the X^2 test; its p-value depends heavily on the number of rows present, even if many of them are inconsequential.