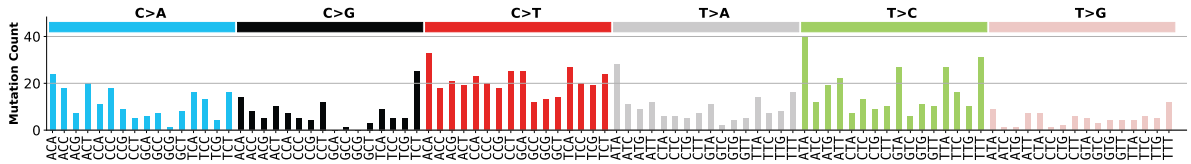
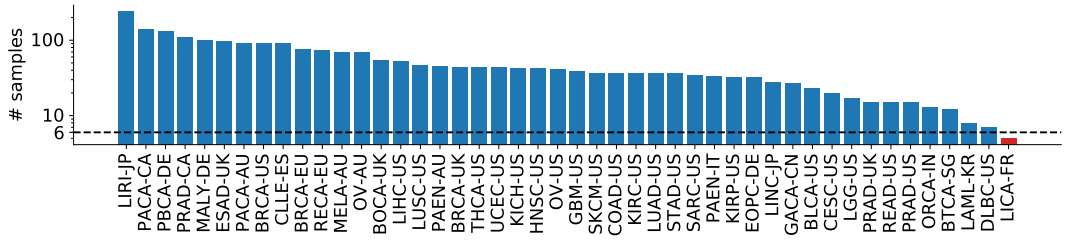
Supplemental Results

Statistics of the mutation count data

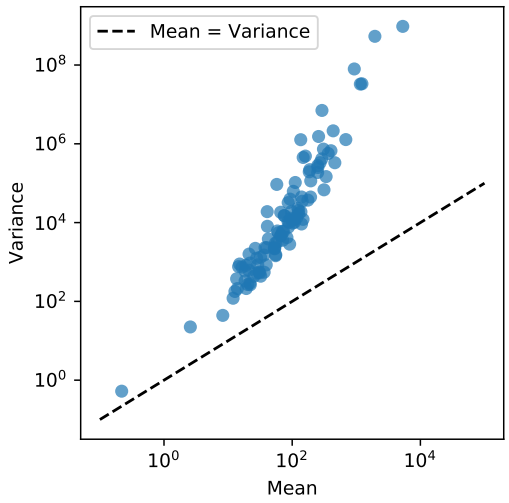


(a)



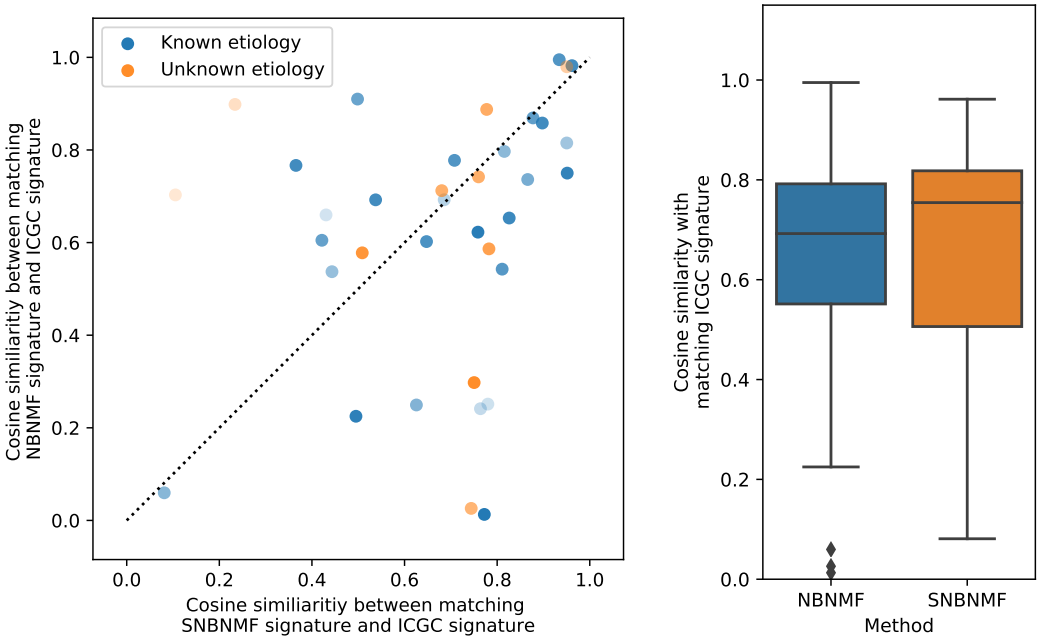
(b)

