SUPPLEMENTARY MATERIALS

A fast exact functional test for directional association and cancer biology applications

Hua Zhong and Mingzhou Song

SUPPLEMENTARY NOTE 1. PEARSON'S CHI-SQUARE TEST OF ASSOCIATION

Pearson's chi-square test [1] determines the independence or association between variables in a contingency table. The null hypothesis is that the row and column variables are statistically independent. Let O be an observed $r \times c$ contingency table with $O_{i,j}$ being the observed count in the cell at row i and column j. Let

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{i,j} \tag{1}$$

stand for the observed sample size. Let $E = [E_{i,j}]$ denote the expected contingency table by chance under the null hypothesis, where $E_{i,j}$ is the expected count in the cell at row i and column j defined by

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \cdot \sum_{m=1}^{r} O_{m,j}}{n} \tag{2}$$

In his seminal work [1], Karl Pearson defined the chi-square statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{3}$$

which asymptotically follows a chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom under the null hypothesis. The p-value of the observed table O can be calculated by the upper tail probability of the distribution.

The chi-square null distribution is asymptotically accurate when the sample size approaches infinity, but it is inexact at small sample sizes when Fisher's exact test is often used. Pearson's chi-square test is symmetrical and cannot detect directional relationships. The chi-square statistic is closely related to other measures of association including G-test and mutual information.

- *H. Zhong is with the Department of Computer Science, New Mexico State University, Las Cruces, NM, 88003, U.S.A. E-mail: huazhong@nmsu.edu.*
- *M. Song is with the Department of Computer Science, New Mexico State University, Las Cruces, NM, 88003, U.S.A. E-mail: joemsong@cs.nmsu.edu.*

SUPPLEMENTARY NOTE 2. FISHER'S EXACT TEST

With the same null hypothesis of Pearson's chi-square test, Fisher's exact test [2], [3] computes the exact p-value of observed contingency table O to test the association between the row and column variables. The exact p-value is based on the multivariate hypergeometric distribution for the null population where all tables have the same row and column sums with the observed table O.

Let $\mathcal{A}$ represent the null population of all tables with fixed row and column sums to those observed in O. We can thus write $\mathcal{A}$ as

$$\mathcal{A} = \left\{ A \;\middle|\; \forall 1 \le i \le r,\ \sum_{j=1}^{c} A_{i,j} = \sum_{j=1}^{c} O_{i,j} \quad \text{and} \quad \forall 1 \le j \le c,\ \sum_{i=1}^{r} A_{i,j} = \sum_{i=1}^{r} O_{i,j} \right\} \tag{4}$$

We denote the row and column sums of A by

$$A_{i\cdot} = \sum_{j=1}^{c} A_{i,j} \tag{5}$$

and

$$A_{\cdot j} = \sum_{i=1}^{r} A_{i,j} \tag{6}$$

Let $\Pr(A)$ represent the null probability of matrix $A \in \mathcal{A}$. Under the null hypothesis, each sample in A is obtained by random sampling without repeat from the given row and column sums. Consequently, the counts in all cells in A follow a multivariate hypergeometric null distribution with mass probability function

$$\Pr(A) = \frac{\prod_{i=1}^{r} A_{i\cdot}! \cdot \prod_{j=1}^{c} A_{\cdot j}!}{n! \cdot \prod_{i=1}^{r} \prod_{j=1}^{c} A_{i,j}!} \tag{7}$$

which is an extension [3] beyond the 2×2 version originally described by Ronald A. Fisher [2].

The p-value of the observed table can be calculated exactly by summing up the probabilities of those tables no less extreme than the observed table O. Table A is no less extreme than O if and only if $\Pr(A) \le \Pr(O)$. An extreme table of low probability supports a strong association between the row and column variables. Therefore the p-value of the test given O is calculated by

$$p\text{-value} = \sum_{A \in \{A \mid \Pr(A) \le \Pr(O) \text{ and } A \in \mathcal{A}\}} \Pr(A) \tag{8}$$

This p-value is exact for all sample sizes giving rise to the name of Fisher's exact test. Due to fast algorithms [4], Fisher's exact test is computationally practical for reasonably large tables and sample sizes, and has become one of the most widely used exact tests.

SUPPLEMENTARY NOTE 3. SYMMETRY OF EXACT FUNCTIONAL TEST ON 2-BY-2 CONTINGENCY TABLES

Lemma 1. *For an observed 2-by-2 contingency table O and another 2-by-2 contingency table A with the same row and column sums of O, $\chi^2_{f^{-1}}(A) \ge \chi^2_{f^{-1}}(O)$ if and only if $\chi^2_f(A) \ge \chi^2_f(O)$, where $\chi^2_{f^{-1}}(A) = \chi^2_f(A^\top)$ and $\chi^2_{f^{-1}}(O) = \chi^2_f(O^\top)$.*

Proof. Let the observed 2-by-2 table O be

$$O = \begin{bmatrix} n_{11}^{obs} & n_{12}^{obs} \\ n_{21}^{obs} & n_{22}^{obs} \end{bmatrix} \tag{9}$$

and let another table A with the same row and column sums of O be

$$A = \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix} \tag{10}$$

Due to the row and column constraints on A, $\chi^2_f(A)$ is a univariate quadratic function and without loss of generality, we choose n_{11} as the free variable. By definition of functional chi-square, we have

$$\frac{\mathrm{d}\chi^2_f(A)}{\mathrm{d}n_{11}} = \frac{8}{n_{1\cdot}n_{2\cdot}}(nn_{11} - n_{1\cdot}n_{\cdot 1}) \tag{11}$$

and

$$\frac{\mathrm{d}\chi^2_{f^{-1}}(A)}{\mathrm{d}n_{11}} = \frac{8}{n_{\cdot 1}n_{\cdot 2}}(nn_{11} - n_{1\cdot}n_{\cdot 1}) \tag{12}$$

Both derivatives are equal to zero at the optimal solution

$$n^*_{11} = \frac{n_{1\cdot}n_{\cdot 1}}{n} \tag{13}$$

which also minimizes both $\chi^2_f(A)$ and $\chi^2_{f^{-1}}(A)$ to zero.

Both $\chi^2_f(A)$ and $\chi^2_{f^{-1}}(A)$ are quadratic functions decreasing before n^*_{11} and increasing afterwards.

Without loss of generality, we suppose $n^{obs}_{11} \le n^*_{11}$.

As both functional chi-square functions $\chi^2_f(A)$ and $\chi^2_{f^{-1}}(A)$ are monotonically decreasing in $(-\infty, n^*_{11}]$, for any $n_{11} \in (-\infty, n^{obs}_{11}]$, both $\chi^2_{f^{-1}}(A) \ge \chi^2_{f^{-1}}(O)$ and $\chi^2_f(A) \ge \chi^2_f(O)$ are true; for any $n_{11} \in (n^{obs}_{11}, n^*_{11}]$, both

$$\chi^2_{f^{-1}}(A) < \chi^2_{f^{-1}}(O) \quad \text{and} \quad \chi^2_f(A) < \chi^2_f(O) \tag{14}$$

are true.

