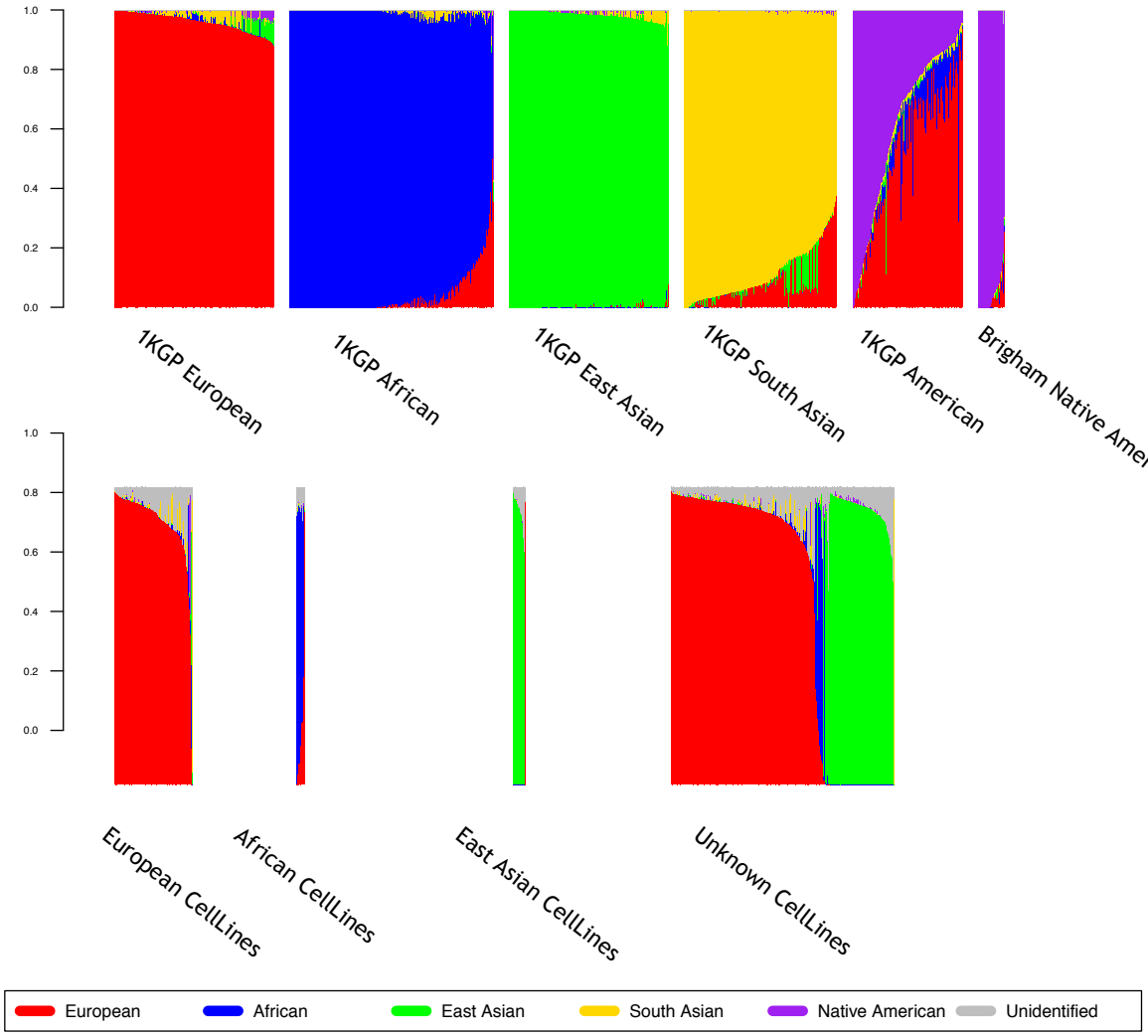
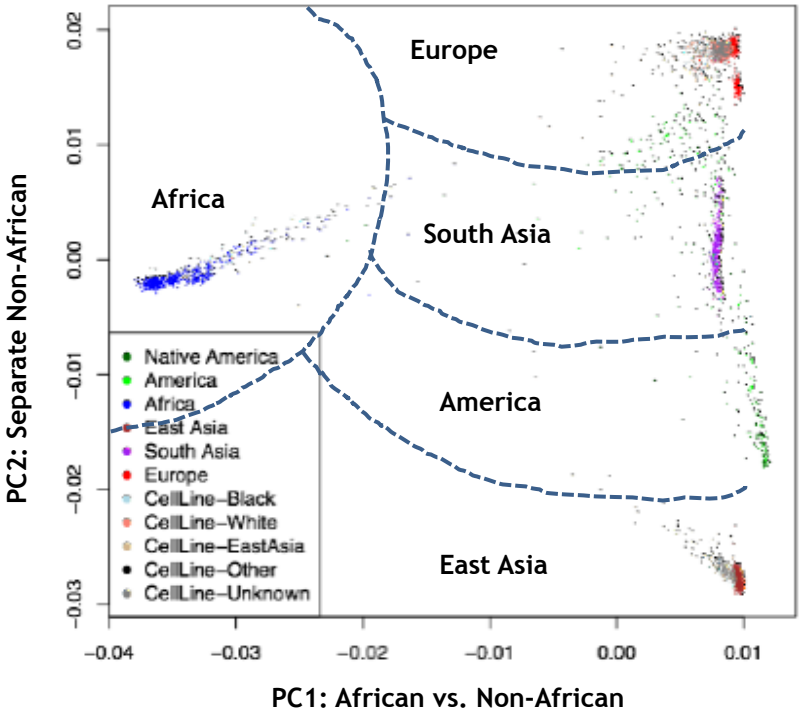


