Supplementary Materials for

Spatial and temporal diversity in genomic instability processes defines lung cancer evolution

Elza C. de Bruin, Nicholas McGranahan, Richard Mitter, Max Salm, David C. Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J. Rowan, Eva Grönroos, Madiha A. Muhammad, Stuart Horswell, Marco Gerlinger, Ignacio Varela, David Jones, John Marshall, Thierry Voet, Peter Van Loo, Doris M. Rassl, Robert C. Rintoul, Sam M. Janes, Siow-Ming Lee, Martin Forster, Tanya Ahmad, David Lawrence, Mary Falzon, Arrigo Capitanio, Timothy T. Harkins, Clarence C. Lee, Warren Tom, Enock Teefe, Shann-Ching Chen, Sharmin Begum, Adam Rabinowitz, Benjamin Phillimore, Bradley Spencer-Dene, Gordon Stamp, Zoltan Szallasi, Nik Matthews, Aengus Stewart, Peter Campbell, Charles Swanton*

*Corresponding author. E-mail: charles.swanton@cancer.org.uk

Published 10 October 2014, *Science* **346**, 251 (2014)
DOI: 10.1126/science.1253462

This PDF file includes

Materials and Methods

Patient Cohort Description

Samples for sequencing were obtained from patients diagnosed with NSCLC who underwent definitive surgical resection prior to receiving any form of adjuvant therapy, such as chemotherapy or radiotherapy. Informed consent allowing for genome sequencing had been obtained. Six samples were collected from University College London Hospital, London (UCLHRTB 10/H1306/42) and one sample was collected from Papworth Hospital, Cambridge (PHRTB 08/H0304/56). All tumors were subjected to pathology review to establish the histological subtype: five tumors were classified with CK7+/TTF1+ adenocarcinoma (L001, L003, L008 and L011) or poorly-differentiated CK7+ carcinoma (L004) histology (LUAD), one tumor (LS01) with squamous cell carcinoma histology (LUSC) and one tumor (L002) with adenosquamous histology (LUAD/LUSC; Fig S2). Tumor stages ranged from IB to IIIB. Two patients presented with disease in two separate lobes of the lung (L003 and L008), patient L001 presented with a germ-line MEN1 mutation. The median patient age was 75.5 years (range 59-84). One patient reported no history of tobacco smoking, three patients reported smoking (current smokers) and three reported previous smoking histories (former smokers). Detailed clinical characteristics are provided in table S1.

Tumor processing

Up to five regions from a single tumor mass, separated by 1cm intervals, and adjacent normal tissue were selected by a pathologist, documented by photography, and snap-frozen. Peripheral blood was collected at the time of surgery from all patients and snap-frozen, except from L001. Approximately 5x5x5mm tumor tissue and 500µl of blood was used for genomic DNA extraction, using the DNeasy kit (Qiagen) according to manufacturer's protocol. DNA was quantified by Qubit (Invitrogen) and DNA integrity was examined by agarose gel eletrophoresis.

DNA ploidy analysis

A small piece of each tumor region (approximately 1x5x5mm) was minced and incubated with 0.5% pepsin in PBS for 30 minutes at 37°C to obtain a single nuclei suspension. Nuclei were washed twice with PBS, fixed in 70% ethanol for 1 hour at room temperature, and again washed twice with PBS. Nuclei were resuspended in a propidium iodide (50µg/ml) solution in PBS and incubated on ice for 30 minutes, washed with PBS and analysed by FACs (488nm) to determine the DNA content using matched normal lung cells as control.

Multi region Whole-Exome Sequencing

For each tumor region and matched germ-line, exome capture was performed on 1-2 µg DNA using the Agilent Human All Exome V4 kit according to the manufacturer's protocol (Agilent). Samples were paired-end multiplex sequenced on the Illumina GAII or HiSeq 2500 at the Advanced Sequencing Facility at the LRI, as described previously (*27*, *28*). Each captured library was loaded on the Illumina platform and paired-end sequenced to the desired average sequencing depth (approximately 100x, detailed coverage information is provided in table S2).

Multi region Whole-Genome Sequencing Illumina platform

For four tumor regions (L002_R1 and R3, and L008_R1 and R3) and matched blood, whole-genome paired-end sequencing was performed by Illumina Cambridge LTD, using 1µg DNA, to the desired average sequencing depth (approximately 100x for the tumor regions and 40x for germ-line, detailed coverage information is provided in table S2).

SNV calling from multi-region WES and WGS data Illumina platform

Raw paired end reads in FastQ format generated by the Illumina pipeline (WES or WGS) were aligned to the full hg19 genomic assembly (including unknown contigs), using bwa 0–5.9 (*38*) with a seed length of 72 bp for data sequenced on the GAII and 100 bp for data sequenced on the HiSeq. Up to 3 or 4 mismatches were allowed per read for the GAII or HiSeq respectively, all other settings were left as default. Picard tools v1.8 was used to merge samples from the same patient region and to remove duplicate reads (http://picard.sourceforge.net) prior to determining sequence coverage (table S2).

Variant calling was performed between tumor and matched germ-line using the "somatic" tool from VarScan2 v2.3.3 (*39*). The input for VarScan2 was the SAMtools (*40*) mpileup output from combined tumor and normal samples generated by skipping bases with a phred score of <20 or reads with a mapping-quality <20. SAMtools BAQ computation was disabled and the coefficient for downgrading mapping quality was set to 50. VarScan2 somatic was run with default settings except for the following: minimum coverage was set to 10 for germline and 6 for tumor regions, the minimum variant sequency for calling a heterozygote was set to 0.01 and tumor purity was set to 0.5. The resulting calls were filtered for false positives using Varscan2's associated fpfilter.pl script, run with settings as described by Ding *et al.* (*41*). Additionally, variants were only accepted if present in ≥ 5% of reads in at least one tumor region and present with ≤2 reads in germ-line and ≥ 2 reads in a tumor region.

Small insertions and deletions (indels) were identified using Pindel version 0.2.4 pindel (*42*) in paired tumor-normal mode as previously described (*27*). All variants were annotated using both ANNOVAR (*43*) and dbNSFP (*44*). All variants identified as nonsilent were manually reviewed using Integrated Genomics Viewers (IGV) 38, and those showing an Illumina specific error profile (*45*) were removed from further analysis. Variants not subjected to ultra-deep orthogonal validation were further filtered using an in-house filter, VarSLR, which models strand-bias, mapping-quality, base-quality and position-in-read in a stepwise logistic regression framework (Salm *et al*., manuscript in preparation), and removed from further analysis. In addition, any substitution identified in dbSNP Build 132 was removed. For the WGS data, variants detected in repetitive regions (RepeatMasker, USCS genomicSuperDups tracks) or blacklisted by the Encode Mappability were removed from the analysis.

Multi region Whole-Genome Sequencing SOLiD™ platform

For LS01, all three tumor regions and matched blood, mate paired libraries (2 x50 bp, insert size 1-3kbp) were prepared and sequenced by Life Technologies (Beverly, MA, USA) on the SOLiD™ sequencer as described for the SOLiD™ Mate-Paired Library Construction Kit. Whole-genome sequencing was performed using the Exact Call

Chemistry (ECC) module as previously reported (*46*). Color space reads were mapped to the hg19 reference genome using SOLiD™ bioscope version 1.3 software and converted to base space. Hard clipping of reads removed an average of 4 bp / 50 bp read mapped. Multiple related bam files were merged and re-headed using samtools (*40*) and duplicate reads were removed with Picard Tools version 1.60 (http://picard.sourceforge.net). Sequence coverage (using Picard Tools version 1.80 (BamIndexStats software)) was estimated after duplicate removal (table S2).

SNV calling from multi-region WGS data SOLiD platform

Variant calling was performed using an in-house substitution-calling algorithm, CaVEMan (Cancer Variants through Expectation Maximisation) (*18, 47, 48*). In brief, this algorithm generates a probability score for each possible genotype at each genomic locus, by comparing tumor and matched normal sequence data to each other and to the reference genome. Copy number and aberrant cell fraction informs the probability score calculation, and was determined for these samples with ASCAT using the whole-genome NGS data combined for the three tumor regions (*21*). A number of filtering criteria were applied including the following: a SNP probability ≤ 0.05, a mutation probability ≥ 0.95, at least 1/3 of variants have a base quality ≥ 25, ≤ 1 variant with base quality ≥ 20 in germ-line, position outside centromeric repeat and more than 5 bases from a simple repeat. Variants reported with the same nucleotide substitution as reported in dbSNP Build 132 and variants present $\geq 2\%$ in matched normal samples (blood or normal lung) were excluded from the dataset.

Ion AmpliSeq™ Custom Validation panel and Comprehensive Cancer Gene Panel sequencing

A total of 1999 mutations (enriched for nonsilent and/or heterogeneous mutations) were subjected to orthogonal validation. For each tumor, an Ion AmpliSeq™ custom panel (Life Technologies) was designed using the online designer (www.ampliseq.com). Multiplex PCRs were performed on DNA from each region of the relevant tumor according to the manufacturer's protocol. Barcoded sequencing libraries were constructed, which were sequenced with 200 bp read length on the Ion Torrent PGM™ sequencer (Life Technologies). For each tumor, a comprehensive cancer gene panel targeting 409 cancer-related genes (Life Technologies) was also used for sequencing on the Ion Torrent PGM™ sequencer by Life Technologies (Beverly, MA & South San Francisco, CA USA). Sequence alignment to target regions from the hg19 genome was performed using the IonTorrent TorrentSuite™ software. Variant allele frequencies (VAFs) for each variant position having a phred score >20 were determined. A variant was considered absent when VAF < 1% while having a read coverage $\geq$ 50x or considered a germ-line variant when VAF > 1% in the germ-line. In total 108 mutations were absent in all tumor regions or identified as germ-line variants (validation rate 94.6%) and were removed from further analysis. Variants with read coverage <50x were considered inconclusive and regional distribution was extracted from exome or genome sequencing data. A total of 1,884 nonsilent and 76,129 silent mutations from all methodologies were included in the analyses.

Mutation clustering

Subclonal clusters of mutations were identified using a previously described Dirichlet process, implemented using a Markov Chain Monte Carlo (MCMC) method (*18, 22*). From the MCMC assignment of mutations to clusters, the most likely configuration of clusters and node assignments was obtained using a stepwise, greedy expectation-maximization (EM) algorithm that alternately added a node and shuffled mutations between nodes until no further improvement in the agreement with the posterior distribution from the MCMC sampling could be made. The best set of clusters was then chosen using the Bayesian information criterion (*49*). Clusters containing less than 1% of the mutations identified in a tumor were excluded from further analysis.

Phylogenetic tree analysis

The subclonal architecture of each tumor was used to construct phylogenetic trees, with ubiquitous clonal mutations representing the most recent common ancestor (trunk) and heterogeneous mutations reflecting later (branch) events. The phylogenetic relationships between tumor regions (and where relevant, subclones) was inferred using both the Maximum Parsimony and the Unweighted Pair Group Methods (UPGMA) as implemented in the MEGA5 package (*50*). In both cases, for each tumor, all identified nonsilent and silent mutations were used as input. In cases where a given tumor region revealed considerable subclonal diversity, this region was divided into a major and a minor subclone based on the variant allele frequencies, prior to phylogenetic tree reconstruction, as previously described (*27*). In addition, in cases where there was evidence for ubiquitous mutations becoming heterogeneous due to regional copy number losses (Fig S4), mutations were considered ubiquitous prior to tree reconstruction.

Maximum parsimony trees were inferred using the max-mini branch-and-bound algorithm (*51*) calculating branch lengths with the average pathway method. The germline sample was designated as the outgroup, and, in the event of multiple optimal topologies, a consensus tree was constructed by collapsing branches reproduced in less than 50% of the trees. For the UPGMA trees, the evolutionary distances were calculated using the number of differences between regions, and uncertainty assessed by a bootstrap test (1000 replicates). Trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree, and the percentage of replicate trees in which regions clustered together in the bootstrap test shown next to the branches. For samples with only two branches (WGS L002 and L008) a straightforward representation is provided with the trunk and branch lengths proportional to the number of variants, as previously described (*28*).

Identification and classification of driver mutations

All identified nonsilent mutations were compared with lists of potential driver genes in NSCLC, containing all genes identified as frequently mutated by large-scale lung cancer sequencing studies (*5, 9, 19, 52*) or large-scale pan-cancer analyses (*20, 53*) using $q < 0.05$ as cut-off, or present in the COSMIC cancer gene census (downloaded February 2013). Genes with a nonsilent mutation identified in our tumors by M-seq that were present in one of these lists, were analyzed in more detail using COSMIC (release v67) to determine whether the amino acid substitution has been previously identified. In addition,

when available, functional prediction scores (SIFT, Polyphen2, LRT and MutationTaster) implemented with dbNSFP (*44*) were used and a mutation was scored as 'deleterious' when at least two out of the four predictors classified the mutation as deleterious. We then classified all nonsilent mutations into 4 categories. Category 1 'high confidence driver mutations' contained all disrupting mutations (nonsense, frameshift, splicing or 'deleterious' missense) in tumor suppressor genes or activating amino acid substitutions in oncogenes as described in the literature. Category 2 'putative driver mutations' contained all other amino acid substitutions located at the same position or up to 5 amino acids away from a substitution present in COSMIC. Category 3 'low confidence driver mutations' contained all other nonsilent mutations in genes that were present in the lists of cancer-related genes described above. Category 4 'unknown significance' contained the remaining nonsilent mutations.

Copy number analysis

Relative copy number was estimated from whole-exome sequencing data using VarScan2 (v2.2.11) (*39*) with default parameters, excluding the sex chromosomes and low mapability regions (ENCODE 'DAC blacklisted' regions) and adjusting for GC-content. To identify genomic segments of constant copy number, logR values were quantile normalized, winsorized using the Median Absolute Deviation, and jointly segmented at the patient level (gamma = 1000) (*54*). Absolute (integer) copy numbers were derived from relative copy numbers using ABSOLUTE (v1.0.6) (*7*). SNVs with ≥50x sequencing coverage were included in the analysis and AmpliSeq derived VAFs were used where possible. Minimum/maximum ploidy was set to within +/-0.5 of the prior ploidy estimate, calculated from the sample's FACS-based DNA-index. Subsequently, the top 5 ABSOLUTE models (ranked by log-likelihood) were retrieved for each exome, and a set of inter-sample models was identified that minimized the total pairwise distance derived from the segments' expected modal copy-number, whilst maximizing the model's posterior log likelihood. Final model solutions were manually reviewed as recommended (*7*). Finally, adjacent segments of equal clonality and absolute copy-number were merged, and annotated with recurrent copy-number changes referenced in the TCGA Copy Number Portal (http://www.broadinstitute.org/tcga/home). Gains and losses for each tumor region were defined relative to ploidy for that region. Gain and loss segments had to overlap with >75% of a TCGA recurrent amplified/lost region to be classified as a putative driver gain/loss.

