Vol 463|14 January 2010|doi:10.1038/nature08629

ARTICLES

A small-cell lung cancer genome with complex signatures of tobacco exposure

Erin D. Pleasance[1], Philip J. Stephens[1], Sarah O'Meara[1,2], David J. McBride[1], Alison Meynert[3], David Jones[1], Meng-Lay Lin[1], David Beare[1], King Wai Lau[1], Chris Greenman[1], Ignacio Varela[1], Serena Nik-Zainal[1], Helen R. Davies[1], Gonzalo R. Ordoñez[1], Laura J. Mudie[1], Calli Latimer[1], Sarah Edkins[1], Lucy Stebbings[1], Lina Chen[1], Mingming Jia[1], Catherine Leroy[1], John Marshall[1], Andrew Menzies[1], Adam Butler[1], Jon W. Teague[1], Jonathon Mangion[2], Yongming A. Sun[4], Stephen F. McLaughlin[5], Heather E. Peckham[5], Eric F. Tsung[5], Gina L. Costa[5], Clarence C. Lee[5], John D. Minna[6], Adi Gazdar[6], Ewan Birney[3], Michael D. Rhodes[4], Kevin J. McKernan[5], Michael R. Stratton[1,7], P. Andrew Futreal[1] & Peter J. Campbell[1,8]

Cancer is driven by mutation. Worldwide, tobacco smoking is the principal lifestyle exposure that causes cancer, exerting carcinogenicity through >60 chemicals that bind and mutate DNA. Using massively parallel sequencing technology, we sequenced a small-cell lung cancer cell line, NCI-H209, to explore the mutational burden associated with tobacco smoking. A total of 22,910 somatic substitutions were identified, including 134 in coding exons. Multiple mutation signatures testify to the cocktail of carcinogens in tobacco smoke and their proclivities for particular bases and surrounding sequence context. Effects of transcription-coupled repair and a second, more general, expression-linked repair pathway were evident. We identified a tandem duplication that duplicates exons 3–8 of *CHD7* in frame, and another two lines carrying *PVT1–CHD7* fusion genes, indicating that *CHD7* may be recurrently rearranged in this disease. These findings illustrate the potential for next-generation sequencing to provide unprecedented insights into mutational processes, cellular repair pathways and gene networks associated with cancer.

More than 1 billion people worldwide smoke tobacco[1]. With 20× greater risk of developing lung cancer than non-smokers and increased risk of many other tumour types, a smoker's lifestyle choice represents the most significant carcinogenic exposure confronting health services today. Tobacco smoke contains more than 60 mutagens that bind and chemically modify DNA[2,3], and these brand the lung cancer genome with characteristic mutational patterns. Point mutations in, for example, *TP53* and *KRAS* show different signatures between smokers and non-smokers with lung cancer[2–4]. However, such studies have been limited to a few genes, and it is unclear how representative these findings are of mutational processes across the whole genome[5]. *In vitro* assays and mouse models have been important tools for testing the mutagenicity of individual chemical constituents of tobacco smoke, but are of limited value for generalizing to the complexity of smoking behaviours, systemic metabolism and cancer development in humans. Massively parallel sequencing technologies promise the capacity to paint a genome-wide portrait of mutation in human cancer. Such data will provide unprecedented insights into the relative contributions of different tobacco carcinogens to mutation *in vivo*, the effects of local DNA structure on mutability and the cellular defence mechanisms against exogenous mutagens.

Lung cancer is the leading cause of cancer-related deaths worldwide, developing in more than a million new patients annually[6]. Small-cell lung cancer (SCLC), representing 15% of cases, is a distinct subtype associated with a typical clinical picture of early metastasis, initial response to chemotherapy but subsequent relapse, and a 2-year survival of <15%[7]. Several tumour suppressor genes are inactivated, including *TP53* (80–90% of cases[8]), *RB1* (60–90% of cases[9,10]) and *PTEN* (13% of cases[11]). Infrequent activating mutations have been found in *PIK3CA*, *EGFR* and *KRAS* (all 10% or lower; http://www.sanger.ac.uk/genetics/CGP/cosmic/), and *MYC* is amplified in 20% of cases.

The development of massively parallel sequencing technologies makes it feasible to catalogue all classes of somatically acquired mutation in a cancer, including base substitutions[12,13], insertions and deletions (indels)[12,13], copy number changes[14] and genomic rearrangements[14]. Reports from high-coverage sequencing of two acute myeloid leukaemia genomes have been published, which have concentrated on detecting point mutations in exons and regulatory regions[12,13]. Here, we report the first detailed analysis of a human cancer classically associated with tobacco smoking, giving unprecedented insights into the mutational burden associated with this lifestyle choice. Such analyses highlight the advances that will be made in our understanding of the pathogenesis of cancer as we sequence hundreds to thousands of human tumours[15].

Sequencing of a SCLC cell line

Most small-cell lung cancers are not surgically resected[7], meaning that cell lines are an indispensable resource for studying this disease. NCI-H209 is an immortal cell line derived from a bone marrow metastasis of a 55-year-old male with SCLC, taken before chemotherapy[16]. The smoking history of the patient is not recorded[16]. However, the specimen showed histologically typical small cells with classic neuroendocrine features: >97% of such tumours are associated with tobacco smoking[17,18]. An Epstein–Barr-virus-transformed

[1]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [2]Life Technologies, Warrington WA3 7QH, UK. [3]European Bioinformatics Institute, Hinxton CB10 1SD, UK. [4]Life Technologies, Foster City, California 94404, USA. [5]Life Technologies, Beverley, Massachusetts 01915, USA. [6]University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [7]Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [8]Department of Haematology, University of Cambridge CB2 2XY, UK.

lymphoblastoid line, NCI-BL209, has been generated from the patient. NCI-H209 has been extensively characterized by spectral karyotyping, capillary sequencing and high-resolution copy-number array (http://www.sanger.ac.uk/genetics/CGP/cosmic/).

Using the SOLiD platform, we generated 25-base-pair (bp) short-read, mate-pair shotgun sequences from the tumour and matched normal genomes. On the basis of detailed power calculations, we estimated that tumour and normal genomes should be sequenced to 30-fold depth to identify somatically acquired genetic variants with high sensitivity and distinguish them from both sequencing errors and germline polymorphisms (Fig. 1a). In total, 112 gigabases (Gb; 39× coverage) from the tumour and 90 Gb (31×) from the normal were aligned to the reference genome (Fig. 1b).

