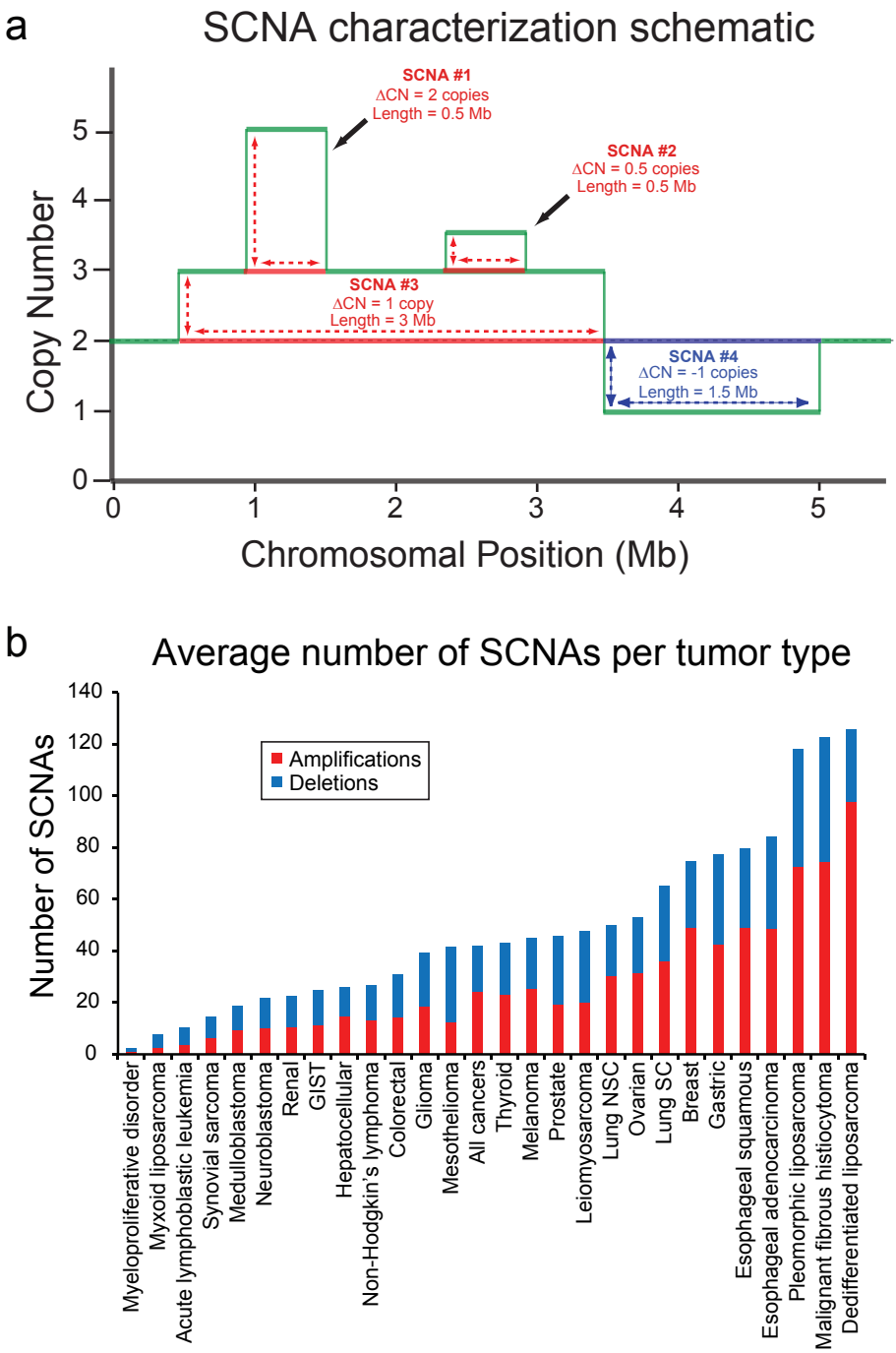
Supplementary Information

The landscape of somatic copy-number alteration across human cancers

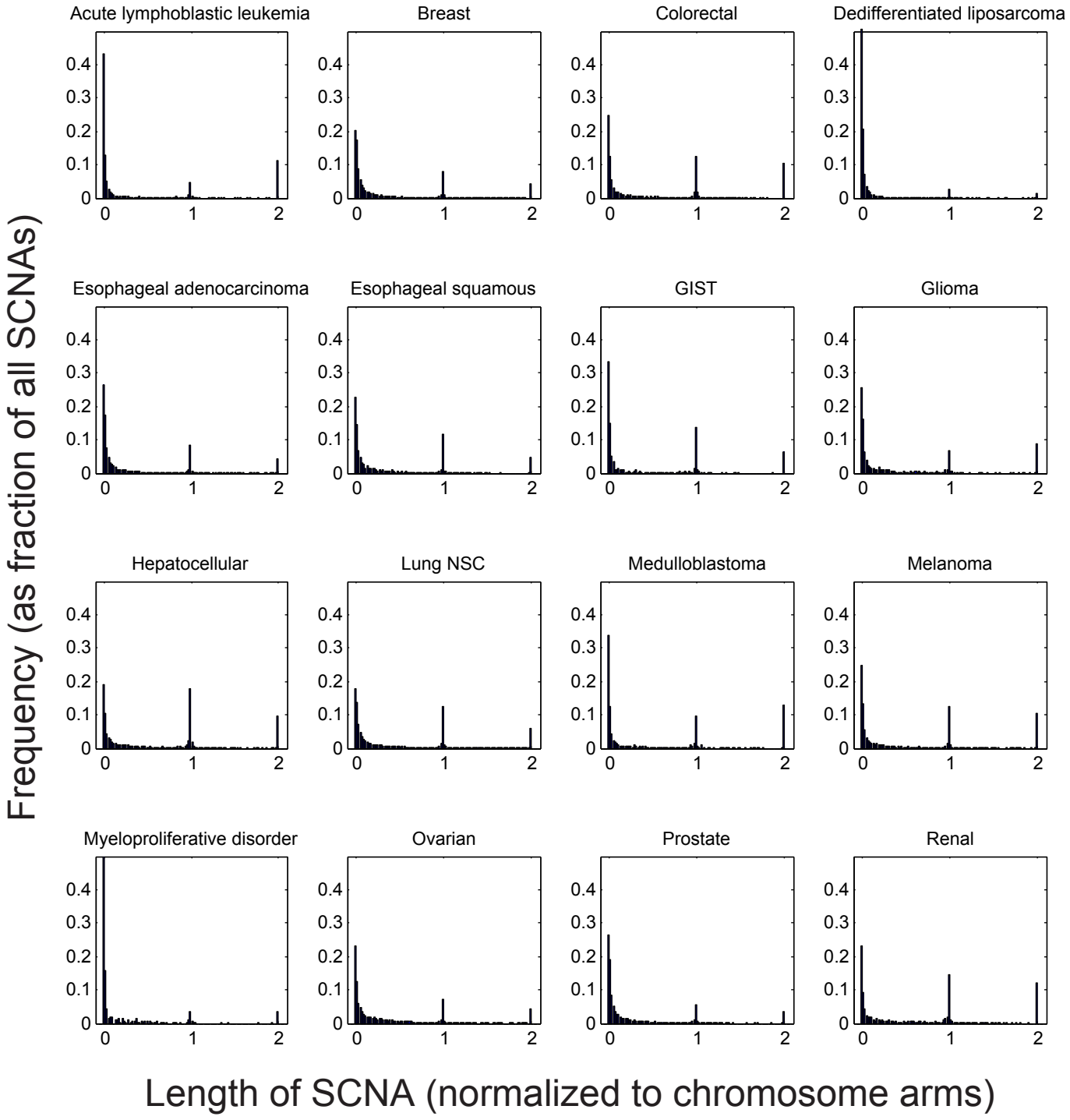
Rameen Beroukhim*, Craig H. Mermel*, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S. Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T. Mc Henry, Reid M. Pinchback, Azra H. Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S. Lawrence, Barbara A. Weir, Kumiko E. Tanaka, Derek Y. Chiang, Adam J. Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J. Kaye, Hidefumi Sasaki, Joel E. Tepper, Jonathan A. Fletcher, Josep Tabernero, Jose Baselga, Ming-Sound Tsao, Francesca DeMichelis, Mark A. Rubin, Pasi A. Janne, Mark J. Daly, Carmelo Nucera, Ross L. Levine, Benjamin L. Ebert, Stacey Gabriel, Anil K. Rustgi, Cristina R. Antonescu, Marc Ladanyi, Anthony Letai, Levi A. Garraway, Massimo Loda, David G. Beer, Lawrence D. True, Aikou Okamoto, Scott L. Pomeroy, Samuel Singer, Todd R. Golub, Eric S. Lander, Gad Getz, William R. Sellers, and Matthew Meyerson

* Contributed equally.

