nature genetics

OPEN

Pan-cancer patterns of somatic copy number alteration

Travis I Zack[1–3,11], Steven E Schumacher[1,2,11], Scott L Carter[1], Andrew D Cherniack[1], Gordon Saksena[1], Barbara Tabak[1], Michael S Lawrence[1], Cheng-Zhong Zhang[1], Jeremiah Wala[1,2,4,5], Craig H Mermel[1], Carrie Sougnez[1], Stacey B Gabriel[1], Bryan Hernandez[1], Hui Shen[6], Peter W Laird[6], Gad Getz[1,12], Matthew Meyerson[1,7–9,12] & Rameen Beroukhim[1,2,7,8,10,12]

Determining how somatic copy number alterations (SCNAs) promote cancer is an important goal. We characterized SCNA patterns in 4,934 cancers from The Cancer Genome Atlas Pan-Cancer data set. Whole-genome doubling, observed in 37% of cancers, was associated with higher rates of every other type of SCNA, *TP53* mutations, *CCNE1* amplifications and alterations of the PPP2R complex. SCNAs that were internal to chromosomes tended to be shorter than telomere-bounded SCNAs, suggesting different mechanisms underlying their generation. Significantly recurrent focal SCNAs were observed in 140 regions, including 102 without known oncogene or tumor suppressor gene targets and 50 with significantly mutated genes. Amplified regions without known oncogenes were enriched for genes involved in epigenetic regulation. When levels of genomic disruption were accounted for, 7% of region pairs were anticorrelated, and these regions tended to encompass genes whose proteins physically interact, suggesting related functions. These results provide insights into mechanisms of generation and functional consequences of cancer-related SCNAs.

SCNAs affect a larger fraction of the genome in cancers than do any other type of somatic genetic alteration[1–5]. SCNAs have critical roles in activating oncogenes and in inactivating tumor suppressors[3,6–12], and an understanding of the biological and phenotypic effects of SCNAs has led to substantial advances in cancer diagnostics and therapeutics[13–16].

A primary challenge in understanding SCNAs is to distinguish the driver events that contribute to oncogenesis and cancer progression from the passenger SCNAs that are acquired during cancer evolution but do not contribute toward it[17–20]. Positively selected SCNAs will tend to recur across cancers at elevated rates[1,4,5]. However, SCNAs may also recur in the absence of positive selection owing to increased rates of generation or decreased negative selection[21,22]. For this reason, it is important to understand how mechanisms of SCNA generation, their temporal ordering and negative selection shape the distribution of SCNAs across the genome[21–25].

A second challenge is to identify the oncogene and tumor suppressor gene targets of driver SCNAs (which often encompass many genes) and elucidate the functional roles of SCNAs. The context of SCNAs can be informative. Positive correlations with other genetic events may indicate functional synergies, whereas anticorrelations may indicate functional redundancies, as redundant events would not be required by the same cancer. Several approaches have been developed to determine the functional effects of genetic events through the analysis of anticorrelation patterns[26–28].

Here we address these challenges through the analysis of 4,934 cancer copy number profiles across 11 cancer types, assembled through The Cancer Genome Atlas Pan-Cancer effort, enabling analysis of large numbers of cancers and comparison of patterns of copy number change across cancer types. We have integrated rigorous statistical approaches into these analyses, including absolute allelic copy number profiling[29], as well as novel computational tools to determine individual SCNA events and their temporal ordering from these profiles and to identify functionally relevant correlations between SCNAs.

RESULTS

Cancer purities and ploidies and rates of copy number alteration within and across cancer types

We analyzed the copy number profiles of 4,934 primary cancer specimens across 11 cancer types (minimum of 136 samples for bladder cancer; maximum of 880 samples for breast cancer; colon and rectal adenocarcinomas were combined; **Supplementary Table 1**). In each cancer, we determined copy numbers at each of 1,559,049 loci relative to the median copy number across the genome, using Affymetrix SNP6 arrays and previously described algorithms[1]. For 3,847 cancers, we also determined purity, ploidy and absolute allelic copy number profiles[29] for the malignant cells using SNP6 array data and, in 1,069 cases, matched whole-exome sequencing data (**Supplementary Table 1**). In the other 1,087 cases, purity and ploidy estimates were ambiguous and were left uncalled. This second group included all cases of acute myeloid leukemia (LAML), which exhibited very few SCNAs.

[1]Broad Institute, Cambridge, Massachusetts, USA. [2]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [3]Biophysics Program, Harvard University, Boston, Massachusetts, USA. [4]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [6]USC Epigenome Center, University of Southern California, Los Angeles, California, USA. [7]Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [8]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [9]Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. [10]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [11]These authors contributed equally to this work. [12]These authors jointly directed this work. Correspondence should be addressed to R.B. (rameen_beroukhim@dfci.harvard.edu)

Received 10 August; accepted 20 August; published online 26 September 2013; doi:10.1038/ng.2760

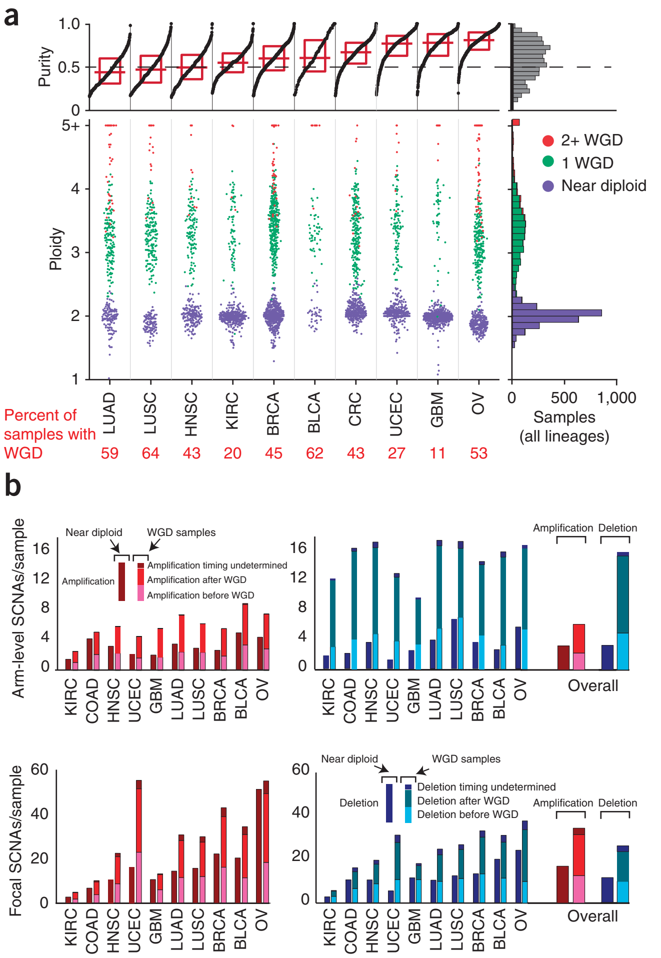

Figure 1 Distribution of SCNAs across lineages. (**a**) Sample purity (top) and ploidy (bottom) across lineages (LUAD, lung adenocarcinoma; LUSC, lung squamous cell; HNSC, head and neck squamous cell; KIRC, kidney renal cell; BRCA, breast; BLCA, bladder; CRC, colorectal; UCEC, uterine cervix; GBM, glioblastoma multiformae; OV, ovary). Box plots show the median, first quartile and third quartile of purity in each lineage. Near-diploid samples are designated in purple; cancers that have undergone one or more than one WGD event are designated in green and red, respectively. Summary data for all lineages are indicated on the right. (**b**) Numbers of arm-level (top) and focal (bottom) amplifications (left) and deletions (right) across lineages. For each lineage, near-diploid samples and those with WGD events are indicated by bars on the left and right, respectively; SCNA in samples with WGD are resolved according to their timing relative to the WGD event.