Both functional chi-square functions $\chi^2_f(A)$ and $\chi^2_{f^{-1}}(A)$ are monotonically increasing in $[n^*_{11}, \infty)$. By symmetry of quadratic functions around $n_{11} = n^*_{11}$, we have exactly $\chi^2_f(A) = \chi^2_f(O)$ and $\chi^2_{f^{-1}}(A) = \chi^2_{f^{-1}}(O)$ at $n_{11} = 2n^*_{11} - n^{obs}_{11}$. Thus, for any $n_{11} \in [n^*_{11}, 2n^*_{11} - n^{obs}_{11})$, both $\chi^2_{f^{-1}}(A) < \chi^2_{f^{-1}}(O)$ and $\chi^2_f(A) < \chi^2_f(O)$ are true; for any $n_{11} \in [2n^*_{11} - n^{obs}_{11}, \infty)$, both

$$\chi^2_{f^{-1}}(A) \ge \chi^2_{f^{-1}}(O) \quad \text{and} \quad \chi^2_f(A) \ge \chi^2_f(O) \tag{15}$$

are true.

The case when $n^{obs}_{11} \ge n^*_{11}$ can be argued analogously.

Therefore, we have proven that $\chi^2_{f^{-1}}(A) \ge \chi^2_{f^{-1}}(O)$ if and only if $\chi^2_f(A) \ge \chi^2_f(O)$. □

Theorem 1. *The p-values of observed exact functional test statistics $\chi^2_f(O)$ and $\chi^2_{f^{-1}}(O)$ are equal for an observed 2-by-2 contingency table O.*

Proof. In computing the p-value of $\chi^2_{f^{-1}}(O)$, any table $A^\top$ that is no less extreme than $O^\top$ will have $\chi^2_{f^{-1}}(A) \ge \chi^2_{f^{-1}}(O)$ if and only if $\chi^2_f(A) \ge \chi^2_f(O)$ by Lemma 1. Additionally $\Pr(A) = \Pr(A^\top)$ by the definition of multivariate hypergeometric distribution. Therefore the p-values of test statistics $\chi^2_f(O)$ and $\chi^2_{f^{-1}}(O)$ are equal. □

SUPPLEMENTARY NOTE 4. A FAST AND EXACT ALGORITHM BY BRANCH AND BOUND

To compute the exact p-value by definition, one must enumerate all tables in the null population $\mathcal{A}$. The run time will be exponential in sample size and table dimension. We present a branch-and-bound algorithm in to speed up the calculation by skipping or including an entire branch of tables.

Just before $A_{i,j}$ is enumerated, we let $U_{l,q} \geq A_{l,q}$ be an upper bound of a not yet enumerated element $A_{l,q}$, constrained by the row and column sums and also those already enumerated cells as follows:

$$U_{l,q} = \begin{cases} \min\left\{O_{i\cdot} - \sum_{s=1}^{j-1} A_{i,s},\ O_{\cdot q} - \sum_{t=1}^{i-1} A_{t,q}\right\} & l = i, q \geq j \\ \min\left\{O_{l\cdot},\ O_{\cdot q} - \sum_{t=1}^{i} A_{t,q}\right\} & l > i, q < j \\ \min\left\{O_{l\cdot},\ O_{\cdot q} - \sum_{t=1}^{i-1} A_{t,q}\right\} & l > i, q \geq j \end{cases} \tag{16}$$

An upper bound of the functional chi-square statistic

We first establish an upper bound $UB(\chi^2_f(A))$ just before the enumeration of $A_{i,j}$. By sorting $U_{i,q}$ for $q = j, \ldots, c$ within row i, we obtain the order statistics

$$U_{i,(1)} \geq U_{i,(2)} \geq \cdots \geq U_{i,(c-j+1)} \qquad (\text{row } i) \tag{17}$$

with corresponding column index $q_{(1)}, q_{(2)}, \ldots, q_{(c-j+1)}$, or

$$U_{l,(1)} \geq U_{l,(2)} \geq \cdots \geq U_{l,(c)} \qquad (\text{row } l > i) \tag{18}$$

with corresponding column index $q_{(1)}, q_{(2)}, \ldots, q_{(c)}$.

To maximize the row functional chi-square

$$\sum_{q=1}^{c} \frac{(A_{l,q} - O_{l\cdot}/c)^2}{O_{l\cdot}/c} = \frac{c}{O_{l\cdot}} \left(\sum_{q=1}^{c} A_{l,q}^2\right) - O_{l\cdot} \tag{19}$$

we show that an optimal solution is, when $l = i$,

$$A^*_{i,(k)} = \begin{cases} \min\left\{U_{i,(1)}, O_{i\cdot} - \sum_{q=1}^{j-1} A_{i,q}\right\} & k = 1 \\ \min\left\{U_{i,(k)}, O_{i\cdot} - \sum_{q=1}^{j-1} A_{i,q} - \sum_{m=1}^{k-1} A^*_{i,(m)}\right\} & k > 1 \end{cases} \tag{20}$$

or, when $l > i$,

$$A^*_{l,(k)} = \begin{cases} \min\left\{U_{l,(1)}, O_{l\cdot}\right\} & k = 1 \\ \min\left\{U_{l,(k)}, O_{l\cdot} - \sum_{m=1}^{k-1} A^*_{l,(m)}\right\} & k > 1 \end{cases} \tag{21}$$

The optimality of the above solution is based on Theorem 2. The proof of the theorem is given in **Supplementary Note 5**.

The optimal solution gives rise to upper bounds on row chi-squares for all unenumerated elements starting at row i:

$$\chi^2_{l,\text{upper}} = \begin{cases} \sum_{k=1}^{c-j+1} \frac{\left(A^*_{i,(k)} - O_{i\cdot}/c\right)^2}{O_{i\cdot}/c} & l = i \\ \sum_{k=1}^{c} \frac{\left(A^*_{l,(k)} - O_{l\cdot}/c\right)^2}{O_{l\cdot}/c} & l > i \end{cases} \tag{22}$$

Thus, we find an upper bound of $\chi^2_f(A)$ before the enumeration of $A_{i,j}$:

$$UB(\chi^2_f(A)) = \sum_{l=1}^{i-1} \sum_{q=1}^{c} \frac{(A_{l,q} - O_{l\cdot}/c)^2}{O_{l\cdot}/c} + \sum_{q=1}^{j-1} \frac{(A_{i,q} - O_{i\cdot}/c)^2}{O_{i\cdot}/c} + \sum_{l=i}^{r} \chi^2_{l,\text{upper}} - \sum_{q=1}^{c} \frac{(O_{\cdot q} - n/c)^2}{n/c} \tag{23}$$

We skip the entire branch starting at $A(i,j)$ if and only if

$$UB(\chi^2_f(A)) < \chi^2_f(O) \tag{24}$$

Otherwise, we will enumerate $A_{i,j}$.

