doi:10.1038/nature22992

Supplementary Note

TBC1D12 hotspot mutations are located in the 5'UTR of the gene at positions -3 and -1 relative to the translation start site within the Kozak consensus region. The sites have been recently described[1,2] and were associated with APOBEC activity in Nik-Zainal et al. The -1 hotspot mutation (G>A/C; Figure 1) occurs in the canonical APOBEC TCW context (AGA/TCT). The -3 hotspot mutation (C>T) does not match the canonical APOBEC context (CCA/TGG), however, APOBEC does have some very low-level activity in the CCA context. Mutational signature analyses[3-5] can assign each mutation (in each patient sample) a probability of being generated by the particular signature (APOBEC, in this case). This probability is estimated based on the context of the mutation and the overall APOBEC activity in the specific sample. Therefore, the same mutation in different patients can be assigned different probabilities of being associated with APOBEC. A mutation that is less frequent in the APOBEC signature may also be attributed to APOBEC, if a sample has very high overall APOBEC activity.

Because *TBC1D12* mutations are close to the coding region, we searched for additional events in exome sequencing data. We found the same hotspot mutations in a large percentage of bladder cancers from TCGA (17.5% overall; 10.8% at -1 and 9.3% at -3 relative to ATG) (Extended Data Fig. 4a; Supplementary table 5). Of these, 15% (4/47) carried alterations at both hotspot positions (Fisher's Exact test p = 0.01), which, as seen in our breast cancer patients, were located on different alleles (Extended Data Fig. 4b). Overall, the statistical evidence for the *TBC1D12* promoter is strong: it is significantly mutated in breast and bladder cancer, with two hotspots adjacent to the translation start site that co-occur in a significant number of patients on homologous chromosomes. Even when analyzing mutations in each of the hotspots separately, the promoter is still significant ($q=5.3 \times 10^{-11}$ and $q=9.44 \times 10^{-10}$, respectively).

The location of the *TBC1D12* hotspot mutations at position -3 and -1 relative to the annotated start codon suggest that these mutations may affect translation. Luciferase reporter assays with constructs containing the

TBC1D12 native promoter sequence and 5'UTR showed a decrease in signal (i.e., lower luciferase protein production) for the C>T but not the G>A mutation (Extended Data Fig. 5e). Neither the wild-type nor mutant sequences showed any binding in the EMSA (Extended Data Fig. 5e), suggesting that the hotspot mutations do not affect transcriptional regulation (consistent with a potential translation-related effect). Inspection of the *TBC1D12* upstream region in the human genome (and several primates) revealed an extended open reading frame and two evolutionary conserved in-frame ATG codons located within the core promoter (Extended Data Fig. 4c). Had translation started at these upstream sites, the hotspot mutations would have created coding alterations, with the C>T forming a nonsense mutation leading to aborted translation and the G>A mutation creating a synonymous variant. Since the promoter in our luciferase assay was based on the genomic promoter, including the alternative start sites, we wanted to determine whether translation of luciferase started upstream of the annotated start codon. We, therefore, compared the sizes of the luciferase proteins encoded by our three vectors. The luciferase protein produced in cells transfected with the wild-type and G>A variant vector was larger than the protein encoded by the C>T variant vector (Extended Data Fig. 4d), suggesting that the translation of luciferase indeed started at one of the upstream ATGs. Based on our data, we, therefore, cannot conclude if the hotspot mutations would have altered protein translation *in vivo*. Further study is required to understand the exact role of these hotspot mutations and whether they promote tumorigenesis.

1 Araya, C. L. *et al.* Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nature genetics*, doi:10.1038/ng.3471 (2015).
2 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).
3 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
4 Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature communications* **6**, 8866, doi:10.1038/ncomms9866 (2015).
5 Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature genetics* **48**, 600-606, doi:10.1038/ng.3557 (2016).

Supplementary Figure 1

Figure 2a

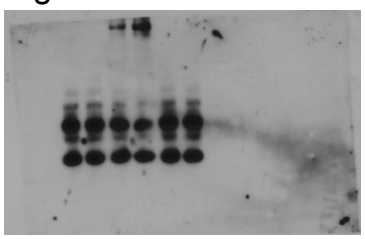
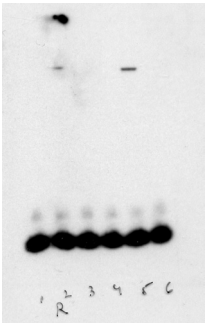


