Supplementary material 1. Estimation of relative cellular abundances in pooled tumor samples using multi-region information.

A. Cell-abundance estimates in tumor samples: definitions and key relationships

Tumors represent a true ecosystem of cell populations where neoplastic and non-neoplastic cells coexist in the same environment. Furthermore, neoplasia itself is defined as a process of abnormal cellular reproduction which results in excessive tissue growth. This continuous multiplication of asexual organisms (i.e. tumor cells) in a resource-constrained environment such as the human body, inevitably creates a situation where cells in the tumor niche compete for nutrients through adaptation; providing the ideal substrate for selection to act. This ongoing evolutionary process has been extensively documented in cancer, and can ultimately lead to the existence of tumor subclones that are fundamentally different from each other. The co-existence of multiple, genetically distinct clones is known as intratumoral heterogeneity (ITH), and its existence can be traced from the genome to the protein products of most cancer types.

The degree of ITH observed in the cancer genomes can be evidenced, at least in part, by the presence of distinct somatic mutations in divergent tumor subclones. Thus, with regards to a somatic mutation of interest, the totality of cells (*C*) in a given tumor tissue sample *i* can be grouped into: normal/stromal cells (randomly intermixed with the tumor), mutant tumor cells (bearing the variant of interest) and wild-type (WT) tumor cells (without the variant).

$$C_i = Normal\ cells + mutant\ tumor\ cells + WT\ tumor\ cells$$

In order to assess the degree of ITH in the tumor, we would need to quantify the frequency of somatic mutations only in the tumor compartment. Therefore, consideration of the tumor/normal relative abundance is needed. Next-generation sequencing (NGS) allows assessment of both parameters with a relatively high degree of accuracy. First, the relative amount of tumor tissue in the sample (i.e. purity, ρ) can be obtained using copy-number algorithms, which rely on single nucleotide polymorphism (SNP) count data and previously-known karyotype frequencies to estimate this parameter.

$$\rho_i = \frac{tumor\ cells_i}{C_i}$$

Secondly, variant allelic frequencies (VAF) obtained from mutation-calling algorithms provide us an estimate of the proportion of mutant DNA in the sample. By combining these parameters with the locus-specific ploidy estimates (also from CN algorithms), investigators have been able to estimate the fraction of tumor cells bearing a particular mutation (*m*). This parameter is known as cancer-cell fraction (CCF) and represents a relative measure of cell abundance (i.e. relative to the subset of tumor cells):

$$CCF_i = \frac{mutant\ tumor\ cells_i}{tumor\ cells_i}$$

By defining these measures in probability terms, we can say that given a random cell drawn from the sample, ρ represents the probability of that cell being neoplastic. However, due to the relative nature of CCF, it must be understood as the conditional probability of mutation presence in a tumor cell from the pool:

$$\rho = P(tumor\ cell)$$

$$CCF = P(mutant\ tumor\ cell \mid tumor\ cell)$$

Finally, if we wanted to estimate the probability of randomly drawing a 'mutant tumor cell' from this sample we would need to consider the above two parameters:

$$P(mutant\ tumor\ cell) = P(tumor\ cell \cap mutant\ tumor\ cell)$$

$$P(mutant\ tumor\ cell) = P(tumor\ cell) \times P(mutant\ tumor\ cell \mid tumor\ cell)$$

By replacing these probabilities for the previously-defined parameters, we can conclude that the mutant cell fraction (MCF) in a sample *i* is a function of its purity and the mutation CCF:

$$MCF_i = \rho_i \times CCF_i$$

These relationships can also be confirmed by assessing the cellular abundances that these parameters represent:

$$MCF_i = \frac{\cancel{tumor\ cells_i}}{C_i} \times \frac{mutant\ tumor\ cells_i}{\cancel{tumor\ cells_i}} = \frac{mutant\ tumor\ cells_i}{C_i}$$

Therefore, if we had knowledge about the total number of cells in a sample we could easily find the number of tumor and mutant cells as follows:

$$tumor\ cells_i = C_i \times \rho_i$$

$$mutant\ tumor\ cells_i = C_i \times \rho_i \times CCF_i$$

B. Cell-abundance inference in multiregional tumor pools using single-region data.

Due to the significant degree of ITH in many cancer types, investigators must rely on the assessment of spatially-distinct regions to capture the full spectrum of aberrations in a tumor. However, multi-regional profiling comes with a significant increase in cost, limiting its use outside of very specific settings. Nevertheless, we could extend the previously-described concepts to infer the cellular composition of more complex cell mixtures (such as a those represented by multiregional DNA tumor pools), providing a cost-effective alternative to ascertain the relative frequency of somatic mutations in spatially-representative tumor mixes. In this scenario, mutation CCF estimates constitute a particularly useful clonality surrogate. Their calculation using NGS data takes into account purity, ploidy and the somatic variants' allelic frequencies, thus providing a normalized estimate which is expressed relative to the totality of cancer cells contributing to the pool.