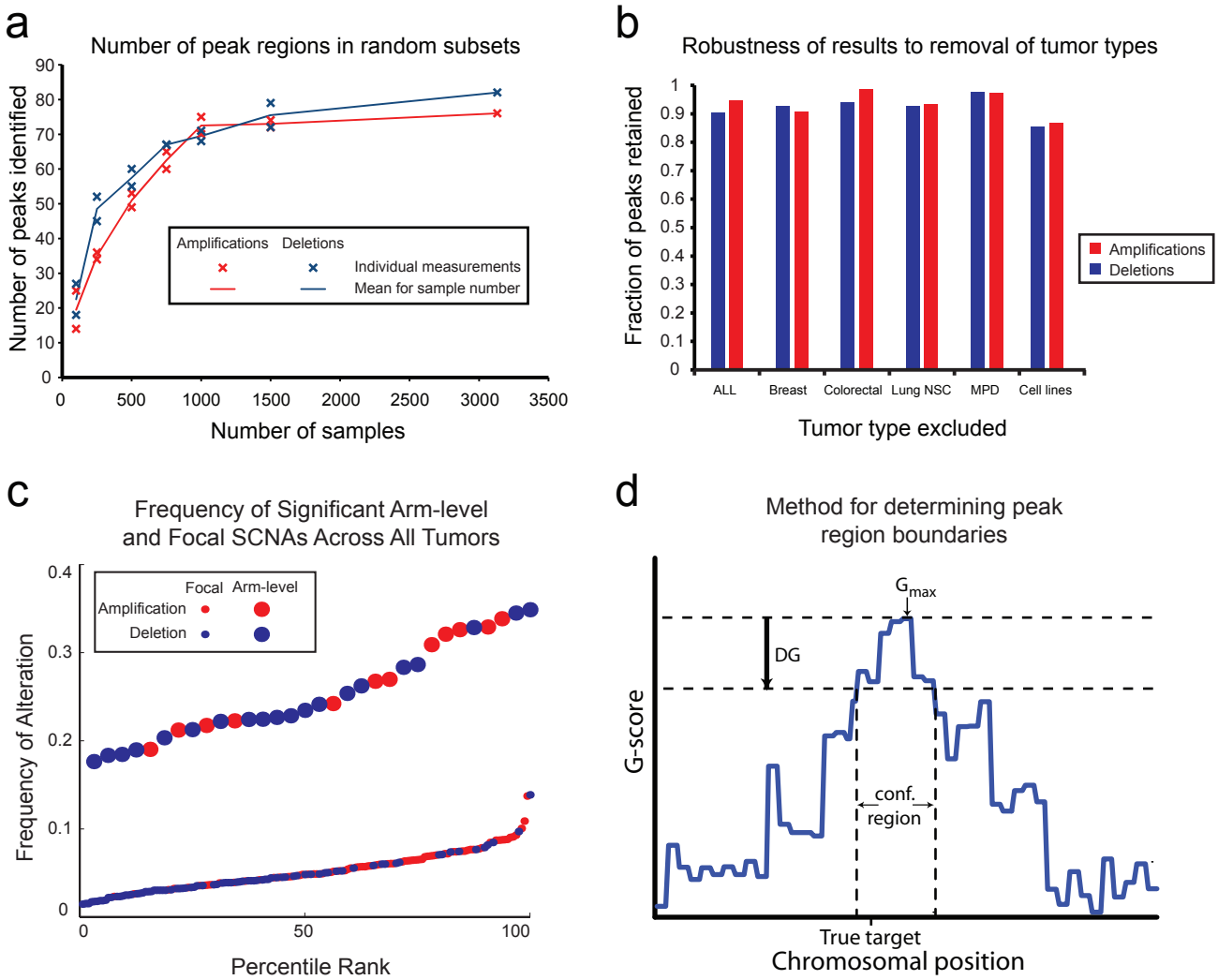
Supplementary Figures and Legends



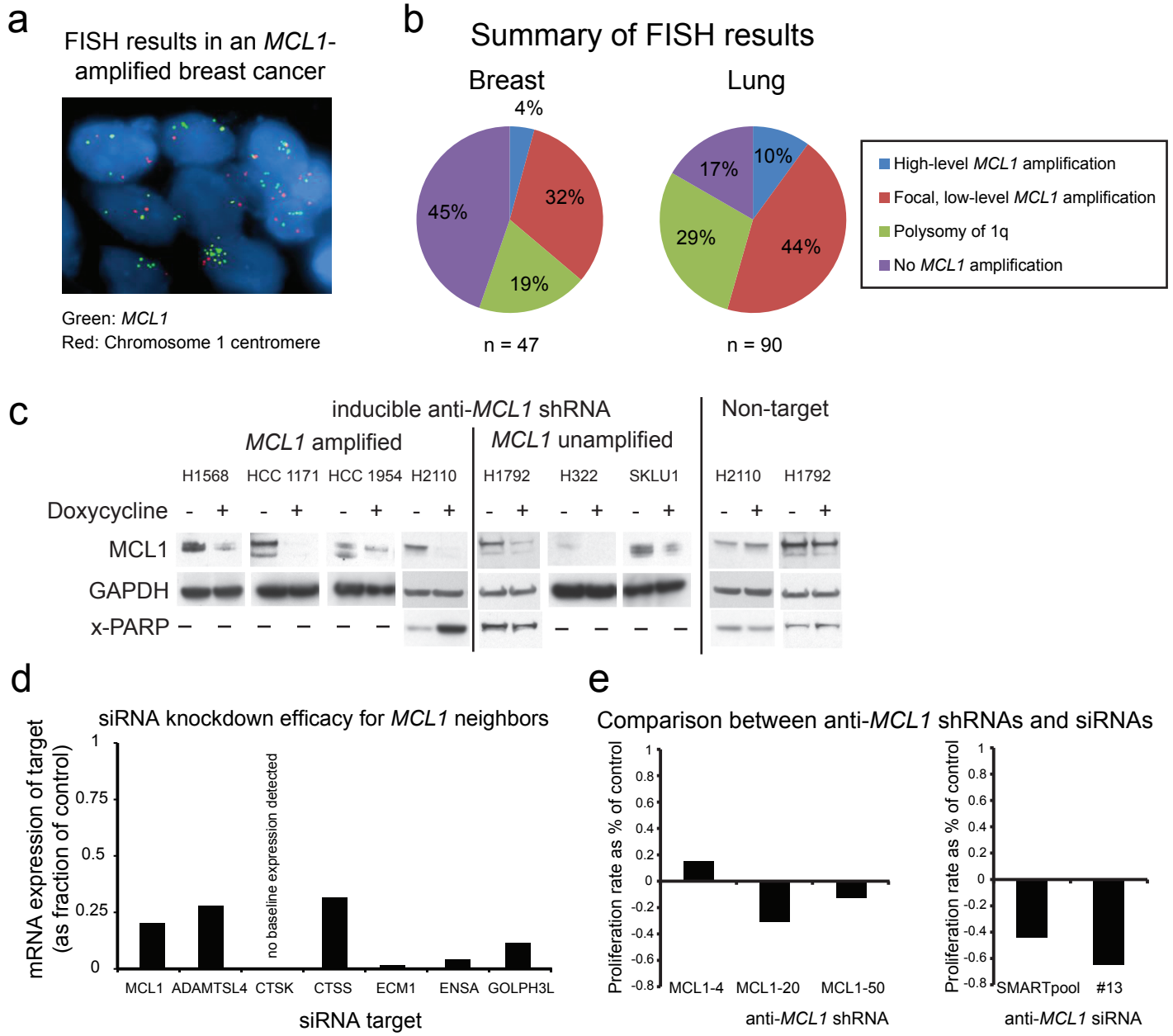
Supplementary Figure 1. Identifying individual SCNAs. a) Schematic of the method for determining SCNA length and amplitudes from complex copy-number profiles. A hypothetical copy-number profile (green) is separated into amplification (red) and deletion (blue) events as shown. b) Bar graph of average number of amplification (red) and deletion (blue) events per tumor for each cancer type with >4 samples. "All cancers" represents the entire dataset.



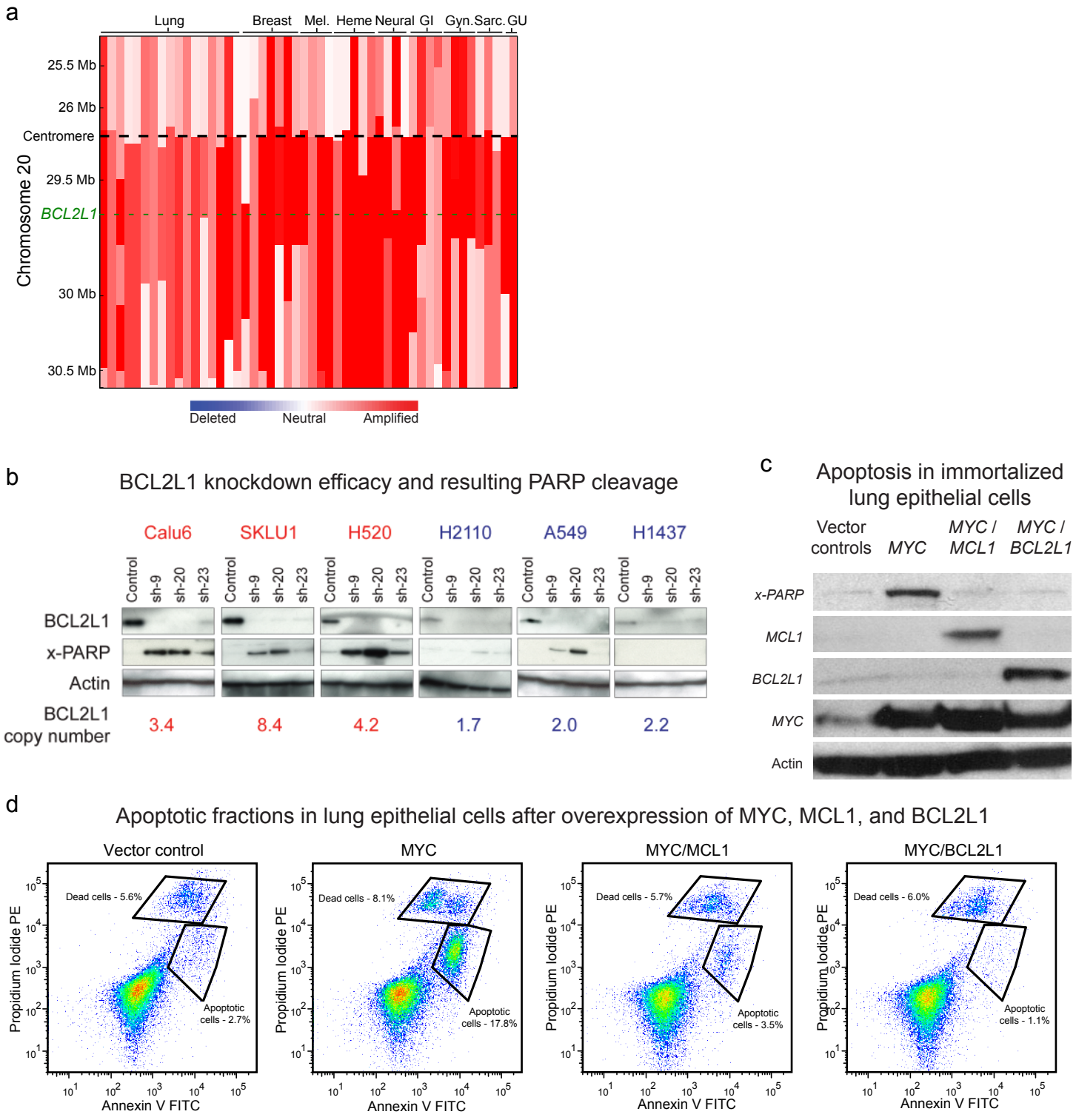
Supplementary Figure 2. Increased prevalence of arm-level SCNAs relative to focal SCNAs of nearly the same length, across the 16 tumor types with >40 samples. Data are presented as in Figure 1a.



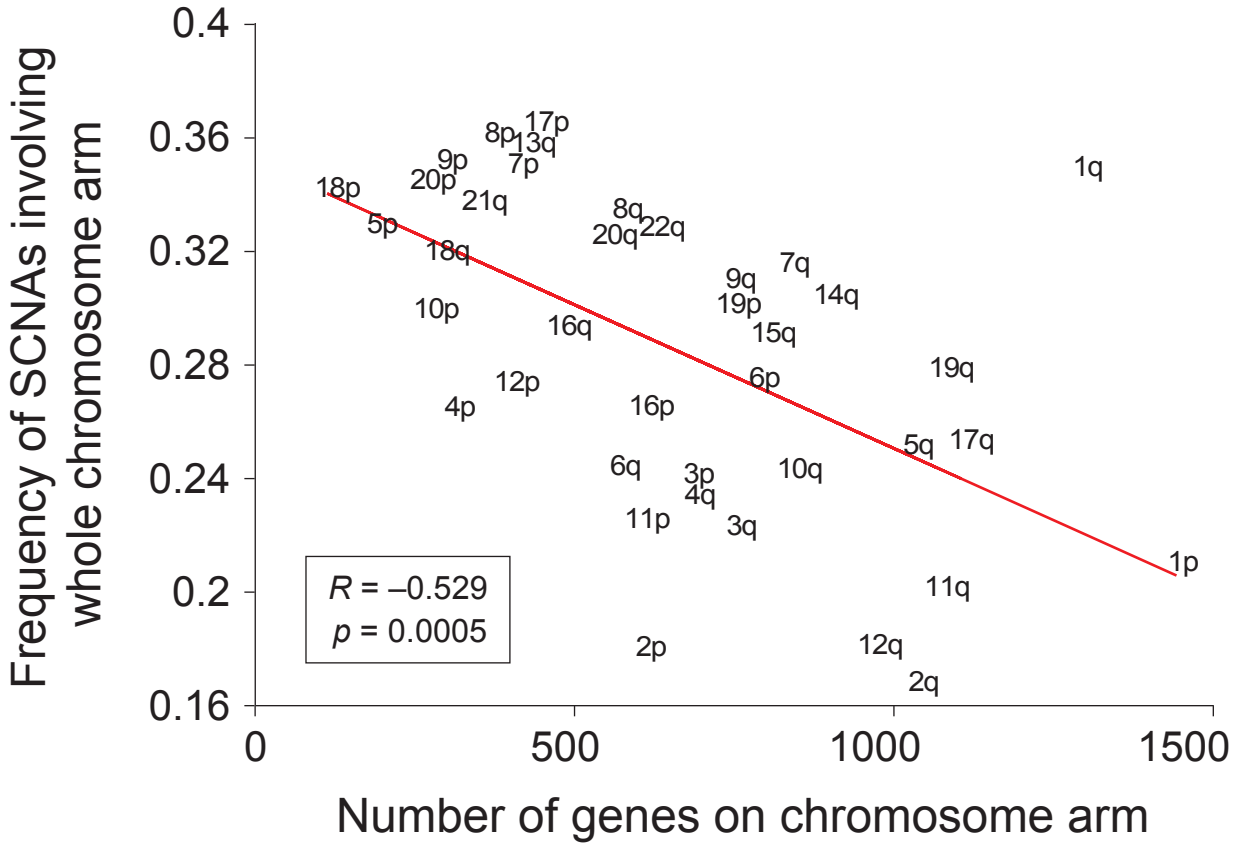
Supplementary Figure 3. Characteristics of focal SCNAs. a) The number of focal amplification (red) and deletion (blue) peaks identified using GISTIC on random subsets of the data. Crosses represent individual randomizations; lines represent averages over all randomizations for a given sample size. b) Robustness of focal SCNA analysis to removal of each of the five most represented tumor types (Lung NSC, acute lymphoblastic leukemia, breast, myeloproliferative disorder, and colorectal) or all cell lines. The fraction of the 76 amplification peaks (red) and 82 deletion peaks (blue) still identified as peak regions when each tumor type is removed is plotted. c) Frequency of significant arm-level (large circles) and focal (small dots) amplifications (red) and deletions (blue), sorted by increasing frequency. d) Method for determining confidence regions likely to include the true target of focal SCNAs. Local maxima in the G-score (G_{max}) correspond to a "minimal common region" of overlap and generally reflect the presence of nearby "target genes" whose alteration plays a role in driving cancer growth. However, the presence of technical and biological noise ("passenger SCNAs") may displace G_{max} from the true target. DG represents the maximum local variation expected in 95% of cases due to such noise. Subtracting DG from G_{max} allows us to determine a confidence region at least 95% likely to contain the gene target (details in Mermel et al, manuscript in preparation). e) Increasing sample size leads to better resolution of likely gene targets. For each of the random tumor subsets in a), we ranked peaks by q-value and computed the median number of genes in each group of 20 peaks, starting with the most significant.



