Supporting Information for

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

Tavor Z. Baharav, David Tse, Julia Salzman

Julia Salzman.
E-mail: julia.salzman@stanford.edu

This PDF file includes:

Supporting Information Text

S.1. Data availability

Mycobacterium tuberculosis data is publicly available under Accession ID PRJEB41116 (1). SARS-CoV-2 data is publicly available under Accession ID PRJNA817806 (2). Both datasets were chosen arbitrarily from the sequence read archive (3) and were analyzed with the same method with the same default parameters. Code for SPLASH (previously called NOMAD) is publicly available at https://github.com/salzman-lab/splash, along with code for OASIS inference. An optimized implementation of SPLASH called SPLASH 2 has been developed (4) using the bounds and techniques described in this work, and is available here https://github.com/refresh-bio/R-SPLASH. *M. tuberculosis* data was generated with this improved implementation of SPLASH. Then, the same optimization procedures were run to generate the optimized p-value bounds.

S.2. 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, 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 $j = 1,\dots,J$ and $k = 1,\dots n_j$. $Z_{j,k}$ denotes the row identity of the k-th observation from the j-th column, so $Z_{j,k} \in [I]$. 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 common probability distribution $\boldsymbol{p}$ over $[I]$. This ordering $k = 1,\dots,n_j$ represents a random ordering of the counts observed and is for analysis purposes only. 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.

S.2.A. Proof of original p-value bound. The original p-value bound proposed in (5) is:

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

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

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

The p-value bound in (5) 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 $\hat{\mu}_j$. We then construct an estimate for the deviation of a column j from the table average by computing $S_j = \sqrt{n_j}(\hat{\mu}_j - \bar{\mu})$, 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:

$$\begin{aligned}
\bar{\mu} &= \frac{1}{M}\sum_{i,j} f_i X_{i,j} = \frac{1}{M}\sum_{j,k} f_{Z_{j,k}} \\
\hat{\mu}_j &= \frac{1}{n_j}\sum_{i,j} f_i X_{i,j} = \frac{1}{n_j}\sum_{k=1}^{n_j} f_{Z_{j,k}} \\
S_j &= \sqrt{n_j}(\hat{\mu}_j - \bar{\mu}) \\
S &= \sum_j c_j S_j
\end{aligned} \qquad [1]$$

To provide a p-value bound, we use Hoeffding's inequality for sums of independent, bounded random variables:

Lemma 1 (Hoeffding's inequality, Prop. 2.1 (6))**.** *Suppose that the random variables X_i, $i = 1, \dots, n$ are independent, and X_i has mean μ_i and $a_i \le X_i \le b_i$ almost surely. Then for all $t \ge 0$ we have:*

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

Taking the true (unknown) distribution of the $\{Z_{j,k}\}$ under the null to be $\boldsymbol{p}$, we define $\mu \triangleq \mathbb{E}_{Z\sim\boldsymbol{p}}[f_Z]$. Then, as in (5),

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

The first inequality comes from a union bound (if neither of the two conditions are met, then $|S| < \epsilon$. (a) assumes that the denominator $\sum_j c_j\sqrt{n_j} \ne 0$, 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,

$$Y_{j,k} \triangleq \frac{c_j}{\sqrt{n_j}}(f_{Z_{j,k}} - \mu) \implies \frac{-c_j\mu}{\sqrt{n_j}} \le Y_{j,k} \le \frac{c_j(1-\mu)}{\sqrt{n_j}} \tag{3}$$

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:

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

This leads to the stated p-value bound in (5) of

$$\mathbb{P}\left(|S(\boldsymbol{f},\boldsymbol{c})| \ge s\right) \le 2\exp\left(-\frac{2(1-\xi)^2s^2}{\sum_{j:n_j>0} c_j^2}\right) + 2\exp\left(-\frac{2\xi^2Ms^2}{\left(\sum_j c_j\sqrt{n_j}\right)^2}\right) \quad \text{with} \quad \xi = \left(1 + \sqrt{\frac{M\sum_j c_j^2}{\left(\sum_j c_j\sqrt{n_j}\right)^2}}\right)^{-1}. \tag{4}$$

S.2.B. Proof of improved p-values. We restate below the finite-sample p-value bound proved in the main text.

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

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

As before, we define $\bar{\mu} = \frac{1}{M}\boldsymbol{f}^\top X \mathbf{1}$ as the estimate of $\mu = \mathbb{E}_{Z\sim\boldsymbol{p}}[f_Z]$, where $\boldsymbol{p}$ is the unknown common row distribution under the null. Due to the structure of the test statistic, the p-value bound in Equation (4) can be improved and simplified. Fixing $\boldsymbol{f}, \boldsymbol{c}$, the test statistic can be simplified as

$$\begin{aligned}
S(\boldsymbol{f},\boldsymbol{c}) &= \sum_j c_j\sqrt{n_j}\,(\hat{\mu}_j - \bar{\mu}) \\
&= \left[\sum_j c_j\sqrt{n_j}\left(\frac{1}{n_j}\sum_{k=1}^{n_j} f_{Z_{j,k}}\right)\right] - \left[\sum_j c_j\sqrt{n_j}\left(\frac{1}{M}\sum_{\ell k} f_{Z_{\ell,k}}\right)\right] \\
&= \sum_{jk}\frac{c_j}{\sqrt{n_j}} f_{Z_{j,k}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\sum_{jk} f_{Z_{j,k}} \\
&= \sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right) f_{Z_{j,k}} \\
&= \sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right). \qquad [5]
\end{aligned}$$

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

Note that $\gamma = 1$ if and only if $\boldsymbol{c} = \pm\sqrt{\boldsymbol{n}/M}$ by Cauchy-Schwarz, in which case 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. Then $S(\boldsymbol{f},\boldsymbol{c}) = 0$ with probability 1.

If $\gamma < 1$, then examining each term in the summation in Eq. (5) yields

$$Y_{j,k} \triangleq \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right) \implies \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right|\left(\min_i f_i - \mu\right) \le Y_{j,k} \le \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right|\left(\max_i f_i - \mu\right). \qquad [6]$$

At right we are simply claiming that $Y_{j,k}$ is almost surely bounded with the provided range. 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.

$$\begin{aligned}
\mathbb{P}\left(|S(\boldsymbol{f},\boldsymbol{c})| \ge \epsilon\right) &= \mathbb{P}\left(\left|\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)\right| \ge \epsilon\right) \\
&\le 2\exp\left(-\frac{2\epsilon^2}{\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\sum_j\left(c_j - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\sqrt{n_j}\right)^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\sum_j\left[c_j^2 - 2c_j\sqrt{n_j}\frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M} + \frac{\left(\sum_\ell c_\ell\sqrt{n_\ell}\right)^2}{M^2}n_j\right]}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\sum_j c_j\sqrt{n_j}\right)^2}\right). \qquad [7]
\end{aligned}$$

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 $\frac{1}{4}$, as we discuss later. Thus, our final p-value bound is

$$\begin{aligned}\mathbb{P}\left(|S(\boldsymbol{f},\boldsymbol{c})| \geq s\right) &\leq 2\exp\left(-\frac{2s^2}{\left(\max_i f_i - \min_i f_i\right)^2\left(\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\sum_j c_j\sqrt{n_j}\right)^2\right)}\right)\\ &= 2\exp\left(-\frac{2s^2}{\left(\max_i f_i - \min_i f_i\right)^2\|\boldsymbol{c}\|^2(1-\gamma)}\right) \qquad [8]\end{aligned}$$

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

Proposition 2. *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*

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

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

S.2.C. Asymptotic distribution of OASIS test statistic. We begin by restating our asymptotic normality result.

Proposition 3 (Restatement of Prop. 2)**.** *Consider any fixed $\boldsymbol{f} \in \mathbb{R}^I, \boldsymbol{c} \in \mathbb{R}^J$, probability distribution $\boldsymbol{p} \in \Delta^I$ with $\sigma_{\boldsymbol{f}}^2 > 0$, and any sequence of column counts $\{\boldsymbol{n}^{(t)}\}_{t=1}^\infty$ where each $\boldsymbol{n}^{(t)} \in \mathbb{N}^J$, with $\min_{j\in[J]} n_j^{(t)} \overset{t\to\infty}{\longrightarrow} \infty$ and $\gamma(\boldsymbol{n}^{(t)},\boldsymbol{c}) < 1$ for all t. Then, the random sequence of OASIS test statistics $\{S_t\}_{t=1}^\infty$, where $S_t = S(X_t,\boldsymbol{f},\boldsymbol{c})$ and $X_t^{(j)} \sim \text{multinomial}(n_j^{(t)},\boldsymbol{p})$ independently across j and t, satisfies*

$$\frac{1}{\sqrt{1-\gamma(\boldsymbol{n}^{(t)},\boldsymbol{c})}} S_t \overset{D}{\to} \mathcal{N}\left(0, \sigma_{\boldsymbol{f}}^2\|\boldsymbol{c}\|^2\right).$$

We show that OASIS's test statistic is asymptotically normally distributed using the Lyapunov Central Limit Theorem.

Theorem 1 (Theorem 27.3 (7))**.** *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 $S_M/s_M \overset{D}{\to} Z$ where $Z \sim \mathcal{N}(0,1)$.

Proof. Continuing from Eq. (5), the test-statistic S can be expressed as

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

for an appropriate reindexing τ which maps j,k 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

$$\begin{aligned}s_M^2 &= \sum_{jk}\mathbb{E}\left[\left(\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)\right)^2\right]\\ &= \sigma_{\boldsymbol{f}}^2\sum_{jk}\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell\sqrt{n_\ell}}{M}\right)^2\\ &= \sigma_{\boldsymbol{f}}^2\left(\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\sum_\ell c_\ell\sqrt{n_\ell}\right)^2\right)\\ &= \sigma_{\boldsymbol{f}}^2\|\boldsymbol{c}\|^2(1-\gamma), \qquad [9]\end{aligned}$$

where we use the definition of $\sigma_{\boldsymbol{f}}^2 = \mathrm{Var}_{Z\sim\boldsymbol{p}}(f_Z) = \sum_i p_i f_i^2 - \left(\sum_i p_i f_i\right)^2$. This is greater than 0 by assumption in the Proposition statement, where we also assumed that γ was bounded away from 1. Thus, our desired result is

$$\frac{1}{\sqrt{1-\gamma}} S(\boldsymbol{f}, \boldsymbol{c}) \xrightarrow{D} \mathcal{N}\left(0, \|\boldsymbol{c}\|^2 \sigma_{\boldsymbol{f}}^2\right). \tag{10}$$

Satisfying Lyapunov's condition: we show that this condition holds for $\delta = 1$, which entails bounding third moments. We assume that $\max_i |f_i - \mu| \leq 1$, which is without loss of generality as $\boldsymbol{f}$ is fixed and everything can simply be rescaled, 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:

$$\left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right| \leq \left|\frac{c_j}{\sqrt{n_j}}\right| + \frac{\|\boldsymbol{c}\|}{\sqrt{M}} \leq \frac{2\|\boldsymbol{c}\|}{\min_j \sqrt{n_j}}. \tag{11}$$

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

$$\begin{aligned}
\sum_{\tau=1}^{M} \mathbb{E}\left[|X_{M\tau}|^3\right] &= \sum_{jk} \mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)\left(f_{Z_{j,k}} - \mu\right)\right|^3\right] \\
&= \sum_{jk} \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right|^3 \mathbb{E}\left[\left|f_{Z_{j,k}} - \mu\right|^3\right] \\
&\leq \sum_{jk} \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right|^3 \sigma_{\boldsymbol{f}}^2 \\
&\leq \left(\frac{2\|\boldsymbol{c}\|}{\min_j \sqrt{n_j}}\right) \sigma_{\boldsymbol{f}}^2 \left(\sum_{jk} \left|\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right|^2\right).
\end{aligned} \tag{12}$$

Verifying Lyapunov's condition:

$$\begin{aligned}
\frac{1}{s_M^3} \sum_{\tau=1}^{M} \mathbb{E}\left[|X_{M\tau}|^3\right] &= \frac{\sum_{jk} \mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right) f_{Z_{j,k}}\right|^3\right]}{\left(\sum_{jk} \mathbb{E}\left[\left|\left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right) f_{Z_{j,k}}\right|^2\right]\right)^{3/2}} \\
&\leq \frac{\left(\frac{2\|\boldsymbol{c}\|}{\min_j \sqrt{n_j}}\right) \sigma_{\boldsymbol{f}}^2 \left(\sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)^2\right)}{\left(\sigma_{\boldsymbol{f}}^2 \sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)^2\right)^{3/2}} \\
&= \frac{2\|\boldsymbol{c}\|}{\min_j \sqrt{n_j}} \left(\sigma_{\boldsymbol{f}}^2 \sum_{jk} \left(\frac{c_j}{\sqrt{n_j}} - \frac{\sum_\ell c_\ell \sqrt{n_\ell}}{M}\right)^2\right)^{-1/2} \\
&= \frac{2\|\boldsymbol{c}\|}{\sigma_{\boldsymbol{f}} \min_j \sqrt{n_j}} \left(\|\boldsymbol{c}\|^2 (1-\gamma)\right)^{-1/2} \\
&= \frac{2}{\sigma_{\boldsymbol{f}} \sqrt{(1-\gamma) \min_j n_j}}
\end{aligned} \tag{13}$$

Since the original quantity is nonnegative, this 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 we have the desired result. The variance of $S(\boldsymbol{f}, \boldsymbol{c})$ has already been computed as s_M^2, and so

$$\frac{1}{\sqrt{1-\gamma}} S(\boldsymbol{f}, \boldsymbol{c}) \xrightarrow{D} \mathcal{N}\left(0, \|\boldsymbol{c}\|^2 \sigma_{\boldsymbol{f}}^2\right). \tag{14}$$

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.

Corollary 3.1 (Restatement of Corollary)**.** *Under the conditions of Proposition 3,*

$$2\Phi\left(-\frac{|S|}{\hat{\sigma}_{\boldsymbol{f}}\|c\|\sqrt{1-\gamma}}\right),$$

is an asymptotically valid p-value.

For comparison with our finite-sample p-value bound, we construct an asymptotically valid p-value bound using standard Gaussian tail bounds as

$$\mathbb{P}(|S(\boldsymbol{f},\boldsymbol{c})| \geq s) = 2\Phi\left(-\frac{s}{\hat{\sigma}_{\boldsymbol{f}}\|c\|\sqrt{1-\gamma}}\right) \leq 2\exp\left(-\frac{s^2}{2\hat{\sigma}_{\boldsymbol{f}}^2\|\boldsymbol{c}\|^2(1-\gamma)}\right), \quad [15]$$

where Φ is the CDF of a standard Gaussian random variable. This holds by Slutsky's theorem, as $\hat{\sigma}_{\boldsymbol{f}}^2$ is an asymptotically consistent estimator of $\sigma_{\boldsymbol{f}} > 0$. Note that this form precisely matches that of the finite-sample p-value bound in Proposition 1, up to bounding $\hat{\sigma}_{\boldsymbol{f}}^2 \leq \frac{1}{4}$ (bounding the variance of a $[0,1]$ random variable as $1/4$). Since $f_{Z_{j,k}} \in [0,1]$, the rate of convergence can be quantified with the Berry-Esseen inequality (8) if desired.

S.2.D. Effect Size. The effect size measure for OASIS is motivated by a simple two group alternative, where each sample originate from either group A or group B. Each group has a characteristic target distribution, $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$ respectively, where observations from a sample in group A (resp. B) are drawn i.i.d. from $\boldsymbol{p}_A$ (resp. $\boldsymbol{p}_B$). Then, denoting Z 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 $\Delta = \left|\mathbb{E}_{Z\sim\boldsymbol{p}_A}\left[f_Z\right] - \mathbb{E}_{Z\sim\boldsymbol{p}_B}\left[f_Z\right]\right|$ the difference between $\mathbb{E}\left[f_Z\right]$ where $Z \sim \boldsymbol{p}_A$ or $Z \sim \boldsymbol{p}_B$. 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 quantity Δ is nonnegative, and is bounded above by the total variation distance between the empirical row distributions of columns where $c_j > 0$ and where $c_j < 0$. This is attained by taking the supremum over all $\boldsymbol{f} \in [0,1]^I$, yielding a value of 0 if and only if the empirical row distributions are the same between the two clusters, and 1 if and only if they are disjoint. Note that due to the rescaling of each point's contribution by $\sqrt{n_j}$ (downweighting) in OASIS's test statistic, for a given $\boldsymbol{c}$ constructing $\boldsymbol{f}$ to minimize the p-value bound will not necessarily maximize the effect size.

In the main text we stated that $0 \leq \hat{\Delta} \leq \delta_{\mathrm{TV}}(\hat{\boldsymbol{p}}_+, \hat{\boldsymbol{p}}_-)$, which we prove below.

Proof of Effect Size Bound. The lower bound on $\hat{\Delta}$ follows trivially from the absolute value. The upper bound follows from the reformulation of $\hat{\Delta}$ as:

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

We can rewrite this as an ℓ_1 norm as 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,

$$\max_{g:\|g\|_\infty \leq 1/2} g^\top x = \frac{1}{2}\|x\|_1,$$

by the dual norm characterization of the ℓ_1 norm. □

Considering general $\boldsymbol{c}$ beyond the simple binary case, there is no clear extension of this effect size measure. 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 centering. One candidate measure which encourages binary $\boldsymbol{f}$ is:

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

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

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

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 these can both be trivially achieved 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 $\tilde{X}$.

