SUPPLEMENTARY METHODS

The normalmixEM() function the mixtools R package (1) was used to find parameter estimates for two normal distributions $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ based on the gene-level Pearson correlation coefficients for DNA copy number and gene expression in each tumor type (CN/GE ρ values), as shown in Supplementary Figure 1. This function applies the expectation-maximization (EM) algorithm (2) to obtain the final parameter estimates $\mu_1, \mu_2, \sigma_1, \sigma_2$ based on the observed data, initial estimates $\mu_{1,0}, \mu_{2,0}, \sigma_{1,0}, \sigma_{2,0}$ of the two means and standard deviations, and initial estimates $\pi_{1,0}, \pi_{2,0}$ of the mixture proportions. In order to distinguish the fitted densities we assume that $\mu_1 < \mu_2$.

The densities shown in Supplementary Figure 1 were computed using the following initial parameter estimates for each tumor type: $\mu_{1,0}$ = the 0.25 quantile of the CN/GE ρ values, $\mu_{2,0}$ = the 0.75 quantile of the CN/GE ρ values, $\sigma_{1,0} = \sigma_{2,0}$ = 0.5 * the standard deviation of the CN/GE ρ values, and $\pi_{1,0} = \pi_{2,0}$ = 0.5. We used the CN/GE ρ values in head and neck squamous cell carcinoma to conduct exploratory analyses to determine the sensitivity of the final parameter estimates to the initial parameter estimates. Changing $\mu_{1,0}$ and $\mu_{2,0}$ to the 0.15 and 0.85 quantiles of the CN/GE ρ values had essentially no effect on the parameter estimates because the values were identical to at least five decimal places, and the same results were found when 0.35 and 0.65 quantiles of the CN/GE ρ values. Changing $\sigma_{1,0}$ and $\sigma_{2,0}$ to either 0.25 * the standard deviation of the CN/GE ρ values or 0.75 * the standard deviation of the CN/GE ρ values for each combination of $\mu_{1,0}$ and $\mu_{2,0}$ yielded similar results. These findings suggest that the parameter estimates produced by the normalEM() function are highly robust to changes in the initial estimates $\mu_{1,0}, \mu_{2,0}, \sigma_{1,0}, \sigma_{2,0}$.

REFERENCES

1. Benaglia R, Chauveau D, Hunter DR, Young D. mixtools: an R package for analyzing finite mixture models. *J. Stat. Soft.* **2009**:32(6): 1 – 29.

2. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc., Ser. B* **1977**:39(1): 1 – 38.