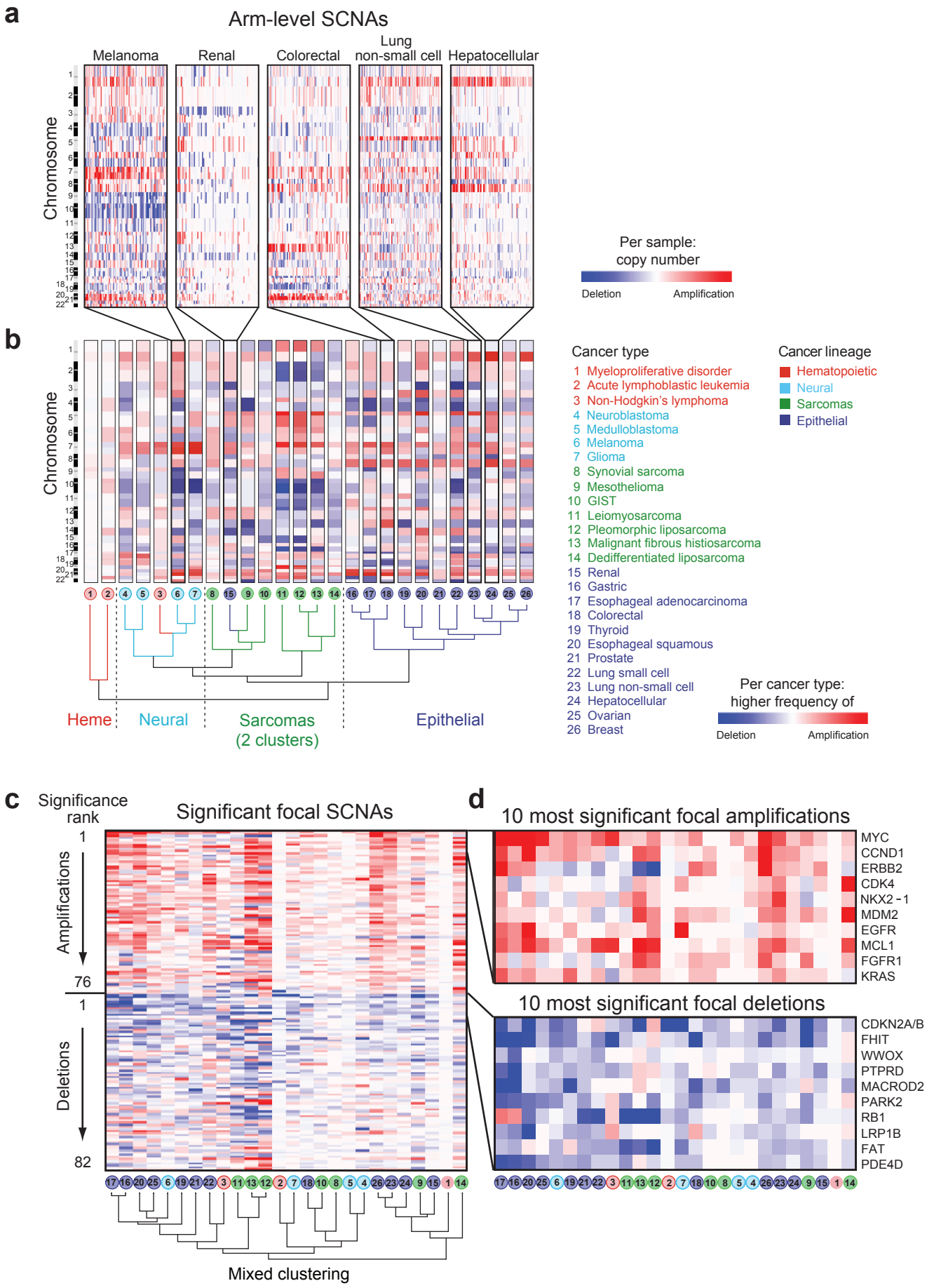
Supplementary Figure 4. The frequency and significance of *MCL1* amplification in human cancers. a) Representative FISH results showing high-level *MCL1* amplification (green signals) in the breast cancer cell line HCC1954. The chromosome 1 centromere is stained red. b) Summary of *MCL1* FISH results in a panel of 47 primary breast cancer and 90 primary lung cancer samples. High-level *MCL1* amplification (blue) was defined as *MCL1* copy number greater than 3x that of the chromosome 1 centromere; focal, low-level *MCL1* amplification (red) was defined as *MCL1* copy numbers less than this but exceeding the centromere; and polysomy of 1q (green) was defined as equal copy numbers of both *MCL1* and the chromosome 1 centromere but exceeding the number of copies of the chromosome 11 centromere. c) Efficacy of doxycycline-inducible *MCL1* knock-down. Western blot analysis of MCL1 protein levels in the 7 cell lines tested in Figure 3c before and after induction of inducible anti-*MCL1* shRNA or non-targeting control. GAPDH was used as a protein loading control. For H2110 (*MCL1* amplified) and H1792 (*MCL1* unamplified), cleaved PARP levels were also determined before and after induced expression of anti-*MCL1* and non-targeting shRNAs. d) siRNA knock-down efficacy for *MCL1* and neighboring genes. Quantitative RT-PCR was used to measure mRNA transcript expression before and after introduction of siRNAs against the 7 non-provisional genes in the *MCL1* peak in H2110 cells (as shown in Figure 3d). The expression of each transcript after knock-down is graphed as a fraction of the expression in mock-treated cell lines. No expression of *CTSK* was detected in mock-transfected H2110 cells. e) Comparison of the effects of multiple anti-*MCL1* shRNAs and siRNAs in H2110 cells. H2110 cells were infected with three independent shRNA constructs against *MCL1*, and treated with an anti-*MCL1* Dharmacon siRNA SMART pool and a single siRNA sequence from that pool. For each treatment, the change in cell number (proliferation rate) over 48 hours (as measured by CellTiterGlo, Promega), relative to non-targeting control, is shown.



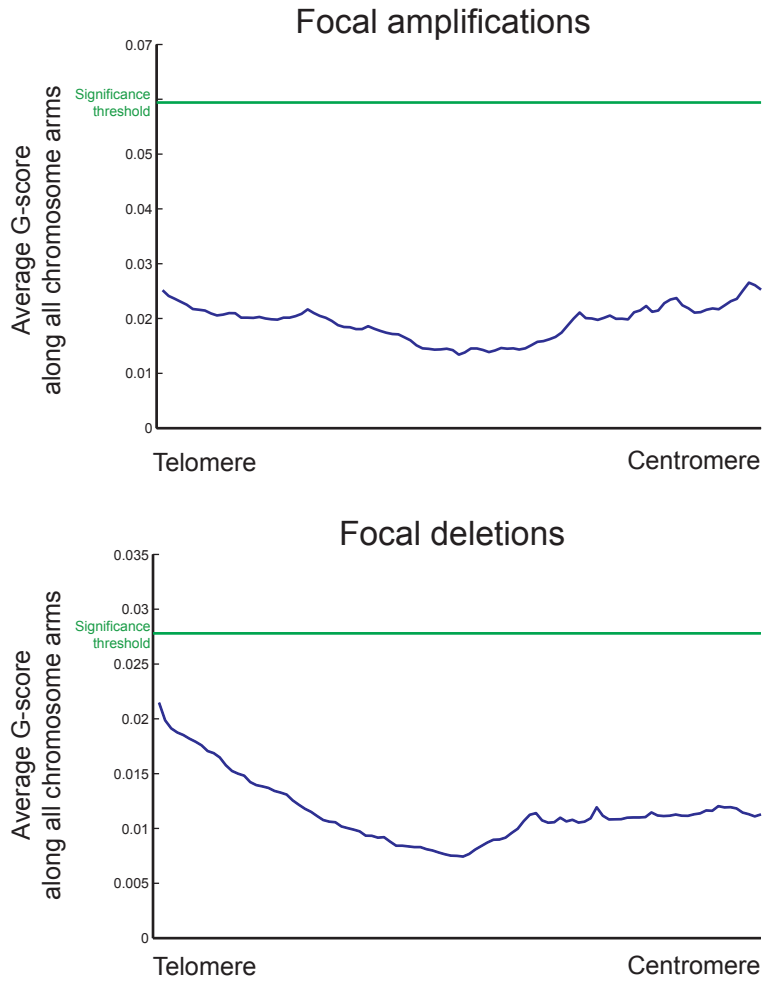
Supplementary Figure 5. Supporting data for *BCL2L1* and *MCL1* experiments. a) Segmented copy-number profiles among 50 tumors of various lineages (shown across the top) with focal amplification of *BCL2L1* are displayed for the region around *BCL2L1* (genomic locations are indicated on the left; distances are proportional to the number of SNP array markers mapping to the region). b) Efficacy of *BCL2L1* knock-down in cell lines. Western blot analysis of BCL2L1 and cleaved PARP protein levels in 6 cell lines tested in figure 3e after infection with anti-*BCL2L1* shRNA or non-targeting control. Actin was used as a protein loading control. c) Increased levels of apoptosis induced by *MYC* expression in immortalized lung epithelial cells [1] are reversed by expression of *MCL1* or *BCL2L1*. Cells transduced with viruses expressing *MYC*, *MYC* and *MCL1*, *MYC* and *BCL2L1*, or vector controls were cultured for 24 hours. Adherent and floating cells were pooled and levels of cleaved PARP, MCL1, BCL2L1, MYC, and actin (as loading control) were assessed by immunoblot. d) In a separate experiment, these cells were washed, stained with anti-Annexin antibody (BioVision) and propidium iodide (Sigma), and analyzed by flow cytometry.



Supplementary Figure 6. The frequency of arm-level SCNAs is negatively correlated with the number of genes covered. The red line represents a sum of least squares fit to the data.


a
Arm-level SCNAs
Melanoma
Renal
Colorectal
Lung non-small cell
Hepatocellular
Chromosome
Per sample: copy number
Deletion
Amplification
b
Chromosome
Heme
Neural
Sarcomas (2 clusters)
Epithelial
Cancer type
1 Myeloproliferative disorder
2 Acute lymphoblastic leukemia
3 Non-Hodgkin's lymphoma
4 Neuroblastoma
5 Medulloblastoma
6 Melanoma
7 Glioma
8 Synovial sarcoma
9 Mesothelioma
10 GIST
11 Leiomyosarcoma
12 Pleomorphic liposarcoma
13 Malignant fibrous histiosarcoma
14 Dedifferentiated liposarcoma
15 Renal
16 Gastric
17 Esophageal adenocarcinoma
18 Colorectal
19 Thyroid
20 Esophageal squamous
21 Prostate
22 Lung small cell
23 Lung non-small cell
24 Hepatocellular
25 Ovarian
26 Breast
Cancer lineage
Hematopoietic
Neural
Sarcomas
Epithelial
Per cancer type: higher frequency of
Deletion
Amplification
c
Significance rank
Significant focal SCNAs
Amplifications
1
76
1
Deletions
82
Mixed clustering
d
10 most significant focal amplifications
MYC
CCND1
ERBB2
CDK4
NKX2-1
MDM2
EGFR
MCL1
FGFR1
KRAS
10 most significant focal deletions
CDKN2A/B
FHIT
WWOX
PTPRD
MACROD2
PARK2
RB1
LRP1B
FAT
PDE4D

Supplementary Figure 7. Clustering of tumor types by arm-level and focal SCNAs. a) Specific arm-level SCNAs can reach high frequencies among individual cancer types. Copy-number profiles (only arm-level SCNAs were included in this view) are displayed for samples selected among five tumor types (arranged across the x-axis) across all autosomes (positions indicated along the y-axis). Red and blue represent gains and losses, respectively. b) Arm-level SCNAs distribute across cancer types by developmental lineage. For each of the 26 cancer types studied, each chromosome arm was assigned an excess amplification score representing the frequency of arm-level gain minus the frequency of arm-level loss. Positive and negative scores are displayed in red and blue, respectively. Tumor types are arranged along the x-axis according to the results of unsupervised hierarchical clustering (see Supplementary Methods) of these scores (dendrogram is on the bottom). Developmental lineage reflects the ICD-O classification scheme except for melanoma, which we designated as of neural lineage due to its derivation from the neural crest. c) All 158 significant focal events (arranged on y-axis according to significance of amplification, followed by significance of deletion) across the 26 cancer types studied in part b), arranged along the x-axis according to the results of unsupervised hierarchical clustering of excess amplification scores (dendrogram is on the bottom). d) Excess amplification scores are displayed for the 10 most significant focal amplifications (upper panel) and deletions (lower panel), ranked top to bottom and denoted by putative target genes from each region. The ordering of the tumor types along the x-axis is the same as in part c).