We then inferred the sequence of SCNA events that led to each copy number profile, using the most parsimonious set of SCNAs that could generate the observed absolute allelic copy numbers (Online Methods and **Supplementary Fig. 1a**). We determined the lengths, locations and numbers of copies changed for each SCNA and, in many cases, allelic structure (**Supplementary Fig. 1b**). We identified a total of 202,244 SCNAs, a median of 39 per cancer sample, comprising 6 categories: focal SCNAs that were shorter than the chromosome arm (median of 11 amplifications and 12 deletions per sample); arm-level SCNAs that were chromosome-arm length or longer (median of 3 amplifications and 5 deletions per sample); copy-neutral loss-of-heterozygosity (LOH) events in which one allele was deleted and the other was amplified coextensively (median of 1 per sample); and whole-genome duplications (WGDs; in 37% of cancers). By amplifications and deletions, we refer to copy number gains and losses, respectively, of any length and amplitude.

Estimated purities and ploidies per cancer varied substantially within and across lineages (**Fig. 1a**). Purity estimates correlated with estimates derived from measurements of leukocyte and lymphocyte contamination using DNA methylation data from the same cancers (**Supplementary Fig. 1c**) (H.S., L. Yao, T. Tiche Jr., T. Hinoue, C. Kandoth *et al.*, unpublished data) but tended to indicate lower purity, consistent with the presence of non-hematopoietic contaminating normal cells. Average ploidies within lineages mirrored WGD frequencies. The average estimated ploidy within samples that had undergone a single WGD was 3.31 (not 4), suggesting that WGD events are associated with large amounts of genome loss. By contrast, samples that had not undergone WGD had an average estimated ploidy of 1.99.

Compared to the near-diploid cancers within each lineage, cancers with WGD had higher rates of every other type of SCNA (**Fig. 1b**) and twice the rate of SCNAs overall. Across lineages, overall SCNA rates largely reflected rates of WGD (**Supplementary Fig. 1d**).

In cancers with WGD, most other SCNAs occurred after WGD (**Fig. 1b** and Online Methods). The fractions of amplifications and deletions that were estimated to occur before WGD were highly correlated across lineages ($R = 0.64$; **Supplementary Fig. 1e**), indicating a consistent estimate for the timing of WGD with respect to other SCNAs. WGD was inferred to occur earliest relative to focal SCNAs among lineages where WGD was common (ovarian, bladder and colorectal cancers) and after most focal SCNAs in lineages in which WGD was least common (glioblastoma and kidney clear-cell carcinoma).

SCNA lengths suggest varied mechanisms of generation

Focal SCNAs for which one boundary is the telomere (telomere bounded) tended to be longer than SCNAs for which both boundaries were internal to the chromosome (median SCNA lengths for telomere-bounded and internal events respectively: amplifications, 19.6 Mb versus 0.9 Mb; deletions, 22.7 Mb versus 0.7 Mb). These differences reflect differences across the entire length distributions of telomere-bounded and internal events. Focal internal SCNAs were observed at frequencies inversely proportional to their lengths (**Fig. 2a** and **Supplementary Fig. 2a,b**), as noted previously[1]. However, telomere-bounded SCNAs tended to follow a superposition of 1/length and uniform length distributions. These distributions were the same whether measuring distance by kilobase, number of array markers or number of genes, indicating that this difference in length does not result from variation in array resolution or gene density across the genome (data not shown). Focal, telomere-bounded SCNAs also accounted for more SCNAs than expected assuming random SCNA locations (12% and 26% of focal amplifications and deletions, respectively; $P < 0.0001$). Both telomere-bounded and internal SCNAs were more likely to end within the centromere than expected given the centromere's length (**Supplementary Fig. 2c**), but differences in their length distributions remained when centromere-bounded events were excluded. Differences between telomere-bounded and internal SCNAs were even more marked for copy-neutral LOH events and displayed no correlation across lineages (**Supplementary Fig. 2d**).

We detected chromothripsis in 5% of samples, ranging from 0% of head and neck squamous cell carcinomas to 16% of glioblastomas (**Fig. 2b** and Online Methods). The rate of chromothripsis was not related to overall rates of SCNA ($R = 0.13$; $P = 0.3$). As previously reported[30], samples with chromothripsis were more likely to have chromothripsis on more than 1 chromosome (14/122 samples with chromothripsis had 2 or 3 such events; $P = 0.003$).

Many chromothripsis events were concentrated in a few genomic regions, often associated with known driver events (**Fig. 2c**). In glioblastomas,

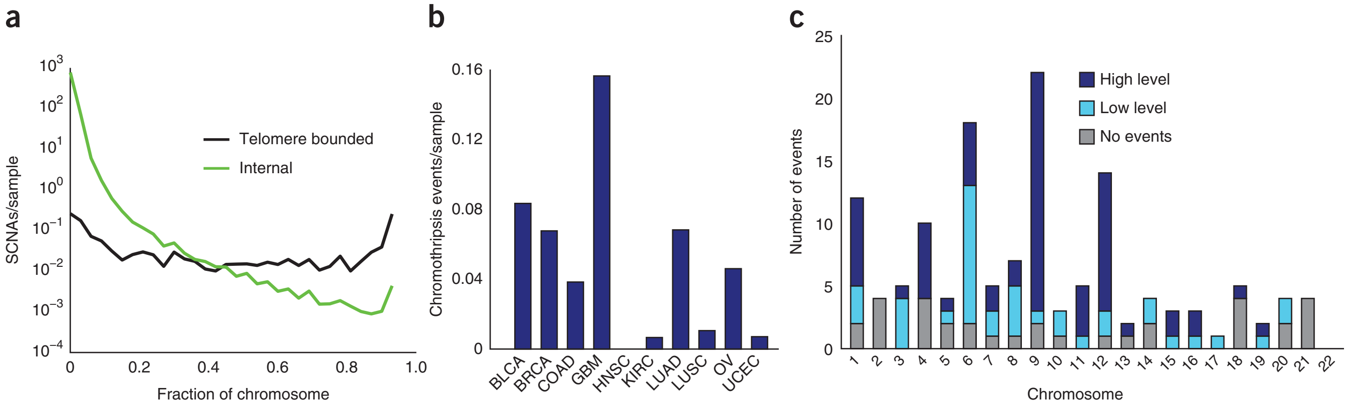

Figure 2 Characteristics of different types of SCNAs. (**a**) Distribution of lengths of SCNAs originating at telomeres compared to those of SCNAs that are internal to the chromosome. (**b**) Rates of chromothripsis across lineages. (**c**) Rates of chromothripsis across chromosomes. Chromothripsis events that involved peak regions of amplification and deletion are indicated in blue (dark blue, amplifications resulting in >4.4 copies or deletions resulting in <1 copy; light blue, low-level events involving smaller changes); events that do not involve peak regions are shown in gray.

chromothripsis events were concentrated on chromosomes 9 and 12 and corresponded, respectively, with homozygous loss of *CDKN2A* (20/22 samples) and coamplification of discontinuous regions containing *CDK4* and *MDM2* (9/12 samples). Across all cancers, 72% of chromothripsis events included a GISTIC peak region (see below).

Recurrent focal SCNAs

We identified 70 recurrently amplified and 70 recurrently deleted regions in a unified 'pan-cancer' analysis across all lineages (**Fig. 3a**, **Supplementary Fig. 2e** and **Supplementary Table 2**). For each of these 140 regions, we identified a 'peak' region that is most likely to contain oncogenes or tumor suppressor genes targeted by these SCNAs. SCNAs involving these regions included 21% of all focal amplifications and 23% of all focal deletions. Focal SCNAs within peak regions tended to be shorter than focal SCNAs elsewhere on the chromosome (median of 12.2 Mb in peak regions versus 19.4 Mb across the genome; $P < 0.0001$) and were more often high-amplitude events ($P < 0.0001$). The number of focal SCNAs involving peak regions per sample tracked the total number of SCNAs ($r = 0.84$; $P < 0.0001$), ranging from 0.4 focal SCNAs in the typical acute myeloid leukemia to 12.3 focal SCNAs in the typical ovarian cancer (mean across all lineages of 5.2).