S.2.E. Statistical validity of data-splitting. The p-value bound outputted by the data-splitting optimization procedure is statistically valid, by a classical data-splitting argument (used in e.g. (9)), which we detail here for completeness. We begin by defining the random table generated by the null for probability vector $\boldsymbol{p}$ with column counts $\{n_j\}_{j=1}^J$. The data generated by the null is $X \sim \texttt{nullTable}(\{n_j\}_{j=1}^J, \boldsymbol{p})$. Data-splitting entails selecting $\{n_j^{\texttt{Train}}\}$ as a function of the given column counts $\{n_j\}_{j=1}^J$, where $n_j^{\texttt{Train}} \leq n_j$ for all j. In this procedure we select uniformly at random $n_j^{\texttt{Train}}$ counts uniformly at random from the n_j total counts in that column and assign them to X_{Train}, with the rest being assigned to X_{Test}. Then, $X_{\text{Train}} \sim \texttt{nullTable}(\{n_j\}_{j=1}^J, \boldsymbol{p})$, $X_{\text{Test}} \sim \texttt{nullTable}(\{n_j - n_j^{\texttt{Train}}\}_{j=1}^J, \boldsymbol{p})$, and the two are independent. This is due to the fact that each count in the multinomial $X^{(j)}$ is an i.i.d. draw from $\boldsymbol{p}$, and so splitting the counts randomly between $X_{\text{Test}}^{(j)}$ and $X_{\text{Train}}^{(j)}$ is simply separating these i.i.d. draws into two groups, resulting in a multinomial for each with the appropriate counts.

As shown by Proposition 1, for any fixed $\boldsymbol{f}, \boldsymbol{c}$, the random variable given by the p-value bound $p(X_{\text{Test}}, \boldsymbol{c}, \boldsymbol{f})$ (where only X_{Test} is random) is a statistically valid p-value bound. We show that even when $\boldsymbol{c} = \boldsymbol{c}(X_{\text{train}})$ and $\boldsymbol{f} = \boldsymbol{f}(X_{\text{train}})$ are functions of the training data, $p(X) = p(X_{\text{Test}}, \boldsymbol{c}(X_{\text{train}}), \boldsymbol{f}(X_{\text{train}}))$ is a valid p-value bound.

$$\begin{aligned}
\mathbb{P}(p(X) \leq u) &\overset{(a)}{=} \mathbb{P}(p(X_{\text{Test}}, \boldsymbol{c}(X_{\text{train}}), \boldsymbol{f}(X_{\text{train}})) \leq u) \\
&\overset{(b)}{=} \sum_{x_{\text{train}}} \mathbb{P}(p\left(X_{\text{Test}}, \boldsymbol{c}(X_{\text{train}}), \boldsymbol{f}(X_{\text{train}})\right) \leq u \mid X_{\text{train}} = x_{\text{train}}) \, \mathbb{P}(X_{\text{train}} = x_{\text{train}}) \\
&\overset{(c)}{=} \sum_{x_{\text{train}}} \mathbb{P}(p\left(X_{\text{Test}}, \boldsymbol{c}(x_{\text{train}}), \boldsymbol{f}(x_{\text{train}})\right) \leq u) \, \mathbb{P}(X_{\text{train}} = x_{\text{train}}) \\
&\overset{(d)}{\leq} \sum_{x_{\text{train}}} u \mathbb{P}(X_{\text{train}} = x_{\text{train}}) \\
&= u
\end{aligned} \tag{16}$$

where (a) holds by definition, (b) by law of total probability, (c) by the fact that X_{Test} is independent of X_{Train}, and (d) as once we condition on X_{Train}, $\boldsymbol{c}$ and $\boldsymbol{f}$ are no longer random, and so the p-value bound from the theorem statement can be used. Finally, the last line follows from summing over the probabilities. The resulting p-value bound is not just random with respect to the data X (which is to be expected), but also with respect to the random splitting procedure. However, since the p-value bound holds for any fixed $\boldsymbol{f}, \boldsymbol{c}$, once we condition on X_{Train}, the p-value bound can be applied.

Note that more generally, there are two sources of randomness used in the algorithm; one from splitting the data into train and test sets, and one from the generation of $\boldsymbol{c}(X_{\text{train}})$ and $\boldsymbol{f}(X_{\text{train}})$. The first source (splitting into train and test sets) 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 exactly solve the inner optimization problem deterministically. However, the randomness in data-splitting is necessary, and can greatly impact performance. If the training data is not representative of the test data, then even if the optimal $\boldsymbol{f}$ and $\boldsymbol{c}$ for the training data are identified, it need not yield a significant p-value bound on the test data. Thus, a natural approach is to generate multiple random splits of the data, perform Bonferroni correction over these multiple splits, and take the minimum as a valid p-value bound. To this end, OASIS generates independent random splits of the data by using the Poisson distribution of the counts under the null (assuming the column counts are Poisson). The Poisson nature of these counts is critically important, as under negative binomial overdispersion the counts observed are no longer independent, and so more sophisticated methods are needed, which do not yield fully independent sets of counts unless the overdispersion parameter is known (10).

S.3. Optimization procedure

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, where we constrain $0 \leq \boldsymbol{f} \leq 1$:

$$2 \exp\left(-\frac{2S^2}{\|\boldsymbol{c}\|^2 (1-\gamma)}\right). \tag{17}$$

For simplicity, we shift and rescale $\boldsymbol{f}$ to be bounded as $|\boldsymbol{f}| \leq 1$, which does not change the structure of the problem: an optimal solution for the original problem can be obtained by modifying the solution to the shifted and rescaled problem.

S.3.A. Reformulating the optimization problem (proof of Lemma 1). Recalling the containment lemma in the main text:

Lemma 2 (Restatement of Lemma 1)**.** *The set of optimal solutions for the p-value bound can be expressed as*

$$\left(\operatorname*{argmax}_{0 \leq \boldsymbol{f} \leq 1, \|c\|_2 \leq 1} \boldsymbol{f}^\top \tilde{X} \boldsymbol{c}\right) \subseteq \operatorname*{argmin}_{0 \leq \boldsymbol{f} \leq 1, \|c\|_2 \leq 1} 2 \exp\left(-\frac{2\left(\boldsymbol{f}^\top \tilde{X} \boldsymbol{c}\right)^2}{1 - \frac{1}{M}\left\langle \boldsymbol{c}, \sqrt{\boldsymbol{n}} \right\rangle^2}\right)$$

We prove this lemma by decomposing $\boldsymbol{c}$ into the part parallel to $\sqrt{\boldsymbol{n}} \triangleq \sqrt{X^\top \mathbf{1}}$ and the part orthogonal to $\sqrt{\boldsymbol{n}}$, i.e. $\boldsymbol{c} = \boldsymbol{c}^{(1)} + \alpha \sqrt{\boldsymbol{n}}$ where $\boldsymbol{c}^{(1)} \perp \sqrt{\boldsymbol{n}}$ and $\alpha \in \mathbb{R}$. This yields an objective value of

$$\begin{aligned}\frac{\left(\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}\right)^2}{\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\boldsymbol{c}^\top\sqrt{\boldsymbol{n}}\right)^2} &= \frac{\left(\boldsymbol{f}^\top\left(\tilde{X}\left(\boldsymbol{c}^{(1)}+\alpha\sqrt{\boldsymbol{n}}\right)\right)\right)^2}{\|\boldsymbol{c}^{(1)}+\alpha\sqrt{\boldsymbol{n}}\|^2 - \frac{1}{M}\left(\alpha\sqrt{\boldsymbol{n}}^\top\sqrt{\boldsymbol{n}}\right)^2} \\ &= \frac{\left(\boldsymbol{f}^\top\left(\tilde{X}\boldsymbol{c}^{(1)}+\alpha\tilde{X}\sqrt{\boldsymbol{n}}\right)\right)^2}{\|\boldsymbol{c}^{(1)}\|^2+\alpha^2\|\sqrt{\boldsymbol{n}}\|^2 - \frac{1}{M}\left(\alpha\|\sqrt{\boldsymbol{n}}\|^2\right)^2} \\ &= \frac{\left(\boldsymbol{f}^\top\tilde{X}\boldsymbol{c}^{(1)}\right)^2}{\|\boldsymbol{c}^{(1)}\|^2+\alpha^2 M - \frac{1}{M}\left(\alpha M\right)^2} \\ &= \frac{\left(\boldsymbol{f}^\top\tilde{X}\boldsymbol{c}^{(1)}\right)^2}{\|\boldsymbol{c}^{(1)}\|^2} \end{aligned} \quad [18]$$

As

$$\tilde{X}\sqrt{\boldsymbol{n}} = \left(X - \frac{1}{\mathbf{1}^\top X\mathbf{1}}X\mathbf{1}\mathbf{1}^\top X\right)\mathrm{diag}(1/\sqrt{\boldsymbol{n}})\sqrt{\boldsymbol{n}} = \left(X - \frac{1}{\mathbf{1}^\top X\mathbf{1}}X\mathbf{1}\mathbf{1}^\top X\right)\mathbf{1} = 0$$

Thus, any component of $\boldsymbol{c}$ in the direction of $\sqrt{\boldsymbol{n}}$ does not contribute to the numerator, and cancels out in the denominator, and so an optimal solution can be obtained by simply optimizing over $\boldsymbol{c}$ orthogonal to $\sqrt{\boldsymbol{n}}$. Since the denominator is equal to $\|\boldsymbol{c}\|^2$, and the numerator and denominator both scale quadratically in $\|\boldsymbol{c}\|$, we can simply optimize the numerator subject to $\|\boldsymbol{c}\| \leq 1$ to identify a pair of optimal $\boldsymbol{f}, \boldsymbol{c}$, as:

$$\underset{\boldsymbol{c}\in\mathbb{R}^J, \boldsymbol{f}\in[0,1]^I}{\mathrm{argmax}} \frac{\left(\boldsymbol{f}^\top\tilde{X}\boldsymbol{c}\right)^2}{\|\boldsymbol{c}\|^2 - \frac{1}{M}\left(\boldsymbol{c}^\top\sqrt{\boldsymbol{n}}\right)^2} \supseteq \underset{\boldsymbol{f}\in[0,1]^I, \|\boldsymbol{c}\|_2\leq 1}{\mathrm{argmax}} \left(\boldsymbol{f}^\top\tilde{X}\boldsymbol{c}\right)^2 \quad [19]$$

$$\supseteq \underset{\boldsymbol{f}\in[0,1]^I, \|\boldsymbol{c}\|_2\leq 1}{\mathrm{argmax}} \boldsymbol{f}^\top\tilde{X}\boldsymbol{c}. \quad [20]$$

where the last line follows as $\|\boldsymbol{c}\| \leq 1 \implies \|-\boldsymbol{c}\| \leq 1$. This proves Lemma 1.

To optimize Eq. (20), we begin by enlarging the constraint set of $\boldsymbol{f}$ to be $[-1,1]^I$ for symmetry (an optimal $\boldsymbol{f}$ for the original problem can be obtained from such a $\tilde{\boldsymbol{f}}$ by considering $\boldsymbol{f} = (1+\tilde{\boldsymbol{f}})/2$ and $\boldsymbol{f} = (1-\tilde{\boldsymbol{f}})/2$, optimizing $\boldsymbol{c} \propto \tilde{X}^\top\boldsymbol{f}$, and comparing their objective values). For fixed $\boldsymbol{c}$, $\boldsymbol{f}$ is then optimized as $\boldsymbol{f} = \mathrm{sign}(\tilde{X}\boldsymbol{c})$, where $\mathrm{sign}(x)$ is $+1$ if $x > 0$, 0 if $x = 0$, and -1 if $x < 0$. $\boldsymbol{c}$ is always optimized as $\boldsymbol{c} \propto \tilde{X}^\top\boldsymbol{f}$. Rolling these updates together, since the sign of an entry doesn't change based on scaling we can write our iterates $\{\boldsymbol{f}^{(t)}\}$ as

$$\boldsymbol{f}^{(t+1)} = \mathrm{sign}\left(\tilde{X}\tilde{X}^\top\boldsymbol{f}^{(t)}\right), \quad [21]$$

where the iterates $\{\boldsymbol{c}^{(t)}\}$ are implicitly computed as $\boldsymbol{c}^{(t)} \propto \tilde{X}^\top\boldsymbol{f}^{(t)}$.

S.3.B. 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 single sufficiently large direction of deviation. We show that a simple extension of OASIS allows for detection of how many ways the contingency table significantly deviates from the null, through a statistical stopping condition.

Consider that one initial partitioning vector $\boldsymbol{c}^{(1)}$ 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 $\boldsymbol{c}^{(1)}$. One way of accomplishing this is by attempting to identify new vectors $\boldsymbol{c}, \boldsymbol{f}$ which yield a significant p-value bound, requiring that this new vector $\boldsymbol{c}$ be orthogonal to the previously found vector $\boldsymbol{c}^{(1)}$. With SARS-CoV-2, this could be a vector separating Omicron BA.1 and BA.2. The constraint $\boldsymbol{c}\perp\boldsymbol{c}^{(1)}$ 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 analogous to identifying the second largest singular value and its corresponding right singular vector in the case of the SVD. 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 bound parallels analyzing the spectrum of $\tilde{X}$ until a significant drop-off occurs. In this iterative setting, the optimizing $\boldsymbol{c}$ is constructed as

$$\boldsymbol{c}^\star = \underset{\substack{\|\boldsymbol{c}\|_2\leq 1,\\ \boldsymbol{c}\perp\boldsymbol{c}^{(1)}}}{\mathrm{argmax}} \boldsymbol{f}^\top\tilde{X}\boldsymbol{c} \implies \boldsymbol{c}^* \propto \tilde{X}^\top\boldsymbol{f} - \frac{\boldsymbol{f}^\top\tilde{X}\boldsymbol{c}^{(1)}}{\|\boldsymbol{c}^{(1)}\|_2^2}\boldsymbol{c}^{(1)}. \quad [22]$$

$\boldsymbol{f}$ is constructed as before, with iterates constructed as $\boldsymbol{f} = \mathrm{sign}(\tilde{X}\boldsymbol{c})$.

Algorithm 1 `OASIS-iter`

Input: Contingency table X
 # Split data randomly into train and test portions
$X_{\text{train}}, X_{\text{test}} \leftarrow$ `splitDataset`(X)
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 $[\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(2)}, \ldots],\ [\boldsymbol{f}^{(1)}, \boldsymbol{f}^{(2)}, \ldots],\ [p^{(1)}, p^{(2)}, \ldots]$

S.3.C. Asymptotic p-value. The asymptotic p-value provided in Corollary 3.1 leads to the following optimization objective:

$$\underset{\boldsymbol{f},\mathbf{c}}{\operatorname{argmax}} \frac{\left(\boldsymbol{f}^\top \tilde{X} \boldsymbol{c}\right)^2}{\|\boldsymbol{c}\|^2 \hat{\sigma}_{\boldsymbol{f}}^2}. \tag{23}$$

Here, $\hat{\sigma}_{\boldsymbol{f}}^2 = \sum_i \hat{p}_i f_i^2 - \left(\sum_i \hat{p}_i f_i\right)^2 = \sum_i \hat{p}_i (f_i - \hat{\boldsymbol{p}}^\top \boldsymbol{f})^2 = \boldsymbol{f}^\top \left(\operatorname{diag}(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^\top\right) \boldsymbol{f}$ and $\hat{\boldsymbol{p}} \triangleq \frac{1}{M} X\mathbf{1}$. Note that $\hat{\boldsymbol{p}} > 0$, as otherwise these rows would not exist in X. For a fixed $\boldsymbol{f}$, the optimal $\boldsymbol{c}$ is easily computed as $\boldsymbol{c}^\star \propto \tilde{X}^\top \boldsymbol{f}^\star$, by Cauchy-Schwarz. This allows us to reduce the optimization problem to that of optimizing over $\boldsymbol{f}$ by dual norm characterizations as

$$\max_{\boldsymbol{f},\boldsymbol{c}} \frac{\left(\boldsymbol{f}^\top \tilde{X} \boldsymbol{c}\right)^2}{\|\boldsymbol{c}\|^2 \hat{\sigma}_{\boldsymbol{f}}^2} = \max_{\boldsymbol{f}} \frac{\left\|\tilde{X}^\top \boldsymbol{f}\right\|^2}{\hat{\sigma}_{\boldsymbol{f}}^2}. \tag{24}$$

We show that an optimal $\boldsymbol{f}$ for this objective can be computed efficiently as an eigenvector problem. Defining $v_{\max}(A)$ as the principal eigenvector of a symmetric matrix A (selecting an arbitrary one if the maximum eigenvalue has multiplicity greater than 1), we state the following Proposition.