Supplementary Figure 8. Average level of focal amplification (top) and deletion (bottom) along a chromosome arm. Each chromosome arm was rescaled to a common length, and the average G-score (in blue; see Supplementary Methods) across all chromosome arms and samples was calculated as a function of distance from the telomere. For comparison, the green line corresponds to our False Discovery Rate q-value threshold of 0.25; G-scores above this line are considered significant. The variations observed in average G-score along the chromosome arm are small compared to this threshold. However, there is a tendency for telomeric regions to be focally deleted. As a result, telomeric deletions have to rise less above the average level to attain significance.

Supplementary Methods

1. DNA isolation and hybridization to arrays

Previously published SNP array datasets were generated as described (Barretina, in review) [1,2,3,4,5,6,7,8,9,10,11,12,13]. For unpublished data, DNA was obtained from cell line pellets or tumors frozen at the time of surgical dissection and maintained at -80C until use, with the exception of 11 gliomas from which sufficiently high-quality DNA could be obtained from paraffin-embedded samples [14]. The majority of tumors were obtained at primary surgery, with the exceptions of 27 prostate tumors obtained through rapid autopsy programs at the Universities of Washington [15] and Michigan [16]. Each sample was genotyped using the Sty I chip of the 500K Human Mapping Array wet (Affymetrix), containing probes to 238,270 SNP loci, according to manufacturer's instructions. In brief, 250 ng of genomic DNA was digested with the StyI restriction enzyme (New England Biolabs), ligated to an adaptor with T4 ligase (New England Biolabs), and PCR-amplified using a 9700 Thermal Cycler I (Applied Biosystems) and Titanium Taq (Clontech) to achieve fragments ranging from 200-1100 bp. These fragments were pooled, concentrated, processed through a clean-up step, and further fragmented with DNaseI (Affymetrix) before being labeled, denatured, and hybridized to arrays. Arrays were then scanned using the GeneChip Scanner 3000 7G (Affymetrix). Samples were processed in batches of 96 on a single plate using a Biomek FX robot with dual 96 and span-8 heads (Beckman Coulter) and a GeneChip Fluidics Station FS450 (Affymetrix) and tracked using 2D barcode racks and single tube readers (ABGene). Raw data are available at www.broad.mit.edu/tumorscape.

2. Generation of segmented data

Probe-level signal intensities were normalized to a common reference array using quantile normalization [17] and combined to form SNP-level signal intensities using the model-based expression (PM/MM) method [18]. For each tumor, genome-wide copy number estimates were obtained using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that is most similar to the tumor (to be described in greater detail in Getz et al, in preparation). This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. However, similar results were also obtained using other previously described methods [19] (data not shown). Normal samples used in this process were confirmed to lack contamination with tumor cells by visual inspection of their copy-number profiles. Copy number profiles were segmented using the Gain and Loss of DNA (GLAD) algorithm [20] with default parameters. Results were robust to modification of these parameters or use of Circulary Binary Segmentation [21] (data not shown). SNP markers within previously mapped CNVs [22] were removed, as were the 10,000 SNPs with the highest absolute G-scores (see below) in our panel of

1480 normal samples and any SNPs that were aberrant in >1% of these normal samples. Segments containing fewer than 6 SNPs were removed.

3. Determination of SCNA lengths and amplitudes

Copy-number profiles were deconstructed into individual SCNAs as shown in Supplementary Figure 1a. The method (to be described in greater detail in Mermel et al, in preparation) determines the minimum number of SCNAs required to reconstruct the copy-number profile. Initially, amplifications are only allowed to overlap amplifications and vice versa for deletions, providing a unique solution to the lengths and amplitudes of these SCNAs. In reality, however, amplifications may overlap deletions, leading to many possible SCNA combinations that could produce a given profile. We applied an iterative optimization algorithm to determine which of these solutions was most likely. Here, the distributions of lengths and amplitudes for SCNAs determined in one iteration were then used to score the likelihood of different possible SCNA combinations in the next iteration. To reduce computation time, the number of possible SCNA combinations was limited by allowing only two SCNAs per chromosome to form basal copy-number levels with which both amplification and deletion SCNAs might overlap. These basal SCNAs were separated by a single breakpoint that might reside anywhere in the chromosome.

4. Length and amplitude thresholds

The length of each SCNA was converted into chromosome-arm units by calculating the fraction of each chromosome arm covered by the SCNA; for SCNAs that cross the centromere, the length is expressed as the sum of the fractions of each chromosome arm covered by the SCNA. This normalization allowed for the comparison of events occurring on chromosome arms of different length and results in length values ranging between 0 and 2. Five chromosomes (13, 14, 15, 21, and 22) have fewer than 8 probes mapping to the short (p) arm; for these chromosomes, only the q-arm is counted, resulting in a maximal SCNA length of 1. Removal of these chromosomes does not substantially affect the distribution of SCNA lengths as shown in Figure 1a or Supplementary Figure 2, nor does it explain the excess of single-arm length SCNAs relative to focal SCNAs of nearly the same size (data not shown).

SCNAs with lengths > 0.98 chromosome arms were used for arm-level analyses and SCNAs with lengths < 0.5 chromosome arms were used for focal analyses. The results of these focal analyses were not significantly different when the focal length threshold was varied from 0.3 to 0.98 (data not shown).

Only SCNAs with copy number changes >0.1 or <-0.1 inferred copies were included in subsequent analyses. These thresholds were achieved in 0.35% and <0.1% of amplifications and deletions in normal samples (representing rare germline CNVs and occasional analytic artifact).

5. Assessing the significance and tissue distribution of arm-level SCNAs

Across the entire dataset, we noted that the frequency with which chromosomal arms are measured to undergo gain or loss is negatively correlated with the size of that arm (Supplementary Figure 6). Two potential explanations for this trend are that longer chromosome arms have a lower background rate of copy number change, or that copy changes affecting larger chromosome arms are subject to a greater degree of negative selection. In either case, deviations from this trend suggest the presence of additional selective pressures. Chromosome arm-level SCNAs which are observed less frequently than predicted likely undergo additional negative selective pressure. Conversely, arm-level SCNAs that are observed more frequently than predicted are likely to be affected by either positive selection, or a relative absence of negative selection.

To determine which arms were significantly enriched/depleted among copy gains and losses, and therefore suggesting the presence of additional selective pressures, we compared the expected frequency of gain and loss for each arm, determined by linear regression (average alteration frequency vs. # genes on chromosome arm), with the actual frequency observed over the entire dataset. Since samples with gain of a chromosome arm cannot have loss of the same arm, we computed the frequency of gains and loss among the undeleted and unamplified samples, respectively. By decoupling the gains and losses in this way, the frequency metric follows a binomial distribution; z-scores for each arm were calculated using the normal approximation to the binomial (Figure 1b), and the resulting p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg FDR method [23].

To assess how these tissue specific arm-level patterns compared across tumor types, we computed the frequency of arm-level gain minus the frequency of arm-level loss for each arm within each tumor type for which we had greater than 20 samples (see Supplementary Figure 7b). Hierarchical clustering of the resulting values was performed using the Pearson correlation distance metric and complete linkage. Replicate clustering with multiple distance metrics and filtering criteria gave broadly similar results (data not shown). To identify the arm-level changes that most significantly differentiated between the resulting major tissue clusters, we utilized the Comparative Marker Selection Tool [24] available in the GenePattern Software Suite [25] (http://www.broad.mit.edu/cancer/software/genepattern/), using the signal-to-noise test statistic (Supplementary Table 6).

6. Identification of Recurrent Focal SCNAs

Significantly recurrent focal SCNAs were identified using the GISTIC methodology [19], with three improvements described below (and to be described in greater detail in Mermel et al, in preparation). The motivation behind GISTIC is to identify regions where SCNAs are observed significantly more frequently than the background rate. In the absence of independent estimates of the background rate, the previous version of GISTIC used the overall frequency of SCNAs across the genome, taking in account the amplitude of copy number change. In part, the improvements described below make use of the large number of segments available in this dataset to refine our estimates of the background rate of SCNA to more accurately reflect its dependence on both amplitude

and length. It should be noted that the existence of widespread positive or negative selective pressure may lead to inaccurately high or low estimates of this background rate. Indeed, as described in the main text, the finding that deletions tend to preferentially avoid gene-dense regions (Figure 2b) is consistent with the presence of widespread negative selective pressures that may lead us to underestimate the background rate of deletion.

A. Scoring of the Genome

Optimally, each marker should be scored (GISTIC uses the "G-score") by the probability of undergoing all the events observed at that marker—either by multiplying the probabilities of each event, or (as is the procedure in GISTIC) adding the logs of those probabilities. With the large dataset available in the current study, we have been able to revise the scoring scheme to reflect these probabilities more accurately. The probability of a marker undergoing a focal SCNA appeared to be approximately equal for SCNAs of all lengths up to the level of a chromosome arm because the frequency of longer SCNAs was inversely proportional to their length (Figure 1a). Therefore, we did not include a length term in the G-score, other than to separate arm-level SCNAs. We found both amplifications and deletions to be exponentially less frequent with increasing amplitude (measured as number of copies); therefore we scored each SCNA proportional to its amplitude. We also found focal deletions (not amplifications) were less frequent in regions with arm-level deletions in the same sample; these were therefore scored with more weight.

Another possible factor determining the background rate of SCNAs is the presence of repeat sequences or segmental duplications. Recombination of homologous DNA sequences such as segmental duplications has been posited to be a mechanism by which focal SCNAs are generated [26]. Although we did observe a statistically significant enrichment of breakpoints in regions of segmental duplication (see Main Text), the effects on the distribution of SCNAs across the genome are not clear. One expectation might be that more SCNAs would be observed near centromeres and telomeres, which are heavily enriched for repetitive sequences. We evaluated this by rescaling each chromosome arm to a single size and summing copy-number profiles across all samples and arms (Supplementary Figure 8). There was little bias toward telomeric or centromeric amplifications. Some excess of telomeric deletions were observed (at approximately 1/3 of the level required to attain significance), but we did not observe excess centromeric deletions. Due to the small magnitude of these effects and the uncertainty as to their source, we did not account for them in our model of the background rate.

An additional modification was implemented for the deletions analysis to account for the fact that deletions affecting any part of a gene are likely to have similar functional consequences. In this new approach, termed 'Gene-GISTIC', each gene (rather than SNP marker) is given a single G-score reflecting the maximal level of deletion seen anywhere in that gene, summed over all samples. One complication is that genes with more SNPs are more likely to score higher by chance alone. Gene-GISTIC corrects for this by using

G-scores generated from similar-sized windows in permuted data as the null distribution when calculating significance values (to be described in detail in Mermel, et al, in preparation).

The Gene-GISTIC approach provides a more accurate weighting of the significance of genes subject to frequent but non-overlapping deletions and an increase in overall power due to a reduction in the number of independent hypotheses tested (from the total number of markers on the array to the number of genes in the genome). A direct comparison of the results of Gene-GISTIC and the traditional SNP-based GISTIC deletions analysis found 82 peaks by Gene-GISTIC compared to 64 by SNP-GISTIC; 62 peaks overlapped (Supplementary Table 7). Known tumor suppressor genes tended to rank higher in the Gene GISTIC results (not shown). One potential drawback to the Gene-GISTIC approach is that regions without known genes (RefSeq genes and miRNAs were included in this study) will not be scored and potentially significant deletions may be missed. Indeed, 11 peaks were more significant according to SNP-GISTIC than Gene-GISTIC (Supplementary Table 7), likely due to the underweighting of deletions occurring outside of known genes.

B. Peak Region Identification

To identify independently significant regions in a single chromosome, GISTIC employed a greedy "peel-off" algorithm approach that identifies the most significant peak, removes all SCNAs spanning that peak, and then rescores the chromosome to identify additional significant peaks. We have modified the algorithm to increase the sensitivity for additional peaks. SCNAs are allowed to contribute to secondary peaks with a weighting proportional to the evidence that the secondary peak represents a separate event from the primary peak. In brief, after removing the SCNAs overlapping the primary peak, the next highest-scoring peak is identified. "Disjoint G-scores" for both the primary and secondary peaks are calculated based only on SCNAs that overlap one or the other peak but not both. SCNAs that overlap both peaks are then allowed to contribute to each peak with a weighting proportional to the disjoint G-score of that peak divided by the sum of disjoint G-scores over both peaks, and the significance of each peak is redetermined. The procedure is performed iteratively until no further significant peaks are identified. The modification improves the sensitivity of the method for identifying known cancer genes without substantially decreasing its specificity (to be described in detail in Mermel et al, in preparation).

C. Peak Region Boundary Determination

We have also modified the method employed by GISTIC to define the boundaries of each peak region, to add an explicit accounting for the likelihood passenger events or other sources of noise have displaced the local G-score peak from the gene target (Supplementary Figure 3d). The variations in G-scores across the genome in permuted data are tabulated to determine the likelihood of observing any given change in G-score (ΔG) over any given distance. We set the boundaries of each peak region such that the

decrease in the G-score from peak to boundary had a likelihood of 5% or less, representing the 95% confidence interval for inclusion of the gene target.