Supplemental Figure 1: (a) An example of mutational profile with trinucleotide sequences as mutation types. (b) Distribution of project codes across all 2,521 patient samples. The dotted line at 6 indicates the inclusion threshold. Excluded samples are summarized in the red bars.



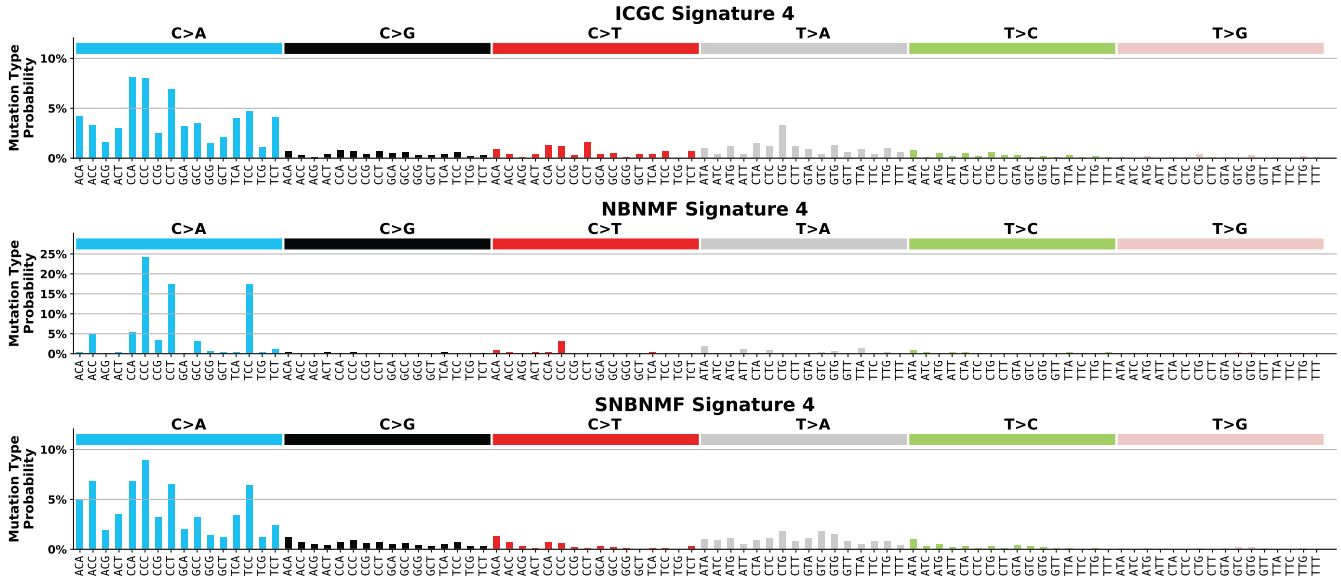
Supplemental Figure 2: Mean and variance of mutational counts of samples with the similar annotation.

