**SUPPLEMENTARY MATERIAL**

**Validation Examples**

***Contribution of the C3 Comparison in the Two-Stage Process for CTC Hypothesis***

The purpose and contribution of C3 is justified based on the consideration of tissue-specific differences in the expression of genes, which have been established in biological studies even for same-condition (homogeneous) population (variability of expression within and across populations). Essentially, the inclusion of C3 aims to alleviate cross-tissue differences appearing in the base populations of control breast and control blood engaged in our study. The following examples are intended to clarify the contribution of C3 in proposed methodology for circulating tumor cell (CTC) hypothesis.

**Examples:**

In an attempt to assess the performance of these three genes, we consider the publically available database GENT (Gene Expression across Normal and Tumor tissue)[[1]](#footnote-1), which presents the variation of published gene expression profiles in datasets of different human tissue types (tumor and normal) in more than 40000 samples measured on the Affymetrix U133A or U133plus2 platforms. We can indeed validate that these three genes show large variability across breast and blood samples and high variation across blood samples.

In order to further check the performance of these 3 genes in an independent test case, we focus on control PB [GSE12517] and breast-tissue populations from two recently published datasets [GSE29431, GSE42568]. We isolate the control populations of the two studies and through pre-processing we exclude samples with extreme outlier expression of the 3 genes of interest; namely we excluded 4 tissue samples from GSE42568 and 2 tissues samples from GSE29431. Overall, the test sample is composed of 23 blood samples from GSE12517, 13 breast samples from GSE42568 and 10 breast samples from GSE29431. Then, we merge the 3 datasets following the same methodology as in the previous comparisons and then examine the distribution of these 3 genes in the individual datasets, including all control samples.

*SNRPF* follows a consistent pattern expressing higher values in blood than in tissue with higher variance. *HNRNPU* shows similar mean values in the blood and tissue datasets, but the variance in blood is significantly higher than that in either tissue dataset. Finally, *PRKAR1A* shows quite different performance in the two tissue sets; one expresses significantly higher values than the other, so that in one case the mean of blood is higher than that of tissue and in the other case the order is reversed. In both cases, though, the variance in blood is much higher than that in tissue. Thus, if this gene is used in the C3 comparison with the blood samples as base, the SAM metric could decrease in comparison to its counterpart in C2 (with the tissue samples as base), owing to the increase of variance in the former. This differential performance in the base populations can then explain the exclusion of the gene in C3 despite its inclusion in C2.

|  |  |  |  |
| --- | --- | --- | --- |
| **GEO Datasets** | **Genes** | **Mean** | **Variance** |
| GSE125171 | HNRNPU | 7.594 | 0.79246 |
| GSE125171 | PRKAR1A | 8.435 | 0.86347 |
| GSE125171 | SNRPF | 7.872 | 0.33797 |
| GSE294312 | HNRNPU | 8.308 | 0.07495 |
| GSE294312 | PRKAR1A | 7.973 | 0.08526 |
| GSE294312 | SNRPF | 7.183 | 0.11145 |
| GSE425683 | HNRNPU | 7.256 | 0.13979 |
| GSE425683 | PRKAR1A | 10.080 | 0.20124 |
| GSE425683 | SNRPF | 7.131 | 0.17067 |

**Mean and Variance for the three genes of C1∩C2∩ intersection:**

**HNRNPU, PRKAR1A and SNRPF from three independent Datasets**

**1**GEO DATASET GSE12517**:** Gene expression profiling on peripheral blood mononuclear cells of 45 adult females. For the validation, we included 11 pre-menopausal, and 12 post-menopausal females. **2**GEO DATASETGSE29431**:** Gene expression profiling on 54 primary breast carcinomas and 12 samples of breast normal tissues from breast cancer patients. **3**GEO DATASETGSE42568**:** Gene expression profiling on 104 cases of primary breast cancer and 17 normal breast tissues. For validation purposes, we included only the control tissues from both Datasets GSE29431 and GSE42568 (12 and 17 samples, respectively).

1. Shin G, Kang TW, Yang S, Baek SJ, Jeong YS, Kim SY. GENT: gene expression database of normal and tumor tissues. *Cancer Inform*. 2011, vol. 10, pp. 149-157. [↑](#footnote-ref-1)