S1 Text: Intra-tumour signalling entropy determines clinical outcome in breast and lung cancer

Christopher R. S. Banerji\textsuperscript{1,2,3,*}, Simone Severini\textsuperscript{2}, Carlos Caldas\textsuperscript{4}, Andrew E. Teschendorff\textsuperscript{1,5}

1. Statistical Cancer Genomics, Paul O’Gorman Building, UCL Cancer Institute, University College London, London WC1E 6BT, UK.
2. Department of Computer Science, University College London, London WC1E 6BT, UK.
3. Centre of Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London WC1E 6BT, UK.
5. CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai Institute for Biological Sciences, 320 Yue Yang Road, Shanghai 200031, China.

* Corresponding Authors: Christopher R.S. Banerji, e-mail: christopher.banerji.11@ucl.ac.uk and Andrew E. Teschendorff, e-mail: a.teschendorff@ucl.ac.uk

Supplementary Materials and Methods

Expression Data

For our initial survival studies in breast cancer and for identification of signalling entropy associations with clinical variables we considered data from 1980 primary breast cancer patients collected by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project [30]. The data consists of microarrays profiled on the Illumina HT 12v3 platform and log normalised as described in reference [30]. This expression data is accompanied by extensive clinical annotation describing tumour grade, size, stage, cellularity, histological subtype, lymph node status, hormone receptor status, Pfam50 subtype and p53 mutation status, as well as patient age at diagnosis, menopausal status, treatment and survival data. As described in [30] the expression data was divided into a discovery set of 997 samples and a validation set of 983 samples, each representing a distribution of clinical variables and molecular subtypes observed in the full data set. Eight further independent breast cancer data sets were located via Oncomine, the GEO database and ArrayExpress [67-69]. All studies considered both ER positive and ER negative breast cancers and were profiled on either Affymetrix or Illumina platforms [33-40]. With the exception of two, all studies were annotated with overall survival data, in the two remaining cases, relapse and distant metastasis were used as proxies for survival. Two studies also lacked ER immunohistochemistry, in these cases, a dip test was utilised to confirm that the distribution of \textit{ESR1} expression was significantly bimodal, before partitioning around medoids was utilised on \textit{ESR1} expression to generate two clusters dividing samples into ER positive and ER negative subgroups. The sample counts, stratified by ER status for all data sets used are provided in S1 Table, alongside GEO and ArrayExpress accession numbers. Normalised data from each study was downloaded from the GEO database or ArrayExpress. Quantile normalisation was subsequently performed across all samples within each study.

For our investigation into lung adenocarcinoma we first considered The Cancer Genome Atlas (TCGA) data set of RNAseq\textsuperscript{2} data profiling 455 tumours (http://cancergenome.nih.gov) and the Director’s Challenge microarray data set profiling 398 tumours, with accompanying clinical annotation. Processed data from both studies was quantile normalised and log transformed. The RNAseq\textsuperscript{2} data was skewed towards low values in a manner that inhibited the computation of signalling entropy, consequently, genes with read counts below
3 were removed from the data set.
Four further independent lung adenocarcinoma data sets with survival data were located via Oncomine, and the GEO database [42, 45, 46, 67, 68, 70]. Normalised data from each study was downloaded from the GEO database and quantile normalised within each study.
For each study separately probes in the microarrays and sequences in the RNAseqv2 data were matched to unique EntrezGene identifiers; probes or sequences mapping to the same identifier were averaged over.

Construction of the Protein Interaction Network
The Protein Interaction Network (PIN) used in this study was constructed as described in [16]. Briefly, we integrated interaction data obtained from Pathway Commons (www.pathwaycommons.org) [71] with the Human Protein Reference Database [72], the National Cancer Institute Pathway Interaction Database (NCI-PID) (pid.nci.nih.gov), the Interactome (Intact) (http://www.ebi.ac.uk/intact/) and the Molecular Interaction Database (MINT) (http://mint.bio.uniroma2.it/mint/). Connections in this network describe a wide array of protein interactions including protein complex formation, post-translational modifications and enzymatic reactions. Notably the PIN includes 20 highly curated immune and cancer signalling pathways from NetPath (www.netpath.org) [73]. Redundant interactions were removed from the PIN and only genes with an EntrezGene identifier annotation were retained. The PIN was subsequently sparsified by the consideration of the cellular localisation of interacting proteins to remove interactions likely to by false positives. The resulting maximally connected PIN contained 8,434 nodes and over 300,000 interactions.

