Large-scale genomic analysis of 21 cancer types points towards saturating cancer gene discovery

Michael S. Lawrence, Petar Stojanov, Craig H. Mermel, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, Gad Getz

Supplementary Information

Supplementary Figures

Supplementary Tables

Supplementary Figure 1

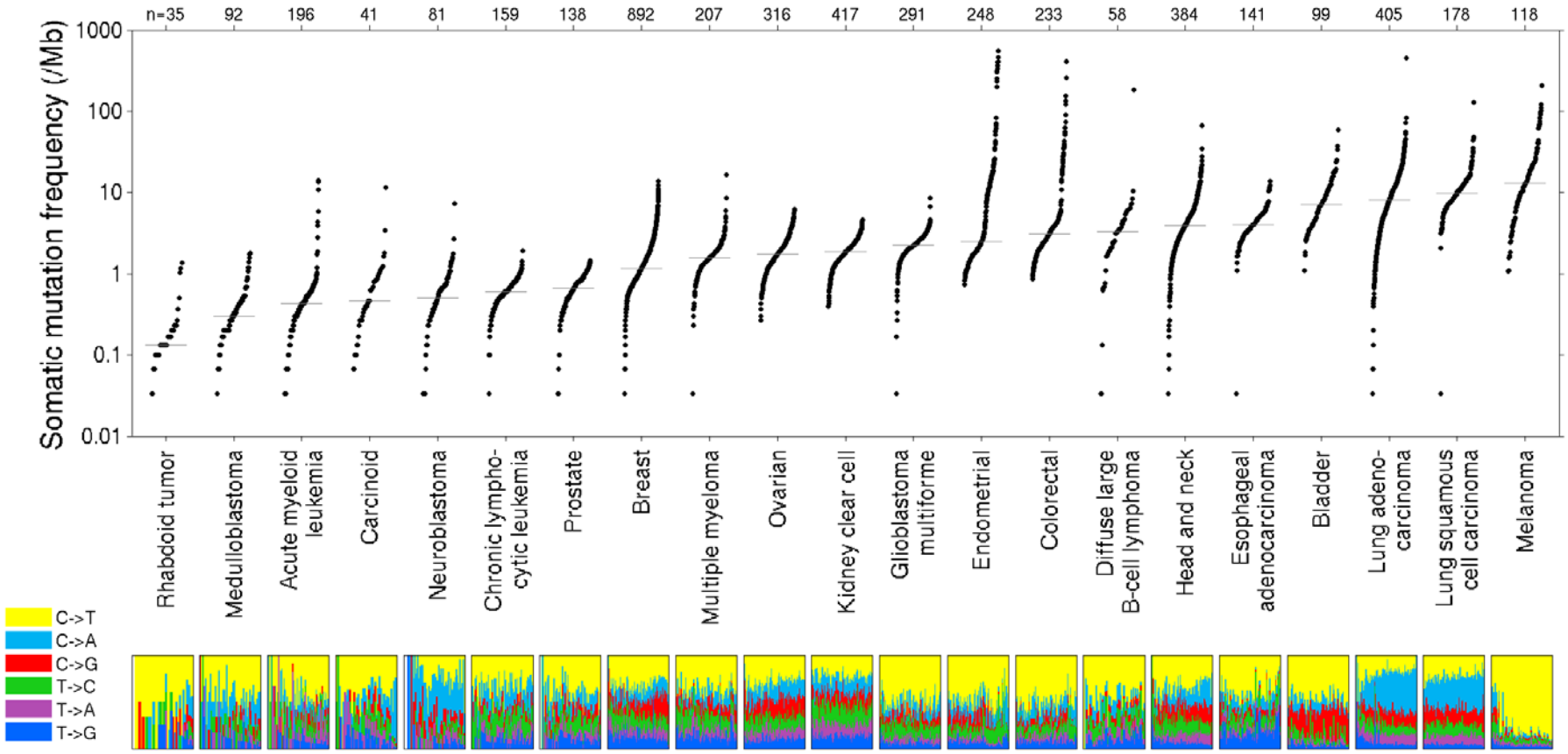


Figure S1. Distribution of mutation rates and spectra across 4,742 tumor-normal (TN) pairs from 21 tumor types, as in Figure 1 from Lawrence et al. *Nature* (2013). Each dot corresponds to a tumor-normal pair, with vertical position indicating the total frequency of somatic mutations in the exome. Tumor types are ordered by their median somatic mutation frequency, with the lowest frequencies (left) found in hematological and pediatric tumors, and the highest (right) in tumors induced by carcinogens such as tobacco smoke and UV light. Mutation frequencies vary more than 1000-fold between lowest and highest mutation rates across cancer and also within several tumor types. The lower panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend on the left.

