1 Clonal diversity index

1.1 Explanation

The clonal diversity index (CDI) is a new score in Pairtree that Quaid and I developed to quantify the amount of clonal heterogeneity present in a tissue sample, with higher CDIs reflecting more diversity. It's a measurement of uncertainty—intuitively, it answers the question, "If I were to reach blindly into this tissue sample and pull out a random cancer cell, how uncertain would I be about what clone this cell originated from?" The more diverse the clonal composition of a sample, the more uncertainty you'd have, and the higher the CDI would be. Since it's a measurement of uncertainty taken from information theory, the CDI is measured in bits of information. It can never be smaller than zero, but can be arbitrarily high. To get a higher CDI, you can have more clones present in a tissue sample, or you can have those clones present in more equal proportions. E.g., a tissue sample composed of 98% of one clone and 2% of another would have a lower CDI than a tissue sample composed of a 50-50 split of the same two clones. A CDI of zero bits means the sample consists of only a single clone, and so you have no uncertainty about what clone each cancer cell belongs to; a CDI of 1 bit corresponds to two clones, each giving rise to 50% of the sample; two clones present in a 98% and 2% mix would give a CDI of 0.14 bits.

The CDI is a direct translation of the Shannon diversity index from ecology into the cancer genomics setting: https://en.wikipedia.org/wiki/Diversity_index#Shannon_index. We would have preferred to call our measure "SDI" rather than "CDI", but the Wolf (2019) melanoma paper (https://www.sciencedirect.com/science/article/pii/S00928674193 defined "SDI" to be something different (which is wrong and misleading, unfortunately), so we decided to call our measure "CDI" to avoid confusion.

1.2 Examples

The CDI is just a concise summary of the diversity you see in the population frequency plots that Steph is including in the paper. If you look at fig. 1 and fig. 2, you can compare the CDI of rXeno 10 BM and CNS, for example, and see that the BM value is much lower (0.01 bits vs. 0.30 bits). This is because the rXeno 10 BM is very nearly pure, but the CNS has small proportions of the pop. 2 and pop. 4 clones in addition to pop. 5. Conversely, the most diverse sample is the original patient sample, with a CDI of 2.37 bits.

