Supplementary Information

Integrative genome analyses identify key somatic driver mutations of small cell lung cancer

Martin Peifer*, Lynnette Fernández-Cuesta*, Martin L Sos, Julie George, Danila Seidel, Lawryn H Kasper, Dennis Plenker, Frauke Leenders, Ruping Sun, Thomas Zander, Roopika Menon, Mirjam Koker, Ilona Dahmen, Christian Müller, Vincenzo Di Cerbo, Hans-Ulrich Schildhaus, Janine Altmüller, Ingelore Baessmann, Christian Becker, Bram de Wilde, Jo Vandesompele, Diana Böhm, Sascha Ansén, Franziska Gabler, Ines Wilkening, Stefanie Heynck, Johannes M Heuckmann, Xin Lu, Scott L Carter, Kristian Cibulskis, Shantanu Banerji, Gad Getz, Kwon-Sik Park, Daniel Rauh, Christian Grütter, Matthias Fischer, Laura Pasqualucci, Gavin Wright, Zoe Wainer, Prudence Russell, Iver Petersen, Yuan Chen, Erich Stoelben, Corinna Ludwig, Philipp Schnabel, Hans Hoffmann, Thomas Muley, Michael Brockmann, Walburga Engel-Riedel, Lucia A Muscarella, Vito M Fazio, Harry Groen, Wim Timens, Hannie Sietsma, Erik Thunnissen, Egbert Smit, Daniëlle AM Heideman, Peter JF Snijders, Federico Cappuzzo, Claudia Ligorio, Stefania Damiani, John Field, Steinar Solberg, Odd Terje Brustugun, Marius Lund-Iversen, Jörg Sänger, Joachim H Clement, Alex Soltermann, Holger Moch, Walter Weder, Benjamin Solomon, Jean-Charles Soria, Pierre Validire, Benjamin Besse, Elisabeth Brambilla, Christian Brambilla, Sylvie Lantuejoul, Philippe Lorimier, Peter M Schneider, Michael Hallek, William Pao, Matthew Meyerson, Julien Sage, Jay Shendure, Robert Schneider, Reinhard Büttner, Jürgen Wolf, Peter Nürnberg, Sven Perner, Lukas C Heukamp, Paul K Brindle, Stefan Haas, and Roman K Thomas

*These authors contributed equally to this work. Correspondence to: Roman Thomas, roman.thomas@uni-koeln.de, Department of Translational Genomics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany, Tel.: +49-221-478-98771

Nature Genetics: doi:10.1038/ng.2396

Supplementary Note

Clinical Samples and Cell Lines

We collected a total of 99 fresh-frozen tumor samples from which 63 were qualified for copy number analysis by Affymetrix SNP 6.0 arrays. Due to the availability of the matched-normal, tumor content, and amount of DNA we sequenced 27 tumor/normal pairs. RNAseq was performed on 15 samples and 2 samples were whole genome sequenced. In addition to patient specimens, 65 cell lines were considered for copy number analysis and we sequenced the exome of two cell lines (where the matched normal was available).
By analyzing the genotype identity using SNP calls derived from the SNP 6.0 arrays we had to eliminate 5 cell lines from the subsequent analysis that show identical genotypes. In this analysis we used Birdsuite[1] to call the genotypes. Among the 65 SNP arrays, raw data of 38 arrays including batch-matched normals were downloaded from the Sanger Institute (http://www.sanger.ac.uk/genetics/CGP/Archive/). SNP arrays of the remaining 27 cell lines were generated using our SNP array pipeline. Note that the entire dataset of the 27 previously uncharacterized SCLC cell lines is presented elsewhere since we only show copy number data for the genes: *CREBBP*, *EP300*, and *SLIT2*. Information about the patient specimens and cell lines is presented in **Supplementary Table 1**.

Pathological Review

All SCLC samples showed a typical SCLC morphology including fragile nuclei with salt and pepper like chromatin, nuclear molding and scant cytoplasm. Further confirmatory immunohistochemistries (IHCs) in addition to the original diagnostic report were performed of the exome/genome sequenced samples if sufficient material was available. All samples showed some nuclear TTF1 staining and were positive for at least one neuroendocrine marker out of synaptophysin, chromogranin A or CD56. In addition p63 and CK5/6 staining was performed to exclude squamous cell differentiation. CK7 staining was either negative or showed a typical perinuclear condensed dot like staining pattern. Thus, all reassessed cases (22 of 27) were confirmed to be SCLC **(Supplementary Table 8)**. SCLC was confirmed by IHC of two additional samples by the provider only and the remaining three samples were *RB1* mutated, making a misclassification of these samples unlikely **(Supplementary Table 8)**.

Nature Genetics: doi:10.1038/ng.2396

Detection of Somatic Mutations

Our approach to identify somatic mutations for whole exome/genome sequencing consists of five major steps:

1) Estimation of Local Copy Number, Purity, Overall Ploidy, Tumor Heterogeneity, and Background Sequencing Error: To determine the local copy number we refined a previously described method[2] in order to correct for outliers and the uneven distribution of coverage across the genome that arises by exon-capture. These efforts lead to substantially fewer fluctuations in the copy number profile. Genome wide copy numbers across all 29 samples are shown in **Supplementary Figure 3a.**

The estimation of purity, ploidy, genome-wide allelic states, tumor heterogeneity, and the sequencing error relies on the analysis of all known SNPs (dbSNP-sites). In detail, we screened for SNPs that are heterozygous in the matched normal and computed a rescaled allelic fraction of these SNPs in the tumor, called theta-value in the following. The theta-value represents the genotypic state of the SNP in the tumor and has the following interpretation: values close to zero indicating heterozygous SNPs, whereas values close to the tumor purity showing that the SNP has lost its heterozygosity. States in between can, together with the copy numbers, discriminate the allelic architecture of the tumor. We next average the theta-value across copy number segments to remove sequencing noise and formulated a mathematical model that relates the observed theta-value and copy number to the underlying allelic state, absolute copy, and the overall tumor purity. Estimates of the absolute copy numbers of all 29 cases are shown in **Supplementary Figure 4b** and purity estimates are given in **Supplementary Table 5**. The genome-wide ploidy is computed from the absolute copy numbers by a weighted average. In total we found 5 triploid and 2 near tetraploid cases **(Supplementary Table 5)**. Fractions of the tumor that show theta-values lying between model predictions can be considered as heterogeneous. The genome-wide fraction of these regions serves as an estimate for the total tumor heterogeneity **(Supplementary Table 5)**. In **Supplementary Figure 9** we show observed and predicted theta-values together with its reconstructed allelic states for two representative cases (one diploid and one triploid).

To determine the overall sequencing error we screened for SNPs that show a variant, which are not in accordance with the two possibilities of the respective germline variant. The total amount of these aberrant signals is related to the total coverage of all SNPs. This yields an estimate of the sequencing errors at a global level.

Nature Genetics: doi:10.1038/ng.2396

2) Mutation Calling and Tumor-Normal Comparison: Since sequencing artifacts for single base substitutions are higher than for insertions and deletions (indels) different detection strategies are used for these two types of mutations. The metrics that have been determined in the first step are only used for substitution calling. Furthermore, we restricted mutation calling only to those portions of the genome with sufficient coverage and discard regions smaller than 15x coverage from the analysis. In case of single nucleotide substitutions, we compute the minimal possible allelic fraction at each sufficiently covered location from the local copy number, purity, and total ploidy. Then we test whether the observed allelic fraction of a putative variant exceeds this quantity. Here, single base quality scores and mapping quality are considered in the allelic fraction by computing a combined phred-score. Those bases that shows a combined quality score < 10 are discarded from the analysis. If a substitution is called in the tumor, we check if the variant is somatic by transforming the allelic fraction of the variant in the normal into a z-score by using the global sequencing error and the coverage of the corresponding genomic position in the normal. If the z-score is smaller than 20, the variant is considered as mutation candidate.

Since on one hand, sequencing errors of indels occur at a lower rate than for single base nucleotides but, on the other hand, their identification by the short-reads is more difficult, we chose the following detection strategy: indels that are present in more than five reads, have an allelic fraction above 5% in the tumor, and are completely absent in the normal are considered as candidate. Given the allelic fraction of the indel, we then test if the absence of the indel in the normal is compatible with chance. If this is not the case, the indel is called as mutation candidate.

3) Filtering Mutation Candidates: Substitution candidates are filtered by testing if the allelic fraction of the variant is significantly different from the maximum between the sequencing error and the observed allelic fraction in the normal. This removes spurious calls at locations of low minimal allelic fraction (e.g., in amplified regions). As second filter, we examine mutation calls that are showing an extreme forward-backward bias (i.e., if the variant is only present in the forward-reads and absent in the backward read, or vice versa). Since the exon-capture enriches for a specific read-direction at the end of the exon, we cannot simply discard variants showing such an extreme forward-backward bias without risking of dropping a significant number of true mutations. On the other hand, a large number of false positives show an extreme forward-backward bias. To take this into account, we examine if an extreme forward-backward bias is also present in the wild-type allele. If this is the case, we keep the mutation as candidate. Finally, a combined mutation score is computed from the underlying statistical model and form filter characteristics.