Supplementary Figure 2

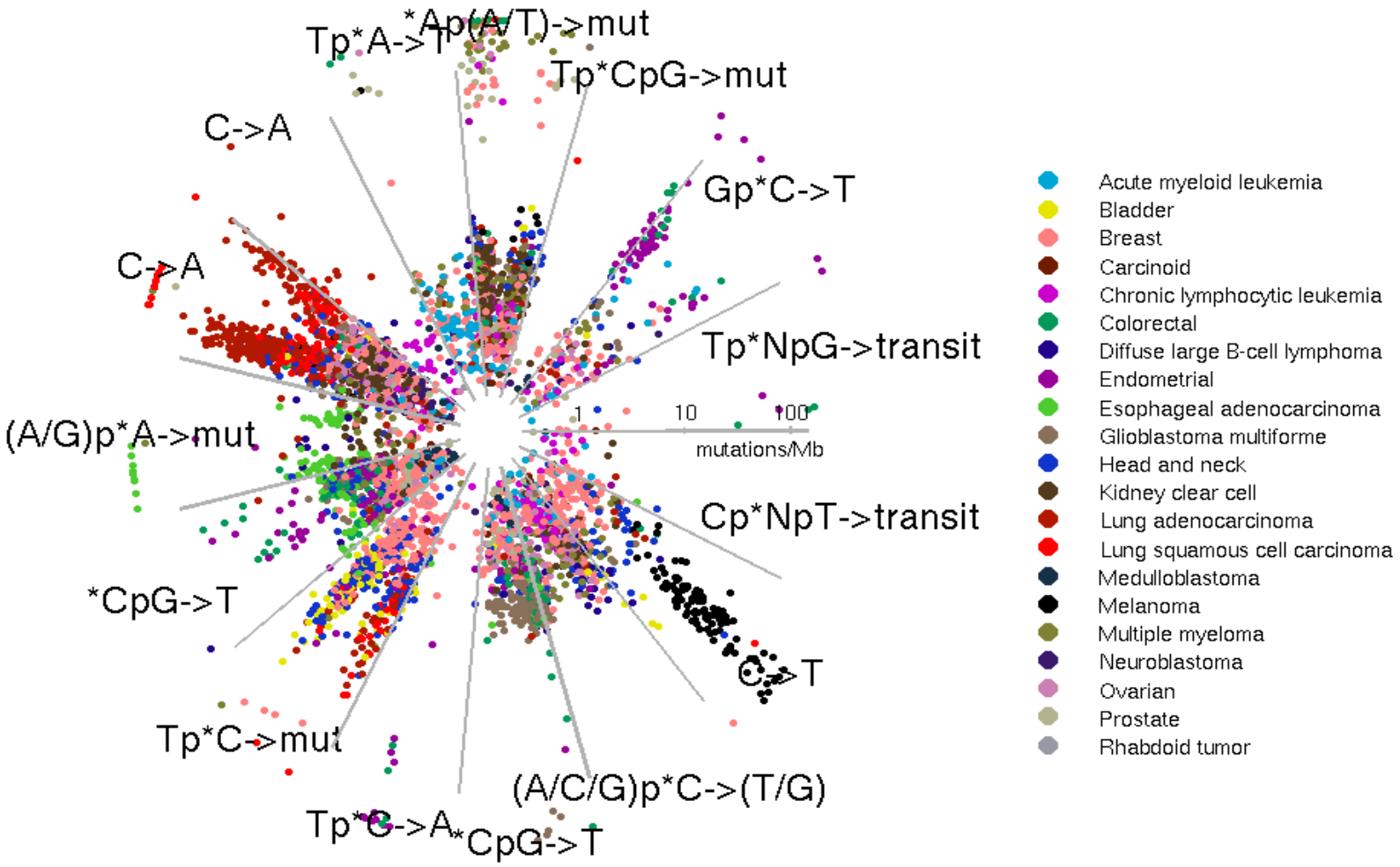
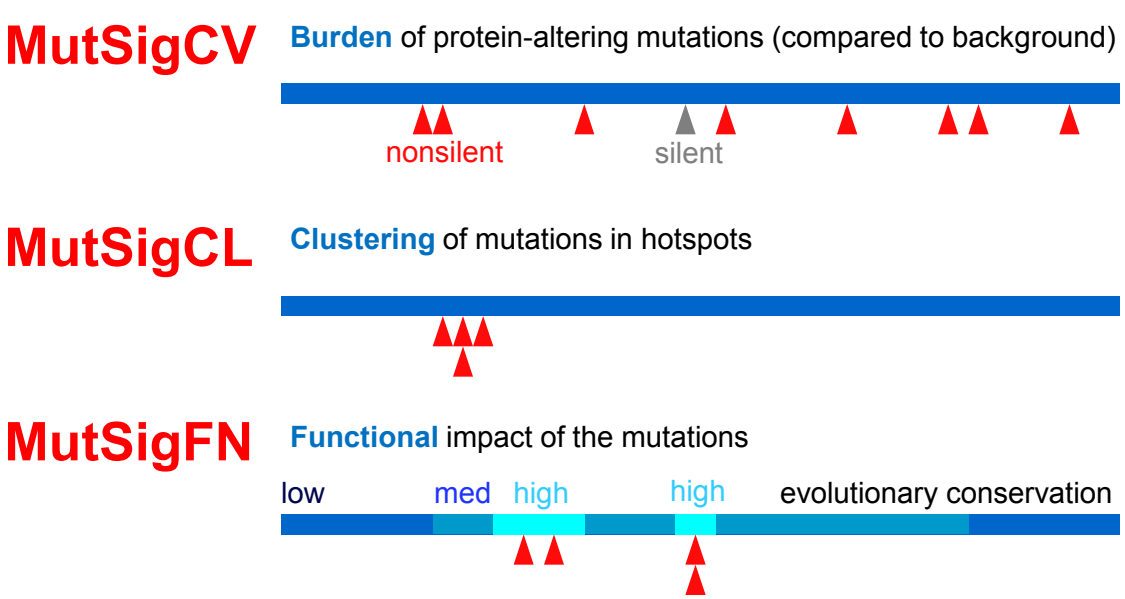


Figure S2. Radial plot representing mutational processes across 4,742 tumor-normal (TN) pairs from 21 tumor types, as in Figure 2 from Lawrence et al. *Nature* (2013). The angular space is compartmentalized into the fifteen different factors discovered by NMF. The distance from the center represents the total mutation frequency. Different tumor types segregate into different compartments based on their mutation spectra. Notable examples are: lung adenocarcinoma and lung squamous carcinoma (red; ~10 o'clock position), melanoma (black; ~4 o'clock position), esophageal (light green), colorectal (dark green), and endometrial cancer (purple) (~8-9 o'clock position), samples harboring mutations of the HPV or APOBEC signature (bladder, and some lung and head and neck samples, marked in yellow, red, and blue respectively; ~7 o'clock position), and AML samples showing a Tp*A->T signature, ~11 o'clock position.

