

The mutation spectrum revealed by paired genome sequences from a lung cancer patient

William Lee¹, Zhaoshi Jiang¹, Jinfeng Liu¹, Peter M. Haverty¹, Yinghui Guan², Jeremy Stinson², Peng Yue¹, Yan Zhang¹, Krishna P. Pant³, Deepali Bhatt², Connie Ha², Stephanie Johnson⁴, Michael I. Kennemer³, Sankar Mohan⁵, Igor Nazarenko³, Colin Watanabe¹, Andrew B. Sparks³, David S. Shames⁵, Robert Gentleman¹, Frederic J. de Sauvage², Howard Stern⁴, Ajay Pandita⁵, Dennis G. Ballinger³, Radoje Drmanac³, Zora Modrusan², Somasekar Seshagiri² & Zemin Zhang¹

Lung cancer is the leading cause of cancer-related mortality worldwide, with non-small-cell lung carcinomas in smokers being the predominant form of the disease^{1,2}. Although previous studies have identified important common somatic mutations in lung cancers, they have primarily focused on a limited set of genes and have thus provided a constrained view of the mutational spectrum^{3–8}. Recent cancer sequencing efforts have used next-generation sequencing technologies to provide a genome-wide view of mutations in leukaemia, breast cancer and cancer cell lines^{9–13}. Here we present the complete sequences of a primary lung tumour (60× coverage) and adjacent normal tissue (46×). Comparing the two genomes, we identify a wide variety of somatic variations, including >50,000 high-confidence single nucleotide variants. We validated 530 somatic single nucleotide variants in this tumour, including one in the *KRAS* proto-oncogene and 391 others in coding regions, as well as 43 large-scale structural variations. These constitute a large set of new somatic mutations and yield an estimated 17.7 per megabase genome-wide somatic mutation rate. Notably, we observe a distinct pattern of selection against mutations within expressed genes compared to non-expressed genes and in promoter regions up to 5 kilobases upstream of all protein-coding genes. Furthermore, we observe a higher rate of amino acid-changing mutations in kinase genes. We present a comprehensive view of somatic alterations in a single lung tumour, and provide the first evidence, to our knowledge, of distinct selective pressures present within the tumour environment.

Most lung cancer cases occur in patients with a history of smoking¹⁴. A non-small-cell lung cancer from a 51-year-old male Caucasian who reported smoking an average of 25 cigarettes per day for 15 years before tumour excision was used in this study. The tumour was characterized by pathology as a poorly differentiated sample with 95% tumour content and focal gland formation, and was TTF1-positive, KRT5-negative by immunohistochemistry, all of which supports a diagnosis of adenocarcinoma^{15,16} (Supplementary Section 1). Substantial data now suggest that smoking-related lung cancer develops through the step-wise accrual of genetic lesions that increase in frequency and size through the progression from cellular hyperplasia to frank malignancy. As such, there are several well-defined genomic alterations that occur with significant frequency in smoking-related non-small-cell lung carcinoma including allelic and copy number losses in 3p, 5q, 6p, 8p, 15q, 17p and 18q, as well as gains in 1q, 8p and Xq^{2,17}. We used a single nucleotide polymorphism (SNP) array and comparative genomic hybridization (CGH)

to identify genomic regions containing allelic imbalance, loss of heterozygosity and copy number variants (Supplementary Section 2). The overall pattern of large copy number alteration seen in this tumour sample (Fig. 1 and Supplementary Section 2) is consistent with these well described alterations, as well as many of those found recently⁸. Examples include copy number loss of *TP53* (Supplementary Fig. 1), *DR4* and *DR5* (also called *TNFRSF10A* and *TNFRSF10B*, respectively), gains of *CDK4* and *KRAS* (Supplementary Fig. 2), and copy-neutral loss of heterozygosity of chromosome 13 including *RB1* (Supplementary Fig. 3). Thus, this tumour sample bears many of the hallmark copy number alterations commonly found in smoking-associated lung cancer.