Bioinformatic algorithms were developed to identify somatically acquired genetic variation from the sequencing data (Supplementary Fig. 1 and Supplementary Tables 1–5), subjected to rigorous validation by polymerase chain reaction (PCR) and capillary sequencing. We had previously identified 29 base substitutions, of which 22

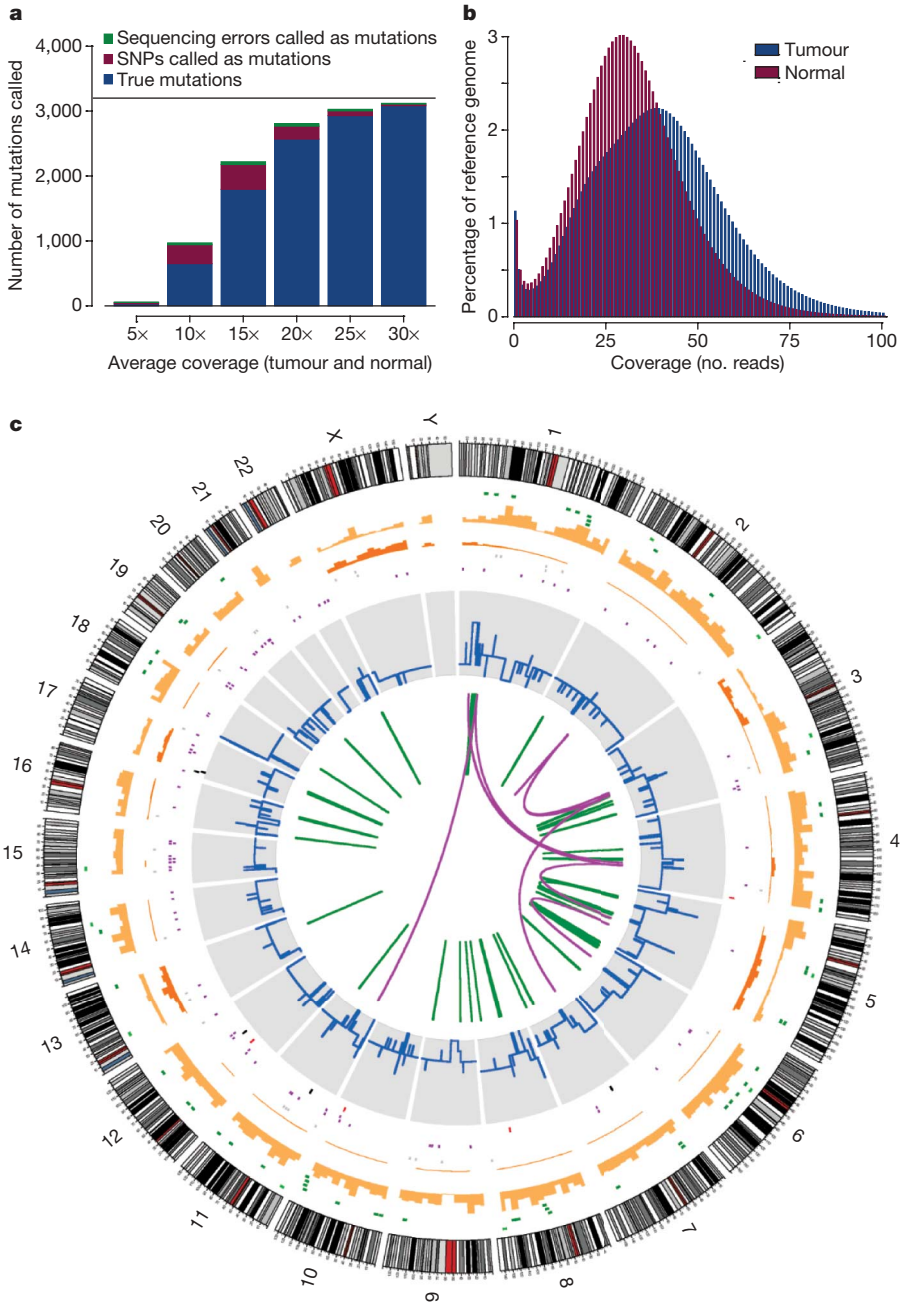

Figure 1 | The compendium of somatic mutations in a small-cell lung cancer genome. **a**, Power calculations showing the number of true somatic substitutions detected (blue) and mis-calls (single nucleotide polymorphisms (SNPs) called as somatic mutations, burgundy, and sequencing errors called as mutations, green) for different levels of sequence coverage. Calculations are based on a true mutation prevalence of 1 per megabase (black line). **b**, Histogram of the actual coverage achieved per base of the tumour (blue) and normal (burgundy) genomes. **c**, Figurative representation of the catalogue of somatic mutations in the genome of NCI-H209. Chromosome ideograms are shown around the outer ring and are oriented pter–qter in a clockwise direction with centromeres indicated in red. Other tracks contain somatic alterations (from outside to inside): validated insertions (light-green rectangles); validated deletions (dark-green rectangles); heterozygous (light-orange bars) and homozygous (dark-orange bars) substitutions shown by density per 10 megabases; coding substitutions (coloured squares; silent in grey, mis-sense in purple, nonsense in red and splice site in black); copy number (blue lines); validated intrachromosomal rearrangements (green lines); and validated interchromosomal rearrangements (purple lines).

(76%) were called by our algorithm from the SOLiD sequencing data (Supplementary Results and Supplementary Table 6). A total of 79 novel coding substitutions and 354 randomly chosen genome-wide variants called by the algorithm were also tested. A total of 77 (97%) of the coding substitutions and 333 (94%) of the random variants were confirmed as genuine somatic mutations (Supplementary Table 7). Under the conditions given here, small indels are difficult to detect and neither of two known indels in coding sequence was identified. Of putative somatic indels that were called, the true-positive rate was 25% by capillary sequencing (Supplementary Results and Supplementary Table 8). Therefore, only somatic indels which were confirmed by capillary sequencing are reported here. All somatic genomic rearrangements called by anomalous read pairs were validated by PCR and capillary sequencing across the breakpoint, as previously described[14].

Repertoire of somatic mutation

Overall, 22,910 somatically acquired substitutions were identified across the NCI-H209 genome, and a further 65 indels, 334 copy number segments and 58 structural variants were confirmed (Table 1, Fig. 1c and Supplementary Tables 1–5).

For point mutations in coding regions, we found the previously described *RB1* C706F mutation, known to abrogate protein function[19], and the mutation that disrupts a splice site in *TP53*. Combined loss of *RB1* and *TP53* is a characteristic feature of SCLC, confirming that NCI-H209 is genetically typical of this disease. One G>T transversion generated a premature stop codon in *MLL2*. We have observed clustering of truncating mutations in this gene, a histone methyltransferase, in renal cancer[20]. Of coding variants, 94 are predicted to change amino acids, and 36 are synonymous. Because cancer is a clonal disease in which the phenotypic consequences of mutation are subject to Darwinian natural selection, accumulation of mutations conferring selective advantage on cancer subclones will manifest as an excess of non-synonymous mutations. However, the observed non-synonymous:synonymous ratio of 2.61:1 is not significantly different from that expected by chance ($P = 0.3$), suggesting that the majority of coding variants do not confer a selective advantage to the cancer.