Tissue types from similar lineages tended to have similar rates of amplification and deletion in peak SCNA regions (**Fig. 3a**). We observed clusters of squamous cell carcinomas (head and neck squamous cell carcinoma, lung squamous cell carcinoma and bladder cancer) and reproductive cancers (ovarian and endometrial cancer) with breast cancer.

The 70 peak regions of amplification contained a median of 3 genes each (including microRNAs), with 60 peaks containing fewer than 25 genes. Twenty-four of these peak regions contained an oncogene known to be activated by amplification (**Supplementary Table 2**), including seven of the top ten regions (*CCND1*, *EGFR*, *MYC*, *ERBB2*, *CCNE1*, *MCL1* and *MDM2*). The ninth and tenth most significant regions (11q14.1 and 8p11.23, respectively) did not contain known oncogenes, but the latter contained the histone methyltransferase *WHSC1L1* and was 18 kb from the known amplified oncogene *FGFR1*. The fourth most significantly amplified peak region (3q26.2) contained *TERC*, which encodes the RNA substrate for the known oncogene *TERT*, which is itself in a peak region of amplification (5p15.33). Another peak with eight genes (9p13.3) contained *RMRP*, another *TERT*-associated RNA[31].

The 70 peak regions of deletion contained a median of 4 genes (including microRNAs), with 52 peaks containing fewer than 25 genes. Twenty-two of these regions contained one of the 100 largest genes in the genome, and 12 contained known tumor suppressor genes (**Supplementary Table 2**; two additional large regions contained the known tumor suppressor genes *ATM* and *NOTCH1*). Four other regions each contained a single gene (*PPP2R2A*, *PTTG1IP*, *FOXK2* and *LINC00290*). We discuss *PPP2R2A* and its binding partner *PPP2R1A* (which is significantly mutated in the same set of cancers (Lawrence *et al.*[32] and M.S.L., P. Stojanov, C.H.M., G. Kryukov, S.E.S. *et al.*, unpublished data) in greater detail below. *LINC00290* is a long noncoding RNA, a member of a group whose role in cancer is increasingly being appreciated[33,34]. Two other regions contained suspected tumor suppressor genes (*ERRFI1* (ref. 35) and *FOXC1* (ref. 36)).

The features most associated with genes in the amplification and deletion peak regions are known to be associated with cancer (**Fig. 3b**). We applied GRAIL[37], which uses literature citations, to find common features of genes in selected regions of the genome. We considered amplifications and deletions separately and analyzed only peaks with fewer than 25 genes.

Of the 37 peak regions of amplification with fewer than 25 genes and without known targets (**Supplementary Table 2**), the most associated features were related to epigenetic and mitochondrial regulation: 'histone', 'cytochrome', 'mitochondrial' and 'acetyltransferase' (**Fig. 3b**). Thirteen of these 37 regions contained chromatin state and histone-modifying genes (**Supplementary Table 2**), reflecting significant enrichment ($P < 0.0001$)[38]. Of these, five (*BRD4*, *KAT6A*, *KAT6B*, *NSD1* and *PHF1*) are subject to recurrent rearrangements in leukemias, sarcomas and midline carcinomas[39–43]. The *BRD4* peak also contained *NOTCH3*, another potential oncogene[44]. Two others, *KDM2A* and *KDM5A*, are reported to regulate the activity of *TP53* and *RB1*, respectively[45,46]. The finding that multiple peak regions of amplification contain epigenetic regulators is consistent with growing evidence suggesting that epigenetic alterations and chromatin remodeling have a critical role in many forms of cancer[47–49]. Ten regions contained genes encoding mitochondria-associated proteins (**Supplementary Table 2**); none of these are subject to recurrent rearrangements in cancer. The 21 peak regions of deletion with fewer than 25 genes and without known tumor suppressor genes or large genes were most associated with 'Phosphatase', 'RNAi', 'PTEN' and 'Prostate'.

Fifty of the 140 peak regions contained a significantly mutated gene, including 23 regions without known oncogene or tumor suppressor gene targets and 32 regions with fewer than 25 genes (**Supplementary**

Figure 3 Significantly recurrent focal SCNAs. (**a**) Frequencies of amplification minus frequencies of deletion (red and blue indicate greater frequencies of amplifications and deletions, respectively) across lineages (*x* axis; see **Supplementary Table 1** for a list of lineage abbreviations) for all 84 significant peak regions of SCNA, arranged in order of significance (*y* axis). The ordering of lineages reflects the results of unsupervised hierarchical clustering of these data. Magnified views of the values for the ten most significant amplification and deletion peaks are shown to the right, alongside candidate targets for these regions. Criteria for selecting the indicated candidates are described in the Online Methods. (**b**) Associated terms in the literature in peak regions containing fewer than 25 genes, according to a GRAIL analysis of all peak regions (top) and peak regions without known cancer genes or large genes (bottom). (**c**) Schematic of the locations of peak regions within chromosomes 4 and 8 (other chromosomes are shown in **Supplementary Fig. 3**) across cancer types (designated by boxes above and below colored as in **a**) and the pan-cancer analysis (right-most column, denoted by a black line). Peaks are designated by candidate targets for each region, selected according to the criteria described in the Online Methods.