Comparison between ICGC, NBNMF and SNBNMF signatures



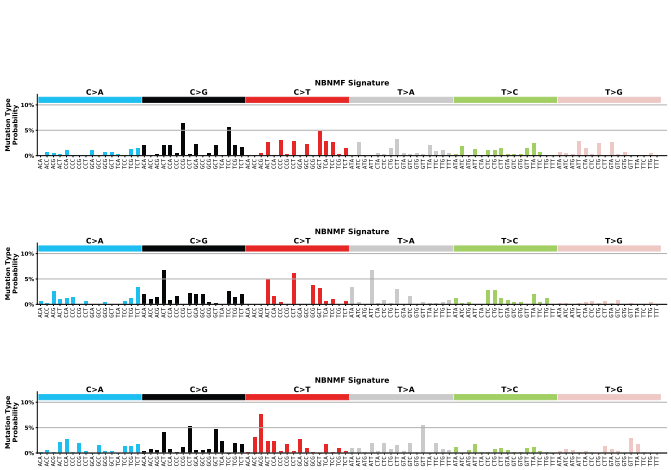
Supplemental Figure 3: Cosine similarity with ICGC signature set. The left figure shows the pairwise comparison between SNBNMF-ICGC signature cosine similarity and NBNMF-ICGC signature cosine similarity, where the points in blue and orange respectively represent signatures with known and unknown etiology and higher signature indexing has higher transparency. Noted that the higher the cosine similarity, the closer the learned signature is to the corresponding ICGC signature. 16 out of 26 signatures with known etiology are more similar to the corresponding ICGC signatures when learned with the proposed method, which also does better in learning signatures with better known etiology (i.e. lower index number.) The right figure compares the distributions of the cosine similarity of NBNMF and SNBNMF, where we observe that our proposed method results in higher median cosine similarity.

Supervised regularization closer to reference signature

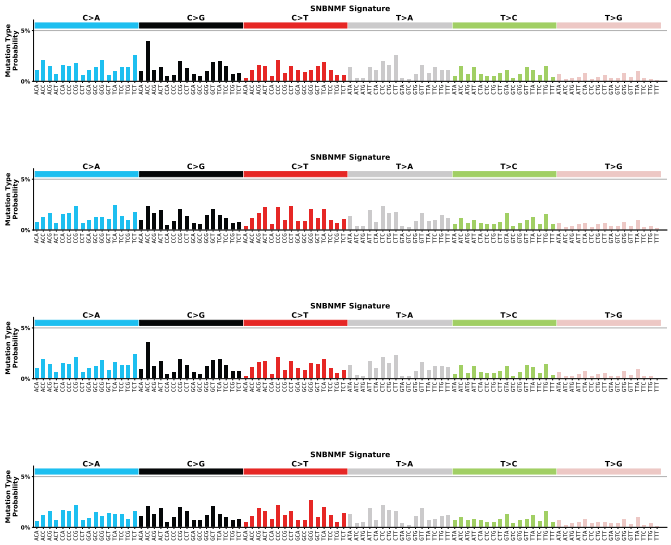


Supplemental Figure 4: Mutational signature 4 generated by the ICGC consortium (Top), a Negative Binomial NMF approach (Middle) and the proposed SNBNMF solution (Bottom).