Owing to the limited throughput of capillary sequencing, there has previously been little attempt to explore regulatory regions of the genome for potential oncogenic mutations. To address this, we extracted somatic substitutions occurring within 2 kilobases (kb) either side of known transcription start sites, which would generally include gene promoters. Mutations were evenly distributed across the 4-kb regions (Supplementary Fig. 2A). We applied hidden Markov models to predict which substitutions might affect transcription factor binding sites. The distribution observed was no different to that seen in random, simulated sets of 'mutations' (Supplementary Fig. 2B), indicating that, analogous to substitutions in coding sequence, most of those found in regulatory regions are selectively neutral to the cancer. Nonetheless, as with coding mutations, there may be a small number that alter transcription factor binding and affect gene regulation, thus providing phenotypic variation for selection to act upon. For example, a T>G mutation 49 bp upstream of the transcription start site of a gene in the RAS oncogene family, *RAB42*, is predicted to have significant disruptive effects on a potential binding motif for the RAS-responsive RREB1 transcription factor ($P = 3 \times 10^{-98}$; Supplementary Fig. 2C).

Taken together, these data indicate that most of the mutations in coding and promoter regions of the NCI-H209 genome are passenger events, conferring no selective advantage to the cells. Ranking algorithms can be useful to prioritize variants for further study, but the key evidence for identifying driver mutations is recurrence in independent tumour samples, supplemented by functional studies.

Table 1 | Somatically acquired genomic variants of all classes in a SCLC genome

Variant	Number
Somatic substitution	22,910
Coding	134 (0.6%)
Nonsense	4
Non-synonymous	94
Synonymous	36
Non-coding, transcribed	182 (0.8%)
Untranslated region	119
Non-coding RNA	63
Intronic	6,463 (28%)
Splice site	5
Other intronic	6,458
Intergenic	16,131 (70%)
Insertions and deletions	65
Coding (frameshift)	2 (3%)
Intronic	25 (38%)
Intergenic	38 (58%)
Genomic rearrangements	58
Deletions	18 (31%)
Tandem duplications	9 (16%)
Other non-inverted intrachromosomal rearrangements	9 (16%)
Inverted intrachromosomal rearrangements	15 (26%)
Interchromosomal rearrangements	7 (12%)
Copy number segments	334

Multiple mutation signatures in NCI-H209

Tobacco smoke contains more than 60 carcinogens which bind and chemically modify DNA, characteristically forming bulky adducts at purine bases (guanine and adenine)[3]. Adducts distort the DNA helix and, if not corrected by nucleotide excision repair or other pathways, allow non-Watson–Crick pairing during DNA replication. The physicochemical properties of the mutagen determine which adduct is formed, what repair mechanism is induced and which mis-pairing is permissible[3]. The substantial mutational load carried in the NCI-H209 genome allows us to discern with great statistical power several distinct mutation signatures—genomic records of the medley of mutagens deposited in the airways and lungs by tobacco smoking.

G>T/C>A transversions were the commonest change observed (34%), followed by G>A/C>T (21%) and A>G/T>C (19%) transitions (Fig. 2a). This distribution is remarkably similar to the pattern of substitutions observed in *TP53* in SCLC cases curated from the published literature (Supplementary Fig. 3). This implies first that the NCI-H209 genome is typical of SCLC, and therefore of tobacco-associated mutational profiles, and second that most mutations were acquired *in vivo*, not during cell culture. G>T transversions caused by polycyclic aromatic hydrocarbons occur more frequently at methylated CpG dinucleotides *in vitro* and in *TP53*[21,22]. To explore this genome-wide, we compared the base preceding G>T mutations with the base before wild-type guanines in NCI-H209 (Fig. 2b). CpG dinucleotides were significantly enriched among the G>T mutation set compared to controls (odds ratio (OR), 1.5; 95% confidence interval (CI), 1.3–1.6; $P < 0.0001$). We can use the fact that only 10–20% of CpG dinucleotides in CpG islands are constitutively methylated compared with 60–70% outside of CpG islands[23] to assess how cytosine methylation affects mutations at the neighbouring guanine (Fig. 2c). G>T mutations at CpG dinucleotides were significantly more likely to be found outside CpG islands than expected by chance (OR, 1.8; 95% CI, 1.1–2.8; $P = 0.02$), suggesting that these transversions do indeed preferentially occur at methylated CpGs.

We next assessed the base preceding the guanine for G>A and G>C mutations (Fig. 2b). For G>A transitions, marked enrichment of CpG dinucleotides was observed in the mutation set compared with wild-type guanines in the genome (OR, 4.0; 95% CI, 3.7–4.3; $P < 0.0001$), and these showed a strong propensity to occur outside CpG islands (OR, 2.6; 95% CI, 1.6–4.1; $P < 0.0001$). This is consistent with the well-described phenomenon of spontaneous deamination of methylated cytosine to thymine. Although G>C transversions showed a similar enrichment for CpG context (OR, 2.2; 95% CI, 1.9–2.5; $P < 0.0001$), these were significantly more likely to occur within CpG islands (OR, 0.6; 95% CI, 0.4–1.0; $P = 0.05$), indicating that the

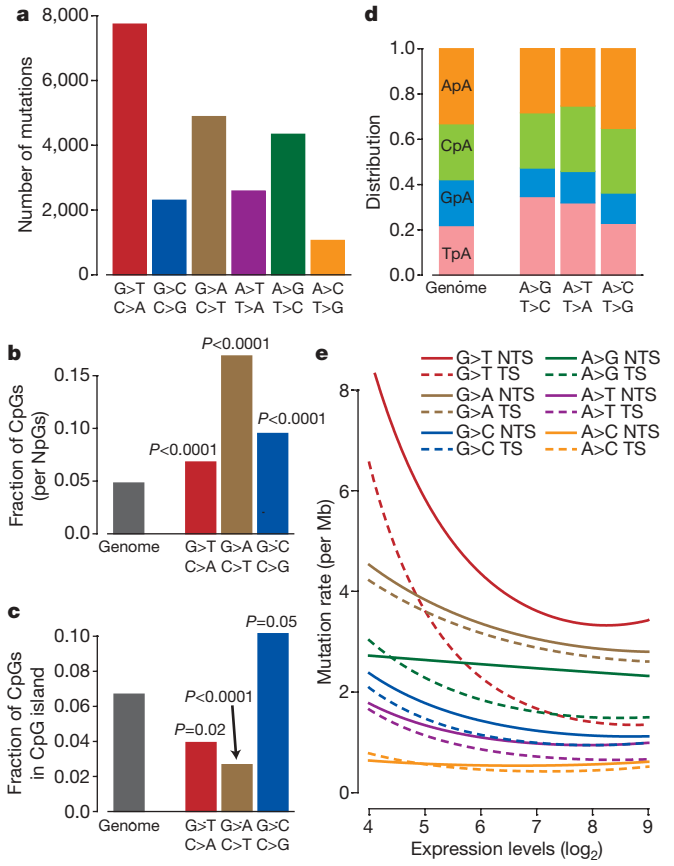

Figure 2 | The mutation profile of NCI-H209. a, Numbers of mutations in each of the six possible mutation classes. **b**, Fraction of the three classes of guanine mutations occurring at CpG dinucleotides in NCI-H209, with *P* values reflecting the comparison with the expected fraction across the genome (grey). **c**, Fraction of guanine mutations at CpGs which are found in CpG islands for each of the three classes of mutation. *P* values reflect comparison with the genome-wide fraction (grey) of CpGs found in CpG islands (and hence more likely to be constitutively unmethylated) versus outside CpG islands (high rates of constitutive methylation). **d**, Distribution of the four NpA dinucleotides for each of the three classes of adenine mutation in NCI-H209, compared to the expected distribution across the genome (left). **e**, Fitted curves showing the effects of gene expression and strand bias on mutation prevalence for the six classes of adenine and guanine mutation in NCI-H209. The *y* axis is expressed as mutations per Mb of at-risk nucleotides, namely mutations/1,000,000 Gs for G>T. NTS, non-transcribed strand; TS, transcribed strand.

carcinogen responsible targets unmethylated CpG dinucleotides. In keeping with previous reports[24,25], we found that the guanine base in G>C transversions was more frequently followed by an adenine than expected by chance (OR, 1.4; 95% CI, 1.3–1.5; $P < 0.0001$).