A lower bound of the functional chi-square statistic

We now establish a lower bound $LB(\chi_f^2(A))$ just before the enumeration of $A_{i,j}$. By sorting $U_{i,q}$ $(q \geq j)$ within row i, we obtain the order statistics

$$U_{i,(1)} \leq U_{i,(2)} \leq \cdots \leq U_{i,(c-j+1)} \qquad (\text{row } i) \tag{25}$$

with corresponding column index $q_{(1)}, q_{(2)}, \ldots, q_{(c-j+1)}$, and

$$U_{l,(1)} \leq U_{l,(2)} \leq \cdots \leq U_{l,(c)} \qquad (\text{row } l > i) \tag{26}$$

with corresponding column index $q_{(1)}, q_{(2)}, \ldots, q_{(c)}$.

Let $d_i(j)$ be the needed sum of counts on row i from column j to c:

$$d_i(j) = \begin{cases} O_{i\cdot} & j = 1 \\ O_{i\cdot} - \sum_{q=1}^{j-1} A_{i,q} & j > 1 \end{cases} \tag{27}$$

If the sum of count upper bounds on row i from column j to c is less than the needed sum $d_i(j)$, i.e.,

$$\sum_{q=j}^{c} U_{i,(q)} < d_i(j) \tag{28}$$

then there is no feasible solution for $A_{i,j}$ and we will return to enumerate another element before cell (i, j).

Otherwise, we claim that assigning

$$A^*_{i,(k)} \quad = \quad \begin{cases} \min\left\{U_{i,(k)}, d_i(j)/c\right\} & k = 1 \\ \min\left\{U_{i,(k)}, \left(d_i(j) - \sum_{q=1}^{k-1} A^*_{i,(q)}\right) / (c-k+1)\right\} & k > 1 \end{cases} \tag{29}$$

into cells from (i, j) to (i, c) achieves a lower bound of chi-square contributed by unenumerated elements in row i. A lower bound of chi-square for row $l > i$ can be achieved by assigning the following elements to row l:

$$A^*_{l,(k)} \quad = \quad \begin{cases} \min\left\{U_{l,(k)}, O_{l\cdot}/c\right\} & k = 1 \\ \min\left\{U_{l,(k)}, \left(O_{l\cdot} - \sum_{q=1}^{k-1} A^*_{l,(q)}\right) / (c-k+1)\right\} & k > 1 \end{cases} \tag{30}$$

The optimality of the assigned elements is based on Theorem 3. We prove the theorem in **Supplementary Note 6**.

The optimal solution gives rise to lower bounds on row chi-squares for all unenumerated elements starting at row i:

$$\chi^2_{l,\text{lower}} \quad = \quad \begin{cases} \sum_{k=1}^{c-j+1} \frac{\left(A^*_{i,(k)} - O_{i\cdot}/c\right)^2}{O_{i\cdot}/c} & l = i \\ \sum_{k=1}^{c} \frac{\left(A^*_{l,(k)} - O_{l\cdot}/c\right)^2}{O_{l\cdot}/c} & l > i \end{cases} \tag{31}$$

Thus, a lower bound of $\chi_f^2(A)$ before $A_{i,j}$ is enumerated can be calculated as

$$LB(\chi_f^2(A)) = \sum_{l=1}^{i-1}\sum_{q=1}^{c} \frac{(A_{l,q} - O_{l\cdot}/c)^2}{O_{l\cdot}/c} + \sum_{q=1}^{j-1} \frac{(A_{i,q} - O_{i\cdot}/c)^2}{O_{i\cdot}/c} + \sum_{l=i}^{r} \chi^2_{l,\text{lower}} - \sum_{q=1}^{c} \frac{(O_{\cdot q} - n/c)^2}{n/c} \tag{32}$$

We will enumerate $A_{i,j}$ if and only if

$$LB(\chi_f^2(A)) \quad \leq \quad \chi_f^2(O) \tag{33}$$

Otherwise, we will keep the entire branch starting at $A_{i,j}$. When elements remaining to be enumerated occupy entire rows

from i to r ($j = 1$), the probability of tables in the entire branch can be computed by a new table formed by the first $i-1$ row of A and a row containing $O_{\cdot q} - \sum_{l=1}^{i-1} A_{l,q}$ as follows:

$$\Pr\left(\{B|B \text{ shares the first } i-1 \text{ enumerated rows of } A \text{ and } B \in \mathcal{A}\}\right) = \frac{\left(\prod_{l=1}^{i-1} O_{l\cdot}!\right) \cdot (\sum_{l=i}^{r} O_{l\cdot})! \cdot \left(\prod_{q=1}^{c} O_{\cdot q}!\right)}{n! \cdot \left(\prod_{l=1}^{i-1}\prod_{q=1}^{c} A_{l,q}!\right) \cdot \left[\prod_{q=1}^{c}\left(O_{\cdot q} - \sum_{l=1}^{i-1} A_{l,q}\right)!\right]} \tag{34}$$

which is added to the extreme table probability, eventually becoming the p-value.

SUPPLEMENTARY NOTE 5. UPPER BOUND OF SUM OF SQUARES WITH LINEAR INEQUALITY CONSTRAINTS

Lemma 2. *Let x' be a sequence of n numbers, which are bounded by $x'_i \le u_i$ for $i = 1, \ldots, n$ with*

$$u_1 \ge u_2 \ge \cdots \ge u_n \tag{35}$$

Let $\hat{x}$ be the sorted version of x' in descending order, and thus

$$\hat{x}_1 \ge \hat{x}_2 \ge \cdots \ge \hat{x}_n \tag{36}$$

Then we must have $\hat{x}_i \le u_i$ for $i = 1, \ldots, n$.

Proof. Suppose $\hat{x}_i > u_i$. There would be at least i numbers $\hat{x}_1, \ldots, \hat{x}_i$ in $\hat{x}$ that are greater than u_i. But since only the first $i-1$ elements in x' are bounded by $u_1 \ge u_2 \ge \cdots \ge u_{i-1} (\ge u_i)$ and can thus be greater than u_i, x' can only provide at most $i-1$ elements greater than u_i. As $\hat{x}$ and x' contain two permutations of exactly the same numbers, it is a conflict to have both at least i and at most $i-1$ numbers greater than u_i. Therefore, we must have $\hat{x}_i \le u_i$ for all $i = 1, \ldots, n$. □

Lemma 3. *Given $x_1, \ldots, x_k \ge 0$ and a constant $s \ge 0$ such that*

$$x_1 + x_2 + \cdots + x_k = s \tag{37}$$

Let $v_1 \ge x_1$ be a given constant and we define recursively

$$z_i = \begin{cases} \min\{v_1, s\} - x_1 & i = 1 \\ \min\{x_2, z_1\} & i = 2 \\ \min\{x_i, z_1 - \sum_{q=2}^{i-1} z_q\} & i \ge 3 \end{cases} \tag{38}$$