Robustness of mutational signature

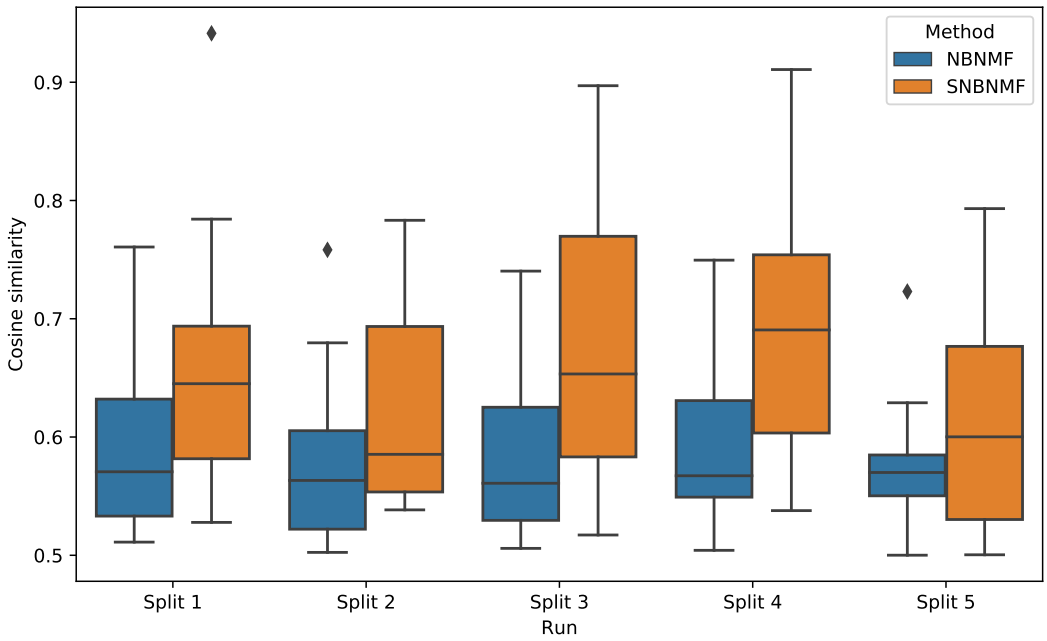


(a)



(b)

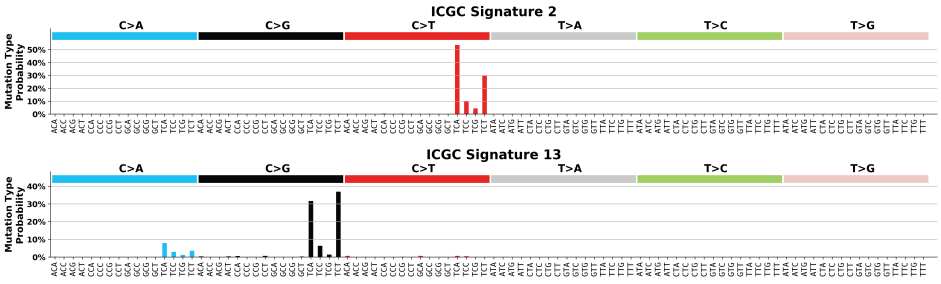
Supplemental Figure 5: NBNMF (a) and SNBNMF (b) signatures learned from the subcohorts with only 6 cancer types that are matched with the COSMIC signature 3.