Gene Set Enrichment Analysis
Gene Set Enrichment Analysis (GSEA) was implemented using a Fisher’s exact test to compare gene lists associated with signalling entropy’s prognostic power against the gene sets defined by the Molecular Signatures Database [50, 74]. Computations were performed via software downloaded from the Molecular Signatures Data Base (www.broadinstitute.org/msigdb) [50, 74].

Breast Cancer Transcriptomic Signatures
We compared the prognostic power of signalling entropy in breast cancer to two transcriptomic signatures, namely the Embryonic Stem (ES) cell based 100 gene signature of Ben-Porath et al. and the 97 gene grade signature of Sotiriou et al. These signatures were selected due to their similarity to signalling entropy in their clinical variable association in breast cancer. Like signalling entropy both signatures are prognostic and correlate with the level of differentiation of a tumour. The Ben-Porath et al. signature also correlates with pluripotency in healthy tissue and associates with ER status and molecular subtype, further properties of signalling entropy. Given this relatedness, it was important to assess the independence of these signatures from our measure.
All genes in the Ben-Porath et al. signature are positively correlated with tumour grade, and thus we computed a score for this signature as the mean log-normalised intensity of these genes in a sample.
The Sotiriou et al. signature contains genes both up-regulated and down-regulated in higher grade breast cancers compared to lower grade. We therefore computed a score for this signature as the statistic of a t-test evaluating the hypothesis that the log-normalised intensities of the up-regulated genes is higher than that of the down-regulated genes.

The Signalling Entropy Prognostic Score in Breast Cancer
For the breast cancer data the signalling entropy prognostic score (SE score) was computed from the METABRIC discovery data set from prognostic genes, which were correlated or anti-correlated with signalling entropy independently of grade and ER status, and whose prognostic power was also independent of grade and ER status. This gene set of 320 genes was refined by fitting a Cox proportional hazards model on 5 year censored data, using all the identified genes as covariates and deleting genes which were not significantly prognostic, independently of others in the gene set. This resulted in a small set of 81 genes, 10 of which were negatively correlated with signalling entropy and 71 of which were positively correlated S2 Table. A
Signalling Entropy prognostic score (SE score) was then defined as the \( t \)-statistic evaluating the hypothesis that the 71 positively correlated genes are expressed more highly than the 10 negatively correlated genes (after \( z \)-score normalising the data for each gene, across samples).

Before survival analysis the SE score was normalised by its standard deviation within each study.

The Signalling Entropy Prognostic Score in Lung Adenocarcinoma

For the lung adenocarcinoma data the SE score was computed analogously to the computation of the breast cancer SE score, using the Director’s Challenge data set as a discovery set. The only differences were that adjustment was made for tumour stage rather than grade and ER status and survival analysis was performed on 3 year censored data rather than 5 year. The basis of the score was a small set of 27 genes, 8 of which were negatively correlated with signalling entropy and 19 of which were positively correlated (S4 Table). The SE score was then computed as the statistic of the \( t \)-test evaluating the hypothesis that the genes negatively correlated with signalling entropy were less expressed than those positively correlated (after \( z \)-score normalising each gene, across samples).

Before survival analysis the SE score was normalised by its standard deviation within each study.

MammaPrint and CADM1 expression

A MammaPrint score was assigned to each sample in the 10 breast cancer datasets and was evaluated from the expression of the 70 genes required to define the signature as the \( t \)-statistic comparing the genes found to positively correlate with survival against those negatively correlated with survival in the study of [24]. Before survival analysis the MammaPrint score was normalised by its standard deviation within each study. CADM1 expression was found in all 6 lung adenocarcinoma data sets and was similarly normalised by its standard deviation within each data set before survival analysis.