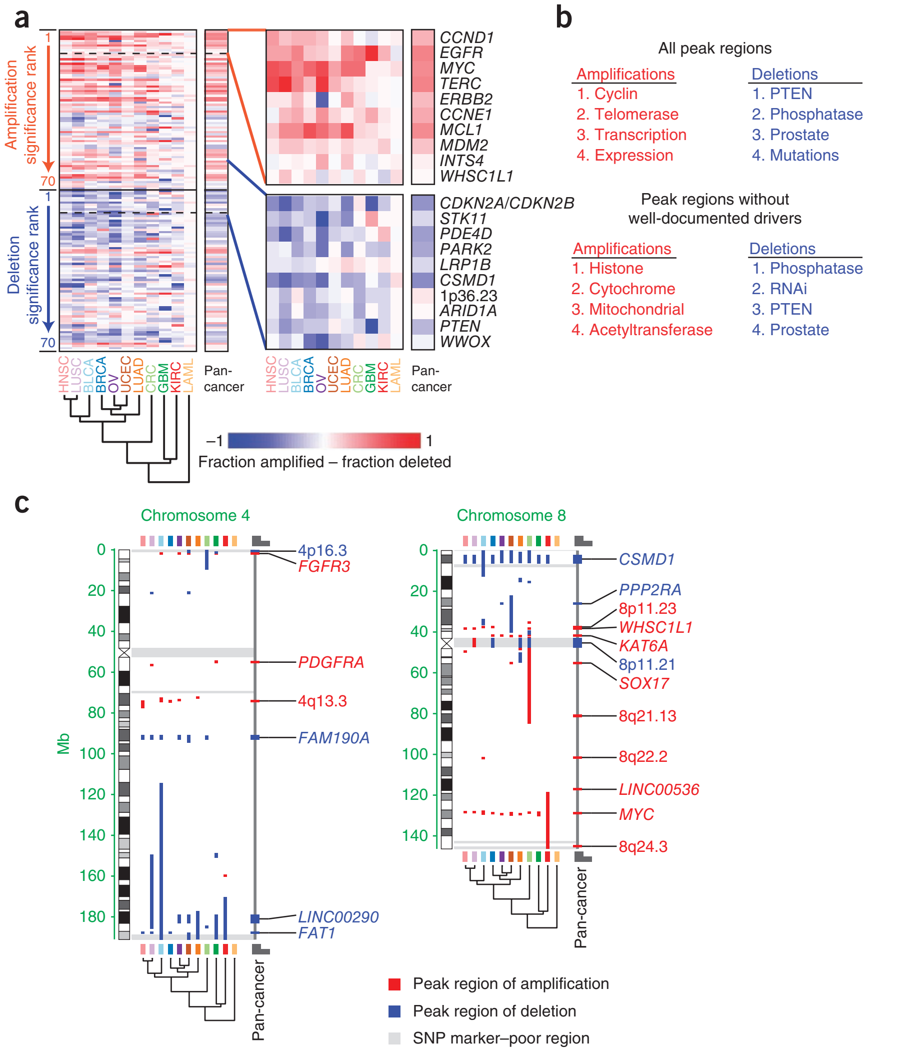

Table 2). We calculated the significance of mutations (including both point mutations and small insertion-deletion events identified in the paired sequencing data) for each gene in each region using our unpublished methods (Lawrence *et al.*[32] and M.S.L., P. Stojanov, C.H.M., G. Kryukov, S.E.S. *et al.*, unpublished data) and corrected for multiple hypotheses reflecting the number of genes in the region. In 3 cases, there were 2 significantly mutated genes per peak, for a total of 35 significantly mutated genes. These 35 genes included 8 of the 23 known amplification-activated oncogenes and all of the 12 known tumor suppressor genes in these peak regions (**Supplementary Table 2**). An additional 2 of the 35 genes (both in amplification peaks) are oncogenes known to be activated by mutations but not by amplifications.

Frameshift and nonsense mutations that are likely to cause loss of function were significantly enriched in genes in deleted regions ($P = 0.0002$), accounting for 19% of these mutations compared to 12% of mutations found in genes in amplified regions. We excluded regions with known oncogenes or tumor suppressor genes or with more than 25 genes from this analysis. These findings are consistent with the prediction that deleted regions without known tumor suppressors are enriched for novel tumor suppressors or genes whose functions are nonessential.

Most peak regions in lineage-specific analyses intersected peak regions in other lineages, and, indeed, in the pan-cancer analysis (**Fig. 3c** and **Supplementary Fig. 3**). We obtained a median of 74 peak regions for each lineage (ranging from 25 in acute myeloid leukemia to 95 in endometrial cancer; 42% were amplification peaks, and 58% were deletion peaks; **Supplementary Table 3**), resulting in a total of 770 peak regions. Of these, 84% intersected peak regions in at least one other lineage ($P < 0.0001$), and 65% intersected peak regions in the pan-cancer analysis. Peak regions tended to be larger in the lineage-specific analyses than in the pan-cancer analysis (1.4 Mb versus 0.7 Mb, respectively), indicating that the pan-cancer analysis has improved resolution.

Nevertheless, some significant SCNAs were identified in lineage-specific analysis but not in the pan-cancer analysis. Across all lineages, we identified 229 peaks not present in the pan-cancer analysis, including amplifications of the known amplified oncogenes *MET*, *CCND2*, *ERBB3* and *MYCN* and deletions of the known tumor suppressor genes *TP53* and *CDKN2C*.

Correlations reflect overall levels of genomic disruption

For each pair of peak regions, we looked for positive and negative correlations between focal SCNAs involving these regions (**Fig. 4a**). We compared the number of samples with SCNAs involving both regions between observed data and permuted data in which SCNAs were randomly assigned to samples while maintaining genomic positions and SCNA structure. We only permuted SCNAs within lineages (and sublineages when available) to avoid lineage-dependent confounders and evaluated correlations between regions on different chromosomes to avoid correlations due to chromosomal structure (Online Methods). We focused on peak regions with less than 25 genes.

We identified significant positive correlations ($q < 0.25$) between 53% of region pairs but no significant anticorrelations (**Fig. 4b**). The high rate of positive correlations results from widely differing levels

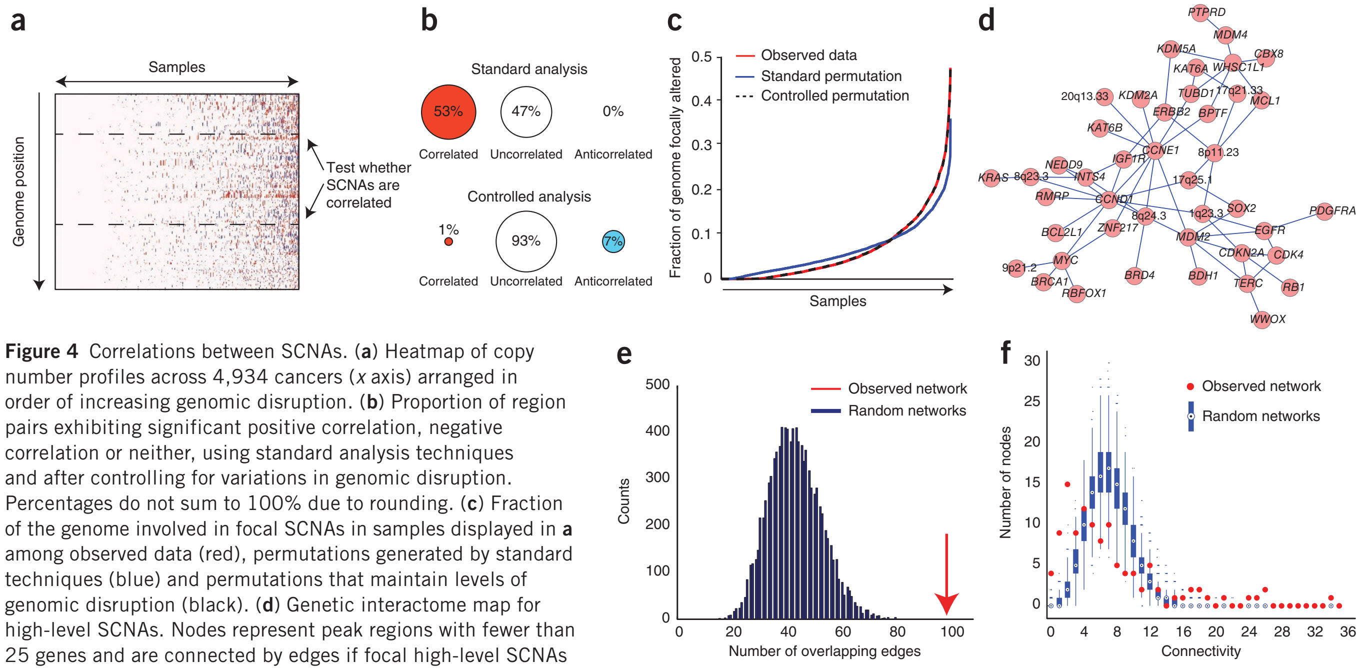

Figure 4 Correlations between SCNAs. (**a**) Heatmap of copy number profiles across 4,934 cancers (*x* axis) arranged in order of increasing genomic disruption. (**b**) Proportion of region pairs exhibiting significant positive correlation, negative correlation or neither, using standard analysis techniques and after controlling for variations in genomic disruption. Percentages do not sum to 100% due to rounding. (**c**) Fraction of the genome involved in focal SCNAs in samples displayed in **a** among observed data (red), permutations generated by standard techniques (blue) and permutations that maintain levels of genomic disruption (black). (**d**) Genetic interactome map for high-level SCNAs. Nodes represent peak regions with fewer than 25 genes and are connected by edges if focal high-level SCNAs (amplification resulting in >4.4 copies and deletions resulting in <1 copy) are significantly anticorrelated. (**e**) Numbers of significant anticorrelations that overlap known protein-protein interactions in the observed genetic interactome network (red arrow) and permuted networks (blue bars). These results are from the analysis of all SCNAs; results from the analysis of high-level alterations are displayed in **Supplementary Figure 4d**. (**f**) Distribution of connectivity values (number of nodes to which each node is connected) for the observed genetic interactome network (red circles) and permuted networks (box plots; boxes indicate first and third quartiles, whiskers represent 99% confidence bounds, and dots represent outliers) in the all-SCNA analysis.

of genomic disruption across samples, which are not maintained in permuted data sets (**Fig. 4c**). Similar results were obtained with other standard statistical approaches such as Fisher's exact tests (data not shown). These findings indicate that varying levels of overall genomic disruption confound analyses of functionally relevant correlations between SCNAs.

