doi:10.1038/nature20123

Supplementary Notes

Supplementary Note 1. Accounting for the impact of technical and batch effects. We used several approaches to ascertain that our transcriptional signatures are observed independently of technical effects. First, different batches are largely indistinguishable with respect to the expression hierarchy, as shown in **Extended Data Fig. S5b**. Second, to minimize the impact of technical effects, namely the differences in complexity (*e.g.* the number of genes detected per cell), we use a weighted version of principal component analysis as described in Methods. Third, the biological clusters we describe are not driven by complexity. As described in Methods, we performed control PCA on shuffled data. Comparison of the PCA on the original and shuffled data (**Extended Data Fig. S2F**) shows that the OC-like and AC-like genes used in our analysis lose their association with PC1 in the shuffled data, indicating that their patterns are not driven by complexity. Similarly, complexity does not account for the PC2/3 stemness program, as PC2 cell scores are positively correlated with complexity (R=0.27), while PC3 cell scores are negatively correlated with complexity (R=-0.24) and stemness genes were defined as those correlated with both PC2 and PC3.

Supplementary Note 2. Assessing the presence of intermediate differentiation states. Technical noise is not expected to distinguish functionally-related from functionally-unrelated sets of genes. Within a given cell, the level of each gene can be over-estimated or under-estimated due to the capture of only a subset of transcripts and their potentially biased amplification; but there is no reason to expect that two functionally related genes will have the same pattern, *i.e.*, commonly over-estimated or commonly under-estimated, except as correlated to their global expression levels. That is, the exception is if the two genes are both highly expressed or both lowly expressed and thus could be commonly affected by the ”complexity” of single cell libraries, such that two lowly expressed genes tend to be undetected in cells with a lower overall number of detected genes. However, this does not affect our lineage scores, both because the set of AC and OC genes are not associated with very different overall expression levels, and because we use “control” gene-sets with comparable expression levels when defining lineage scores. In each of the three tumors that we profiled at high depth, and within each of the two lineages we find significant co-expression patterns that suggest distinct differentiation states (**Extended Data Fig. 5C**). For example, within the AC lineage, we find significant co-expression patterns in the range of 0.5 to 1, as well as within the range of 1 to 2. However, in more limited ranges we typically do not detect significant co-expression patterns (*e.g.*, in the range 1.5 to 2, we detect significant co-expression only in one of the three tumors). We conclude that cells likely exist in distinct stages of differentiation although the number of distinct states may be limited.