For M-seq WGS regions ASCAT was used for allele specific copy number estimation (*21*). Subclonal copy number analysis was performed using the Battenberg algorithm and was used as the input for the mutation clustering (*22*).

Hierarchical clustering of copy number profiles was performed at cytoband resolution using the hclust function in R, based on euclidean distances. Cytoband coordinates were retrieved from the UCSC Genome Browser database (http://genome.ucsc.edu/)

Large-scale structural variations

Inter- and intra-chromosomal translocations, inversions, and large (≥10 kb) insertions/deletions were identified in WGS data using CREST (v1.0.1) (*55*) on tumor

regions jointly, after pre-processing BAMs with the GATK IndelRealigner (v2.1-13-g1706365). To reduce the false positive rate, breakpoint junctions of putative structural variants were de novo assembled using TIGRA (v0.3.7) (*56*) in germline and tumor BAMs individually, and aligned to hg19 using BLAT (v35) (*57*); breakpoints that were re-constituted from tumor BAMs only were considered valid. Furthermore, structural variants (SVs) with breakpoints mapping to low-copy repeats (i.e. UCSC genomicSuperDups track) were removed. Breakpoints were annotated with gene information, presence of nearby (<100 bp) polymorphic SV breakpoints (2013-07-23, http://dgv.tcag.ca/dgv/app/home) and high-copy repeat content (RepBase v17.03) using ANNOVAR. Breakpoint mechanism classification was attempted according to the criteria defined in Yang *et al.* (*58*) with additional strict manual review. In order to assign SVs to specific tumor regions, soft-clipped and discordant paired-end reads consistent with each SV in each tumor region were identified; an SV was called present if at least one of these read-level events was identified. All results were plotted using Circos (*59*).

For the breakpoint clustering analysis, structural variant breakpoints (nodes) within 10kb of one another were linked by edges representing both inter-breakpoint distance and chromosome. To simplify the network, nodes connected to fewer than 3 edges were omitted. Breakpoint homology profiling as performed as previously described (*23*).

TCGA exome data sets

TCGA LUAD and LUSC exome data sets were obtained from: https://confluence.broadinstitute.org/display/GDAC/DCC+MAFs. Mutations calls were further filtered by removing any mutations that did not pass the following filters: variant allele frequency, >0.01; sequencing depth at mutated base, >10; mutant reads, >6. For a number of LUSC samples a likely sequencing artefact was identified (C>A mutations at CCA sites, see below for detailed description). These samples were removed from further analysis. Only patient samples where SNP6.0 data was also available were utilized. In total 294 LUAD and 124 LUSC TCGA samples were used in the analysis.

Estimating the cancer cell fraction of TCGA mutations

For TCGA samples, the cancer cell fraction, defined as the proportion of cancer cells harbouring a given mutation, was estimated by integrating the wild-type and mutant allele counts, absolute major and minor copy numbers, and tumor purity estimates as previously described (*60*). In brief, for a given mutation we first calculated the observed mutation copy number, n_{mut}, describing the fraction of tumor cells carrying a given mutation multiplied by the number of chromosomal copies at that locus using the following formula:

$$n_{mut} = VAF \frac{1}{p}[pCN_t + CN_n(1-p)]$$

where *VAF* corresponds to the variant allele frequency at the mutated base, and p, CN_t, CN_n are respectively the tumor purity, the tumor locus specific copy number, and the normal locus specific copy number. We then calculated the expected mutation copy number, n_{chr}, using the *VAF* and assigning a mutation to one of the possible copy numbers using maximum likelihood. The cancer cell fraction of a given mutation was then calculated as n_{mut}/n_{chr} (i.e. the observed mutation copy number divided by the

expected mutation copy number if present in 100% of tumor cells). Confidence intervals were obtained by bootstrap resampling the wild-type and mutant reads at each mutated locus (n=10,000). A mutation was classified as either clonal or subclonal based on the confidence interval of the mutation copy number. Any mutation whose mutation copy number upper 95% confidence did not overlap 1 was classified as subclonal, with all other mutations classified as clonal.

Integer major and minor copy numbers were estimated using SNP6.0 data, normalized with the aroma R package (*61-63*), and processed with OncoSNP (*64*). Tumor purity estimates were obtained using ASCAT (*21*).

Temporal dissection of mutations of M-seq tumors

For each M-seq tumor, we classified each mutation as 'early' or 'late' based on whether it was located on the trunk or branch of the phylogenetic tree, with all truncal mutations classified as 'early' and any branch mutation as 'late'.

In cases where genome doubling was observed, mutation copy number estimates were used to time mutations relative to the doubling. Only regions with at least two copies of the major allele were used, whilst regions where the minor allele was equal to one were excluded as these could reflect mutations on the minor allele that occurred before doubling or mutations on one copy of the major allele occurring after doubling. Any mutations that had a ploidy ≥2 were considered to have occurred before doubling, and all ploidy of 1 mutations as after doubling. For tuncal dissection, mutations with a variant allele frequency less than 5% or less than 2 tumor reads were excluded. Temporal dissection of mutations occurring before and after genome doubling was performed independently for each tumor region. For L008, where whole-genome sequencing data was available, the consensus between the two regions was used. For L001, one region (R1) displayed considerably higher cellularity and was therefore used.

Chi-square tests were used to compare the mutation spectra of the six mutations types (C>A, C>G, C>T, T>A, T>C, T>G). To compare the relative frequency of specific mutation types a two-sided Fisher's exact test was used. APOBEC enrichment was assessed as described below.

Temporal dissection of mutations in TCGA samples

For each single-region TCGA sample, it was not possible to construct phylogenetic trees as described above. We therefore classified mutations as early or late based on their clonal status and when possible we timed mutations relative to copy number events. In brief, for timing mutations relative to copy number events, we restricted our analysis to mutations occurring in regions with at least two copies of the major allele. For any such region, mutations at ploidy ≥2 were classified as 'before event' and any mutations with a ploidy of 1 were classified as 'after event'. Combining this with our cancer cell fraction estimates (see above), all clonal mutations that were not classified as 'after event' were aggregated as 'early', whilst all subclonal or 'after event' mutations were aggregated as 'late'. Significance in mutation spectra between 'early' and 'late' mutations were then compared using a paired t-test.

Estimating smoking strand bias

Strand bias was calculated based on annotating each C>A mutation as to whether it fell on the transcribed or untranscribed strand in the UCSC hg19 gene track (available from http://genome.ucsc.edu/cgi-bin/hgTables). TCGA samples, grouped according to histology and smoking status, were considered in aggregate. Two-sided Fisher's exact tests were used to determine significance.

Detecting an APOBEC mutation pattern

To detect an APOBEC mutation pattern the methods outlined by Roberts *et al*. (*13*) were adopted. In brief, the enrichment E_{TCW} relating to the strength of mutagenesis at the TCW motif across the genome was calculated as follows:

$$E_{TCT} = \frac{mutations_{TCW} \times context_{CorG}}{mutations_{CorG} \times context_{TCW}}$$

where $mutations_{TCW}$ is the number of mutated cytosines (and guanines) falling in a TCW (or WGA) motif, $mutations_{C\ (or\ G)}$ is the total number of mutated cytosines (or guanines), $context_{TCW}$ is the total number of TCW (or WGA) motifs within a 41-nucleotides region centered on the mutated cytosines (and guanines) and $context_{C\ (or\ G)}$ is the total number of cytosines (or guanines) within the 41-nucleotides region centered on the mutated cytosines (or guanines). Only specific base subsitutions were included (TCW to TTW or TGW, WGA to WAA or WCA, C to T or G, and G to A or C). Over-representation of APOBEC signature mutations in each sample was determined using a two-sided Fisher's exact test comparing the ratio of the number of cytosine-to-thymine or cytosine-to-guanine substitutions and guanine-to-adenine or guanine-to-cytosine substitutions that occurred in and out of the APOBEC target motif (TCW or WGA) to an analogous ratio for all cytosines and guanines that reside inside and outside of the TCW or WGA motif within 41-nucleotide region centered on the mutation cytosine (and guanine). P-values were corrected using Benjamin-Hochberg multiple testing correction, and a significance threshold of $q < 0.05$ was used. For each sample, APOBEC mutation enrichment was determined for all mutations, 'early' mutations and 'late' mutations separately. When temporally dissecting APOBEC mutation patterns in TCGA data, only samples with a significant APOBEC enrichment were used. Comparisons between early and late APOBEC mutation enrichment was performed using a paired t-test.

Identifying a likely sequencing artefact signature in TCGA LUSC samples

To identify patients that exhibited an enrichment C<A mutations occurring within a CCA or TCG mutation context, we adapted the methods described above. Thus, we defined enrichment $E_{CCAorTCG}$ as follows:

$$E_{CCAorTCG} = \frac{mutations_{CCAorTCG} \times context_{CorG}}{mutations_{CorG} \times context_{CCAorTCG}}$$

where $mutations_{CCAorTCG}$ is the number of mutated cytosines-to-adenine (and guanine-to-thymidine) falling in a CCA (or TGG) or TCG (or CGA) motif, $mutations_{C\ (or\ G)}$ is the total number of mutated cytosines-to-adenine (or guanine-to-thymidine), $context_{CCAorTCG}$ is the total number of CCA (or TGG) or TCG (or CGA) motifs within a

41-nucleotides region centered on the mutated cytosines (and guanines) and $context_{C\ (or\ G)}$ is the total number of cytosines (or guanines) within the 41-nucleotides region centered on the mutated cytosines (or guanines). Only specific base substitutions involving C>A were considered.

Over-representation of CCA mutations in each sample was determined using a two-sided Fisher's exact test comparing the ratio of the number of cytosine-to-adenine and guanine-to-thymidine substitutions that occurred in and out of the target motif (CCA or TGG) to an analogous ratio for all cytosines and guanines that reside inside and outside of the CCA or TGG motif within 41-nucleotide region centered on the mutation cytosine (and guanine). P-values were corrected using Benjamin-Hochberg multiple testing correction, and a significance threshold of $q < 0.05$ was used.

A specific sequencing plate (9898) was found to show highly significant enrichment of CCA mutations ($P < 0.001$). Any samples sequenced on this plate as well as any samples showing CCA enrichment were therefore removed from further analysis. In total, 28 samples were removed.

Fluorescent In Situ Hybridization analysis

A small section of snap-frozen tumor material was fixed o/n in 10% neutral buffered formalin and paraffin-embedded. Dual color FISH was carried out using 2 centromeric probes (CEP2 and CEP16 labeled with spectrum orange and spectrum green, respectively; Abbott Laboratories), selected on the basis of infrequent copy number alterations in the TCGA LUAD and LUSC dataset for these chromosomes (Fig 2A). Briefly, 5-micron sections were used. Following dewaxing and rehydration using ethanol gradients, slides were placed in SPoTLight Pretreatment buffer (Invitrogen) at 98°C for 15 minutes and washed. Digestion enzyme was added to each slide and incubated for 22 minutes at room temperature. Slides were washed and dehydrated. 1.5µL of each centromeric probe was mixed with 10 µL hybridization buffer, denatured for 5 minutes at 95°C and placed on each dehydrated section. Slides were incubated overnight at 37°C in a moisturized chamber. After washing with 0.5x SSC at 75°C, slides were stained and mounted using Vectashield mounting media with DAPI (Vector Labs). Slides were scanned and images were captured using a 40x objective on the Applied Imaging Ariol System (Applied Imaging), with seven 0.5-mm z-stacks. For each region, sixty nuclei were scored manually in an unbiased manner.

APOBEC3B mRNA analysis by qRT-PCR

Total RNA was isolated from tumor regions of which fresh frozen material was available (L001, L002, L003, L004 and L011), using the AllPrep DNA/RNA kit from Qiagen according to the manufacturer's instruction, and used to synthesize cDNA. The cDNA was then amplified using APOBEC3B Taqman Assay or Taqman Assays for the housekeeping gene TBP (Applied Biosystems) on a 7500 FAST Real Time PCR machine (Applied Biosystems). APOBEC3B expression was normalised towards TBP, and the fold-change in expression was determined against the expression in the adjacent normal lung.

Supplementary Figures

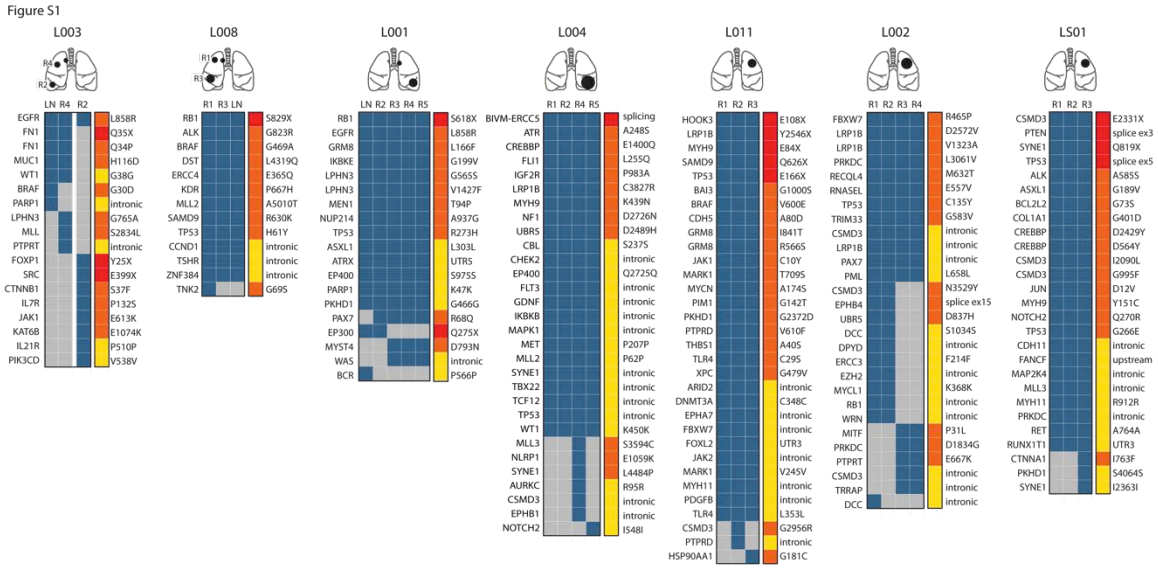

Fig. S1.

Intratumor heterogeneity of somatic mutations detected by a comprehensive cancer gene panel. Heatmaps show the regional distribution of all point mutations detected by sequencing a panel of 409 cancer-related genes. Presence (blue) or absence (grey) is indicated for each mutation per tumor region. The right column displays the severity of each mutation; nonsense or splice site (red), missense (orange) or silent (yellow). Cartoons above each heatmap depict the location of each tumor in the lung.