Then, the following properties are true regarding z_i:

$$z_i \ge 0 \quad i = 1, \ldots, k \tag{39}$$

$$z_i \le x_i \quad i = 2, \ldots, k \tag{40}$$

$$z_1 = \sum_{i=2}^{k} z_i \tag{41}$$

Proof. Let $\widehat{i}$ be the smallest integer from 2 to k such that

$$x_{\widehat{i}} \ge z_1 \quad (\widehat{i} = 2) \tag{42}$$

or

$$x_{\widehat{i}} \geq z_1 - \sum_{q=2}^{\widehat{i}-1} z_q \quad (2 < \widehat{i} \leq k) \tag{43}$$

First we show that $\widehat{i}$ must exist. If such $\widehat{i}$ does not exist, we must have

$$x_k < z_1 - \sum_{q=2}^{k-1} z_q \tag{44}$$

and

$$z_i = x_i \quad i = 2, \ldots, k \tag{45}$$

The above two equations imply that

$$z_1 - \sum_{q=2}^{k} x_q > 0 \tag{46}$$

However, by definition we have

$$z_1 - \sum_{q=2}^{k} x_q \tag{47}$$

$$= z_1 - (s - x_1) \tag{48}$$

$$= (z_1 + x_1) - s \tag{49}$$

$$= \min\{v_1, s\} - s \tag{50}$$

$$\leq 0 \tag{51}$$

Evidently, Eq. (46) and (51) constitute a conflict. Therefore, such $\widehat{i}$ must exist.

If $\widehat{i} = 2$, we have $z_2 = z_1 \geq 0$ and $z_i = 0$ $(3 \leq i \leq k)$, which imply that all claims in Eqs. (39, 40, 41) of this lemma are true.

If $\widehat{i} > 2$, we have

$$z_i = x_i, \quad i = 2, \ldots, \widehat{i} - 1 \tag{52}$$

$$z_{\widehat{i}} = z_1 - \sum_{q=2}^{\widehat{i}-1} x_q \tag{53}$$

$$z_i = 0, \quad i = \widehat{i} + 1, \ldots, k \tag{54}$$

It follows that

$$\sum_{i=2}^{k} z_i = z_1 \tag{55}$$

By definition of z_i in Eq. (38), we have

$$z_i \leq x_i, \quad i = 2, \ldots, k \tag{56}$$

For $i = \widehat{i} - 1$, we must have by definition of $\widehat{i}$

$$x_{\widehat{i}-1} \leq z_1 - \sum_{q=2}^{\widehat{i}-2} x_q \tag{57}$$

which leads to

$$z_1 - \sum_{q=2}^{\widehat{i}-1} x_q \geq 0 \tag{58}$$

The left hand size of the inequality coincides with $z_{\widehat{i}}$ in Eq. (53) and thus

$$z_{\widehat{i}} \geq 0 \tag{59}$$

As $v_1 \geq x_1$ and $s \geq x_1$, we must have $z_1 \geq 0$. Together with Eqs. (52), (54), and (59), we have

$$z_i \geq 0, \quad i = 1, \ldots, k \tag{60}$$

With Eqs. (55), (56), and (60), we complete the proof of Lemma 3. □

Lemma 4. *Given n constants $\hat{x}_1 \geq \ldots \geq \hat{x}_n \geq 0$, and an objective function*

$$f(\hat{x}, z) = (\hat{x}_1 + z_1)^2 + \sum_{i=2}^{n} (\hat{x}_i - z_i)^2 \tag{61}$$

the following constrained quadratic program

$$\begin{cases} \min\limits_{z_1,\ldots,z_n} & f(\hat{x}, z) \\ \textit{subject to} & \sum\limits_{i=2}^{n} z_i = z_1 \\ & z_i \geq 0,\ i = 1, \ldots, n \\ & z_i \leq \hat{x}_i,\ i = 2, \ldots, n \end{cases} \tag{62}$$

is solved at $z_i^* = 0$ for $1 \leq i \leq n$.

Proof. We can re-write the problem following the convention used in [5]:

$$\begin{cases} \min\limits_{z_1,\ldots,z_n} & (\hat{x}_1 + z_1)^2 + \sum\limits_{i=2}^{n} (\hat{x}_i - z_i)^2 \\ \text{subject to} & \sum\limits_{i=2}^{n} z_i - z_1 = 0 \\ & z_i \geq 0,\ i = 1, \ldots, n \\ & \hat{x}_i - z_i \geq 0,\ i = 2, \ldots, n \end{cases} \tag{63}$$

We define Lagrangian multipliers $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_n)^\top$ and $\gamma = (\gamma_2, \ldots, \gamma_n)^\top$, and the Lagrangian function

$$L(z, \lambda, \gamma) = (\hat{x}_1 + z_1)^2 + \sum_{i=2}^{n} (\hat{x}_i - z_i)^2 - \lambda_0 \left(\sum_{i=2}^{n} z_i - z_1 \right) - \sum_{i=1}^{n} \lambda_i z_i - \sum_{i=2}^{n} \gamma_i (\hat{x}_i - z_i) \tag{64}$$

Then the gradient of $L(z, \lambda, \gamma)$ with respect to z is

$$\nabla_z L(z, \lambda, \gamma) = \begin{bmatrix} 2(\hat{x}_1 + z_1) + \lambda_0 - \lambda_1 \\ 2(z_2 - \hat{x}_2) - \lambda_0 - \lambda_2 + \gamma_2 \\ \vdots \\ 2(z_n - \hat{x}_n) - \lambda_0 - \lambda_n + \gamma_n \end{bmatrix} \tag{65}$$

The Karush-Kuhn-Tucker (KKT) first-order optimality condition for an optimal solution z^* and the corresponding λ^* and

γ^* of Eq. (63) is

$$\nabla_z L(z^*, \lambda^*, \gamma^*) = 0 \tag{66}$$

$$\lambda_0^* \left(\sum_{i=2}^{n} z_i^* - z_1^* \right) = 0 \tag{67}$$

$$\lambda_i^* \geq 0, 1 \leq i \leq n \tag{68}$$

$$\gamma_i^* \geq 0, 2 \leq i \leq n \tag{69}$$

$$\lambda_i^* z_i^* = 0, 1 \leq i \leq n \tag{70}$$

$$\gamma_i^* (\hat{x}_i - z_i^*) = 0, 2 \leq i \leq n \tag{71}$$

Next we can verify that the following values of z^*, λ^*, and γ^* satisfy the KKT first order optimality condition:

$$\begin{aligned}
&z_i^* = 0,\ 1 \leq i \leq n \\
&\lambda_0^* = -2\hat{x}_1 \\
&\lambda_1^* = 0 \\
&\lambda_i^* = 2(\hat{x}_1 - \hat{x}_i),\ 2 \leq i \leq n \\
&\gamma_i^* = 0,\ 2 \leq i \leq n
\end{aligned}$$

As the Hessian matrix of the objective function is $\nabla_{zz}^2 f(\hat{x}, z) = 2I$ and thus positive definite, z^* is a global minimizer of $f(\hat{x}, z)$ according to Theorem 16.4 [5]. We thus have solved the original quadratic program defined in Eq. (62). □