For mutations involving adenines, fewer substitutions of all classes were seen at GpA dinucleotides than expected by chance ($P < 0.0001$; Fig. 2d), and A>T and A>G occurred significantly more frequently at TpA than expected ($P < 0.0001$). Among somatically acquired indels, single-base-pair insertions were more likely to be gains of A or T nucleotides than C or G (8:1). Curiously, single base deletions favoured loss of C/G nucleotides, rather than A/T (26:12), and there was a propensity for the C/G deletions to occur at CC or GG dimers or longer (18 out of 26). In contrast to the frequency of indels at runs of A or T nucleotides, deletions at C or G tracts are not well described, and our findings may reflect a distinct mutation signature.

Thus, the sequence context of the ~23,000 mutations in the NCI-H209 genome provides tremendous power to identify multiple distinctive mutation signatures, not evident from targeted re-sequencing studies of limited genomic regions.

Imprint of two DNA repair pathways

Several pathways can repair DNA lesions caused by exogenous carcinogens. Bulky adducts on purines are the predominant form of DNA damage induced by tobacco carcinogens, and can be sufficiently disruptive to impede RNA polymerase when they occur on the transcribed strand of genes. Stalled RNA polymerases can recruit the nucleotide excision repair machinery, leading to excision of the altered nucleotide, preventing mutation. In studies of *TP53* mutations in lung cancer, G>T transversions occur more frequently on the non-transcribed strand[2,5], suggesting that many of the same lesions occurring on the transcribed strand are correctly identified and removed by the cell. We found that guanine and adenine substitutions are generally less frequent on the transcribed than the non-transcribed strand (Supplementary Fig. 4), confirming that purines seem to be the major target of carcinogens in tobacco smoke.

We next correlated mutation prevalence to gene expression (Fig. 2e). For a given level of gene expression, the effects of transcription-coupled repair are revealed by the significant separation of curves for mutations on the transcribed and non-transcribed strands. We found evidence for significant transcription-coupled repair for G>T transversions ($P < 0.0001$), as well as A>G ($P = 0.003$) and A>T ($P = 0.03$), possibly G>C ($P = 0.08$), but not G>A ($P = 0.3$) or A>C ($P = 0.8$) mutations. Thus, the extent of transcription-coupled repair differs for the various classes of mutation, presumably reflecting differences in the ability of the transcription-coupled repair machinery to recognize and/or repair different adduct lesions.

For most mutations, there seems to be another novel expression-linked repair pathway that operates on both strands and is at least as numerically important as transcription-coupled repair. Thus, significantly lower mutation prevalence, on both transcribed and non-transcribed strands, was observed in more highly expressed genes for G>T ($P < 0.0001$), G>A ($P < 0.0001$), G>C ($P < 0.0001$) and A>T ($P < 0.0001$). Again, there are some interesting differences across mutation classes in the relative contributions of the two repair pathways. For A>G mutations, only transcribed strand mutations decreased with higher gene expression, suggesting that transcription-coupled repair is the more important pathway for preventing such events. In contrast, G>A mutations occurred equally on transcribed and non-transcribed strands, but mutations on both strands were significantly reduced in more highly expressed genes, indicating that the novel expression-linked repair pathway is more important than transcription-coupled repair here.

Taken together, these data imply that at least two separate DNA repair pathways have been enlisted for protection of the NCI-H209 genome, notwithstanding the difficulties in extrapolating cell line expression levels to *in vivo* expression during cancer progression. The fact that the two pathways have operated with differing efficacy across the six classes of mutation implies that the lesions have distinct physicochemical effects on DNA structure, with variable recognition and excision by the genome surveillance machinery.

Genomic rearrangements and copy number

We identified 58 somatically acquired genomic rearrangements in the NCI-H209 genome. These include 18 (31%) deletions and 9 (16%) tandem duplications. The majority of rearrangements, however, cannot be ascribed to classical structural variant patterns, due to the considerably greater complexity of somatically acquired rearrangements compared to germline events. This is exemplified by a set of rearrangements incorporating regions from chromosomes 1p32-36 and 4q25-28 (Fig. 3). Here, most of the intrachromosomal rearrangements are in inverted orientation, but cannot be classical inversions because they demarcate copy number changes and do not have reciprocal breakpoints. By similar reasoning, most interchromosomal rearrangements also seem to be unbalanced. Other clusters of unbalanced rearrangements were found in NCI-H209, including chromosomes 3q and 5q, and we have seen this phenomenon in many other solid tumour genomes.

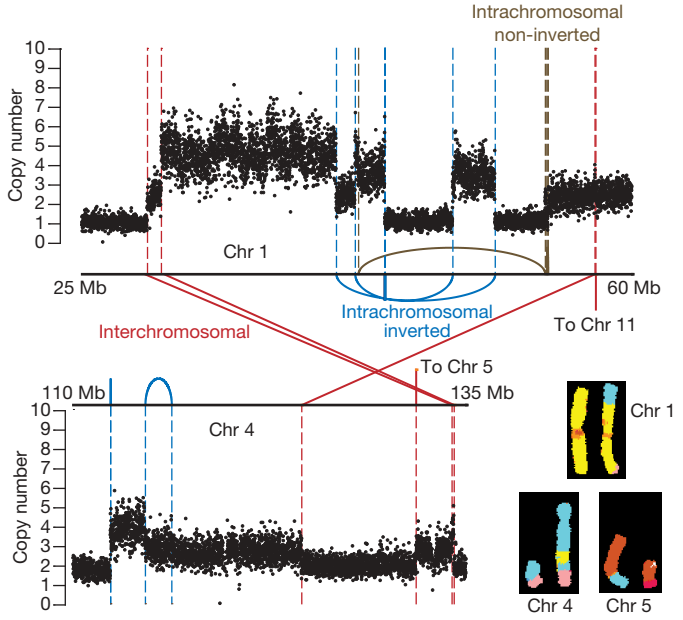

Figure 3 | Localized complexes of somatically acquired genomic rearrangements in NCI-H209. Copy number plots across regions on chromosomes 1 and 4 are shown. Inverted intrachromosomal rearrangements (blue), non-inverted intrachromosomal rearrangements (brown) and interchromosomal rearrangements (red) are shown in relation to copy number changes. The inset shows the representative chromosomes on spectral karyotyping. There are three breakpoints between chromosome 1 (yellow) and 4 (light blue), and a translocation between chromosomes 4 and 5 (tan).