Figure 2b



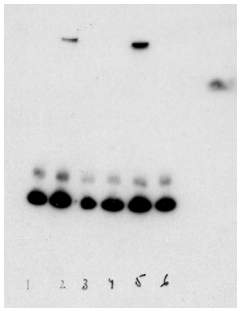
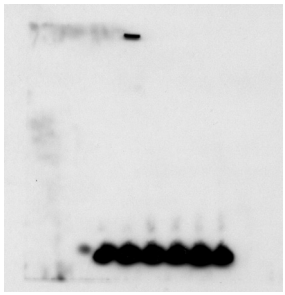
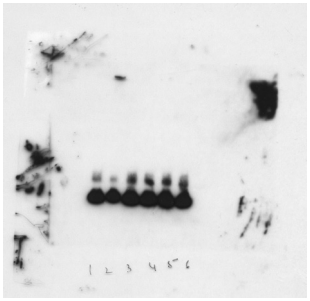
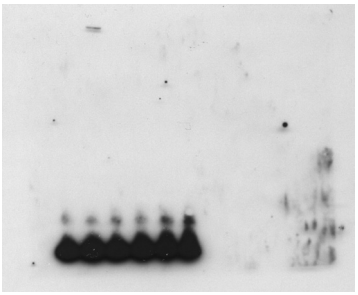


Figure 2c







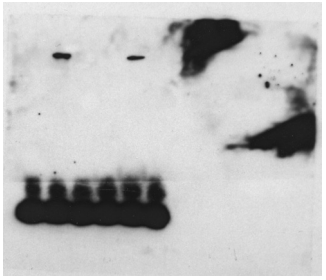
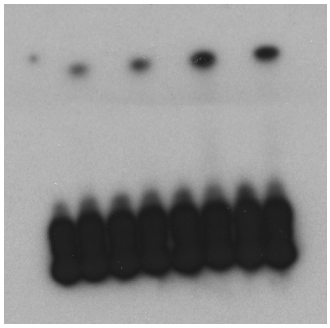
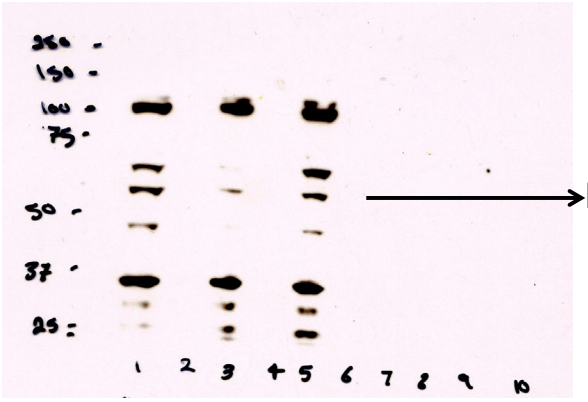


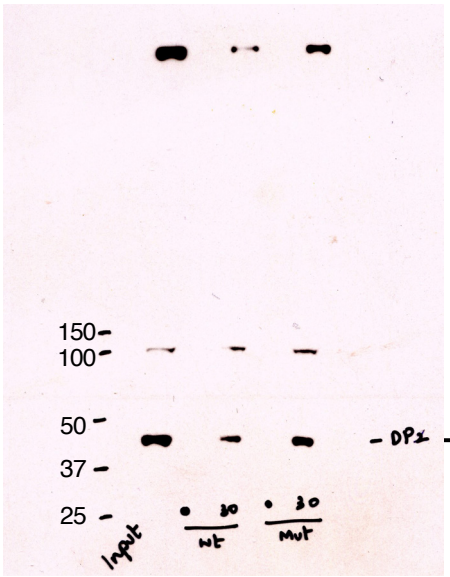
Figure 3b



No molecular markers were run for EMSA.

