Supporting Text

Analytical Calculation

Under the model assumptions we can write down an expression for $P_{n,\alpha}(f)$ that reflects a process of drawing N_g independent "true" correlations Z_t from the distribution $q(Z_t)$, each of which is submitted to a Gaussian noise of variance ${\sigma_n}^2$, and identifying the $N_{TOP}(=\alpha N_g)$ top genes. Submitting the N_g true values to another realization of the noise, we obtain another list of N_{TOP} genes. For finite n, the lists are expected to be different due to noise (nonvanishing ${\sigma_n}^2$). The probability to obtain an overlap f between two PGLs, $P_{n,\alpha}(f)$, is given by EQ. **2** in the main text which is specified here in more details:

$$
\begin{aligned}
P_{n,\alpha}(f) = \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \sum_{h,l\in\{0,1\}^{N_g}} & \Bigg\{ \delta\Big(\sum_{j=1}^{N_g} h_j - N_{TOP}\Big)\delta\Big(\sum_{j=1}^{N_g} l_j - N_{TOP}\Big)\delta\Big(\sum_{j=1}^{N_g} h_j l_j - fN_{TOP}\Big) \\
\prod_{j=1}^{N_g}\Bigg[(1-h_j)\int_{-x_1}^{x_1} dZ^1_{mj}\frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{(Z^1_{mj}-Z_j)^2}{2{\sigma_n}^2}} &+ h_j\Big(1-\int_{-x_1}^{x_1} dZ^1_{mj}\frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{(Z^1_{mj}-Z_j)^2}{2{\sigma_n}^2}}\Big)\Bigg] \\
\prod_{k=1}^{N_g}\Bigg[(1-l_k)\int_{-x_2}^{x_2} dZ^2_{mk}\frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{(Z^2_{mk}-Z_k)^2}{2{\sigma_n}^2}} &+ l_k\Big(1-\int_{-x_2}^{x_2} dZ^2_{mk}\frac{1}{\sqrt{2\pi}\sigma_n}e^{-\frac{(Z^2_{mk}-Z_k)^2}{2{\sigma_n}^2}}\Big)\Bigg]\Bigg\}. \qquad (1)
\end{aligned}
$$

Here $\delta(.)$ is the Kronecker delta, Z_j is the true correlation of the jth gene, and Z^1_{mj}, Z^2_{mj} are the measured correlations of the jth gene in the first and second realizations, respectively. Nr is a normalization factor. $h = (h_1, \ldots, h_{N_g})$ and $l = (l_1, \ldots, h_{N_g})$ are binary vectors of size N_g whose nonzero elements correspond to the genes included in the PGLs of the first and the second realizations, respectively. The integration variables x_1, x_2 can be thought of as artificial "thresholds" which separate the N_{TOP} top correlations from the rest in the two realizations. The density of x_1, x_2 has no effect in the large N_g limit, and is thus omitted here. Replacing the delta functions in EQ. **1** by their integral representations one obtains:

$$
\begin{aligned}
P_{n,\alpha}(f) \quad = \quad & \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \sum_{h,l\in\{0,1\}^{N_g}} \\
& \Bigg\{\int_{-\pi}^{\pi}\frac{dydzdw}{(2\pi)^3}e^{iy\left(\sum_{j=1}^{N_g} h_j - N_{TOP}\right)+iz\left(\sum_{j=1}^{N_g} l_j - N_{TOP}\right)+iw\left(\sum_{j=1}^{N_g} h_j l_j - fN_{TOP}\right)} \\
& \prod_{j=1}^{N_g}\Big[(1-h_j)P(x_1,Z_j) + h_j(1-P(x_1,Z_j))\Big]\Big[(1-l_j)P(x_2,Z_j) + l_j(1-P(x_2,Z_j))\Big]\Bigg\}, \qquad (2)
\end{aligned}
$$

where $P(x,Z) \equiv P(x,Z,\sigma_n) = \int_{-x}^{x} dZ_m \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(Z_m-Z)^2}{2\sigma_n^2}}$. (From now on, we shall omit the dependence on σ_n