Theorem 2. *Given n variables $x_1, \ldots, x_n$ and constants $a > 0$ and $u_1 \geq \cdots \geq u_n \geq 0$ satisfying $\sum_{i=1}^{n} u_i \geq a$, the maximization problem*

$$\begin{cases}
\max\limits_{x_1,\ldots,x_n} & \sum\limits_{i=1}^{n} x_i^2 \\
\textit{subject to} & \sum\limits_{i=1}^{n} x_i = a \\
& x_i \geq 0,\ i = 1, \ldots, n \\
& x_i \leq u_i,\ i = 1, \ldots, n
\end{cases} \tag{72}$$

can be solved by

$$x_i^* = \begin{cases} \min\{u_i, a\} & i = 1 \\ \min\{u_i, a - \sum_{t=1}^{i-1} x_t^*\} & i > 1 \end{cases} \tag{73}$$

Proof. Let $x = (x_1, \ldots, x_n)^\top$. The objective function is equivalent to $\|x\|^2$, the square of the L_2 norm of x. We prove the theorem by constructing an optimal solution x^* element by element from x_1^* to x_n^*. The solution x^* is analogous to assigning a guests to n rooms with a total capacity no less than a, by filling the rooms in decreasing order of their capacity $u_1 \geq u_2 \geq \cdots \geq u_n$ until all a guests are assigned.

As the constrained problem is feasible and the objective function is bounded from above, there must exist a global optimal solution

$$x' = (x_1', \ldots, x_n')^\top \tag{74}$$

We next show that the solution

$$x^* = (x_1^*, \ldots, x_n^*)^\top \tag{75}$$

defined in Eq. (73) must also be a global optimal solution.

Let $\hat{x}$ be the sorted version of x' in descending order and thus

$$\hat{x}_1 \geq \hat{x}_2 \geq \cdots \geq \hat{x}_n \tag{76}$$

By Lemma 2, We must have $\hat{x}_i \leq u_i$ for all $i = 1, \ldots, n$. As the objective function is insensitive to the order of the n variables and $\hat{x}$ satisfies all constraints, $\hat{x}$ must be an optimal solution of the original problem.

By iteratively transforming $\hat{x}$ to match each element in x^*, we will demonstrate that x^* will not reduce the objective function value. We use vector $\hat{x}^i$ to represent the updated vector from $\hat{x}^{i-1}$ after x_i^* has replaced $\hat{x}_i^{i-1}$. Let $\hat{x}^0 = \hat{x}$, which is already an optimal solution.

We prove our claim by induction. At iteration i, we sort the last $n - i + 1$ elements in $\hat{x}^{i-1}$, i.e., $\hat{x}_i^{i-1}, \ldots, \hat{x}_n^{i-1}$, in descending order such that

$$\hat{x}_i^{i-1} \geq \cdots \geq \hat{x}_n^{i-1} \tag{77}$$

By Lemma 2, this transformed $\hat{x}^i$ is an optimal solution to the original problem since $\hat{x}^{i-1}$ before the transformation is an optimal solution.

Now we compute an adjustment vector z^i as follows:

$$z_1^i = 0 \tag{78}$$

$$\vdots$$

$$z_{i-1}^i = 0 \tag{79}$$

$$z_i^i = x_i^* - \hat{x}_i^{i-1} \tag{80}$$

$$z_{i+1}^i = \min\left\{\hat{x}_{i+1}^{i-1},\ z_i^i\right\} \tag{81}$$

$$z_{i+2}^i = \min\left\{\hat{x}_{i+2}^{i-1},\ z_i^i - z_{i+1}^i\right\} \tag{82}$$

$$\vdots$$

$$z_n^i = \min\left\{\hat{x}_n^{i-1},\ z_i^i - \sum_{q=i+1}^{n-1} z_q^i\right\} \tag{83}$$

Letting

$$v_1 = u_i \tag{84}$$

$$s = a - \sum_{t=1}^{i-1} x_t^* \tag{85}$$

$$k = n - i + 1 \tag{86}$$

and mapping z_q^i to z_{q-i+1} (as defined in Lemma 3) and $\hat{x}_q^{i-1}$ to x_{q-i+1} (as defined in Lemma 3) for $q = i, \ldots, n$, we can apply Lemma 3 and obtain

$$z_q^i \geq 0,\ q = i, \ldots, n \tag{87}$$

$$z_q^i \leq \hat{x}_q^{i-1},\ q = i+1, \ldots, n \tag{88}$$

$$z_i^i = \sum_{q=2}^{n} z_q^i \tag{89}$$

This leads to the update from $\hat{x}^{i-1}$ to $\hat{x}^i$ as follows:

$$\hat{x}_1^i = \hat{x}_1^{i-1} = x_1^* \tag{90}$$

$$\vdots$$

$$\hat{x}_{i-1}^i = \hat{x}_{i-1}^{i-1} = x_{i-1}^* \tag{91}$$

$$\hat{x}_i^i = \hat{x}_i^{i-1} + z_i^i = x_i^* \tag{92}$$

$$\hat{x}_{i+1}^i = \hat{x}_{i+1}^{i-1} - z_{i+1}^i \tag{93}$$

$$\vdots$$

$$\hat{x}_n^i = \hat{x}_n^{i-1} - z_n^i \tag{94}$$

The objective function is

$$\left\|\hat{x}^i\right\|^2 = \sum_{q=1}^{n} \left(\hat{x}_q^i\right)^2 \tag{95}$$

$$= \sum_{q=1}^{i-1} \left(\hat{x}_q^{i-1}\right)^2 + f\left(\left(\hat{x}_i^{i-1}, \ldots, \hat{x}_n^{i-1}\right)^\top, \left(z_i^i, \ldots, z_n^i\right)^\top\right) \tag{96}$$

$$\geq \sum_{q=1}^{i-1} \left(\hat{x}_q^{i-1}\right)^2 + f\left(\left(\hat{x}_i^{i-1}, \ldots, \hat{x}_n^{i-1}\right)^\top, (0, \ldots, 0)^\top\right) \qquad (\because \text{ Lemma 4}) \tag{97}$$

$$= \sum_{q=1}^{i-1} \left(\hat{x}_q^{i-1}\right)^2 + \sum_{q=i}^{n} \left(\hat{x}_q^{i-1}\right)^2 \tag{98}$$

$$= \sum_{q=1}^{n} \left(\hat{x}_q^{i-1}\right)^2 \tag{99}$$

$$= \left\|\hat{x}^{i-1}\right\|^2 \tag{100}$$

which establishes that $\hat{x}^i$ is also an optimal solution.