Figure S1 | Admixture of COSMIC cell lines with reference samples. Genomic ancestry estimates are shown for 1009 cell lines from the COSMIC database and for the reference samples used to identify genomic clusters. Each vertical bar represents a different sample, and the height of the bar (y-axis) represents the total genomic ancestry proportion. The height of each color represents the proportion of the ancestry represented by that color (see legend). Groups labeled with the text "1KGP" represent reference samples from the 1000 Genomes Project, and the Native American reference samples are from Bigham et al. (2010). Panels labeled with the text "CellLines" represent cell line samples. Cell lines reported as African ("African CellLines", n=26) show predominantly African ancestry (blue), as well as a gradation of European ancestry proportion (red). Cell lines reported as European ("European CellLines", n=244) are mostly comprised of European ancestry (red), with some cell lines showing predominantly African ancestry (blue), and some showing small amounts of South Asian ancestry (gold). Cell lines reported as East Asian ("East Asian CellLines, n=38) are almost exclusively of East Asian ancestry, except for one inaccurately reported cell line that shows exclusively European ancestry (red). Among the 701 cell lines for whom ancestry was reported as unknown ("Unknown CellLines"), 453 were of predominantly European ancestry (red), 30 were of predominantly African ancestry (blue), and 215 were of predominantly East Asian ancestry. Amongst this group of cell lines with previously unknown ancestry, the predominantly African cell lines were the most admixed, followed by the predominantly European cell lines, and then the mostly non-admixed cell lines of East Asian origin.



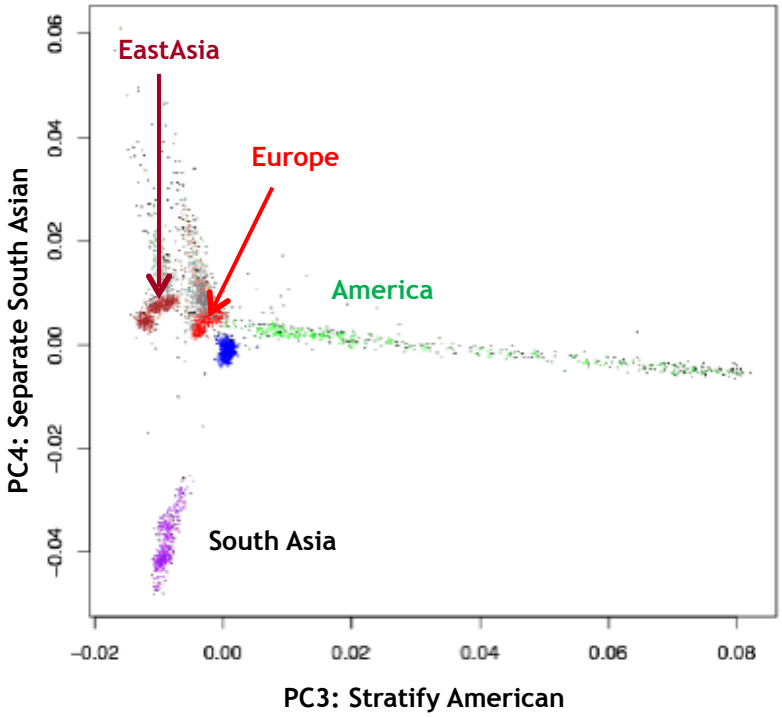


Figure S2 | Principal Component Analysis of COSMIC cell lines. Using our combined data set after filtering for linkage disequilibrium and allele frequency, we can observe clustering on the basis of ancestry between cell lines and reference populations. The first principal component (PCs) split African populations from those that migrated out of Africa, and the second PC identifies the differences between European, Asian, and Native American groups. The third PC identifies Native American ancestry from Asian, and the forth simultaneously identifies unidentified cluster signal (>0) and South Asian populations (<0). Grey dots represent individual cell lines that do not have ancestry or race reported in COSMIC. Lighter colors represent cell lines with ancestry reported, while darker colors represent reference samples from 1000 Genomes Project and Native Americans from Bigham et al. (2010).