Supplementary Figure 3

a



b

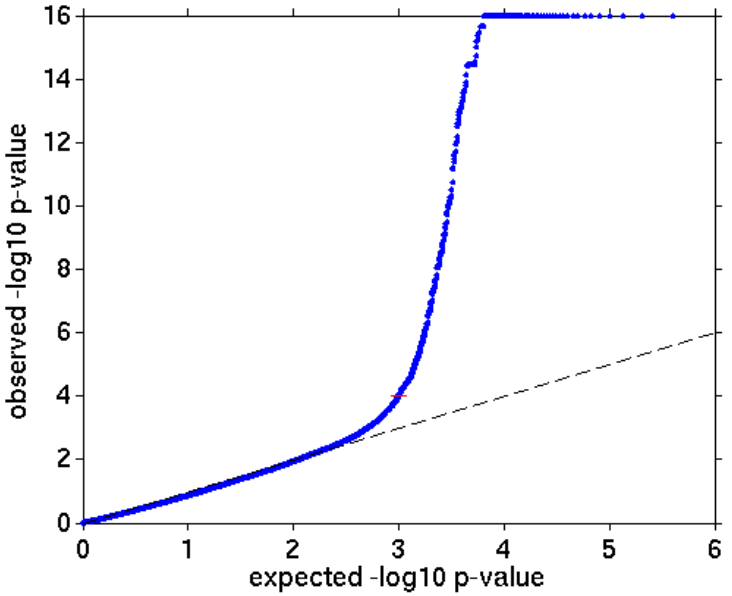
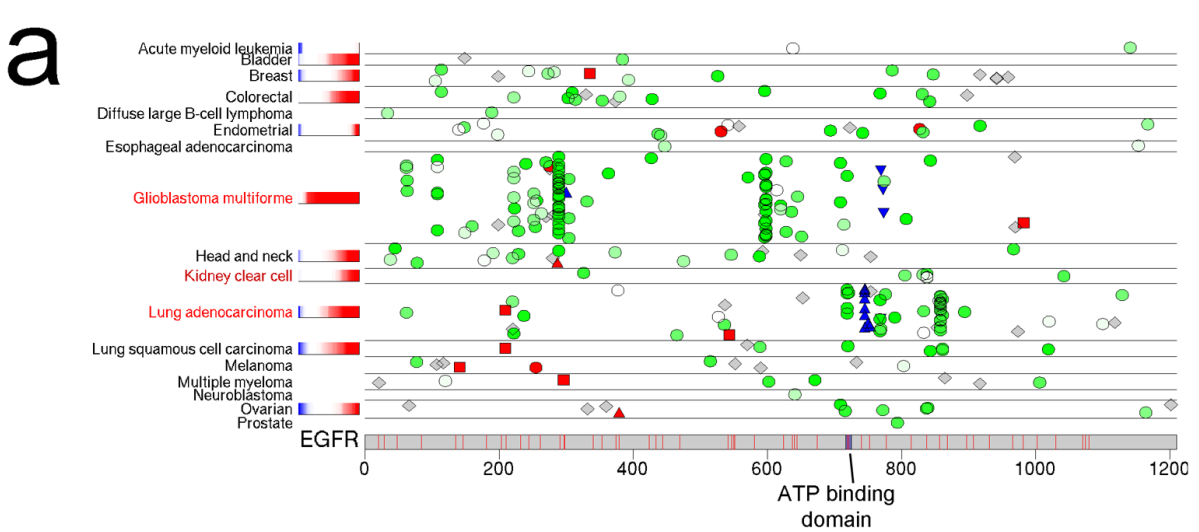
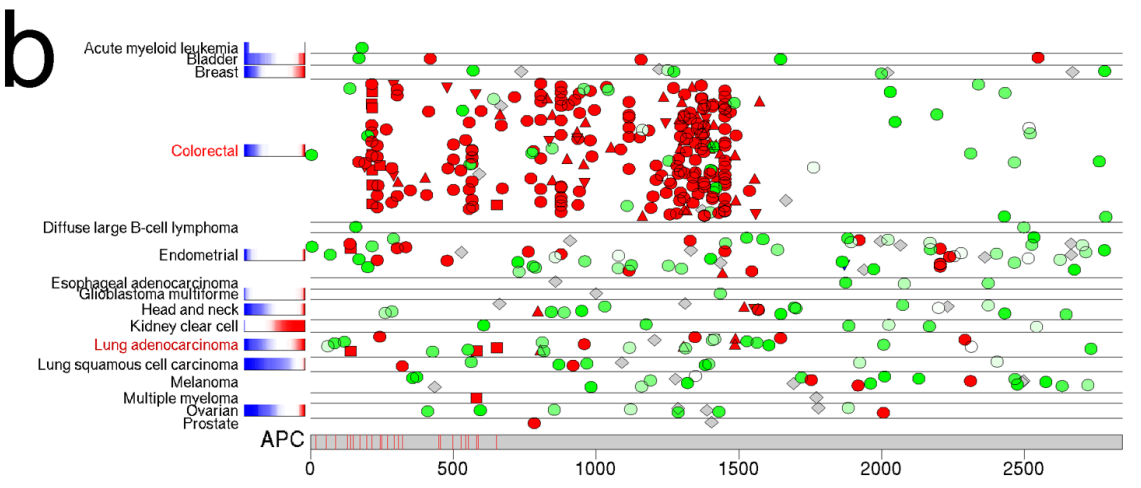
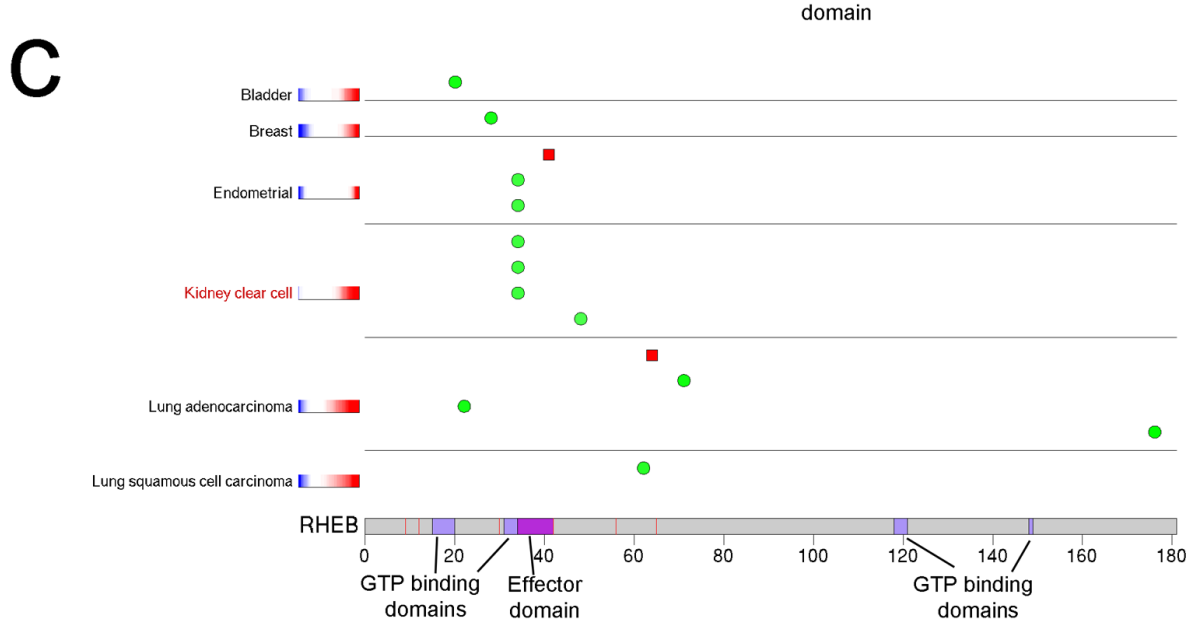