OncotypeDX and Kratz et al. score approximations

OncotypeDX and the Kratz et al. scores are both derived from RT-PCR rather than from microarrays like MammaPrint and CADM1 expression. Consequently, the comparison of the SE score to these scores is not conclusive as we only consider microarray and RNAseq data, which though correlated, is normalised differently to RT-PCR data. Moreover some of the array platforms considered lack expression of certain genes required to compute the OncotypeDX and Kratz et al. scores.

We therefore approximate these scores from our data to allow a rough comparison. In the case of OncotypeDX, a score was computed from the expression of 21 genes (or as many as were represented in the data set) in each breast cancer microarray sample using the formula defined by reference [73]. In the case of the Kratz et al. score, this was similarly approximated from the expression of 14 genes (or as many as were represented in the data set) in each lung adenocarcinoma microarray or RNAseq sample, using the formula defined by reference [22].

Meta-analysis of prognostic scores

Meta-analysis was performed to combine survival statistics for signalling entropy, the SE score, MammaPrint, CADM1 expression and the approximations of OncotypeDX and the Kratz et al. score. Concordance indices were computed for each prognostic measure in each data set considered and combined using a random effects model. The \( p \)-values were combined using Fisher's combined test. Forest plots were generated using the \textit{survcomp} package in R [76].

Evaluation of random gene expression signatures

The 3 random gene expression signatures described by Venet et al. were obtained from the supplementary information provided in reference [32]. Following the analysis described by Venet et al., the probes in the METABRIC data sets mapping to genes in each random gene set were extracted, genes which were associated with cell cycling as described by Ben-Porath et al. [12] were removed. The expression values were then median polished and a principal component analysis was performed. Samples were then partitioned into two groups,
for each score via the median PC1 value. A Cox regression on 5 year censored data was then performed to evaluate the prognostic power of each random gene expression signature in each METABRIC data set. Only KRISHNAN2007DEFEAT was significantly prognostic in both METABRIC data sets, however partition of the datasets into ER+ and ER- samples mitigated this prognostic association.

Supplementary Results

Signalling entropy correlates with measures of tumour differentiation in breast cancer

Previously we demonstrated that signalling entropy correlated with differentiation potential in healthy tissue, was elevated in cancerous tissue and was higher in CSCs than in the tumour bulk [16]. We therefore anticipated that signalling entropy would correlate with transcriptomic and histological measures of the level of tumour differentiation.

We consider this hypothesis in the METABRIC breast cancer data sets. Two appropriate transcriptomic measures of tumour differentiation in breast cancer are the Ben-Porath 100 gene Embryonic Stem (ES) cell signature (found to associate with both breast cancer and lung adenocarcinoma tumour grade [12,13]) and the 97 gene Sotiriou grade signature [31].

As expected, we found that both signatures were strongly correlated with signalling entropy in the METABRIC data sets (Ben-Porath: \( p < 2.2\times 10^{-16} \), S1 Figure) indicating that the plastic signalling regime measured by signalling entropy is indeed associated with external, transcriptomic measures of the stemness of a tumour.

As anticipated, signalling entropy also strongly correlated with histological tumour grade in breast cancer, being significantly higher in grade 3 tumours compared with grade 2 and significantly higher in grade 2 tumours as compared with grade 1 (\( p < 7.3\times 10^{-15} \) and \( p < 2.6\times 10^{-4} \) respectively, S2A Figure). This result is consistent with the findings of Ben-Porath et al., which demonstrated that high grade tumours displayed enrichment of an ES cell transcriptomic signature [12]. Importantly, however, we found that unlike signalling entropy the Ben-Porath signature was unable to distinguish between grade 1 and grade 2 tumours in the discovery set of METABRIC (Ben-Porath signature: \( p = 0.4 \), signalling entropy: \( p < 2.6\times 10^{-4} \), S2B Figure). This result suggests that our measure is capable of discerning between tumours of varying levels of differentiation, and indeed is more sensitive to such variation than an ES cell transcriptomic signature. We also note that tumours with a high signalling entropy display a bi-modality of enrichment for the Ben-Porath ES cell signature. High ES cell enriched, high signalling entropy tumours are generally grade 3, whereas the low ES cell enriched, high signalling entropy tumours can be of lower grade (S1 Figure). This is indicative that a high signalling entropy is related to but not solely determined by the level of differentiation of a tumour, and that other factors, such as inter-cellular heterogeneity may cause high signalling entropy in certain samples.