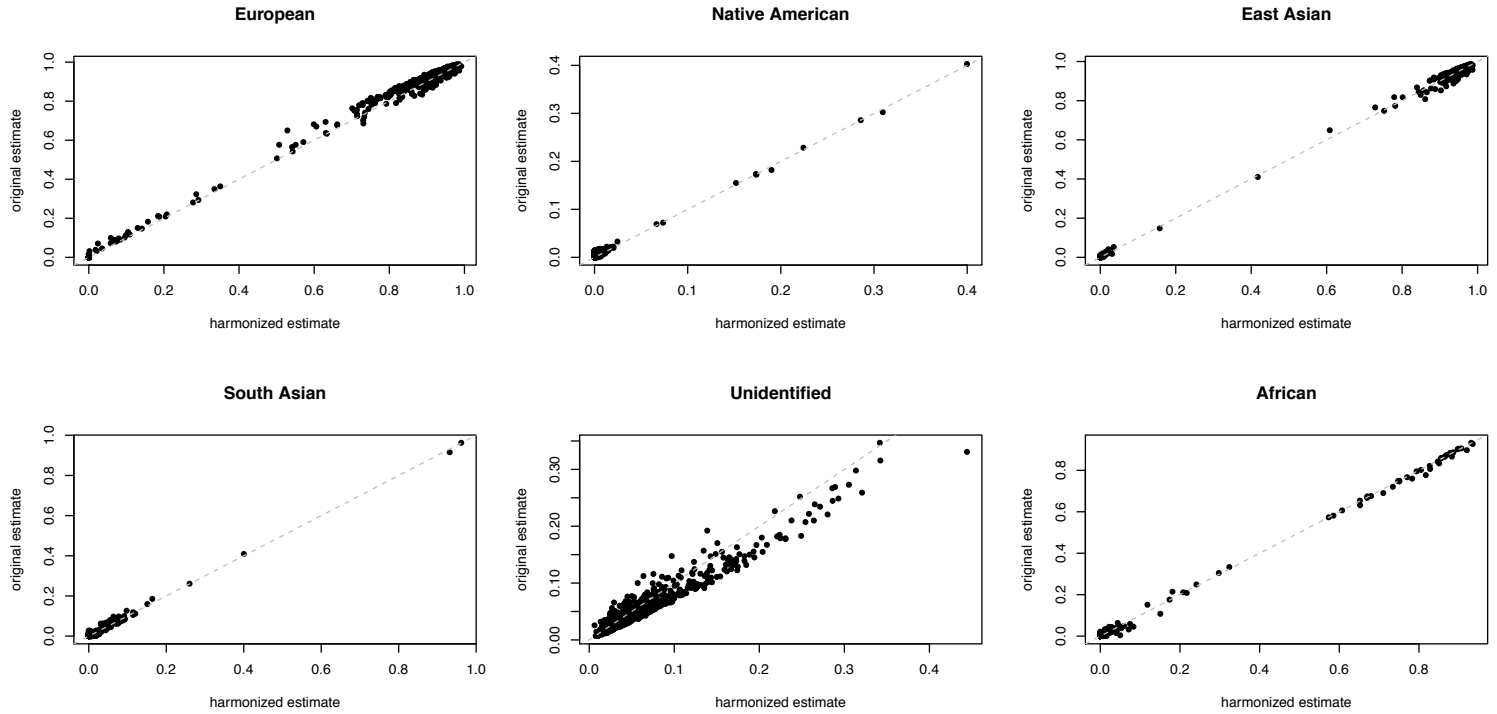


Figure S3 | Ancestry estimates from admixture analysis are unaffected by tumorous copy number changes. Harmonized estimates, or ancestry estimates derived from only the genotypes whose allelic ratio is not effected by copy number (x-axis), are highly concordant with our original admixture estimates (y-axis). This supports that our original estimates are not significantly influenced by copy number changes.