Proposition 4. *An optimal $\boldsymbol{f}^\star$ for* Eq. (24) *can be constructed as*

$$\boldsymbol{f}^\star = \operatorname{diag}(\hat{\boldsymbol{p}})^{-1/2} v_{max}\left(\operatorname{diag}(\hat{\boldsymbol{p}})^{-1/2} \tilde{X} \tilde{X}^\top \operatorname{diag}(\hat{\boldsymbol{p}})^{-1/2}\right).$$

Proof. Observe that the matrix used in the computation of $\hat{\sigma}_{\boldsymbol{f}}^2$ is positive semidefinite, and that restricting $\boldsymbol{f} \perp \hat{\boldsymbol{p}}$ the resulting operator is positive definite:

$$\begin{aligned} \operatorname{diag}(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^\top &\succeq 0 \\ \operatorname{diag}(\hat{\boldsymbol{p}}) &\succ 0 \end{aligned}$$

To identify an optimal $\boldsymbol{f}$ for Eq. (24), we show that the objective is scale and shift invariant with respect to $\boldsymbol{f}$. Formally, defining the maximization objective of Eq. (24) as

$$h(\boldsymbol{f}) \triangleq \frac{\left\|\tilde{X}^\top \boldsymbol{f}\right\|^2}{\hat{\sigma}_{\boldsymbol{f}}^2}, \tag{25}$$

we claim that for any $\beta \in \mathbb{R}$, $h(\boldsymbol{f} + \beta\mathbf{1}) = h(\boldsymbol{f})$. It is clear that

$$\hat{\sigma}_{\boldsymbol{f}+\beta\mathbf{1}}^2 = \sum_i \hat{p}_i (f_i + \beta - \hat{\boldsymbol{p}}^\top (\boldsymbol{f} + \beta\mathbf{1}))^2 = \sum_i \hat{p}_i (f_i - \hat{\boldsymbol{p}}^\top \boldsymbol{f})^2 = \hat{\sigma}_{\boldsymbol{f}}^2, \tag{26}$$

and since $\mathbf{1}^\top X \mathbf{1} = M$,

$$\begin{aligned} (\boldsymbol{f} + \beta\mathbf{1})^\top \tilde{X} &= \boldsymbol{f}^\top \tilde{X} + \beta \mathbf{1}^\top \left(X - \frac{1}{M} X\mathbf{1}\mathbf{1}^\top X\right) \operatorname{diag}(1/\sqrt{X^\top \mathbf{1}}) \\ &= \boldsymbol{f}^\top \tilde{X} + \beta \left(\mathbf{1}^\top X - \frac{\mathbf{1}^\top X \mathbf{1}}{M} \mathbf{1}^\top X\right) \operatorname{diag}(1/\sqrt{X^\top \mathbf{1}}) \\ &= \boldsymbol{f} \tilde{X}. \end{aligned} \tag{27}$$

Combining eqs. (26) and (27) together yields that

$$h(\boldsymbol{f} + \beta\mathbf{1}) = h(\boldsymbol{f}). \tag{28}$$

Additionally, the maximization objective is scale invariant. For any scalar $\kappa \neq 0$:

$$\begin{aligned} h(\kappa \boldsymbol{f}) &= \max_{\boldsymbol{f}} \frac{\|\tilde{X}^\top(\kappa \boldsymbol{f})\|^2}{(\kappa \boldsymbol{f})^\top \left(\operatorname{diag}(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^\top\right)(\kappa \boldsymbol{f})} \\ &= \max_{\boldsymbol{f}} \frac{\kappa^2 \|\tilde{X}^\top \boldsymbol{f}\|^2}{\kappa^2 \boldsymbol{f}^\top \left(\operatorname{diag}(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^\top\right) \boldsymbol{f}} \\ &= h(\boldsymbol{f}) \end{aligned} \tag{29}$$

Since the numerator and denominator both scale quadratically in the magnitude of $\boldsymbol{f}$, the denominator ($\hat{\sigma}^2_{\boldsymbol{f}}$) can instead be treated as a constraint, i.e.

$$\max_{\boldsymbol{f}} \frac{\|\tilde{X}^\top \boldsymbol{f}\|^2}{\hat{\sigma}^2_{\boldsymbol{f}}} = \max_{\boldsymbol{f}:\hat{\sigma}^2_{\boldsymbol{f}} \leq 1} \|\tilde{X}^\top \boldsymbol{f}\|^2 \tag{30}$$

$$= \max_{\boldsymbol{f}:\boldsymbol{f}^\top \operatorname{diag}(\hat{\boldsymbol{p}})\boldsymbol{f} \leq 1, \boldsymbol{f}^\top \hat{\boldsymbol{p}} = 0} \boldsymbol{f}^\top \tilde{X}\tilde{X}^\top \boldsymbol{f} \tag{31}$$

The right hand side of Eq. (30) is clearly at most the left hand side, and an optimal $\boldsymbol{f}$ for the left hand side can be rescaled so that $\hat{\sigma}^2_{\boldsymbol{f}} = 1$, yielding a feasible $\boldsymbol{f}$ for the right hand side with the same objective value. Eq. (31) follows from the shift invariance in Equations (26) and (27).

We now show that this can be reformulated as an eigenvector problem. Defining $D = \operatorname{diag}(\hat{\boldsymbol{p}})$ and $\boldsymbol{y} = D^{1/2}\boldsymbol{f}$:

$$\begin{aligned} \max_{\substack{\boldsymbol{f}:\boldsymbol{f}^\top \operatorname{diag}(\hat{\boldsymbol{p}})\boldsymbol{f} \leq 1 \\ \boldsymbol{f}^\top \hat{\boldsymbol{p}} = 0}} \boldsymbol{f}^\top \tilde{X}\tilde{X}^\top \boldsymbol{f} &= \max_{\substack{\boldsymbol{y}:\|\boldsymbol{y}\|^2 \leq 1 \\ \boldsymbol{y}^\top D^{-1/2}\hat{\boldsymbol{p}} = 0}} \boldsymbol{y}^\top D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\boldsymbol{y} \\ &\leq \max_{\boldsymbol{y}:\|\boldsymbol{y}\|^2 \leq 1} \boldsymbol{y}^\top D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\boldsymbol{y} \\ &= \|D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\|. \end{aligned} \tag{32}$$

The upper bound is due to removing the orthogonality constraint. The final equality is attained by taking $\boldsymbol{y}$ as the principal eigenvector of the given matrix scaled to have unit norm, with objective value equal to the largest eigenvalue $\lambda^\star$ as below:

$$\boldsymbol{y}^\star = v_{\max}\left(D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\right), \tag{33}$$

$$\lambda^\star = \lambda_{\max}\left(D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\right). \tag{34}$$

This upper bound in Eq. (32) is attained with equality, as for this optimizing $\boldsymbol{y}^\star$, $\boldsymbol{y}^\star \perp \hat{\boldsymbol{p}}$:

$$\begin{aligned} \hat{\boldsymbol{p}}^\top D^{-1/2}\boldsymbol{y}^\star &= \hat{\boldsymbol{p}}^\top D^{-1/2}\left(\frac{1}{\lambda^\star} D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\boldsymbol{y}^\star\right) \\ &= \frac{1}{\lambda^\star}\hat{\boldsymbol{p}}^\top D^{-1}\tilde{X}\tilde{X}^\top D^{-1/2}\boldsymbol{y}^\star \\ &= \frac{1}{\lambda^\star}\mathbf{1}^\top \tilde{X}\tilde{X}^\top D^{-1/2}\boldsymbol{y}^\star \\ &= 0. \end{aligned} \tag{35}$$

Thus, since the unconstrained optimum lies within the constraint set, it is also the constrained optimum. An optimizing $\boldsymbol{f}$ for Eq. (30) can then be constructed as

$$\boldsymbol{f}^\star = \operatorname{diag}(\hat{\boldsymbol{p}})^{-1/2} v_{\max}\left(\operatorname{diag}(\hat{\boldsymbol{p}})^{-1/2}\tilde{X}\tilde{X}^\top \operatorname{diag}(\hat{\boldsymbol{p}})^{-1/2}\right). \tag{36}$$

Since the constructed $\boldsymbol{y}$ has $\|\boldsymbol{y}\| = 1$, $\hat{\sigma}^2_{\boldsymbol{f}} = 1$. Thus, this $\boldsymbol{f}^\star$ is also optimal for the original p-value objective in Eq. (23). □

To construct an optimal solution for Eq. (23), we observe that by Cauchy-Schwarz for any fixed $\boldsymbol{f}$ the maximum is attained when $\boldsymbol{c} \propto \tilde{X}^\top \boldsymbol{f}$. Thus, using the computed $\boldsymbol{f}^\star$ above and taking $\boldsymbol{c}^\star$ as

$$\boldsymbol{c}^\star = \tilde{X}^\top \boldsymbol{f}^\star \tag{37}$$

yields an optimal solution for Eq. (23).

A previous version of this work stated that the optimal $\boldsymbol{f}$ could be computed as a principal right eigenvector of the matrix $D^{-1}\tilde{X}\tilde{X}^\top$. This is equivalent (assuming without loss of generality that the principal eigenvalue has multiplicity 1):

$$\begin{aligned}
\boldsymbol{f} &= D^{-1}\tilde{X}\tilde{X}^\top \boldsymbol{f} \\
\boldsymbol{f} &= D^{-1/2}(D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2})D^{1/2}\boldsymbol{f} \\
D^{1/2}\boldsymbol{f} &= (D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2})D^{1/2}\boldsymbol{f} \\
D^{1/2}\boldsymbol{f} &= v_{\max}\left(D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\right) \\
\boldsymbol{f} &= D^{-1/2}v_{\max}\left(D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}\right)
\end{aligned}$$

as desired. We provide a method for modifying this vector to improve its finite sample p-value bound performance in Section S.3.C.1, and provide computational details on optimizing $\boldsymbol{f}$ in Section S.3.D.1.

S.3.C.1. Modification for finite-sample bound. The $\boldsymbol{f}$ optimizing the asymptotic p-value, derived in the previous section, will in general be suboptimal for the finite sample p-value bound. However, if the objective is to optimize the finite-sample bound, the asymptotically optimal $\boldsymbol{f}^\star$ can be used as an initialization. One simple approach would be to binarize the vector as $f_i' = 1$ if $f_i^\star \geq 0$, and 0 otherwise. For this $\boldsymbol{f}'$, the optimal $\boldsymbol{c}$ can be identified as $\boldsymbol{c} \propto \tilde{X}^\top \boldsymbol{f}'$. Introducing randomness, $\boldsymbol{f}^\star$ can be randomly rounded to $0, 1^I$ with probabilities proportional to the corresponding entry of $\boldsymbol{f}$ (after normalizing $\boldsymbol{f}^\star$ vector to $[0,1]^I$), allowing for multiple trials and a better final objective value.

S.3.D. More computationally intensive optimization methods. Analyzing the reformulated finite-sample optimization problem, observe that the form of the finite-sample p-value bound objective is simply a quadratic program:

$$\max_{\boldsymbol{f}\in[0,1]^I} \boldsymbol{f}^\top \tilde{X}\tilde{X}^\top \boldsymbol{f}. \qquad [38]$$

Since $\tilde{X}\tilde{X}^\top \gtrsim 0$ is positive semi-definite, this is an integer constrained quadratic program. This is a known combinatorial optimization problem, and specialized algorithms for optimizing it have been devised. One well-known approach is based on semidefinite programming (SDP) relaxations, relaxing the vector $\boldsymbol{f}$ into an $I \times I$ matrix optimization variable A and optimizing

$$\max_{A\gtrsim 0, A_{ii}=1\forall i} \operatorname{Tr}\left(\tilde{X}\tilde{X}^\top A\right), \qquad [39]$$

where Tr denotes the matrix Trace operator. The objective value of Eq. (39) upper bounds that of Eq. (38). We can convert the solution to Eq. (39) to an integer solution via Goemans-Williamson rounding (11), which picks a random vector $\boldsymbol{v}$ on the unit sphere in $\mathbb{R}^I$, and assigns $\boldsymbol{f}_i = 1$ if $A_i\boldsymbol{v} \geq 0$ and $\boldsymbol{f}_i = -1$ otherwise. This attains an expected approximation ratio of at least $2/\pi$, for Eq. (38). Multiple random $\boldsymbol{v}$ can be selected to slightly improve the resulting solution.

In practice however, solving an SDP is computationally intensive; in the biological setting of interest, contingency tables have many rows due to sequencing error in observing the targets. This makes methods whose sample complexities scale superlinearly in the number of rows too computationally intensive to use. Future work on computationally efficient methods with provable approximation guarantees can be readily utilized in this optimization framework; the primary contribution of OASIS is in its formulation of minimizing the p-value as an optimization problem, for which we propose an alternating maximization-based solution which empirically yields good performance in an efficient manner.

S.3.D.1. Empirical performance. We study the performance of these different optimization methods by simulating the "strong signal" setting described in Section S.6.A. The tables generated have 12 rows and 10 columns, and a clear two group structure, with signal focused in the first two rows of the tables. When run on 1000 random tables, the aforementioned SDP relaxation followed by rounding attained OPT 88% of the time. Alternating maximization yields similarly good performance, attaining OPT over 82% of the time. Computing the asymptotically optimal $\boldsymbol{c}, \boldsymbol{f}$ and rounding (Section S.3.C.1) attains OPT 72% of the time on this synthetic dataset. Computing $\boldsymbol{f}, \boldsymbol{c}$ from the SVD of $\tilde{X}$ yields OPT 0% of the time, even if we postprocess $\boldsymbol{f}$ by rounding its entries

While solving the SDP relaxation yields slightly better performance than alternating maximization, it is significantly more computationally intensive and requires more complicated machinery (cvxpy (12)). To test the different methods' computational complexities, we increase the counts per column from 10 to 100 to make sure that all rows have counts. We then test for various numbers of rows, to show the poor scaling of SDP solvers. Results are described in Table S2. The asymptotic approach identifies the principal eigenvector of $D^{-1/2}\tilde{X}\tilde{X}^\top D^{-1/2}$ where $D = \operatorname{diag}(\boldsymbol{p})$, which we naively compute with an eigendecomposition. However, using power iteration, only matrix vector products with $\tilde{X}, \tilde{X}^\top$ are required, which can be much more efficient.

S.3.E. Comparison with the SVD. The 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 computes

$$\operatorname*{argmax}_{\boldsymbol{c},\boldsymbol{f}} \frac{\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}}{\|\boldsymbol{f}\|_2\|\boldsymbol{c}\|_2},$$

as opposed to the p-value bound objective of

$$\operatorname*{argmax}_{\boldsymbol{c},\boldsymbol{f}} \frac{\boldsymbol{f}^\top \tilde{X}\boldsymbol{c}}{\|\boldsymbol{f}\|_\infty\|\boldsymbol{c}\|_2}.$$

An optimal $\boldsymbol{f}$ exists 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 binarizing $\boldsymbol{f}$ according to the entries of $\tilde{X}\boldsymbol{c}$. As discussed in the previous section, computing $\boldsymbol{f}, \boldsymbol{c}$ from the SVD of $\tilde{X}$ yields OPT 0% of the time, even if we postprocess $\boldsymbol{f}$ by rounding (binarizing) its entries, as it is still not optimizing the desired objective. We take a closer look at a toy example where an SVD does not yield the desired partitioning is shown in Figure S8.

S.4. SARS-CoV-2 analysis details

The SARS-CoV-2 analysis was based on dehosted sequencing data of nasopharyngeal swabs with coinfections available under accession PRJNA817806 (2). SPLASH contingency tables were tested using `OASIS-opt` with a 25% train/test data split. We then utilize anchors whose BY corrected p-value bound 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. Beyond the aforementioned thresholds, there are additional potential avenues towards filtering that we did not utilize or fine-tune, 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 splits 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 (13). We used four SARS-CoV-2 reference genomes for the strains circulating during this time period. We manually constructed "archetype" genomes for Delta, Omicron BA.1, and Omicron BA.2 by editing the Original (Wuhan) reference (NCBI assembly NC_045512.2) to contain all (and only) the defining mutations of that strain, detailed and made publicly available in (5). A bowtie index comprised of the four SARS-CoV-2 genomes was constructed (bowtie-build on a FASTA file containing four 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 file is aligned to the bowtie index using (-v 0 -a) options to ensure that we only get exact matches, and record all possible matches.

S.4.A. Embedding-aggregation. To aggregate the $\boldsymbol{c}^{(a)}$ vectors obtained from experiments on different anchors a (different tables), one natural objective is

$$\begin{aligned}\boldsymbol{c}^\star &= \underset{\|\boldsymbol{v}\|\le 1}{\operatorname{argmax}} \sum_a \langle \boldsymbol{c}^{(a)}, \boldsymbol{v}\rangle^2 \\ &= \underset{\|\boldsymbol{v}\|\le 1}{\operatorname{argmax}} \|\boldsymbol{C}\boldsymbol{v}\|_2 .\end{aligned}$$

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 principal right singular vector of $\boldsymbol{C}$. To visualize the data further, we also plot the second singular vector, which can be understood as finding a vector with maximal sum of squared inner products with $\boldsymbol{c}^{(a)}$ subject to being orthogonal to the first $\boldsymbol{c}$ identified, and see that this separates Omicron BA.1 from BA.2.

We provide additional analysis in Figure S9, where we show the spectrum of the embedding-aggregation matrix, the p-values generated by OASIS-iter, and

S.4.B. Counts-aggregation. Counts-aggregation iteratively identifies orthogonal $\boldsymbol{c}$ vectors on the stacked contingency tables. To merge the tables, we discard all targets that have fewer than 10 counts in a given table. Then, we normalize each table by $1/\sqrt{M}$, its total number of counts, to ensure that high count tables do not completely outweigh low count ones, and stack the resulting tables together. This yields a visually similar clustering to embedding-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).