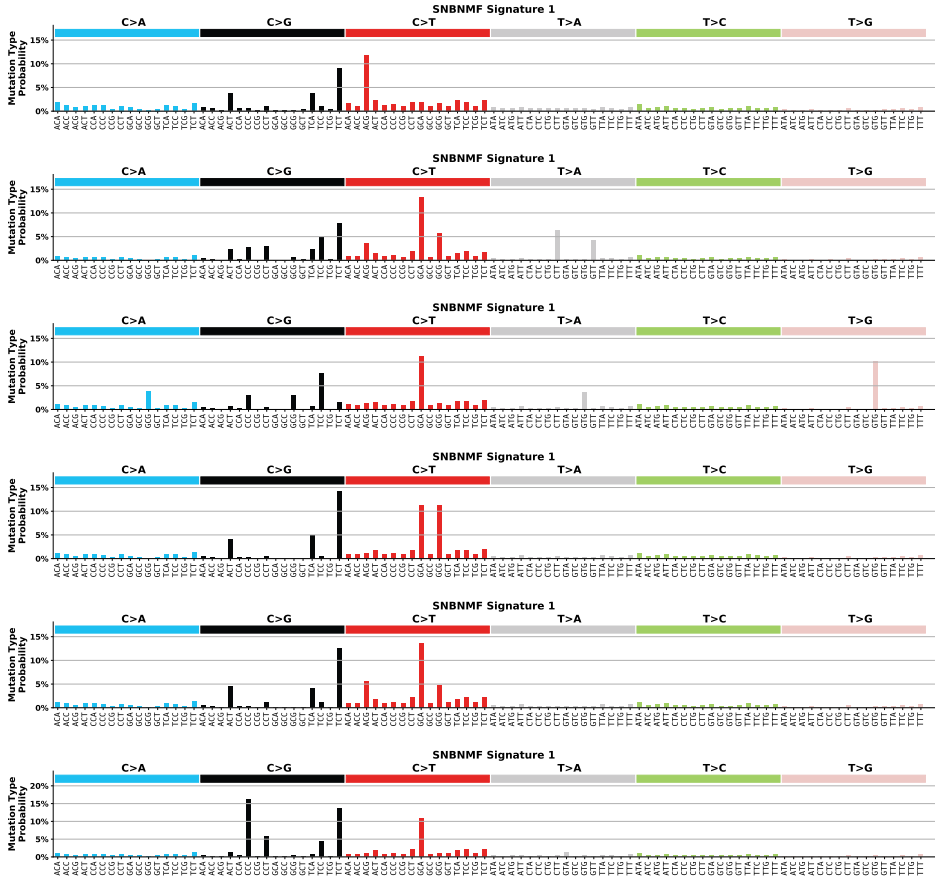
Supplemental Figure 6: Cosine similarity to all signatures using NBNMF (Blue) and SNBNMF (Orange) across multiple runs using different random subsets. Using the supervised regularization term generally improves the cosine distance to the reference signatures using subset of our cohort.

Molecular labels in mutational signature learning

APOBEC expression



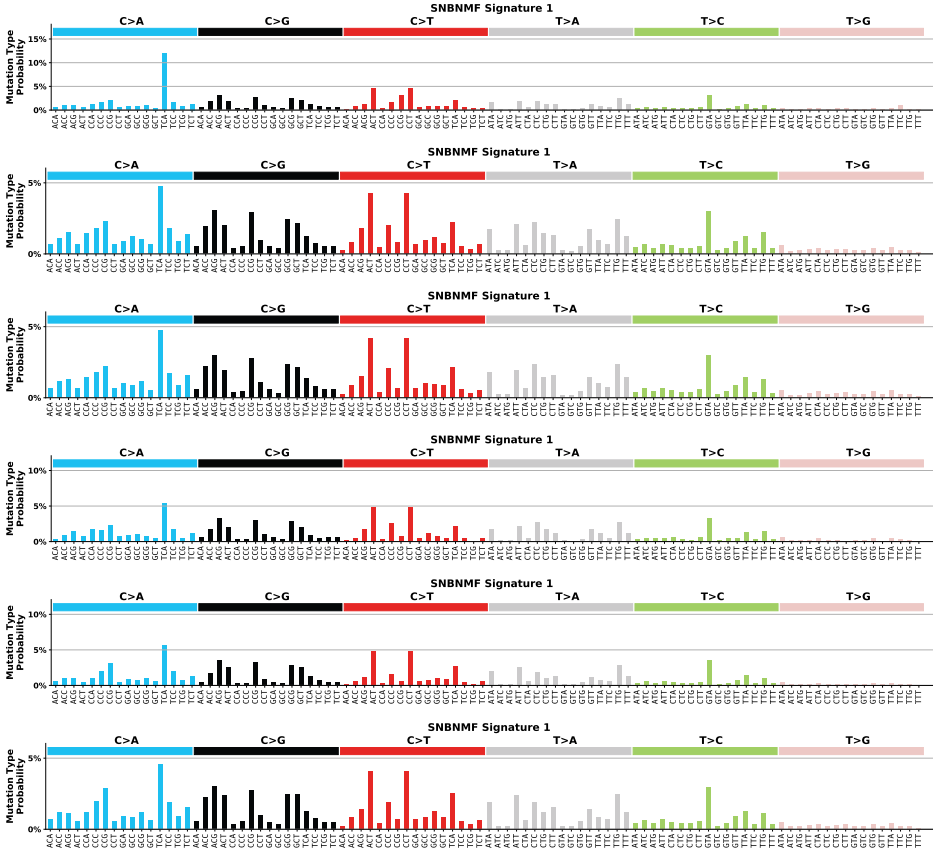
(a)