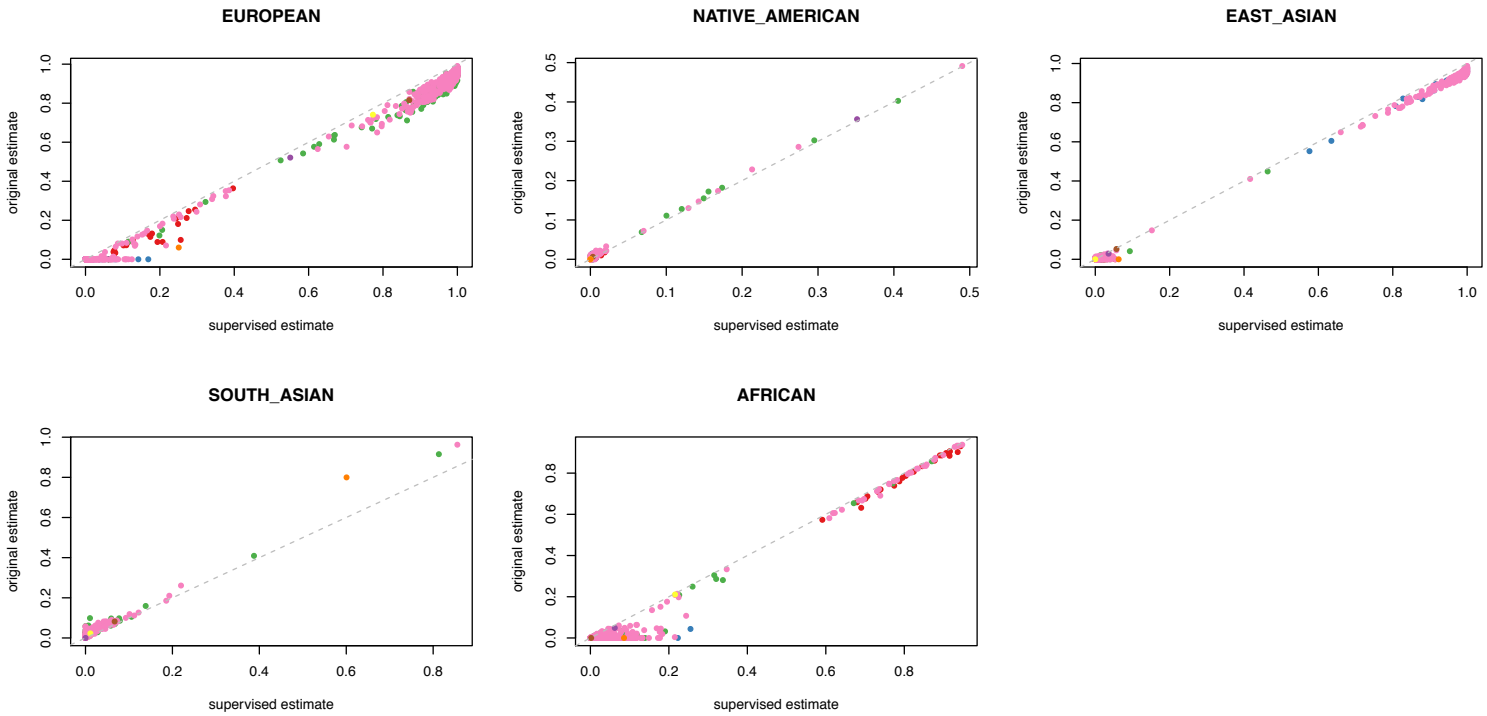


Figure S4 | Supervised vs unsupervised learning admixture analysis. Ancestry estimates are shown from an admixture analysis using a supervised learning approach (x-axis) and an admixture analysis using an unsupervised learning approach (y-axis, our original estimates). The results from the supervised approach, which does not estimate the membership proportion of any clusters (i.e. the 6th unknown cluster) beyond the 5 inherent in our reference samples, are concordant with our original estimates.