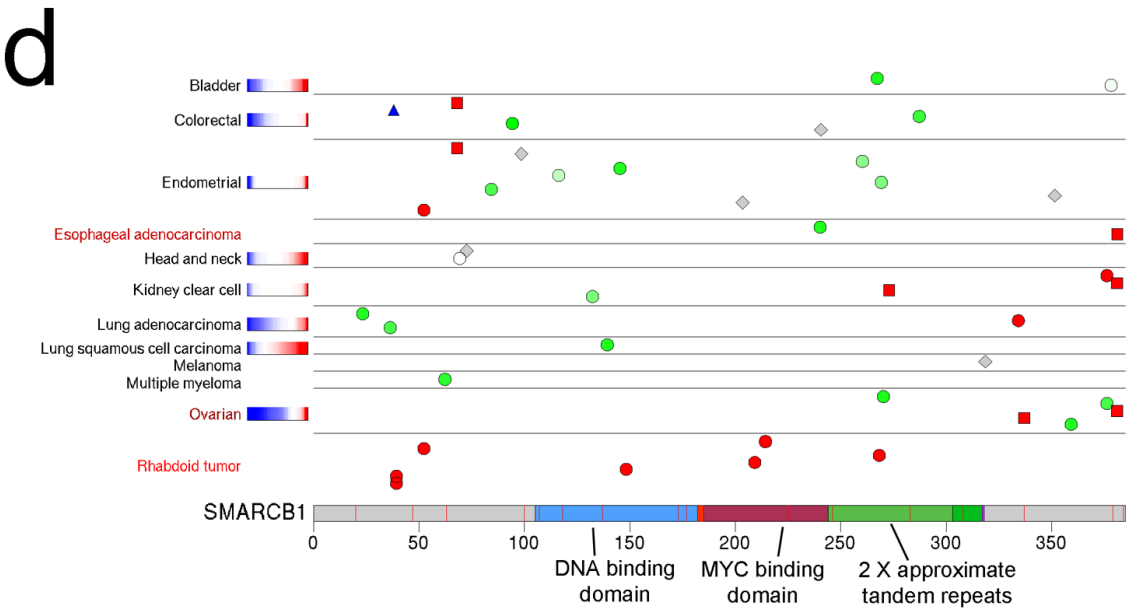
Figure S3. **a.** Three statistical tests were used to detect cancer genes: mutation burden test (calculated by MutSigCV); mutation clustering (calculated by MutSigC); and enrichment in functional sites (calculated by MutSigFN). **b.** Q-Q plot of the p-values obtained after combining the three tests, when applied on the combined set of 4,742 TN pairs.