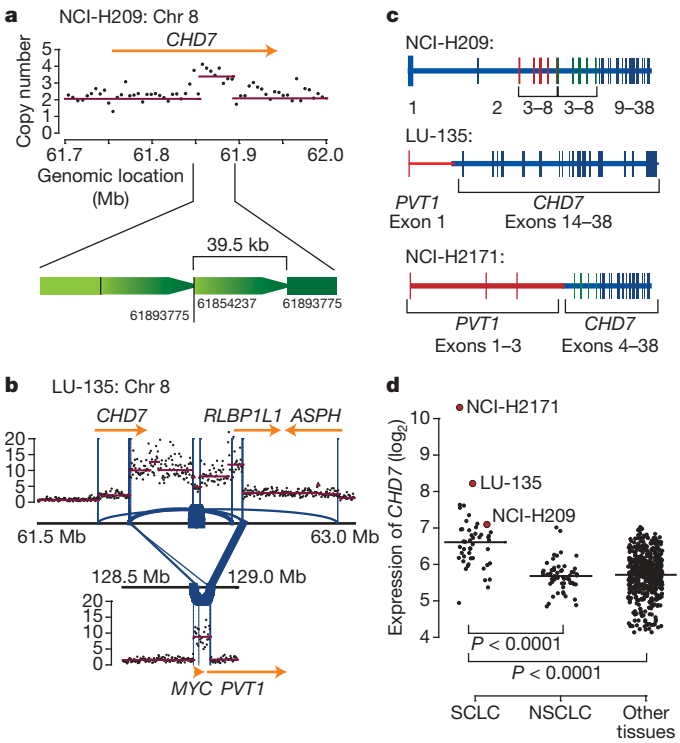

Figure 4 | *CHD7* rearrangements in SCLC cell lines. a, A somatically acquired 39.5-kb tandem duplication is found in NCI-H209. **b**, The LU-135 cell line shows co-amplification of the 3′ portion of *CHD7* together with *MYC* and the 5′ portion of *PVT1*. Blue lines show locations of genomic rearrangements observed in the amplicons, with the thickness of the line proportional to the number of reads spanning the breakpoint. **c**, Transcripts resulting from *CHD7* rearrangements are an in-frame duplication of exons 3–8 in NCI-H209 and two amplified *PVT1–CHD7* fusion genes in NCI-H2171 and LU-135. **d**, *CHD7* is overexpressed in SCLC compared to both non-small-cell lung cancer and other tumour types. LU-135 and NCI-H2171 show massive overexpression of *CHD7* in keeping with the genomic amplification present in these cell lines.

Chromosomal rearrangements can juxtapose two genes: if they are in the same orientation with an intact open reading frame, an oncogenic fusion gene may result. In NCI-H209, a predicted in-frame fusion gene was created by a 240-kb deletion on chromosome 16, adjoining the first two exons of *CREBBP* with the 3′ portion of *BTBD12*, a gene involved in repair of double-stranded DNA breaks[26,27]. Notably, in acute myeloid leukaemia, *CREBBP* is recurrently fused with *MYST3* (ref. 28). PCR with reverse transcription (RT–PCR) showed that the predicted *CREBBP–BTBD12* fusion transcript is expressed in NCI-H209, but not in 55 other SCLC cell lines. The significance of the predicted fusion gene with respect to cancer development is therefore unclear.

CHD7 rearrangements in SCLC cell lines

Intrachromosomal rearrangements can also result in internal rearrangements of genes, through loss or duplication of exons. A 39-kb tandem duplication was found in *CHD7*, predicted to lead to in-frame duplication of exons 3–8 (Fig. 4a). We previously identified a massively amplified and highly expressed fusion gene comprising exons 1–3 of *PVT1*, a non-coding RNA gene immediately downstream of *MYC*, and exons 4–38 of *CHD7* in another SCLC cell line, NCI-H2171[14]. This raises the possibility that *CHD7* rearrangements may be recurrent in SCLC. Using multiplex ligation-dependent probe amplification, we identified a further SCLC cell line (LU-135) with internal exon copy number alterations, among 63 lines screened (Supplementary Fig. 5). LU-135 was therefore studied by mate-pair sequencing (Fig. 4b). This demonstrated that, as for NCI-H2171, the *CHD7* amplicon was linked to *MYC* amplification. One breakpoint predicted the existence of a fusion gene between exon 1 of *PVT1* and exons 14–38 of *CHD7* (Fig. 4c), and as demonstrated by RT–PCR across the breakpoint, this transcript is expressed. In keeping with genomic amplification and active expression of the *PVT1* locus, NCI-H2171 and LU-135 show particularly elevated levels of *CHD7* transcripts (Fig. 4d). SCLC cell lines on average show a $\log_2$ greater expression of *CHD7* than both non-small-cell lung cancer lines and other tumour types ($P < 0.0001$).

Thus, *CHD7* is rearranged in three SCLC cell lines. Two carry a *PVT1–CHD7* fusion gene in the setting of *MYC* amplification. *PVT1* is a non-coding gene immediately downstream of *MYC*, and may itself be a transcriptional target of the MYC protein[29]. Insertion of *CHD7* into this locus with subsequent amplification gives the double hit of increased gene copy number and regulatory elements for a co-amplified transcription factor, explaining the massive overexpression seen in these cell lines. *PVT1* is recurrently rearranged in variant Burkitt's lymphoma translocations[30], and may be oncogenic[31]. The NCI-H209 rearrangement is predicted to duplicate one of the two chromodomains. CHD7 is a chromatin remodeller, promoting enhancer-mediated transcription through association with histone H3K4 methylation[32]. Histone modifiers have been implicated as cancer genes[33], and a family member, *CHD5*, may function as a tumour suppressor gene[34]. Recurrent rearrangements of *CHD7* in SCLC would be an interesting extension of this theme if functional studies and genomic analyses of primary samples confirm our data.

Discussion

The compendium of somatic alterations in a cancer genome is shaped by multiple intrinsic and extrinsic processes, including exposure to mutagens, selective pressures active in the tissue microenvironment, genomic instability and DNA repair pathways[15]. The advent of massively parallel sequencing heralds an era in which unbiased, genome-wide mutation screens allow the consequences of these processes to be discerned and decoded. Even in this single lung cancer genome, we

can identify several distinctive point mutation patterns, reflecting the cocktail of carcinogens present in cigarette smoke, as well as signatures of the partially successful attempts of the cell's surveillance machinery to repair DNA damage. The complete catalogue of somatically acquired mutations in a given cancer harbours the subset of variants that drive the neoplastic phenotype, and a likely candidate, *CHD7* rearrangement, has emerged from the NCI-H209 genome.