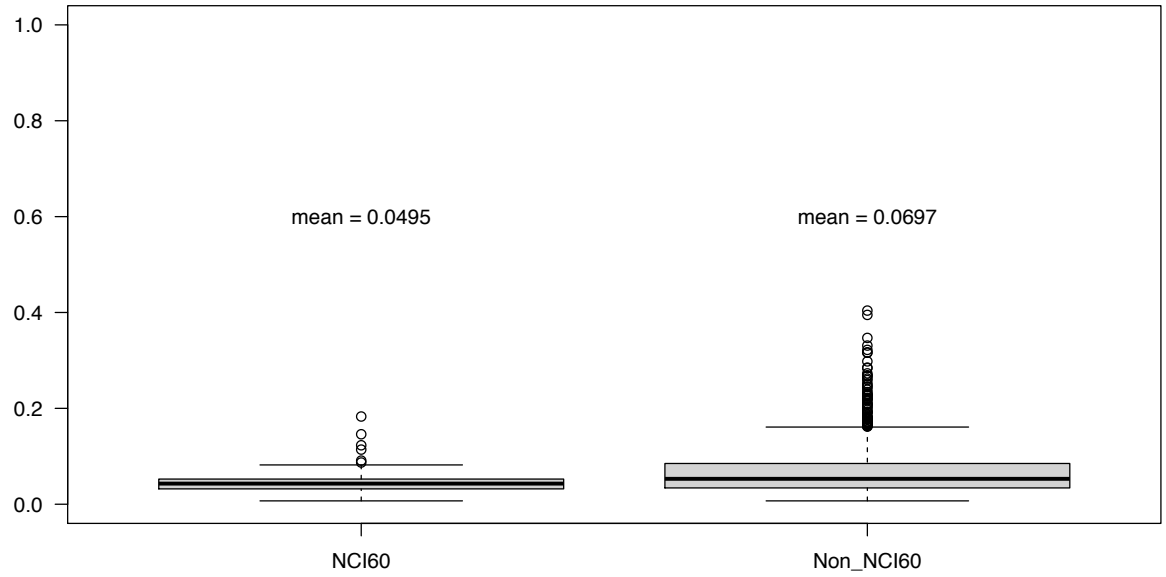


Figure S5 | Unidentified cluster proportion in NCI60 and non-NCI60 cell lines. The mean genomic proportion of the unidentified 6th cluster across NCI60 cell lines is significantly lower than the mean proportion across non-NCI60 cell lines (p-value=1.9x10-5).