7. The effect of gene size and density on observed SCNA frequency

To determine whether large genes are associated with peak regions of amplification or deletion, we ranked genes according to the genomic footprint of their coding sequence, defined as the largest difference between transcription start and stop sites over all annotated transcripts in genome build hg18. We computed the local gene density around each gene by counting the number of annotated genes residing within a 4 Mb window centered around the midpoint of the gene and dividing by the average number of genes in the 4 Mb window around all genes in the genome.

To determine the relationship between SCNA frequency and gene density, we first discretized each copy number profile based on the following 7 copy number ranges: < 1, 1-1.5, 1.5-1.75, 1.75-2.3, 2.3-3, 3-4, and > 4. The gene density within each of these copy number ranges was calculated by dividing the total number of genes residing within each copy number bin across all samples by the number of SNP markers covered by those regions; these density values (in genes per SNP; similar values were obtained using genes per Mb) were normalized against the average gene density across the genome in Figure 2b. We computed the significance of deviations from the average gene density by comparing the gene density for each copy number bin to the distribution of gene densities in 1e6 random permutations of identically sized regions across the genome. The green lines in Figure 2b correspond to the gene densities giving Bonferonni-corrected p-values of .01. These lines spread outward at extreme copy numbers because the number of segments residing within these bins is smaller.

8. GRAIL Analysis

To compare the functional relatedness of the genes identified by our focal SCNA analysis, we utilized the GRAIL algorithm [27] (full methods and algorithm available at www.broad.mit.edu/mpg/grail) on amplification and deletion peak regions separately, using the default parameters. In brief, GRAIL determines the relatedness between any two genes in different peak regions based upon the frequency with which the same terms are found in PubMed abstracts citing each gene (all PubMed abstracts until December 2006 are used). Each gene is scored by its level of relatedness to all genes in all other peak regions, and assigned a p-value reflecting the likelihood of achieving such a score by chance. Each peak region is assigned the p-value of its most significant gene with a multiple hypothesis correction to reflect the number of genes in the peak. The literature terms most associated with the top genes in each peak region are noted. To confirm the p-values assigned to the peak regions were not overestimates of significance, we compared them to similar p-values generated using 1000 permutations of the locations of the peak regions ("permuted controls" in Figure 2c).

9. GO Term Analysis

The latest Gene Ontology annotations were downloaded from The Gene Ontology website (http://www.geneontology.org/GO.downloads.ontology.shtml). We associated each GO term with all genes that are annotated with that term or any of its descendent terms in the GO hierarchy. We assessed enrichment of each GO term by comparing the number of genes associated with that term and present in our amplification and deletion peak regions to the number expected if these genes were distributed at random across the genome. Peak regions with greater than 25 genes were eliminated from the analysis to maximize power, and at most 2 genes from each peak region were allowed to count towards the enrichment score to eliminate confounding due to genomic clustering of close homologues. GO terms with fewer than 10 associated genes were excluded from the analysis to avoid significant enrichments based only on very small numbers of genes. The significance of the enrichment for each peak was calculated using the G-test, with an FDR correction to account for the number of hypotheses being tested.

10. Peak Region Overlap

To quantify the degree of overlap among peak regions identified in different datasets, we counted two peaks as being the same if their 95% confidence intervals overlap. P-values, representing the likelihood of obtaining the observed levels of overlap if peak regions were randomly distributed, were determined by permuting the locations of the peak regions in each dataset 1,000 times and determining the fraction of peaks that overlap in each permutation.

To count the total number of non-overlapping peak regions identified across all cancer sets, we first removed peaks that overlapped with any of the 158 peaks in the pooled analysis. The remaining peak regions were sorted by size (smallest to largest); starting with the smallest peak, we examined each peak for overlap with any smaller peak. If overlap was observed, the larger of the two peaks was removed.

11. Fluorescence *in-situ* hybridization (FISH)

Four-micron tissue microarray (TMA) sections were mounted on standard glass slides and baked at 60°C for at least two hours, then de-paraffinized and digested using methods described previously [28].

The following DNA probes were co-hybridized: RP11-54A4 (SpectrumGreen), which maps to 1q21.2 and includes *MCL1*; D1Z5 (SpectrumOrange), which maps to 1p11-q11 (SpectrumOrange); and D11Z1 (SpectrumAqua), which maps to 11p11.11-q11.11. The D1Z5 and D11Z1 probes were purchased from Abbott Molecular/Vysis, Inc. The *MCL1* probe was obtained from CHORI (www.chori.org), direct-labeled using nick translation and precipitated using standard protocols. Final probe concentration was approximately 50-100 ng/ul. D1Z5 and D11Z1 final probe concentrations followed manufacturer's recommendations.

TMA sections and probes were co-denatured, hybridized at least 16 hrs at 37°C in a darkened humid chamber, washed in 2X SSC at 70°C for 10 min, rinsed in room temperature 2X SSC, and counterstained with DAPI (4',6-diamidino-2-phenylindole, Abbott Molecular/Vysis, Inc.). Slides were imaged using an Olympus BX51 fluorescence microscope. Individual images were captured using an Applied Imaging system running CytoVision Genus version 3.9.

12. Cell culture conditions

NCI-H2110, HCC 1954, HCC 1171, NCI-H1568, NCI-H322, NCI-H1792, SKLU1, NCI-H647, NCI-H520, NCI-H2228, LCLC-97TM1, PC9, NCI-H1437, and NCI-H3122 cells were maintained in RPMI 1640 plus 2 mM L-glutamine (Cellgro) supplemented with 10% fetal bovine serum (Gemini Bio-Products), 1 mM sodium pyruvate, and penicillin/streptomycin (Cellgro). For A549 and Calu6 cells, F12K and DMEM respectively were substituted for RPMI. Immortalized lung epithelial cells [29] were maintained in SAGM small airway cell basal medium with supplements (Lonza).

13. Quantitative PCR

Quantitative real-time PCR was performed with an ABI 7900 HT Sequence Detection System (Applied Biosystems) using the QuantiTect SYBR Green kit (Qiagen). Copy-numbers were quantified relative to the repetitive sequence element Line-1 as previously described [30]. For *MCL1*, the forward and reverse primer sequences were CTTCCAAGGTAAGGGGGTTC and ACTGACTCGTTTCGGTTTCC, respectively; for *BCL2L1* the forward and reverse primer sequences were CCTCTCCCGACCTGTGATAC and CTTCCTCGGAAAGTCACTCC, respectively.

14. RNAi and cDNA expression

Inducible shRNA vectors were generated as previously described [31] using sequences targeted against *MCL1* (GCATTGGCATCTTTGGATTTC) and scrambled control (GTGGACTCTTGAAAGTACTAT) [32]. Stable shRNA vectors were provided by The RNAi Consortium [33] and sequences were inserted to target *MCL1* (GCTAAACACTTGAAGACCATA, GGATTGTGACTCTCATTTCTT, and GCAGGATTGTGACTCTCATTT), and *BCL2L1* (CGTGTCTGTATTTATGTGTGA, CCACCAGGAGAACCACTACAT, and TGGCCTCAGAATTGATCATTT), as well luciferase and LacZ (CGCGATCGTAATCACCCGAGT and CTCTGGCTAACGGTACGCGTA) controls. Lentiviruses were made by transfection of 293T packaging cells with a three plasmid system [34,35]. Target cells were incubated with lentivirus for one hour in the presence of 8 μg/ml polybrene. Infections leading to >30% decreases in proliferation due to viral toxicity were repeated at lower titer. Cells were selected using puromycin at 2 mg/ml over 2-3 days or until all of the non-infected cells died.

Knockdown of *MCL1*, *ADAMTSL4*, *CTSK*, *CTSS*, *ECM1, ENSA*, *GOLPH3L*, and a non-targeting control was also achieved by transfection with siRNA siGenome SMARTpools

(Dharmacon), and the single sequence #13 (GAUUGUGACUCUCAUUUCUUU) from the *MCL1* SMARTpool as previously described [36].

Retroviral vectors were used to introduce specific genes into immortalized lung epithelial cells [37]. *MCL1* and *BCL2L1* cDNAs were each introduced into pWZL-BLAST; *MYC* cDNA was introduced into pBABE-Puro.

15. Cell proliferation assays

Proliferation of cells in inducible *MCL1* knockdown experiments was measured using the xCELLigence RTCA machine (Acea Biosciences). Cells were seeded at 1500 cells/well in 96-well plates and doxycycline (100 ng/ml) or vehicle control was added after 24 hours. Electrical impedence was measured every 30 minutes for 48 hours post-induction. Each measurement was performed in triplicate on at least two separate occasions. Proliferation of cells in all other experiments was measured using CellTiterGlo reagent (Promega). with measurements taken at 0 and 48 hours post-infection or -transfection (for *MCL1* shRNA and siRNA experiments, respectively) or at 3 and 7 days post-infection (*BCL2L1* shRNA experiments). Cell proliferation assays performed on cells infected with stable shRNA vectors were performed immediately after lentiviral infection, in parallel on aliquots treated with and without puromycin. The data presented represent cells not treated with puromycin, although similar results obtained in both cases (data not shown).

16. Xenografts

Female nu/nu mice maintained in pathogen-free facilities were implanted subcutaneously with 5e6 cells infected with inducible shRNA vectors against *MCL1* or scrambled control. Tumor size was assessed by calipers twice weekly. When tumors reached 100 mm^3 (11 days post-implant), eight mice in each group were fed doxycycline 25 mg/kg po qd and eight additional control mice were fed D5W for an additional 11 days.

17. Immunoblot analysis

Both adherent and floating cells were harvested after incubation overnight and lysed using 2x SDS sample buffer (125 mM Tris-base, 138 mM SDS, 10% β-mercaptoethanol, 20% glycerol, bromophenol blue, pH 6.8). Lysates were boiled for 10 min., cleared of insoluble material by centrifugation at 16,000 x g, and subjected to SDS-10% polyacrylamide gel electrophoresis (PAGE). Blots were probed with antibodies against MYC (ab32, Abcam), MCL1 (ab32087, Abcam), BCL2L1 (2762, Cell Signaling), cleaved PARP (9541, Cell Signaling), GAPDH (MAB374, Chemicon), and actin (ab8227, Abcam).

18. Flow cytometry

Adherent and floating cells were harvested after incubation overnight and stained with Annexin V-FITC (Sigma) and propidium iodide (BioVision). Flow cytometric analysis was performed on 3e4 cells using the BD LSR II flow cytometer (BD Biosciences).

Supplementary Note 1: Background and Terminology

a) Somatic vs. Germline Copy Number Changes

Throughout this paper, we use the term somatic copy number alteration (SCNA) to refer to *somatic changes* in the number of copies of a DNA sequence that arise during the process of cancer development. SCNA should not be confused with two similar terms, copy number variation (CNV) and copy number polymorphism (CNP), which refer to copy number changes in DNA segments present in an individual's germline DNA. Definitions of these terms, as used throughout the manuscript, are as follows:

Somatic Copy Number Alteration *(SCNA):* A sequence that is found at different copy numbers in an individual's germline DNA and in the DNA of a clonal sub-population of cells.

Copy Number Variation *(CNV):* A DNA sequence that is found at different copy numbers in the germline DNA of two different individuals.

Copy Number Polymorphism *(CNP):* A locus that exhibits CNV above some specified frequency (typically 1-5%) among individuals within a population.

Because not all of the cancer DNA specimens in our dataset are matched to normal DNA specimens, we cannot be entirely confident that any given copy number change observed in a cancer DNA sample was not present in the germline of the patient. To avoid confounding our analysis of somatic CNAs with germline CNVs, we have masked from our dataset all markers covering previously annotated CNPs [22], as well as those markers found to be altered in at least 1% of the normal samples in our dataset (see Supplementary Methods, above).

The amplitude of copy number change

In the cytogenetics literature, "gains" has traditionally referred to increases of one or a small number of copies of a DNA segment, typically spanning a large genomic region. In contrast, "amplifications" has referred to more focal events that can reach much higher copy numbers. A similar distinction has been made between "losses" and "deletions". Current analytical methods do not allow the determination of absolute copy number from array-based platforms, rendering these distinctions less clear. For consistency, we refer to arm-level events as "gains" or "losses" because of their large genomic extent and tendency to involve limited copy number changes, and focal events as "amplifications" and "deletions" due to their more limited extent and propensity to reach higher copy numbers.

b) Background Rates and Selection of SCNAs

Oncogenesis is an evolutionary process [38]. DNA alterations are acquired at random according to a rate of generation that is determined by the competing processes of mutation and repair, and which may vary according to the type of aberration and the genetic and cellular context. Once acquired, these alterations may be neutral, or may be subject to positive selection (if they promote oncogenesis) or negative selection (if they have deleterious effects on the cell). In the absence of selective pressure, an alteration will be observed at a "background rate" equal to its generation rate times the number of cell divisions. The frequency with which an alteration is observed in cancer specimens is determined by both this background rate and the degree of selective advantage or disadvantage it confers.