Figure S2

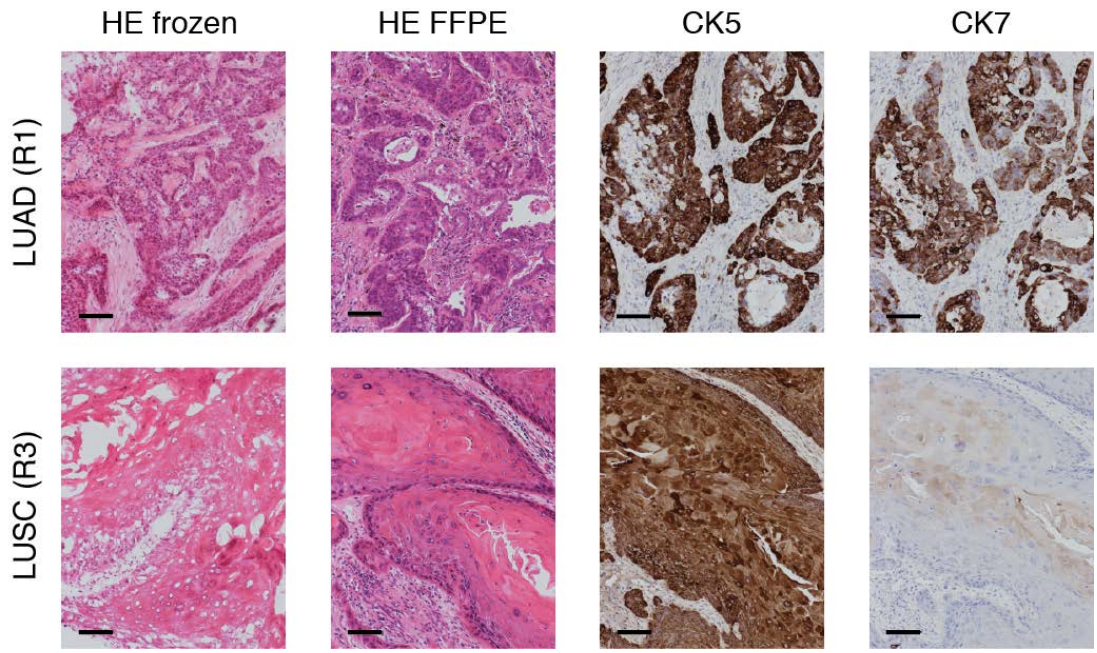

Fig. S2
L002 histopathology staining. Sections from LUAD (upper) or LUSC (lower) regions are shown. Left section is taken from frozen material adjacent to material used for DNA extraction for sequencing. The remaining sections are taken from FFPE representing LUAD or LUSC tumor regions and stained with Hematoxylin and Eosin or with an antibody detecting CK5 or CK7, as indicated. Scale bars, 100 µm.

Figure S3

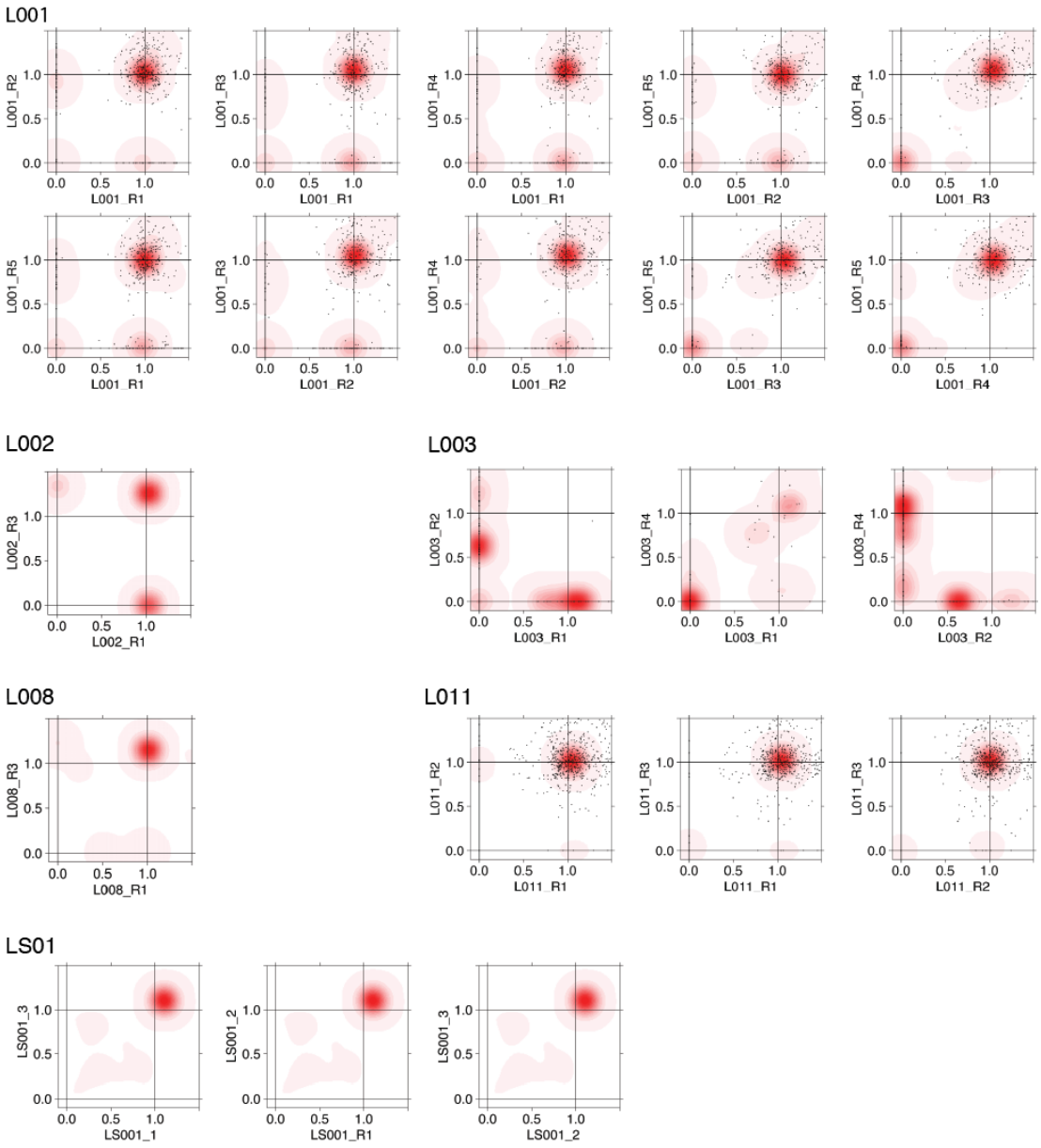

Fig. S3

Clonal architecture of NSCLC tumors. 2D-dirichlet plots show the cancer cell fraction of the mutations in regions of each NSCLC tumor; increasing intensity of red indicates the location of a high posterior probability of a cluster. For the M-seq WES tumors, identified mutations are indicated in black dots. L004 is presented in Fig.1B.

Figure S4

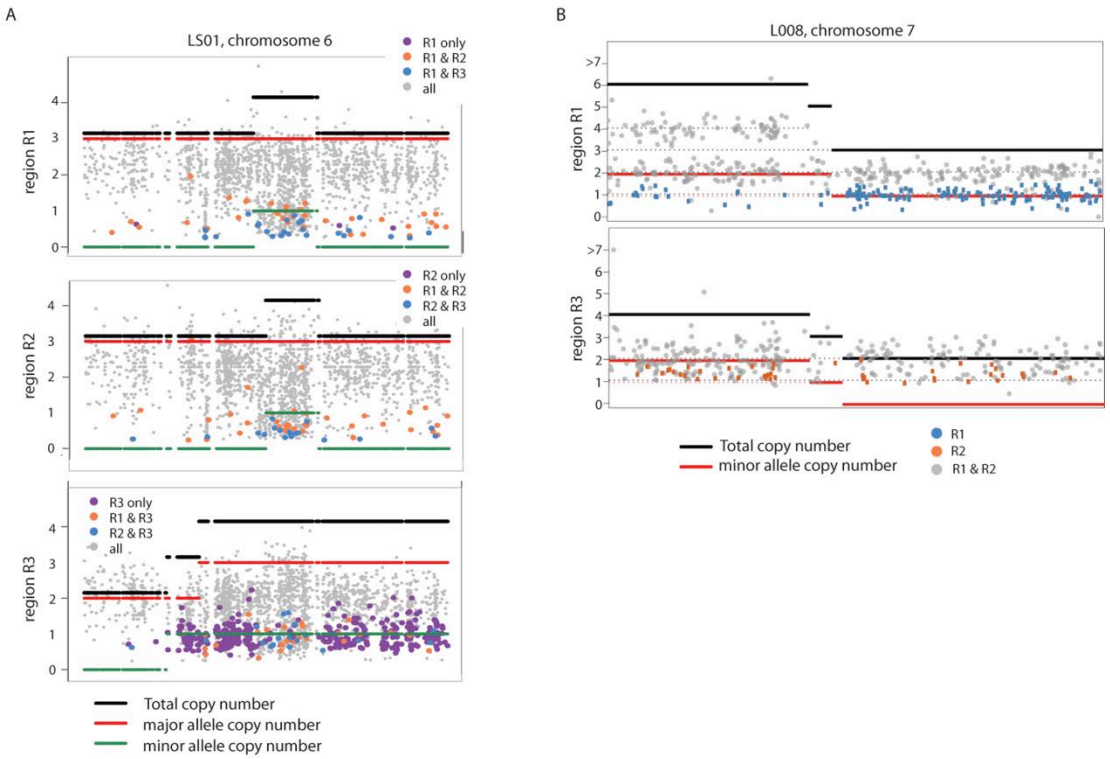

Fig. S4

Copy number events leading to mutational heterogeneity.
A) In LS01, loss of one allele of chromosome 6 in regions R1 and R2 results in private mutations in R3, creating mutational intra-tumor heterogeneity. Total copy number is shown with a black line, major allele copy number with a red line and minor allele copy number with a green line. Shared and private mutations are indicated in blue, orange, purple and grey.
B) Loss of one allele of chromosome 7q specifically in region R3 of L008 also results in loss of mutations carried on the allele, creating mutational diversity between regions R1 and R3. Total copy number is shown with a black line and the minor allele copy number with a red line. Mutations only detected in region R1 are depicted in blue, mutations only detected regionR3 are depicted in orange, with shared mutations in grey.

Figure S5

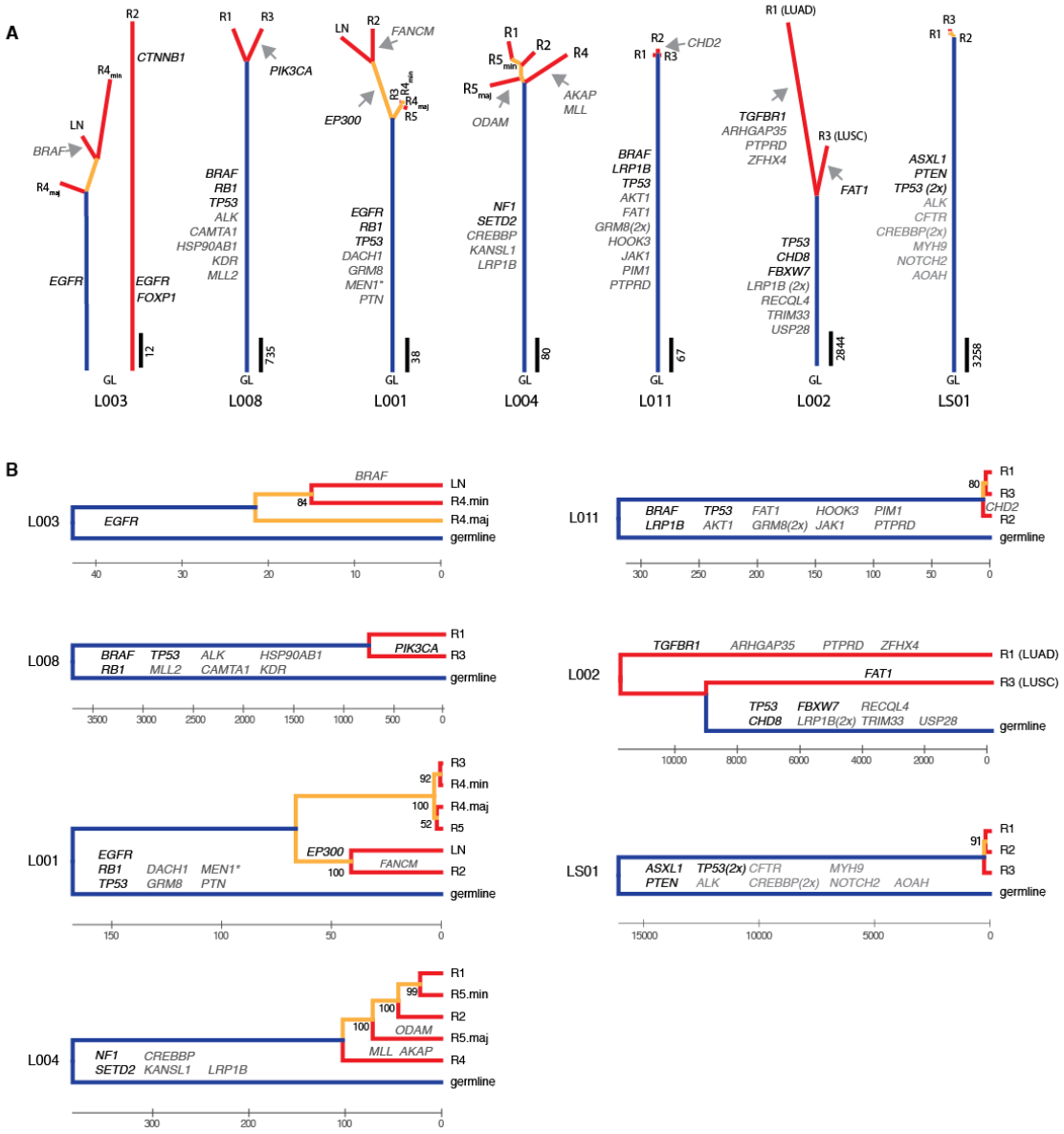

Fig. S5.

A) Phylogenetic trees generated by a maximum parsimony approach based on the distribution of all detected mutations. Scale is indicated for each sample next to the trunk with the number of mutations; trunk and branch lengths are proportional to the number of mutations acquired. GL indicates germ-line. Trees of L002, L008 and LS01 are based on M-seq WGS data. Categories 1 and 2 driver mutations are indicated next to the trunk or with an arrow pointing to the branches where they were acquired.

B) Dendrograms were inferred using the UPGMA method. The dendrograms are presented to scale, with the number of mutations as evolutionary distance. Uncertainties assessed by bootstrap tests are indicated next to the nodes. Categories 1 and 2 driver mutations are indicated above the trunk or branch where they were acquired.