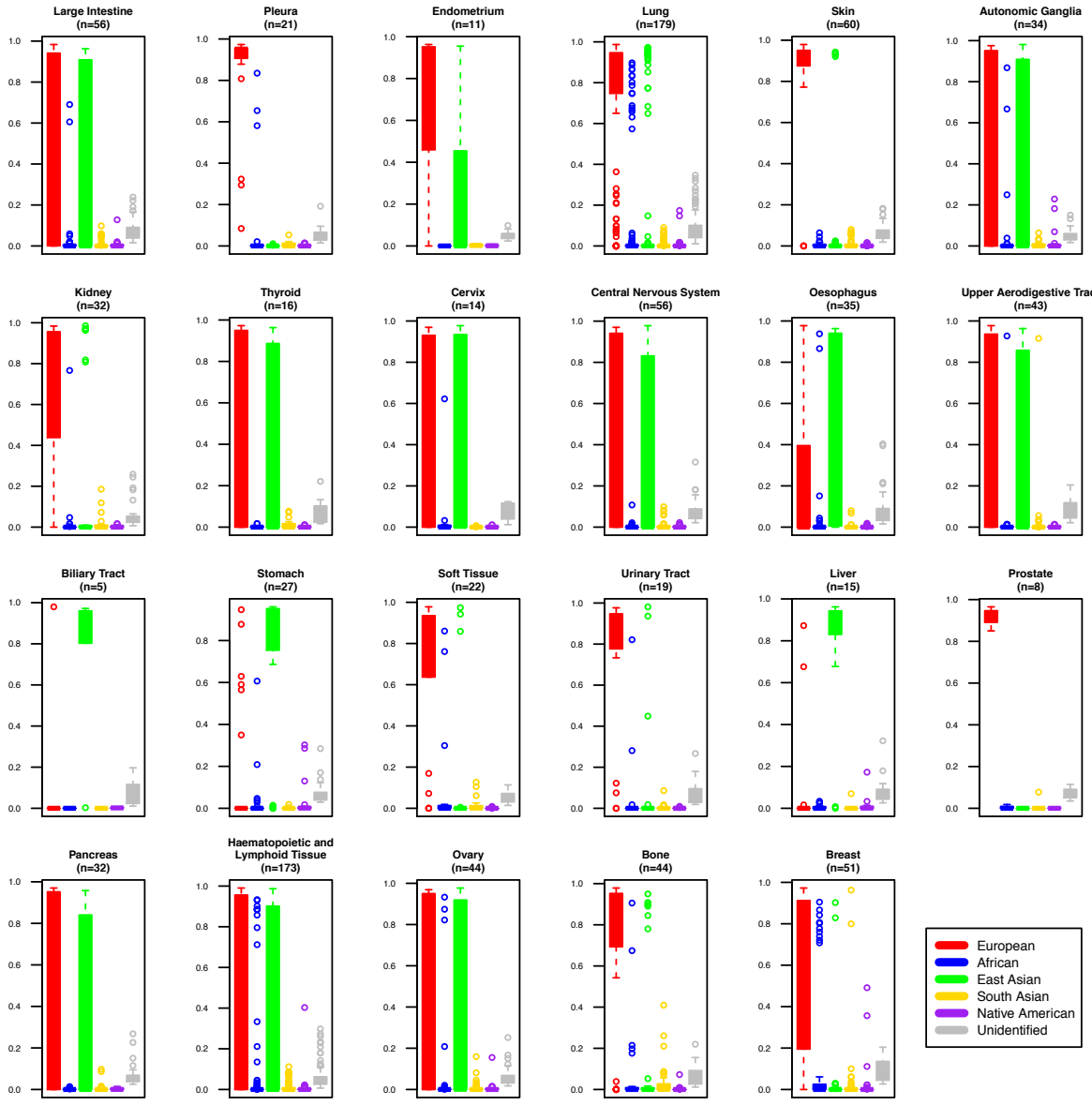


Figure S6 | Distributions of ancestral estimates across tumor type. Boxplots are shown that represent the distributions of ancestry proportion estimates for cell lines belonging to each of 23 cancer types. The number of cell lines belonging to each tumor type that are represented in each panel are shown below each tumor type name. Most cancer types are of predominantly European (red), Asian (green), or European and Asian ancestry. Exceptions to this are lung, hematopoietic and lymphoid, and breast tumors, which have significant numbers of cells lines with predominantly African ancestry (blue). Cell lines from prostate cancer, one of the most common cancers in all men, have only European ancestry. Few or no cell lines from cancers with significantly high incidence rates in African ancestry individuals, like stomach cancer, liver cancer, pancreatic cancer, and kidney cancer, have African ancestry (blue).

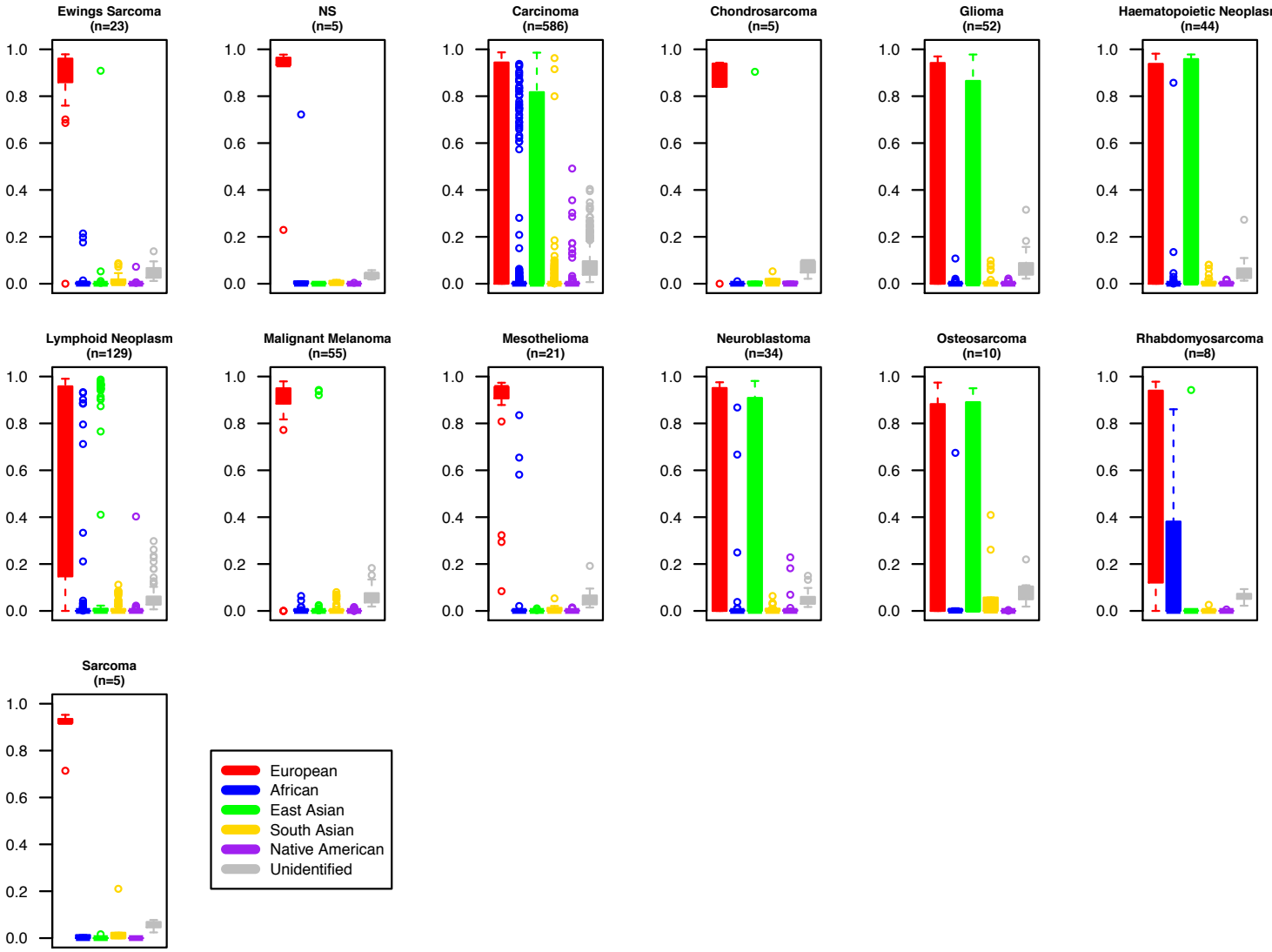


Figure S7 | Distributions of ancestral estimates across primary histology type. Boxplots are shown that represent the distributions of ancestry proportion estimates for cell lines belonging to each of 13 primary histology types. The number of cell lines belonging to each tumor type that are represented in each panel are shown below each tumor type name. Most cancer types are of predominantly European (red), Asian (green), or European and Asian ancestry. One exception to these patterns is within cell lines from rhabdomyosarcomas, which have significant African ancestry and lack East Asian ancestry. Carcinoma cell lines are the most ancestrally diverse, and show significant ancestral representation from Europeans (red), Africans (blue), East Asians (green), and even South Asians (gold) and Native Americans (purple).