Alterations that promote oncogenesis (often referred to as "driver events"), in particular, should be present at above the background rate. Alterations that do not contribute to the cancer phenotype (often referred to as "passenger events") may nevertheless be observed in the bulk of a cancer sample if a subsequent beneficial alteration (driver event) provides the cell a net fitness advantage. This process is often referred to as "hitch-hiking" [39]. Indeed, even somewhat deleterious alterations may achieve fixation through hitch-hiking if the subsequent driver events confer a net fitness advantage to the cell. The process by which a cell is able to reach fixation through a less fit intermediate has been described as "stochastic tunneling" [40]. The result of hitch-hiking and stochastic tunneling is that many alterations observed in cancer genomes do not promote oncogenesis.

Systematic efforts to discover all oncogenic somatic genetic alterations therefore require both an accurate model of the background rate and a sufficiently large collection of cancer samples to provide sufficient power to detect alterations occurring above this frequency [19,41,42,43]. For point mutations, reasonable estimates of the background rates are provided by the synonymous and intergenic mutation frequencies, which are believed to be selectively neutral [44]. By contrast, no clear distinction has been defined between selected and neutral SCNAs, making precise estimation of the background rates difficult. A common approach to making these estimates is to assume the background rates are similar to the overall rate of SCNA within each chromosome [45,46] or across the entire genome [19,47].

While this approach of estimating the background mutation rate from the observed data is statistically unbiased, the fact that the observed data has already been subjected to a selective process *in vivo* makes it is impossible to precisely distinguish between variation due to differences in mutation rates from variation due to differences in the level of selective advantage or disadvantage conferred by each mutation. An additional complication is that the background rate estimated from the data will also include false positive events (due to technical noise from the measuring platform) and false negative events (that occur below the detection limit of the measuring platform). Therefore, somatic alterations may appear to occur at a significantly elevated frequency across samples for at least four reasons: (i) they are generated in that region at a rate significantly above the genome-wide average, (ii) they occur in a region subject to significantly less negative selection than the typical genomic region, (iii) they give a selective advantage to cells harboring them (i.e. they are driver alterations), or (iv) they

represent systematic artifact. While the statistical background rate estimated from the observed data is useful in the identification of regions altered at statistically significant frequencies – potentially suggesting the presence of positive selection – one should not simply equate this rate with the biological background mutation rate, or assume that all mutations occurring at an elevated frequency are drivers. Conversely, one should not assume that all mutations occurring at rates equal to or lower than the estimated background rate are passengers.

The interpretation of the significance of a frequent mutation therefore depends on our understanding of its particular background rate. This rate may vary according to specific features of the mutation, such as the type of base pair substitution for point mutations or the length, magnitude, and surrounding sequence for copy number alterations. Naïve analyses which do not account for these features – by assuming, for example, that SCNAs are equally likely to occur anywhere in the genome or to be of any size – will be biased towards regions with high background mutation rates and away from regions with low background mutation rates. For example, it is known that point mutation rates vary significantly according to the type of substitution (e.g. transition vs. transversion) and sequence context (e.g. CpG vs. non CpG); various statistical methods for the analysis of point mutations take this variation into account to avoid biasing the results towards genes or regions with many mutable bases [41,42,43]. The background mutation rate may also be underestimated if many mutations confer negative selective pressure and therefore are observed less commonly than they occur. In this case, a neutral mutation observed at the true background rate may appear to be significantly enriched in cancer.

One of the goals in the analysis of SCNAs is to identify features that correlate with the frequency with which these SCNAs are observed. Whether these features influence SCNA frequencies through mechanistic effects on background mutation rates, through selective pressure, or through association with technical artifact should be determined by appropriate validation experiments.

Supplementary Note 2: The impact of sample size on focal SCNA analysis

In this paper, we have utilized the large sample collection generated by analyzing DNA specimens across multiple cancer types to increase our power to identify and resolve the targets of significant regions of focal SCNA. To understand the effects of sample size on the ability to discover targets of focal SCNA, we must separately consider the two critical steps in our focal SCNA analysis: 1) identifying that a region is undergoing SCNA significantly above the background rate, and therefore is likely be subject to positive selection; and 2) given that a region of SCNA is undergoing selection, resolving the genomic region most likely to contain the target gene(s).

Step 1: Identifying that a region of SCNA is undergoing positive selection

The GISTIC G-score at each marker locus is constructed to estimate the probability of observing the set of SCNAs covering that locus by chance, taking into account both the frequency and mean amplitude of SCNA (see Supplementary Methods). To compute the significance of each region, the G-score is compared to the distribution of G-scores expected if the SCNAs in the region were all random events generated at the background rate. GISTIC estimates this background rate using the overall rate of focal SCNA across the genome.

For a focal SCNA occurring at a fixed frequency and average amplitude, the power to detect that region generally increases with sample size. However, the relationship between detection power and sample size is complicated by several additional issues. For one thing, combining heterogeneous sample sets can reduce the power to detect SCNAs that are primarily enriched in a single subset (by reducing the frequency of the region of interest in the combined dataset). That is, mixing cancer specimens across tissue types will diminish the power to detect true lineage restricted SCNAs. Mixing samples with different background rates of alteration can similarly affect the statistical power of the combined analysis in ways that obscure the effect of sample size alone.

Across the 17 individual cancer types studied in our dataset, there is a weak but significant association between sample size and the number of significant focal SCNAs detected ($r = 0.51$, $p = .04$; data not shown). Of course, because the total number of 'true' driver SCNAs in each cancer type is unknown, the number of significant SCNAs identified in any given cancer type is not a direct measure of statistical power. A more informative measure of the relationship between sample size and statistical power is demonstrated by an analysis of randomly selected subsets of the entire dataset (Supplementary Fig 3a). Since each subset is drawn from the same total dataset, the expected frequency and background rate of each subset is, on average, the same. As can be seen, increasing the number of samples increased the number of peaks identified over all subset sizes, indicating that our increased sample size led to increased power overall. However, it is also clear that the number of peaks appears to be saturating by 3131 samples, suggesting that adding additional samples will not greatly increase our power to detect novel SCNA targets (at least for a similarly composed dataset).

Step 2: Resolving the genomic region most likely to contain the target gene(s)

Once a region of significant focal SCNA has been identified, the next step is to define the genomic boundaries likely to contain the target gene(s) of that SCNA. Most approaches to resolving this region directly or indirectly compute the minimal common region (MCR) of overlap among the SCNAs covering the significant locus, as this is the region most likely to contain a targeted gene. However, due to both technical and biological noise (e.g. segmentation artifacts or random "passenger" SCNAs that confer no selective advantage to the cell), the MCR may be displaced from the actual location of the gene targets. We have developed a statistically based approach (see above and Mermel et al., manuscript in preparation) that models the expected variations in the G-score using the observed level of noise across the genome to determine a wider region than the MCR for which we are 95% confident contains the true target gene.

The two major determinants to how narrowly a significant region of SCNA can be refined are the size of the MCR due to SCNAs overlapping the target gene (here called "driver SCNAs") and the noise level contributed by SCNAs that do not necessarily overlap the target gene (here called "passenger SCNAs"; these may represent real SCNAs or analytic artifact).

By definition, the size of the MCR can never increase with the addition of samples containing driver SCNAs, and will more typically decrease. Indeed, under the simplifying assumptions that driver SCNA breakpoints are random with a uniform distribution between 0 and some maximal distance L units away from a target gene, the minimum distance to a breakpoint will scale as $1/(n+1)$ [48], where n represents the number of samples with driver SCNAs. This reduces to 1/n when n is large, implying that ***the expected size of the MCR is inversely proportional to the number of samples harboring the driver SCNA***. In reality, the assumptions behind this derivation do not hold exactly, as there is a minimal observed SCNA length that depends on the resolution of the measuring platform, and SCNA breakpoints are likely to be scattered non-uniformly across the genome. Nonetheless, for the vast majority of focal peak regions, the model does a reasonably good job of approximating actual MCR sizes in random subsets of the dataset (data not shown), suggesting that number of samples remains the major factor limiting the resolution of most focal peaks. The fact that the MCR resolution scales inversely with the absolute number of driver SCNAs, rather than the overall frequency of aberration, implies that once a region of significant SCNA has been detected, the addition of extra samples (even if they contain a low frequency of alteration at a given locus) will only help to resolve the target gene. In particular, doubling the number of samples with the driver SCNA will halve the expected size of the MCR.

The relationship between the noise level due to passenger SCNAs and sample size is difficult to model as it depends on the particular mix of samples in the dataset as well as the underlying error model of the measuring platform and analytical methods. Insofar as the noise around a given locus is unbiased, the errors from additional samples with passenger SCNAs will tend to cancel, whereas the signal contributed by samples with driver SCNAs will tend to add. Overall, this will result in more confident boundary estimation with greater numbers of samples. In fact, according to the central limit theorem, ***the error in boundary estimation will decrease as $1/\sqrt{N}$,*** where N represents the total number of samples (including those with driver and passenger SCNAs)**.** This result, like the one above, suggests that increasing numbers of samples will tend to provide more precise estimates of the location of the target gene.

Empirically, we observe that for all but the most frequent regions of SCNA (where we are likely saturating the resolution limit of the array), our ability to resolve the target region is roughly inversely proportional to the size of a randomly chosen subset (see Supplementary Figure 3e), as predicted by the models above. The median number of genes per peak region roughly halves when increasing the sample size from 1600 to 3131, suggesting that further improvements in resolution could be achieved with further increases in the sample size.

Supplementary Note 3: Data sources

The 250K SNP array data used in this study were obtained from several sources, including previously published data from our laboratory[1,2,3,4,5,6,7,8,9,13,28,58,68] (Barretina et al, in review; Brachmann et al, in preparation; Bass et al, in preparation) and other groups[10,11,12] and previously unpublished data from cancer and normal specimens (Supplementary Table 1). The published cancer copy-number profiles include 510 non-small cell lung cancers[9,10,13,58,68], 388 acute lymphoblastic leukemias[10,11,12], 130 breast cancers[1,2,10], 215 myeloproliferative disorders[8], 151 colorectal cancers[10,28], 2 medulloblastomas[10], 111 renal cancers[7,10], 106 hepatocellular cancers[3,10], 77 melanomas[4,10], 7 ovarian cancers[10], 54 prostate cancers[6,10], 73 esophageal adenocarcinomas (Bass et al, in preparation), 52 dedifferentiated liposarcomas (Barretina et al, in review), 40 esophageal squamous cell cancers[10,13], 21 gastrointestinal stromal tumors (GISTs; Barretina et al, in review), 10 gliomas[10], 21 small cell lung cancers[10], 36 myxofibrosarcomas (Barretina et al, in review), 32 leiomyosarcomas[10] (Barretina et al, in review), 31 neuroblastomas[5,10], 25 synovial sarcomas (Barretina et al, in review), 26 mesotheliomas (Brachmann et al, in preparation), 24 pleomorphic liposarcomas (Barretina et al, in review), 23 gastric cancers[10] (Bass et al, in preparation), 4 thyroid cancers[10], 21 non-Hodgkin's lymphomas[10], and 115 miscellaneous other types of cancer[10] (Barretina et al, in review).

Supplementary Note 4: A Pooled analysis of arm-level SCNAs

Several previous studies have analyzed arm-level SCNAs in large numbers of cancer samples characterized by low-resolution array or cytogenetic technologies [49,50]. These studies have identified arm-level SCNAs observed frequently both within and across cancer subtypes. Moreover, these arm-level SCNAs have been shown to segregate by cancer type, with cancers of similar developmental origin showing similar patterns of SCNA.

In parallel to our approach to focal SCNAs, we compared frequencies of arm-level SCNA to estimates of their background rates. In many ways, this analysis serves to highlight certain broad similarities between arm-level and focal SCNAs.

As with our focal SCNA analysis, our analysis of arm-level SCNAs began with a systematic evaluation of the observed rate of these events across the genome. We observed that arm-level alterations are more common in short rather than long chromosome arms (Supplementary Figure 6). The correlation is stronger when the length of the chromosome arm is measured by number of genes rather than megabases ($p = 0.0005$). This trend is observed in separate analyses of 25 of the 26 cancer types most represented in our dataset. The sole exception is hepatocellular carcinoma, which shows no trend in either direction—in part due to a very high frequency of amplification of the longest chromosome arm, 1q. In 13 of these 26 cancer types, including examples from all developmental lineages, this trend reached statistical signficance within a single type (data not shown). Although both focal and arm-level SCNAs exhibit decreasing

frequency with length, the strength of the trend differs in the two cases. Several possibilities may account for this, including differences in the mechanisms by which these SCNAs are generated, the effects of selection, and experimental artifact.