We also found that signalling entropy correlated with tumour cellularity and was highest in those samples with the greatest proportion of cancerous cells (\( p < 7.7\times 10^{-15} \)). This result is consistent with our previous finding that cancer cells display a higher level of signalling promiscuity than their healthy counterparts [16,20].

Signalling Entropy correlates with levels of tumour differentiation in lung adenocarcinoma

We next considered whether signalling entropy also associated with transcriptomic and histological assessments of tumour differentiation in lung adenocarcinoma. The Ben-Porath signature was investigated by Hassan et al. in the context of lung adenocarcinoma and found to show association with histological grade and clinical outcome [13].

By considering the Director’s Challenge data set we found that, as in breast cancer, signalling entropy correlated strongly with the Ben-Porath stem cell signature in lung adenocarcinoma (\( p < 2.2\times 10^{-16} \), S5A Figure). Moreover, we found that signalling entropy correlated strongly with histological assessments of tumour differentiation in lung adenocarcinoma, being highest in poorly differentiated tumours, then moderately differentiated tumours (\( p < 1.4\times 10^{-4} \) poorly vs. moderately differentiated) and lowest in well differentiated tumours (\( p < 2.5\times 10^{-5} \) moderately vs. well differentiated, S5B Figure). These results further support the notion that signalling entropy is a measure of the stemness of a tumour.
Signalling entropy associates with breast cancer heterogeneity and luminal B breast cancer displays the highest signalling entropy, yet among the weakest enrichment for ES cell genes

We next examined whether signalling entropy was associated with breast cancer subtypes to ascertain whether a stem cell like signalling regime, or inter-cellular diversity could be related to breast cancer heterogeneity as previously suggested [12,14].

We divided the samples in both the discovery and validation sets of METABRIC by their Pfam50 intrinsic subtype classifications [77] and found that signalling entropy was strongly associated with molecular subtype. Normal tumours displayed the lowest signalling entropy, followed by luminal A tumours (luminal A vs normal $p < 2.5e-8$), then HER2 tumours (HER2 vs luminal A $p < 1.9e-12$), and lastly luminal B and basal tumours (luminal B vs HER2 $p < 0.02$ basal vs luminal B $p = 0.3$), which displayed statistically equivalent signalling entropies. This result is in concordance with that of Ben-Porath et al. [12], who demonstrated that basal breast cancers displayed a stronger enrichment of an ES cell transcriptomic signature than luminal A breast cancer.

To more fully compare the Ben-Porath signature with signalling entropy, we examined its ability to discriminate between the intrinsic subtypes of breast cancer. As reported in the presenting paper, the signature is most enriched in the basal breast cancer subtype and least in luminal A. However the overall ordering of the intrinsic subtypes by Ben-Porath signature enrichment is quite dissimilar from the ordering by signalling entropy. Notably luminal B breast cancers are significantly less enriched for the Ben-Porath signature than basal ($p < 2.2e-16$), however, signalling entropy suggests that these subtypes share a highly promiscuous signalling regime. This result is intriguing as it implies that the severity of luminal B breast cancer may be linked to a plastic signalling regime, rather than the enrichment of genes typically over-expressed in ES cells. One may conclude from this that the high signalling entropy of the luminal B subtype may be driven by inter-cellular heterogeneity, rather than high intra-cellular signalling promiscuity.

Criticism of molecular subtyping has derived from the lack of diversity in the histological subtype of tumours used to define the classifications [78]. Consequently, we also examined the association between signalling entropy and histological subtype. This revealed that medullary carcinomas have the highest signalling entropy, consistent with these cancers generally being of higher grade and displaying a basal phenotype [79]. Invasive ductal carcinomas of no special type (IDCs-NST) held the second highest signalling entropy, significantly higher than invasive lobular carcinomas ($p < 2.3e-5$) mixed ductal and lobular carcinomas ($p < 1e-3$) and tubular carcinomas, which all displayed a statistically identical signalling entropy. This result suggests that the different histological subtypes display significant differences in the promiscuity of their signalling regimes, a result which may reflect differences in the potency of their cell of origin, or in their degree of de-differentiation from a common originator.