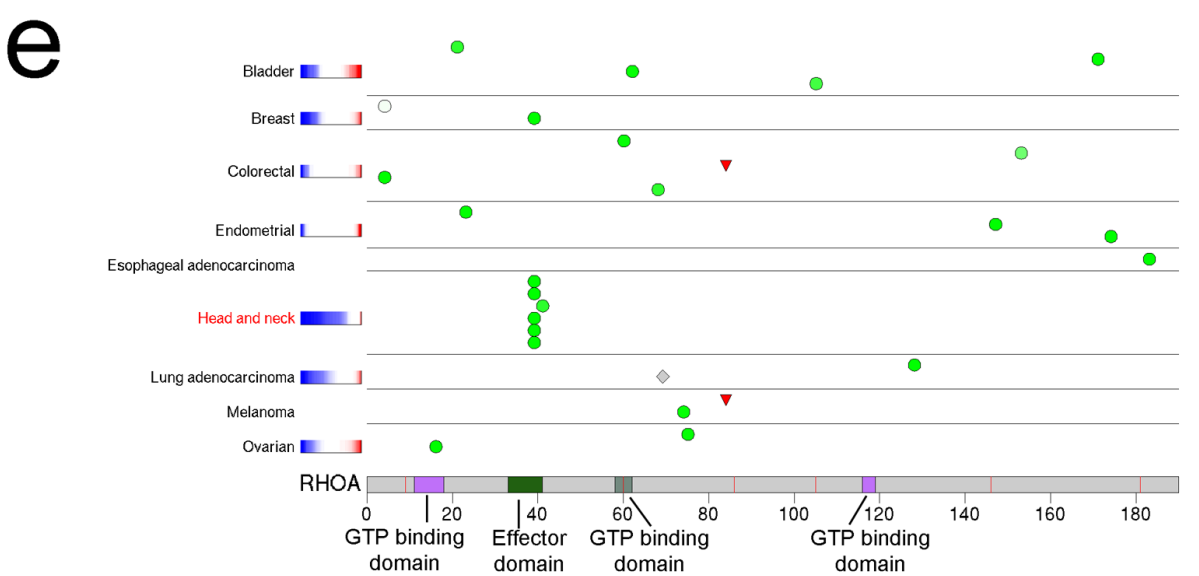
Supplementary Figure 4











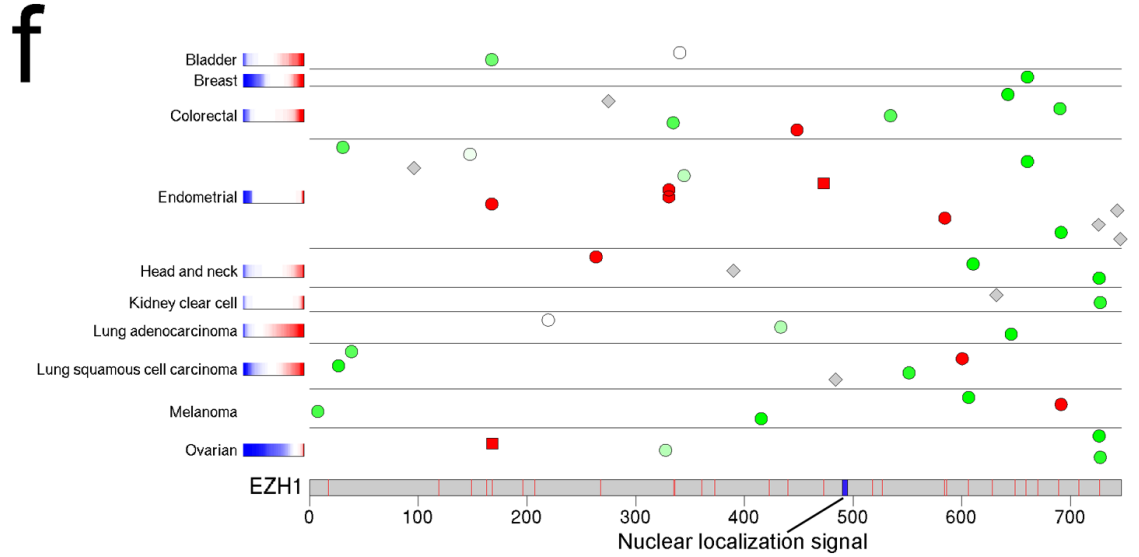


Figure S4. Mutation patterns for six known and novel cancer genes. Similar diagrams for all genes are available at http://www.tumorportal.org. **a**. *EGFR* shows distinctive tumor-type-specific concentrations of mutations at different regions of the gene. **b**. *APC* shows mutation throughout the gene in many tumor types, but colorectal cancer shows a distinctive concentration of truncating mutations in the N-terminal half of the protein. **c**. *RHEB*, which encodes a small GTPase in the *RAS* superfamily, shows a mutational hotspot in the effector domain. **d**. *SMARCB1*, which is significant based on its mutations in rhabdoid tumor and esophageal adenocarcinoma, also has truncating mutations in several other tumor types. **e**. *RHOA*, analogously to *RHEB*, encodes a small GTPase and shows a mutational hotspot in the effector domain. **f**. *EZH1*, which encodes a histone methyltransferase, does not achieve significance in any one tumor type, but does so when analyzing the cohort as a whole.

Supplementary Figure 5

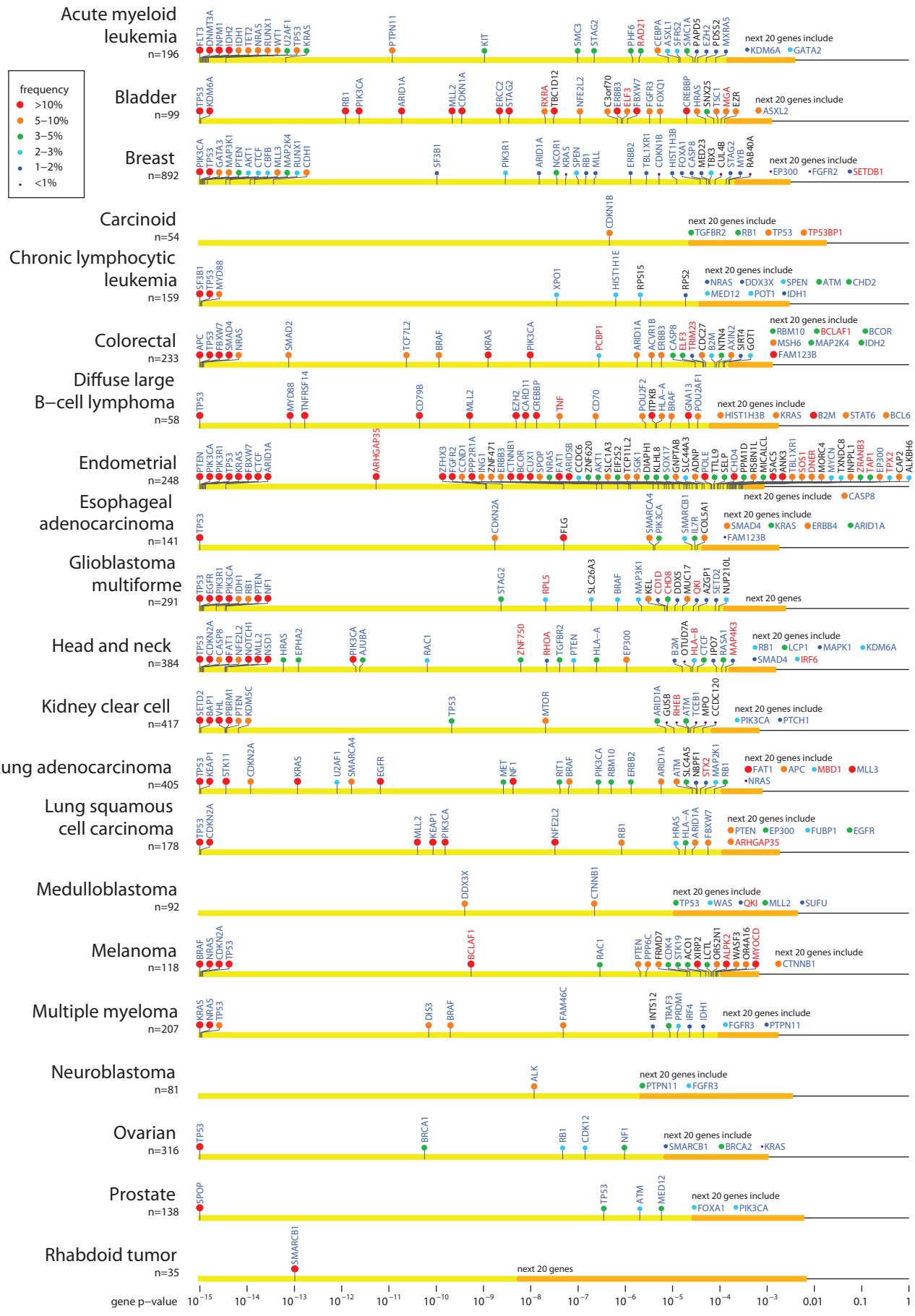


Figure S5. Cancer genes in 21 tumor types. For each tumor type, genes are arrayed on the horizontal line according to p-value (combined value for three tests in MutSig). Yellow region contains genes that achieve FDR q≤0.1. Orange interval contains p-values for the next 20 genes. The color of the gene's name indicates whether the gene is a known cancer gene (blue), a novel gene with clear connection to cancer (red; discussed in text), or an additional novel gene (black). The color of the circle associated with each gene name indicates the frequency (percent of patients carrying non-silent somatic mutations) in the corresponding tumor type. Total number of samples analyzed (n) is indicated beneath the name of each tumor type.