We therefore re-evaluated correlations between SCNAs after controlling for genomic disruption by maintaining in the permuted data the fractions of the genome affected by each of the amplifications and deletions in each sample (**Fig. 4c**, Online Methods and **Supplementary Fig. 4a**,**b**). We performed the analysis in two ways: evaluating all SCNAs (**Supplementary Table 4**) and evaluating only high-level amplifications and homozygous deletions (Online Methods and **Supplementary Table 4**). In many cases, high-level amplification or homozygous deletion may be necessary to activate an oncogene or to inactivate a tumor suppressor gene[16], and, in such cases, correlated features may be masked by noise in lower level events.

When evaluating all SCNAs, we identified significant positive correlations between <1% of region pairs (40 interactions; **Supplementary Table 4**) and anticorrelations between 7% of region pairs (396 interactions; **Fig. 4b** and **Supplementary Table 4**). Correcting for genomic disruption altered the estimated significance of these interactions and also changed the rank ordering of the significance estimates (**Supplementary Fig. 4c**). High-level amplifications and homozygous deletions were relatively rare, limiting our power to detect anticorrelations in the analysis of high-level alterations. Of the 1,094 interactions we were powered to detect, we observed positive correlations between <1% of region pairs (3 interactions; **Supplementary Table 4**) and anticorrelations between 10% of region pairs (108 interactions; **Fig. 4d** and **Supplementary Table 4**). The three correlations included deletions of *CDKN2A* with amplifications of *EGFR*, amplifications of *PDGFR* with amplifications of *CDK4*, and deletions of *PPP2RA* with amplifications of 19p13.2.

We predicted that anticorrelated SCNAs would often indicate functional redundancies, and, therefore, genes in the affected regions would often be in similar pathways and interact physically. We tested this hypothesis by comparing networks representing significantly anticorrelated SCNAs (anticorrelation networks) with DAPPLE, a set of curated protein-protein interactions (PPIs)[37] (Online Methods).

Networks formed by our anticorrelation analyses and by PPIs significantly overlapped ($P < 0.0001$ and 0.006 for all-SCNA analysis and analysis of high-level alterations, respectively; **Fig. 4e** and **Supplementary Fig. 4d**). For example, in the analysis of all SCNAs, we observed 100 overlapping edges, a 2-fold increase over the 43.4 overlapping edges expected by chance. This significance was not observed for correlated events ($P = 1$ for analyses of both all SCNAs and high-level alterations). These results suggest that the observed anticorrelations are related to biological interactions.

Anticorrelation networks were enriched for both isolated nodes and highly connected 'hub' regions (**Fig. 4f**). To analyze the structure of these networks, we generated control anticorrelation networks representing the most significant edges from permuted data in which we had randomized the SCNA sample assignments within each lineage. In the all-SCNA analysis, 28 regions were anticorrelated with fewer than 3 other regions, compared to 3 isolated nodes in the average permutation ($P < 0.01$).

The isolated nodes in the all-SCNA analysis were enriched for regions containing large genes (including 10 of 28 such regions; $P = 0.004$). Conversely, they trended toward excluding regions with known oncogenes or tumor suppressor genes (5 of 35 such regions; $P = 0.06$). Most peak regions exhibited fewer anticorrelations in the analysis of high-level alterations, possibly owing to decreased power. The most extreme exception involved *CDKN2A*, which anticorrelated with 14 regions in the analysis of high-level alterations and with only

9 regions in the all-SCNA analysis. Consistent with these findings, *CDKN2A* is often inactivated by homozygous deletions.

We applied a similar analysis to identify events associated with WGD. We included both SCNAs and mutations, using the 200 most significantly mutated genes across The Cancer Genome Atlas Pan-Cancer data set (Online Methods, refs. 32,50 and M.S.L., P. Stojanov, C.H.M., G. Kryukov, S.E.S. *et al.*, unpublished data). Three SCNA peak regions and two significantly mutated genes correlated with WGD (**Supplementary Table 4**). *TP53* mutations and *CCNE1* amplifications correlated with WGD; both have been functionally associated with tolerance of tetraploidy in experimental models[51–54]. Our findings indicate that these associations apply to human tumors across multiple lineages. We also found that deletions of *PPP2R2A* and mutations of its binding partner *PPP2R1A* were correlated with WGD. These two genes belong to phosphoprotein phosphatase complex 2 (PPP2), which regulates mitotic spindle formation and can lead to chromosomal mis-segregation and abnormal mitoses when depleted[55,56].

Eleven genetic events anticorrelated with WGD, including two amplifications, five deletions and four mutations. (**Supplementary Table 4**). The deletions included *CDKN2A*, *PTEN* and *NF1*, and three of the four mutations also involved genes known as or proposed to be tumor suppressors (*CTCF*[57], *MAP3K1* (ref. 9) and *ATM*). The anticorrelations of these tumor suppressors may result from a greater difficulty in biallelically inactivating tumor suppressor genes in samples with extra copies subsequent to WGD[29].

DISCUSSION

This study represents the largest analysis so far of high-resolution copy number profiles generated using a single platform and the first large-scale analysis of absolute allelic copy number data across cancer types. We identified common patterns of SCNA across cancer types, including a tendency for telomeric events to be longer and more frequent than SCNAs within chromosomes and for duplications of large regions of the genome (through WGD or polysomy) to lead to subsequent increases in the numbers of SCNAs (especially deletions) in the duplicated regions. SCNAs also tend to reside in the same regions of the genome across different cancer types.

A primary challenge in the analysis of somatic genetic data is distinguishing between patterns of alteration that reflect the mechanism by which those alterations were generated, positive selection and negative selection. An underlying assumption of our analyses is that patterns of alteration that are observed across all chromosomes are likely to reflect mechanistic biases, whereas deviations from these patterns at individual loci are likely to reflect selective pressures.

The differences between telomere-bounded and internal SCNAs across all chromosomes suggest that different mechanisms underlie their generation. Internal SCNAs have been proposed to occur as a result of the apposition of the two breakpoints in three-dimensional space. Chromatin is arranged as a 'fractal globule' during interphase[58,59], in which the likelihood that two breakpoints would be apposed decreases proportional to the linear distance between them, implying a 1/length distribution. Conversely, SCNAs that start at the telomere may be related to telomere shortening and telomere crisis and may be associated with a single double-strand break that could occur anywhere within the chromosome[60].

Of the 140 peak regions in the pan-cancer analysis, only 35 contained known amplified oncogenes or tumor suppressor genes. SCNAs in some of the remaining regions may recur because these regions are subject to relatively small amounts of negative selection[21] or because of mechanistic biases favoring the generation of SCNAs in these regions[61], as has been suggested for deletions involving large genes[1,5,62]. Indeed, we found that SCNAs involving large genes often did not anticorrelate with any other genetic events, suggesting that the genes in these regions may have limited functional roles in oncogenesis. However, it remains likely that many additional oncogenes and tumor suppressor genes are within these regions. Moreover, these 140 regions and the additional 229 peak regions identified in the lineage-specific analyses are likely to constitute a subset of the regions that are significantly altered in cancer. Analyses of other cancer types have identified additional peak regions[1,4], and the limited resolution of the array platform may have obscured detection of some SCNAs.