Sequencing of the samples was performed using unchained combinatorial probe anchor ligation (cPAL) chemistry on arrays of self-assembling DNA nanoballs (DNBs)¹⁸ (Supplementary Section 3), resulting in a total of 171.25 gigabases (Gb) of mapped sequence (60× average coverage) for the tumour sample and 131.3 Gb (46× coverage) for the matched normal (Table 1). Reads were aligned to the reference genome (National Center for Biotechnology Information (NCBI) Build 36) and variants were called and scored using a local *de novo* assembly approach¹⁸. High variant call accuracy was confirmed by genotyping (Supplementary Section 2). Regions that could not be called in one or both genomes accounted for 12.86% of the total reference genome and 8.4% of protein-coding regions. We identified widespread somatic variations throughout the genome and on all scales (Fig. 1), including point mutations (Supplementary Tables 1 and 2), small insertions and deletions (Supplementary Section 4, Supplementary Fig. 4 and Supplementary Table 3) and somatic structural variations (Supplementary Table 4). Examination of sequence read coverage confirmed gains of *KRAS* and *CDK4* as well as loss of *TP53* (Supplementary Fig. 5).

Single-nucleotide variants (SNVs) were called independently for both the tumour and the normal genomes and filtered to obtain over 83,000 candidate new somatic mutations (Supplementary Section 3 and Supplementary Fig. 6). These include 540 candidate non-synonymous mutations and 195 synonymous mutations in the protein-coding regions of the genome. Using nucleic acid mass spectrometry we were able to validate 70% of the predicted protein-coding region changes including 302 (of 418 tested) non-synonymous and 90 (of 143 tested) synonymous mutations (Supplementary Tables 1 and 2). A large fraction of the candidate SNVs that were not validated were only partially called in the normal sample or were possibly a result of low frequency in the tumour (Supplementary Section 3). Comparing our results to those

¹Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, California 94080, USA. ²Department of Molecular Biology, Genentech Inc., South San Francisco, California 94080, USA. ³Complete Genomics Inc., Mountain View, California 94043, USA. ⁴Department of Pathology, Genentech Inc., South San Francisco, California 94080, USA. ⁵Department of Oncology Diagnostics, Genentech Inc., South San Francisco, California 94080, USA.