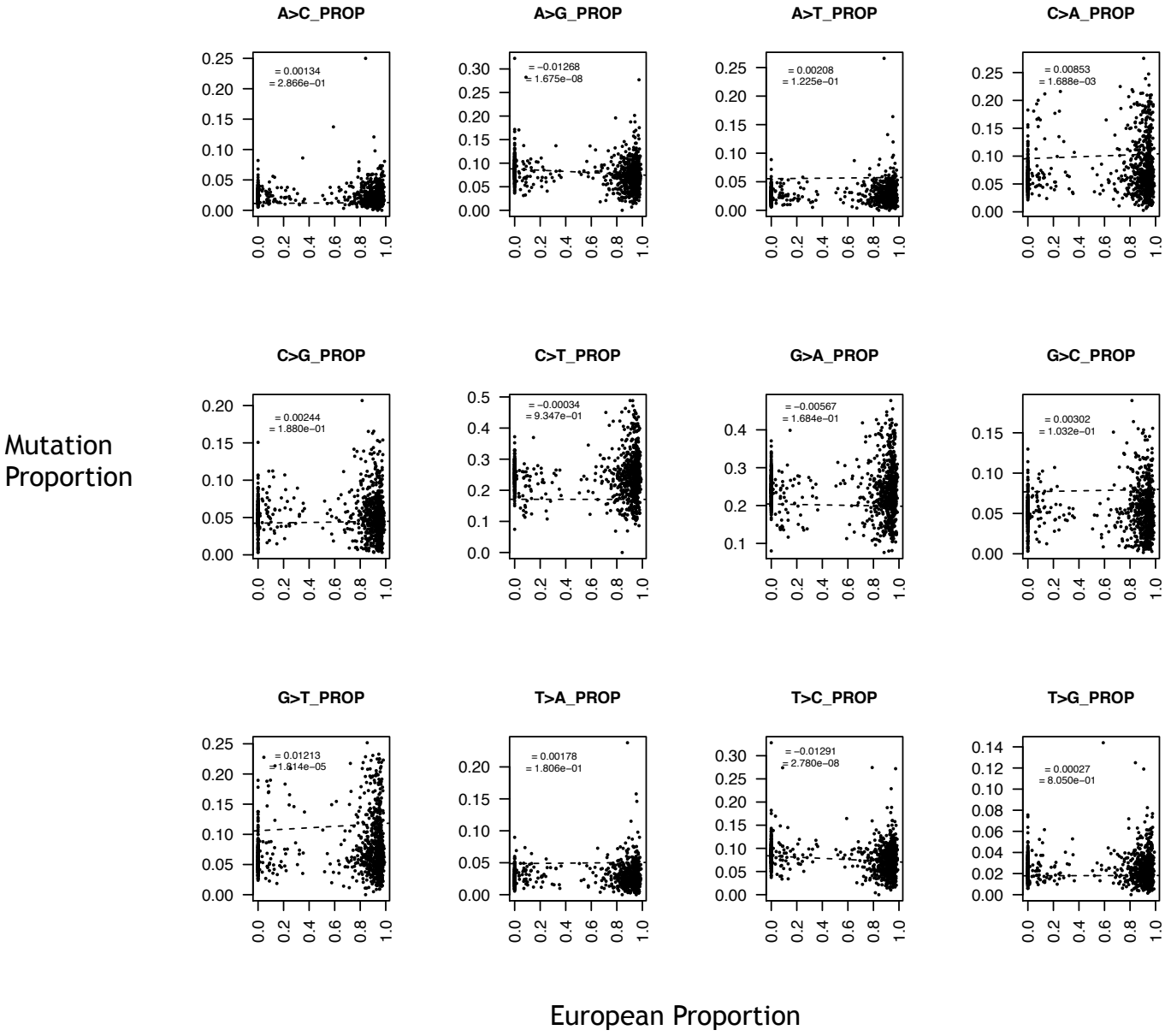


Figure S8 | Scatterplots of European ancestry proportion compared with mutation proportion. The relationship is shown between European ancestry proportion and the proportion of coding single nucleotide mutations made up by each of twelve possible mutation types. Significant relationships are seen for four mutation types (i.e. A>G/T>C and C>A/G>T, Table 1), which represent two independent mutation classes after accounting for reverse complementation. β_a and p-values are shown in each figure panel, and represent the association between European ancestry and mutation proportion for each mutation type after accounting for primary site and histology. Dotted black lines represent regression lines.