Varying levels of genomic disruption across cancers are likely to engender biases in analyses of correlations, not only between SCNAs, but also between SCNAs and other features of these cancers. For example, increased genomic disruption has been associated with poor prognosis in multiple cancer types[63,64]. Poor prognosis is therefore likely to be associated with increased rates of SCNA across much of the genome. Controlling for this tendency will be required to identify SCNAs that are functionally associated with progression. It will also be important to account for other possible confounders such as mechanistically linked events (for example, chromothripsis or SCNAs that encompass multiple peak regions).

Whole-genome sequencing data can indicate the specific rearrangements that contributed to each SCNA[11,24], and assessment of genetic heterogeneity within tumors can also distinguish early from late events[23,29]. Both of these approaches are likely to inform the mechanisms by which SCNAs are generated and the selective pressures that shape them.

Results from this study are available at http://www.broadinstitute.org/tcga/, including segmented copy number data (viewable using the Integrative Genomics Viewer[65]) and the frequency and significance of copy number changes across and within cancer types.

URLs. GRAIL, http://www.broadinstitute.org/mpg/grail/; UCSC Golden Path database (27 February 2012), ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/; R software, http://cran.r-project.org/; Vegan: Community Ecology Package, http://vegan.r-forge.r-project.org/; Firehose, http://www.broadinstitute.org/cancer/cga/firehose.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully acknowledge contributions from the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group (contributing consortium members are listed in the **Supplementary Note**). We thank A. Shehata, P. Bandopadhayay, W. Gibson, R. Ferrer and P. Horowitz for their help and comments during construction of the manuscript. This work was conducted as part of The Cancer Genome Atlas Research Network, with funding from grants U24CA143867 (Genome Characterization Center: M.M., R.B. and G.G.) and U24CA143845 (Genome Data Analysis Center: G.G. and M.M.) of the US National Institutes of Health. Additional support was from US NIH/National Cancer Institute grants U54CA143798 (R.B.), U54HG003067 (S.B.G.) and U24CA143882 (P.W.L.) and from the V Foundation and Pediatric Low-Grade Astrocytoma Foundation (R.B.).

AUTHOR CONTRIBUTIONS

T.I.Z., S.E.S., S.L.C., M.M., G.G. and R.B. conceived of the study. Analytic methods were developed by T.I.Z., S.E.S., S.L.C., B.T., J.W., C.H.M., M.S.L., G.G. and R.B. DNA methylation analyses were provided by H.S. and P.W.L. Segmented copy number data were provided by G.S. Analyses were performed by T.I.Z., S.E.S., S.L.C., A.D.C., C.-Z.Z., C.S., S.B.G. and B.H. The manuscript was written by T.I.Z., S.E.S., M.M., G.G. and R.B.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
2. Baudis, M. Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer* **7**, 226 (2007).
3. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
4. Kim, T.M. *et al.* Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.* **23**, 217–227 (2013).
5. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
6. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
7. Weir, B.A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–U22 (2007).
8. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
9. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
10. Xue, W. *et al.* A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl. Acad. Sci. USA* **109**, 8212–8217 (2012).
11. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
12. Harada, T. *et al.* Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene* **27**, 1951–1960 (2008).
13. Tsao, M.S. *et al.* Erlotinib in lung cancer—molecular and clinical predictors of outcome. *N. Engl. J. Med.* **353**, 133–144 (2005).
14. Cheang, M.C.U. *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst.* **101**, 736–750 (2009).
15. Kim, E.S. *et al.* Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. *Lancet* **372**, 1809–1818 (2008).
16. Lowe, S.W. *et al.* p53 status and the efficacy of cancer therapy *in vivo*. *Science* **266**, 807–810 (1994).
17. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA* **104**, 20007–20012 (2007).
18. Taylor, B.S. *et al.* Functional copy-number alterations in cancer. *PLoS ONE* **3**, e3179 (2008).
19. Krasnitz, A., Sun, G., Andrews, P. & Wigler, M. Target inference from collections of genomic intervals. *Proc. Natl. Acad. Sci. USA* **110**, E2271–E2278 (2013).
20. Mullighan, C.G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
21. Solimini, N.L. *et al.* Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* **337**, 104–109 (2012).
22. Nijhawan, D. *et al.* Cancer vulnerabilities unveiled by genomic loss. *Cell* **150**, 842–854 (2012).
23. Landau, D.A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
24. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
25. Notta, F. *et al.* Evolution of human *BCR-ABL1* lymphoblastic leukaemia–initiating cells. *Nature* **469**, 362–367 (2011).
26. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
27. Vaske, C.J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
28. Vandin, F., Upfal, E. & Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
29. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
30. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
31. Maida, Y. *et al.* An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461**, 230–235 (2009).
32. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **489**, 214–218 (2013).
33. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* **20**, 908–913 (2013).
34. Cheetham, S.W., Gruhl, F., Mattick, J.S. & Dinger, M.E. Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* **108**, 2419–2425 (2013).
35. Ying, H. *et al.* Mig-6 controls EGFR trafficking and suppresses gliomagenesis. *Proc. Natl. Acad. Sci. USA* **107**, 6912–6917 (2010).
36. Du, J. *et al.* FOXC1, a target of polycomb, inhibits metastasis of breast cancer cells. *Breast Cancer Res. Treat.* **131**, 65–73 (2012).
37. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
38. Arrowsmith, C.H., Bountra, C., Fish, P.V., Lee, K. & Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discov.* **11**, 384–400 (2012).
39. French, C.A. *et al.* Midline carcinoma of children and young adults with *NUT* rearrangement. *J. Clin. Oncol.* **22**, 4135–4139 (2004).
40. Borrow, J. *et al.* The translocation t(8;l6)(p11,p13) of acute myeloid leukaemia fuses a putative acetyltransferase to the CREB binding protein. *Nat. Genet.* **14**, 33–41 (1996).
41. Champagne, N. *et al.* Identification of a human histone acetyltransferase related to monocytic leukemia zinc finger protein. *J. Biol. Chem.* **274**, 28528–28536 (1999).
42. Jaju, R.J. *et al.* A novel gene, *NSD1*, is fused to *NUP98* in the t(5;11)(q35;p15.5) in *de novo* childhood acute myeloid leukemia. *Blood* **98**, 1264–1267 (2001).
43. Micci, F., Panagopoulos, I., Bjerkehagen, B. & Heim, S. Consistent rearrangement of chromosomal band 6p21 with generation of fusion genes *JAZF1/PHF1* and *EPC1/PHF1* in endometrial stromal sarcoma. *Cancer Res.* **66**, 107–112 (2006).
44. Park, J.T. *et al.* *Notch3* gene amplification in ovarian cancer. *Cancer Res.* **66**, 6312–6318 (2006).
45. Garkavtsev, I., Kazarov, A., Gudkov, A. & Riabowol, K. Suppression of the novel growth inhibitor p33^{ING1} promotes neoplastic transformation. *Nat. Genet.* **14**, 415–420 (1996).
46. Beshiri, M.L. *et al.* Coordinated repression of cell cycle genes by KDM5A and E2F4 during differentiation. *Proc. Natl. Acad. Sci. USA* **109**, 18499–18504 (2012).
47. Gargalionis, A.N., Piperi, C., Adamopoulos, C. & Papavassiliou, A.G. Histone modifications as a pathogenic mechanism of colorectal tumorigenesis. *Int. J. Biochem. Cell Biol.* **44**, 1276–1289 (2012).
48. Berman, B.P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. *Nat. Genet.* **44**, 40–46 (2012).
49. Füllgrabe, J., Kavanagh, E. & Joseph, B. Histone onco-modifications. *Oncogene* **30**, 3391–3403 (2011).
50. Imielinski, N. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
51. Andreassen, P.R., Lohez, O.D., Lacroix, F.B. & Margolis, R.L. Tetraploid state induces p53-dependent arrest of nontransformed mammalian cells in G1. *Mol. Biol. Cell* **12**, 1315–1328 (2001).
52. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with *TP53* mutations. *Cell* **148**, 59–71 (2012).
53. Ho, C.C., Hau, P.M., Marxer, M. & Poon, R.Y.C. The requirement of p53 for maintaining chromosomal stability during tetraploidization. *Oncotarget* **1**, 583–595 (2010).
54. Dalton, W.B., Yu, B. & Yang, V.W. p53 suppresses structural chromosome instability after mitotic arrest in human cells. *Oncogene* **29**, 1929–1940 (2010).
55. Tang, Z. *et al.* PP2A is required for centromeric localization of Sgol and proper chromosome segregation. *Dev. Cell* **10**, 575–585 (2006).
56. Khanna, K.K. & Jackson, S.P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.* **27**, 247–254 (2001).
57. Filippova, G.N. *et al.* Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity. *Cancer Res.* **62**, 48–52 (2002).
58. Fudenberg, G., Getz, G., Meyerson, M. & Mirny, L.A. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29**, 1109–1113 (2011).
59. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
60. Artandi, S.E. *et al.* Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. *Nature* **406**, 641–645 (2000).
61. Yunis, J.J. & Soreng, A.L. Constitutive fragile sites and cancer. *Science* **226**, 1199–1204 (1984).
62. Smith, D.I., Zhu, Y., McAvoy, S. & Kuhn, R. Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett.* **232**, 48–57 (2006).
63. Pinto, A.E. *et al.* DNA ploidy is an independent predictor of survival in breast invasive ductal carcinoma: a long-term multivariate analysis of 393 patients. *Ann. Surg. Oncol.* **20**, 1530–1537 (2013).
64. Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
65. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