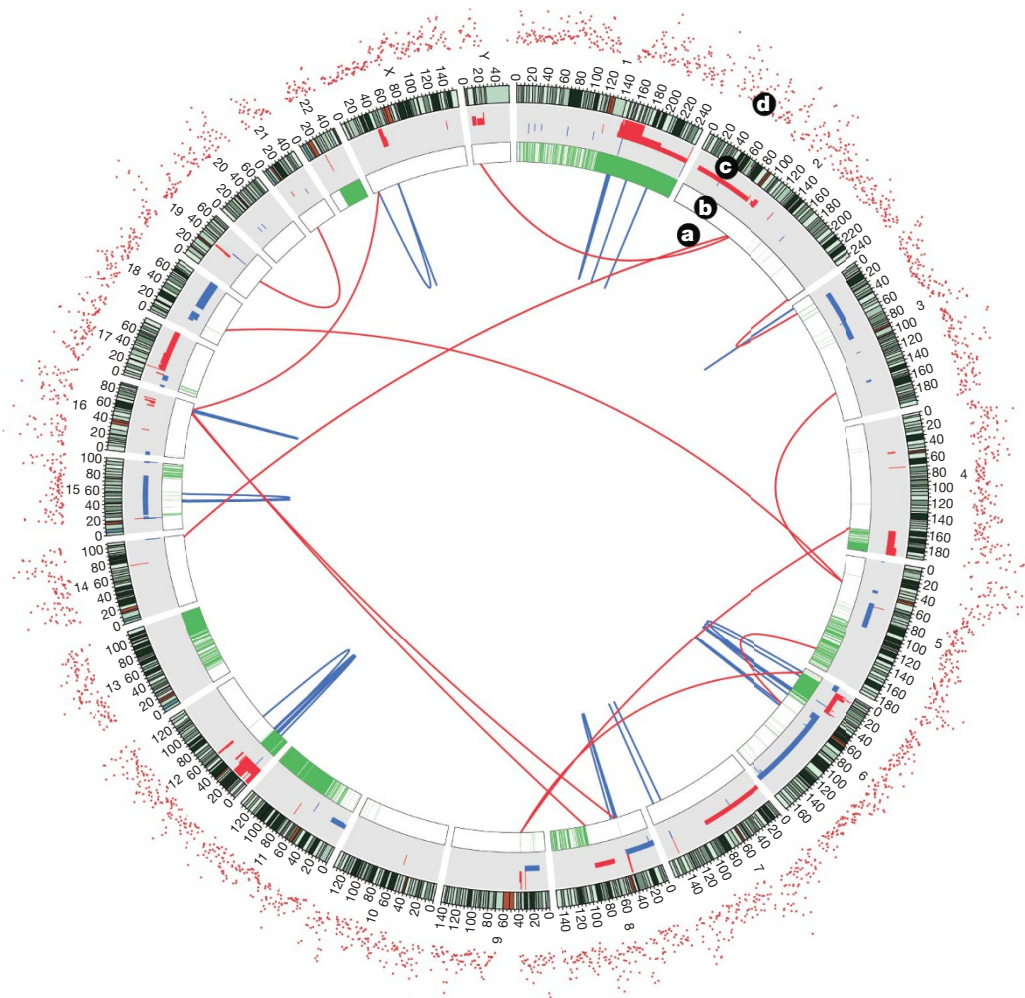


Figure 1 | The genomic landscape of somatic alterations. **a–d**, Various types of genomic profiles of the adenocarcinoma sample in this study. **a**, Experimentally confirmed somatic structural variations. Red lines indicate interchromosomal structural variations whereas blue lines represent intrachromosomal structural variations. **b**, Regions of loss of heterozygosity and allelic imbalance are in green and were based on the

Affymetrix SNP 6.0 array data. **c**, Copy number profiles were derived from the Agilent array data with red indicating copy number gain and blue representing copy number loss (scale range: 0 to 4 copies). **d**, Each red dot represents the number of high-confidence somatic SNVs in a 1 Mb window. This figure was created using the Circos program²⁸.

from sequencing surveys performed previously^{5,19,20}, we found *KRAS* Gly12Cys to be the only previously known mutation, although we identified new non-synonymous mutations in 13 genes previously observed to be mutated in a large-scale study of lung adenocarcinoma, including frequently mutated genes *LRP1B* and *NF1* (Supplementary Table 1)⁵. Protein sequence analysis using bioinformatic tools was also used to evaluate the potential effect of the validated somatic variations (Supplementary Section 5 and Supplementary Table 1). Through the use of several different computational methods, this approach showed that somatic mutations as a group have a significantly different score distribution compared to germline variations and are thus more likely to affect function (Supplementary Fig. 7). Notable among the somatic alterations in protein-coding genes is the Asp194His change in *NEK9*

that is located within a frequently mutated region in the protein kinase domains (Asp-Phe-Gly motif). The Asp residues in the analogous location in other protein kinases are frequently mutated in many diseases (Supplementary Fig. 8). Interestingly, a previous study of lung carcinoma samples also identified a mutation in *NEK9*, although not in the same position⁶. Furthermore, several genes with multiple protein-altering mutations were identified within this sample, including four distinct non-synonymous somatic mutations in *MUC16*, a known ovarian cancer antigen¹⁹.

We then tested a further 231 non-coding SNVs (Supplementary Table 2) to estimate the sensitivity of our method. We developed a score for each mutation that represents its likelihood to be a true somatic variation (Supplementary Section 6) and set a cutoff score that provided 90% precision and 82% sensitivity among the set of validated mutations. Further genome-wide analysis was performed with the set of 50,675 high-confidence mutations that exceeded this score threshold (Supplementary Table 5).