For predicting target strain, 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, constituting 2,479 of the 2,495 anchors tested.

S.5. OASIS extensions

In this section we discuss several extensions of OASIS.

S.5.A. Metadata-based construction of $\boldsymbol{f}, \boldsymbol{c}$. 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 $\{A,T,C,G\}^{27}$ (5)). 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. From a different direction, to improve interpretability, regularization can be imposed on $\boldsymbol{f}$ in the optimization formulations to yield sparser vectors resulting in more parsimonious descriptions of deviations between samples.

Similarly, in the presence of sample metadata (e.g. cell-type) in the form of a vector $\tilde{\boldsymbol{c}} \in [-1,1]^J$, the user may wish to incorporate this information into constructing $\boldsymbol{c}$. In the simplest case, the metadata is fully trusted, and we want to see if this sample partitioning yields a significant p-value bound. In this case, $\tilde{\boldsymbol{c}}$ is utilized, along with the optimizing $\boldsymbol{f}^\star$ from the split data, 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:

$$\boldsymbol{c}^\star = \underset{\substack{c \,:\, \|c\| \le 1, \\ c_j \tilde{c}_j \ge 0 \;\forall\; j}}{\operatorname{argmax}} \left| \boldsymbol{f}^\top \tilde{X} \boldsymbol{c} \right|. \qquad [40]$$

While the overall objective is non-convex, it is still biconvex. The maximization of $\boldsymbol{f}$ is the same as before, and $\boldsymbol{c}$ is now maximized as, defining $\vec{S} = \boldsymbol{f}^\top \tilde{X}$:

$$c_j \propto \begin{cases} S_j & \text{if } \tilde{c}_j S_j \ge 0, \\ 0 & \text{if } \tilde{c}_j S_j < 0. \end{cases} \quad \text{or} \quad c_j \propto \begin{cases} -S_j & \text{if } \tilde{c}_j S_j \le 0, \\ 0 & \text{if } \tilde{c}_j S_j > 0. \end{cases} \qquad [41]$$

This enables efficient approximate optimization via alternating maximization. Many other forms of metadata (e.g. category of cell-type beyond just binary) can yield different metadata-aided approaches, and can be framed as optimization problems that admit efficient alternating maximization solutions.

S.5.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 non-parametric measure of dependence was proposed, xicor (14), with good theoretical properties and empirical performance. OASIS can also be used to test independence 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 bound against the null of independence. This empirically performs quite well, and can yield an effect size estimate comparable to xicor, as shown in Figure S16. The example functions used are detailed in (14) and are plotted at left, showing that OASIS has additional power against balanced, structured alternatives, such as circular and heteroscedastic.

S.5.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_Z where $Z \sim \boldsymbol{p}$, and simply rely on the boundedness of f_Z. One natural concentration inequality that utilizes both the boundedness of a random variable as well as its variance is Bernstein's inequality, of which an empirical variant was derived in (15). An improved p-value bound can potentially be obtained by bounding the true variance $\sigma^2_{\boldsymbol{f}}$ with high probability by some function of the empirical variance $\hat{\sigma}^2_{\boldsymbol{f}}$ on the training set, and utilizing Bernstein's inequality instead of Hoeffding's, applied to the same test statistic. Preliminary analysis shows that in low count regimes this approach yields much worse bounds, as the variance estimate concentrates much slower than the mean estimate, but improved bounds and analyses could yield better performance.

S.5.D. Tensor analysis. In computational genomics, it is of high priority to discover effects or partitionings that are reproducible. In the setting of (5) this motivates going beyond treating each table independently, and instead analyzing the anchor by sample by target tensor. OASIS's framework enables this, where $\boldsymbol{f}$ and $\boldsymbol{c}$ can be generated and optimized 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.

S.6. Additional discussion

For brevity and flow we defer some additional plots from the main text to here.

S.6.A. Comparison of OASIS and X^2 in planted setting. In this section we define and provide additional simulations (Figure S1) for the example in main text Figure 1. 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). In our uncorrupted planted model with $\epsilon = 0$, 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 column's probability distribution with another probability distribution $\boldsymbol{q}^{(j)}$, which is generated independently for each column. Concretely,

$$\boldsymbol{p}^{(j)} = \begin{cases} e_1 + \epsilon \boldsymbol{q}^{(j)} & \text{if } j \le J/2, \\ e_2 + \epsilon \boldsymbol{q}^{(j)} & \text{if } j > J/2. \end{cases} \qquad [42]$$

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 independently generated by taking each entry to be independently uniformly distributed between $[0, 1]$, and then normalized to a valid probability distribution.

Considering in more detail the specifics of the simulations, in subplot C) we showed two tables highlighting the shortcomings of Pearson's X^2 test. In both examples, power is hampered by the large number of rows. If the number of rows with nonzero counts doubles, and the value of the test statistic changes minimally, then the p-value X^2 provides will decrease dramatically (square root of the original p-value, $10^{-10} \to 10^{-5}$. Additionally, X^2 normalizes by the square root of the row sums, downweighting abundant rows and upweighting rare rows (which often arise from sequencing errors).

This is a critical drawback, as seen in the strong signal setting: because the first two rows are abundant, X^2 downweights their deviations, and yields only a moderately significant p-value. In the weak signal setting, even though no one row constitutes a large deviation, X^2 upweights the deviation in each row, and sums, yielding a similarly significant p-value. In contrast, OASIS identifies the strong and reproducible deviation in the the first setting, yielding an extremely significant p-value bound by focusing on the difference in expression of the first two rows. In the weak signal setting, OASIS struggles to find a strong separation between the samples, and only yields a slightly significant p-value bound. A spectral analysis of X_{corr} for the two tables is shown in Figure S1c, highlighting the flat spectrum of the weak signal table and concentrated spectrum of the strong signal table.

S.6.B. Robustness against simulated noise. OASIS provides robustness against many undesirable alternatives, beyond just negative binomial overdispersion. We provide simulations showing OASIS's robustness against corruption of each individual column's probability distribution, and additionally show OASIS's improved robustness to deviation of one sample in our approximate power calculation in Section S.7. In all settings, `OASIS-opt` (OASIS-opt in legends) was run with 5 random splits, each with 25% training data. `OASIS-rand` (OASIS in legends) was run with 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.

In our simulations we need to generate distributions over targets. The simplest option is to simply take a uniform distribution over the targets. Modelling the more realistic setting where some targets are more likely than others, we can also generate a vector $\tilde{\boldsymbol{p}}$ where each entry is drawn independently from some distribution, and renormalize this to a probability distribution, $\boldsymbol{p} = \tilde{\boldsymbol{p}}/\|\tilde{\boldsymbol{p}}\|_1$. In our simulations we utilize two base distributions. One obvious parameter free choice is a uniform distribution. To model the fact that some rows are much more abundant than others we also simulate with an exponential distribution. We select a rate parameter of 5 for this, arbitrarily.

S.6.B.1. Negative binomial overdispersion. As discussed, OASIS is robust to negative binomial overdispersion, as modelled in (16). 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 (16), sequencing data is overdispersed in practice, and counts should be modelled as negative binomials. For an overdispersion parameter θ (where $\theta = 0$ corresponds to the true null), we thus generate synthetic data as $X_{ij} \sim \text{Pois}(\Gamma(\lambda p_i/\theta, \theta))$. Following the trends depicted in (16), we model our overdispersion parameter as a function of the expected number of counts to be observed n, approximated as $\theta(n) = 3/n$ if $x < 150$, and $\theta(n) = 0.02$ otherwise. In these plots, the number of observations per column was varied from 10 to 1000 in log-scale through 10 equally spaced points. For the tables we show one index which shows the full dynamic range, index 6/10 with a value of $\exp(\ln(10) + \frac{5}{9}(\ln(1000) - \ln(10))) \approx 129$.

In both settings, with uniform or exponential target distributions (Figures S3 and S4), OASIS provides significantly better control of the discovery rate against this uninteresting alternative stemming from overdispersion.

S.6.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,

$$\boldsymbol{p}^{(j)} = (1 - \epsilon)\frac{1}{r}\mathbf{1} + \epsilon \boldsymbol{q}^{(j)}, \qquad [43]$$

where the synthetic data is generated by taking a Poisson number of observations from each column, multinomial with that column's probability vector $\boldsymbol{p}^{(j)}$, independently for each column. To avoid additional parameters, a uniform target distribution was used as the true null. $\boldsymbol{q}^{(j)}$ were generated independently by taking each entry to be independent and uniformly distributed between $[0, 1]$ and normalizing. This is illustrated in Figure S5.

In simulations, we observe that OASIS is extremely robust to these types of perturbations; whereas X^2 immediately calls that these tables deviate from the null, OASIS does not prioritize these unstructured deviations and maintains control of the discovery rate for much larger corruption magnitudes, especially for larger tables.

S.6.C. Power against simulated alternatives. OASIS has power against alternatives with unique per-sample expression (motivated by V(D)J recombination) like those in Figure S7, and can be shown to have more power against splicing-type alternatives (Figure S5) than the X^2 test in some settings.

S.6.C.1. Two-group setting. Here we show OASIS's power against alternative splicing type alternatives across a broader range of parameters. Alternative tables were generated by taking two groups of samples of equal size, and defining the two group's probability distributions as, for $k = 1, 2$

$$p_i^{(k)} = \begin{cases} 1 - \epsilon + \epsilon/I & \text{if } i = k, \\ \epsilon/I & \text{otherwise.} \end{cases} \qquad [44]$$

This indicates that $\boldsymbol{p}^{(1)}$ has most of its counts in the first row, and $\boldsymbol{p}^{(2)}$ in the second. The number of observations per sample is drawn independently from a Poisson distribution with mean 20 for each sample. Figure S6 shows that OASIS can have more power than X^2 in a variety of parameter regimes.

S.6.C.2. Unique per-sample expression. In this setting, the probability distribution of column $j \in [J]$ is $\boldsymbol{p}^{(j)} \in [0,1]^I$, with $I = J$:

$$p_i^{(j)} = \begin{cases} (1-\epsilon) + \epsilon/r & \text{if } i = j, \\ \epsilon/r & \text{otherwise.} \end{cases} \quad [45]$$

This setting is illustrated in Figure S7, where a corruption of $\epsilon = 0$ 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 S7, where in most settings Pearson's X^2 has more power than OASIS, but not by a significant margin.

S.7. Power analysis

Providing exact power calculations of OASIS and X^2 under different alternatives is difficult, as the random multinomial draws affect both X and E. Even asymptotic analyses would require more advanced tools from random matrix theory. To provide intuition for the power profile of OASIS relative to X^2, we instead perform approximate power calculations in a toy setting. Concretely, we assume that we observe a scaled version of the underlying alternative matrix, and compute OASIS and X^2 p-values for this observed X. We show that OASIS indeed has substantial power in many common 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 all classes of alternatives, e.g. 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 our power computations, 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 bound is continuous 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 bound.

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,

$$P_{X^2} \approx \exp\left(-\frac{X^2}{4IJ}\right). \quad [46]$$

S.7.A. Two-group alternative. Here we assume for simplicity that all n_j are equal, and that the clusters are of equal sizes. 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 $\boldsymbol{p}^{(1)}$, and for the second half of columns $\boldsymbol{p}^{(2)}$. In this setting we have:

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

and thus

$$E = n\left(\frac{\boldsymbol{p}^{(1)} + \boldsymbol{p}^{(2)}}{2}\right)\mathbf{1}_J^\top$$

To simplify our results, we utilize the symmetric chi-squared distance between two probability distributions $\chi^2(\boldsymbol{p}, \boldsymbol{q}) = 2\sum_i \frac{(p_i - q_i)^2}{p_i + q_i}$. The Pearson X^2 statistic is then

$$\begin{aligned} X^2 &= \sum_{i,j} \frac{(X_{i,j} - E_{i,j})^2}{E_{i,j}} \\ &= n \sum_{i,j<J/2} \frac{\left(p_i^{(1)} - \frac{p_i^{(1)}+p_i^{(2)}}{2}\right)^2}{\left(p_i^{(1)} + p_i^{(2)}\right)/2} + n \sum_{i,j\geq J/2} \frac{\left(p_i^{(2)} - \frac{p_i^{(1)}+p_i^{(2)}}{2}\right)^2}{\left(p_i^{(1)} + p_i^{(2)}\right)/2} \\ &= \frac{nJ}{2} \sum_i \frac{\left(p_i^{(1)} - p_i^{(2)}\right)^2}{p_i^{(1)} + p_i^{(2)}} \\ &= \frac{M}{4}\chi^2\left(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}\right) \end{aligned} \quad [47]$$

For disjoint $\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}$, the χ^2 distance is equal to 1. From the $(I-1) \times (J-1)$ degrees of freedom, this yield an approximate p-value of

$$P_{X^2} \approx \exp\left(-\frac{M^2\left(\chi^2\left(\boldsymbol{p}^{(1)},\boldsymbol{p}^{(2)}\right)\right)^2}{16I\times J}\right)$$

For OASIS, a pair of optimal $\boldsymbol{c}$ and $\boldsymbol{f}$ can be constructed where $\boldsymbol{c}$ is $+1$ for the first $J/2$ components, and -1 for the latter half, and $\boldsymbol{f}$ is sign $\left(\boldsymbol{p}^{(1)}-\boldsymbol{p}^{(2)}\right)$. Then, the OASIS test statistic is

$$\begin{aligned}
S &= \sum_{j=1}^{J/2}\sqrt{n_j}c_j(\boldsymbol{f}^\top\boldsymbol{p}^{(1)}-\bar{\mu}) + \sum_{j=J/2+1}^{J} c_j\sqrt{n_j}(\boldsymbol{f}^\top\boldsymbol{p}^{(2)}-\bar{\mu}) \\
&\overset{(a)}{=} \sum_{j=1}^{J/2}\sqrt{n}\boldsymbol{f}^\top\left(\boldsymbol{p}^{(1)}-\boldsymbol{p}^{(2)}\right) \\
&= \frac{\sqrt{MJ}}{2}\boldsymbol{f}^\top\left(\boldsymbol{p}^{(1)}-\boldsymbol{p}^{(2)}\right) \\
&\overset{(b)}{=} \frac{\sqrt{MJ}}{2}\delta_{\mathrm{TV}}(\boldsymbol{p}^{(1)},\boldsymbol{p}^{(2)})
\end{aligned} \tag{48}$$

where in (a) we used the fact that all $n_j = n$ were the same and that the clusters were equally balanced, and so the opposing signs of c_j cancelled the effect of $\bar{\mu}$. In (b) we used the optimal $\boldsymbol{f} = \mathrm{sign}(\boldsymbol{p}^{(1)} - \boldsymbol{p}^{(2)})$.

S.7.A.1. Summary. The X^2 p-value can be approximated as

$$P_{X^2} \approx \exp\left(-\frac{M^2}{I\times J}\left(\chi^2\left(\boldsymbol{p}^{(1)},\boldsymbol{p}^{(2)}\right)\right)^2\right). \tag{49}$$

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 bound of

$$P_{\mathrm{OASIS}} \le 2\exp\left(-\frac{M}{4}\delta_{\mathrm{TV}}^2\left(\boldsymbol{p}^{(1)},\boldsymbol{p}^{(2)}\right)\right). \tag{50}$$

Assuming the total variation distance and χ^2 distance between the distributions are constant, OASIS has more power than X^2 when $I \times J \ge M$; this is often the case we are in, where there are many rows with very low counts. Assumptions on the balanced classes can be alleviated 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 $c_j < 0$, requiring a different theoretical analysis.