Figure 3d





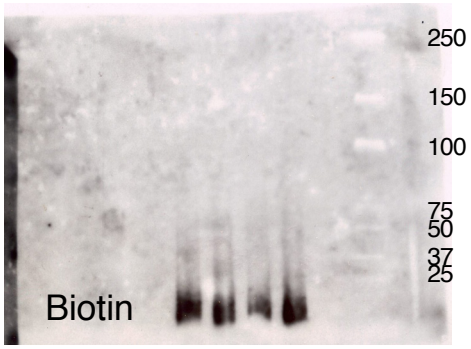
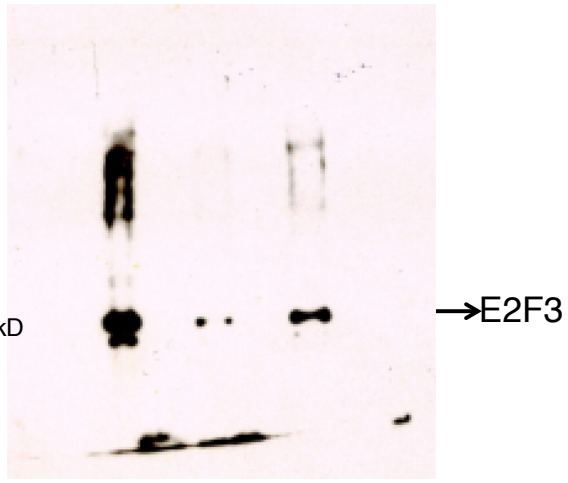
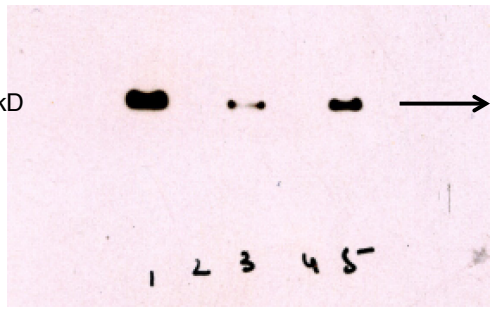
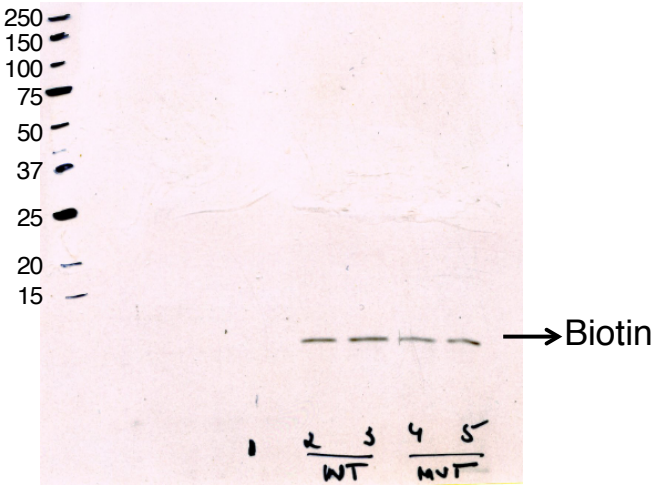


Figure 3e

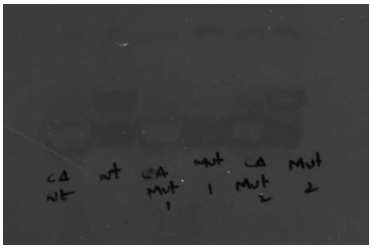




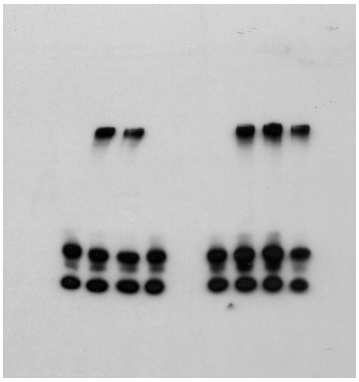


Extended Data Figure 4d

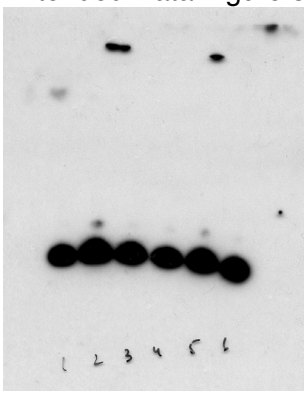
Adjustments made for brightness and contrast