Nature Genetics: doi:10.1038/ng.2396

4) *Down-Sample Approach to Overcome Model Misspecifications:* Since regions of high coverage are more sensitive to violations of model assumptions we apply a down-sample approach to gain more sensitivity. Model misspecifications can, e.g., arise by inaccuracies on the quantities of our statistical model or by tumor heterogeneity. At each position, where no substitution was detected we synthetically down-sample the coverage in the tumor to 15x and repeating the calling and filtering procedure. A mutation that is then detected is reported with the lowest possible mutation score of zero.

5) Accounting for Tumor Heterogeneity and further Model Inaccuracies: The down-sample approach can further be used to assess possible tumor heterogeneity. To this end, the computed minimal allelic fraction is multiplied by a constant factor to lower the threshold globally. This factor is determined by the following procedure: we vary the factor form 1 to 0.7 (thus corresponding to 0% – 30% tumor heterogeneity) and count the number of mutations that show a mutation score of zero. We then chose the factor corresponding to the first local minimum of the mutation score count.

Annotation of Somatic Mutations

Annotation of the detected somatic mutations is based on the Consensus CDS database (CCDS) in order to obtain a consistent annotation of protein coding regions. We report only those mutations that are within the reading frame of a protein coding sequence plus its splice-site. To assess a potential functional impact of a mutation we include PFAM-domains[3] into the annotation. PFAM-domains are recomputed from our gene model in order to eliminate misclassifications due to different splice variants. Each mutation is annotated regarding the following parameters:

1) Transcript accession number
2) Position in the protein sequence
3) Mutation type: missense, nonsense, silent, frame-shift indel, in-frame indel, splice-site
4) Predicted amino acid change
5) If a variant is at a dbSNP site
6) PFAM-domain

Detected somatic mutations including their validation status are listed in **Supplementary Table 9**.

Nature Genetics: doi:10.1038/ng.2396

Validation of Somatic Mutations and Frequently Mutated Genes

Validation of Mutations Detected in the Discovery Screen

Since the high mutation rate yielded more than 8,000 candidates for somatic mutations across the 29 samples we only validated an assorted set of genes by conventional dideoxy sequencing. Among these mutations in all identified driver genes driver are validated: *TP53*, *RB1*, *PTEN*, *CREBBP*, *EP300*, *SLIT2*, *MLL, COBL*, and *EPHA7*. To assess the false-discovery rate of our mutation-calling algorithm in an unbiased fashion we validated a total of 343 unselected mutations in 26 samples of a pediatric tumor and achieved a false-discovery rate of 9% (data will be published elsewhere). All of these samples of a pediatric tumor were processed by the same sequencing workflow and infrastructure.

By the independence of the sequencing runs, mutations that are detected by exome sequencing and by whole genome sequencing are considered as being validated. In addition, two samples (S00050, S00356) were also run on exon sequencing at the Broad Institute. Due to a lack of a sufficient quantity of tumor DNA, whole genome amplification (WGA) was performed on those samples. This, however, should only increase the false-positive rate due to WGA artifacts. Concordant mutation calls can therefore still be considered as being validated. All validated mutations including their validation status and validation method are given in **Supplementary Table 9**. The number of overlapping and discordant mutations between all independent sequencing runs are shown for all four samples in **Supplementary Figure 11**. Note that due to the variability of the library, different sequencing coverage, different sources of low-level contaminations, and false-positive mutation calls, the overlap between two independent runs should not be considered as sensitivity of the mutation caller.

Extended Mutation Screen

In order to assess the mutation frequency of *CREBBP*, *EP300*, and *SLIT2* we extended our sequencing efforts to an independent validation set.

1) *CREBBP*/*EP300*: We sequenced the region around the histone acetylation domain (HAT-domain, exon 18-30) on 26 patient specimens and 45 cell lines with dideoxy sequencing. The observed clustering of the mutations in exon sequencing motivates the restriction to the location around the HAT-domain. All detected mutations are checked if they are absent in the matched normal in case of patient specimens. Since there is no matched normal available for almost all cell lines we only report variants that are not in the SNP-database in those cases. All discovered and validated mutations in

Nature Genetics: doi:10.1038/ng.2396

CREBBP/*EP300* of the extended mutation screen are given in **Supplementary Table 10**.

2) *SLIT2*: We sequenced the full-length gene using a quantitative PCR based capture technique followed by 454 sequencing on the Roche GS FLX Titanium. 26 patient specimens and 34 cell lines were sequenced.

Primers were designed using our in-house developed primer design pipeline PrimerXL (Levefer et al. manuscript in preparation) using tiling settings, taking into account SNP positions. We aim for a target annealing temperature of 60°C. Primers where controlled for quality by performing a qPCR reaction on human genomic DNA (Roche) using the Kapa 2G Robust mastermix (Kapa biosystems). All qPCR reactions were performed according to manufacturers' protocols on a CFX384 qPCR instrument (Bio-Rad) in a 5μl volume containing 0.25μM of primer and 10ng of input DNA. PCR assays successfully amplified the targeted sequence with a Cq of 24.1±0.7 on average. Primers were ordered in 23 fold, each set with a different MID tag attached to the 5' end. Then, we carried out target capture by performing a qPCR on tumor DNA. Amplicons were pooled for samples with different MIDs by inverted spinning of the PCR plate in 3 pools (1 with 23 samples, 2 with 22 samples). 500μl of pooled PCR product was purified using the Qiaquick PCR purification kit (Qiagen) and DNA quality of the pools was assured by capillary electrophoresis on a Bioanalyzer 2100 machine using the DNA 7500 kit (Agilent). Next, DNA was purified using the Roche High purification PCR cleanup microkit (Roche) and for library production we used the NEBNext DNA sample prep mastermix Set 2 (New England Biolabs). Emulsion PCR and sequencing was carried out according to manufacturers' protocol. Each pool of samples was sequenced on 1/8 of a 454 sequencing flow cell, generating a mean number of 4370±1646 reads per sample with a maximum of 8353 and a minimum of 2208 reads.

We mapped the 454 reads to the human reference genome (NCBI build 37/hg19) using BWA[4] in BWASW[5] mode. Recalibration of the read quality scores was done using the Genome analysis toolkit (GATK)[6] and duplicate reads were removed using Picard tools (http://picard.sourceforge.net). An average coverage of 47.2±18.5 fold was achieved across the genomic targets in all samples, which is sufficiently high for reliable variant calls. A minimal coverage of 10 fold for 80% of the target locations was achieved in 94% of the samples **(Supplementary Table 11).** Positions of the variants were extracted on all samples simultaneously from the mapped BAM files using Samtools[7] mpileup with default settings. All variants were then imported into the NXTVAT web tool (De Wilde, manuscript in preparation) to provide mutation filtering. We validated all detected variants that were not present in the SNP database by dideoxy sequencing (in case of patient specimens the matched-normal was additionally sequenced) and mapped validated mutations to NCBI build 36/hg18 using Galaxy.[8] Annotations of the

Nature Genetics: doi:10.1038/ng.2396

detected mutations were done by same module we used in exon sequencing. All validated variants of *SLIT2* are shown in **Supplementary Table 12**.