S.7.B. Time series. Consider a simple time series setting, where at time $t = 0$ cells have target distribution $\boldsymbol{p}^{(1)}$, at time $t = T$ distribution $\boldsymbol{p}^{(2)}$, and their distributions vary smoothly in between, i.e. at time t they have the mixture distribution $\boldsymbol{p}^{(1)} + t(\boldsymbol{p}^{(2)} - \boldsymbol{p}^{(1)}) = (1-t)\boldsymbol{p}^{(1)} + t\boldsymbol{p}^{(2)}$. This can model microbial evolution in bulk RNA-seq data, cell-cycle in single cell RNA-seq, and many other settings. For $T+1$ evenly spaced time intervals (T is assumed to be odd), 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 expected matrix E are

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

Pearson's X^2 test statistic is:

$$\begin{aligned}
X^2 &= M\left\|\mathrm{diag}(1/\sqrt{X\mathbf{1}})(X-E)\,\mathrm{diag}(1/\sqrt{X^\top\mathbf{1}})\right\|_F^2 \\
&= nM\left\|\mathrm{diag}(1/\sqrt{X\mathbf{1}})\left(\boldsymbol{p}^{(1)}-\boldsymbol{p}^{(2)}\right)\left(\frac{1}{2}\mathbf{1}-\boldsymbol{t}\right)^\top\right\|_F^2 \\
&= nM\sum_{i=1}^{I}\frac{\left(p_i^{(1)}-p_i^{(2)}\right)^2}{\frac{M}{2}\left(p_i^{(1)}+p_i^{(2)}\right)}\left\|\frac{1}{2}\mathbf{1}-\boldsymbol{t}\right\|_2^2 \\
&= \frac{M}{12}\chi^2\left(\boldsymbol{p}^{(1)},\boldsymbol{p}^{(2)}\right) + o(T),
\end{aligned} \tag{51}$$

simplifying the Frobenius norm of a rank one matrix. $\left\|\frac{1}{2}\mathbf{1}-\boldsymbol{t}\right\|_2^2 = \frac{(T+1)(T+2)}{12T} = \frac{T}{12} + o(T)$. This leads to an approximate p-value of

$$P_{X^2} \approx \exp\left(-\frac{M^2}{48I \times T}\left(\chi^2\left(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}\right)\right)^2\right)$$

We claim that a pair of optimal $(\boldsymbol{c}, \boldsymbol{f})$ can be constructed as

$$\begin{aligned}\boldsymbol{c} &= \frac{1}{2} - \boldsymbol{t},\\ \boldsymbol{f} &= \operatorname{sign}(\boldsymbol{p}^{(1)} - \boldsymbol{p}^{(2)}.\end{aligned}$$

This yields

$$\begin{aligned}S &= \sqrt{n}\boldsymbol{f}^\top\left(\boldsymbol{p}^{(2)} - \boldsymbol{p}^{(1)}\right)\left(\boldsymbol{t} - \frac{1}{2}\mathbf{1}\right)^\top \boldsymbol{c}\\ &= 2\sqrt{n}\delta_{\mathrm{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})\left\|\boldsymbol{t} - \frac{1}{2}\mathbf{1}\right\|^2\\ &= 2\sqrt{n}\delta_{\mathrm{TV}}(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})\frac{(T+1)(T+2)}{12T}\end{aligned}$$

This gives OASIS a p-value bound of

$$\begin{aligned}P_{\mathrm{OASIS}} &\le 2\exp\left(-\frac{2S^2}{\|\boldsymbol{c}\|^2\|\boldsymbol{f}\|_\infty^2}\right)\\ &\le 2\exp\left(-n(T+1)\delta_{\mathrm{TV}}^2(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})/6\right)\\ &= 2\exp\left(-M\delta_{\mathrm{TV}}^2(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})/6\right)\end{aligned} \quad [52]$$

S.7.B.1. Summary. X^2 provides an approximate asymptotic p-value of

$$\begin{aligned}P_{X^2} &\approx \exp\left(-\frac{M^2\left(\chi^2\left(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}\right)\right)^2}{I \times T}\right)\\ &\approx \exp\left(-M\left(\chi^2\left(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}\right)\right)^2 \times \frac{n}{I}\right).\end{aligned}$$

OASIS yields a p-value bound of

$$P_{\mathrm{OASIS}} \le \exp\left(-M\delta_{\mathrm{TV}}^2(\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)})/6\right) \quad [53]$$

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

S.7.C. 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 assumed to be the same and c is even. Formally, the alternative matrix A is as below:

$$\begin{aligned}A &= \left(1 - \frac{c}{c-1}\alpha\right)I_c + \frac{\alpha}{c-1}\mathbf{1}\mathbf{1}^\top\\ X &= A\operatorname{diag}(\boldsymbol{n})\\ E &= \frac{1}{c}\mathbf{1}\mathbf{1}^\top\operatorname{diag}(\boldsymbol{n})\end{aligned}$$

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

$$\begin{aligned}X^2 &= \sum_{ij}\frac{(X_{ij} - E_{ij})^2}{E_{ij}}\\ &= \sum_j \frac{n_j^2((1-\alpha) - \frac{1}{c})^2}{n_j/c} + \sum_{i\neq j}\frac{n_j^2\left(\frac{\alpha}{c-1} - \frac{1}{c}\right)^2}{n_j/c}\\ &= c\left(1 - \alpha - \frac{1}{c}\right)^2\sum_j n_j + c\left(\frac{\alpha}{c-1} - \frac{1}{c}\right)^2\sum_{i\neq j} n_j\\ &= cM(1-\alpha-1/c)^2 + \underbrace{c(c-1)M\left(\frac{\alpha}{c-1} - \frac{1}{c}\right)^2}_{\text{negligible}}.\end{aligned} \quad [54]$$

Thus, for fixed c, the X^2 p-value can be approximated as

$$P_{X^2} \approx \exp(-M^2(1-\alpha)^4)$$

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

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

constitute an optimal pair. Since $n_j = M/c$ for all j, $\bar{\mu} = 1/2$, and

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

This yields

$$\begin{aligned} S &= \sum_j c_j \sqrt{n_j}\,(\hat{\mu}_j - \bar{\mu}) \\ &= \sum_j \sqrt{M/c}\frac{1}{2}\left(1 - \frac{c}{c-1}\alpha\right) \\ &= \sqrt{cM}\frac{1}{2}\left(1 - \frac{c}{c-1}\alpha\right) \end{aligned} \quad [55]$$

and a corresponding p-value bound of

$$P_{\text{OASIS}} \le \exp\left(-\frac{S^2}{\|\boldsymbol{c}\|_2^2}\right) = \exp\left(-\frac{M}{4}\left(1 - \frac{c}{c-1}\alpha\right)^2\right) \approx \exp\left(-M\,(1-\alpha)^2\right)$$

S.7.C.1. Summary. In this toy-setting of unique per-sample expression, X^2 will yield a p-value of

$$P_{X^2} \approx \exp\left(-M^2(1-\alpha)^4\right). \quad [56]$$

Comparatively, assuming OASIS identifies an optimal pair $(\boldsymbol{c}, \boldsymbol{f})$, its p-value bound can be approximated as

$$P_{\text{OASIS}} \le \exp\left(-\frac{M}{4}\left(1 - \frac{c}{c-1}\alpha\right)^2\right) \approx \exp\left(-M(1-\alpha)^2\right). \quad [57]$$

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

S.7.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 $X_{00}, X_{10}, X_{0,1:}, X_{1,1:}$

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

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

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

Comparatively, OASIS will choose $\boldsymbol{f} = [1\ 0]^\top$, which yields $S_j = \sqrt{n}/J$ for $j = 2, \dots, n$ and $S_1 = -\sqrt{n}(1-1/J)$. This yields an optimizing $\boldsymbol{c} = [-(1-1/J)\ 1/J\ \dots\ 1/J]$ with $\|\boldsymbol{c}\|^2 = 1 - 1/J$. The test statistic is then $S = \sqrt{n}\|\boldsymbol{c}\|^2$ Thus,

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

S.7.D.1. Summary. X^2 provides a p-value of

$$P_{X^2} \approx \exp\left(-\frac{(nJ)^2}{2(2J)^2}\right) = \exp\left(-n^2/8\right), \quad [58]$$

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

$$P_{\text{OASIS}} \leq \exp(-S^2/\|\boldsymbol{c}\|^2) = \exp(-n). \quad [59]$$

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

S.7.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 the table is sparse, i.e. $\frac{M}{I \times J}$ is small. This highlights one of the difficulties with the X^2 test; its p-value depends heavily on the number of rows present, even if many of them are inconsequential.

S.8. Supplemental figures

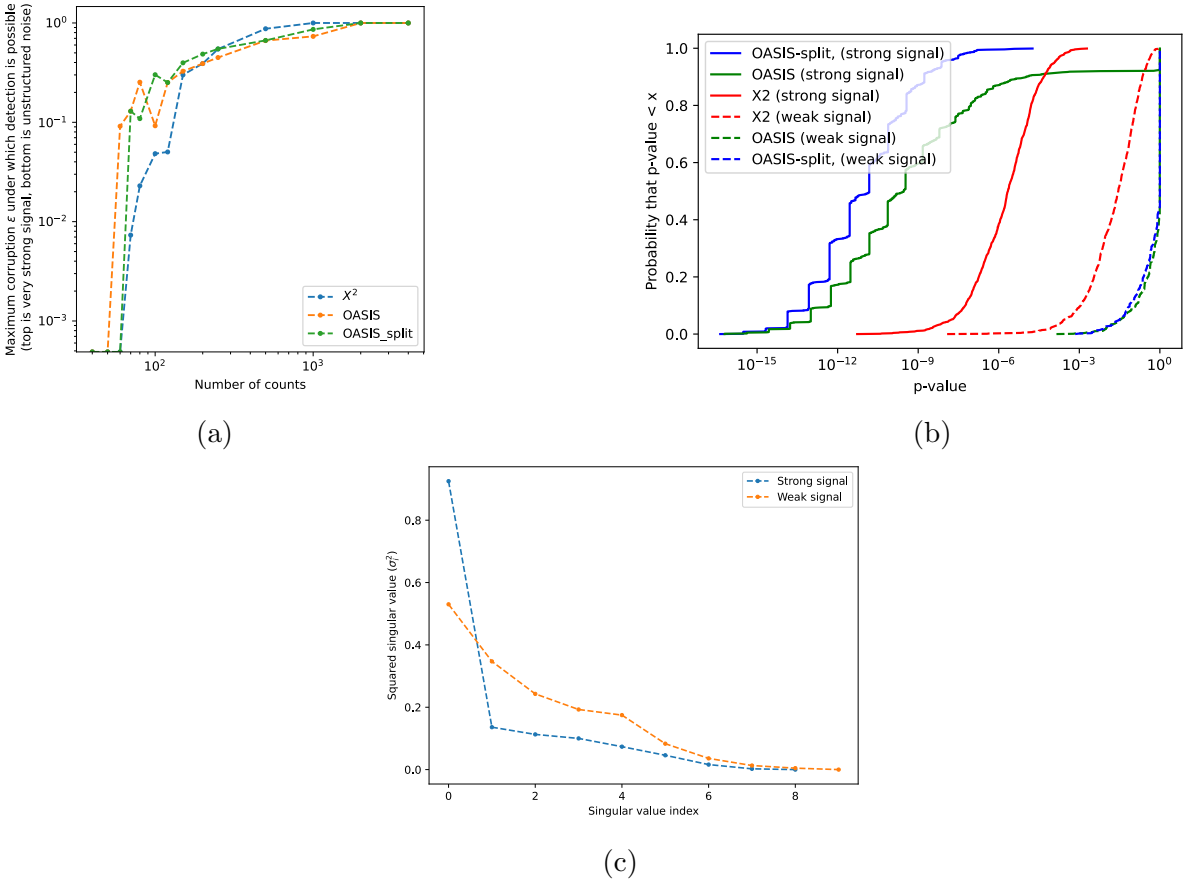


Fig. S1. Following Figure 1D, this plot shows the performance of `OASIS-opt` and X^2 over the class of 2-group alternatives defined in Eq. (42). (a) shows an empirical achievability curve, where the y-axis indicates how structured versus unstructured the alternative is, and the x-axis indicates the total number of counts in the table. The more up and to the left a curve is, the better. For any alternative (any $\epsilon \in [0, 1]$), a statistical test should reject the null with enough observations. For a given number of counts, we performed binary search to identify the largest ϵ for which over half of the p-values (bounds) were less than $1E-10$. 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` is run with 5 random splits as in Figure 1D, which we denote as OASIS-split, and we additionally show results (labelled OASIS) when only 1 random split is used. (b) plots an ECDF of p-values in the two settings for different methods. Performing 5 random splits removes the tail of the p-value bounds, generally improving performance. The flat behavior between adjacent data points for `OASIS-opt` is due to the data-splitting: an increase in the number of observations does not necessarily translate to an increase in the number of observations in the test data. (c) shows the spectrum of X_{corr} for the two example tables in main text Figure 1c. X^2 yields a similar p-value for the two tables, as the sum of the squared singular values is similar. OASIS prioritizes the "strong signal" table with the more concentrated spectrum.

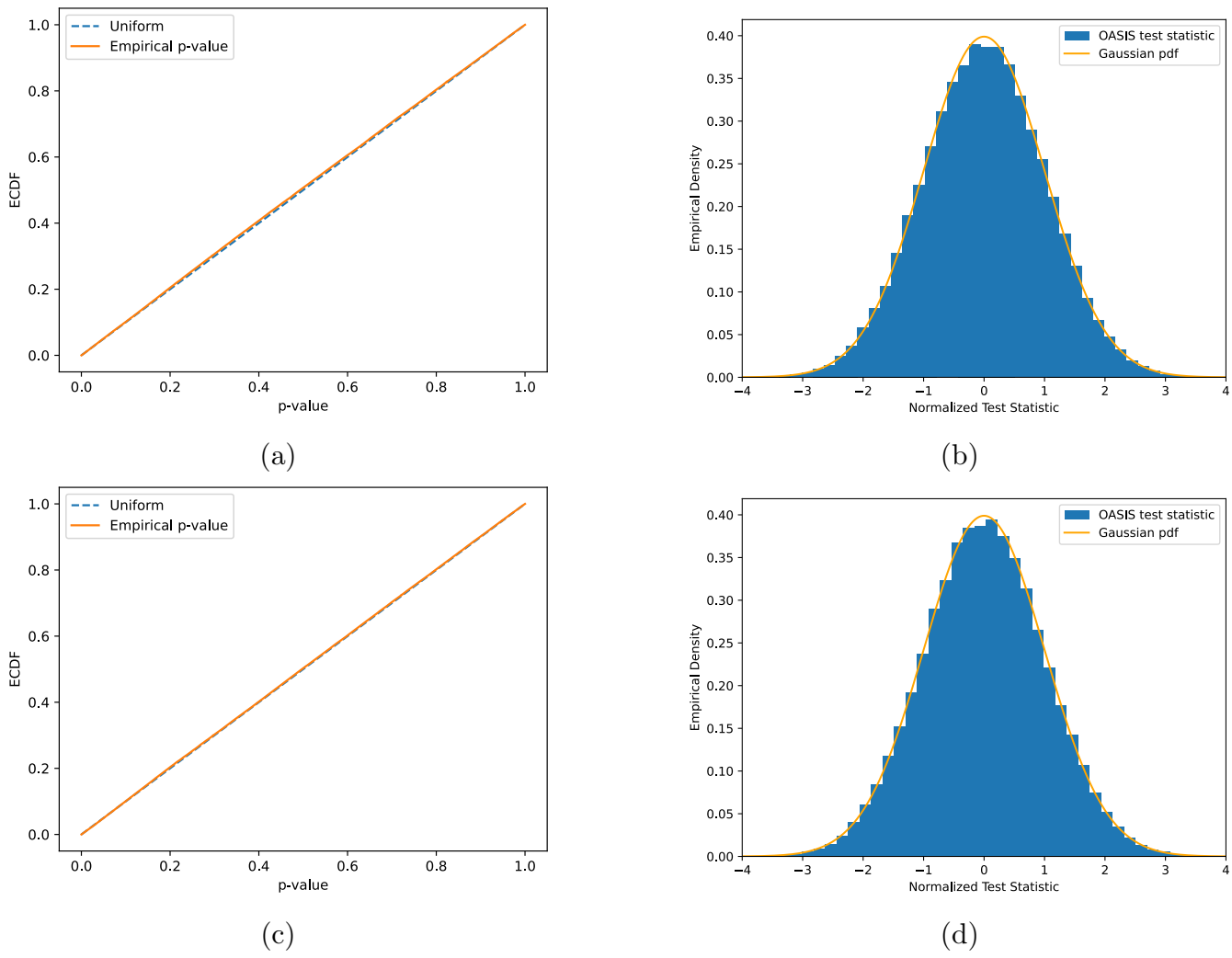


Fig. S2. Distribution of the normalized OASIS test statistic $\frac{S}{\hat{\sigma}_{\boldsymbol{f}}\|\boldsymbol{c}\|\sqrt{1-\gamma}}$ (right). As predicted by Proposition 3 this follows a Gaussian distribution, leading to uniformly distributed asymptotic p-values (left). The synthetic tables generated have few counts, but the approximation is still extremely good. Each plot shows 100k randomly generated tables with 5 rows, 8 columns, and a random target distribution (each entry i.i.d. uniform, normalized). Column counts for simulations in the first row are independently drawn from a Poisson with mean 10, while the second row has mean 30. For each trial a random $\boldsymbol{p}$ was generated, $\{n_j\}$ were drawn randomly from Poisson distributions, and X was generated from this null. Then, $\boldsymbol{c}$ and $\boldsymbol{f}$ were independently randomly generated, where the coordinates of $\boldsymbol{f}$ were i.i.d. Ber$(1/2)$, and the coordinates of $\boldsymbol{c}$ were i.i.d. $U([-1,1])$. This process was repeated until a table with $\hat{\sigma}_{\boldsymbol{f}} > 0$ and $\gamma < 1$ was generated to yield a valid run (as otherwise the test statistic is identically 0). KS distance (maximum distance between OASIS's asymptotic p-value's ECDF and the uniform distribution of .0079 for $n = 10$ counts per column, $.0038$ for $n = 30$ counts per column.