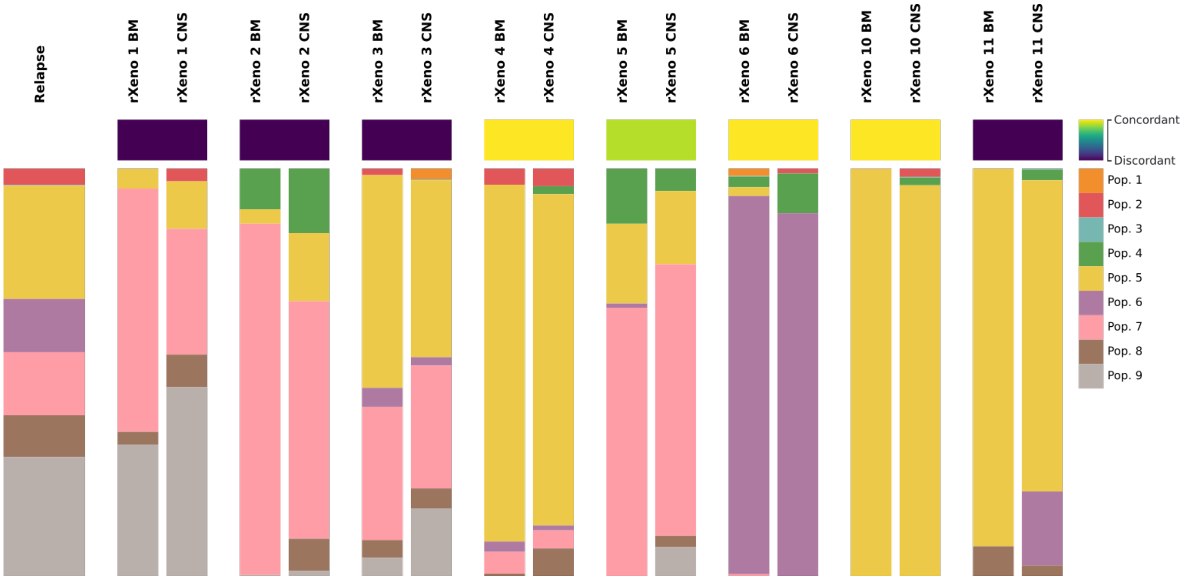

Figure 1: SJMLL039 population frequencies

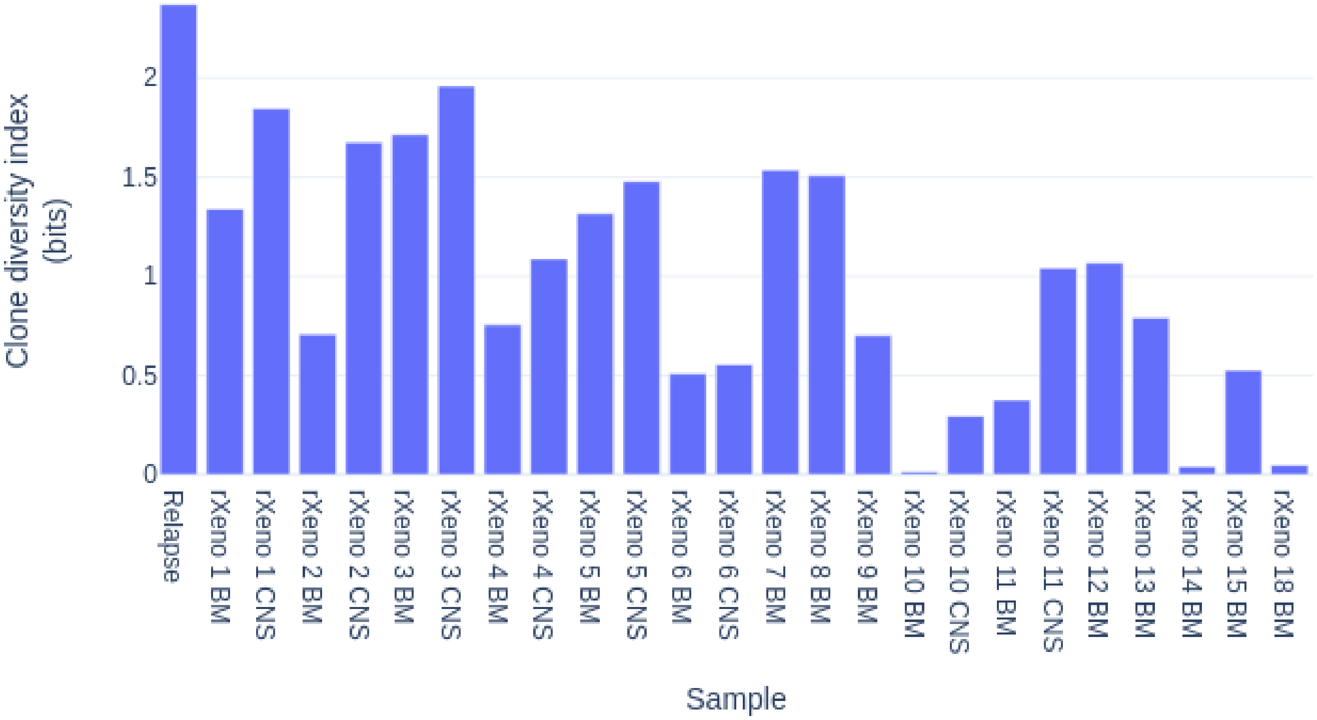

Figure 2: SJMLL039 CDI values in relapse samples

CDI is interesting in the context of this study because we can compare CDI values across all BM/CNS pairs. When we look at the diagnosis samples in either MLL, we don't see a clear relationship. However, in the relapse pairs (fig. 3), we see in SJMLL026 that BM CDI is always greater than the CNS CDI from the same mouse (i.e., under the diagonal), and that in SJMLL039 the BM CDI is always less (i.e., above the diagonal). You could intuit this relationship by looking at the population frequency bars for each pair (as in the rXeno 10 BM/CNS example above), but the CDI gives us a means of quantifying it.

We also did a Wilcoxon signed-rank test using the CDI values for each BM/CNS relapse pair in the two cancers. In SJMLL026, we could reject the null hypothesis of "relapse BM CDI is less than relapse CNS CDI" with $p = 0.0005$, supporting the alternative hypothesis that "relapse BM CDI is greater than relapse CNS CDI". In SJML039, we could reject the null hypothesis of "relapse BM CDI is greater than relapse CNS CDI" with $p = 0.004$, supporting the alternative hypothesis that "relapse BM CDI is less than relapse CNS CDI". This is an interesting result—we consistently see the relapse CNS samples being less diverse than BM in SJMLL026, but more diverse in SJMLL039.

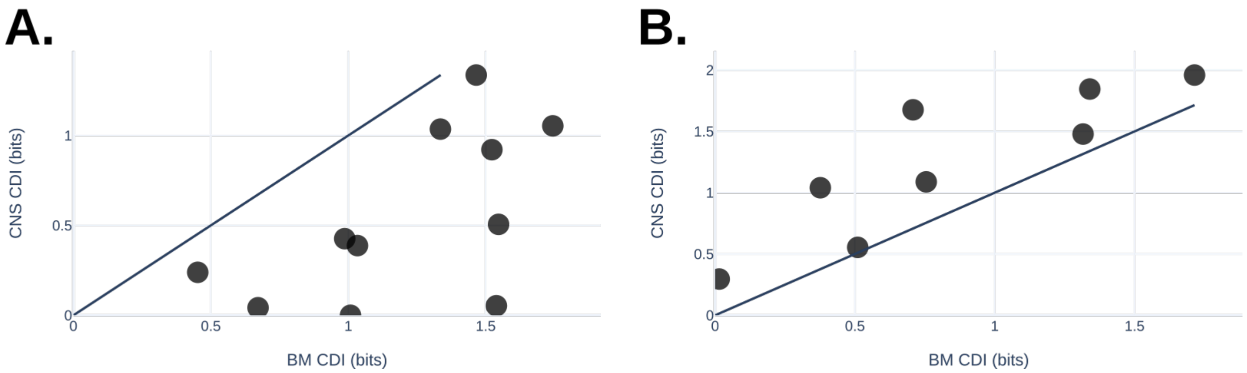

Figure 3: A. SJMLL026 BM vs. CNS CDI. in relapse xenos. B. SJMLL039 BM vs. CNS CDI. in relapse xenos