in P.) Simple manipulations yield

$$\begin{aligned} P_{n,\alpha}(f) &= \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \int_{-\pi}^{\pi}\frac{dydzdw}{(2\pi)^3}\sum_{h,l\in\{0,1\}^{N_g}}\Bigg\{\prod_{j=1}^{N_g}\Big[(1-h_j)P(x_1,Z_j)+h_j(1-P(x_1,Z_j))\Big] \\ &\qquad \Big[(1-l_j)P(x_2,Z_j)+l_j(1-P(x_2,Z_j))\Big]e^{(iyh_j+izl_j+iwh_jl_j)}e^{(-iyN_{TOP}-izN_{TOP}-iwfN_{TOP})}\Bigg\}, \end{aligned} \tag{3}$$

which results in

$$P_{n,\alpha}(f) = \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \int_{-\pi}^{\pi}\frac{dydzdw}{(2\pi)^3}\prod_{j=1}^{N_g}B(x_1,x_2,y,z,w,Z_j), \tag{4}$$

where

$$\begin{aligned} B(x_1,x_2,y,z,w,Z) &= P(x_1,Z)P(x_2,Z)+P(x_1,Z)(1-P(x_2,Z))e^{iz} \\ &+ (1-P(x_1,Z))P(x_2,Z)e^{iy}+(1-P(x_1,Z))(1-P(x_2,Z))(e^{iy}+e^{iz}+e^{iw}). \end{aligned} \tag{5}$$

For $N_g \gg 1$, one can approximate summation over the Z_j by integrating $dq(Z)$, which for symmetric $q(Z)$ gives

$$\begin{aligned} P_{n,\alpha}(f) &= \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \int_{-\pi}^{\pi}\frac{dydzdw}{(2\pi)^3} \\ &\qquad \exp\Big(-N_g\left(i\alpha y+i\alpha z+i\alpha f w\right)\Big)\exp\Big(2N_g\int_0^\infty dZq(Z)\ln\left(B(x_1,x_2,y,z,w,Z)\right)\Big). \end{aligned} \tag{6}$$

Defining A as

$$\begin{aligned} A(x_1,x_2,y,z,w,Z) &= P(x_1,Z)P(x_2,Z)\left(e^{-i(y+z+w)}-e^{-i(y+w)}-e^{-i(z+w)}+1\right) \\ &+ P(x_1,Z)(e^{-i(y+w)}-1)+P(x_2,Z)(e^{-i(z+w)}-1)+1, \end{aligned} \tag{7}$$

yields

$$P_{n,\alpha}(f) = \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \int_{-\pi}^{\pi}\frac{dydzdw}{(2\pi)^3}$$

$$\exp\Big(-N_g\left(i\left(1-\alpha\right)y-i(1-\alpha)z-i(1-\alpha f)w\right)\Big)\exp\Big(2N_g\int_0^\infty dZq(Z)\ln\left(A(x_1,x_2,y,z,w,Z)\right)\Big). \tag{8}$$

The above integral can be written as

$$P_{n,\alpha}(f) = \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \int_{-\pi}^{\pi}\frac{dydzdw}{(2\pi)^3}\exp\left(-N_gF\right), \tag{9}$$

where

$$F(x_1,x_2,y,x,w;f) = -i(1-\alpha)y-i(1-\alpha)z-i(1-\alpha f)w-2\int_0^\infty q(Z)dZ\ln(A(x_1,x_2,y,z,w,Z)). \tag{10}$$

The Saddle Point (SP) equations $\nabla F = 0$ (where f is treated as a parameter) are

$$\begin{aligned}
1-\alpha &= 2\int_0^\infty q(Z)dZ\frac{e^{-i(y+w)}(e^{-iz}-1)P(x_1,Z)P(x_2,Z)+e^{-i(y+w)}P(x_1,Z)}{A(x_1,x_2,y,z,w,Z)} \\
1-\alpha &= 2\int_0^\infty q(Z)dZ\frac{e^{-i(z+w)}(e^{-iy}-1)P(x_1,Z)P(x_2,Z)+e^{-i(z+w)}P(x_1,Z)}{A(x_1,x_2,y,z,w,Z)} \\
1-\alpha f &= 2\int_0^\infty q(Z)dZ\frac{e^{-iw}(e^{-i(y+z)}-e^{-iy}-e^{-iz})P(x_1,Z)P(x_2,Z)+e^{-i(y+w)}P(x_1,Z)+e^{-i(z+w)}P(x_2,Z)}{A(x_1,x_2,y,z,w,Z)} \\
0 &= \int_0^\infty q(Z)dZ\frac{P_{x_1}(x_1,Z)\left[(e^{-i(y+z+w)}-e^{-i(y+z)}-e^{-i(z+w)}+1)P(x_2,Z)+e^{-i(y+w)}-1\right]}{A(x_1,x_2,y,z,w,Z)} \\
0 &= \int_0^\infty q(Z)dZ\frac{P_{x_2}(x_2,Z)\left[(e^{-i(y+z+w)}-e^{-i(y+z)}-e^{-i(z+w)}+1)P(x_1,Z)+e^{-i(z+w)}-1\right]}{A(x_1,x_2,y,z,w,Z)},
\end{aligned} \tag{11}$$