ONLINE METHODS

Generation of copy number profiles. The pipeline used to generate relative copy number estimates will be described elsewhere (B.T., G.S., S. Monti, J. Gentry, B.H. *et al.*, unpublished data). In brief, probe-level signal intensities from Affymetrix SNP6 .CEL files were normalized to a uniform brightness across arrays and merged to form intensity values for each probe set using SNPFileCreator, a Java implementation of dChip[66,67]. These intensities were mapped to copy number levels using Birdseed[68] in the case of SNP markers and on the basis of experiments with cell lines with varying dosage of X in the case of copy number markers[1]. Recurrent germline copy number variations (CNVs) were identified across all DNA samples from normal tissue, and markers within these regions (representing ~15% of all markers) were removed from further analysis. Noise was further reduced by application of Tangent normalization followed by circular binary segmentation[69,70]. Quality control metrics were applied at various stages in the pipeline, resulting in the removal of data representing 23 cancers out of 4,957 primary cancers that had been profiled by SNP6 arrays.

HAPSEG (S.L.C., M.M. and G.G., unpublished data) and ABSOLUTE[29], running on Firehose[71], were applied to data from 4,870 of these cancers, including both the SNP6 data and, when available, whole-exome sequencing data from the same cancers (1,069 samples). Of these, purity and ploidy estimates and genome-wide absolute allelic copy numbers were called in 3,847 cancers (**Supplementary Table 1**). The 200 acute myeloid leukemia samples were not called by ABSOLUTE because they exhibited copy number alterations across small fractions of their genomes, resulting in insufficient data for accurate calls by the algorithm.

Determination of SCNAs. We determined the most likely series of SCNAs that led to the copy number profiles generated by ABSOLUTE for each homologous chromosome (henceforth, 'allele'). Each SCNA was characterized by its length, amplitude, genomic position and, when determinable, allele and the timing of its generation relative to neighboring segments. We deconstructed each chromosome individually into two sequential steps (to be described in greater detail; T.I.Z., J.W., S.E.S., C.-Z.Z., S.L.C. *et al.*, unpublished data):

(1) Finding a set of the most parsimonious arrangements of copy levels on the two parental alleles (allelic partitioning).
(2) Finding the most likely set of SCNA events that would give rise to these copy number profiles (allele deconstruction).

Allelic partitioning. Our data consist of integer copy numbers of each allele at each locus. The data are segmented, with infrequent changes in copy number between adjacent markers on the array (fewer than 1 breakpoint per 1,000 markers). We started with no information about which copy levels or breakpoints belonged on the same chromosome. The purpose of allele partitioning was to find a set of the most parsimonious partitions of copy levels between the two alleles.

There is some information inherent in the structure of segmentation. Because breakpoints are rare, introducing breakpoints that are not necessary to explain our observations adds complexity to our model. There are only two situations in which this does not determine partitioning between the two alleles: (i) when the two alleles are at the exact same copy level at a particular locus or (ii) when both alleles have a breakpoint at the exact same SNP marker. The first situation is common; we expect the second situation to be rare. In either case, we lose the ability to confidently say whether segments preceding that position occurred on the same or on the opposite allele as segments subsequent to this position. We call these loci 'flex points', as we are free to swap segments between the two alleles only in these regions. We labeled regions between adjacent flex points 'contigs', as the partitioning of these segments relative to one another is fixed. The total number of possible arrangements of a given chromosome is 2^f, where f is the number of flex points on the chromosome.

If there were fewer than eight flex points, we enumerated all possible permutations of the contigs across the two alleles. If there were eight or more flex points, such enumeration was computationally prohibitive, and we focused on the most likely allelic partitions. We assume that the most likely partitions will tend to assign unlikely copy levels (which vary widely from the chromosome-wide average) to the same allele, so that they can be accounted for by a single unlikely event rather than requiring separate unlikely events on each allele.

Allele deconstruction. Once the segments were fixed to each allele, SCNA determination was performed in similar fashion to methods described previously[1,71], which identify the combination of SCNAs that would result in the observed copy number profile and have maximum likelihood of having occurred. The likelihood of an SCNA occurring was estimated according to the observed frequencies of SCNAs with similar lengths and amplitudes of copy number change across the entire data set.

Here, however, we considered absolute allelic copy number levels, which are discrete numbers, whereas previous methods focused on continuous total copy ratios. The discretized data allow enumeration of more possible SCNA combinations (including multiple overlapping amplifications and deletions) than is computationally possible in continuous data. The absolute copy numbers also require that we distinguish SCNA likelihoods in near-diploid samples from SCNA likelihoods in samples that have undergone WGD, which tend to have higher rates of other types of SCNAs (**Fig. 1b**).

SCNA timing relative to WGD and chromosome duplication. We determined the temporal relations of individual SCNAs to WGD using different approaches for deletions and amplifications.

We considered deletions that involved a change from two copies to zero copies of an allele in WGD samples to have likely occurred before WGD. Similarly, deletions that involved a change from two copies to one copy of an allele were considered to have occurred after WGD. Other deletions were left uncalled because of ambiguities introduced by surrounding alterations. When determining the timing of genome doubling, we did not include arm-level or whole-chromosome events, as events of this size are too common to rule out two sequential events that appear to have the same breakpoints.

Amplifications are more ambiguous than deletions because the extra copies of DNA may end up elsewhere in the genome and be affected by subsequent events in those regions. However, because WGD affects the whole genome simultaneously, we expect estimates of WGD timing based on amplifications to be similar overall to estimates based on deletions. We called events with an even total copy change as occurring before WGD and events with odd copy change as occurring after WGD.

The same metrics were used to determine events before or after chromosome duplication (**Fig. 2b**). Again, amplifications are more uncertain than deletions because they may involve disparate regions of the genome.

Chromothripsis detection. Chromothripsis results from different mechanisms from most focal events and has a very different distribution across lineages[30,72]. We identified chromothripsis events in diploid samples based on three features that are observable in copy number profiles and that have been associated with chromothripsis previously[72]:

(1) A single chromosome exhibits an unexpectedly large number of SCNAs given the observed frequency of SCNAs within the sample.
(2) SCNAs on this chromosome tend to be more abnormally closely spaced than we would expect by chance.
(3) SCNAs are non-overlapping (because they occurred simultaneously) and lead to copy number changes of +1 or –1.