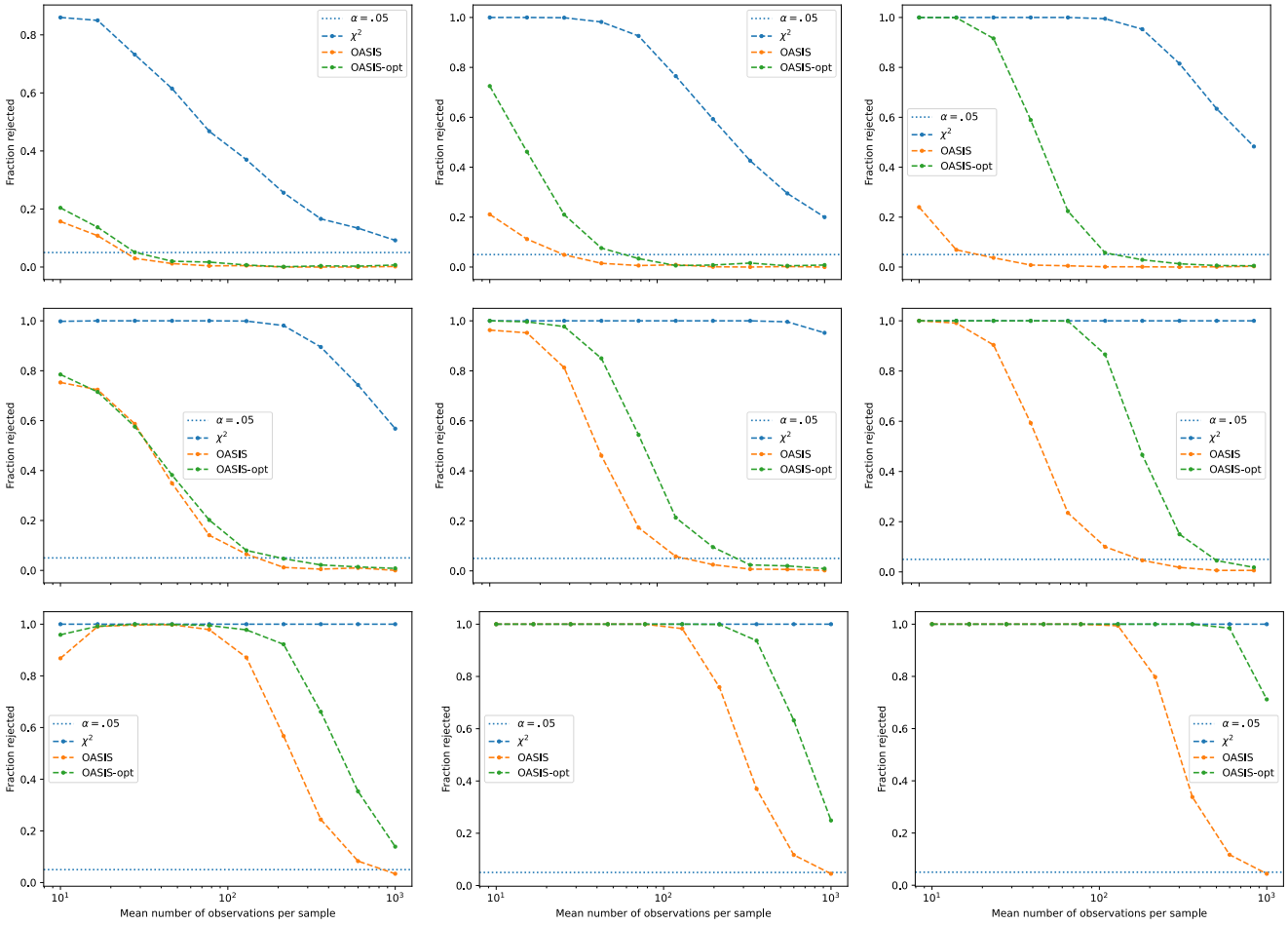


Fig. S3. Simulated data for negative binomial overdispersion. Null target distribution generated from exponential distributions, described in Section S.6.B.1. OASIS has significantly better control of the false discovery rate than the X^2 test, and requires substantially fewer samples per column to control the FDR. The 3x3 grid of plots displays results for tables varying in number of columns (10, 50, 400 from left to right) and number of rows (5, 20, 100 from top to bottom).

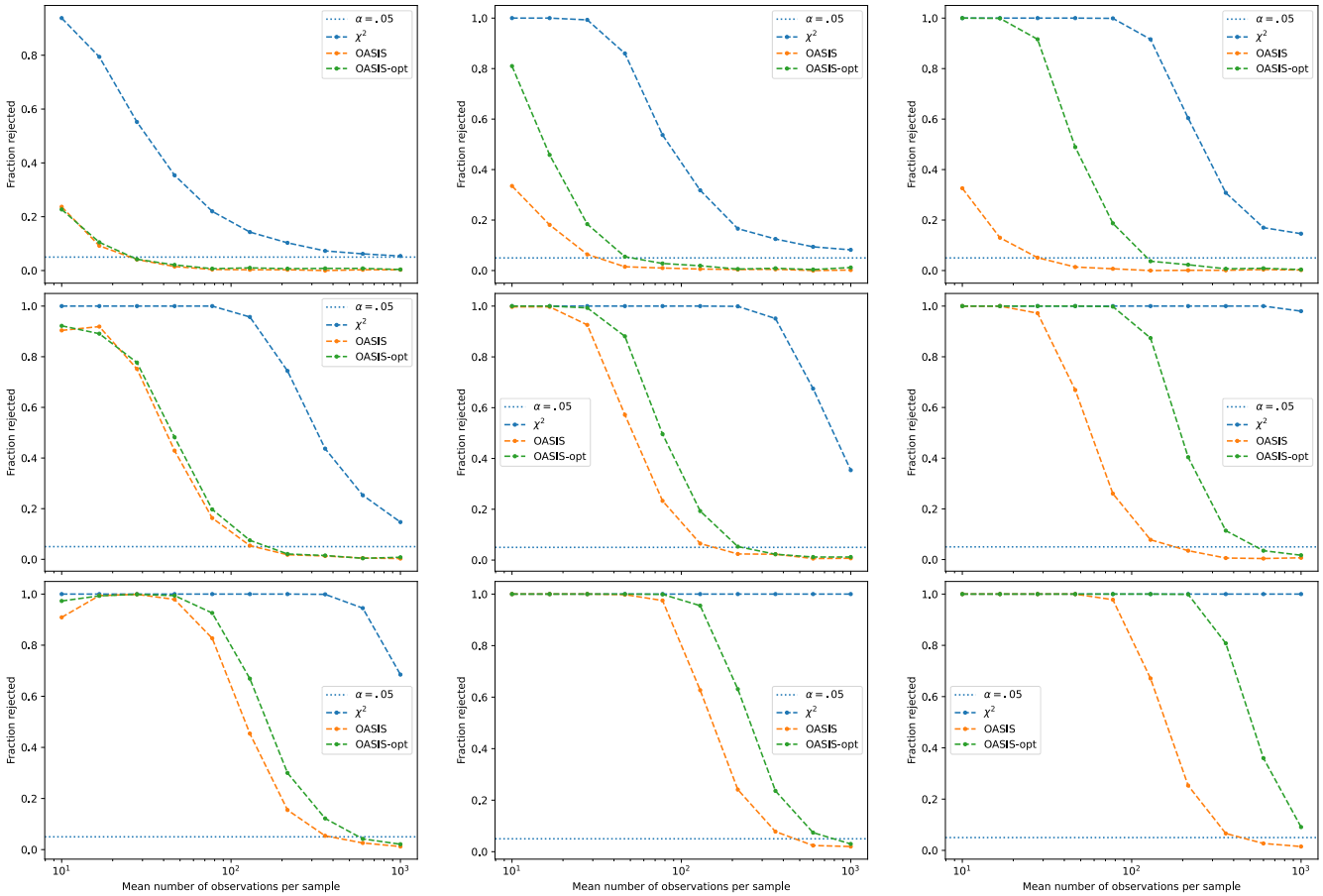


Fig. S4. Simulated data for negative binomial overdispersion. Null target distribution is uniform over rows. OASIS has significantly better control of the false positive rate than the X^2-test, and requires substantially fewer samples per column to control the FDR. The 3x3 grid of plots displays results for tables varying in number of columns (10, 50, 400 from left to right) and number of rows (5, 20, 100 from top to bottom).

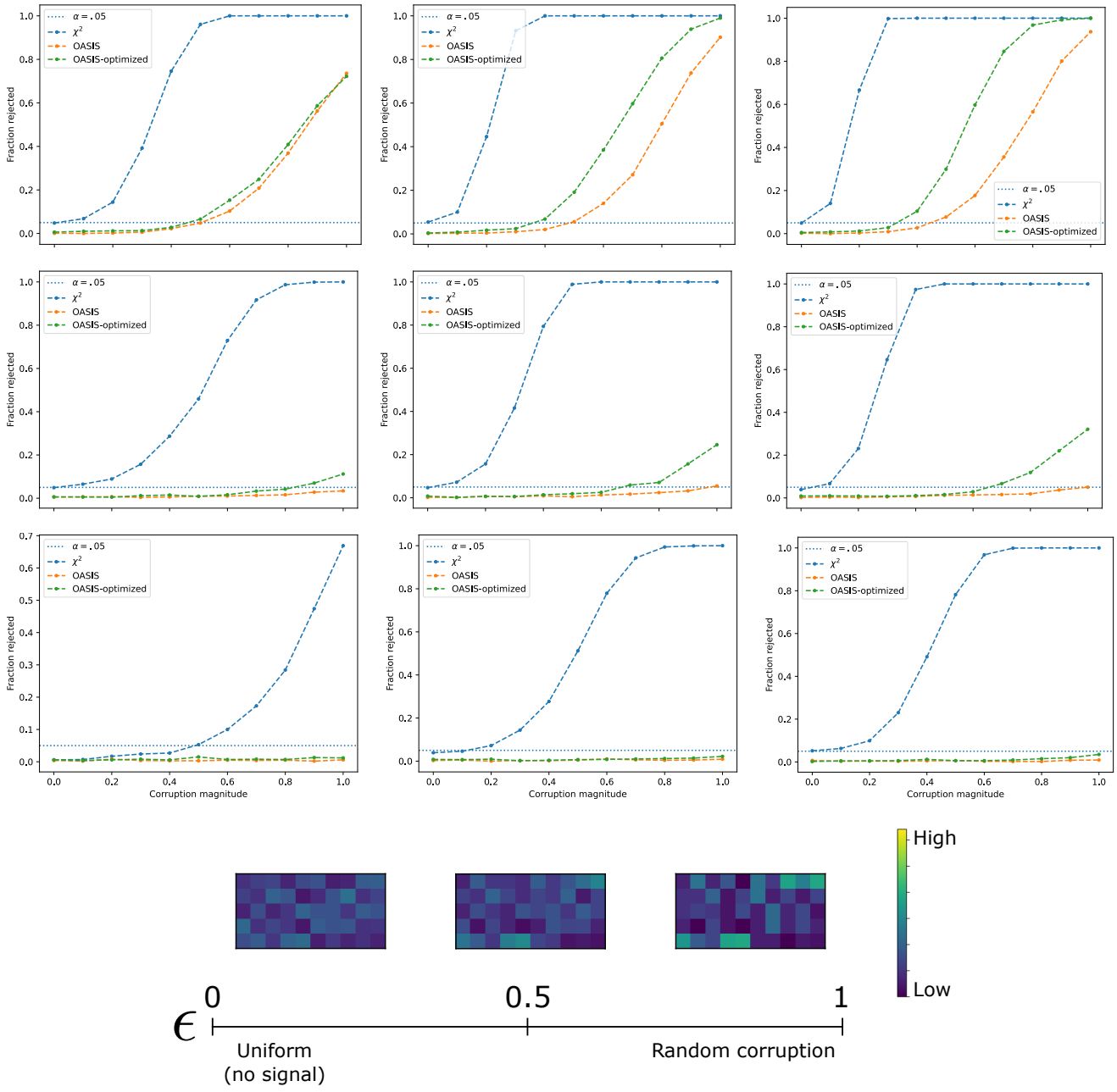


Fig. S5. Simulated data for independent ℓ_1 corruption of each column (Eq. (43)). At bottom, a diagram showing tables under different corruption magnitudes ϵ. Number of observations in each column is Poisson distributed with mean 100, independent across columns. Observations are drawn independently, and are multinomially distributed for each column. The 3x3 grid of plots displays results for tables varying in number of columns (10, 50, 100 from left to right) and number of rows (20, 10, 500 from top to bottom). X^2 has significant power against this alternative, even though it represents contamination and a "biological null".

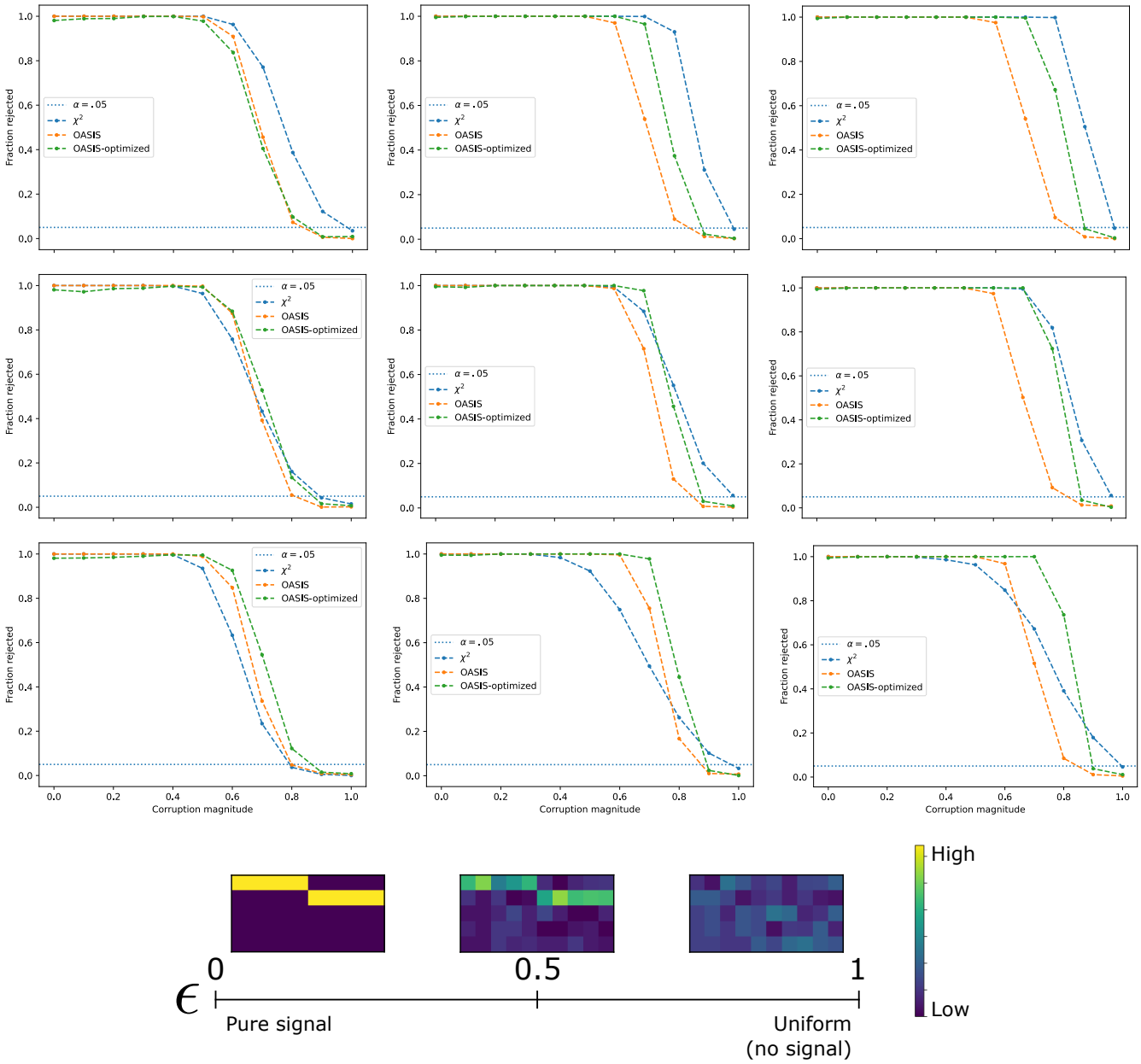


Fig. S6. Simulated data for alternative splicing-type alternative (Eq. (44)). Number of observations in each column is Poisson distributed with mean 20. The 3x3 grid of plots displays results for tables varying in number of columns (10, 50, 100 from left to right) and number of rows (20, 100, 400 from top to bottom). OASIS generally has more power than X^2.