where $P_x(x,Z)$ is the derivative of $P(x,Z)$ with respect to x. For Nr the SP equations are those shown in **11** and an additional equation, $w=0$, which is obtained from $\frac{\partial F}{\partial f}=0$. Substituting $w=0$ in the last two SP equations in **11** one obtains $y,z=0$, and the first three SP equations become:

$$\begin{aligned}
1-\alpha &= 2\int_0^\infty q(Z)dZP(x_1,Z) \\
1-\alpha &= 2\int_0^\infty q(Z)dZP(x_2,Z) \\
1-\alpha f &= 2\int_0^\infty q(Z)dZ\Big(P(x_1,Z)+P(x_2,Z)-P(x_1,Z)P(x_2,Z)\Big).
\end{aligned} \tag{12}$$

Note that the last equation can be written in a more meaningful way as

$$\alpha f = 2\int_0^\infty q(Z)dZ\left(1-P(x_1,Z)\right)\left(1-P(x_2,Z)\right). \tag{13}$$

For very large N_g, the SP expansion gives

$$P_{n,\alpha}(f) \sim \sqrt{\frac{N_g \det R}{2\pi \det H}}e^{-N_g(F(f)-F(f_n^*))}, \tag{14}$$

where $R_{5\times5}$ and $H_{6\times6}$ are the second derivative matrices of F at the saddle point with respect to (x_1,x_2,y,z,w) and (x_1,x_2,y,z,w,f), respectively, det denotes matrix determinant, and f_n^* is the value of f minimizing F. Thus, f_n^* is obtained by taking the value of f in the solution of the set of the SP EQS. **12** (which are easily solved numerically). For large N_g, $P_{n,\alpha}(f)$ gets a sharp maximum at $f=f_n^*$, and as $N_g\to\infty$, $P_{n,\alpha}(f)$ tends to a delta function at $f=f_n^*$. The meaning of this result is that for $N_g\to\infty$, and for a given finite number of samples n, the values of both x and f are independent of the specific selection of the n samples. Expanding EQ. **14** into a series around f_n^* and keeping the leading term one obtains our final expression for $P_{n,\alpha}(f)$

$$P_{n,\alpha}(f) \sim \frac{1}{\sqrt{2\pi}\Sigma_n}e^{-\frac{(f-f_n^*)^2}{2\Sigma_n^2}}, \tag{15}$$

where the variance Σ_n^2 is given by:

$$\Sigma_n^2 = \frac{detH}{N_g detR}. \tag{16}$$

Simulations

Adjusting $\mu_g(i)-\mu_p(i)$ to fit the true distribution

As described in the main text, the two Gaussians G($\mu_g(i)$,$\sigma_g(i)$) and G($\mu_p(i)$,$\sigma_p(i)$) are approximating the probability distribution of the expression of gene i for $n=N_s$. However, we are interested in the true distributions, namely those corresponding to infinite N_s. Therefore, we have to rescale $\Delta\mu(i) \equiv \mu_g(i) - \mu_p(i)$ so that the distribution of the resulting correlations will fit $q(Z_t)$. The rescaling can be done, for example, by keeping $\mu_g(i)$ (and $\sigma_g(i), \sigma_p(i)$), and changing $\mu_p(i)$ such that we get

$$\Delta\mu(i) = Z_{mi}\sqrt{\frac{V_t}{V_t+\sigma_n^2}\frac{P_L\sigma_g(i)^2+(1-P_L)\sigma_p(i)^2}{P_L(1-P_L)(1-Z_i^2)}}, \tag{17}$$

where P_L is the relative fraction of good outcome patients in the data set.