(b)

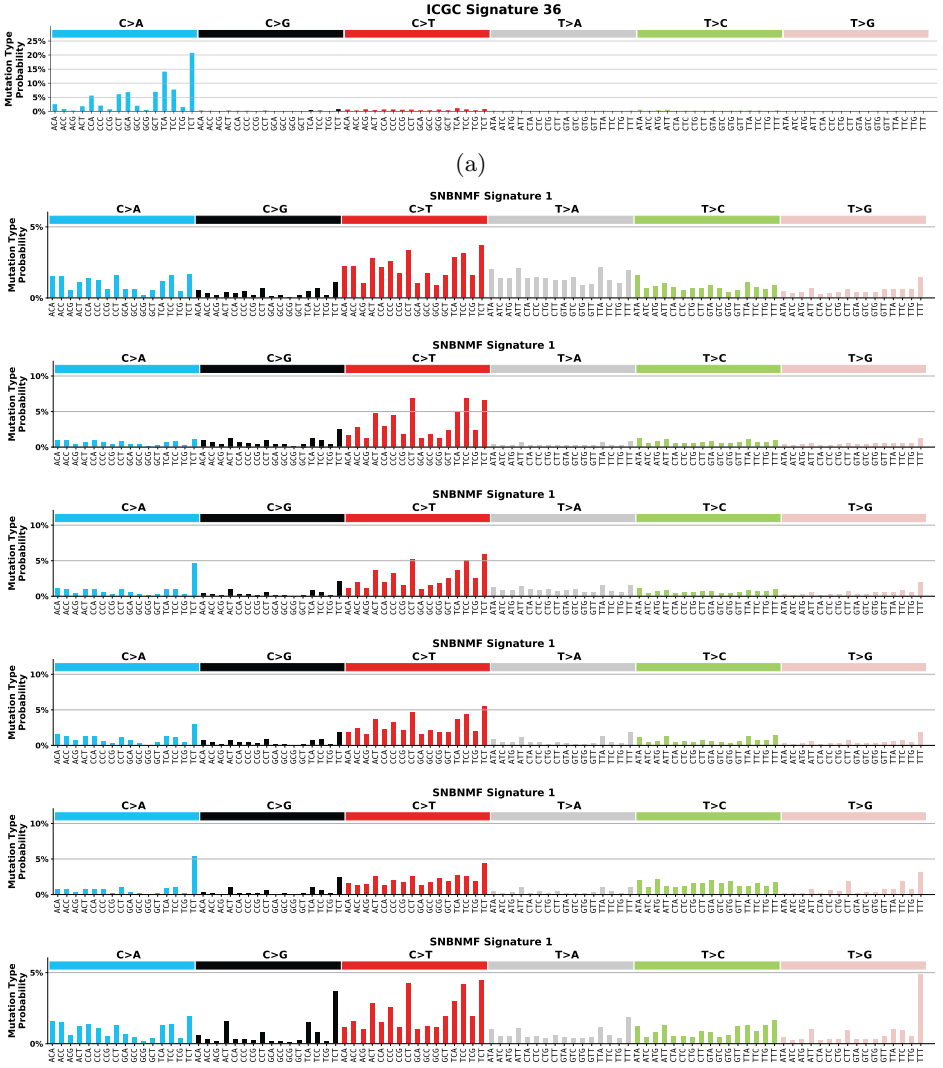
Supplemental Figure 7: The top two signatures are reference signature 2 and 13 with known relationships to APOBEC expression. The remaining signatures show the SNBNMF signature 1 (constrained and trained to be predictive of APOBEC expression) using different number of factors (40-35), with the top one using 40 factors and the bottom one representing the signature trained using only 35 factors.