The cellular composition of a pool of cancer cells can be expressed as:

$$C_{pool} = C_1 + C_2 + C_3 \ldots + C_i = \sum_{i=1}^{N} C_i$$

Similarly, the number of tumor cells and mutant tumor cells in the pool can be expressed as:

$$tumor\ cells_{pool} = \sum_{i=1}^{N} C_i \times \rho_i$$

$$mutant\ tumor\ cells_{pool} = \sum_{i=1}^{N} C_i \times \rho_i \times CCF_i$$

Given a tumor with distinct somatic mutations in spatially-separated regions, we can obtain *n* tissue samples, each composed of a certain number of cells. Assuming that each sample included in the pool contributed the same number of cells $(C_i = C)$ [a], the total amount of cells in the pool and in each sample will be:

$$C_{pool} = C_i \times n$$

$$C_i = \frac{C_{pool}}{n} = \frac{1}{n}$$

[a] *equal cellular contents from distinct regions are equivalent to mixing in equal proportions the total DNA extracted from different tumor regions*

We could also express it as the probability that a random cell drawn from the pool originated from a given region, which would be inversely related to the total number in the mix:

$$P(cell\ from\ region\ i) = \frac{1}{n}$$

Therefore, the probability of drawing a tumor cell from region *i* in this pool of *n* regions, will be a function of the mixture itself as well as the amount of tumor cells in the region (i.e. its purity). In this context, it represents the conditional probability of drawing a tumor cell from the subset appertaining to that region: [b]

$$P(tumor\ cell\ )_i = \ \ P(cell\ from\ region\ i\ \cap\ tumor\ cell\ \ )$$

$$P(tumor\ cell)_i = \ \ P(cell\ from\ region\ i)\ \times P(tumor\ cell\ |\ cell\ from\ region\ i)$$

$$P(tumor\ cell)_i = \frac{1}{n} \times\ \rho_i = \frac{\rho_i}{n}$$

Since cells in the pool can only originate from one tumor region (intersection probability = 0), in order to estimate the probability of finding any tumor cell in the pool we would need to consider the union of the probabilities in each region and conclude that the purity of the pool will be a near-average of the individual purities:

$$P(tumor\ cell)_{pool} = P(tumor\ cell)_1 + \ P(tumor\ cell)_2 \ldots + \ P(tumor\ cell)_i$$

$$P(tumor\ cell)_{pool} = \sum_{i=1}^{N} P(tumor\ cell)_i = \rho_{pool}$$

$$\rho_{pool} = \sum_{i=1}^{N} \frac{\rho_i}{n}$$

[b] *these derivations also apply to sample material mixed in different proportions. The relative contribution of each sample must be known in advance (i.e. different volumes and/or DNA concentrations).*

We could follow the same reasoning to estimate the probability of finding a ‘mutant tumor cell’ in the pool, regardless of its tumor/normal status $(i.e. MCF_{pool})$:

$$P(mutant\ tumor\ cell)_{pool} = \sum_{i=1}^{N} P(mutant\ tumor\ cell)_{i} = MCF_{pool}$$

$$P(mutant\ tumor\ cell)_{pool} = \sum_{i=1}^{N} P(cell\ from\ region\ i\ \cap mutant\ tumor\ cell)$$

$$MCF_{pool} = \sum_{i=1}^{N} P(cell\ from\ region\ i)\ \times P(mutant\ tumor\ cell\ |\ cell\ from\ region\ i)$$

$$MCF_{pool} = \sum_{i=1}^{N} \frac{MCF_i}{n} = \sum_{i=1}^{N} \frac{CCF_i\ \times \rho_i}{n}$$

Finally, since the CCF of a given mutation represents the conditional probability of a cell being both neoplastic and mutant, we can conclude:

$$P(mutant\ tumor\ cell\ |\ tumor\ cell) = \frac{P(mutant\ tumor\ cell)}{P(tumor\ cell)}$$

$$CCF_{pool} = \frac{MCF_{pool}}{\rho_{pool}}$$

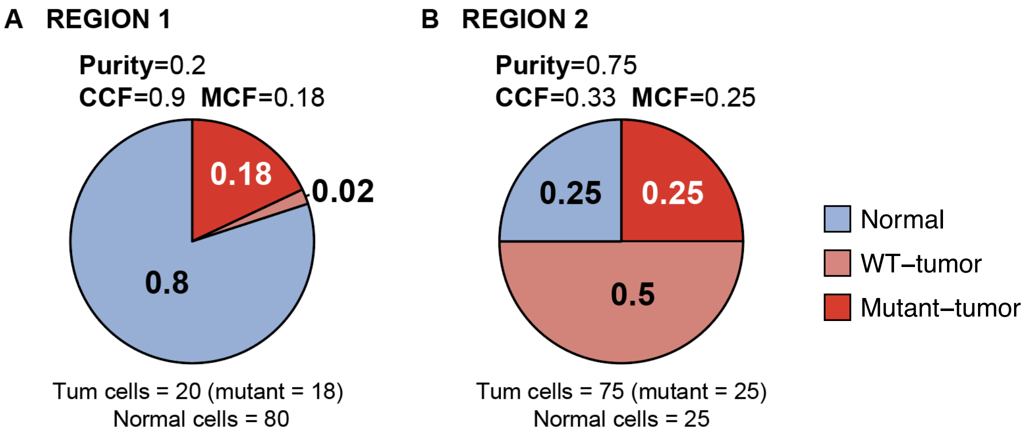
$$CCF_{pool} = \frac{\sum_{i=1}^{N} \frac{CCF_i\ \times \rho_i}{\cancel{n}}}{\sum_{i=1}^{N} \frac{\rho_i}{\cancel{n}}}$$