Supplementary Figure 6

a

b

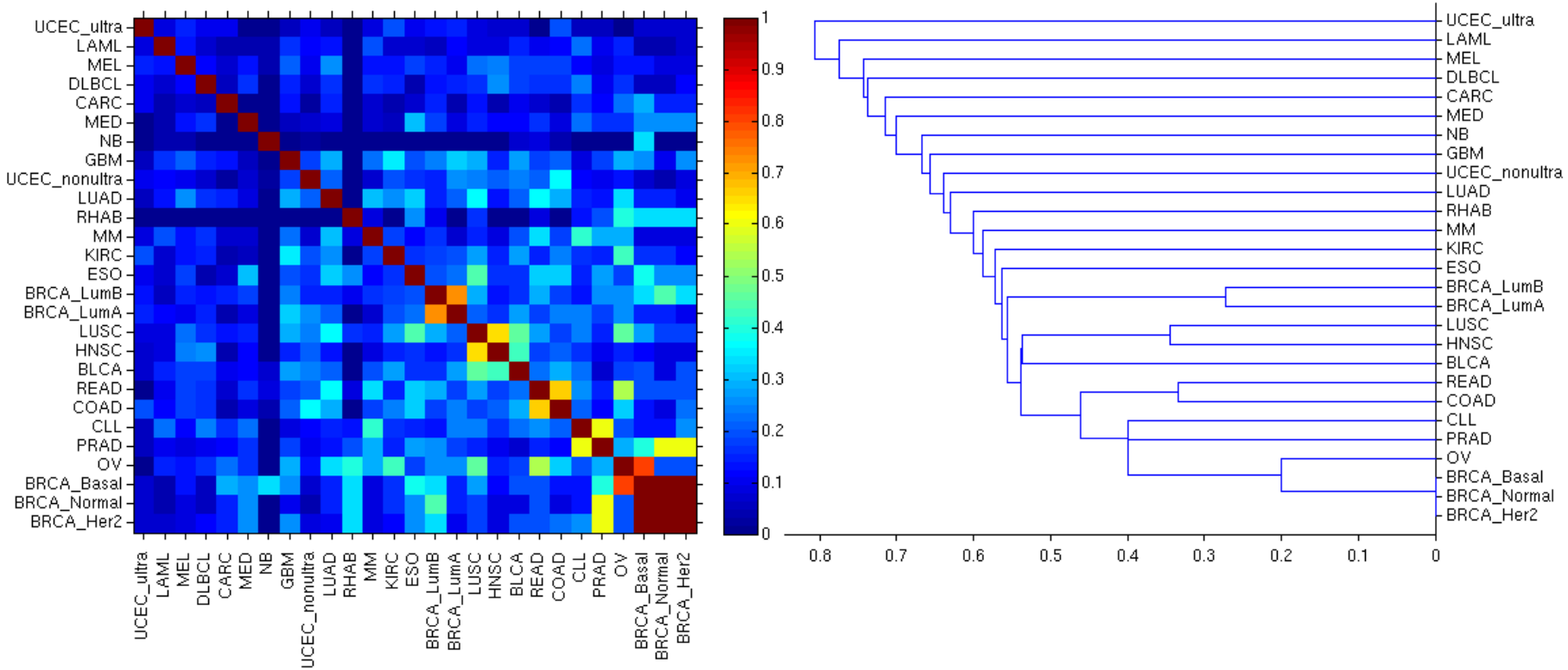


Figure S6. **a.** Heatmap showing the similarity between each pair of tumor types. Similarity scores were defined for each pair of tumor types as the number of genes significantly mutated in *both* tumor types, divided by the number of genes significantly mutated in *either* tumor type. Heatmap shows the color scale from blue (similarity=0) to red (similarity=1). **b.** Dendrogram based on pairwise distances between tumor types, equal to 1 minus their similarity score.