We also found that signalling entropy could discriminate grade matched ER positive tumours from ER negative tumours ($p < 0.01$). Given the finding by Ben-Porath et al. that ER negative tumours were enriched for an ES cell signature [12], this result suggests that a stem cell-like, promiscuous signalling regime is more prevalent in ER negative tumours, regardless of histological grade.

Tumours carrying a mutation in p53 have also been demonstrated to be enriched for an ES cell gene expression signature [80]. In line with this result we found that grade matched p53 mutated tumours displayed a higher signalling entropy as compared their wild type counterparts ($p = 0.005$). This result is consistent with the hypothesis that p53 mutations in breast cancer can facilitate de-differentiation [80].

Signalling Entropy associates with smoking history and tumour stage in lung adenocarcinoma

We next investigated whether exposure to cigarette smoke was associated with a plastic signalling regime in lung adenocarcinoma. Hassan et al., previously reported that lung adenocarcinoma patients with a positive smoking history displayed an increased expression of stem cell genes. In line with these findings, we found a strong association between signalling entropy and smoking history in lung adenocarcinoma (TCGA adenocarcinoma: never smoked vs. currently smoking $p = 1.4e-7$, Director’s Challenge adenocarcinoma: never smoked vs. currently smoking $p < 0.03$, S6 Figure). These results suggest that cigarette smoking
induces an increased signalling promiscuity in lung adenocarcinoma. We also found that signalling entropy significantly correlated with tumour stage in both the TCGA and Director’s Challenge lung adenocarcinoma data sets (p < 4.8e − 8), suggesting that signalling entropy associates with current leading assessments of prognosis.

**Signalling entropy is elevated in heterogeneous samples on average**

**Preliminaries**

Let $G = (V, E)$ be an undirected graph, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and $E = \{(i, j)|i, j \in V\}$ a set of edges; we denote the adjacency matrix of $G$ by $A = (a_{ij})_{i,j \in V}$. In our analysis $G$ represents the undirected topology of the interactome.

To each vertex $i \in V$ we assign a variable $x_i \in \mathbb{R}_{>0}$, and denote the vector containing all such variables by $x = (x_i)_{i=1}^n \in \Omega \subset \mathbb{R}_{>0}$, where $\Omega$ is some bounded domain. In our analysis $x$ will represent the vector of log normalised gene expression values for a homogeneous sample, we note that as the expression of genes cannot be infinite we bound $x$ within a finite domain $\Omega$.

We consider a random walk, on the graph $G$; with transition probability matrix $P(x) = (p_{ij}(x))_{i,j \in V}$ defined via

$$
p_{ij}(x) = \frac{a_{ij}x_j}{\sum_{k \in V} a_{ik}x_k}.
$$

We define the following measures

1. The local entropy of vertex $i \in V$, defined by

$$
S_i(x) := -\sum_{j \in V} p_{ij}(x) \log p_{ij}(x).
$$

2. The entropy rate of $P(x)$, defined by

$$
S_R(x) := \sum_{i \in V} \pi_i(x) S_i(x).
$$

where $\pi_i$ denotes the stationary distribution of $P(x)$ and satisfies

$$
\pi_j(x) = \sum_{i \in V} p_{ij}(x) \pi_i(x).
$$

We will refer to $S_R(x)$ as the signalling entropy of $x$, as it is equivalent up to a normalisation factor.

Let us also define the following functions

$$
W_{ij}(x) := a_{ij}x_i x_j
$$

$$
W_i(x) := \sum_j W_{ij}(x)
$$

$$
W(x) := \sum_i W_i(x).
$$

We also note that by multiplying $p_{ij}(x)$ by $1 = x_i/x_i$, we see that

$$
p_{ij} = \frac{W_{ij}(x)}{W_i(x)}
$$

Thus $P(x)$ describes a weighted random walk on an undirected graph and it thus follows that

$$
\pi_i(x) = W_i(x)/W(x).
$$
The proof follows from simple substitution into (3):

\[ LHS = \sum_i p_{ij}(x)\pi_i(x) \]
\[ = \sum_i \frac{W_{ij}(x)W_i(x)}{W(x)} \]
\[ = \frac{W_j(x)}{W(x)} \]
\[ = RHS. \]