At $i = n$, we must have

$$\|\hat{x}\|^2 = \left\|\hat{x}^0\right\|^2 \leq \left\|\hat{x}^1\right\|^2 \leq \cdots \leq \left\|\hat{x}^n\right\|^2 = \|x^*\|^2 \tag{101}$$

Therefore, we have proven that x^* is another optimal solution to the problem defined in Eq. (72). This completes the proof of Theorem 2. □

SUPPLEMENTARY NOTE 6. LOWER BOUND OF SUM OF SQUARES WITH LINEAR INEQUALITY CONSTRAINTS

Given n variables $x_1, \ldots, x_n$, and constants $a, b > 0$ and $0 \leq u_1 \leq \cdots \leq u_n$ satisfying $\sum_{i=1}^n u_i \geq a$, our goal is to solve the quadratic programming problem with inequality constraints:

$$\begin{cases} \min\limits_{x_1,\ldots,x_n} & \sum\limits_{i=1}^{n} (x_i - b)^2 \\ \text{subject to} & \sum\limits_{i=1}^{n} x_i = a \\ & x_i \geq 0,\ i = 1, \ldots, n \\ & x_i \leq u_i,\ i = 1, \ldots, n \end{cases} \tag{102}$$

and we prove the following solution

$$x_i^* = \begin{cases} \min\{u_i, a/n\} & i = 1 \\ \min\{u_i, (a - \sum_{t=1}^{i-1} x_t^*)/(n-i+1)\} & i > 1 \end{cases} \tag{103}$$

satisfies a second-order sufficient condition for constrained optimization.

We first establish a property of the solution x^* in the following lemma.

Lemma 5. *For $x_1^*, \ldots, x_n^*$ defined in Eq. (103), there exists $0 \le \widehat{i} < n$ such that*

$$\begin{aligned} & x_1^* = u_1 \le x_2^* = u_2 \le \cdots \le x_{\widehat{i}}^* = u_{\widehat{i}} \\ \le\ & x_{\widehat{i}+1}^* = x_{\widehat{i}+2}^* = \cdots = x_n^* = \frac{a - \sum_{t=1}^{\widehat{i}} x_t^*}{n - (\widehat{i}+1) + 1} \end{aligned}$$

Proof. Let $\widehat{i}+1$ be the first time that

$$u_{\widehat{i}+1} \ge \frac{a}{n}\ (\widehat{i} = 0) \quad \text{or} \quad u_{\widehat{i}+1} \ge \frac{a - \sum_{t=1}^{\widehat{i}} x_t^*}{n - (\widehat{i}+1) + 1}\ (\widehat{i} \ge 1) \tag{104}$$

in the iteration defined in Eq (103). Such $\widehat{i}$ must exist because of the condition $\sum_{i=1}^n u_i \ge a$. It thus follows that

$$x_{\widehat{i}+1}^* = \frac{a - \sum_{t=1}^{\widehat{i}} x_t^*}{n - (\widehat{i}+1) + 1} \tag{105}$$

which implies

$$a - \sum_{t=1}^{\widehat{i}} x_t^* = x_{\widehat{i}+1}^* [n - (\widehat{i}+1) + 1] \tag{106}$$

Now we compute $x_{\widehat{i}+2}^*$ by definition

$$\begin{aligned} x_{\widehat{i}+2}^* &= \min\left\{ u_{\widehat{i}+2}, \frac{a - \sum_{t=1}^{\widehat{i}+1} x_t^*}{n - (\widehat{i}+2) + 1} \right\} \\ &= \min\left\{ u_{\widehat{i}+2}, \frac{a - \sum_{t=1}^{\widehat{i}} x_t^* - x_{\widehat{i}+1}^*}{n - (\widehat{i}+2) + 1} \right\} \\ &= \min\left\{ u_{\widehat{i}+2}, \frac{x_{\widehat{i}+1}^* [n - (\widehat{i}+1) + 1] - x_{\widehat{i}+1}^*}{n - (\widehat{i}+2) + 1} \right\} \\ &\qquad (\because \text{ Eq (106)}) \\ &= \min\left\{ u_{\widehat{i}+2},\ x_{\widehat{i}+1}^* \right\} \\ &= x_{\widehat{i}+1}^* \quad (\because x_{\widehat{i}+1}^* \le u_{\widehat{i}+1} \le u_{\widehat{i}+2}) \end{aligned}$$

Therefore we have

$$x_{\widehat{i}+1}^* = x_{\widehat{i}+2}^* = \cdots = x_n^* = \frac{a - \sum_{t=1}^{\widehat{i}} x_t^*}{n - (\widehat{i}+1) + 1} \tag{107}$$

By the definition of $\widehat{i}$, we must have for any $i \le \widehat{i}$

$$u_i < \frac{a}{n} \ \text{ for } i = 1; \qquad u_i < \frac{a - \sum_{t=1}^{i-1} x_t^*}{n - i + 1} \ \text{ for } 2 \le i \le \widehat{i} \tag{108}$$

and thus by definition of x_i^*, we also have

$$x_1^* = u_1 \leq x_2^* = u_2 \leq \cdots \leq x_{\widehat{i}}^* = u_{\widehat{i}} \tag{109}$$

Finally, we establish the relationship between $x_{\widehat{i}+1}^*$ and $x_{\widehat{i}}^*$. By definition of $x_{\widehat{i}}^*$, we have

$$x_{\widehat{i}}^* = u_{\widehat{i}} \leq \frac{a - \sum_{t=1}^{\widehat{i}-1} x_t^*}{n - \widehat{i} + 1} \tag{110}$$

which implies

$$a - \sum_{t=1}^{\widehat{i}-1} x_t^* \geq (n - \widehat{i} + 1) u_{\widehat{i}} = (n - \widehat{i} + 1) x_{\widehat{i}}^* \tag{111}$$

By Eq. (107), we have

$$\begin{aligned} x_{\widehat{i}+1}^* &= \frac{a - \sum_{t=1}^{\widehat{i}} x_t^*}{n - \widehat{i}} && (112)\\ &= \frac{a - \sum_{t=1}^{\widehat{i}-1} x_t^* - x_{\widehat{i}}^*}{n - \widehat{i}} && (113)\\ &\geq \frac{(n - \widehat{i} + 1) x_{\widehat{i}}^* - x_{\widehat{i}}^*}{n - \widehat{i}} \quad (\because Eq.\ (111)) && (114)\\ &= x_{\widehat{i}}^* && (115) \end{aligned}$$

Eqs. (107), (109), and (115) complete the proof of Lemma 5. □

We can re-write the original problem in Eq. (102) in a canonical form [5] by

$$\begin{cases} \min\limits_{x_1,\ldots,x_n} & \sum\limits_{i=1}^{n} (x_i - b)^2 \\ \text{subject to} & \sum\limits_{i=1}^{n} x_i - a = 0 \\ & x_i \geq 0,\ i = 1, \ldots, n \\ & -x_i + u_i \geq 0,\ i = 1, \ldots, n \end{cases} \tag{116}$$

Let vector $x = (x_1, \ldots, x_n)^\top$. We define the Lagrangian multiplier vector

$$\lambda = (\lambda_0, \lambda_{1,1}, \lambda_{1,2}, \ldots, \lambda_{n,1}, \lambda_{n,2})^\top \tag{117}$$