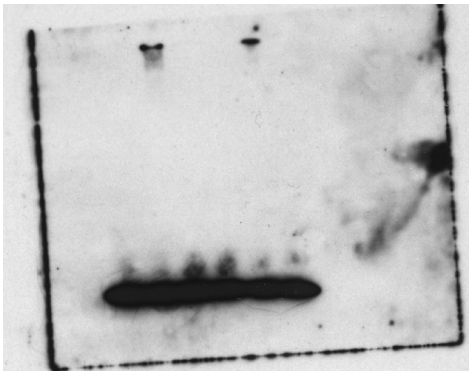
Extended Data Figure 5a

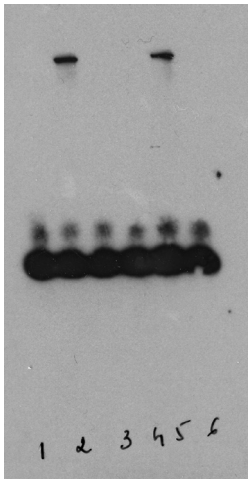


Extended Data Figure 5b

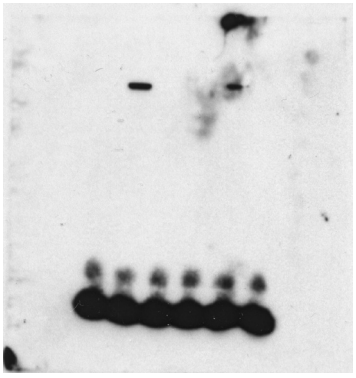


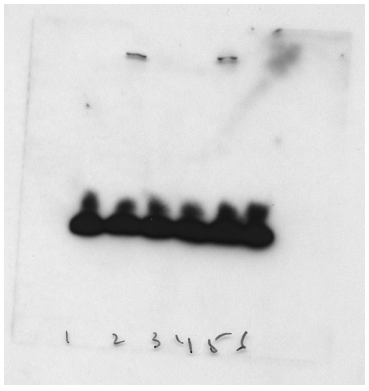
Extended Data Figure 5c



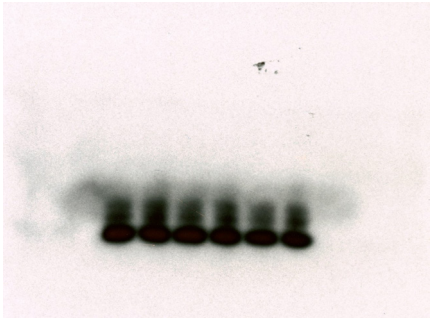


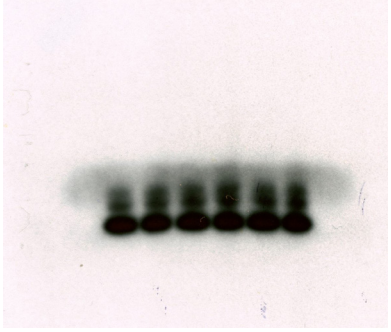
Extended Data Figure 5d



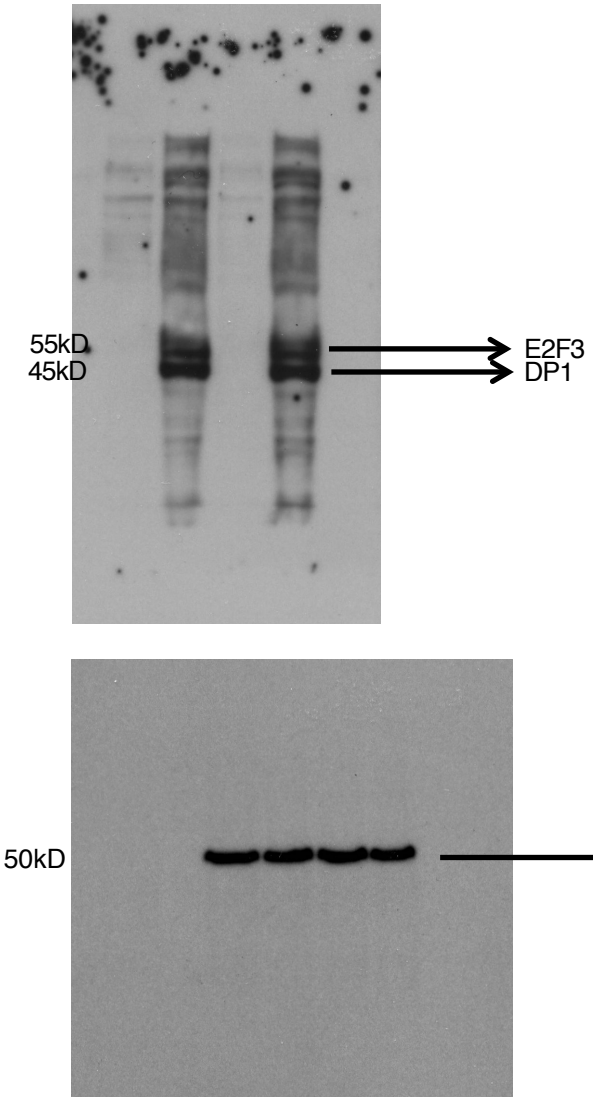


Extended Data Figure 5e

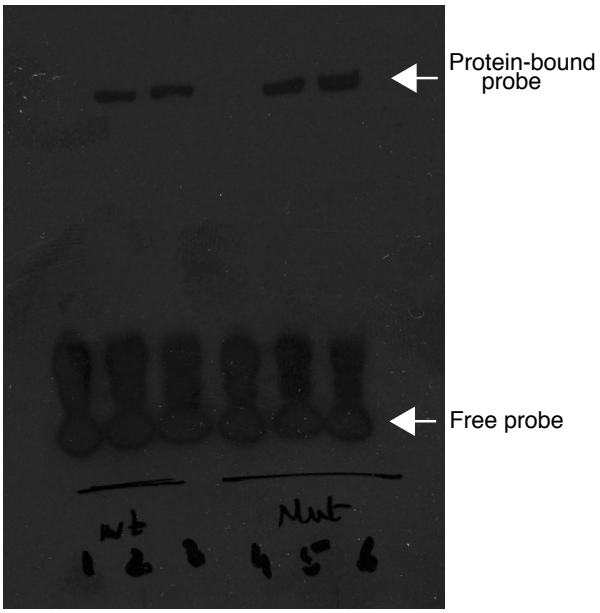




Extended Data Figure 6a



Extended Data Figure 6b
Adjustments made for brightness



Extended Data Figure 8