$$CCF_{pool} = \frac{\sum_{i=1}^{N} CCF_i\ \times \rho_i}{\sum_{i=1}^{N} \rho_i}$$

C. Working example

To illustrate these concepts we've provided an example below. The pie charts represent two tumor regions that represent hypothetical tissue samples. Both are represented by a group of tumor (red) and normal (blue) cells. A given subclonal mutation of interest may be differentially represented in the two samples (dark red).

In one of the regions (A) the mutation is seen in 90% of the tumor cells (CCF=0.9), while on the second one (B), only a third of the tumor cells are mutant (CCF=0.33). It is clear that depending on the sample examined we would have reached radically different conclusions regarding the clonality of the mutation. Region 1 suggests it is early and present clonally throughout the tumor, while region 2 shows the variant present at a relatively low frequency (in about a third of the cancer cells).



To illustrate our pooling approach, we are going to simulate a mixture of the cells composing both samples. To simplify calculations while assuming a mixture in equal proportions, we assigned an equal amount of total cells to each sample (n=100 each); the number of cells in each category is also shown.

We can now use our previously-described formulations to infer the cellular composition of this hypothetical mixture containing all the cells sampled (n=200). Given that purity and MCF are parameters whose definitions are relative to the totality of cells, when combining both cell repertoires, these values are effectively averaged:

$$\rho_{pool} = \frac{\sum_{i=1}^{N} \rho_i}{n} = \frac{0.2 + 0.75}{2} = \mathbf{0.475}$$

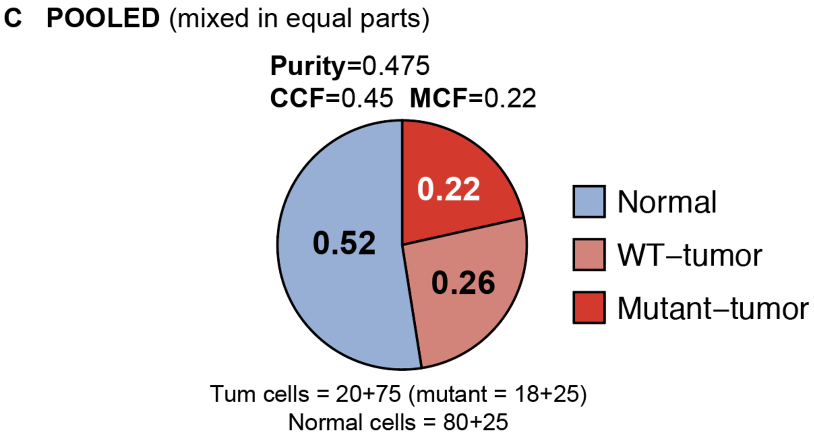
$$MCF_{pool} = \frac{\sum_{i=1}^{N} MCF_i}{n} = \frac{0.18 + 0.25}{2} = \mathbf{0.215}$$

However, the same is not true for CCF, which is a measure expressed relative to the totality of tumor cells. Neoplastic cells represent a subset of the sample that we cannot easily separated from the rest of the tissue. For these reasons, we cannot simple average these values ($\overline{CCF}$ = 1.07, or 107% in this case) and need to account for the purity in each region to be able to estimate it:

$$CCF_{pool} = \frac{\sum_{i=1}^{N} CCF_i \times \rho_i}{\sum_{i=1}^{N} \rho_i}$$

$$CCF_{pool} = \frac{(0.9 \times 0.2) + (0.33 \times 0.75)}{0.2 + 0.75} = \mathbf{0.45}$$

Panel C shows our final multiregional pool, which is composed of the sum of cells from all regions (A+B = 200, at the bottom). The final cell quantities are listed on the bottom. We can finally confirm that our calculations were correct by going back to our previous definitions:



$$\rho_{pool} = \frac{Tumor\ cells}{all\ cells} = \frac{20 + 75}{200} = \mathbf{0.475}$$

$$MCF_{pool} = \frac{mutant\ cells}{all\ cells} = \frac{18 + 25}{200} = \mathbf{0.215}$$

$$CCF_{pool} = \frac{mutant\ cells}{Tumor\ cells} = \frac{18 + 25}{20 + 75} = \mathbf{0.45}$$