We define the Lagrangian function

$$L(x, \lambda) = \sum_{i=1}^{n} (x_i - b)^2 - \sum_{i=1}^{n} \lambda_{i,1} x_i - \sum_{i=1}^{n} \lambda_{i,2} (-x_i + u_i) - \lambda_0 \left[\left(\sum_{i=1}^{n} x_i \right) - a \right] \tag{118}$$

whose gradient with respect to x can be calculated by

$$\nabla_x L(x, \lambda) = \begin{bmatrix} 2(x_1 - b) - \lambda_{1,1} + \lambda_{1,2} - \lambda_0 \\ \vdots \\ 2(x_i - b) - \lambda_{i,1} + \lambda_{i,2} - \lambda_0 \\ \vdots \\ 2(x_n - b) - \lambda_{n,1} + \lambda_{n,2} - \lambda_0 \end{bmatrix} \tag{119}$$

We next prove that x^* and an implied λ^* satisfy the optimal KKT conditions for inequality-constrained quadratic programming problems [5] (page 464) stated in the following lemma.

Lemma 6 (Satisfaction of Karush-Kuhn-Tucker condition)**.** *There exists λ^* associated with x^* such that the following KKT condition is satisfied:*

$$\nabla_x L(x^*, \lambda^*) = 0, \tag{120}$$

$$\sum_{i=1}^{n} x_i^* - a = 0, \tag{121}$$

$$x_i^* \geq 0, \quad i = 1, \ldots, n \tag{122}$$

$$-x_i^* + u_i \geq 0, \quad i = 1, \ldots, n \tag{123}$$

$$\lambda_{i,1}^* \geq 0, \quad i = 1, \ldots, n \tag{124}$$

$$\lambda_{i,2}^* \geq 0, \quad i = 1, \ldots, n \tag{125}$$

$$\lambda_0^* \left(\sum_{i=1}^{n} x_i^* - a \right) = 0, \tag{126}$$

$$\lambda_{i,1}^* x_i^* = 0, \quad i = 1, \ldots, n \tag{127}$$

$$\lambda_{i,2}^* (-x_i^* + u_i) = 0, \quad i = 1, \ldots, n \tag{128}$$

Proof. By definition of x_i^* in Eq. (103), Eq. (122) and Eq. (123) are always true. Using Lemma 5, we can derive that $\sum_{i=1}^{n} x_i^* = a$, leading to satisfaction of both Eq. (121) and Eq. (126).

We choose $\lambda_{i,1}^* = 0$ for all i to satisfy Eq. (124) and Eq. (127).

We also choose $\lambda_{i,2}^* = 0$ to satisfy Eq. (128) for $i > \widehat{i}$ ($\widehat{i}$ is defined in Lemma 5). When $i \leq \widehat{i}$, we have $x_i^* = u_i$ which satisfy Eq. (128) for $i \leq \widehat{i}$.

When $i > \widehat{i}$, Eq. (120) is true if and only if

$$2(x_i^* - b) - \lambda_0^* = 0, \quad \text{for } i > \widehat{i}$$

and by Lemma 5 we have

$$\lambda_0^* = 2(x_{\widehat{i}+1}^* - b) = \frac{2a - 2\sum_{t=1}^{\widehat{i}} x_t^*}{n - (\widehat{i}+1) + 1} - 2b \tag{129}$$

When $i \leq \widehat{i}$, Eq. (120) is true if and only if

$$2(x_i^* - b) + \lambda_{i,2}^* - \lambda_0^* = 0, \quad \text{for } i \leq \widehat{i}$$

which implies by Lemma 5

$$\lambda_{i,2}^* = 2(x_{\widehat{i}+1}^* - x_i^*) \geq 0, \quad \text{for } i \leq \widehat{i} \tag{130}$$

Together with $\lambda_{i,2}^* = 0$ chosen for $i > \widehat{i}$ earlier, $\lambda_{i,2}^*$ now satisfies Eq. (125). Therefore, x^* and λ^* meet all required KKT conditions. This completes the proof of Lemma 6. □

Theorem 3. *Given n variables $x_1, \ldots, x_n$ and constants $a, b > 0$ and $0 \leq u_1 \leq \cdots \leq u_n$ satisfying $\sum_{i=1}^{n} u_i \geq a$, the quadratic*

programming problem

$$\begin{cases} \min\limits_{x_1,\ldots,x_n} & \sum\limits_{i=1}^{n}(x_i-b)^2 \\ \textit{subject to} & \sum\limits_{i=1}^{n} x_i = a \\ & x_i \geq 0,\ i=1,\ldots,n \\ & x_i \leq u_i,\ i=1,\ldots,n \end{cases} \tag{131}$$

can be solved by

$$x_i^* = \begin{cases} \min\{u_i, a/n\} & i=1 \\ \min\{u_i, (a-\sum_{t=1}^{i-1} x_t^*)/(n-i+1)\} & i>1 \end{cases} \tag{132}$$

Proof. By definition of $L(x,\lambda)$, the Hessian matrix of $L(x,\lambda)$ is

$$\nabla_{xx}^2 L(x,\lambda) = 2I \tag{133}$$

where I is the $n \times n$ identity matrix, and it always satisfies a second-order condition

$$w^\top \nabla_{xx}^2 L(x^*,\lambda^*) w > 0 \quad \text{for } w \neq 0$$

Together with the KKT condition established in Lemma 6, x^* and λ^* satisfy a second-order sufficient condition for constrained optimization by Theorem 12.6 [5]. Therefore, x^* is a strict local solution. The objective function can rewritten as

$$\sum_{i=1}^{n}(x_i-b)^2 = x^\top I x - 2b x^\top \mathbf{1} + nb^2$$

where $\mathbf{1} = [1,\ldots,1]^\top$. By Theorem 16.4 [5], the positive definite identity matrix I in the quadratic function implies that x^* is also a global solution of the constrained optimization problem defined in Eq. (131). □