A caveat to this analysis is that we do not distinguish between whole-chromosome and single-arm-level SCNAs, although the mechanisms and rates between these may differ. Indeed, in separate analyses of these two types of SCNA, both trend towards fewer events in SCNAs covering more genes. However, this trend was significant only for whole-chromosome SCNAs ($p = 0.003$), not single-arm-level events ($p = 0.28$) (data not shown). This may be due to the ambiguities inherent in attempting to separate these two types of SCNA: namely, any whole-chromosome SCNA is equivalent to concordant SCNAs in both of its arms. Single-arm-level SCNAs can only be detected when the two arms are discordant (as is frequently observed with deletion of 8p and amplification of 8q). As a result, fewer single-arm-level SCNAs will be detected, reducing the power available to identify significant trends. Moreover, any SCNA of an acrocentric chromosome (chromosomes 13, 14, 15, 21, and 22) is inherently ambiguous, as it is simultaneously a whole-chromosome and single-arm SCNA. For these reasons, we present a unified analysis of arm-level SCNAs that includes whole-chromosome SCNAs.

The prevalence of specific arm-level SCNAs, however, is not fully explained by the number of genes present in each of these arms. Indeed, the high frequency of specific arm-level gains and losses suggests enrichment due to selective pressure, as has been noted in many prior publications [50,51,52]. To our knowledge, however, none of these prior publications has determined the statistical significance of arm-level SCNA by explicitly comparing the frequencies of arm-level SCNAs to the expected rate given their gene number (see Supplementary Methods, above). Across all cancers, 11 of the 39 autosomal chromosome arms exhibit copy number gains and 17 exhibit copy number losses significantly more often than predicted by the number of genes they contain (Figure 1b; see Supplementary Methods). The vast majority of these are strikingly significant, with the most prominent being amplifications of 1q, 20q, and 7p ($p < 1e\text{-}85$ in each case), and deletions of 17p, 9p, and 13q ($p < 1e\text{-}33$ for each). Interestingly, the most significantly deleted arms contain some of the most frequently mutated tumor suppressor genes, including *TP53* (17p), *CDKN2A/B* (9p), and *RB1* (13q), suggesting that the striking enrichment of loss of these arms may be due largely to these genes (Supplementary Table 8). Only nine of the 39 chromosome arms are neither significantly gained nor lost. Despite the finding that most chromosome arms exhibit significant gains or losses, only one (14q) shows both ($p = 0.003$).

Indeed, the striking significance of these arm-level SCNAs across cancer reflects a directional consistency across many different cancer types. In particular, we analyzed arm-level SCNAs separately in each of the 17 cancer types represented by greater than 40 samples (Supplementary Table 8). The 11 significantly gained chromosome arms identified in the pooled analysis were found to be independently gained in a median of 8 cancer types (range 2-11); these same arms were only rarely found to undergo significant loss in any cancer type (median 0, range 0-2 types). Similarly, the 17 significantly deleted arms in the pooled analysis were found to be independently lost in a median of 4 cancer

types (range 2-9), and were only sporadically gained in specific subtypes (median 1, range 0-2 types; note that these gains were predominantly seen in hematopoietic cancers). Chromosome 14q, the only arm found to be both gained and lost in the pooled analysis, was significantly gained in 4 cancer types (acute lymphoblastic leukemia, non-small cell lung carcinoma, small cell lung cancer, and prostate carcinoma) and lost in 3 cancer types (GIST, melanoma, and renal carcinoma). The mutually exclusive gains or losses observed for nearly all chromosome arms across large numbers of cancer types suggest that the selective pressures that shape these events operate in tissues throughout the body rather than being confined to limited, tissue-specific microenvironments.

We were also interested in the extent to which the significant arm-level SCNAs are shared across tissue boundaries. Prior studies have shown many arm-level SCNAs to be prevalent in multiple cancer types [50,51,52]. We compared the arm-level SCNAs identified as significant in each of the 17 well-represented cancer types to those identified in their "complement" (i.e. the entire dataset excluding the cancer type in question). Similar to focal SCNAs, we observed that the large majority (median of 87%) of the arm-level SCNAs identified in any cancer type were also significant in the complement (versus 37% overlap expected by chance). Across all the cancer types, we identified 26 'lineage-restricted' events not found in the complementary pooled analysis (19 arm-level gains and 7 arm level losses), for an average of 1.6 new arm-level SCNAs per tissue type (range 0-7). Nine of these arm-level gains are identified exclusively among hematopoietic cancers. These lineage-restricted arm-level SCNAs may reflect important lineage-specific biology. An interesting example is 13q, which is frequently lost across most cancer types, but is gained in 50% of colorectal cancers, possibly due to the oncogenic effects of *CDK8* and the unique requirement for intact *RB1* (both on 13q) observed in colorectal cancer [28,53]. Chromosome 2 is the only chromosome not significantly altered in at least one cancer type.

Supplementary Note 5: Comparison of focal peak regions to 18 prior publications

To compare our focal peak regions to the results of prior high-resolution cancer copy-number analyses, we compared these regions to a set of 18 publications which reported copy-number regions of interest determined through the use of oligonucleotide arrays on at least 40 samples within any of the 17 major cancer types in our dataset [1,4,6,7,11,19,54,55,56,57,58,59,60,61,62,63,64,65].

Among the 76 peak regions of amplification reported here, 18 had not been identified in any of the prior publications (Supplementary Table 3). For each region of interest, most of these publications reported the minimal common region of overlap across their sample set; here we report a more conservative peak region that is much wider than the minimal common region of overlap to account for the effects of biological and technical noise. Nevertheless, of the 58 amplified regions identified in both this study and at least one of the prior 18 publications, 33 were found to be narrower (and therefore better-resolved) in the present analysis. The size of these regions was a median of 30% of the minimum size of the overlapping regions of interest in any of these prior 18 publications. For example,

the peak region including *GRB2* was identified in one of these 18 publications, but is only 2% of the size of the region in that publication, engendering a much greater ability to focus in on *GRB2* as a possible target. Indeed, *GRB2* is a member of the molecular adaptor family of genes, which we find to be highly enriched among the peak regions of amplification (see Main Text) and, although not known to be an oncogene, is known to play a central role in cancer cell cycle and motility [66].

Among the 82 peak regions of deletion reported here, 18 had also not been identified in any of the prior publications. Our deletion analysis was performed at gene-level resolution to achieve greater power in detecting non-overlapping deletions affecting large genes (see Supplementary Methods), whereas all the prior publications extended to marker-level resolution. Nevertheless, among the 64 regions identified in both this study and at least one of the prior publications, 21 were found to be narrower in the present analysis, with a median size of 10% of the minimum size from the prior publications. A more comparable marker-level analysis of our data (SNP-GISTIC, see Supplementary Methods) exhibited narrower peak regions than in 73% of those regions that had been previously reported (data not shown).

Supplementary Note 6: Tissue-type clustering of arm-level and focal SCNAs

We were interested in examining how the SCNAs identified in the pooled analysis vary across individual cancer types, focusing on the 26 cancer types represented by at least 20 samples in our collection. Some of the arm-level SCNAs occur at very high frequencies within individual subtypes (Supplementary Figure 7a). Indeed, 13 of the 26 cancer types exhibited at least one arm-level SCNA that was present in the majority of samples of that tumor type. By contrast, focal SCNAs were rarely present in the majority of samples of a given cancer type, with only 6 of 26 types exhibiting a focal SCNA present in a majority of samples.

We were also interested in quantifying the extent to which arm-level and focal SCNAs are shared between cancers of similar developmental lineage. Prior studies have demonstrated a tendency for cancers of similar developmental lineage to cluster together on the basis of overall copy number [67], but did not separate out the contributions of these two types of events. Therefore, for each cancer type, we generated an aggregate SCNA profile by subtracting the frequency of loss from the frequency of gain for each significant arm-level and focal SCNA. We then clustered the resulting "consensus" SCNA profiles for each cancer type.

This particular clustering metric attempts to capture the net balance of arm-level changes rather than their absolute frequency; for example, a tumor type with 50% gains and 50% losses of a particular locus would receive the same score as a tumor type with no gains or losses of that locus. However, the clustering results were largely robust to the use of alternative clustering metrics, including scoring each cancer type according to the absolute frequency of gain and loss at each locus, and different clustering parameters

such as complete vs. average linkage and Euclidean vs. Correlation Distance metrics. Also, the high degree of variability within cancer types suggests that this analysis will be influenced by the particular sampling of cancers within each type. For this reason we restricted the analysis to cancer types with >20 tumors (most were represented by >45 tumors) and looked for general features driving the major clusters rather than the specific placement of each cancer type.

Hierarchical clustering of cancer types based on arm-level SCNA profiles (Supplementary Figure 7b) revealed a pattern that closely mimicked the developmental lineage of the tissue types. Three major sub-clusters are readily apparent: a major division between hematopoietic cancers and all other cancer types, followed by a division between epithelial and non-epithelial solid tumors. Within these latter two groups, there are distinct sub-clusters of related tumors, including gastrointestinal (gastric, esophageal adenocarcinoma, and colorectal), gynecologic (ovarian, breast), sarcomas (plus renal cancer), and neural tumors (plus non-Hodgkin's lymphoma). The segregation of cancer types by developmental lineage is highly non-random ($p < 1e\text{-}5$; see Supplementary Methods), and more consistent than the previous attempts using overall SCNA profiles [49]. Specific arm-level SCNAs that distinguish these major subclusters, such as gain of chromosome arm 8q and loss of 17p in epithelial tumors, were identified through comparative marker selection analysis [24] and are listed in Supplementary Table 6.

In contrast, hierarchical clustering of cancer types based on focal SCNAs does not recapitulate developmental lineage as closely (Supplementary Figure 7c). Although there was a tendency for tumors of similar lineages to cluster together ($p = .01$), all three major clusters contained several representatives of each lineage. Consistent with this observation, the ten most significant amplified regions (Supplementary Figure 7d, top panel) and deleted regions (Supplementary Figure 7d, bottom panel) frequently exhibit significant levels of focal SCNA in cancers across diverse lineages. For example, both *EGFR* and *MDM2* amplifications are frequently observed in gliomas (neural) and non-small cell lung cancers (epithelial), but not in medulloblastomas (neural) or small cell lung cancers (epithelial).

The finding that arm-level CNAs, but not focal CNAs, appear to cluster predominantly on the basis on developmental lineage suggests that developmentally encoded selective pressures shape the pattern of these events within specific cancer types. By contrast, such pressures appear to be less important in shaping the pattern of focal CNAs observed within and between individual cancer types.

Supplementary Note 7: How to use the cancer copy number web portal

The cancer copy number portal accompanying this paper (**www.broadinstitute.org/tumorscape**) was designed to facilitate interpretation of this copy number dataset for the general research community. In addition to allowing download and visualization of both the raw and segmented copy number data, we have integrated a web service that allows for rapid querying of pre-processed analyses of the

copy number data for all the well-represented cancer subtypes in the dataset, as well as several defined aggregated datasets (such as all cancers, all epithelial cancers and all sarcoma cancers).

At present, there are two primary modes for querying these analyses: by gene and by cancer type. Below, we summarize the typical use case for each of these modes and present an outline for how to approach and interpret the portal data.

1) By Gene Analysis:

   The "By Gene" analysis mode is designed to quickly summarize the evidence that any given gene is the target of SCNA within a given cancer subtype. It is based on GISTIC analyses of 17 individual cancer types and an additional 6 aggregated datasets, as described in the Supplementary Methods above.

   To access, first click on the 'Analyses' tab on the navigation bar on the left side of the portal, then click on the 'by Gene' sub-tab. Enter the HUGO gene symbol (e.g. KRAS, MYC, CDK4) of any Refseq gene, then hit "Search". After a few seconds, the results from your gene should be loaded. You will see three tabs ("Summary", "Amplifications", and "Deletions"), followed by the gene symbol you queried and its genomic coordinates (in genome build hg18).

   Below that, you will see two paragraphs separately summarizing the evidence for that gene being a target of amplifications (first paragraph) and deletions (second paragraph). The first sentence of this summary paragraph states whether or not the gene is significantly amplified or deleted across the entire cancer copy number dataset, and whether or not the gene is present within a peak region of amplification or deletion in the entire dataset. A gene may be significantly altered but fail to reside within the peak region of alteration; although we cannot rule out the possibility that the gene is targeted by focal SCNAs, the fact that it is not in the peak region means that there is greater evidence for at least one other region on the same chromosome. Conversely, a gene may reside in a peak region of alteration but be insignificantly altered; this is usually due to an inability to confidently resolve the peak region and provides very little evidence that the gene is an actual target of SCNA. For genes that lie within a peak region of alteration, the number of additional genes in that peak are also listed; the fewer the genes in the peak, the more likely it is that that gene is the actual target.