Supplementary Figure 7

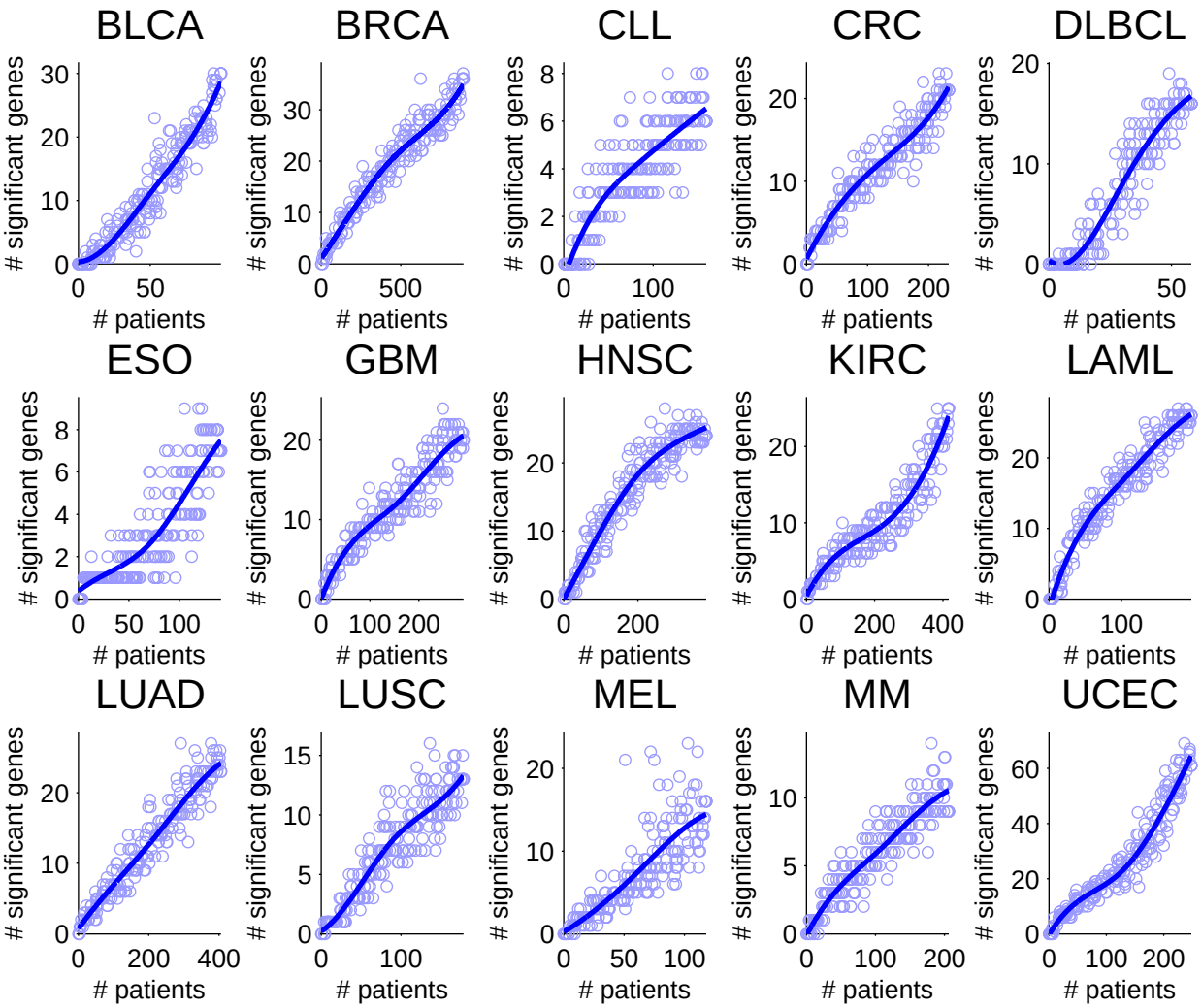


Figure S7. Down-sampling analysis shows that candidate cancer gene discovery is continuing to accumulate within each tumor type. Each point represents a random subsampling of the patients. X-axis indicates the number of patients in the sample; Y-axis indicates the number of significant genes in the sample. Blue line is a smoothed fit.

Supplementary Figure 8

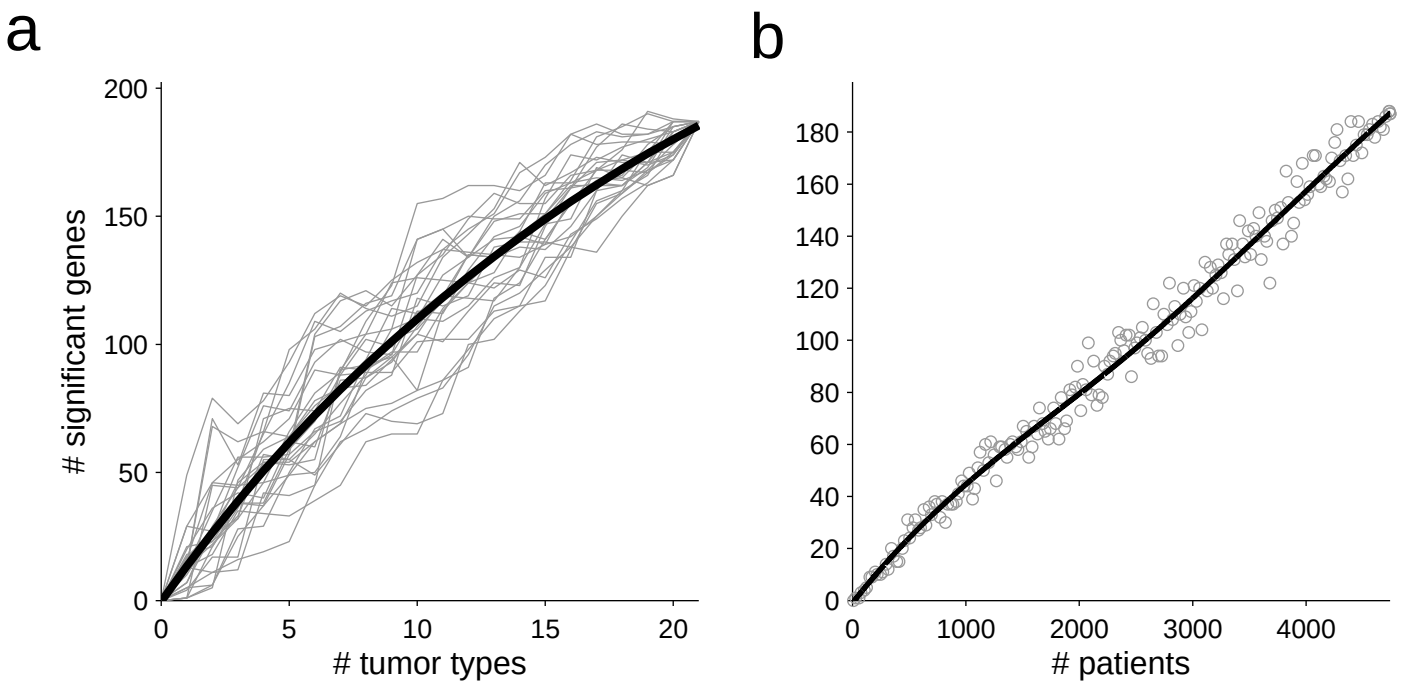


Figure S8. Down-sampling of combined cohort, with FDR correction applied by subtracting the expected number of false positives at each downsampling point.