Analysis of Significantly Mutated Genes

To assess the significance of recurrently mutated genes we adapt a previously described procedure[9] to account for requirements arising from a high mutational background. The method we propose consists of the following three major steps:

1. Estimation of the expected background mutation rate for each gene
2. Assessing the set of sufficiently expressed genes by RNAseq
3. Q-value computation and correction for accumulation of synonymous mutations

1) Estimation of the Expected Background Mutation Rate for Each Gene

Similar to Ding et al.[9] we first determine background mutation rates from synonymous mutations since these mutations are assumed to represent the neutral mutation rate. The mutation rates are then lifted to the background mutation rate of non-synonymous mutations by using the gene-specific codon usage. To this end, the global mutation rate r_i for each sample *i* is determined, the 6 neighbor independent mutation rates (assuming strand symmetry): $q_{A:T>C:G}$, $q_{A:T>T:A}$, $q_{C:G>G:C}$, $q_{C:G>A:T}$, $q_{A:T>G:C}$, $q_{G:C>A:T}$, and the neighbor dependent rate $q_{CpG>TpG:CpA}$ accounting for the elevated mutation rate of 5-methylcytosine by spontaneous hydrolytic deamination. Then for each gene *j* the non-synonymous to synonymous ratio R_j is computed form the previously determined rates and the gene-specific codon usage. Due to genes that are negatively selected, the theoretical value of R_j turns out of being slightly too small such that we correct the genome wide average of R_j by the observed non-synonymous to synonymous ratio over all samples and genes: leading to a global correction factor x. Together with the total amount of sequenced bases for each sample *i* and gene *j*: n_{ij}, the gene-specific expected background mutation rate l_j is given by

$$\lambda_j = \xi R_j \sum_i \rho_i n_{ij}.$$

2) Assessing the Set of Sufficiently Expressed Genes by RNAseq

Robust identification of genes that are frequently mutated across the tumor samples is massively hampered by the high abundance of passenger mutations (i.e., random mutations that are not contributing to tumor development or progression) as a result of

Nature Genetics: doi:10.1038/ng.2396

the high mutation rate. Due to the absence of negative selection pressure, genes that are not expressed in tumor cells accumulate more passenger events than genes that encode proteins of important cellular functions. To avoid that such genes are identified as being significantly mutated, we consider only those genes in the following that show in more than half of the RNAseq analyzed samples an expression larger than 1 FPKM (fragments per kilobase of exon per million fragments mapped). In case of different splice variants the maximal gene expression was chosen. This gene-based method has the advantage that the samples that were RNA sequenced do not necessarily have to match the exome/genome-sequenced samples. Expression values of all samples are shown in **Supplementary Table 13.**

3) Q-value Computation and Correction for Accumulation of Synonymous Mutations

In addition to the absence of gene expression, genes that show an extraordinary high number of silent mutations are rather passenger events. To incorporate the accumulation of silent mutations into the significance score, we first compute the probability that the observed number of samples that show only silent mutations ns_j for gene j is larger than expected. Given the parameters determined in 1) this probability is given by

$$p_j^s = pois\left(ns_j, \frac{\lambda_j}{\xi R_j}\right),$$

where *pois* is the cumulative Poisson distribution. If ns_j is larger than zero, the product between p^s_j and ns_j then defines an estimate for the expected gene-specific number of synonymous mutations. The correction factor c_j defined by ratio between the gene-specific expected number of silent mutations and $\lambda_j/\xi R_j$:

$$c_j = max\left(1, ns_j p_j^s \frac{\xi R_j}{\lambda_j}\right).$$

In case of a local mutation rate that is higher than the genome-wide level and which is reflected by an accumulation of synonymous mutations, the correction factor c_j is larger than one. Given the number of mutated samples n_j, the p-value for gene j is finally computed by

$$p_j = 1 - pois\left(n_j, \lambda_j c_j\right).$$

To account for multiple hypothesis testing, the false discovery rate (q-values) is determined by the Benjamini-Hochberg method.[10] Significantly mutated genes that have a q-value ≤ 0.1 including different modes of importance filtering (without filtering,

Nature Genetics: doi:10.1038/ng.2396

correcting for the accumulation of synonymous mutations, and filtering for gene expression) are shown in **Supplementary Table 4**.

Analysis of RNAseq Data

RNAseq is performed on cDNA libraries prepared from PolyA+ RNA extracted from tumor cells. We aim for a library with an insert size of 250bp that allows us to sequence 95bp paired-end reads without overlap. RNAseq affords robust detection of expressed fusion transcripts in addition to enabling integrated analyses of orthogonal genomics datasets by providing gene expression data.

Detection of Chimeric Transcripts

For the analysis of RNAseq data, we have developed a pipeline that affords accurate and efficient mapping and downstream analysis of transcribed genes in cancer samples (manuscript in preparation). Briefly, paired-end RNAseq reads are aligned against the human reference genome (hg18) using spliced mappers such as TopHat[11] or GSNAP.[12] Unique paired-end alignments that are within the expected mapping distance are used to estimate the transcriptional abundance of annotated genes or exons and are used to reconstruct alternatively spliced isoforms of known genes using Cufflinks.[13] By contrast, uniquely aligning read pairs that are not in accordance with the expected mapping distance in combination with singleton reads (i.e., only one end can be mapped) are selected for a de-novo assembly using Velvet[14] and Oases (a transcriptome assembler by Daniel Zerbino and Marcel Schulz, unpublished). The aim for this procedure is to accurately reconstruct rearranged transcripts. By comparing the assembled transcripts with the Refseq-database and with the reference genome, we query for those candidates that show a partial alignment onto two different genes. These alignments are thereby representing a putative chimeric transcript. For each candidate, we detect fusion-point spanning reads from the initially unmapped read pairs to localize the breakpoint within the transcript. To allow confident predictions of chimeric transcripts, we subsequently filter candidate chimeras by their read distribution around the potential fusion point. We finally consider fusion candidates for experimental validation where at least one read-pair is uniquely mapped to the human genome (to the two different genes), at least one 95bp read unambiguously spanned a junction between two exons of the two genes, and the coverage is at least 5x.

Validation of Chimeric Transcripts

Total RNA was extracted from frozen tumor sections containing more than 90% tumor cells, using Qiagen RNeasy Mini Kit (Qiagen), and 1µg RNA was reverse transcribed

Nature Genetics: doi:10.1038/ng.2396

using the SuperScript III Reverse Transcriptase kit (Invitrogen), according to the manufacturer's instructions. cDNA was treated with RNA-H (Invitogen) and cleaned by using the Nucleo Spin Extract II Kit (Macherey-Nagel) according to the manufacturer's instructions. Candidate-specific fusion-point encompassing primers were designed in order to amplify by RT-PCR the region over the fusion-point. All validated chimeric transcripts are shown in **Supplementary Table 3** and **Supplementary Figures 3, 12**.

Mutation Calling in RNAseq Data

We implemented a mutation caller that allows us to identify missense, nonsense, and nonstop mutations in RNAseq data. Due to limitations of transcriptome sequencing and the spliced aligner we cannot detect splice-site mutations and indels. Since a detailed model of the minimal allelic fraction, as for genomic sequencing, is not available, we applied a different strategy to call mutations. We first estimated the sequencing error as described above (see page 7) and then, similar to the tumor normal comparison (see page 8), we transformed the observed allelic fractions into a z-value and called a mutation if the z-value exceeds 20. To obtain reliable mutation calls, we analyzed only portions of the transcriptome that are least 10x covered. Some highly expressed genes are showing a coverage of several 1000x, to prevent an explosion of false-positives in those regions (e.g., driven by low-level contaminations of the library or sequencing artifacts) we synthetically limited the coverage to 100x but keeping the allelic fractions as they were observed. In those 12 samples that were exome and transcriptome sequenced we called only 29% (650/2246) of the missense, nonsense, and nonstop mutations detected by exome sequencing. Matching mutations between transcriptome and genomic sequencing are highlighted in **Supplementary Table 9.**

CREBBP and *EP300* FISH Break Apart Assay

Metaphase spreads were prepared by treating cells with colcemid (Roche, Switzerland) at a concentration of 0.1 μg/ml. The culture flasks were incubated at 37°C for 1 hour. Mitotic arrest was followed by treatment with 0.075M KCl. The cells were then repeatedly washed with modified Carnoy's fixative (3:1 methanol to acetic acid and then dropped onto pre-cleaned slides and observed for metaphase spreads. Metaphase spreads were pre-treated with 2x SSC solution at 37C° for 60 min. and digested with Digest-All III (dilution 1:2) at 37°C for 6 min.; FISH probes were denatured at 73°C for 5 min. and immediately placed on ice. Subsequently, the metaphase spreads and FISH probes were co-denatured at 85°C for 4 min. and hybridized overnight at 37°C. Post hybridization washing was done with 0.5x SSC at 75°C for 5 min., and the fluorescence detection was carried out using streptavidin-Alexa-594 conjugates (dilution 1:200) and anti-digoxigenin-FITC (dilution 1:200). Slides were then counterstained with 4',6-Diamidin-2' phenylindoldihydrochlorid (DAPI) and mounted. The samples were

Nature Genetics: doi:10.1038/ng.2396

analysed under a 63x oil immersion objective using a fluorescence microscope (Zeiss, Jena, Germany) equipped with appropriate filters, a charge-coupled device camera and the FISH imaging and capturing software Metafer 4 (Metasystems, Altlussheim, Germany). Assessment of the experiments was done independently by two evaluators (R.M and S.P).

Rearrangement of either of the genes was defined a-priori as following: A split of a signal pair, resulting in a single red and single green signal for at least one allele is referred to as a translocation. A loss of a signal, resulting in either a single red or single green signal for at least one allele is referred to as a rearrangement through deletion. A wild-type allele displays a juxtaposed red and green signal (mostly forming a yellow signal).