   After the summary for the entire cancer dataset, we provide a summary of the results across the different independent cancer subtypes. In particular, we list the number of independent subtypes in which that gene was significantly altered and the number of subtypes in which the gene was located in a peak region of alteration. Because looking across many different datasets increases the likelihood that a gene will be in a peak region by chance alone, care must be taken before interpreting the significance of these numbers. For comparison, we list the fraction of genes in the genome which are significantly altered or located in a peak region of alteration in at least as many subtypes as the current gene of

interest. This allows some estimation of the likelihood that the gene in question is a false positive arising due to the number of hypotheses being tested.

To see more detailed information on the Amplifications or Deletions affecting this gene, click on the "Amplifications" or "Deletions" tab above the summary statements. This will load a table of the GISTIC results, where each row corresponds to one of the analyzed subsets. The rows are color-coded to quickly summarize the significance of SCNA for that gene and whether it is located in a peak region. For each row, we list the coordinates of the nearest peak region in that subtype (this will include the gene if it is located within a peak region) along with the number of genes in the peak and the false-discovery rate (FDR) q-value for the queried gene. The smaller the number of genes and the smaller the q-value, the more likely it is that the given gene is actually the target of SCNA in that cancer type. Note that when there are no peak regions identified in the chromosome in question in a cancer type, no peak region is listed and the number of genes is set to 0 by default.

We also list three different measures of the frequency of SCNA for the gene in each cancer type. Overall frequency measures the fraction of cancers that exhibit any SCNA at that gene. Focal frequency measures the fraction of cancers that exhibit SCNAs spanning less than half a chromosome arm in length. High-level frequency measures the fraction of cancers that exhibit SCNAs of greater than 1 copy. All these numbers are likely to be underestimates due to the effects of contaminating normal cells in many of the cancer samples and the limited resolution of the copy number platform.

There are several additional navigation features that can be unveiled by clicking on various parts of the table. Clicking on any underlined cancer subtype name will take you to the "By Cancer Type" analysis page for that subtype (see below). Clicking on the underlined coordinates for any peak region will open the copy number data in that region for that cancer type in the integrated genome viewer (IGV) (Robinson et al, in preparation). Finally, clicking anywhere else in any row with at least 1 gene in the nearest peak region will cause the gene symbols for the all genes in that peak to be listed in the sidebar to the right of the table. Clicking on any gene in this sidebar will load the "By Gene" analysis page for that gene.

2) By Cancer Type Analysis

The "By Cancer Type" analysis mode is designed to quickly summarize the significant regions of focal CNA within each cancer subtype. It is based on the same GISTIC analyses of 17 individual cancer types and an additional 6 aggregated datasets, as described in the Supplementary Methods and "By Gene" analysis section above.

To access, first click on the "Analyses" tab on the navigation bar on the left side of the portal, then click on the "By Cancer Type" sub-tab. By default, the

"all_cancers" subtype (representing all 3,131 cancer DNA samples present in our dataset) is selected first. By convention, aggregated tumor subsets are denoted by the prefix "all_" to distinguish them from individual cancer subtypes. To select a new cancer subtype, simply click the down arrow next to the name of the cancer type, select the cancer type of interest from the drop-down list, and hit "Search". After a few seconds, the data from that cancer type should be loaded.

The first tab you will see is the "Summary" tab, which contains a summary of the samples comprising the selected subset. In particular, we list the total number of DNA samples and cell lines for each subtype contained within that subset; for aggregated datasets, we also list the total number of samples and subtypes contained within the subset. Finally, we list the number of peak regions of focal SCNA identified in the dataset.

To view the regions of SCNA in more detail, click on the "Amplifications" or "Deletions" tab. This will load a table of the GISTIC results for that subset, sorted from most to least significant according to the FDR q-value. For each significant region of SCNA (represented by a single row in the table), we list the genomic coordinates of the peak region boundaries, the number of genes contained in the peak, the residual q-value for that peak (a measure of the likelihood that the peak was falsely discovered), and three different measures of the frequency of that event (as in the "By Gene" analysis described above). Note that the residual q-value for a peak will tend to differ from the overall q-value for genes in that peak, for two reasons: 1) the peak region may extend over genes with varied q-values, and 2) unlike the overall q-value, the residual q-value accounts for the possibility that a single SCNA may extend across more than one peak region by penalizing each of those peak regions (see Supplementary Methods).

As with the "By Gene" tables, clicking on any row with more than one gene in the peak will result in a list of the genes in that peak region appearing in the right-hand sidebar. Clicking on one of these genes will load the corresponding "By Gene" Analysis page. Clicking on the underlined peak region will load the copy number data for that region in the selected cancer subtype in the integrated genome viewer (IGV).

References

1 Haverty, P. M. *et al.* High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes Chromosomes Cancer* **47**, 530-542 (2008).

2 Nikolsky, Y. *et al.* Genome-wide functional synergy between amplified and mutated genes in human breast cancer. *Cancer Res* **68**, 9532-9540 (2008).

3 Chiang, D. Y. *et al.* Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* **68**, 6779-6788 (2008).

4 Lin, W. M. *et al.* Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer Res* **68**, 664-673 (2008).

5 George, R. E. *et al.* Genome-wide analysis of neuroblastomas using high-density single nucleotide polymorphism arrays. *PLoS ONE* **2**, e255 (2007).

6 Demichelis, F. *et al.* Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosomes Cancer* **48**, 366-380 (2009).

7 Beroukhim, R. *et al.* Patterns of gene expression and copy-number alterations in VHL disease-associated and sporadic clear cell carcinoma of the kidney. *Cancer Res* **69**, 4674-4681 (2009).

8 Kilpivaara, O. *et al.* A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nat Genet* **41**, 455-459 (2009).

9 Ramos, A. *et al.* Amplification of PDGFRA and KIT in Non-small Cell Lung Cancer. *Cancer Biology and Therapy* **8**, Epub ahead of print (2009).

10 GlaxoSmithKline. *GSK Cancer Cell Line Genomic Profiling Data*, <https://cabig.nci.nih.gov/tools/caArray_GSKdata> (2008).

11 Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758-764 (2007).

12 Mullighan, C. G. *et al.* BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**, 110-114 (2008).

13 Bass, A., Watanabe, H., Yu, S. & al, e. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinoma. *Nat Genet* **Epub ahead of print** (2009).

14 Maher, E. A. *et al.* Marked Genomic Differences Characterize Primary and Secondary Glioblastoma Subtypes and Identify Two Distinct Molecular and Clinical Secondary Glioblastoma Entities. *Cancer Res* **66**, 11502-11513 (2006).

15 Roudier, M. P. *et al.* Phenotypic heterogeneity of end-stage prostate carcinoma metastatic to bone. *Hum Pathol* **34**, 646-653 (2003).

16 Rubin, M. A. *et al.* Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res* **6**, 1038-1045 (2000).

17 Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2**, research0032.0031 - research0032.0011 (2001).

18 Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS* **98**, 31-36 (2001).

19 Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104**, 20007-20012 (2007).

20 Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F. & Barillot, E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413-3422 (2004).

21 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat* **5**, 557-572 (2004).

22 McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-1174 (2008).

23 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc, Ser B* **57**, 289-300 (1995).

24 Gould, J., Getz, G., Monti, S., Reich, M. & Mesirov, J. P. Comparative gene marker selection suite. *Bioinformatics* **22**, 1924-1925 (2006).

25 Reich, M. *et al.* GenePattern 2.0. *Nat Genet* **38**, 500-501 (2006).

26 Dorschner, M. O., Sybert, V. P., Weaver, M., Pletcher, B. A. & Stephens, K. NF1 microdeletion breakpoints are clustered at flanking repetitive sequences. *Hum Mol Genet* **9**, 35-46 (2000).

27 Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).

28 Firestein, R. *et al.* CDK8 is a colorectal cancer oncogene that regulates [bgr]-catenin activity. *Nature* **455**, 547-551 (2008).

29 Lundberg, A. S. *et al.* Immortalization and transformation of primary human airway epithelial cells by gene transfer. *Oncogene* **21**, 4577-4586 (2002).

30 Zhao, X. *et al.* An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Res* **64**, 3060-3071 (2004).

31 Wee, S. *et al.* PTEN-deficient cancers depend on PIK3CB. *Proc Natl Acad Sci U S A* **105**, 13057-13062 (2008).

32 Wiederschain, D. *et al.* Single-vector inducible lentiviral RNAi system for oncology target validation. *Cell Cycle* **8**, 498-504 (2009).

33 Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283-1298 (2006).

34 Naldini, L. *et al.* In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**, 263-267 (1996).

35 Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery in vivo. *Nat Biotechnol* **15**, 871-875 (1997).

36 Gaither, A. *et al.* A Smac mimetic rescue screen reveals roles for inhibitor of apoptosis proteins in tumor necrosis factor-alpha signaling. *Cancer Res* **67**, 11493-11498 (2007).

[37] Morgenstern, J. P. & Land, H. Advanced mammalian gene transfer: high titre retroviral vectors with multiple drug selection markers and a complementary helper-free packaging cell line. *Nucleic Acids Res* **18**, 3587-3596 (1990).

[38] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).

[39] Merlo, L. M. F., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924-935 (2006).

[40] Iwasa, Y., Michor, F. & Nowak, M. A. Stochastic tunnels in evolutionary dynamics. *Genetics* **166**, 1571-1579, doi:166/3/1571 [pii] (2004).

[41] Getz, G. *et al.* Comment on "The Consensus Coding Sequences of Human Breast and Colorectal Cancers:. *Science* **317**, 1500 (2007).

[42] Rubin, A. F. & Green, P. Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500 (2007).

[43] Forrest, W. F. & Cavet, G. Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500; author reply 1500 (2007).

[44] Davies, H. *et al.* Somatic Mutations of the Protein Kinase Gene Family in Human Lung Cancer. *Cancer Res* **65**, 7591-7595 (2005).

[45] Diskin, S. J. *et al.* STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res* **16**, 1149-1158 (2006).

[46] Guttman, M. *et al.* Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet* **3**, e143 (2007).

[47] Taylor, B. S. *et al.* Functional copy-number alterations in cancer. *PLoS ONE* **3**, e3179 (2008).

[48] Hogg, R. V., Craig, A. & McKean, J. W. *Introduction to Mathematical Statistics*. (Prentice Hall, 2004).

[49] Myllykangas, S. *et al.* DNA copy number amplification profiling of human neoplasms. *Oncogene* **25**, 7324-7332 (2006).

[50] Baudis, M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer* **7**, 226 (2007).

[51] Mitelman Database of Chromosome Aberrations in Cancer (2009). Mitelman F, Johansson B and Mertens F (Eds.), http://cgap.nci.nih.gov/Chromosomes/Mitelman

[52] NCI and NCBI's SKY/M-FISH and CGH Database (2001), http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi

[53] Morris, E. J. *et al.* E2F1 represses [bgr]-catenin transcription and is antagonized by both pRB and CDK8. *Nature* **455**, 552-556 (2008).

[54] Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A* **102**, 9625-9630 (2005).

[55] Wozniak, A. *et al.* Array CGH analysis in primary gastrointestinal stromal tumors: cytogenetic profile correlates with anatomic site and tumor aggressiveness, irrespective of mutational status. *Genes Chromosomes Cancer* **46**, 261-276 (2007).

[56] Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* **16**, 1465-1479 (2006).

[57] Torring, N. *et al.* Genome-wide analysis of allelic imbalance in prostate cancer using the Affymetrix 50K SNP mapping array. *Br J Cancer* **96**, 499-506 (2007).

[58] Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-898 (2007).

[59] Martin, E. S. *et al.* Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. *Cancer Res* **67**, 10736-10743 (2007).

[60] Cancer_Genome_Atlas_Research_Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).

[61] Paulsson, K. *et al.* Microdeletions are a general feature of adult and adolescent acute lymphoblastic leukemia: Unexpected similarities with pediatric disease. *Proc Natl Acad Sci U S A* **105**, 6708-6713 (2008).

[62] Zender, L. *et al.* An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135**, 852-864 (2008).

[63] Etemadmoghadam, D. *et al.* Integrated genome-wide DNA copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. *Clin Cancer Res* **15**, 1417-1427 (2009).

[64] Northcott, P. A. *et al.* Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. *Nat Genet* **41**, 465-472 (2009).

[65] Sheffer, M. *et al.* Association of survival and disease progression with chromosomal instability: a genomic exploration of colorectal cancer. *Proc Natl Acad Sci U S A* **106**, 7131-7136 (2009).

[66] Giubellino, A., Burke, T. R., Jr. & Bottaro, D. P. Grb2 signaling in cell motility and cancer. *Expert Opin Ther Targets* **12**, 1021-1033 (2008).

[67] Myllykangas, S., Bohling, T. & Knuutila, S. Specificity, selection and significance of gene amplifications in cancer. *Semin Cancer Biol* **17**, 42-55 (2007).

[68] Sos, M. L. *et al.* Predicting drug activity in non-small cell lung cancer based on genetic lesions. *JCI* **119**, 1727-1740 (2009).