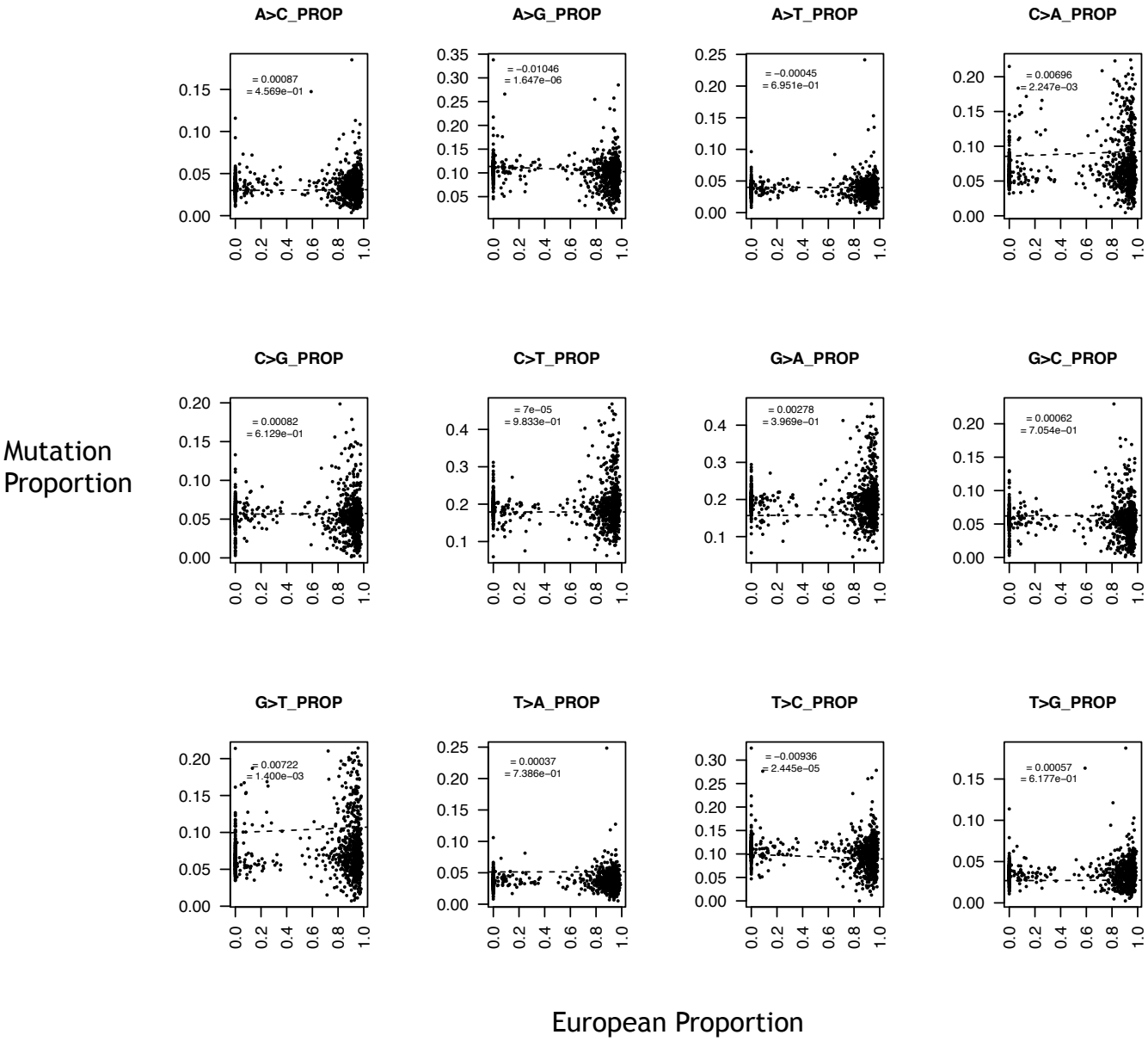


Figure S9 | Scatterplots of European ancestry proportion compared with mutation proportion in noncoding regions. The relationship is shown between European ancestry proportion and the proportion of noncoding single nucleotide mutations made up by each of twelve possible mutation types. As in coding regions, significant relationships are seen for four mutation types (i.e. A>G/T>C and C>A/G>T, Table 1), which represent two independent mutation classes after accounting for reverse complementation. β_a and p-values are shown in each figure panel, and represent the association between European ancestry and mutation proportion for each mutation type after accounting for primary site and histology. Dotted black lines represent regression lines.