Figure S6

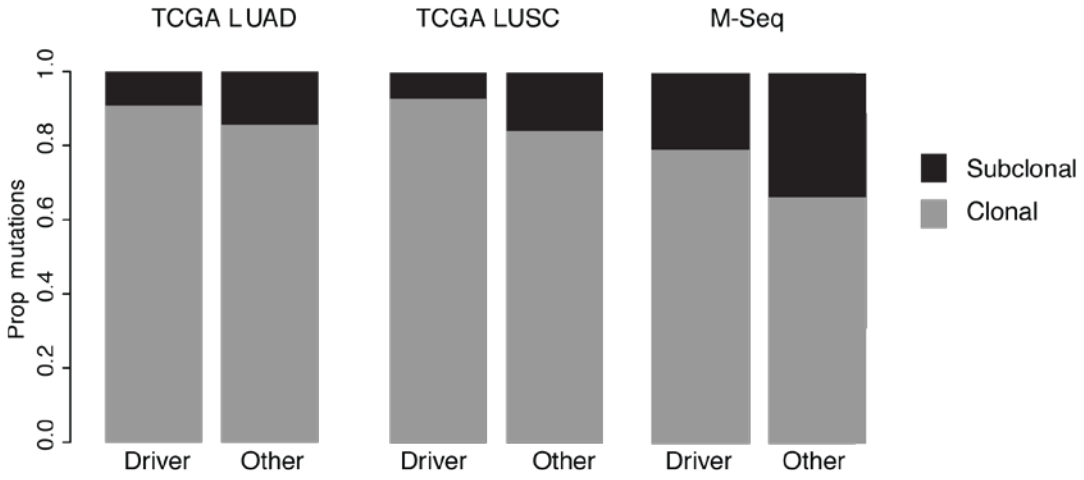

Fig. S6.

Clonality of driver and non-driver mutations in TCGA LUAD and LUSC tumors as well as M-seq samples. Driver genes are significantly more often clonally mutated compared to non-driver genes. For TCGA samples, driver gene status is based on (*1-3*). For M-seq tumors, category 1-2 drivers are displayed *(16).*

Figure S7

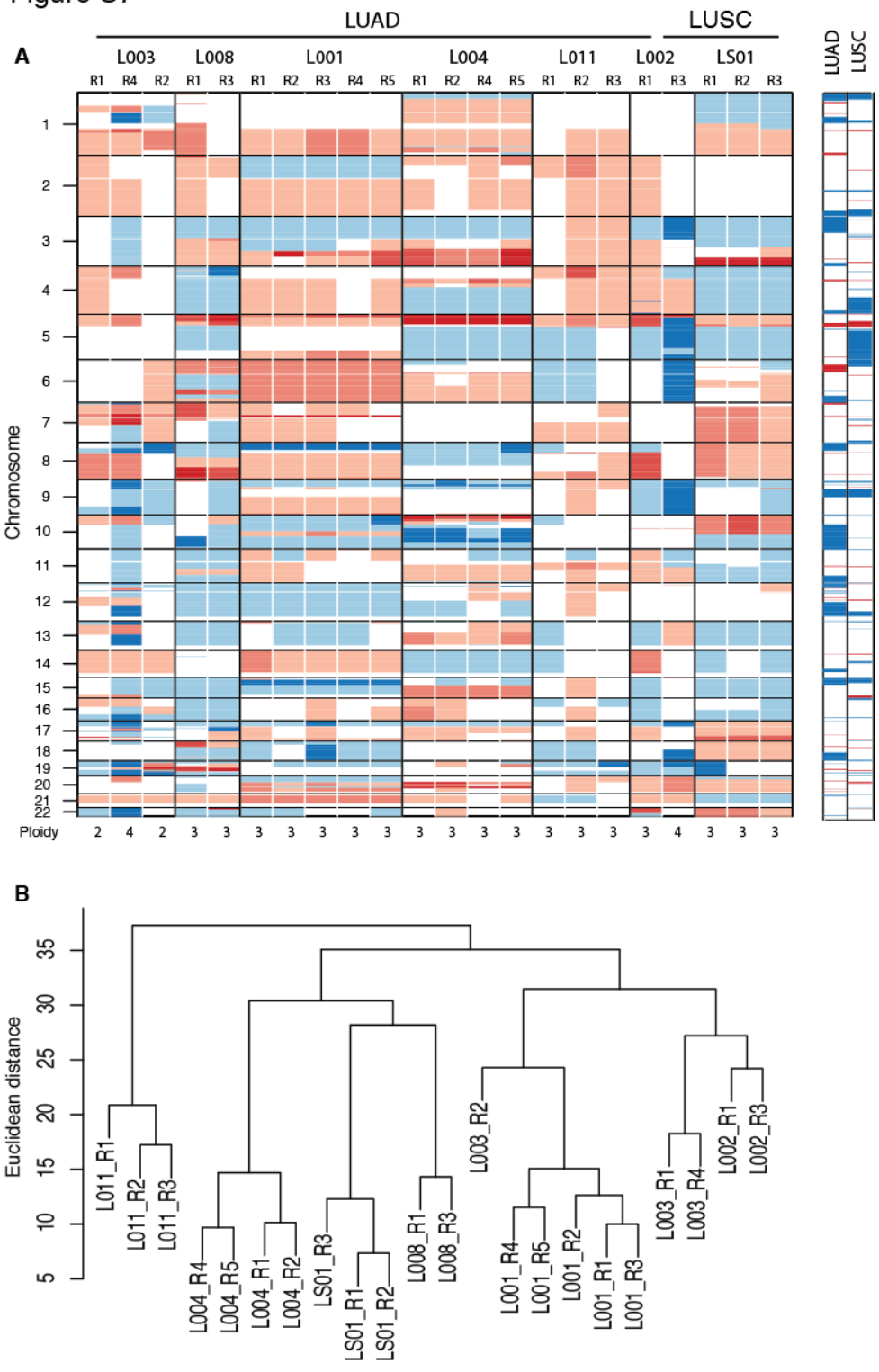

Fig. S7.

A)Copy number states across the genome for every tumor region. Losses (blue) and gains (red) are depicted relative to mean ploidy of the tumor region, with darker colors representing increased deviations from ploidy. Mean ploidy of each tumor region is

shown at the bottom of the plot. On right, recurrent gains (red) and losses (blue) are shown for TCGA LUAD and LUSC samples.
B) Hierarchical clustering of copy number profiles at cytoband resolution, using Euclidian distance. Copy number profiles of tumor regions from the same patient clearly cluster together, with the exception of L003, for which two clonally distinct tumors were observed.

Figure S8

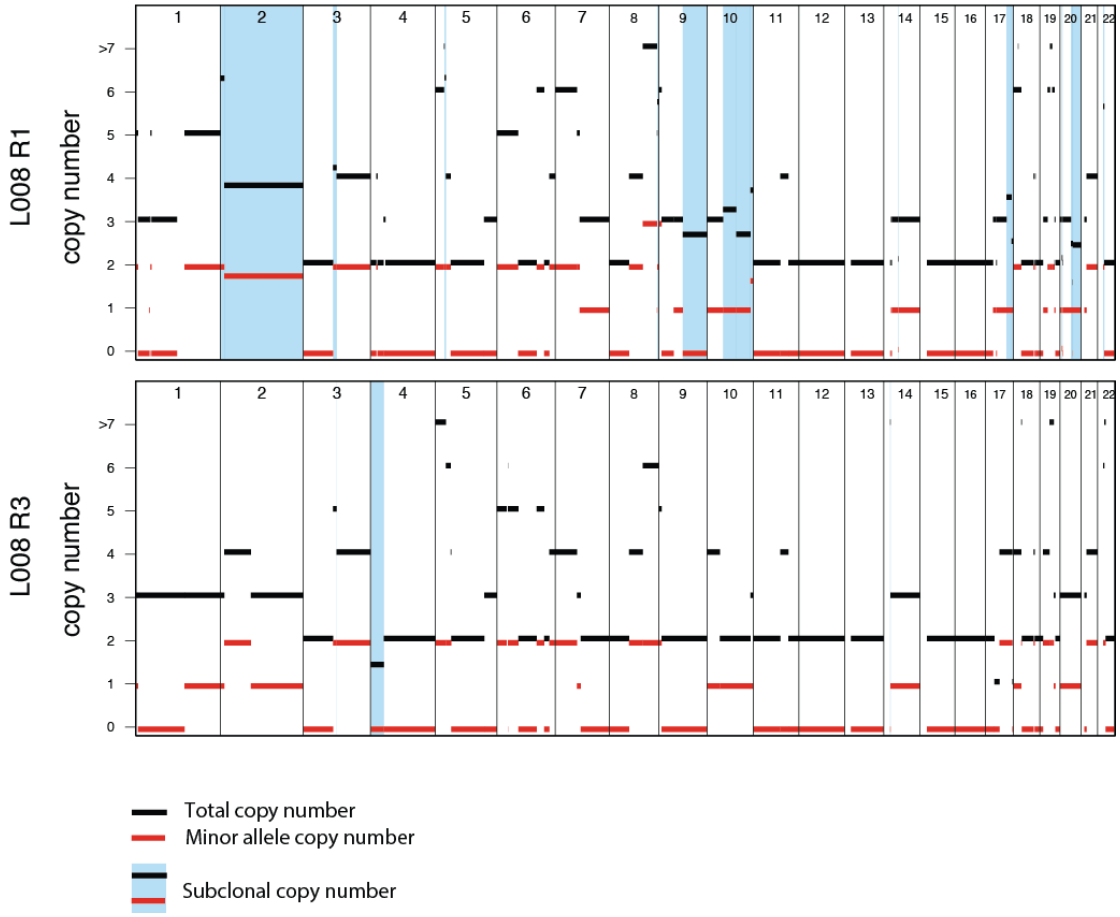

Fig. S8.

Copy number profile of L008. Chromosomal segments with subclonal copy number aberrations are highlighted in blue.

Figure S9

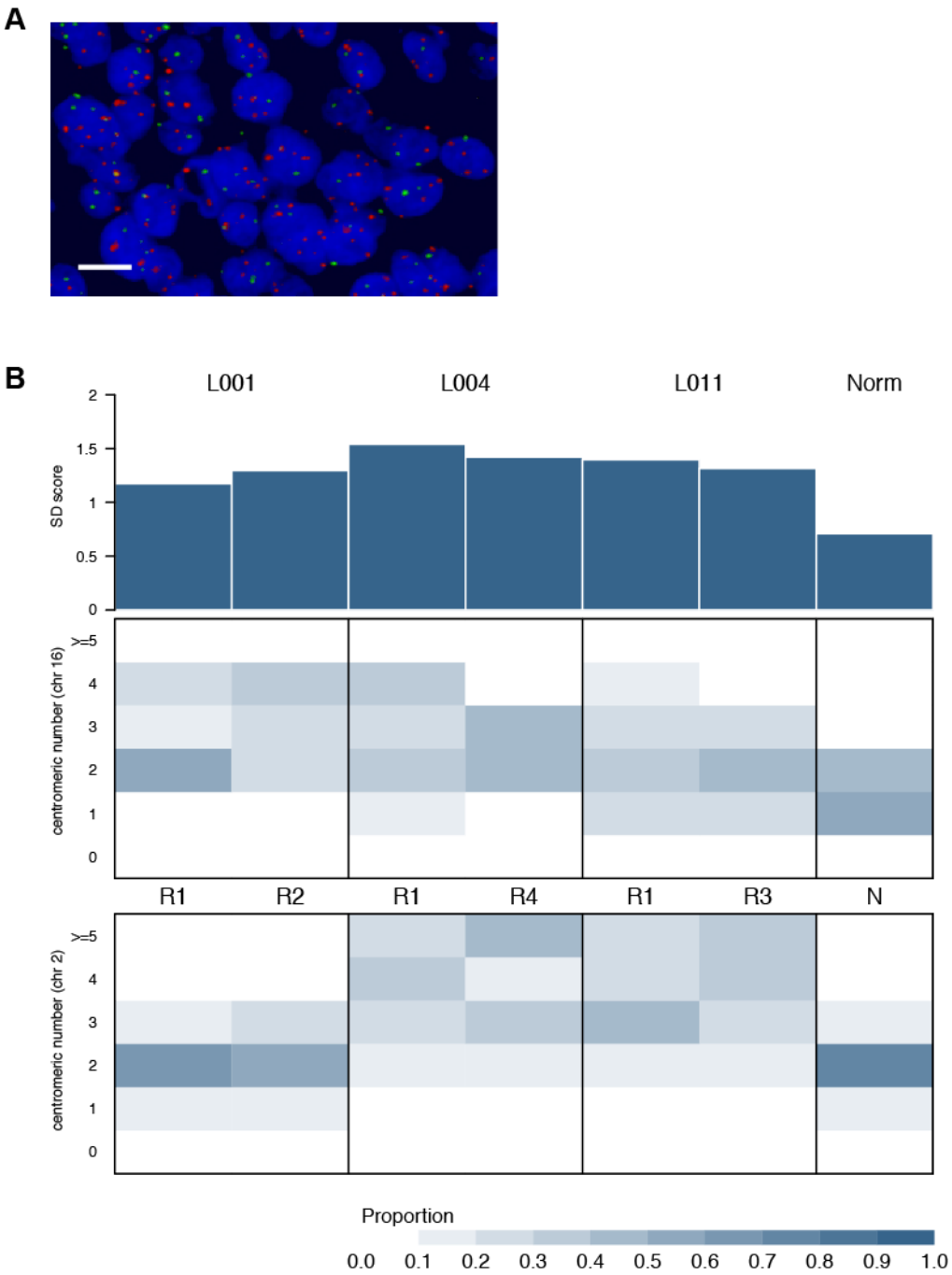

Fig. S9.
Centromeric FISH for M-seq tumors. **A)** Cenotromeric FISH for L004 region R4. Centromeric probes for chromosome 2 are shown in red, and for chromosome 16 in green. Scale bar, 10µm. **B)**The top panel shows the Shannon diversity index for each tumor region. The bottom panels show the proportion of different centromeric counts observed for chromosomes 2 and 16.