Analysis of Significantly Amplified or Deleted Regions

Many algorithms to detect significantly amplified or deleted regions in cancer samples have been proposed recently, e.g., GISTIC,[15,16] JISTIC,[17] and RAE.[18] In case of GISTIC and JISTIC a global threshold is used to discriminate between amplifications and deletions which ignores the variability in tumor purity. RAE takes this into account by using a variable threshold. We implemented a novel algorithm that is entirely distribution driven by transforming raw copy numbers (un-segmented) for each sample to ranks across genomic positions (manuscript in preparation). Since the rank transformation is invariant to monotonous transformations the method automatically corrects for differences in tumor purity, array saturation effects, and differences in the baseline level. In brief, the user provides upper and lower quantiles to distinguish between deletions and amplifications. These quantiles control the focality of the identified peaks in a manner that a very narrow quantile adjustments lead to highly focal peaks. Ranks that are accounting for amplifications are independently processed from deletions. Next, rank sums are computed for each genomic position and subsequently smoothed to remove the noise from the data. As smoothing algorithm we used a kind of segmentation procedure to facilitate peak finding by the resulting piecewise constant function. The additional advantage of rank sums is that the underlying statistical model for the null hypothesis can analytically be determined. Here, long-range correlations due to the local constant nature of copy numbers are taken into account by a linear propagation of the correlations into the final statistics. Adjustment of multiple hypothesis testing on the identified segments is carried out by the Benjamini-Hochberg approach[10]. To filter out only reliable copy number alterations, we subsequently removed samples having the strongest impact on each identified peak region and repeated the analysis. Only those peaks are finally considered as driver copy number

Nature Genetics: doi:10.1038/ng.2396

alterations that are still significant after removal of the corresponding sample. All identified copy number alterations in SCLC are given in **Supplementary Table 2**.

References

1. Korn, J.M. *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature genetics* **40**, 1253-60 (2008).
2. Chiang, D.Y. *et al*. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods* **6**, 99-103 (2009).
3. Finn, R.D. *et al*. The Pfam protein families database. *Nucleic acids research* **36**, D281-8 (2008).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
5. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
6. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-303 (2010).
7. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
8. Giardine, B. *et al*. Galaxy: a platform for interactive large-scale genome analysis. *Genome research* **15**, 1451-5 (2005).
9. Ding, L. *et al*. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75 (2008).
10. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).
11. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
12. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-81 (2010).
13. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511-5 (2010).
14. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-9 (2008).
15. Beroukhim, R. *et al*. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007-12 (2007).
16. 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 biology* **12**, R41 (2011).
17. Sanchez-Garcia, F., Akavia, U.D., Mozes, E. & Pe'er, D. JISTIC: identification of significant targets in cancer. *BMC bioinformatics* **11**, 189 (2010).
18. Taylor, B.S. *et al*. Functional copy-number alterations in cancer. *PloS one* **3**, e3179 (2008).
19. Karro, J.E., Peifer, M., Hardison, R.C., Kollmann, M. & von Grunberg, H.H. Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. *Molecular biology and evolution* **25**, 362-74 (2008).
20. Hecht, S.S. Progress and challenges in selected areas of tobacco carcinogenesis. *Chemical research in toxicology* **21**, 160-71 (2008).

Nature Genetics: doi:10.1038/ng.2396

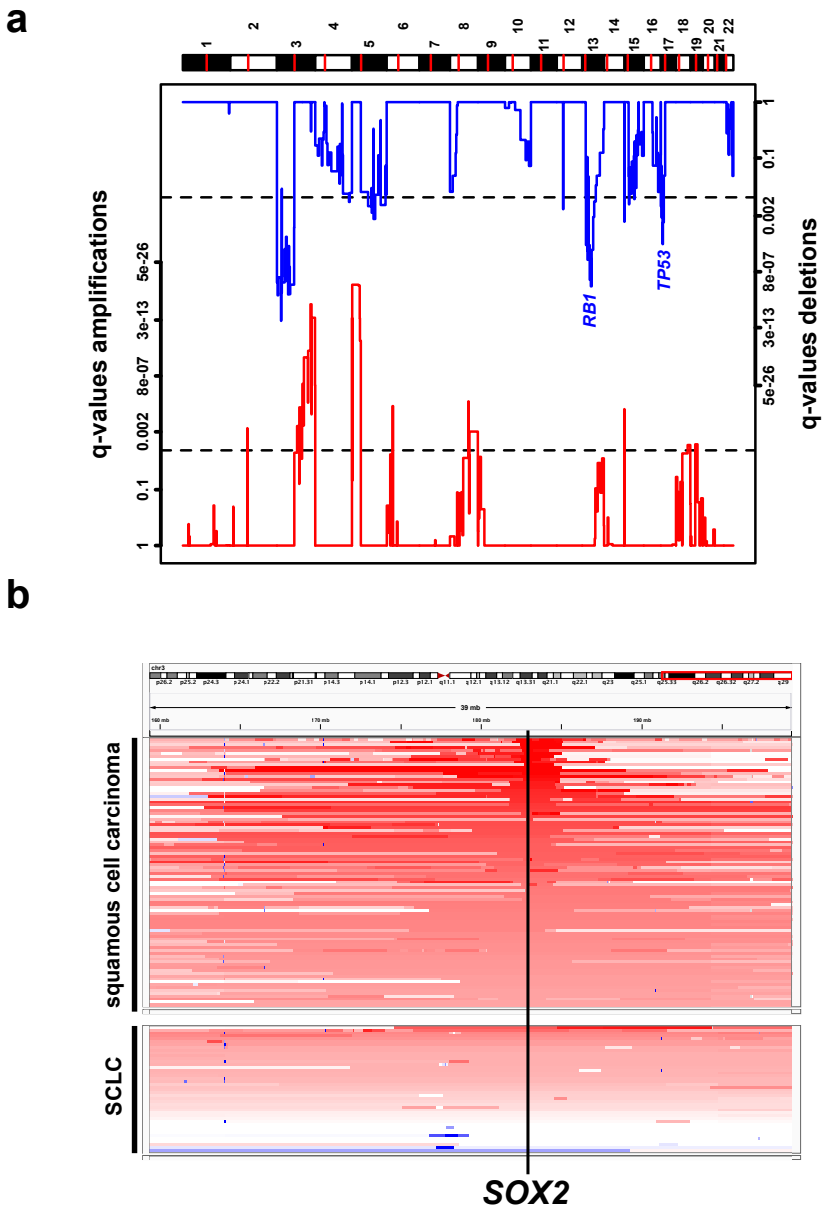
21. Pleasance, E.D. *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-90 (2010).
22. Rodin, S.N. & Rodin, A.S. Origins and selection of p53 mutations in lung carcinogenesis. *Seminars in cancer biology* **15**, 103-12 (2005).

Nature Genetics: doi:10.1038/ng.2396

List of Figures and Tables

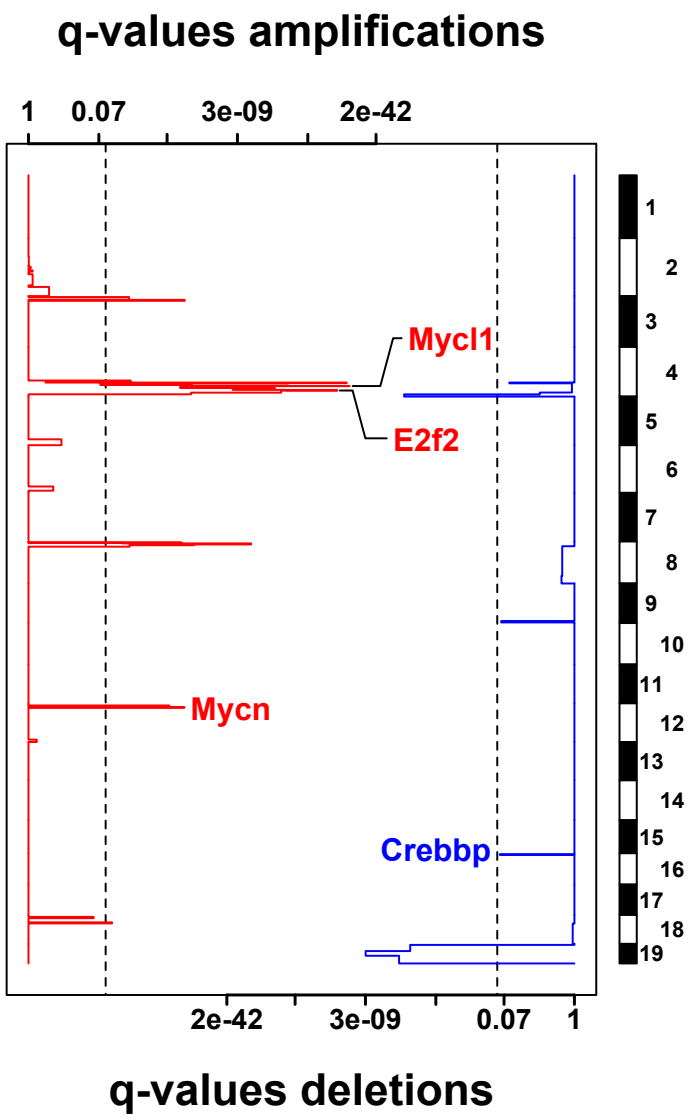
Nature Genetics: doi:10.1038/ng.2396

Supplementary Figures



Supplementary Figure 1. Analysis of broad copy number alterations. a) Copy number analysis of the 63 human SCLC samples using our rank-sum based algorithm. Here, thresholds are adjusted to extract broad copy number events (upper quantile: 40% ; lower quantile: 40%). b) A comparison of the region containing the lineage transcription factor *SOX2* between squamous cell carcinoma and SCLC.