Tobacco smoke deposits many hundreds of chemicals in the airways and lungs. Each carcinogen-associated mutation represents the consequence of three processes: chemical modification of a purine by a mutagen, failure to repair the lesion by genome surveillance pathways and incorrect nucleotide incorporation opposite the distorted base during DNA replication. G>T transversions are the commonest substitution in NCI-H209, mutations previously linked to polycyclic aromatic hydrocarbons[3] and acrolein[22] in tobacco smoke. We found enrichment of G>T mutations at CpG dinucleotides, especially outside CpG islands, supporting *in vitro* evidence that these carcinogens preferentially bind methyl-CpG dinucleotides[21,22]. Polycyclic aromatic hydrocarbons containing a cyclopentane ring have been associated with G>C transversions[35]. We found them enriched at CpG dinucleotides, but, in contrast to G>T and G>A mutations, our data indicate that unmethylated CpGs are the target here, underscoring the remarkable statistical power genome-wide mutation screens give for delineating mutation spectra.

We can also infer signatures of DNA repair in cancer genomes. Transcription-coupled repair is induced by stalling of RNA polymerase at bulky adducts on the transcribed strand, and we saw evidence for this process in NCI-H209. We have also discovered the imprint of a novel and more general form of expression-linked repair, through which mutation frequency is reduced on both strands in highly expressed genes. The expression-linked decrease in mutation frequency may reflect global genomic nucleotide excision repair, in which distorting adducts are corrected genome-wide[36]. Why this pathway should be more effective in highly transcribed regions is unclear. One possibility is that single-stranded (ss)DNA formed on both strands during transcription facilitates recognition of the adduct, and there is some evidence that components of the nucleotide excision repair pathway can recognize adducts in ssDNA[37]. Strikingly, some mutation types were repaired almost exclusively by transcription-coupled repair (A>G), some showed evidence for only the more general expression-linked repair (G>A), whereas others had features of both mechanisms (G>T, A>T). Such differences are presumably determined by the physicochemical properties of the multitudinous adducts induced by carcinogens present in tobacco smoke.

On average, lung cancer develops after 50 pack-years of smoking[38] (where a pack-year is 7,300 cigarettes, representing the number smoked in a pack a day for a year). Candidate gene re-sequencing studies suggest that the mutation prevalence in NCI-H209 is similar to that of primary lung cancers[39,40]. If the majority of mutations derive from the mélange of mutagens present in tobacco smoke, the clone of cells that ultimately becomes cancerous would acquire, over its lifetime, an average of one mutation for every 15 cigarettes smoked. If this is the case in a localized cluster of cells, then the number of mutations acquired across the whole bronchial tree from even one cigarette must be substantial. The data presented here demonstrate the power of whole-genome sequencing to disentangle the many complex mutational signatures found in cancers induced by tobacco smoke.

METHODS SUMMARY

Massively parallel, shotgun sequencing of genomic DNA from NCI-H209 and the matched EBV-transformed B-cell line was performed on the SOLiD platform (Life Technologies) to a target >30-fold depth. Mate-pair 25-bp reads were mapped back to the NCBI36 reference genome in 2-base encoded colour space. Bespoke bioinformatic algorithms were developed to identify all classes of somatically acquired genetic variant from the sequencing data. PCR and capillary sequencing on DNA from both tumour and normal lines was performed to confirm all putative somatically acquired genomic rearrangements and indels, together with random samples of coding and non-coding substitutions. To identify potential regulatory consequences of point mutations, substitutions in promoter regions were compared against random samples of simulated variants for effects on predicted transcription-factor binding sites. The sequence context of mutation positions was compared against randomly sampled genomic positions which had sufficient sequence coverage by chi-squared tests. Gene expression levels for NCI-H209 were determined on the Affymetrix U133A microarray, and analysed for mutation prevalence by Poisson regression.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 September; accepted 30 October 2009.
Published online 16 December 2009.

1. Jha, P. Avoidable global cancer deaths and total deaths from smoking. *Nature Rev. Cancer* **9,** 655–664 (2009).
2. Hecht, S. S. Progress and challenges in selected areas of tobacco carcinogenesis. *Chem. Res. Toxicol.* **21,** 160–171 (2008).
3. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21,** 7435–7451 (2002).
4. DeMarini, D. M. Genotoxicity of tobacco smoke and tobacco smoke condensate: a review. *Mutat. Res.* **567,** 447–474 (2004).
5. Rodin, S. N. & Rodin, A. S. Origins and selection of p53 mutations in lung carcinogenesis. *Semin. Cancer Biol.* **15,** 103–112 (2005).
6. Toh, C. K. The changing epidemiology of lung cancer. *Methods Mol. Biol.* **472,** 397–411 (2009).
7. Sher, T., Dy, G. K. & Adjei, A. A. Small cell lung cancer. *Mayo Clin. Proc.* **83,** 355–367 (2008).
8. Wistuba, I. I., Gazdar, A. F. & Minna, J. D. Molecular genetics of small cell lung carcinoma. *Semin. Oncol.* **28,** 3–13 (2001).
9. Horowitz, J. M. *et al.* Frequent inactivation of the retinoblastoma anti-oncogene is restricted to a subset of human tumor cells. *Proc. Natl Acad. Sci. USA* **87,** 2775–2779 (1990).
10. Mori, N. *et al.* Variable mutations of the RB gene in small-cell lung carcinoma. *Oncogene* **5,** 1713–1717 (1990).
11. Yokomizo, A. *et al.* PTEN/MMAC1 mutations identified in small cell, but not in non-small cell lung cancers. *Oncogene* **17,** 475–479 (1998).
12. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
13. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361,** 1058–1066 (2009).
14. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40,** 722–729 (2008).
15. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
16. Carney, D. N. *et al.* Establishment and identification of small cell lung cancer cell lines having classic and variant features. *Cancer Res.* **45,** 2913–2923 (1985).
17. Barbone, F., Bovenzi, M., Cavallieri, F. & Stanta, G. Cigarette smoking and histologic type of lung cancer in men. *Chest* **112,** 1474–1479 (1997).
18. Lubin, J. H. & Blot, W. J. Assessment of lung cancer risk factors by histologic category. *J. Natl Cancer Inst.* **73,** 383–389 (1984).
19. Kaye, F. J., Kratzke, R. A., Gerster, J. L. & Horowitz, J. M. A single amino acid substitution results in a retinoblastoma protein defective in phosphorylation and oncoprotein binding. *Proc. Natl Acad. Sci. USA* **87,** 6922–6926 (1990).
20. Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* doi:10.1038/nature08672 (in the press).
21. Denissenko, M. F., Chen, J. X., Tang, M. S. & Pfeifer, G. P. Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proc. Natl Acad. Sci. USA* **94,** 3893–3898 (1997).
22. Feng, Z., Hu, W., Hu, Y. & Tang, M. S. Acrolein is a major cigarette-related lung cancer agent: Preferential binding at p53 mutational hotspots and inhibition of DNA repair. *Proc. Natl Acad. Sci. USA* **103,** 15404–15409 (2006).
23. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genet.* **38,** 1378–1385 (2006).
24. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
25. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321,** 1801–1806 (2008).
26. Fekairi, S. *et al.* Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases. *Cell* **138,** 78–89 (2009).
27. Svendsen, J. M. *et al.* Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. *Cell* **138,** 63–77 (2009).
28. Rozman, M. *et al.* Type I MOZ/CBP (MYST3/CREBBP) is the most common chimeric transcript in acute myeloid leukemia with t(8;16)(p11;p13) translocation. *Genes Chromosom. Cancer* **40,** 140–145 (2004).
29. Carramusa, L. *et al.* The PVT-1 oncogene is a Myc protein target that is overexpressed in transformed cells. *J. Cell. Physiol.* **213,** 511–518 (2007).