SUPPLEMENTARY NOTE 7. THE HOUSE NOISE MODEL

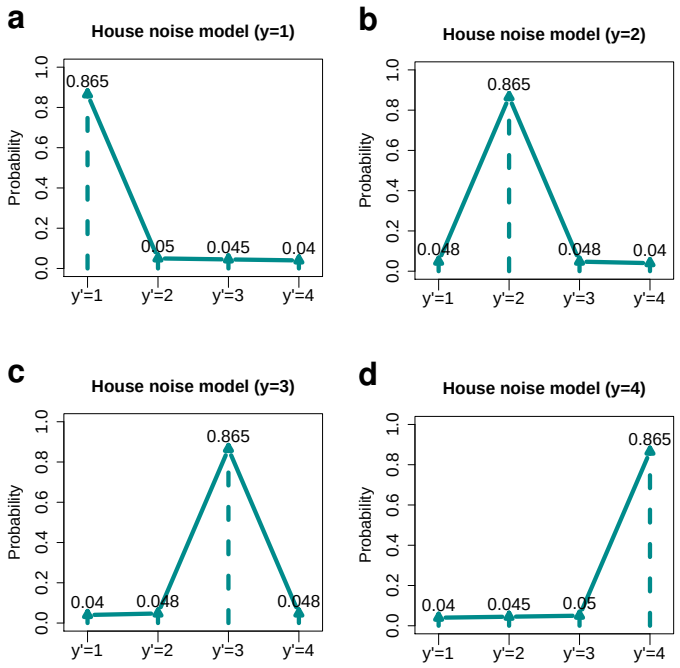
The model specifies a conditional probability function to transform the value $Y=y$ of a discrete random variable Y to a noisy version $Y'=y'$ defined by

$$P_{Y\prime|Y}(y\prime|y,\theta) \quad = \quad \begin{cases} \left[\left(1-\frac{|y\prime-y|}{\sum_{d=1}^{c}|d-y|}\right)\frac{\theta}{c-1}\right](1-\theta)+\frac{\theta}{c}, & y\prime \neq y \\ \left[\frac{\theta}{c-1}+1-\theta\right](1-\theta)+\frac{\theta}{c}, & y\prime = y \end{cases} \tag{134}$$

where c is the number of discrete levels of $Y, Y' \in \{1,\ldots,c\}$, and parameter $\theta \in [0,1]$ represents the noise level. In this model, the conditional probability decreases as y' moves further away from y. Supplementary Figure 1 illustrates the house noise model for a random variable Y with four levels at the noise level of $\theta = 0.1$.

SUPPLEMENTARY NOTE 8. SCREENING PUTATIVE NONCODING RNAS ON WHICH LUNG CANCER PHENOTYPES FUNCTIONALLY DEPENDS

From a total of $91,213$ unannotated transcription start sites (TSSs) in the phase 1 data of FANTOM5 [6], [7] with annotation based on Gencode V10 [8], we tested the functional dependency of lung cancer phenotypes on each TSS. We also evaluated 589 experimentally validated cancer driver genes obtained from Network of Cancer Genes 5.0 [9] originally taken from



Supplementary Figure 1. **An example of the house noise model.** The noise level is $\theta = 0.1$. Noise-free discrete variable Y has four levels from 1 to 4. The horizontal axis is its noisy version Y'. The vertical axis is the probability of observing $Y' = y'$ conditioned on $Y = y$ at the given noise level. The conditional probability mass functions of $\Pr(Y' = y'|Y = y)$ are shown for (a) $y = 1$, (b) $y = 2$, (c) $y = 3$ and (d) $y = 4$.

Cancer Gene Census [10]. Using the TSS abundance of these known genes as a reference, we selected 1049 unannotated TSSs as candidates for lung cancer markers.

We performed the exact functional test on 32 samples related to lung in the FANTOM5 TSS expression data. They include 17 cell line samples covering 11 lung cancer subtypes:

1) adenocarcinoma,
2) giant cell carcinoma,
3) squamous cell carcinoma,
4) large cell carcinoma,
5) small cell carcinoma,
6) alveolar cell carcinoma,
7) bronchioalveolar carcinoma,
8) bronchogenic carcinoma,
9) fibrous histiocytoma,
10) bronchial squamous cell carcinoma, and
11) non-small cell cancer.

They also include 15 normal lung samples from three normal lung tissues

1) adult lung cell,
2) fetal lung cell, and
3) the right lower lobe of adult lung cell

and four lung primary cells

1) small airway epithelial cell,

2) alveolar epithelial cell,
3) bronchial epithelial cell, and
4) smooth muscle cell.

We predicted novel putative noncoding RNA TSSs directionally associated with lung cancer in the following steps:

1) Normalize the data for each TSS by nonlinear transform. We iteratively applied logarithm transform on the expression values (tags per million) of each TSS from all 899 FANTOM5 samples, until all the values of each TSS follow a normal distribution as indicated by Shapiro test statistic $W \geq 0.9$. If the iterative logarithm transform could not increase W, we used the original data. This step is important for the exact functional test to operate at the correct scale for each TSS;
2) Discretize the log-transformed expression values for each TSS on all 899 samples using R package `Ckmeans.1d.dp` [11]. The level of discretization was selected by R package `mclust` [12]. Using all 899 samples for discretization, we took the dynamic range of each TSS into account. Thus, the discretized values obtained were more robust than using only the 32 lung samples;
3) Perform the exact functional test to examine the functional dependency of lung cancer on each TSS and obtain p-values. The tests were applied only on the discrete expression levels of each TSS from the 32 lung samples;
4) Select those unannotated TSSs which have stronger functional effect than known cancer genes on lung cancer phenotypes: the p-value must be no greater than 95% of all 589 well-known cancer genes from Cancer Gene Census.

Finally, we obtained 1049 unannotated TSSs, with a p-value threshold of 0.015 taken from step 4 above. They are listed in

Supplementary File 6 `"Novel_unannotated_cancer_TSSs.tsv"`

Although some TSSs may originate from novel coding exons previously unannotated, most of them are likely to be from novel noncoding RNAs. These TSSs constitute novel hypotheses for potential directional association with lung cancers.

REFERENCES

[1] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine Series 5*, vol. 50, no. 302, pp. 157–175, 1900.

[2] R. A. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of P," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[3] G. Freeman and J. H. Halton, "Note on an exact treatment of contingency, goodness of fit and other problems of significance," *Biometrika*, vol. 38, no. 1/2, pp. 141–149, 1951.

[4] C. R. Mehta and N. R. Patel, "A network algorithm for performing Fisher's exact test in r×c contingency tables," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 427–434, 1983.

[5] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006.

[6] A. R. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. L. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. Jorgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. A. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. C. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistraci, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drablos, A. S. B. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J.-i. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofmann,

I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. J. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R.-i. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. D. Prendergast, O. J. L. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. C. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verado, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki, "A promoter-level mammalian expression atlas," *Nature*, vol. 507, no. 7493, pp. 462–470, Mar 2014.

[7] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C. J. Mungall, E. Arner, J. K. Baillie, N. Bertin, H. Bono, M. de Hoon, A. D. Diehl, E. Dimont, T. C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T. F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E. A. Schultes, P. A. C. 't Hoen, Z. Tatum, M. Thompson, T. Toyoda, D. W. Wright, C. O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A. R. R. Forrest, and H. Kawaji, "Gateways to the FANTOM5 promoter level mammalian expression atlas," *Genome Biology*, vol. 16, no. 1, p. 22, 2015.

[8] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[9] O. An, G. M. Dall'Olio, T. P. Mourikis, and F. D. Ciccarelli, "NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings," *Nucleic Acids Research*, vol. 44, no. D1, pp. D992–D999, 2015.

[10] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[11] H. Wang and M. Song, "Ckmeans.1d.dp: Optimal *k*-means clustering in one dimension by dynamic programming," *The R Journal*, vol. 3, no. 2, pp. 29–33, 2011.

[12] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, 2012.