Nature Genetics: doi:10.1038/ng.2396



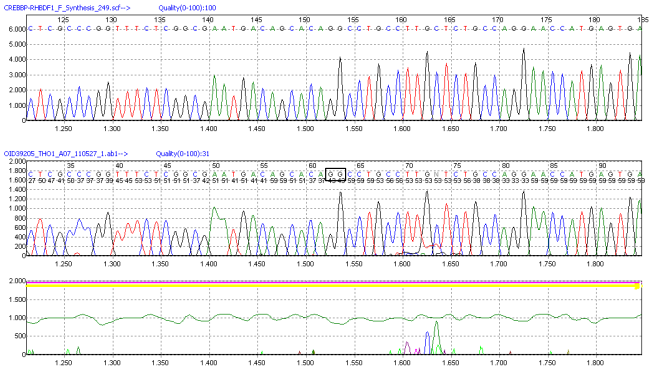
Supplementary Figure 2. Copy number analysis of 15 primary SCLC mouse tumors. Thresholds are chosen to identify focal events (upper quantile: 15%; lower quantile: 15%) and a significance level of 5% is used (vertical dashed lines).

Nature Genetics: doi:10.1038/ng.2396

a

CREBBP-RHBDF1

RNAseq_transcript:<CCCCCCCCGGCCGGGCCCTGGCCGGCCCGGCCGCCCGTGCCCGGGGCTGTTTTCGCGAGCAGGTGAAA
ATGGCTGAGAACTTGCTGGACGGACCGCCCAACCCCAAAAGAGCCAAACTCAGCTCGCCCGGTTTCTCGGCGAATGACAGCACA
GGCCTGCCTTGCTCTGCCAGGAACCATGAGTGAGGCCCGCAGGGACAGCACGAGCAGCCTGCAGCGCAAGAAGCCACCCTGG
CTAAAGCTGGACATTCCCTCTGCGGTGCCCCTGACGGCAGAAGAGCCCAGCTTCCTGCAGCCCCTGAGGCGACAGGCTTTCCTG>

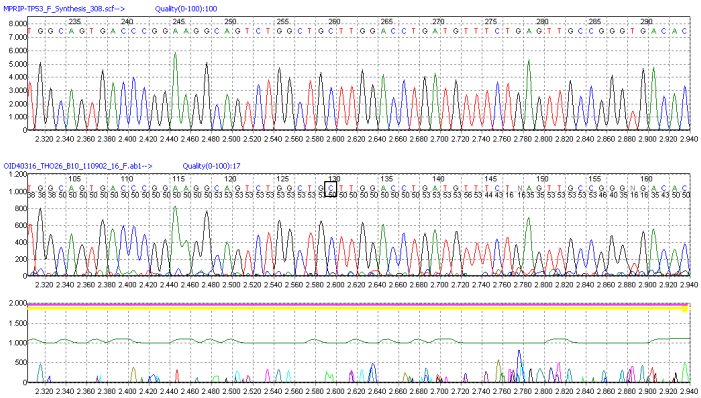


Translation<MAENLLDGPPNPKRAKLSSPGFSANDSTGLPCSARNHEStop
CREBBP ***RHBDF1***

b

MPRIP-TP53

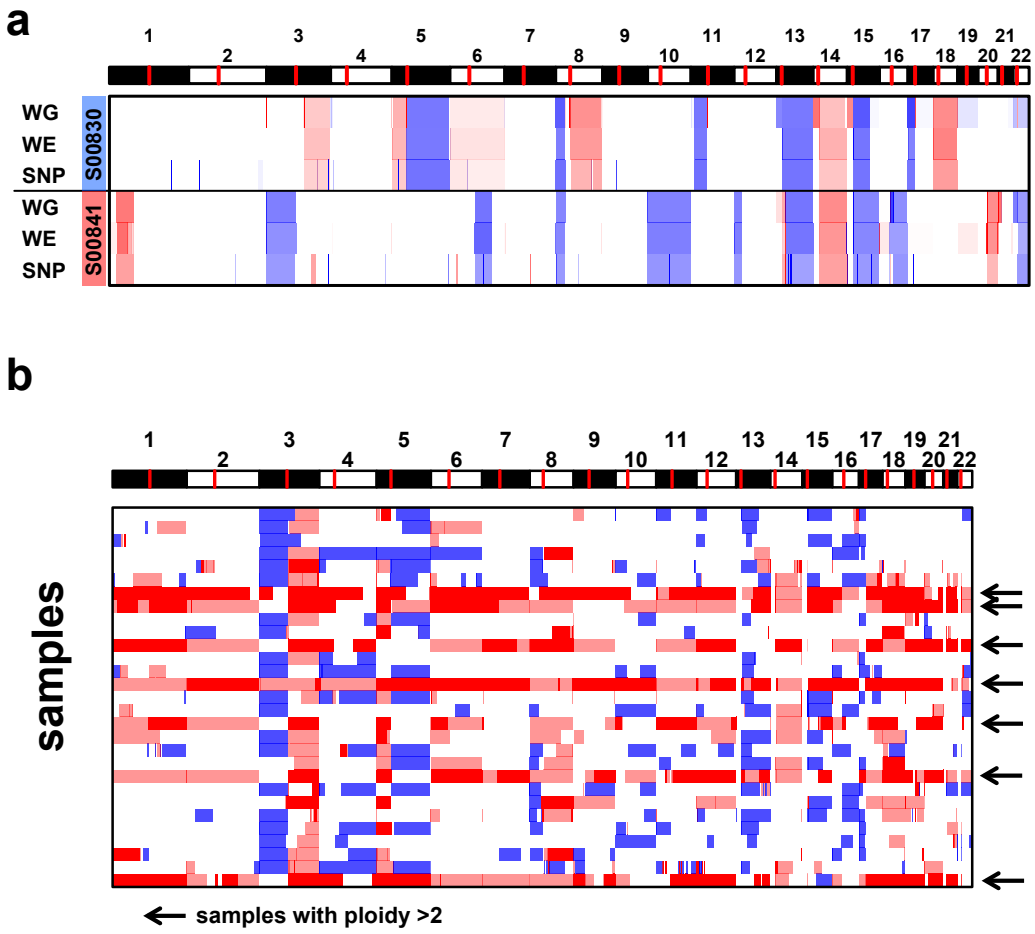
RNAseq_transcript:<TGGTGCAGGGGCCGCCGGTGTAGGAGCTGCTGGTGCAGGGGCCACGCGGGGAGCAGCCTCTGGCATTC
TGGGAGCTTCATCTGGACCTGGGTCTTCAGTGAACCATTGTTCAATATCGTCCGGGGACAGCATCAAATCATCCATTGCTTGGGAC
GGCAAGGGGGACAGAACGTTGTTTTCAGGAAGTAGTTTCCATAGGTCTGAAAATGTTTCCTGACTCAGAGGGGGCTCGACGCTAG
GATCTGACTGCGGCTCCTCCATGGCAGTGACCCGGAAGGCAGTCTGGCTG**C**TTGGACCTGATGTTTCTGAGTTGCCGGGTGACA
CAGGATCTGTCTTTCTTCAAGAAGTCAGGGTTGCTTTTAGATTTCATTATATCATATCCGGACACCGTGGCACTGTCAGGGGACTTC>