Motivation for Creating the Simulation Model

The most straightforward way to perform simulations is the following: For each n, divide the data set into the maximal number of nonoverlapping training sets of size n, $K(n) \equiv \lfloor \frac{N_s}{n} \rfloor$. (Clearly, allowing overlaps between the training sets will result in an overestimate of f.) For each training set generate a PGL, ending up with $K(n)$ lists. Then calculate the overlaps between the $K(n)/2$ independent pairs of lists. Repeat this procedure T times to obtain $T \cdot K(n)/2$ overlap values whose mean and variance have to match the analytical prediction of f_n^*, and Σ_n^2 respectively. We have found that performing the simulations in this way gives strong data-dependent fluctuations in the estimates of f_n^* (for a fixed value of n), resulting in sometimes a nonmonotonic behavior of f_n^* in n. This was observed both for the biological and simulated data sets (data not shown). We note that this instability in the estimate of f_n^* cannot be attributed to the lack of computational resources (i.e., too small number of repeats T) and may occur even if one enumerates all possible $\binom{N_s}{n\,n\,..\,n}$ partitions for a given data set.

To overcome this problem, we created our model, which in addition to eliminating the aforementioned phenomenon, allows to produce simulation results for unlimited n as opposed to the aforementioned procedure, which is limited to $n = N_s/2$.

Checking the Model Assumptions on the Real Data Sets

Our analytical calculation is based on four main assumptions:

Assumption 1: The distributions of the measured Z's are Gaussian, centered for each gene around its Z_t.

Assumption 2: The variance σ_n^2 is the same for all genes.

Assumption 3: The noise variables $Z - Z_t$ are independent (i.e. uncorrelated noise for different genes).

Assumption 4: $q(Z_t)$, the distribution of the true correlations, can be approximated by a Gaussian with variance V_t. This assumption is easily generalized to represent $q(Z_t)$ as a mixture of Gaussians.

The successful application of our method to real data sets depends on the extent to which our assumptions hold. We checked these assumptions on the six data sets analyzed in this work. The validity of *Assumption 1* is demonstrated in Fig. 4, for five randomly peaked genes from each data set. We have checked it also for many other genes, and the Z's distributions of almost all genes were very well approximated by Gaussians.

Results for *Assumption* 2 appear in Fig. 5. The histograms were generated by selecting 1,000 random pairs of nonoverlapping training sets of size $n = N_s/2$. The correlation of each gene with survival was calculated in each training set, and the variance of its transformed correlation, Z, within each pair was recorded. This resulted in 1,000 variances for each gene as obtained from the 1,000 randomly generated pairs of training sets. The average of these 1,000 values was the estimate for the noise variance of each gene. The variance histograms of the six data sets are tightly centered around the mean value $\hat{\sigma}_n^2$ (red vertical lines) which is very close to the analytical value $1/(n-3)$ (green vertical lines), implying that the data sets can be well described using *Assumption 2*. A relatively less centered histogram is obtained for lung cancer (1) which may explain the relatively high deviations between analytical prediction and simulations observed in this data set.

Results for *Assumption 3* appear in Table 2. In most data sets the fitted a, b values satisfy $a \approx 1$ and $b \approx -1$, which implies that the noise of the genes is uncorrelated (see explanation in *Materials and Methods*).

Results for *Assumption 4* are exhibited in Fig. 6. Since we used $n = N_s$, our sampling noise σ_n^2 is rather small. The nice match between the fits and the real histograms of the measured q_n therefore reflects the validity of this assumption also for the true distribution q.

References

[1] Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., *et al.* (2002) *Nat. Med.* **8**, 816-824.

[2] van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., *et al.* (2002) *N. Engl. J. Med.* **347**, 1999-2009.

[3] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., *et al.* (2005) *Lancet* **19-25;365(9460)**, 671-679.

[4] Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., *et al.* (2002) *N. Engl. J. Med.* **346**, 1937-1947.

[5] Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A., Tabuchi, H., *et al.* (2003) *Lancet* **361**, 923929.

[6] Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., *et al.* (2001), *Proc. Natl. Acad. Sci. USA* **98(24)**, 13790-13795.