MUTYH mutation status



Supplemental Figure 8: Signature 1 (constrained to be predictive of MUTYH mutation status) using different number of factors (40-35), with the top one using 40 factors and the bottom one representing the signature trained using only 35 factors. The reference signature is derived from the MUTYH-associated polyposis colorectal cancer patient.

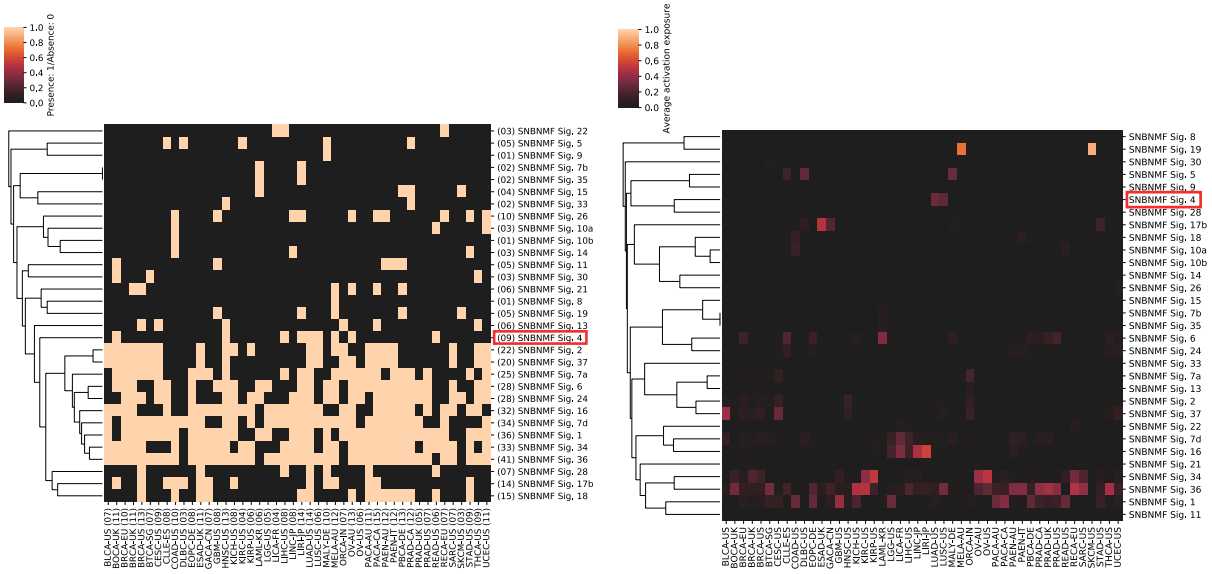
OxoG score



(a)