Translation<...KDRSCVTRQLRNIRSKQPDCLPGHCHGGAAVRSStop
MPRIP ***TP53***

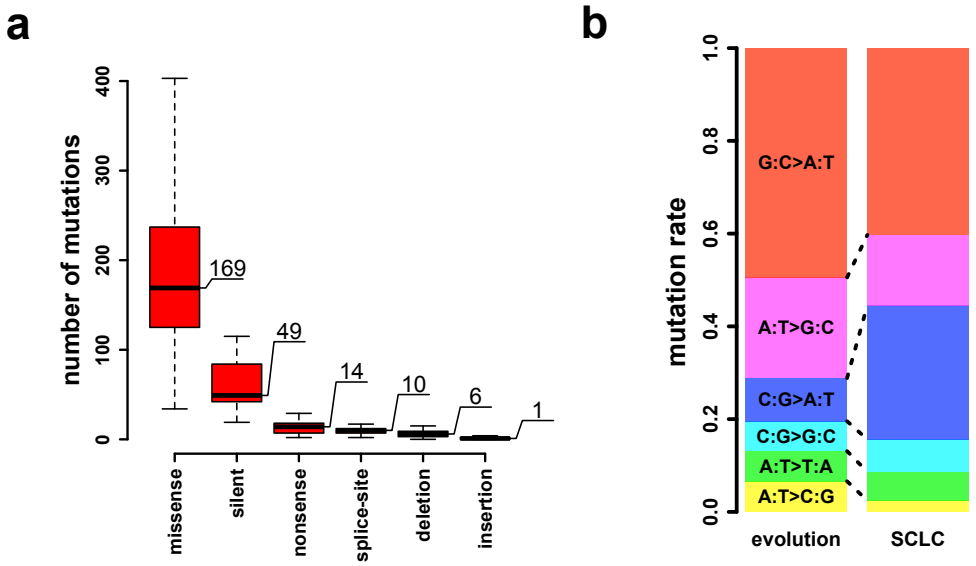
Supplementary Figure 3. Validation of *CREBBP-RHBDF1* and *MPRIP-TP53* fusions. Dideoxy validation sequencing results including the sequence across the fusion-point of the chimeric transcripts a) *CREBBP-RHBDF1* and b) *MPRIP-TP53*. The fusion-point is marked by a black rectangle in the electropherograms.

Nature Genetics: doi:10.1038/ng.2396



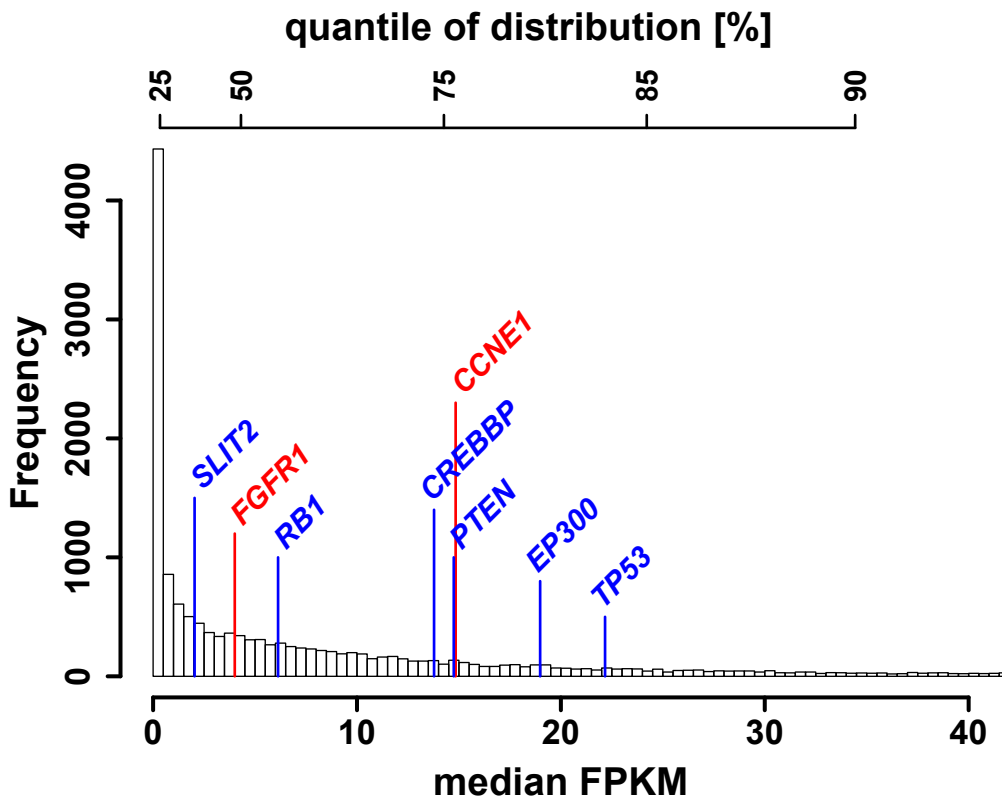
Supplementary Figure 4. Copy number status of genome/exome sequenced samples. a) A cross-platform comparison of copy numbers derived from whole genome sequencing (WG), whole exome sequencing (WE), and SNP 6.0 arrays. Centromeres are marked by vertical red lines in the genome annotation shown at the upper part of a and b. b) Absolute copy number segments inferred from whole exome sequencing. Arrows mark samples having a ploidy larger than two.

Nature Genetics: doi:10.1038/ng.2396



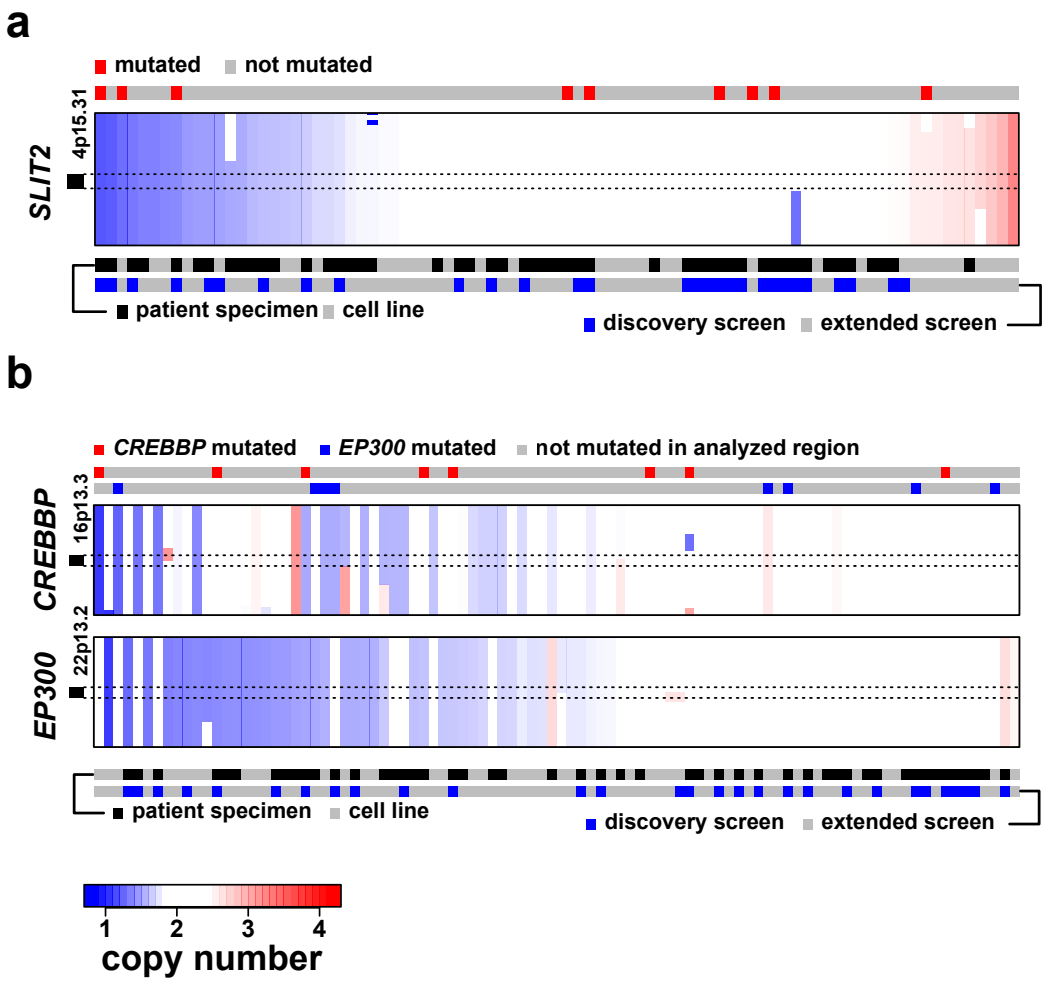
Supplementary Figure 5. Mutation spectrum of SCLC. a) Distribution of detected mutations by exome sequencing. b) A comparison of context independent transversion and transition rates (an overall strand symmetry is assumed) between rates derived from molecular evolution[19] and rates derived from the SCLC exome sequencing. All rates are scaled such that their overall sum is one for comparability reasons. The elevated rate of the C:G>A:T transversion that can be linked to the exposure of carcinogens in tobacco smoke.[20-22]