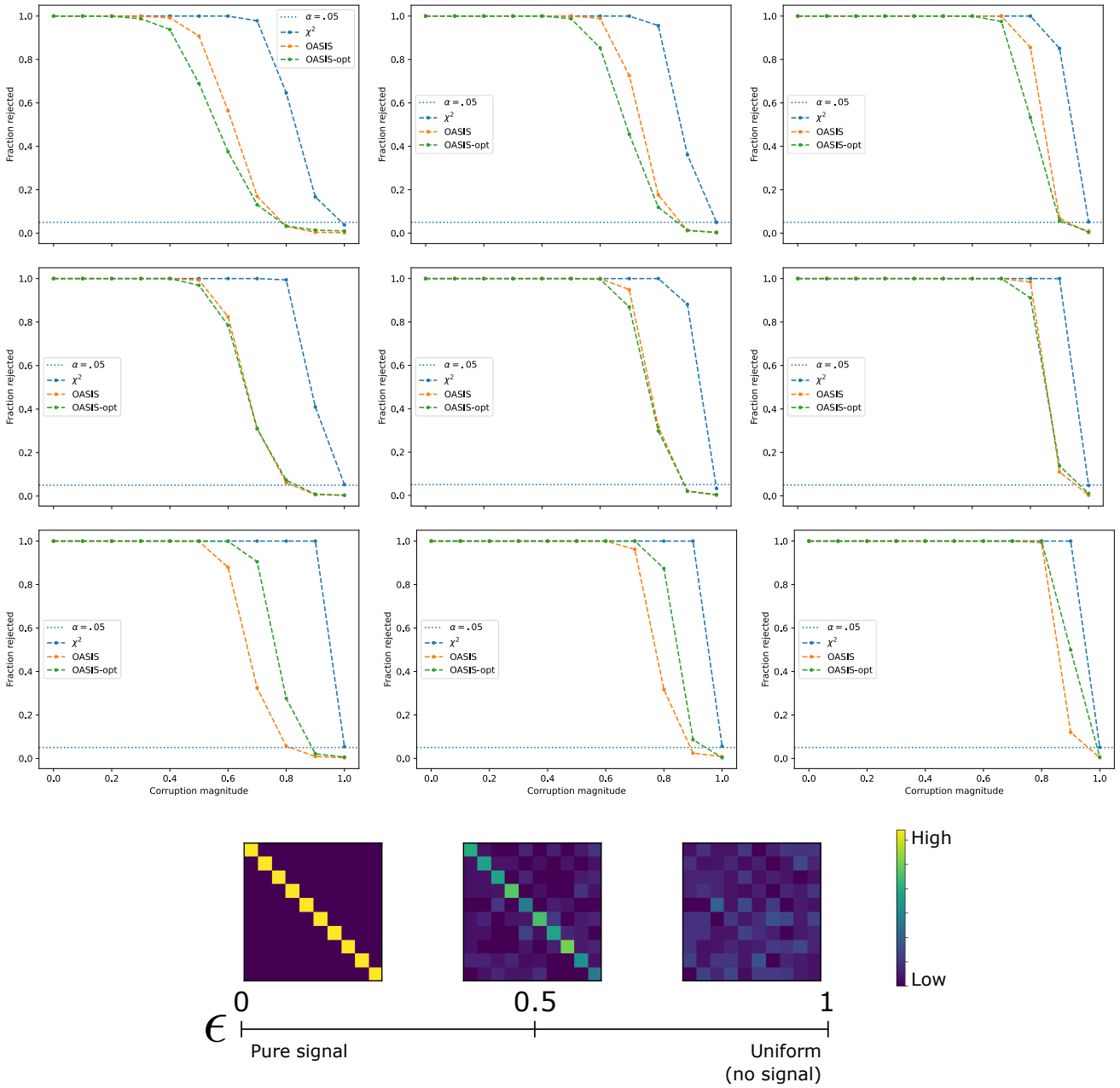
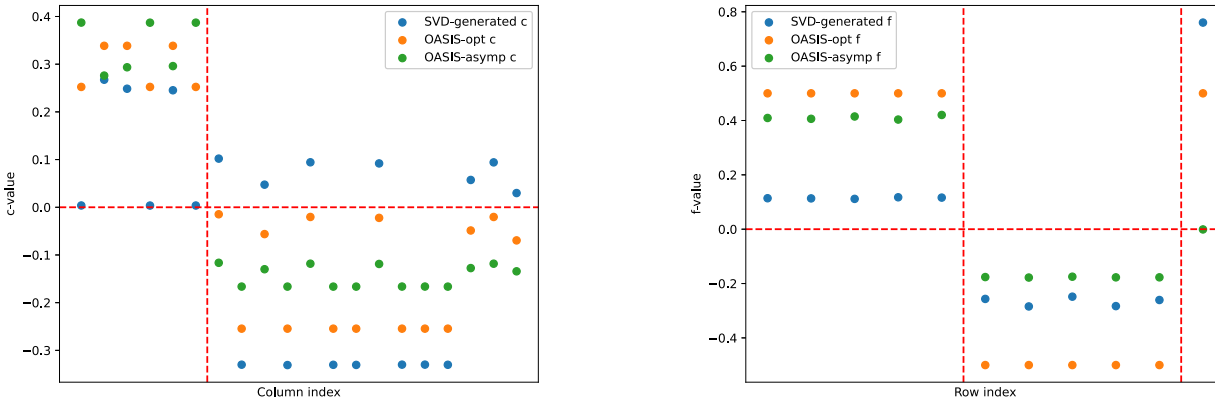
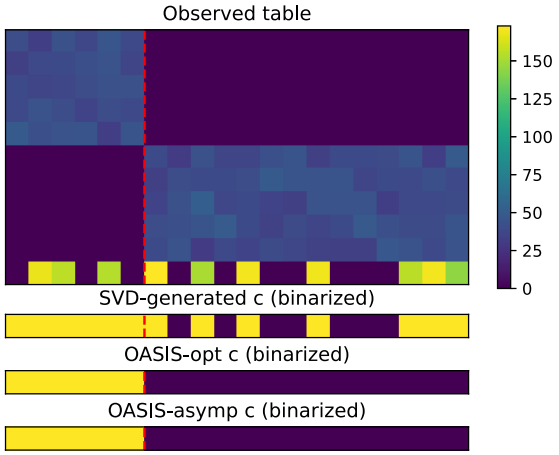


Fig. S7. Simulated data for unique per-sample expression alternative (Eq. (45)). Number of observations in each column is Poisson distributed with the given mean, independent across columns. Detailed probabilistic model in Eq. (45). The 3x3 grid of plots displays results for tables varying in number of rows (10, 25, 100 from top to bottom) and mean number of observations per column (10, 20, 50 from left to right). X^2 has more power, but all methods perform well.



(a) (b)



(c)

Fig. S8. SVD does not identify the visually clear planted structure in the toy example, while OASIS does. Shown in (c) is the contingency table X with two groups of columns, one which expresses rows 1-5 (the left 6 columns), and the other which expresses rows 6-10 (the right 14 columns). 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. The last row should be considered noise, as this signal does not match any other row's expression. Multinomial sampling provides slight noise to make this more realistic. Both methods call this table as significant: an extreme example chosen for clarity of exposition. The latent structure is identified by an embedding $\boldsymbol{f}$ which takes value 1 on the first 5 rows, -1 on the next 5, and 0 (i.e., disregarding), the final row. OASIS, both from its asymptotic-optimized $\boldsymbol{c}$ (OASIS-asymp) and its finite-sample bound optimized $\boldsymbol{c}$ (OASIS-opt), perfectly identify the latent clustering. 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 $\boldsymbol{f}$. (b) plots the embeddings identified by the two methods, subtracting .5 from OASIS-opt's $\boldsymbol{f}$, so that it centers around 0. The SVD-based $\boldsymbol{f}$ does identify the row groupings, but utilizes only half the dynamic range for the first 10 rows, assigning the last row a value of 0.76, where all other components have magnitude at most 0.28. OASIS more clearly identifies the row structure, with the asymptotic objective forcing the bottom row to a value less than 9×10^{-4} in magnitude where all other components have a magnitude of at least 0.17. OASIS-opt utilizes the full dynamic range for the first 10 rows. (a) plots $\boldsymbol{c}$, highlighting that the SVD-identified $\boldsymbol{c}$ cannot identify the latent structure. OASIS-opt identifies the clustering in a well separated manner, and the asymptotic optimized $\boldsymbol{c}$ yields even greater separation (as it fully discounts the last "noisy" row).

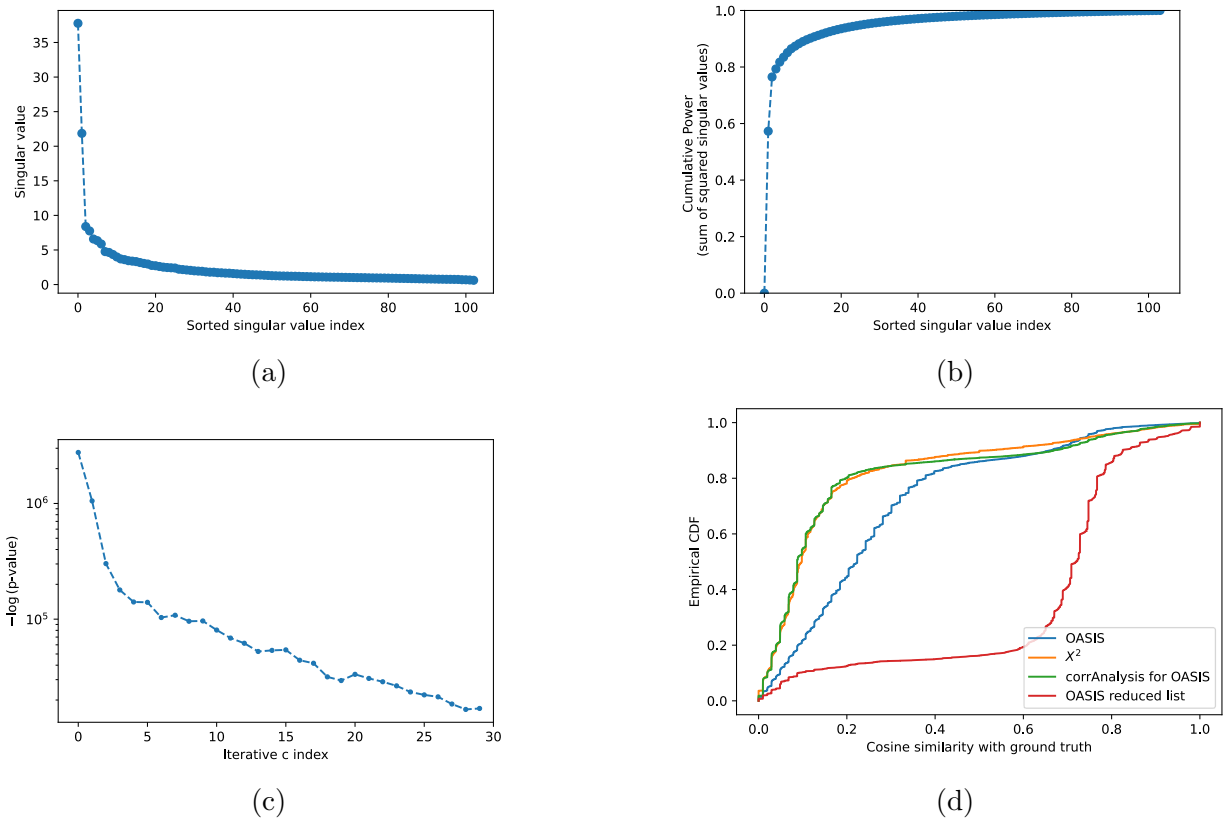


Fig. S9. Analysis of 100,914 tables generated by SPLASH (5) from SARS-CoV-2 data (2). (a,b) show the spectrum of the C matrix from embedding-aggregation for SARS-CoV-2 data, illustrating why a 2D embedding is sufficient. (c) plots the negative log of the p-value bounds for counts-aggregation for SARS-CoV-2. All are 0, up to machine precision, due to the large number of counts. Thus, while for smaller matrices significance testing provides a good cutoff, for larger tables one may still need to find an "elbow" in the curve. As can be seen, after the second index (the elbow) the remaining points follow a linear trend, indicating that only 2 components should be utilized. This threshold can be programatically identified by looking at the ratio of successive log p-value bounds, and finding the maximum. (d) Plots the empirical CCDF of absolute cosine similarities between the binarized sample embeddings per table (c for OASIS, principal right singular vector from correspondence analysis for X^2) and ground truth vector indicating whether a patient (sample) has Delta or not. Only tables that a method declares significant after multiple hypothesis correction are used. The tables that X^2 rejects do not yield signal that correlates well with the ground truth; for example, 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. This plot additionally shows 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 the tables OASIS rejects, similar results are obtained, but when analyzing all calls made by Pearson's X^2, many have worse correspondence with sample metadata.

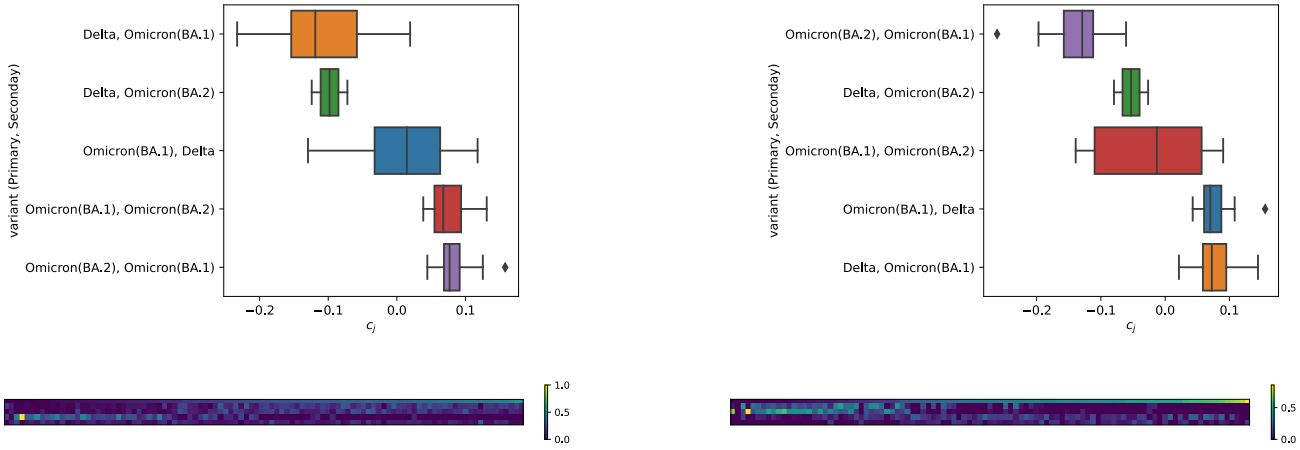


Fig. S10. Top 2 tables in terms of `OASIS-opt` p-value bound, not called by `OASIS-rand`. Both tables yield a vector $\boldsymbol{c}$ with a binarized absolute cosine similarity of 0.76 with ground truth metadata (was the patient infected with Delta or not).

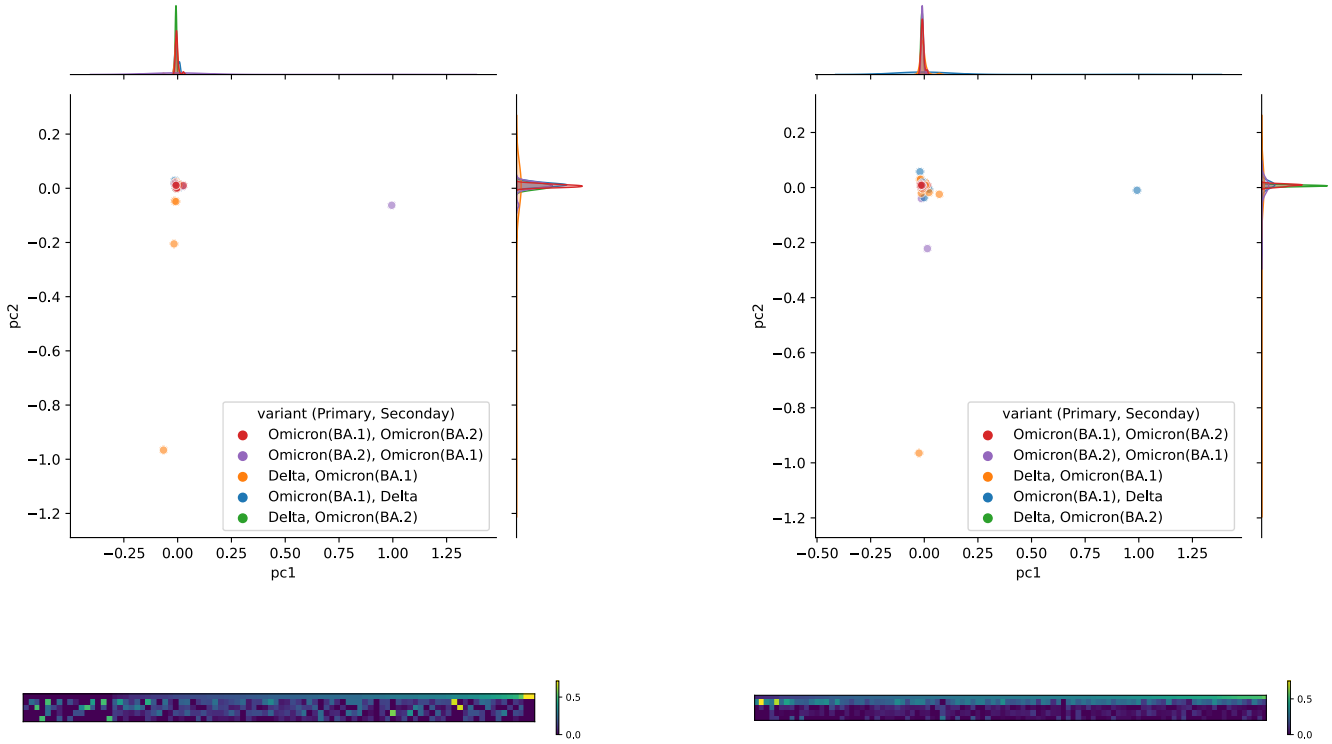


Fig. S11. 2 tables with X^2 q-values of 0, not called by `OASIS-opt` (q-value > 0.05). Since there were more than 2 tables with a X^2 q-value of 0, we chose the two with the smallest `OASIS-opt` p-value bounds. The output 1D embedding by correspondence analysis yields an absolute cosine similarity of 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.