The composition of somatic variations in this sample is distinct from that of germline variations. Across the entire genome, somatic variations occurred predominantly at G•C base pairs (78%), the most prevalent changes being G•C → T•A transversions (46%). A similar pattern was previously observed in a small-cell lung cancer cell line¹³. We find that this enrichment for G•C → T•A transversions is a genome-wide trend for somatic variations and forms a sharp

Table 1 Sequence coverage summary for normal and tumour genomes		
	Normal	Tumour
Mapped sequence (Gb)	131.30	171.25
Average haploid coverage	46×	60×
Per cent of genome covered	99.4	99.2
Per cent of genome with 10× or greater coverage	92.0	91.1
Per cent of genome called	91.1	89.1
SNVs	2,952,347	2,720,041

Coverage percentage and variations are with respect to NCBI Build 36 of the human genome reference assembly.

contrast to germline variations, where $A \cdot T \rightarrow G \cdot C$ and $G \cdot C \rightarrow A \cdot T$ transitions account for 69% of the variations (Fig. 2a). This direct comparison between germline and somatic variations underscores a strong influence of smoking-related DNA damage¹⁴. Furthermore, $G \cdot C \rightarrow C \cdot G$ somatic changes are strongly enriched at GpA/TpC dinucleotides, with A accounting for 52% of the nucleotides after $G \rightarrow C$ variations. This dinucleotide mutational preference, also observed in coding regions across many cancer types⁶, is not observed in the germline collection (chi-squared test, $P < 1 \times 10^{-16}$; Supplementary Fig. 9).

Previous studies have estimated a wide range of mutation rates in lung tumours, but those estimates are primarily based on mutations observed within subsets of protein-coding genes^{5,6}. The 50,675 high-confidence somatic SNVs in this sample translate into an approximate rate of 17.7 mutations per megabase (Mb) throughout the genome. A previous study sequenced 623 genes in 188 lung tumours

and observed up to 49 mutations in a single smoker sample whereas non-smokers had fewer than five mutations⁵. We identified 17 mutations in our sample in the same set of genes, placing our sample within the range of previously observed lung-tumour mutation rates.

The distribution of somatic mutations in the genome is highly non-uniform, as protein-coding exons have a substantially lower rate of 12.5 per Mb ($P = 7 \times 10^{-13}$). Furthermore, we found that expressed genes were much less likely to have non-synonymous mutations ($P = 2.2 \times 10^{-5}$, Fisher's exact test; Supplementary Fig. 10) and the set of expressed genes has a much lower mutation rate of 8.3 per Mb (including introns and untranslated regions (UTRs)). Genes that are not expressed, on the other hand, are mutated in a manner similar to the genome-wide average (Fig. 2b and Supplementary Section 7). The transcribed strand has recently been shown to have lower rates of mutation compared to the non-transcribed strand in a small-cell cancer cell line¹³. We observe a similar trend in our sample (Supplementary Fig. 11), with a much more pronounced strand bias in expressed genes (Fig. 2b). This is consistent with the pattern that would result from transcription-coupled DNA repair processes. The ratio of overall non-synonymous substitution rate (K_a) to synonymous substitution rate (K_s) in this sample is 0.97, suggesting that most mutations are 'passengers'. However, we observe a significantly lower K_a/K_s ratio in expressed genes compared to non-expressed genes (0.81 versus 1.26), suggesting selective pressure against protein-altering mutations in expressed genes. In contrast, within the kinase genes we observed ten non-synonymous mutations and only a single synonymous mutation, resulting in a high K_a/K_s ratio. This indicates a potential selective advantage for non-synonymous mutations in kinases.

Looking beyond protein-coding regions of the genome, we identified somatic variations in four RNA genes (*tRNA^{Tyr}-GTA*, *MIR598*, *SNORD12C* and *SNORA74B*), 397 variations in 331 pseudogenes, 649 variations in conserved transcription factor binding sites, and none in predicted microRNA target sites (Supplementary Section 6). We observed substantially lower mutation rates in regions immediately upstream from transcription start sites (Fig. 2c). Within 2 kilobases (kb) upstream of transcription start sites, the mutation rate is 10.5 per Mb, 40% lower than the genome-wide average ($P < 1 \times 10^{-16}$, Fisher's exact test). These regions have higher GC content (mean = 49.3%, compared with 40.9% for the entire genome), but even after accounting for the coverage effects of GC content we still observed a significantly lower somatic mutation rate in the regions upstream from transcription start sites (Supplementary Fig. 12). This suggests that these regions are also under purifying selection similar to exonic sequence, possibly due to their important role in regulating transcription. This observation in combination with the lower mutation rates observed in protein-coding and expressed genes is indicative of global selective pressures against alterations in specific genomic regions.