Nature Genetics: doi:10.1038/ng.2396



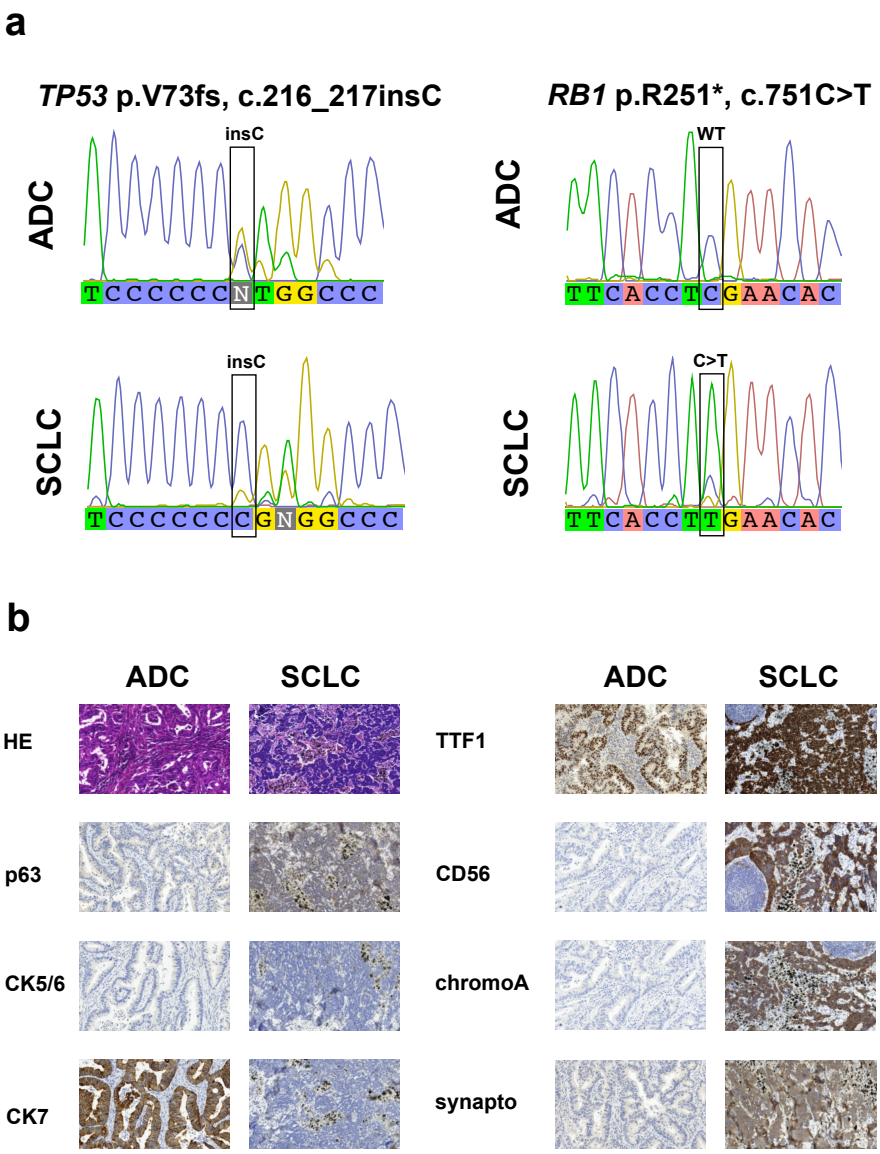
Supplementary Figure 6. Distribution of gene expression. Distribution of the median expression values. Median expression of bona-fide tumor suppressor genes such as *TP53*, *RB1*, and *PTEN* together with a selection of identified driver genes in SCLC is indicated in the overall expression profile.

Nature Genetics: doi:10.1038/ng.2396



Supplementary Figure 7. Mutational status of *SLIT2*, *CREBBP*, and *EP300*. Copy number and mutation status of a) *SLIT2* and b) *CREBBP*/*EP300* across all samples analyzed.

Nature Genetics: doi:10.1038/ng.2396



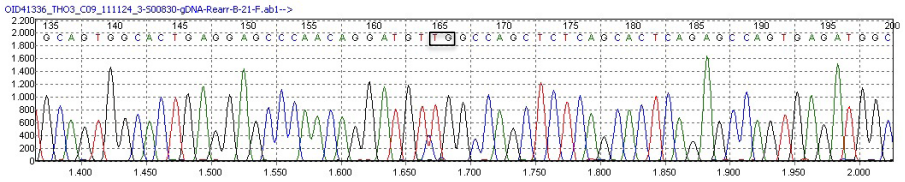
Supplementary Figure 8. Transdifferentiation of an adenocarcinoma to SCLC. a) Mutation analysis of the *TP53* (V73fs) and *RB1* (R251*) mutation between SCLC and lung adenocarcinoma (ADC) from the identical patient. Three years prior diagnosis of SCLC, an ADC tumor was surgically resected. b) Immune histochemistry of the ADC and SCLC tumor. To clearly discriminate the histology of the two tumors, 7 makers including: p63, CK5/6, CK7, TTF1, CD56, chromogranine A (chromoA), and synaptophysin (synapto) were stained.

Nature Genetics: doi:10.1038/ng.2396

a

S00830: chr3:71164157; chr3:71249957

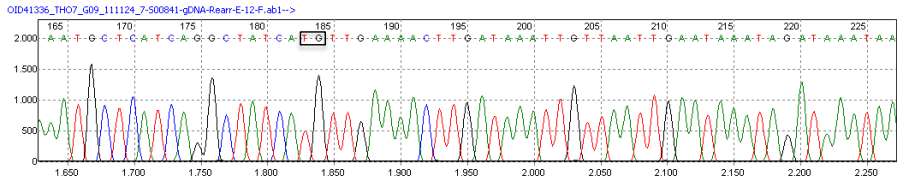
GTGAATGGGACAGATTGCCATATGGATGAAGACACTGAACAAGCTGTAAAACTCTTG
GCGACATGTGCCCGGGGGGATTGTCCCATATTAACACCACCTGGGCTGTGGCAATT
GTGCTCTGTCTGTGCAGTGGCACTGAGGAGCCCAACAGGATGT**TG**GCCAGCTCTCA
GCACTCAGAGCCAGTGAGATGGCCCTCTCGCTGCTGGCTCGGTAGGTCCTGTTTTC
ATACTAACTTGGAACCTGGACATTTTTGACACACCTTCCATTCTATTGGTCCAAGATTA
CTTCATTCTCATCAGTGCCTTCATATTCTTAGTAAATAAGTGCCATCTGTTATTTACAAA!