From this result it follows that

\[ S_R(x) = -\frac{1}{W(x)} \sum_{ij} W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)} \quad (4) \]

**Motivation**

In this section we wish to demonstrate that the signalling entropy of a heterogeneous sample generated from a 50:50 mixture of two cell types is greater, on average, than the signalling entropy of a homogeneous sample. This amounts to proving the following proposition:

**Proposition.** Let \( x, y \in \Omega \), then

\[ \int_{\Omega} \int_{\Omega} \left( S_R\left( \frac{x+y}{2} \right) - S_R(x) \right) dxdy > 0. \quad (5) \]

Let us consider the following claim

**Claim (Super-additivity).** Let \( x, y \in \Omega \) then

\[ S_R\left( \frac{x+y}{2} \right) > \frac{S_R(x)}{2} + \frac{S_R(y)}{2}. \quad (6) \]

It is clear that if the claim is true then the proposition must be true. Notice first that if the claim is true then as it is a strict bound \( \exists \epsilon > 0 \) such that \( S_R\left( \frac{x+y}{2} \right) > \frac{S_R(x)}{2} + \frac{S_R(y)}{2} + \epsilon \). Whence

\[ \int_{\Omega} \int_{\Omega} \left( S_R\left( \frac{x+y}{2} \right) - S_R(x) \right) dxdy > \int_{\Omega} \int_{\Omega} \left( \frac{S_R(y)}{2} - \frac{S_R(x)}{2} + \epsilon \right) dxdy \]
\[ = |\Omega|^2 \epsilon \]
\[ > 0, \]

and thus the proposition is true.

We will therefore derive in a sufficient condition for the claim to be true. We will then demonstrate numerically that this condition holds for over 528 distinct in silico mixtures of homogeneous healthy differentiated tissues.

The results in this section thus provide evidence that signalling entropy is raised in heterogeneous biological samples comprising of a 50:50 mix of two cell types, as compared to homogeneous samples, at the population level.

**A sufficient condition for the claim to be true**

Here we will prove the following theorem

**Theorem.** Let \( x, y \in \Omega \), let \( a = \max_i \left( \frac{x_i}{y_i} \right) \), and let \( b = \min_i \left( \frac{x_i}{y_i} \right) \). A sufficient condition for the claim above to be true is

\[ \text{sign}(1 - 1/b + 2/a) + \text{sign}(1 - a + 2b) = 2 \quad (10) \]
**Proof:** From (4)

\[ S_R((x+y)/2) = -\frac{1}{W((x+y)/2)} \sum_{ij} W_{ij}((x+y)/2) \log \frac{W_{ij}((x+y)/2)}{W_i((x+y)/2)} , \]  

(11)

it is therefore prudent to first consider \(W_{ij}((x+y)/2)\):

\[ W_{ij}((x+y)/2) = \frac{a_{ij}}{4} (x_i + y_i)(x_j + y_j) \]  

(12)

\[ = \frac{1}{4}(W_{ij}(x) + W_{ij}(y) + a_{ij}(x_iy_j + x_jy_i)). \]  

(13)

We will define

\[ \hat{W}_{ij}(x,y) = a_{ij}(x_iy_j + x_jy_i) \]  

(14)

\[ \hat{W}_i(x,y) = \sum_j \hat{W}_{ij}(x,y) \]  

(15)

\[ W(x,y) = \sum_j \hat{W}_i(x,y) \]  

(16)

for notational ease. Note that as \(y, x > 0\)

\[ \hat{W}_{ij}(x,y) = \frac{x_i}{y_i}W_{ij}(y) + \frac{y_i}{x_i}W_{ij}(x). \]  

(17)

\[ \hat{W}_i(x,y) = \frac{x_i}{y_i}W_i(y) + \frac{y_i}{x_i}W_i(x). \]  

(18)

(19)

It thus follows that:

\[ S_R((x+y)/2) = -\frac{1}{W(x) + W(y) + W(x,y)} \sum_{ij} (W_{ij}(x) + W_{ij}(y) + \hat{W}_{ij}(x,y)) \log \frac{W_{ij}(x) + W_{ij}(y) + \hat{W}_{ij}(x,y)}{W_i(x) + W_i(y) + \hat{W}_i(x,y)} . \]  