Figure S10

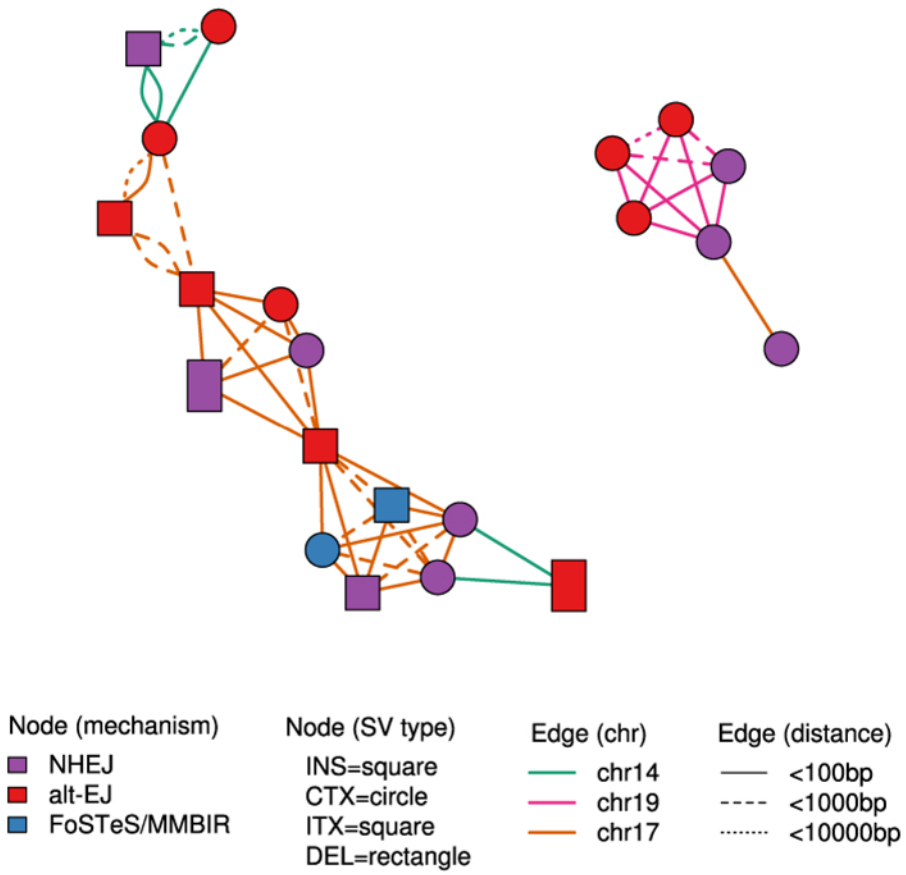

Fig. S10.

Structural variant breakpoints (nodes) within 10kb of one another are linked by edges representing both inter-breakpoint distance (line type) and chromosome (line color). Breakpoint homology profiling indicates that these highly localised breakpoint clusters involved either nonhomologous end-joining (purple) or alternative end-joining (red), indicative of double-strand break events, and are concentrated on chr17 (brown line) and chr19 (pink line).

Figure S11

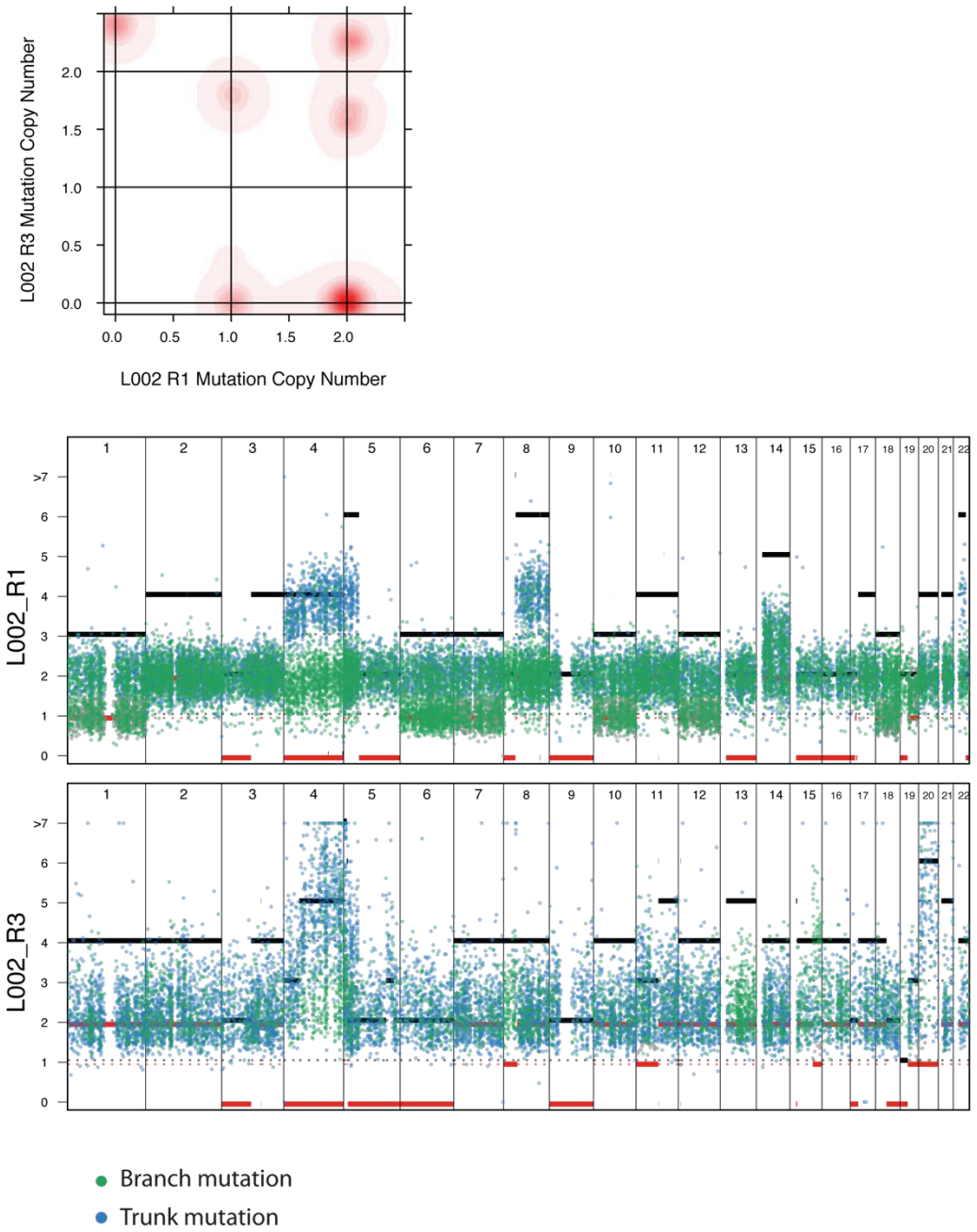

Fig. S11.

2D-dirichlet process plot (upper) using mutation copy numbers; increasing intensity of red indicates the location of a high posterior probability of a cluster. Private mutations clustered at a copy number 2 are observed in both regions R1 and R3, suggesting two independent genome-doubling events. Copy number profiles of L002 regions R1 and R3 (lower). Trunk mutations are showed in blue, whilst private mutations are depicted in green. Total copy number is depicted as a black line, with minor allele as a red line.

Figure S12

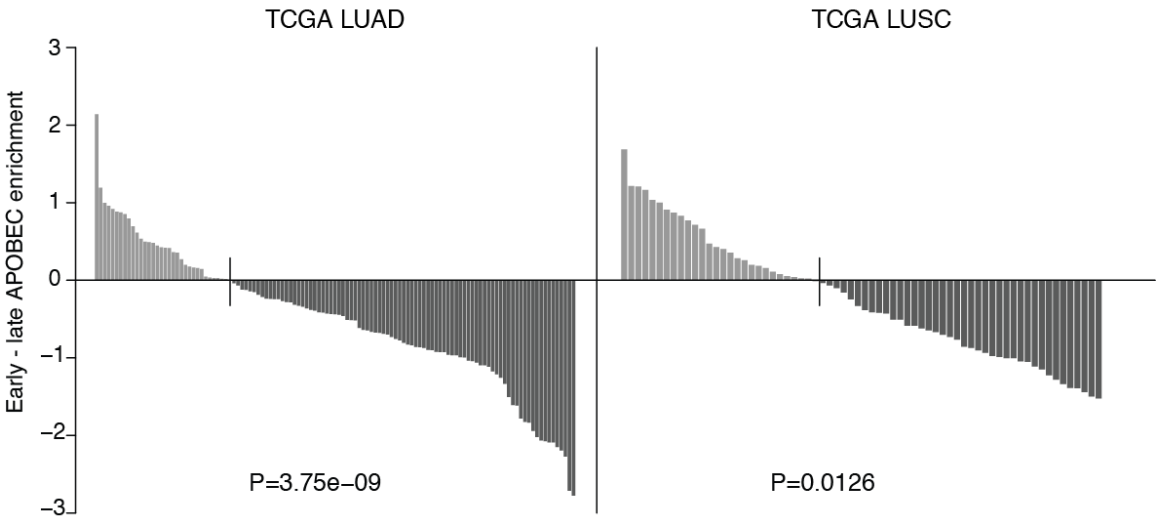

Fig. S12.

Rainfall plot showing difference in early and late APOBEC mutation enrichment for TCGA LUAD and TCGA LUSC. Each bar represents one TCGA tumour and its height corresponds to the difference in early versus late APOBEC enrichment. Only samples with significant APOBEC enrichment as well harbouring both early and late mutations are shown.

Figure S13

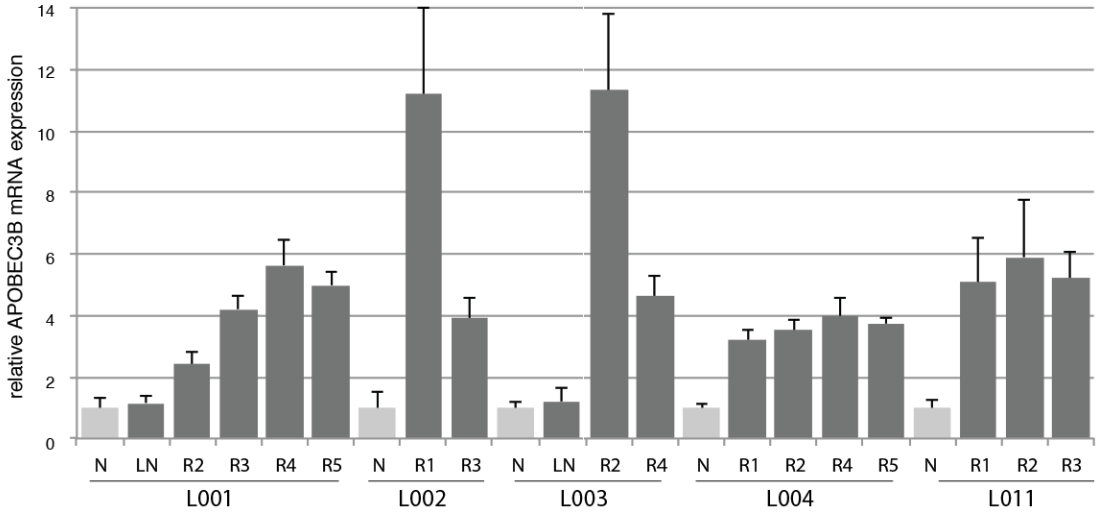

Fig. S13.

Graph showing APOBEC3B mRNA expression in tumor regions relative to the adjacent normal lung for each tumor, using TBP mRNA expression for normalization.

Figure S14

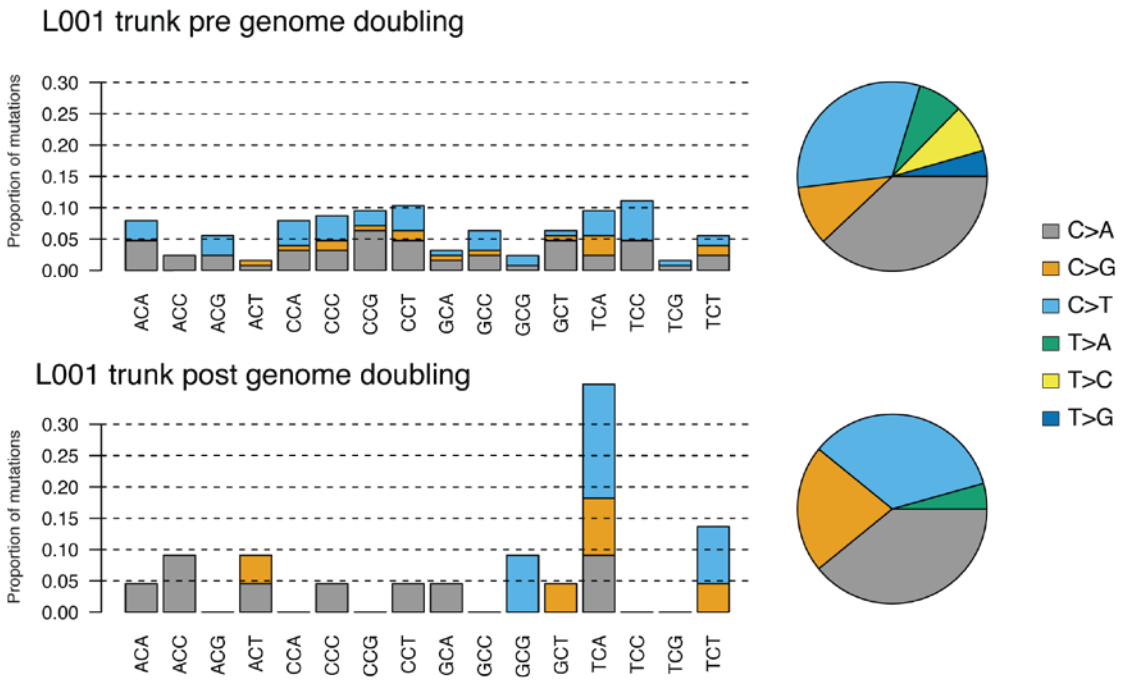

Fig. S14.

Truncal mutations pre and post genome doubling in L001. Stacked bar-charts (left) and pie charts (right) showing prevalence of different mutation types for truncal mutations occuring before and after genome doubling in L001.

Supplementary tables

Table S1.

Detailed patient characteristics.

Patient ID	Age (years)	Gender	Histology	Tumour Location /Max Diameter (mm)	Lymph Node(s)/ Location	Histological Stage	Stage (I-IV)	Regions Sequenced	Smoking Status (Pack-Years*)	Smoking Cessation (years)
L003	84	F	LUAD	RLL/22 & RUL/25	2/Station 4	pT4N2M0	IIIB	R2 (RLL), R4 (RUL), LN	Never	N/A
L008	75	M	LUAD	RUL/15 & RML/24	2/Hilar	pT4N1M0	IIIA	R1 (RUL), R3 (RML), LN	Former (25)	20
L001†	59	F	LUAD	LLL/30	3/Hilar	pT2aN1M0	IIA	R1-R5, LN	Former (10)	23
L004‡	73	M	Undiff. NSCLC	LLL/100	none	pT3N0M0	IIB	R1-R4	Current (50)	N/A
L011	49	F	LUAD	LUL/55	none	pT2aN0M0	IB	R1-R3	Current (45)	N/A
L002	78	M	LUAD/ LUSC	LUL/90	2/Station 5	pT3N2M0	IIIA	R1-R4	Current (>50)	N/A
LS01	69	M	LUSC	LLL/38	none	pT2aN0M0	IB	R1-R3	Former (50)	Unknown

Abbreviations: LLL, lower lobe; LUL, left upper lobe; RLL, right lower lobe; RUL, right upper lobe; RML, right middle lobe; R, region; LN, lymph node; Undiff, undifferentiated. * a pack-year is defined as the number of packs of cigarettes smoked per day multiplied by the number of years the person has smoked. †L001 presented a synchronous MEN1 syndrome-associated tumor, classified as separate tumor based on histological morphology, biochemical profile and octreotide scan imaging. ‡L004 presented a synchronous oesophageal adenocarcinoma, classified as separate primary tumors based on histological morphology and immunohistochemistry marker profile.