Supplementary Figure 9

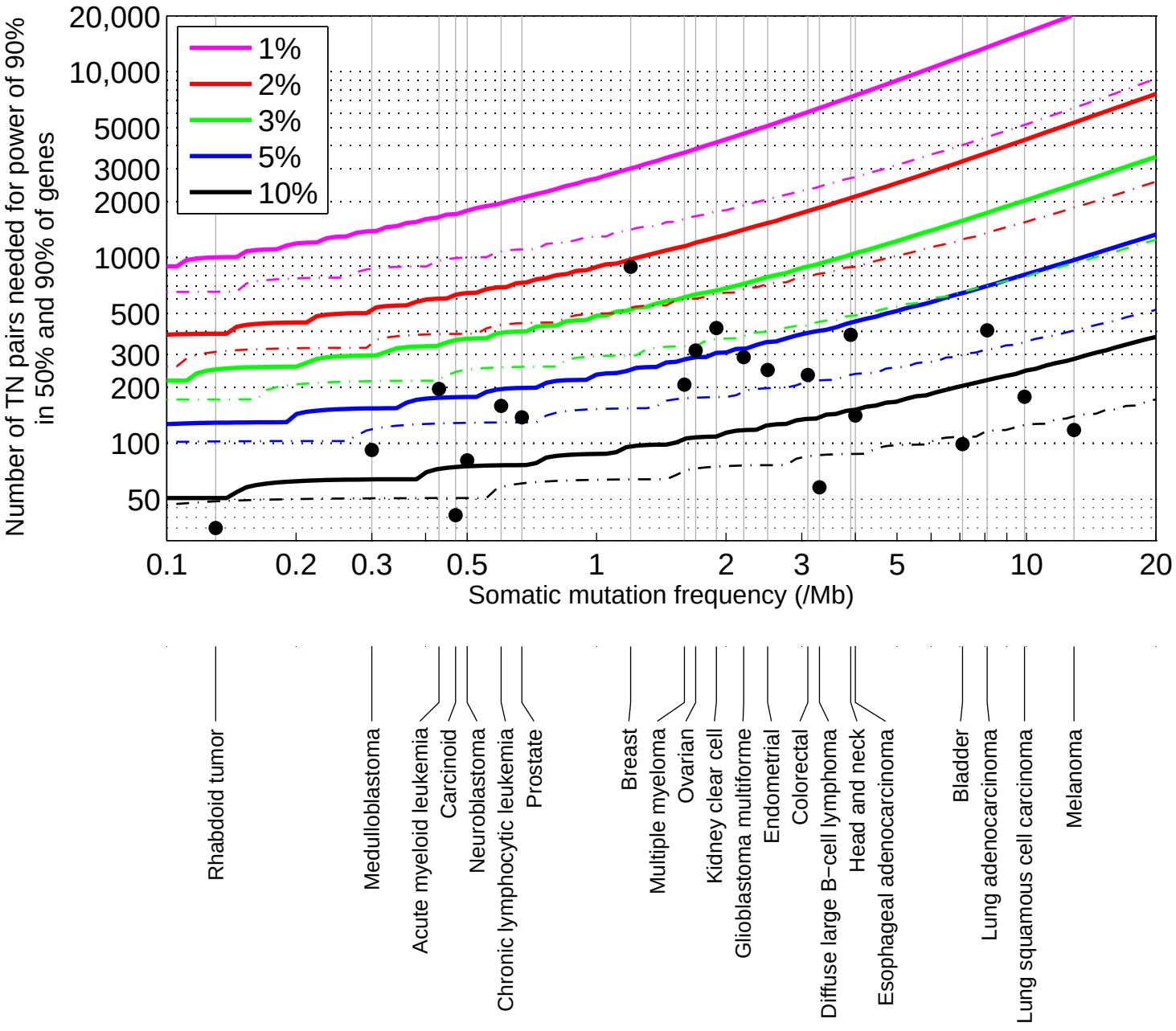


Figure S9. Number of tumor-normal pairs required to achieve 90% power for 90% of genes (solid) and 90% power for the median gene (dashed). Other details of figure are exactly as described in Figure 5 of the main text.

Supplementary Tables

Table S1. List of source datasets analyzed in this work, and references to the corresponding publications.

Table S2. The 260 significantly mutated cancer genes found by analysis with the MutSig suite. The first set of columns (A-B) list the gene symbol and its full name. The second set of columns (D-I) indicate the membership of the genes in various sets. Column D contains a “1” if the gene is listed in the Cancer Gene Census (v65) as being recurrently somatically mutated. Column E contains a “1” if the gene was not previously reported to be somatically mutated in cancer at a significant rate, either in the CGC or in a recent publication (Supplementary Table 3). Columns F and G report whether the gene is a member of the Cancer5000 and Cancer5000-S sets, respectively. Column H contains a "1" for those thirty genes that were identified as significantly mutated only when analyzing the combined cohort of 4742 samples. Column I contains a "1" for the 33 novel genes with clear and consistent connections to cancer hallmarks that were discussed in the text. The remaining sets of columns diverge in the four tabs of the spreadsheet. The first tab ("Individual q-values") reports the q-values from analyzing each tumor type separately and correcting for only the 18,388 genes—this corresponds to the Cancer5000 set. The second tab ("q-values when testing 400K hyp.") reports the q-values after the more stringent correction for 18,388 genes x 22 analyses—this corresponds to the Cancer5000-S set. The third tab ("q-value from RHT") reports the q-values obtained when performing the Restricted Hypothesis Testing (RHT) as described in the text. The fourth tab ("Frac. of non-silent mutations") lists the number (and percentage) of patients of each tumor type carrying a non-silent mutation in the gene. The fifth and last tab (“P-values”) lists the raw p-values obtained from each statistical test (MutSigCV, MutSigCL, MutSigFN), as well as each pairwise combination, , and the final overall (all three tests) combined p-value, for each gene.

Table S3. List of the 21 tumor types studied, and the significantly mutated genes found by the MutSig suite in each tumor type. Column D lists the genes that were found to be significant when directly analyzing the full list of 18,388 genes. Column E lists the additional genes found to be significant when performing the Restricted Hypothesis Testing (RHT) as described in the text. In parentheses after each gene name are two numbers: the number and percentage of patients carrying a non-silent mutation in the gene.

Table S4. List of references reporting the identification of candidate cancer genes. This table lists the genes that are not yet listed in the Cancer Gene Census (CGC) but have been reported in previous publications.

Table S5. List of references to biological literature supporting the 33 novel candidate cancer genes with clear and compelling connections to cancer biology.

Table S6. Summary of the analysis comparing the performance of each of the three MutSig metrics separately, in pairwise combinations, and all three combined as in the main analysis. The table lists the number of candidate cancer genes in the Cancer 5000 list and the Cancer 5000-S list, as well as the number of genes identified from the list of CGC genes that contain somatic mutations in tumor types we studied, and the number of genes identified from the list of 33 novel candidate cancer genes with compelling connections to cancer biology.