(20)

We will now appeal to the log sum inequality:

**Theorem** (Log-sum inequality). Let \(a_1, ..., a_n, b_1, ..., b_n\) be non-negative numbers then

\[ \sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} . \]  

(21)

If we denote \(a_1 = W_{ij}(x), a_2 = W_{ij}(y), a_3 = \hat{W}_{ij}(x,y)\) and \(b_1 = W_i(x), b_2 = W_i(y), b_3 = \hat{W}_i(x,y)\), and apply the log-sum inequality to the summand of (22) we obtain:

\[ S_R((x+y)/2) \geq -\frac{1}{W(x) + W(y) + W(x,y)} \sum_{ij} W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)} + \]  

(22)

\[ W_{ij}(y) \log \frac{W_{ij}(y)}{W_i(y)} + \hat{W}_{ij}(x,y) \log \frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)} \]  

(23)

\[ = \frac{W(x)S_R(x) + W(y)S_R(y) - \sum_{ij} \hat{W}_{ij}(x,y) \log \frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)}}{W(x) + W(y) + W(x,y)} . \]  

(24)

Now consider the term \(-\sum_{ij} \hat{W}_{ij}(x,y) \log \frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)}\) and apply the log sum inequality again:

\[ -\sum_{ij} \hat{W}_{ij}(x,y) \log \frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)} = -\sum_{ij} \left( \frac{x_i}{y_i}W_{ij}(y) + \frac{y_i}{x_i}W_{ij}(x) \right) \log \frac{x_i}{y_i}W_{ij}(y) + \frac{y_i}{x_i}W_{ij}(x) \]  

\[ \geq -\sum_{ij} \frac{x_i}{y_i}W_{ij}(y) \log \frac{W_{ij}(y)}{W_i(y)} + \frac{y_i}{x_i}W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)} . \]  

(25)
We notice that

\[ \frac{1}{2(W(x) + W(y) + W(x,y))} \left( S_R(x)(W(x) - W(y) - \hat{W}(xy)) + S_R(x)(W(y) - W(x) - \hat{W}(xy)) - 2 \sum \frac{x_i W_{ij}(y)}{y_i} \log \frac{W_{ij}(y)}{W_i(y)} + \frac{y_i W_{ij}(x)}{x_i} \log \frac{W_{ij}(x)}{W_i(x)} \right). \]  

Returning to (22), it now follows that

\[ \frac{1}{2(W(x) + W(y) + W(x,y))} \left( S_R(x)(W(x) - W(y) - \hat{W}(xy)) + S_R(x)(W(y) - W(x) - \hat{W}(xy)) - 2 \sum \frac{x_i W_{ij}(y)}{y_i} \log \frac{W_{ij}(y)}{W_i(y)} + \frac{y_i W_{ij}(x)}{x_i} \log \frac{W_{ij}(x)}{W_i(x)} \right). \]  

So the claim is true if

\[ \begin{align*} 
0 < \frac{1}{2(W(x) + W(y) + W(x,y))} & \left( S_R(x)(W(x) - W(y) - \hat{W}(xy)) + S_R(x)(W(y) - W(x) - \hat{W}(xy)) - 2 \sum \frac{x_i W_{ij}(y)}{y_i} \log \frac{W_{ij}(y)}{W_i(y)} + \frac{y_i W_{ij}(x)}{x_i} \log \frac{W_{ij}(x)}{W_i(x)} \right) \\
& > (S_R(x) + S_R(y))(W(x) + W(y) + \hat{W}(xy)) 
\end{align*} \]  

We note that if (27) holds then

\[ 2 \left( W(x)S_R(x) + W(y)S_R(y) - \sum \frac{x_i W_{ij}(y)}{y_i} \log \frac{W_{ij}(y)}{W_i(y)} + \frac{y_i W_{ij}(x)}{x_i} \log \frac{W_{ij}(x)}{W_i(x)} \right) > (S_R(x) + S_R(y))(W(x) + W(y) + \hat{W}(xy)) \]  