30. Zeidler, R. *et al.* Breakpoints of Burkitt's lymphoma t(8;22) translocations map within a distance of 300 kb downstream of *MYC*. *Genes Chromosom. Cancer* **9,** 282–287 (1994).
31. Guan, Y. *et al.* Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin. Cancer Res.* **13,** 5745–5755 (2007).
32. Schnetz, M. P. *et al.* Genomic distribution of CHD7 on chromatin tracks H3K4 methylation patterns. *Genome Res.* **19,** 590–601 (2009).
33. van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene *UTX* in human cancer. *Nature Genet.* **41,** 521–523 (2009).
34. Bagchi, A. *et al.* CHD5 is a tumor suppressor at human 1p36. *Cell* **128,** 459–475 (2007).
35. Jackson, M. A., Lea, I., Rashid, A., Peddada, S. D. & Dunnick, J. K. Genetic alterations in cancer knowledge system: analysis of gene mutations in mouse and human liver and lung tumors. *Toxicol. Sci.* **90,** 400–418 (2006).
36. Shuck, S. C., Short, E. A. & Turchi, J. J. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. *Cell Res.* **18,** 64–72 (2008).
37. Liu, Y. *et al.* Interactions of human replication protein A with single-stranded DNA adducts. *Biochem. J.* **385,** 519–526 (2005).
38. Lubin, J. H. *et al.* Cigarette smoking and cancer risk: modeling total exposure and intensity. *Am. J. Epidemiol.* **166,** 479–489 (2007).
39. Davies, H. *et al.* Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res.* **65,** 7591–7595 (2005).
40. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Wellcome Trust (grant reference 077012/Z/05/Z). P.J.C. is a Kay Kendall Leukaemia Fund Intermediate Clinical Fellow. I.V. is supported by the Human Frontiers Science Programme. J. Minna. and A.G. are supported by an NCI grant (NCI P50CA70907).

Author Contributions E.D.P. undertook the development and implementation of bioinformatic algorithms for analysis of the sequencing data, assisted by A. Meynert, D.J., D.B., K.W.L., C.G., G.R.O., L.S., L.C., M.J., C. Leroy, J. Marshall, A. Menzies, A.B., J.W.T., J. Mangion, Y.A.S., S.F.M., H.E.P., E.F.T., G.L.C., C.C.L., E.B., M.D.R., K.J.M. and P.J.C. P.J.S., S.O.M. and D.J.M. were responsible for generating libraries and running sequencers, together with downstream validation analyses, assisted by M.-L.L., I.V., S.N.-Z., H.R.D., L.J.M., C. Latimer and S.E. J.D.M. and A.G. generated the cell lines. E.D.P., M.R.S., P.A.F. and P.J.C. directed the research and wrote the manuscript, which all authors have approved.

Author Information Genome sequence data have been deposited at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00000000051. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.J.C. (pc8@sanger.ac.uk).

METHODS

Power calculations. To estimate the levels of coverage required for accurate detection of somatically acquired base substitutions, we make the following assumptions. (1) Tumour DNA is approximately diploid, with 3.2 Gb per cell. (2) All reads can be mapped back correctly to the genome, even if they have a mutation, SNP or sequencing error. (3) Mutations and SNPs are heterozygous (homozygous variants are much easier to detect accurately). (4) Mutation prevalence 1 per Mb; SNP prevalence 1 per kb; sequencing error rate averaged 1 per 100 bases, with moderate degrees of systematic error (allowing different bases in the genome to have different error rates); overdispersion 2.0 (ratio of the variance of coverage as reads per base exceeds the mean reads per base). (5) We use the following strategy for calling a mutation at a given base, which is covered by C_T reads in the tumour DNA and C_N reads in the normal DNA: of the C_T reads in tumour DNA, the substitution is seen in a greater number of reads than the other two possible substitutions, with at least as many reads as a particular threshold, $\kappa(C_T)$; of the C_N reads in the normal DNA, the number of reads containing a particular substitution must be fewer than a threshold, $\kappa(C_N)$. For example, for a base covered 10× in the tumour and normal, we might require at least four reads of the tumour to have a variant base call which was not seen in any of the reads from the normal. The optimal choice of strategy (that is, the set of $\kappa(C_T)$ and $\kappa(C_N)$ for all C_T and C_N) will depend on the sequencing error rates, systematicity of sequencing errors, level of contamination, mutation rate and SNP rate. For these calculations, the optimal strategy for each set of conditions was defined as the set of $\kappa(C_T)$ and $\kappa(C_N)$ which minimized the miscall rate (either failing to call a mutation or calling a SNP or non-variant base a mutation).

The probability of calling a true mutation and the probabilities of mis-calls of sequencing errors and SNPs as mutations are calculated for each level of coverage of tumour and normal genomes, based on the $\kappa(C_T)$ and $\kappa(C_N)$ thresholds above, a binomial distribution of alleles (for heterozygous variants) and sequencing errors following a beta-binomial distribution. The distribution of coverages for the genome is then modelled by a negative binomial distribution to account for overdispersion, and the overall true- and false-positive rates summed.

Sequencing. Genomic DNA was used to construct mate-pair libraries of different insert sizes, the smallest insert libraries were from 600 bp to 800 bp and the largest were 3–4 kb. Library preparation, emulsion PCR, slide preparation and sequencing were all performed according to the manufacturer's protocol (Applied Biosystems SOLiD Library Preparation Protocol). Sequencing generated 25-bp tags in colour space from each end of the DNA fragment. Primary data analysis including image analysis and base-calling was carried out with the corona pipeline (Applied Biosystems). Alignment of reads to the NCBI36 reference genome (translated into colour space) was performed with corona-lite (version 0.32), allowing two mismatches in mapping and up to four in the pairing step. Duplicate read pairs with identical coordinates were removed, and only uniquely mapping reads were used for analysis.