b

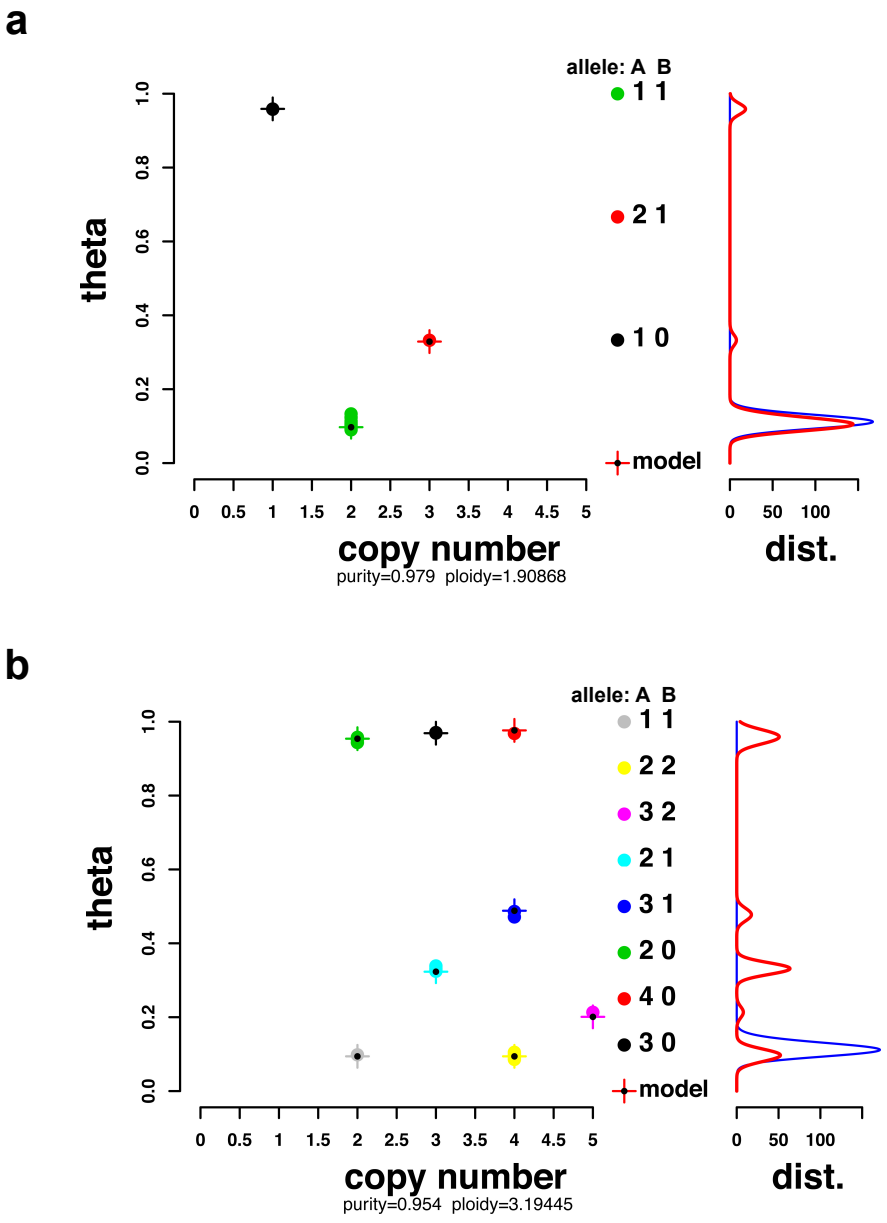
S00841:chr6:51387729 ; chr6:51076977

CTCCTACCTAAAAGGACTTTACAACTTTGAGTTCCCTGTTAATGCCTACTCAGTCTTTA
ATGCTCAATTCCATTATCACCTCCAATGGGAAGCCTAAAGTGACAAAGCCAACTGTTC
TGATGTGTTTAATTTCTCTCTTTTCAAATGCTCATCAGGCTATCA**TG**TTGAAAACTTGA
TAAATTGTTAATTGAATAAATAGATAAATAATTGCACAAACTAAAGAATACAGAGCTTCT
TATGAGAAAAGCTCTACTATTAAGCTTGTATGAGAATATGGTTATCTTTCTGAATATTAAT



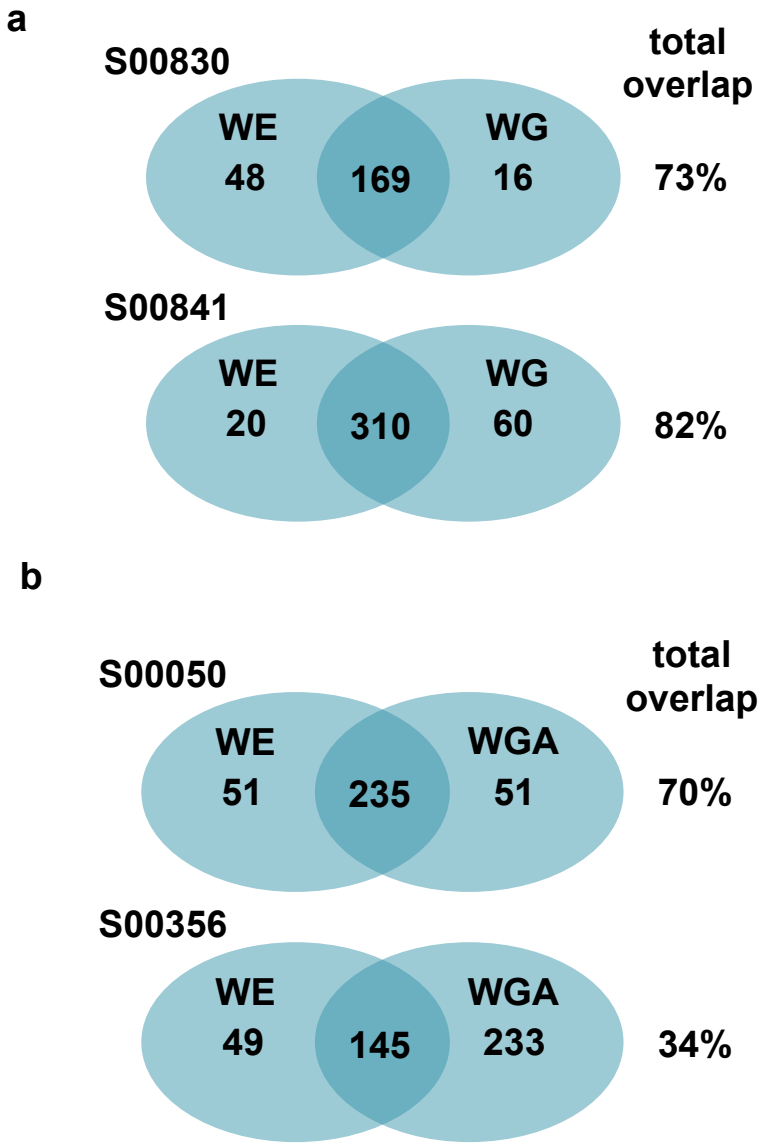
Supplementary Figure 9. Validation of genomic rearrangements. Validation sequencing of the rearrangements detected by whole genome sequencing. The genomic breakpoint is marked by a black rectangle in the electropherograms of the rearrangement in sample a) S00830 and b) S00841.

Nature Genetics: doi:10.1038/ng.2396



Supplementary Figure 10. Reconstruction of allelic states. Observed and modeled theta-values in case of a diploid a) and a triploid b) tumor. The different allelic states are indicated by distinct colors and the distribution of the threta-values between tumor (red) and normal (blue) is shown in the right panel. A cross marks model predictions of the theta-values.

Nature Genetics: doi:10.1038/ng.2396

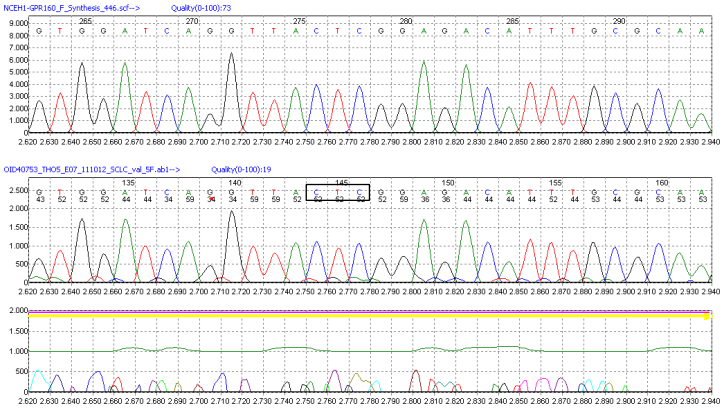


Supplementary Figure 11. Comparison between genome and exome sequencing. a) Overlap between mutation calls from whole exome sequencing (WE) and whole genome sequencing (WG). The total overlap between the two cases differs due to variations in the library. b) A comparison of mutation calls between exome sequencing of whole-genome amplified DNA (WGA) and non-amplified DNA (WE). The presence of WGA introduced artifacts is supported by the large proportion of oligonucleotide changes, which are present in the discordant set between WGA and non-WGA samples.

Nature Genetics: doi:10.1038/ng.2396

GPR160-NCEH1

RNAseq_transcript:<TGTAGTTTGAGCTTATTTTTTAGGCTGGCATCTTGAGTAAACTGTTGTCCAAGGGCAGCAGCCAGATTTCCA
CCAGCACTGTCACCAGAAATGCAAATTCTGCCTGGATCAACCATATACTTCTGTAAGACTTCTGGCTTCAGGAAATACTTTGTGGCC
CGTGCTCAGGAAAATAAACCTTTGGAACTAGCCTGTATTCAATGGAAACAATGACAGCATTCAATTCCTCAGCCATTGCTGTACACA
GCTCATCATAATACCTGATTTTTGCACTTGCCAAGGCCCAGCCTCCTCCGTGGATATAAACGACGCTGCGTTTCAGTGGCTCTTCG
GGCTTCGGAGGGCCTTCAAACACTCTGACTTCCACACCATCAAAGTCTGTGTCGGTCACCTTCACTTGGGCAGAAGACCACGCG
CTTTTTTTGCCAAAAGAAACAATGATAAAATTCAGTGCCAGCAGGTGATGGCTCAGTCCCAGGTAGTGGATCAGGTTA**CTC**GGAGA
CATTTGCGCAAGTGTAGGACGAAGCGAGAGCGGTTCCTTCCAGCCGAACCTGCCTCCGGAGGCCCGGTCCGTGCGTTCGTCTC
CGGACTGCAGGGGGCGGCCCTGCCACCACCTCGAGGCCACCCGAGGATCGATGGCTGTCGGCTCCGCAAGGTTGAGGCCCCG
ACACCTGCAGCGCCCGTGCCCAGGTGACCGCCCGAGGGAGGGGCGAG>



Supplementary Figure 12. Validation of *GPR160-NCEH1*. Same as in **Supplementary Figure 9** but for the *GPR160-NCEH1* chimeric transcript.

Nature Genetics: doi:10.1038/ng.2396