We note that RHS of (28) satisfies

\[ 2(W(x)S_R(x) + W(y)S_R(y) - \sum \frac{x_i W_{ij}(y)}{y_i} \log \frac{W_{ij}(y)}{W_i(y)} + \frac{y_i W_{ij}(x)}{x_i} \log \frac{W_{ij}(x)}{W_i(x)} > 2W(x)S_R(x)(1 + \min \frac{y_i}{x_i}) + 2W(y)S_R(y)(1 + \min \frac{x_i}{y_i}) \]  

We notice that

\[ \hat{W}(xy) = \sum \frac{x_i W_i(y) + y_i W_i(x)}{x_i y_i} < \max \frac{x_i}{y_i} W(y) + \max \frac{y_i}{x_i} W(x) \]  

whence the LHS of (28) satisfies

\[ (S_R(x) + S_R(y))(W(x) + W(y) + \hat{W}(xy)) < S_R(x)W(x)(1 + \max \frac{y_i}{x_i}) + S_R(y)W(y)(1 + \max \frac{x_i}{y_i}). \]  

Whence it follows that if

\[ S_R(x)W(x)(1 + \max \frac{y_i}{x_i}) + S_R(y)W(y)(1 + \max \frac{x_i}{y_i}) < 2W(x)S_R(x)(1 + \min \frac{y_i}{x_i}) + 2W(y)S_R(y)(1 + \min \frac{x_i}{y_i}) \]  

\[ S_R(x)W(x)(1 - \max \frac{y_i}{x_i} + 2 \min \frac{y_i}{x_i}) + S_R(y)W(y)(1 - \max \frac{x_i}{y_i} + 2 \min \frac{x_i}{y_i}) > 0 \]  

\[ S_R(x)W(x)(1 - 1/b + 2/a) + S_R(y)W(y)(1 - a + 2b) > 0 \]
then \( LHS < RHS \). Where
\[
a = \max_i \left( \frac{x_i}{y_i} \right),
\]
and
\[
b = \min_i \left( \frac{x_i}{y_i} \right),
\]
as in the claim. We note that as \( S_R(x)W(x) > 0 \) and \( S_R(y)W(y) > 0 \) than the condition will always hold provided
\[
\text{sign}(1 - 1/b + 2/a) + \text{sign}(1 - a + 2b) = 2.
\]
Hence the theorem is correct. \( \square \)

We note that the condition can be computed numerically for a range of values and holds for the majority of biologically plausible ranges (S8 Figure).

**Empirical validation that the claim is true on \( \Omega \)**

If the condition derived above holds in \( \Omega \), the space of biologically admissible homogeneous sample expression regimes, then the claim and hence the proposition explained at the start of this section are true and we have proven our postulate.

To investigate whether this is the case we consider the data set described in GSE2361 [29], which profiles 33 distinct adult tissues and 3 foetal tissues. We will disregard the foetal tissues as the tissue types overlap with adults tissues and thus cannot be considered distinct homogeneous samples. We first computed the value of \( \text{sign}(1 - 1/b + 2/a) + \text{sign}(1 - a + 2b) \) for every pairwise combination of the 33 tissue types (528 pairwise combinations). We found that for every combination the condition \( \text{sign}(1 - 1/b + 2/a) + \text{sign}(1 - a + 2b) = 2 \) was satisfied and thus the claim was correct for this data set (S9 Figure). While it is certainly true that not every tissue, when mixed with another, will display an increased signalling entropy, we do see a trend towards this happening, with the reproductive tissues (testes, uterus, ovaries, breast) providing the most significant increases (S10 Figure).

Finally, as expected, the proposition is indeed correct for this data set, and the signalling entropy of the mixed samples is higher than the homogeneous samples on average; this increase is also significant \( p = 0.012 \) (paired Wilcoxon test) (S11 Figure).

**Summary**

To conclude, we have derived a condition which if satisfied by homogeneous samples guarantees that signalling entropy will be higher in heterogeneous 50:50 mixtures of two homogeneous samples, as compared to the unmixed samples, on average. We then verified that for a large number of homogeneous adult tissues the condition was indeed satisfied and signalling entropy was higher in 50:50 mixtures of homogeneous tissues on average.

These result suggest signalling entropy represents a quantifier of intra-sample sample heterogeneity, being correlated with heterogeneity at the population level.

**References**