Substitution detection. The corona-lite pipeline (version 0.32) was used to generate a preliminary list of variant bases from the uniquely aligning reads. We used the optimal thresholds defined in point 5 of the power calculations above (based on a mutation prevalence of 8 per Mb, as estimated from capillary sequence data in COSMIC) to determine whether there was sufficient evidence for calling a somatic substitution or not at each base in this preliminary list. Resulting tumour-specific substitutions were further filtered to remove (1) those residing in regions of loss of heterozygosity (LOH) in the normal cell line; (2) those potentially due to misalignment in segmental duplications and near sequence gaps; (3) those corresponding to polymorphic positions in dbSNP; (4) those potentially due to misalignment or miscalls as they are adjacent to SNPs or within 5 bp of insertions and deletions; and (5) those where all supporting reads contained the putative variant in the first or last 5 bp of the read (to reduce effects of misalignment across indels). Substitutions were annotated using Ensembl version 52.

The bioinformatic algorithm initially compares the tumour genome against the published reference genome to identify variants, before comparing it against the EBV-transformed line to determine whether the variant is somatically acquired or germ line. Thus, a mutation in the B-cell line induced by EBV transformation will not be falsely detected as a variant in the SCLC line. However, one potential artefact resulting from changes induced by EBV transformation is in regions where there is loss of heterozygosity—this could potentially lead to false mutation calls in the tumour because the EBV line contains only one allele at heterozygous SNP positions. For this reason, we excluded regions of LOH in the EBV line from the mutation-calling algorithm (filter criterion (1) above).

LOH in the normal cell line was determined by analysis of the EBV-transformed B-cell line on an Affymetrix SNP6 array, using hidden Markov model algorithms for identification of allele-specific copy number.

Insertion and deletion detection. Small insertions (up to 3 bp) and deletions (up to 11 bp) were called using corona-lite (version 0.4). Indels found in the tumour and not in the normal were further filtered to require (1) minimum three supporting tumour reads; (2) minimum one read on each strand; (3) no LOH in the normal; (4) maximum 100× coverage (to remove regions of misalignment); and (5) minimum 30× normal coverage (to reduce the number of germline indels in the set).

Structural variant detection. Abnormal read pairs that mapped to the genome with MAQ at an unexpected distance or orientation were identified, grouped and filtered as described[14]. To define somatic variants, at least eight independent read pairs were required in the tumour, and a maximum number in the normal were allowed as defined based on the same optimal thresholds used in substitution calling.

Confirmation by capillary sequencing. A set of 79 novel coding mutations and 354 randomly chosen, genome-wide substitutions predicted by the algorithm, and 262 small insertions and deletions, were taken forward for independent validation by capillary sequencing across the region of the mutation. Structural variants were confirmed by PCR across the breakpoint and capillary sequencing the breakpoint. All confirmations were done in both the tumour and normal to determine if the variants were somatic or germ line.

Copy number determination. Reads were counted in windows across the genome, corrected for genome uniqueness as described[14]. Circular binary segmentation was used to segment the data, using a Bioconductor package originally designed for array-CGH data[41]. The adaptation of the algorithm for shotgun sequencing data has been described in detail elsewhere[14].

Substitution analysis. For mutation context, the bases ±10 bp from the mutation were extracted and the number of each base counted. Context of equivalent changes (for example, C>T and G>A) were combined. The background context was determined from 200,000 randomly sampled bases from the genome at positions in which sufficient coverage of the tumour and normal genomes was achieved. Strand bias was calculated based on annotating each mutation as to whether it fell on the transcribed or untranscribed strand in Ensembl 52. Gene expression data were derived from the Affymetrix U133A array. Chi-squared testing was performed to detect departures of observed mutation context from that expected in the covered genome. Poisson regression was used for the analysis of the effects of gene expression on mutation prevalence, incorporating the number of at-risk bases in each gene footprint as the offset, allowing quadratic terms for the relationship between expression and mutation prevalence, and using a dummy variable for transcribed versus non-transcribed strand mutations. The image of the entire NCI-H209 genome was produced using Circos[42].

Analysis of promoter mutations. The promoter region of each gene was defined as the 2,000 bp each side of the transcription start site (TSS) for its representative transcript. Some of the promoters defined this way overlapped, giving 63.9 Mb of unique sequence. We mapped the substitutions to the gene with the nearest TSS, and confirmed that the reference bases provided matched the corresponding bases in the reference genome. We investigated the impact of the somatic substitutions on transcription factor binding to the promoter regions using a Sunflower model (M. M. Hoffmann and E. Birney, submitted) based on the JASPAR databases'[43] set of vertebrate transcription factors. The database contains two matrices for the transcription factor Lhx3: one based on human data, and one from mouse. The mouse factor was excluded and the remaining 100 matrices used to build the model with an unbound prior probability of 0.99; leaving 1×10^{-4} for each factor. The G+C content was close to 50% for the promoters for both lines, so we used a uniform distribution for the unbound emission probabilities (A = C = G = T = 0.25).

The reference sequences for each promoter, plus 300 bp of padding sequence on each end (4,600 bp per promoter in total), were downloaded from Ensembl 53. The padding sequence is necessary for two reasons. First, the posterior probability decoding of Sunflower is unreliable near the beginning and end of a sequence (M. M. Hoffmann and E. Birney, submitted). Second, if a transcription factor binding site overlaps the beginning or end of the sequence, it cannot be correctly modelled as only part of its state loop will be traversed.

For the set of observed substitutions, we generated the variant promoters from the reference sequences. We also randomly sampled the unobserved substitutions according to the mutation distributions of the observed substitutions. As a further control, these were restricted to the same mutation type per promoter as the observed substitutions. For example, if a promoter had a C>T substitution at position −160, the set of possible unobserved substitutions for that promoter would be all of the other possible C>T and G>A substitutions. If a promoter had multiple observed substitutions, any of those types were allowed in the sampling. Variant sequences for 1,000 sets of random unobserved substitutions were generated. The set size was the same as the observed set size, for a total of 315,000 unobserved substitutions. Using the pairwise comparison functionality of Sunflower and the 100-factor JASPAR model, we compared the variant

sequences to the reference sequences, obtaining a relative entropy value for each. To ensure that the results were related to real binding sites and not random noise, we randomly shuffled the columns of all the JASPAR position weight matrices (PWMs) to create new sets of matrices. From 100 such sets, we built Sunflower models using the same parameters as for the real JASPAR data, and compared the variant promoter sequences to their references for the observed substitutions.

***CHD7* analysis.** Multiplex ligation-dependent probe amplification (MLPA) for *CHD7* was performed using the commercially available assay (MRC-Holland b.v.) according to the manufacturer's instructions. Identification of rearrangements in LU-135 was performed using a 3–4-kb mate-pair library sequenced on the SOLiD platform, as described above. Rearrangements were confirmed by PCR and capillary sequencing across the breakpoint. RT–PCR across aberrant exon junctions was used to assess expression of the *PVT1–CHD7* fusion transcripts (NCI-H2171 and LU-135) and the internal duplication of exons 3–8 (NCI-H209). RT–PCR products were capillary sequenced to confirm their veracity. Gene expression data were derived from the Affymetrix U133A array for the core cell lines listed on our website (http://www.sanger.ac.uk/genetics/CGP/CellLines/).

41. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23,** 657–663 (2007).
42. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).
43. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34,** D95–D97 (2006).