Table S2.

Detailed coverage information.

	Tumor	Region	Coverage	
			Mean	Median
M-seq WES	**L003**	B	131	103
		R1	103	76
		R2	100	77
		R4	107	76
	L008	B	111	87
		R1	106	78
		R3	91	72
		R5	102	79
	L001	N	57	44
		R1	99	72
		R2	119	87
		R3	67	49
		R4	115	86
		R5	131	96
	L004	B	60	47
		R1	108	76
		R2	188	140
		R4	100	76
		R5	174	129
	L011	B	99	78
		R1	92	70
		R2	95	73
		R3	92	70
	L002	B	132	121
		R1	91	70
		R2	94	74
		R3	148	117
		R4	95	74
M-seq WGS	**L008**	B	38	38
		R1	90	86
		R3	102	102
	L002	B	38	38
		R1	96	95
		R3	97	98
	LS01	B	20	NA
		R1	22	NA
		R2	17	NA
		R3	24	NA

Table S3.

Detailed information of candidate driver mutations

Tumor	Gene	Substitution		In Cosmic (v67)	Cancer-related gene			dbNSFP	Tumor region	Driver Category
		cDNA	AA		NSCLC	pan-cancer	TUSON			
L003	CTNNB1	C110T	S37F	YES	x	x	OG	4xD	R2	1
	EGFR	T2573G	L858R	YES	x	x	OG	4xD	all	1
	FOXP1	C75A	Y25X	STOP		x		STOP	R2	1
	BRAF	G89A	G30D	YES	x	x	OG	1xD	R1 (LN)	2
	IL7R	C394T	P132S	5aa		x	-	4xD	R2	3
	JAK1	G1837A	E613K	5aa		x		0xD	R2	3
	MUC1	C346G	H116D	NO		x		NA	R1,R4	3
	MLL	C8492T	S2831L	NO		x		4xD	R4	3
L008	BRAF	G1406C	G469A	YES	x	x	OG	4xD	all	1
	PIK3CA	G1624A	E542K	YES	x	x	OG	3xD	R3	1
	RB1	C2486G	S829X	YES	x	x	TS	STOP	all	1
	TP53	C577T	H193Y	YES	x	x	TS	4xD	all	1
	ALK	G2467A	G823R	5aa		x	OG	4xD	all	2
	CAMTA1	A1481C	Q494P	5aa		x		4xD	all	2
	HSP90AB1	C52T	Q18X	YES		x		STOP	all	2
	KDR	C2000A	P667H	5aa	x	x		4xD	all	2
	MLL2	G15028A	A5010T	5aa	x	x		NA	all	2
	ERCC4	G1093C	E365Q	5aa		x		1xD	all	3
	FANCM	A2068G	R690G	NO			TS	1xD	all	3
	FANCM	A2152C	N718H	5aa			TS	1xD	all	3
	KANSL1	C3260G	P1087R	5aa			TS	1xD	all	3
	MAGI2	G1723A	D575N	5aa			OG	2xD	R3	3
L001	EGFR	T2573G	L858R	YES	x	x	OG	4xD	all	1
	EP300	C823T	Q275X	STOP		x	TS	STOP	LN, R2	1
	RB1	C1853G	S618X	STOP	x	x	TS	STOP	all	1
	TP53	G818A	R273H	YES	x	x	TS	3xD	all	1
	DACH1	C1389A	S463R	5aa	x		TS	3xD	all	2
	FANCM	G2578T	E860X	NO			TS	STOP	R2	2
	GRM8	A498C	L166F	5aa	x			4xD	all	2
	MEN1	A280C	T94P	5aa		x	TS	3xD	all	2
	PTN	G258T	K86N	YES			OG	4xD	all	2
	ATP5B	G1342A	E448K	5aa		x		3xD	all	3
	ELF4	G964T	A322S	NO		x		0xD	R2	3
	NUP214	C2810G	SP exon21	NO		x		splicing	all	3
	PAX7	G203A	R68Q	NO		x		4xD	R2-5	3
	RBM10	G970C	A324P	NO	x		TS	1xD	all	3
L004	NF1	G8239A	D2747N	5aa	x	x	TS	2xD	all	1
	SETD2	G5009C	R1670T	5aa	x	x	TS	3xD	all	1
	AKAP9	G3589A	E1197K	5aa		x		3xD	R4	2
	CREBBP	G4198C	E1400Q	YES*		x	TS	2xD	all	2
	KANSL1	A2167T	R723W	5aa			TS	4xD	all	2
	LRP1B	T11479C	C3827R	5aa	x			4xD	all	2
	MLL3	C10781G	S3594C	5aa		x		4xD	R4	2
	ODAM	G775T	D259Y	5aa		x		3xD	R5	2
	ANO3	A58T	S20C	5aa			TS	4xD	all	3
	ATAD2	A3394G	T1132A	NO			TS	0xD	all	3
	DNAH12	G7181A	W2394X	STOP		x		STOP	R5	3
	EWSR1	G1766A	R589K	YES *		x		1xD	R5	3
	FLI1	T863A	L288Q	5aa		x		NA	all	3
	MYH9	G1317C	K439N	5aa		x		3xD	all	3
	NCOA1	G2494C	E832Q	5aa		x		0xD	R4	3
	ODAM	C268A	Q90K	NO		x		0xD	all	3
	TAF15	A215G	N72S	NO		x		0xD	all	3
	TRIM23	A772T	I258F	NO			OG	3xD	all	3
L011	BRAF	T1799A	V600E	YES	x	x	OG	3xD	all	1
	LRP1B	C7638G	Y2546X	5aa	x			STOP	all	1
	TP53	G892T	E298X	YES	x	x	TS	STOP	all	1
	AKT1	G44C	R15P	5aa		x	OG	3xD	all	2

	CHD2	A3065T	K1022M	5aa			TS	3xD	R2	2	
	FAT1	T3320G	L1107R	5aa		x	TS	NA	all	2	
	GRM8	T2522C	I841T	5aa	x			4xD	all	2	
	GRM8	C1696A	R566S	YES	x			2xD	all	2	
	HOOK3	G322T	E108X	5aa		x		STOP	all	2	
	JAK1	G29A	C10Y	5aa		x		2xD	all	2	
	PIM1	G142T	G48C	5aa		x		4xD	all	2	
	PTPRD	G3076T	V1026F	5aa	x			4xD	all	2	
	CLTCL1	G2914T	D972Y	5aa		x	-	NA	all	3	
	MYCN	G520T	A174S	NO		x	OG	0xD	all	3	
	MYH9	G250T	E84X	5aa		x		STOP	all	3	
	PLEKHA 6	G2854A	E952K	5aa			TS	4xD	all	3	
	TBX3	C1872A	H624Q	5aa			TS	1xD	all	3	
	XPC	G1325T	G442V	5aa		x		NA	all	3	
L002	FAT1	C76T	R26X	STOP		x	TS	STOP	R3,R4	1	
	TGFBR1	C1310G	S437X	STOP	x			STOP	R1,R2	1	
	TP53	G404A	C135Y	YES	x	x	TS	4xD	all	1	
	ARHGAP35	G2915-	R972fs*55	FS		x	TS	FS	R1,R2	2	
	CHD8	G1232C	G411A	5aa		x	TS	NA	all	2	
	FBXW7	G1394C	R465P	YES *	x	x	TS	4xD	all	2	
	LRP1B	A7715T	D2572V	5aa	x			3xD	all	2	
	LRP1B	T3968C	V1323A	5aa	x			2xD	all	2	
	PTPRD	G1999A	E667K	5aa	x			3xD	R1,R2	2	
	RECQL4	T1895C	M632T	5aa		x		NA	all	2	
	TRIM33	G1748T	G583V	5aa		x		3xD	all	2	
	USP28	A1971T	SP exon 16	splice			TS	splicing	all	2	
	ZFHX4	C6394A	P2132T	YES	x			NA	R1,R2	2	
	ARHGAP5	A38G	Y13C	NO			TS	3xD	all	3	
	HOXA11	G688A	E230K	5aa		x		1xD	R1,R2	3	
	KCNQ5	G1247T	SP exon 9	splice			OG	splicing	all	3	
	LIMCH1	A333T	R111S	5aa			TS	1xD	all	3	
	MBOAT2	A498T	L166F	NO			OG	2xD	R1,R2	3	
	MGA	C72G	F24L	5aa		x	TS	NA	R3,R4	3	
	MITF	C92T	P31L	NO		x		2xD	R3,R4	3	
	NTRK1	C47A	A16D	NO	x	x		1xD	R1-3	3	
	PRDM16	C1967T	P656L	5aa		x		0xD	R3,R4	3	
	SMC3	T2457G	I819M	5aa		x		2xD	R1-3	3	
	SMO	G385A	V129I	NO		x		2xD	all	3	
	WHSC1L1	G1375A	E459K	NO		x		1xD	R1,R2	3	
	ZFHX4	C8014A	Q2672K	5aa	x			NA	all	3	
	ZFHX4	C5131A	P1711T	5aa	x			NA	R1,R2	3	
	ZFHX4	G9271T	G3091C	5aa	x			NA	R3,R4	3	
	ZNF292	A1436T	Y479F	NO			TS	NA	all	3	
LS01	ASXL1	G566T	G189V	splice		x	TS	splicing	all	1	
	PTEN	G209+1A	SP exon3	splice	x	x	TS	splicing	all	1	
	TP53	A97-2G	SP exon 5	splice	x	x	TS	splicing	all	1	
	TP53	G797A	G266E	YES	x	x	TS	4xD	all	1	
	ALK	G1753T	A585S	YES *	x	x	OG	1xD	all	2	
	CFTR	C1057A	Q353K	5aa	x			NA	all	2	
	CREBBP	G7285T	D2429Y	YES		x	TS	3xD	all	2	
	CREBBP	G1690T	D564Y	5aa		x	TS	4xD	all	2	
	MYH9	A452G	Y151C	YES		x		4xD	all	2	
	NOTCH2	A809G	Q270R	5aa		x	TS	1xD	all	2	
	AOAH	T688A	C230S	5aa			TS	NA	all	3	
	ASXL2	G1448A	C483Y	NO		x	TS	NA	all	3	
	COL1A1	G1202A	G401D	splice		x		splicing	all	3	
	ETV6	G206T	W69L	5aa		x		NA	all	3	
	JUN	A35T	D12V	5aa		x		3xD	all	3	
	MLL	A1699C	T567P	NO		x		NA	R3	3	
	MLL3	C4209G	F1403L	NO		x		NA	R3	3	

Abbreviations: AA, amino acid; subst, substitution; FS, frameshift; 5aa, within 5 amino acids; Cat, category driver mutation; SP, splicing; *different substitution in COSMIC database v67. x indicates that gene is

identified as potential driver gene in NSCLC sequencing or pan-cancer data; TS or OG indicate identification as Tumor Suppressor or OncoGene by TUSON (16).

Table S4

Chromosomal rearrangements identified in L002 and L008.

Sample	SV type	left position	right position	mechanism	R1	R3
L002	CTX	chr2:155998950	chr22:29065839	alt-EJ	present	present
		chr2:236363124	chr3:167098659	VNTR	present	present
		chr18:32544412	chr8:118344509	VNTR	present	present
		chr1:31404868	chr6:47733848	alt-EJ	present	-
		chr5:34614984	chr22:42503929	alt-EJ	present	-
		chr6:13191304	chr18:22134677	NHEJ	present	-
		chrX:11952758	chr7:28756841	alt-EJ	present	-
		chrX:11953158	chr5:11979024	alt-EJ	present	-
		chr11:93465300	chr4:124476626	NHEJ	-	present
		chr12:116282296	chr8:38486564	NHEJ	-	present
		chr5:147360434	chr2:233682520	alt-EJ	-	present
		chr21:47629746	chr15:72570211	NHEJ	-	present
		chr8:38486329	chr12:116281769	NHEJ	-	present
		chr18:34666016	chr21:45437425	alt-EJ	-	present
	ITX	chr8:116143651	chr8:118139173	FoSTeS	present	present
		chr10:54090613	chr10:54104578	NHEJ	present	-
		chr10:54322851	chr10:54355481	NHEJ	present	-
		chr3:195620398	chr3:190198721	NHEJ	present	-
		chr6:13191101	chr6:76640812	alt-EJ	present	-
		chrX:11951845	chrX:80996702	alt-EJ	present	-
		chr12:102694542	chr12:104762458	alt-EJ	-	present
	DEL	chr4:141659964	chr4:142707155	alt-EJ	present	present
		chr5:5321894	chr5:13526542	NHEJ	present	present
		chr4:187325363	chr4:189488078	alt-EJ	present	-
		chr2:235046417	chr2:237576417	NHEJ	-	present
		chr5:21812115	chr5:45755186	alt-EJ	-	present
		chr8:33575222	chr8:122792951	alt-EJ	-	present
	INS	chr1:209720908	chr1:209401475	alt-EJ	-	present
		chr10:53461830	chr10:53401722	alt-EJ	-	present
		chr11:65387961	chr11:65255262	NHEJ	-	present
L008	CTX	chr1:41420591	chr6:153820131	NHEJ	present	present
		chr1:41429899	chr6:138283492	NHEJ	present	present
		chr1:41522450	chr6:138288122	alt-EJ	present	present
		chr3:142094843	chr19:675705	VNTR	present	present
		chr4:21087179	chr18:66132623	alt-EJ	present	present
		chr6:139208791	chr1:41385823	alt-EJ	present	present
		chr8:118172674	chr4:111859631	VNTR	present	present
		chr14:21641322	chr19:10070782	alt-EJ	present	present
		chr14:21659076	chr17:29633312	FoSTeS	present	present
		chr14:22931256	chr17:27047855	NHEJ	present	present
		chr14:22945889	chr17:40582022	NHEJ	present	present
		chr14:23865070	chr20:9485468	alt-EJ	present	present
		chr14:43202845	chr17:27041499	NHEJ	present	present
		chr14:43221709	chr17:30598522	alt-EJ	present	present
		chr14:43226786	chr17:40484004	alt-EJ	present	present
		chr14:43983534	chr17:32557240	NHEJ	present	present
		chr14:45606236	chr17:29633089	NHEJ	present	present
		chr14:45607709	chr17:40582243	NHEJ	present	present
		chr17:32006130	chr14:21438822	NHEJ	present	present
		chr17:32531855	chr14:45639647	alt-EJ	present	present
		chr17:33970201	chr14:43983604	alt-EJ	present	present
		chr17:40583254	chr19:10080082	NHEJ	present	present

		chr18:3374296	chr11:93333146	VNTR	present	present
		chr18:63100198	chr4:20528186	NHEJ	present	present
		chr18:63106534	chr3:88076969	alt-EJ	present	present
		chr19:10077608	chr14:45881885	alt-EJ	present	present
		chr19:10077623	chr14:23858748	alt-EJ	present	present
		chr19:10077841	chr17:32532990	NHEJ	present	present
		chr20:6037223	chr17:27048797	FoSTeS	present	present
		chr20:7462891	chr17:33945011	alt-EJ	present	present
		chr20:7463341	chr17:32564592	alt-EJ	present	present
		chr20:7469802	chr14:43913262	NHEJ	present	present
	ITX	chr6:138283919	chr6:116907520	alt-EJ	present	present
		chr14:45624623	chr14:22934184	NHEJ	present	present
		chr17:27047629	chr17:32563709	alt-EJ	present	present
		chr17:27048551	chr17:30619424	FoSTeS	present	present
		chr17:32556109	chr17:40483031	alt-EJ	present	present
		chr18:1257044	chr18:54722755	NHEJ	present	present
		chr2:40299203	chr2:41619314	NHEJ	present	-
		chr2:41618561	chr2:40304683	NA	present	-
		chr4:21144053	chr4:21639085	NHEJ	present	-
		chr4:37953544	chr4:21638626	alt-EJ	present	-
	DEL	chr14:22934146	chr14:43204483	alt-EJ	present	present
		chr17:29659502	chr17:32564783	NHEJ	present	present
		chr18:25060027	chr18:53994682	NHEJ	present	present
	INS	chr1:45843008	chr1:39747926	alt-EJ	present	present
		chr2:42346589	chr2:42187285	alt-EJ	present	present
		chr11:103580980	chr11:80028837	NHEJ	present	present
		chr17:39957797	chr17:27041102	NHEJ	present	present
		chr18:25357067	chr18:25063438	alt-EJ	present	present
		chr18:25382005	chr18:1168244	NHEJ	present	present
		chr18:66410779	chr18:63101330	alt-EJ	present	present

Abbreviations: CTX, interchromosomal rearrangement; ITX, intrachromosomal rearrangement; DEL, deletion; INS, insertion.