Recurrent genomic rearrangements are characteristic features of many human cancers^{3,8,21–23}. Gene translocations, such as *MLL4-ALK*, have been shown to be prevalent in non-small-cell lung cancer²⁴ as well as a range of other cancer types^{25–27}. We utilized the mate-paired nature of the sequence library to identify potential large-scale structural variations (see Supplementary Section 8 and Supplementary Fig. 13 for details). This analysis resulted in 344 putative structural variations, 79 of which were not present with the required coverage in the normal sample and were therefore designated as somatic candidates. To validate these predicted somatic structural variations, we amplified (by polymerase chain reaction (PCR)) and sequenced the 65 cases that had sufficiently unique sequences within their read cluster regions for PCR primer design (Supplementary Section 9 and Supplementary Table 6). These sequences were then mapped to the reference genome to confirm the breakpoints at single nucleotide resolution (Supplementary Section 9). Of these 65 cases, 43 (66%) were found to be somatic changes owing to their strong and specific PCR products in the tumour sample but not in the matched normal sample (Supplementary Fig. 14 and Supplementary Table 4).

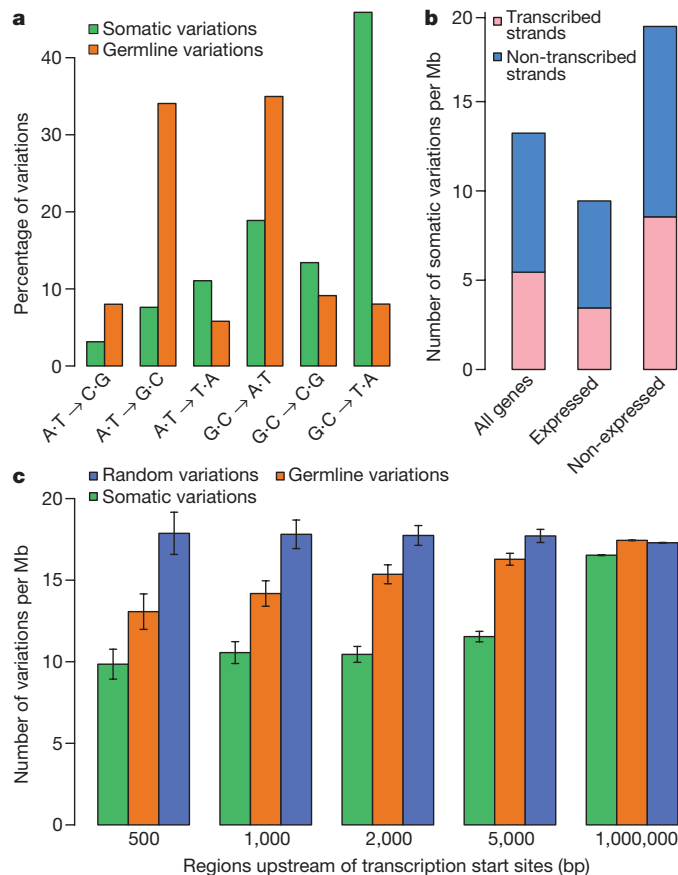


Figure 2 | Somatic single-nucleotide mutation trends and patterns.

a, Somatic mutations are primarily $G \cdot C \rightarrow T \cdot A$ transversions. Distribution of specific nucleotide changes among germline and somatic variations in the lung genome. $G \cdot C \rightarrow T \cdot A$ transversions account for 46% of high-confidence somatic mutations, whereas most germline variations are $A \cdot T \rightarrow G \cdot C$ or $G \cdot C \rightarrow A \cdot T$ transitions. **b**, Expressed genes have a lower mutation rate than non-expressed genes. Genes that are expressed in the tumour sample, as determined by microarray data, have a mutation rate of 8.3 per Mb (including introns and 3' and 5' UTRs) that is substantially lower than the mutation rate of 17.5 per Mb observed in unexpressed genes. Mutation rates in transcribed strands (pink bars) are lower than those in non-transcribed strands (blue bars). **c**, Promoters are depleted for somatic mutations. To obtain the mutation rates in the promoter regions, we examined the 50,675 high-confidence somatic variations and the same number of randomly selected germline and simulated mutations, and calculated the number of variations per Mb in the regions immediately upstream of transcription start sites. In regions up to 5 kb upstream of transcription start sites, there are significantly fewer somatic mutations than germline variations or random variations. Error bars represent standard deviation of the mutation rates from 1,000 random samplings.