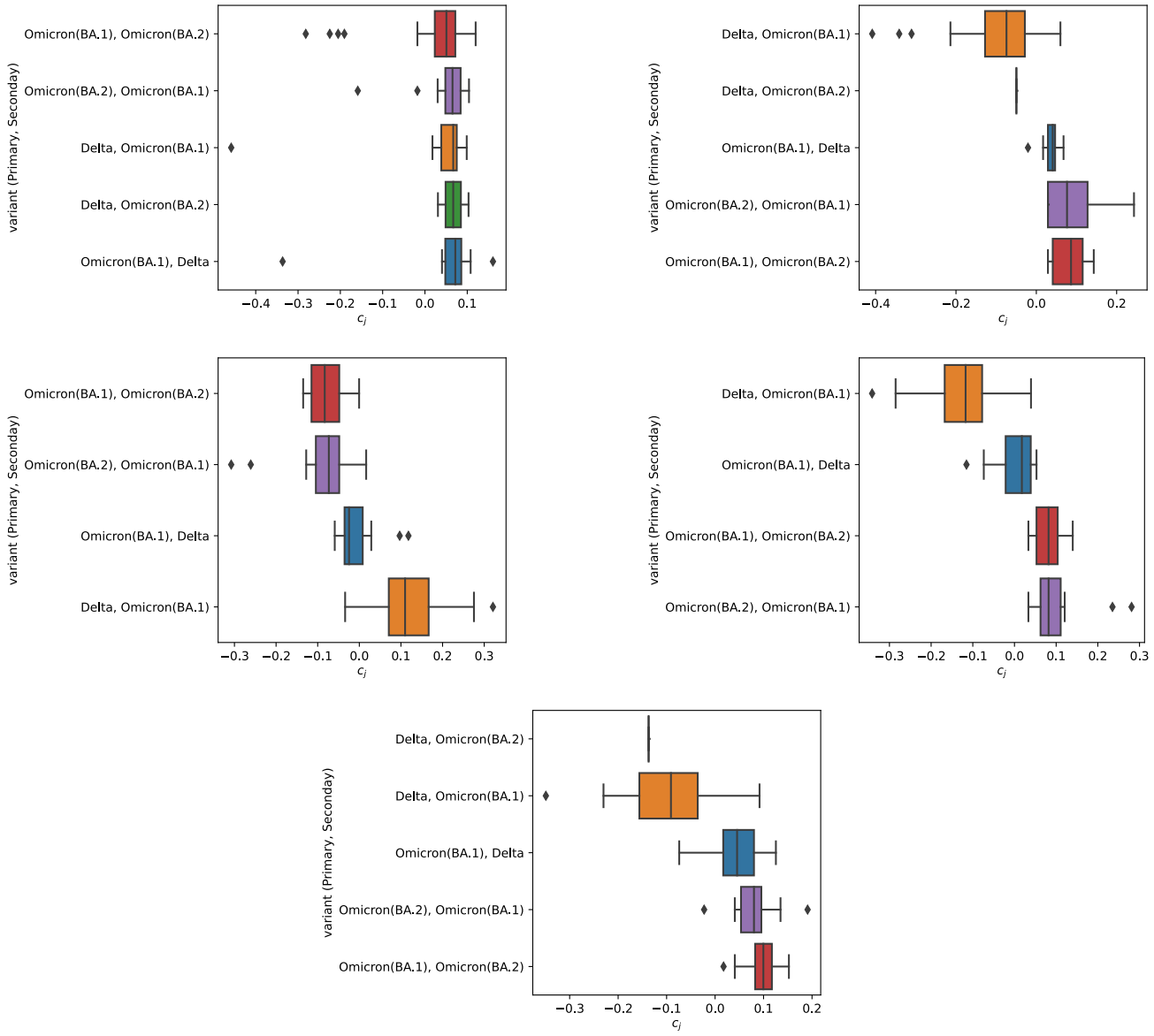


Fig. S12. All 5 tables in `OASIS-opt`'s reduced list of anchors (effect size in the top 10%, $M > 1000$, significant post BY correction `OASIS-opt` p-value bound), not called by X^2. First is 482×100, 3688 counts, cosine similarity of 0.2. Second is 98×77, with 2540 counts, cosine similarity of 0.61. 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.

Fig. S13. BLAT (17) to SARS-CoV-2 genome via UCSC genome browser (18). 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 (5)). 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, target 1 is predicted to be Delta, and target 2 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. This deletion is not annotated in the UCSC genome browser, but corresponds to a known Omicron deletion. 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.

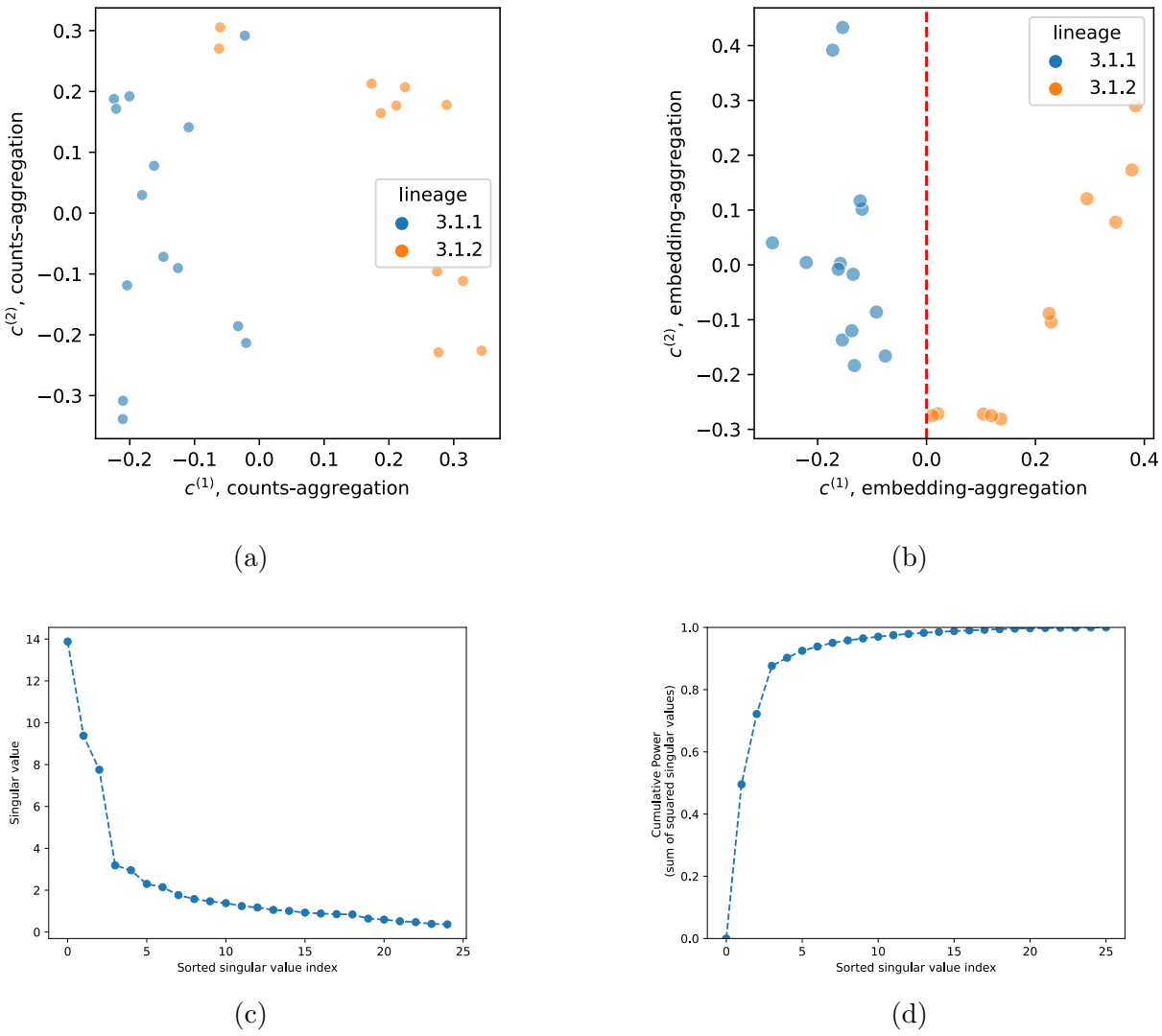


Fig. S14. Subplots (a,b) show 2D embeddings of *M. tuberculosis* data. The second dimension does not help with sub-sub-lineage classification. Plots (c,d) show the spectrum of the C matrix from embedding-aggregation.

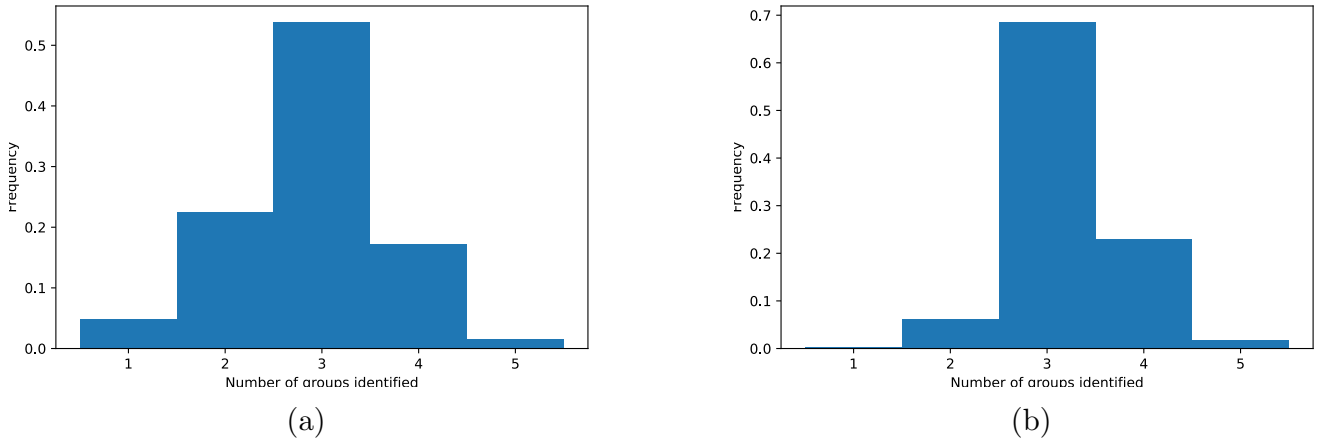


Fig. S15. OASIS, unaware of the underlying multi-group structure, is able to give a statistically valid stopping criterion for subclustering. We show evaluation in a planted setting with 3 groups, 10 rows, and 20 columns. Counts per column are Poisson distributed: (a) shows mean 20 observations per column, (b) shows mean 30. We plot the number of identified $\boldsymbol{c}$ vectors using `OASIS-iter`; finding orthogonal $\boldsymbol{c}^{(k)}$ until p-value is no longer significant. OASIS correctly identifies the number of clusters in 55% of simulations, increasing to 69% with mean 30 observations per column. OASIS is only derived as a test on contingency tables, but this simple analysis shows its utility in other applications such as subclustering.

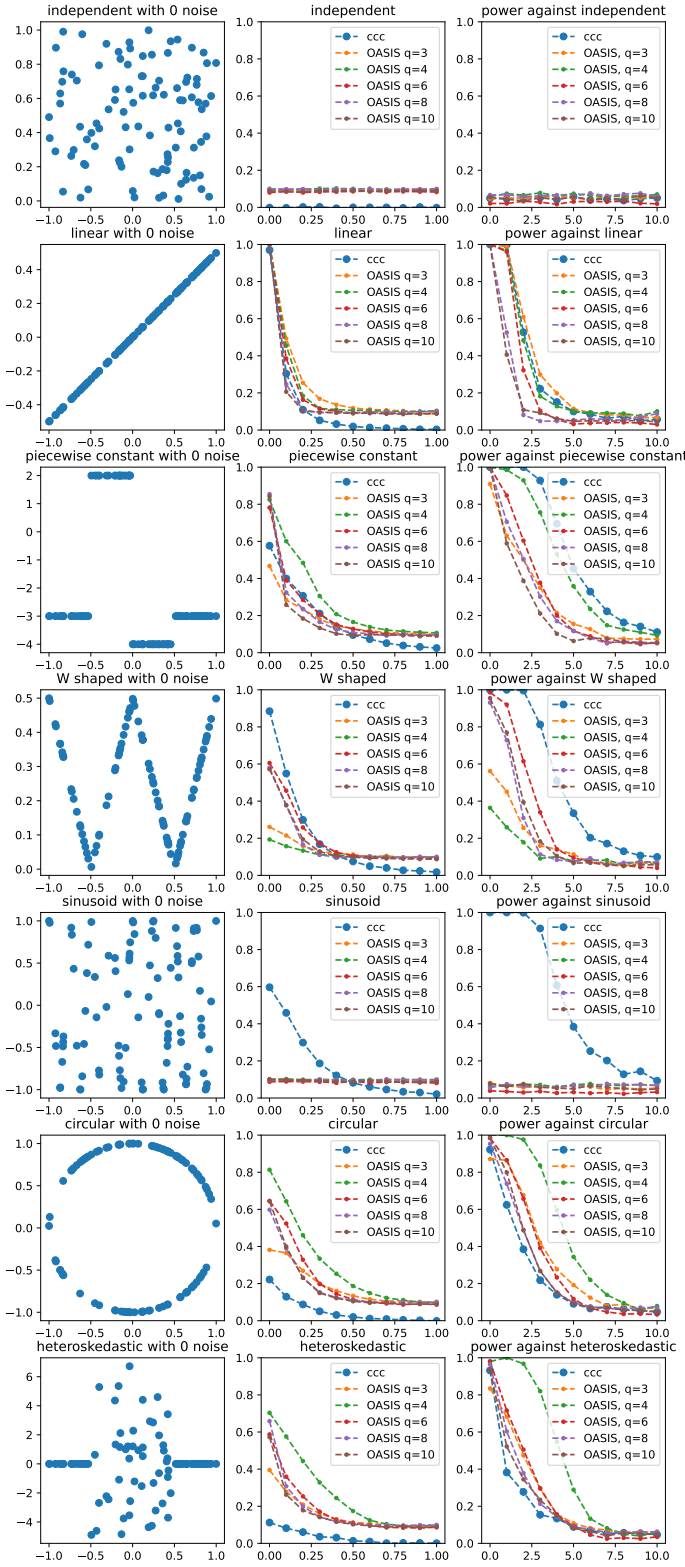


Fig. S16. Comparison of `OASIS-opt` and xicor (14). OASIS asymptotic p-value used, 25% train test split. Each row represents a different setting, where the left column shows the noiseless problem instance $y = f(x)$, the middle column shows the magnitude of the correlation coefficient averaged over 1000 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. The quantization level q, which maps the input continuous-valued random variable to a discrete categorical one, is varied across trials. This is performed independently for the row and column random variables, and binned into q bins of equal counts ($1/q$ quantiles).

S.9. Supplemental tables

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 S1. Power at $\lambda \approx 129$, noise model as in (16). Exponentially distributed target distribution, described in text. Full plots in Figure S3.

number of rows (I)	SDP+rounding	Alternating maximization	Asymptotically-optimal (eigenvector + rounding)
12	14.5	0.86	0.20
24	194	1.2	0.25
48	-	1.26	0.59
96	-	1.31	5.9
192	-	1.48	16.7

Table S2. Computational time (in seconds) for computing $\boldsymbol{c}, \boldsymbol{f}$ using the stated method, on 1000 matrices. As can be seen, the SDP solver quickly becomes infeasible. Forming the matrix $\tilde{X}\tilde{X}^\top$ also becomes costly as the number of rows increases. Alternating maximization stays efficient and performant. Timing results generated on a MacBook Air (M2, 2022), runs declared as a failure after 5 minutes.

References

1. 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).
2. 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).
3. R Leinonen, H Sugawara, M Shumway, INSD Collaboration, The sequence read archive. *Nucleic acids research* **39**, D19–D21 (2010).
4. M Kokot, R Dehghannasiri, TZ Baharav, J Salzman, S Deorowicz, Splash2 provides ultra-efficient, scalable, and unsupervised discovery on raw sequencing reads. *bioRxiv* pp. 2023–03 (2023).
5. K Chaung, et al., Splash: a statistical, reference-free genomic algorithm unifies biological discovery. *Cell* **186**, 5440–5456 (2023).
6. MJ Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. (Cambridge University Press) Vol. 48, (2019).
7. P Billingsley, *Probability and measure*. (John Wiley & Sons), (2008).
8. AC Berry, The accuracy of the gaussian approximation to the sum of independent variates. *Transactions american mathematical society* **49**, 122–136 (1941).
9. F Chen, S Roch, K Rohe, S Yu, Estimating graph dimension with cross-validated eigenvalues. *arXiv preprint arXiv:2108.03336* (2021).
10. A Neufeld, J Popp, LL Gao, A Battle, D Witten, Negative binomial count splitting for single-cell rna sequencing data. *arXiv preprint arXiv:2307.12985* (2023).
11. MX Goemans, DP Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM (JACM)* **42**, 1115–1145 (1995).
12. S Diamond, S Boyd, CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* **17**, 1–5 (2016).
13. 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).
14. S Chatterjee, A new coefficient of correlation. *J. Am. Stat. Assoc.* **116**, 2009–2022 (2021).
15. A Maurer, M Pontil, Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* (2009).
16. 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).
17. WJ Kent, Blat—the blast-like alignment tool. *Genome research* **12**, 656–664 (2002).
18. WJ Kent, et al., The human genome browser at ucsc. *Genome research* **12**, 996–1006 (2002).