References and Notes

1. World Health Organization, www.who.int/cancer/en/ (2013).

2. R. Siegel, D. Naishadham, A. Jemal, Cancer statistics, 2013. *CA Cancer J. Clin.* **63**, 11–30 (2013). Medline doi:10.3322/caac.21166

3. H. Tanaka, K. Yanagisawa, K. Shinjo, A. Taguchi, K. Maeno, S. Tomida, Y. Shimada, H. Osada, T. Kosaka, H. Matsubara, T. Mitsudomi, Y. Sekido, M. Tanimoto, Y. Yatabe, T. Takahashi, Lineage-specific dependency of lung adenocarcinomas on the lung development regulator TTF-1. *Cancer Res.* **67**, 6007–6011 (2007). Medline doi:10.1158/0008-5472.CAN-06-4774

4. B. A. Weir, M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y. Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M. S. Tsao, D. Twomey, R. G. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, I. I. Wistuba, M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, M. Meyerson, Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007). Medline doi:10.1038/nature06358

5. L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, L. Fulton, R. S. Fulton, Q. Zhang, M. C. Wendl, M. S. Lawrence, D. E. Larson, K. Chen, D. J. Dooling, A. Sabo, A. C. Hawes, H. Shen, S. N. Jhangiani, L. R. Lewis, O. Hall, Y. Zhu, T. Mathew, Y. Ren, J. Yao, S. E. Scherer, K. Clerc, G. A. Metcalf, B. Ng, A. Milosavljevic, M. L. Gonzalez-Garay, J. R. Osborne, R. Meyer, X. Shi, Y. Tang, D. C. Koboldt, L. Lin, R. Abbott, T. L. Miner, C. Pohl, G. Fewell, C. Haipek, H. Schmidt, B. H. Dunford-Shore, A. Kraja, S. D. Crosby, C. S. Sawyer, T. Vickery, S. Sander, J. Robinson, W. Winckler, J. Baldwin, L. R. Chirieac, A. Dutt, T. Fennell, M. Hanna, B. E. Johnson, R. C. Onofrio, R. K. Thomas, G. Tonon, B. A. Weir, X. Zhao, L. Ziaugra, M. C. Zody, T. Giordano, M. B. Orringer, J. A. Roth, M. R. Spitz, I. I. Wistuba, B. Ozenberger, P. J. Good, A. C. Chang, D. G. Beer, M. A. Watson, M. Ladanyi, S. Broderick, A. Yoshizawa, W. D. Travis, W. Pao, M. A. Province, G. M. Weinstock, H. E. Varmus, S. B. Gabriel, E. S. Lander, R. A. Gibbs, M. Meyerson, R. K. Wilson, Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008). Medline doi:10.1038/nature07423

6. Z. Kan, B. S. Jaiswal, J. Stinson, V. Janakiraman, D. Bhatt, H. M. Stern, P. Yue, P. M. Haverty, R. Bourgon, J. Zheng, M. Moorhead, S. Chaudhuri, L. P. Tomsho, B. A. Peters, K. Pujara, S. Cordes, D. P. Davis, V. E. Carlton, W. Yuan, L. Li, W. Wang, C. Eigenbrot, J. S. Kaminker, D. A. Eberhard, P. Waring, S. C. Schuster, Z. Modrusan, Z. Zhang, D. Stokoe, F. J. de Sauvage, M. Faham, S. Seshagiri, Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010). Medline doi:10.1038/nature09208

7. S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson, G. Getz, Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012). Medline doi:10.1038/nbt.2203

8. T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C. Z. Zhang, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013). Medline doi:10.1038/ng.2760

9. R. Govindan, L. Ding, M. Griffith, J. Subramanian, N. D. Dees, K. L. Kanchi, C. A. Maher, R. Fulton, L. Fulton, J. Wallis, K. Chen, J. Walker, S. McDonald, R. Bose, D. Ornitz, D. Xiong, M. You, D. J. Dooling, M. Watson, E. R. Mardis, R. K. Wilson, Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121–1134 (2012). Medline doi:10.1016/j.cell.2012.08.024

10. E. D. Pleasance, P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert, D. Jones, M. L. Lin, D. Beare, K. W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H. R. Davies, G. R. Ordoñez, L. J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J. W. Teague, J. Mangion, Y. A. Sun, S. F. McLaughlin, H. E. Peckham, E. F. Tsung, G. L. Costa, C. C. Lee, J. D. Minna, A. Gazdar, E. Birney, M. D. Rhodes, K. J. McKernan, M. R. Stratton, P. A. Futreal, P. J. Campbell, A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010). Medline doi:10.1038/nature08629

11. W. Lee, Z. Jiang, J. Liu, P. M. Haverty, Y. Guan, J. Stinson, P. Yue, Y. Zhang, K. P. Pant, D. Bhatt, C. Ha, S. Johnson, M. I. Kennemer, S. Mohan, I. Nazarenko, C. Watanabe, A. B. Sparks, D. S. Shames, R. Gentleman, F. J. de Sauvage, H. Stern, A. Pandita, D. G. Ballinger, R. Drmanac, Z. Modrusan, S. Seshagiri, Z. Zhang, The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010). Medline doi:10.1038/nature09004

12. G. P. Pfeifer, P. Hainaut, On the origin of G —> T transversions in lung cancer. *Mutat. Res.* **526**, 39–43 (2003). Medline doi:10.1016/S0027-5107(03)00013-7

13. S. A. Roberts, M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena, S. Harris, R. R. Shah, M. A. Resnick, G. Getz, D. A. Gordenin, An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013). Medline doi:10.1038/ng.2702

14. M. B. Burns, N. A. Temiz, R. S. Harris, Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013). Medline doi:10.1038/ng.2701

15. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V.

Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. StrattonAustralian Pancreatic Cancer Genome InitiativeICGC Breast Cancer ConsortiumICGC MMML-Seq ConsortiumICGC PedBrain, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). Medline doi:10.1038/nature12477

16. Detailed information on methods is available on *Science* online.

17. S. V. Sharma, D. W. Bell, J. Settleman, D. A. Haber, Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7**, 169–181 (2007). Medline doi:10.1038/nrc2088

18. N. Bolli, H. Avet-Loiseau, D. C. Wedge, P. Van Loo, L. B. Alexandrov, I. Martincorena, K. J. Dawson, F. Iorio, S. Nik-Zainal, G. R. Bignell, J. W. Hinton, Y. Li, J. M. Tubio, S. McLaren, S. O' Meara, A. P. Butler, J. W. Teague, L. Mudie, E. Anderson, N. Rashid, Y. T. Tai, M. A. Shammas, A. S. Sperling, M. Fulciniti, P. G. Richardson, G. Parmigiani, F. Magrangeas, S. Minvielle, P. Moreau, M. Attal, T. Facon, P. A. Futreal, K. C. Anderson, P. J. Campbell, N. C. Munshi, Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014). Medline doi:10.1038/ncomms3997

19. M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. de Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansén, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. A. Jänne, V. A. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. A. Garraway, G. Getz, M. Meyerson, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012). Medline doi:10.1016/j.cell.2012.08.029

20. M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, G. Getz, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014). Medline doi:10.1038/nature12912

21. P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A. L. Børresen-Dale, V. N. Kristensen, Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16910–16915 (2010). Medline doi:10.1073/pnas.1009843107

22. S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Jönsson, O.

Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerød, S. A. Aparicio, A. Tutt, A. M. Sieuwerts, Å. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A. L. Børresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell; Breast Cancer Working Group of the International Cancer Genome Consortium, The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012). Medline doi:10.1016/j.cell.2012.04.023

23. A. Malhotra, M. Lindberg, G. G. Faust, M. L. Leibowitz, R. A. Clark, R. M. Layer, A. R. Quinlan, I. M. Hall, Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res.* **23**, 762–776 (2013). Medline doi:10.1101/gr.143677.112

24. S. F. Bunting, A. Nussenzweig, End-joining, translocations and cancer. *Nat. Rev. Cancer* **13**, 443–454 (2013). Medline doi:10.1038/nrc3537

25. C. Z. Zhang, M. L. Leibowitz, D. Pellman, Chromothripsis and beyond: Rapid genome evolution from complex chromosomal rearrangements. *Genes Dev.* **27**, 2513–2530 (2013). Medline doi:10.1101/gad.229559.113

26. S. C. Baca, D. Prandi, M. S. Lawrence, J. M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T. Y. MacDonald, M. Ghandi, E. Van Allen, G. V. Kryukov, A. Sboner, J. P. Theurillat, T. D. Soong, E. Nickerson, D. Auclair, A. Tewari, H. Beltran, R. C. Onofrio, G. Boysen, C. Guiducci, C. E. Barbieri, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, A. H. Ramos, W. Winckler, M. Cipicchio, K. Ardlie, P. W. Kantoff, M. F. Berger, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, O. Elemento, G. Getz, F. Demichelis, M. A. Rubin, L. A. Garraway, Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013). Medline doi:10.1016/j.cell.2013.03.021

27. M. Gerlinger, S. Horswell, J. Larkin, A. J. Rowan, M. P. Salm, I. Varela, R. Fisher, N. McGranahan, N. Matthews, C. R. Santos, P. Martinez, B. Phillimore, S. Begum, A. Rabinowitz, B. Spencer-Dene, S. Gulati, P. A. Bates, G. Stamp, L. Pickering, M. Gore, D. L. Nicol, S. Hazell, P. A. Futreal, A. Stewart, C. Swanton, Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014). Medline doi:10.1038/ng.2891

28. M. Gerlinger, A. J. Rowan, S. Horswell, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N. Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C. R. Santos, M. Nohadani, A. C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P. A. Futreal, C. Swanton, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012). Medline doi:10.1056/NEJMoa1113205

29. M. Reck, D. F. Heigener, T. Mok, J. C. Soria, K. F. Rabe, Management of non-small-cell lung cancer: Recent developments. *Lancet* **382**, 709–719 (2013). Medline doi:10.1016/S0140-6736(13)61502-0

30. H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B. A. Gusterson, C. Cooper, J. Shipley, D.

Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G. J. Riggins, D. D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J. W. Ho, S. Y. Leung, S. T. Yuen, B. L. Weber, H. F. Seigler, T. L. Darrow, H. Paterson, R. Marais, C. J. Marshall, R. Wooster, M. R. Stratton, P. A. Futreal, Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002). Medline doi:10.1038/nature00766

31. S. Kang, A. G. Bader, P. K. Vogt, Phosphatidylinositol 3-kinase mutations identified in human cancer are oncogenic. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 802–807 (2005). Medline doi:10.1073/pnas.0408864102

32. H. Shi, W. Hugo, X. Kong, A. Hong, R. C. Koya, G. Moriceau, T. Chodon, R. Guo, D. B. Johnson, K. B. Dahlman, M. C. Kelley, R. F. Kefford, B. Chmielowski, J. A. Glaspy, J. A. Sosman, N. van Baren, G. V. Long, A. Ribas, R. S. Lo, Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. *Cancer Discov.* **4**, 80–93 (2014). Medline doi:10.1158/2159-8290.CD-13-0642

33. C. J. Lord, A. Ashworth, The DNA damage response and cancer therapy. *Nature* **481**, 287–294 (2012). Medline doi:10.1038/nature10760

34. S. M. Dewhurst, N. McGranahan, R. A. Burrell, A. J. Rowan, E. Grönroos, D. Endesfelder, T. Joshi, D. Mouradov, P. Gibbs, R. L. Ward, N. J. Hawkins, Z. Szallasi, O. M. Sieber, C. Swanton, Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014). Medline doi:10.1158/2159-8290.CD-13-0285

35. R. Sotillo, J. M. Schvartzman, N. D. Socci, R. Benezra, Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal. *Nature* **464**, 436–440 (2010). Medline doi:10.1038/nature08803

36. A. J. Lee, D. Endesfelder, A. J. Rowan, A. Walther, N. J. Birkbak, P. A. Futreal, J. Downward, Z. Szallasi, I. P. Tomlinson, M. Howell, M. Kschischo, C. Swanton, Chromosomal instability confers intrinsic multidrug resistance. *Cancer Res.* **71**, 1858–1870 (2011). Medline doi:10.1158/0008-5472.CAN-10-3604

37. J. C. Soria, C. Cruz, R. Bahleda, J. P. Delord, L. Horn, R. S. Herbst, D. Spigel, A. Mokatrin, G. Fine, S. Gettinger, Clinical activity, safety and biomarkers of PD-L1 blockade in non-small cell lung cancer (NSCLC). *Eur. J. Cancer* **49** (suppl. 2), 3408 (2013).

38. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). Medline doi:10.1093/bioinformatics/btp324

39. D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, R. K. Wilson, VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012). Medline doi:10.1101/gr.129684.111

40. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). Medline doi:10.1093/bioinformatics/btp352

41. L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan, J. F. McMichael, J. W. Wallis, C. Lu, D.

Shen, C. C. Harris, D. J. Dooling, R. S. Fulton, L. L. Fulton, K. Chen, H. Schmidt, J. Kalicki-Veizer, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, M. C. Wendl, S. Heath, M. A. Watson, D. C. Link, M. H. Tomasson, W. D. Shannon, J. E. Payton, S. Kulkarni, P. Westervelt, M. J. Walter, T. A. Graubert, E. R. Mardis, R. K. Wilson, J. F. DiPersio, Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012). Medline doi:10.1038/nature10738

42. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009). Medline doi:10.1093/bioinformatics/btp394

43. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010). Medline doi:10.1093/nar/gkq603

44. X. Liu, X. Jian, E. Boerwinkle, dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011). Medline doi:10.1002/humu.21517

45. K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, S. Kanaya, Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011). Medline doi:10.1093/nar/gkr344

46. K. J. McKernan, H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu, E. F. Tsung, C. R. Clouser, C. Duncan, J. K. Ichikawa, C. C. Lee, Z. Zhang, S. S. Ranade, E. T. Dimalanta, F. C. Hyland, T. D. Sokolsky, L. Zhang, A. Sheridan, H. Fu, C. L. Hendrickson, B. Li, L. Kotler, J. R. Stuart, J. A. Malek, J. M. Manning, A. A. Antipova, D. S. Perez, M. P. Moore, K. C. Hayashibara, M. R. Lyons, R. E. Beaudoin, B. E. Coleman, M. W. Laptewicz, A. E. Sannicandro, M. D. Rhodes, R. K. Gottimukkala, S. Yang, V. Bafna, A. Bashir, A. MacBride, C. Alkan, J. M. Kidd, E. E. Eichler, M. G. Reese, F. M. De La Vega, A. P. Blanchard, Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009). Medline doi:10.1101/gr.091868.109

47. E. Papaemmanuil, M. Cazzola, J. Boultwood, L. Malcovati, P. Vyas, D. Bowen, A. Pellagatti, J. S. Wainscoat, E. Hellstrom-Lindberg, C. Gambacorti-Passerini, A. L. Godfrey, I. Rapado, A. Cvejic, R. Rance, C. McGee, P. Ellis, L. J. Mudie, P. J. Stephens, S. McLaren, C. E. Massie, P. S. Tarpey, I. Varela, S. Nik-Zainal, H. R. Davies, A. Shlien, D. Jones, K. Raine, J. Hinton, A. P. Butler, J. W. Teague, E. J. Baxter, J. Score, A. Galli, M. G. Della Porta, E. Travaglino, M. Groves, S. Tauro, N. C. Munshi, K. C. Anderson, A. El-Naggar, A. Fischer, V. Mustonen, A. J. Warren, N. C. Cross, A. R. Green, P. A. Futreal, M. R. Stratton, P. J. Campbell; Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium, Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N. Engl. J. Med.* **365**, 1384–1395 (2011). Medline doi:10.1056/NEJMoa1103283

48. I. Varela, P. Tarpey, K. Raine, D. Huang, C. K. Ong, P. Stephens, H. Davies, D. Jones, M. L. Lin, J. Teague, G. Bignell, A. Butler, J. Cho, G. L. Dalgliesh, D. Galappaththige, C. Greenman, C. Hardy, M. Jia, C. Latimer, K. W. Lau, J. Marshall, S. McLaren, A.

Menzies, L. Mudie, L. Stebbings, D. A. Largaespada, L. F. Wessels, S. Richard, R. J. Kahnoski, J. Anema, D. A. Tuveson, P. A. Perez-Mancera, V. Mustonen, A. Fischer, D. J. Adams, A. Rust, W. Chan-on, C. Subimerb, K. Dykema, K. Furge, P. J. Campbell, B. T. Teh, M. R. Stratton, P. A. Futreal, Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–542 (2011). Medline doi:10.1038/nature09639

49. G. Schwarz, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978). doi:10.1214/aos/1176344136

50. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011). Medline doi:10.1093/molbev/msr121

51. R. McLendon *et al*.Cancer Genome Atlas Research Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008). Medline doi:10.1038/nature07385

52. P. S. Hammerman, M. S. Lawrence, D. Voet, R. Jing, K. Cibulskis, A. Sivachenko, P. Stojanov, A. McKenna, E. S. Lander, S. Gabriel, G. Getz, C. Sougnez, M. Imielinski, E. Helman, B. Hernandez, N. H. Pho, M. Meyerson, A. Chu, H.-J. E. Chun, A. J. Mungall, E. Pleasance, A. Gordon Robertson, P. Sipahimalani, D. Stoll, M. Balasundaram, I. Birol, Y. S. N. Butterfield, E. Chuah, R. J. N. Coope, R. Corbett, N. Dhalla, R. Guin, A. He, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, K. Mungall, K. Ming Nip, A. Olshen, J. E. Schein, J. R. Slobodan, A. Tam, N. Thiessen, R. Varhol, T. Zeng, Y. Zhao, S. J. M. Jones, M. A. Marra, G. Saksena, A. D. Cherniack, S. E. Schumacher, B. Tabak, S. L. Carter, N. H. Pho, H. Nguyen, R. C. Onofrio, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, P. S. Hammerman, G. Getz, M. Meyerson, A. Protopopov, J. Zhang, A. Hadjipanayis, S. Lee, R. Xi, L. Yang, X. Ren, H. Zhang, S. Shukla, P.-C. Chen, P. Haseley, E. Lee, L. Chin, P. J. Park, R. Kucherlapati, N. D. Socci, Y. Liang, N. Schultz, L. Borsu, A. E. Lash, A. Viale, C. Sander, M. Ladanyi, J. Todd Auman, K. A. Hoadley, M. D. Wilkerson, Y. Shi, C. Liquori, S. Meng, L. Li, Y. J. Turman, M. D. Topal, D. Tan, S. Waring, E. Buda, J. Walsh, C. D. Jones, P. A. Mieczkowski, D. Singh, J. Wu, A. Gulabani, P. Dolina, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, B. D. O'Connor, J. F. Prins, J. Liu, D. Y. Chiang, D. Neil Hayes, C. M. Perou, L. Cope, L. Danilova, D. J. Weisenberger, D. T. Maglinte, F. Pan, D. J. Van Den Berg, T. Triche Jr., J. G. Herman, S. B. Baylin, P. W. Laird, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C.-J. Wu, S. Yingchun Liu, M. S. Lawrence, L. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, M.-D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, N. Schultz, R. Sinha, G. Ciriello, E. Cerami, B. Gross, A. Jacobsen, J. Gao, B. Arman Aksoy, N. Weinhold, R. Ramirez, B. S. Taylor, Y. Antipin, B. Reva, R. Shen, Q. Mo, V. Seshan, P. K. Paik, M. Ladanyi, C. Sander, R. Akbani, N. Zhang, B. M. Broom, T. Casasent, A. Unruh, C. Wakefield, R. Craig Cason, K. A. Baggerly, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, J. Zhu, C. Szeto, G. K. Scott, C. Yau, S. Ng, T. Goldstein, P. Waltman, A. Sokolov, K. Ellrott, E. A. Collisson, D. Zerbino, C. Wilks, S. Ma, B. Craft, M. D. Wilkerson, J. Todd Auman, K. A. Hoadley, Y. Du, C. Cabanski, V. Walter, D. Singh, J. Wu, A. Gulabani, T. Bodenheimer, A. P. Hoyle, J. V.

Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, J. S. Marron, Y. Liu, K. Wang, J. Liu, J. F. Prins, D. Neil Hayes, C. M. Perou, C. J. Creighton, Y. Zhang, W. D. Travis, N. Rekhtman, J. Yi, M. C. Aubry, R. Cheney, S. Dacic, D. Flieder, W. Funkhouser, P. Illei, J. Myers, M.-S. Tsao, R. Penny, D. Mallery, T. Shelton, M. Hatfield, S. Morris, P. Yena, C. Shelton, M. Sherman, J. Paulauskis, M. Meyerson, S. B. Baylin, R. Govindan, R. Akbani, I. Azodo, D. Beer, R. Bose, L. A. Byers, D. Carbone, L.-W. Chang, D. Chiang, A. Chu, E. Chun, E. Collisson, L. Cope, C. J. Creighton, L. Danilova, L. Ding, G. Getz, P. S. Hammerman, D. Neil Hayes, B. Hernandez, J. G. Herman, J. Heymach, C. Ida, M. Imielinski, B. Johnson, I. Jurisica, J. Kaufman, F. Kosari, R. Kucherlapati, D. Kwiatkowski, M. Ladanyi, M. S. Lawrence, C. A. Maher, A. Mungall, S. Ng, W. Pao, M. Peifer, R. Penny, G. Robertson, V. Rusch, C. Sander, N. Schultz, R. Shen, J. Siegfried, R. Sinha, A. Sivachenko, C. Sougnez, D. Stoll, J. Stuart, R. K. Thomas, S. Tomaszek, M.-S. Tsao, W. D. Travis, C. Vaske, J. N. Weinstein, D. Weisenberger, D. A. Wigle, M. D. Wilkerson, C. Wilks, P. Yang, J. John Zhang, M. A. Jensen, R. Sfeir, A. B. Kahn, A. L. Chu, P. Kothiyal, Z. Wang, E. E. Snyder, J. Pontius, T. D. Pihl, B. Ayala, M. Backus, J. Walton, J. Baboud, D. L. Berton, M. C. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. A. Kigonya, S. Alonso, R. N. Sanbhadti, S. P. Barletta, J. M. Greene, D. A. Pot, M.-S. Tsao, B. Bandarchi-Chamkhaleh, J. Boyd, J. E. Weaver, D. A. Wigle, I. A. Azodo, S. C. Tomaszek, M. Christine Aubry, C. M. Ida, P. Yang, F. Kosari, M. V. Brock, K. Rogers, M. Rutledge, T. Brown, B. Lee, J. Shin, D. Trusty, R. Dhir, J. M. Siegfried, O. Potapova, K. V. Fedosenko, E. Nemirovich-Danchenko, V. Rusch, M. Zakowski, M. V. Iacocca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, Z. Fan, N. Todaro, J. Eckman, J. Myers, W. Kimryn Rathmell, L. B. Thorne, M. Huang, L. Boice, A. Hill, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, C. Morrison, C. Gaudioso, J. M. S. Bartlett, S. Kodeeswaran, B. Zanke, H. Sekhon, K. David, H. Juhl, X. Van Le, B. Kohl, R. Thorp, N. Viet Tien, N. Van Bang, H. Sussman, B. Duc Phu, R. Hajek, N. Phi Hung, K. Z. Khan, T. Muley, K. R. Mills Shaw, M. Sheth, L. Yang, K. Buetow, T. Davidsen, J. A. Demchok, G. Eley, M. Ferguson, L. A. L. Dillon, C. Schaefer, M. S. Guyer, B. A. Ozenberger, J. D. Palchik, J. Peterson, H. J. Sofia, E. Thomson, P. S. Hammerman, D. Neil Hayes, M. D. Wilkerson, N. Schultz, R. Bose, A. Chu, E. A. Collisson, L. Cope, C. J. Creighton, G. Getz, J. G. Herman, B. E. Johnson, R. Kucherlapati, M. Ladanyi, C. A. Maher, G. Robertson, C. Sander, R. Shen, R. Sinha, A. Sivachenko, R. K. Thomas, W. D. Travis, M.-S. Tsao, J. N. Weinstein, D. A. Wigle, S. B. Baylin, R. Govindan, M. Meyerson; Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012). Medline doi:10.1038/nature11404

53. T. Davoli, A. W. Xu, K. E. Mengwasser, L. M. Sack, J. C. Yoon, P. J. Park, S. J. Elledge, Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013). Medline doi:10.1016/j.cell.2013.10.011

54. G. Nilsen, K. Liestøl, P. Van Loo, H. K. Moen Vollan, M. B. Eide, O. M. Rueda, S. F. Chin, R. Russell, L. O. Baumbusch, C. Caldas, A. L. Børresen-Dale, O. C. Lingjaerde, Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012). Medline doi:10.1186/1471-2164-13-591

55. J. Wang, C. G. Mullighan, J. Easton, S. Roberts, S. L. Heatley, J. Ma, M. C. Rusch, K. Chen, C. C. Harris, L. Ding, L. Holmfeldt, D. Payne-Turner, X. Fan, L. Wei, D. Zhao, J. C.

Obenauer, C. Naeve, E. R. Mardis, R. K. Wilson, J. R. Downing, J. Zhang, CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011). Medline doi:10.1038/nmeth.1628

56. K. Chen, L. Chen, X. Fan, J. Wallis, L. Ding, G. Weinstock, TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.* **24**, 310–317 (2014). Medline doi:10.1101/gr.162883.113

57. W. J. Kent, BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). Medline doi:10.1101/gr.229202. Article published online before March 2002

58. L. Yang, L. J. Luquette, N. Gehlenborg, R. Xi, P. S. Haseley, C. H. Hsieh, C. Zhang, X. Ren, A. Protopopov, L. Chin, R. Kucherlapati, C. Lee, P. J. Park, Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013). Medline doi:10.1016/j.cell.2013.04.010

59. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). Medline doi:10.1101/gr.092759.109

60. P. J. Stephens, P. S. Tarpey, H. Davies, P. Van Loo, C. Greenman, D. C. Wedge, S. Nik-Zainal, S. Martin, I. Varela, G. R. Bignell, L. R. Yates, E. Papaemmanuil, D. Beare, A. Butler, A. Cheverton, J. Gamble, J. Hinton, M. Jia, A. Jayakumar, D. Jones, C. Latimer, K. W. Lau, S. McLaren, D. J. McBride, A. Menzies, L. Mudie, K. Raine, R. Rad, M. S. Chapman, J. Teague, D. Easton, A. Langerød, M. T. Lee, C. Y. Shen, B. T. Tee, B. W. Huimin, A. Broeks, A. C. Vargas, G. Turashvili, J. Martens, A. Fatima, P. Miron, S. F. Chin, G. Thomas, S. Boyault, O. Mariani, S. R. Lakhani, M. van de Vijver, L. van 't Veer, J. Foekens, C. Desmedt, C. Sotiriou, A. Tutt, C. Caldas, J. S. Reis-Filho, S. A. Aparicio, A. V. Salomon, A. L. Børresen-Dale, A. L. Richardson, P. J. Campbell, P. A. Futreal, M. R. Stratton; Oslo Breast Cancer Consortium (OSBREAC), The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012). Medline

61. H. Bengtsson, P. Wirapati, T. P. Speed, A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* **25**, 2149–2156 (2009). Medline doi:10.1093/bioinformatics/btp371

62. H. Bengtsson, P. Neuvial, T. P. Speed, TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics* **11**, 245 (2010). Medline doi:10.1186/1471-2105-11-245

63. M. Ortiz-Estevez, A. Aramburu, H. Bengtsson, P. Neuvial, A. Rubio, CalMaTe: A method and software to improve allele-specific copy number of SNP arrays for downstream segmentation. *Bioinformatics* **28**, 1793–1794 (2012). Medline doi:10.1093/bioinformatics/bts248

64. C. Yau, D. Mouradov, R. N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O. Sieber, C. C. Holmes, A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.* **11**, R92 (2010). Medline