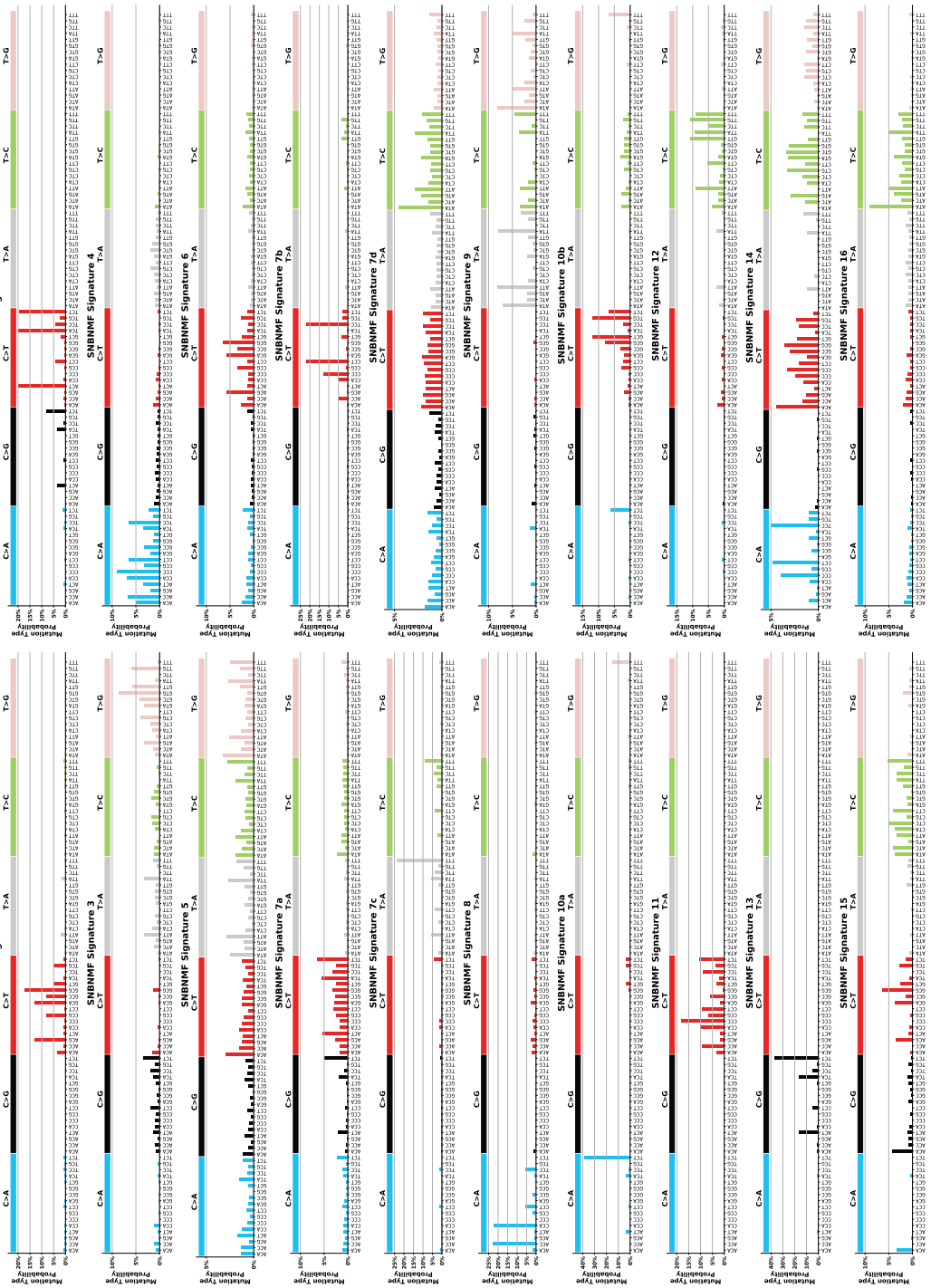
Supplemental Figure 9: Signature 1 (constrained to be predictive of OxoG scores) using different number of factors (40-35), with the top one using 40 factors and the bottom one representing the signature trained using only 35 factors.

Signature analysis



Supplemental Figure 10: Left: Binary matrix showing the activation of mutational signature in samples with individual project codes. Right: Instead of a binary matrix, this heatmap shows the average exposure for a signature within the samples of each project code. To gain a better understanding of the behavior of the SNBNMF approach, we summarized the activation of mutational signatures within individual project codes. This analysis enables us to relate the labels used in the SNBNMF process to the activation of specific signatures. Here, we observe that signature 4, attributed to smoking, is activated in patients with Lung Cancer (the corresponding project codes are LUSC and LUAD). This analysis can help attribute signatures with unknown etiologies to individual labels and thus improve our understanding of the mutational signature process.

Mutational Signatures Catalogues



Supplemental Figure 11: Mutational signatures 1-16 learned by SNBNMF.