Until recently, it has not been possible to interrogate the complete mutational spectrum of a tumour. Thus, it has been difficult to interpret how the various changes within a single tumour may work together to fulfil the hallmark traits of the malignant phenotype. Our results show that lung cancers can harbour large numbers of new mutations from the single-nucleotide level up to chromosome-scale alterations. Indeed, we found that at least eight genes in the

METHODS SUMMARY

Validation of candidate SNVs was performed using the Sequenom MassARRAY platform with standard protocols (Supplementary Section 3). Structural variations were validated by PCR and Sanger sequencing (Supplementary Section 9). Further validation of two specific structural variations was performed by FISH using bacterial artificial chromosome (BAC)-based probes (Supplementary Section 9).

Received 7 December 2009; accepted 10 March 2010.

1. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer J. Clin.* **55**, 74–108 (2005).
2. Herbst, R. S., Heymach, J. V. & Lippman, S. M. Lung cancer. *N. Engl. J. Med.* **359**, 1367–1380 (2008).
3. 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).
4. Davies, H. *et al.* Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res.* **65**, 7591–7595 (2005).
5. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).
6. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
7. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
8. Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007).

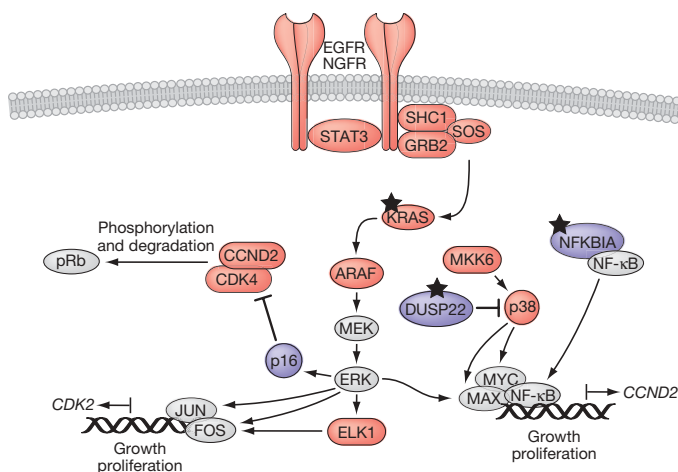


Figure 3 | A model for how the multiplicity of mutations within the MAPK cascade may act together to drive constitutive pro-growth signalling. Red shapes indicate amplification, purple ovals indicate loss of heterozygosity and/or deletion, black stars indicate mutation, grey ovals indicate no detected changes. The tumour harbours an activating point mutation in *KRAS* as well as copy number gains in *KRAS* and *EGFR*. Furthermore, there are high-level copy number gains of *SHC1*, *GRB2*, *SOS*, *ARAF*, *MAP3K3* and *ELK1*, suggesting that there are at least eight potentially activating genetic lesions within this particular pathway²⁹. MAP3K3 (not displayed in the figure) can act as a branch point between MAPK and SAPK signalling. This tumour exhibits multiple activating signals via high-level copy number gains within the p38 pathway including *MKK6* and *p38* itself. Also, there is a point mutation in *DUSP22*, a negative regulator of p38 signalling. ERK and p38 directly affect the transition from G1 to S phase by transcriptionally activating *MYC* and *MAX*, which regulate cyclin D2 (*CCND2*) transcription. The p38 cascade activates *NFKB*, which also activates *CCND2* transcription. This tumour harbours a potentially inactivating mutation in *NFKBIA*, which normally prevents NF- κ B from entering the nucleus. Active p38 signal transduction plus loss of *NFKBIA* could lead to aberrant activation of *CCND2* transcription via *MYC/MAX* and NF- κ B. Furthermore, activated *ELK1* (via MAPK) leads to transcription of *FOS*. *MYC/MAX* and *FOS/JUN* (AP1) collaborate to transcribe *CDK4*. Cyclin D2 binds to and activates *CDK4*, a complex regulated by the cyclin-dependent kinase inhibitor 4A (p16). This sample showed copy number losses on 9p21, which contains the tumour suppressor gene *p16*. Thus, this tumour has multiple activating and inactivating hits that may drive oncogenic signal transduction through MAPK and related pathways, thereby overwhelming the G1/S cell cycle checkpoint and leading to unregulated cellular proliferation.

9. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
10. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
11. Shah, S. P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
12. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2009).
13. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
14. Hecht, S. S. Tobacco smoke carcinogens and lung cancer. *J. Natl Cancer Inst.* **91**, 1194–1210 (1999).
15. Chu, P. G. & Weiss, L. M. Expression of cytokeratin 5/6 in epithelial neoplasms: an immunohistochemical study of 509 cases. *Mod. Pathol.* **15**, 6–10 (2002).
16. Tan, D. *et al.* Thyroid transcription factor-1 expression prevalence and its clinical implications in non-small cell lung cancer: a high-throughput tissue microarray and immunohistochemistry study. *Hum. Pathol.* **34**, 597–604 (2003).
17. Wistuba, I. I. & Gazdar, A. F. Lung cancer preneoplasia. *Annu. Rev. Pathol.* **1**, 331–348 (2006).
18. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
19. Forbes, S. A. *et al.* The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* doi:10.1002/0471142905.hg1011s57 (2008).
20. Stenson, P. D. *et al.* The human gene mutation database: 2008 update. *Genome Med.* **1**, 13 (2009).
21. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res.* **16**, 1465–1479 (2006).
22. Beroukhi, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104**, 20007–20012 (2007).
23. Bignell, G. R. *et al.* Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* **17**, 1296–1303 (2007).
24. Soda, M. *et al.* Identification of the transforming *EML4–ALK* fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566 (2007).
25. Lin, E. *et al.* Exon array profiling detects *EML4–ALK* fusion in breast, colorectal, and non-small cell lung cancers. *Mol. Cancer Res.* **7**, 1466–1476 (2009).
26. Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–293 (1973).
27. Tomlins, S. A. *et al.* Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
28. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
29. Dhillon, A. S., Hagan, S., Rath, O. & Kolch, W. MAP kinase signalling pathways in cancer. *Oncogene* **26**, 3279–3290 (2007).

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

Acknowledgements We thank T. Wu for critical reading of manuscript, C. Santos for sample handling, M. Vasser and the DNA Synthesis Group for oligonucleotide synthesis, J. Turcotte and G. Cavet for coordination, G. Nilsen for data submission, J. Fitzgerald and A. Baucom for data storage, J. Lee for laboratory support, A. Bruce for graphical assistance, and T. Bhangale, S. Jhunhunwala and A. Halpern for discussion.

Author Contributions W.L., project coordination, SNV and overall data analysis and preparation of manuscript; Z.J., structural variation analysis and preparation of manuscript; J.L., mutation pattern and trend analysis, loss of heterozygosity analysis, expression analysis and preparation of manuscript; P.M.H., copy number/loss of heterozygosity analysis, pathway analysis, expression analysis and preparation of manuscript; P.Y., mutation analysis and preparation of manuscript; Y.G. and Z.M., PCR validation of structural variations; J.S., D.B. and S.S., MassArray mutation validation; Y.Z., bioinformatic prediction of mutations and data processing; K.P.P., M.I.K., I.N. and A.B.S., DNA nanoball preparation and sequencing, base calling, quality control and structural variation mapping; C.H. and Z.M., microarray data production; S.J. and H.S., sample handling and pathology analysis; C.W., structural variation breakpoint mapping; D.S.S., pathway analysis and data interpretation; R.G., manuscript critiques and statistical analysis; F.J.d.S., project coordination and manuscript commenting; A.P. and S.M., FISH analysis; R.D. and D.G.B., project coordination, data interpretation and manuscript commenting; Z.Z., project design, data interpretation and preparation of manuscript.

Author Information Sequence data has been submitted to the NCBI Short Read Archive under accession number SRA012097. Microarray data has been submitted to the NCBI Gene Expression Omnibus under accession number GSE20585. 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 Z.Z. (zemin@gene.com).