Previous estimates of rates of chromothripsis have been complicated by uncertainty as to the absolute numbers of copies of change. In our application of these criteria, we evaluated the absolute allelic copy number data to identify chromosomes that contained more non-overlapping SCNAs that involved a single-copy change than we would expect by chance, given the number of SCNAs within the sample and using the binomial distribution. From these chromosomes, we applied the additional criterion that these SCNAs should be more tightly distributed within the chromosome than we would expect given a random selection of non-overlapping SCNAs within our data set. If this criterion was not met, we applied a recursive algorithm to remove the SCNA furthest from the centroid location of the SCNAs potentially derived from chromothripsis and recomputed these two statistics.

Further details of the method will be described separately (T.I.Z., J.W., S.E.S., C.-Z.Z., S.L.C. *et al.*, unpublished data).

doi:10.1038/ng.2760

Impurity-corrected GISTIC. In cases where we were able to estimate purity and ploidy from ABSOLUTE, we 'corrected' total copy ratios for signal dampening due to cancer cell impurity (i.e., contamination with normal DNA). We called this *In Silico* Admixture Removal (ISAR).

The observed copy ratio $R(x)$ at locus x is a function of the purity α, cancer cell ploidy τ (representing the average copy number across the genome) and integer copy number (in the cancer cells) $q(x)$[29], with

$$R(x) = (\alpha \times q(x) + 2(1-\alpha))/D$$

where D represents the average ploidy across all cells in the cancer:

$$D = \alpha\tau + 2(1-\alpha)$$

From this, we can determine $q(x)$ as

$$q(x) = D \times R(x)/\alpha - 2(1-\alpha)/\alpha$$

We assume that the functionally relevant number is the copy ratio within cancer cells, representing the integer number of copies $q(x)$ divided by the overall ploidy of the cell τ, calculated as

$$R'(x) = q(x)/\tau = R(x)/\alpha - 2(1-\alpha)/(\alpha\tau)$$

Use of $R'(x)$ has the effect of amplifying the signal from low-purity samples to be equivalent to that of higher purity samples. For samples for which ABSOLUTE calls were not available, we used $R(x)$.

To determine significantly recurrent regions of SCNA, we used GISTIC 2.0 (ref. 71) applied to the transformed copy number data. We used a noise threshold of 0.3, a broad length cutoff of 0.5 chromosome arms, a confidence level of 95% and a copy-ratio cap of 1.5.

For some lineage-specific analyses, dozens of regions on a single chromosome arm were identified as significant peaks because of the presence in many samples of discontinuous SCNAs (such as chromothripsis) on those chromosome arms. This phenomenon has been observed previously[1]. We narrowed these regions by applying in all lineage-specific analyses an 'arm-level peel-off' correction that considers all SCNAs on a chromosome arm in a single sample to be part of a single event when determining whether multiple significantly recurrent events exist on that chromosome arm. This approach has also been used in previous analyses[73].

The genes listed in each peak region include all protein-coding genes and microRNAs and additional noncoding RNAs as listed in the files refGene.txt, refLink.txt, refSeqStatus.txt and wgRna.txt from the UCSC Golden Path database as of 27 February 2012.

Significance of chromatin-modifying genes among peak regions of amplification without known driver genes. To determine whether epigenetic regulators were enriched in peak regions, we compared the number of regions with epigenetic regulators (using a published list[38]) to permuted data sets in which each gene in each region was replaced by a gene randomly selected from elsewhere in the genome.

Correlation analysis. To determine the significance of SCNA co-occurrences, we compared the observed rate of co-occurrences to the rate of co-occurrences in 5,000 permuted copy number profiles for which we had randomized the sample assignment for each chromosome, while maintaining genomic position and lineage and sublineage assignments. We only considered SCNAs in different chromosomes to avoid confounding due to geographic proximity. This analysis generated the permuted distribution in **Figure 4c** (blue line) and **Supplementary Figure 4a**,**b**, and the FDR-corrected[74] P values in **Figure 4b** (top).

To control for variable rates of genomic disruption across samples, we modified the permutations so that they maintained both the numbers of amplified and deleted markers A_j^0 and D_j^0 in each sample j. After randomizing sample assignments for each chromosome as described above, we applied simulated annealing[75,76] in which we picked a chromosome at random and swapped it between two randomly chosen samples within the same lineage at each step and accepted the step with a probability of $1 - E_{\text{total}}$, where

$$E_{\text{total}} = T_{\text{amp}} \times \sum_j \frac{(A_j^{t+1} - A_j^0)}{A_j^0 + 1} + T_{\text{del}} \times \sum_j \frac{(D_j^{t+1} - D_j^0)}{D_j^0 + 1}$$

and A_j^t and D_j^t represent the numbers of amplified and deleted markers in sample j and step t. T_{amp} and T_{del} are temperature factors that were slowly increased during the annealing, and the 1 in the denominator of each value is to avoid dividing by 0 in samples without any events. This approach generated the distributions shown in **Figure 4c** (dashed line) and the FDR-corrected[74] P values in **Figure 4b** (bottom). This procedure was applied in two separates analyses: one in which we looked at all SCNAs that passed the noise thresholds we used for our GISTIC significance analyses (above) and one in which we only considered loci with copy number of <1 or >4.4. The second analysis we termed our 'high-level' analysis.

Intersection between mutual exclusivity network and DAPPLE network. To validate the functionality of our network, we looked at the overlap between our network and DAPPLE, a curated data set of PPIs[77]. Of the >400,000 PPI pairs, we took only pairs with a score equal to 1 (indicating highest confidence). Two peak regions had an edge between them in the PPI network under two conditions;

(1) A protein within the first peak was a direct interaction with a protein in the second peak.
(2) A protein in the first peak had at least three distinct paths of length 2 in the PPI network to a protein in the second peak.

To improve specificity, we only tested regions containing fewer than 25 genes. We determined whether the similarity between the PPI network and the anticorrelation network was significant by comparing the extent of overlap with permutations in which the edges in the anticorrelation network were randomly reassigned while maintaining the overall connectivity of the graph. By comparing both observed and anticorrelation networks to the same PPI network, we controlled for the propensity of regions with many genes to map to more PPIs.

Somatic genetic correlates with WGD. To determine which of the 200 most significant somatic mutations correlate with WGD, we used the permmatswap function in the R[78] package vegan with the quasifit handle (M.S.L., P. Stojanov, C.H.M., G. Kryukov, S.E.S. *et al.*, unpublished data) to produce a series of independent assignments for mutations on each gene within each sample. This function maintained the number of mutations per gene per lineage, as well as the number of mutations per sample.

To determine which of the peak regions had SCNAs that correlate with WGD, we compared the number of times each SCNA was observed in WGD samples in our observed data to the number of times the SCNA was observed in WGD samples in the permutations created by our simulated annealing approach.

Overlap of peak regions of SCNA. Two regions were considered to overlap if their 95% confidence intervals intersected. To determine significance of overlap, we compared the number of peak regions that overlapped across at least 2 lineages in the observed data to 100,000 permutations in which the locations of each peak region were randomly shuffled within its chromosome arm (disallowing extension past the telomere or centromere).

GRAIL analysis. We used GRAIL[37] to find common functional terms in the literature for the genes in peak regions of SCNA. We used only PubMed abstracts through December 2006. We removed the following non-informative keywords from those GRAIL terms found to be most significant: 'growth', 'cancer', 'cancers', 'tumor', 'tumors', 'proliferation', 'suppressor', 'factors', 'loss', 'like', 'rich', 'cell', 'cells', 'yeast', 'system', 'family', 'deletions', 'elegans' and 'national'.

66. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2**, RESEARCH0032 (2001).
67. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).

68. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
69. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
70. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
71. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
72. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
73. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
74. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
75. Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
76. Cerny, V. Thermodynamical approach to the traveling salesman problem—an efficient simulation algorithm. *J. Optim. Theory Appl.* **45**, 41–51 (1985).
77. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
78. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2012).

doi:10.1038/ng.2760