LETTERS

Characterizing the cancer genome in lung adenocarcinoma

Barbara A. Weir[1,2]*, Michele S. Woo[1]*, Gad Getz[2]*, Sven Perner[3,4], Li Ding[5], Rameen Beroukhim[1,2], William M. Lin[1,2], Michael A. Province[6], Aldi Kraja[6], Laura A. Johnson[3], Kinjal Shah[1,2], Mitsuo Sato[8], Roman K. Thomas[1,2,9,10], Justine A. Barletta[3], Ingrid B. Borecki[6], Stephen Broderick[11,12], Andrew C. Chang[14], Derek Y. Chiang[1,2], Lucian R. Chirieac[3,16], Jeonghee Cho[1], Yoshitaka Fujii[18], Adi F. Gazdar[8], Thomas Giordano[15], Heidi Greulich[1,2], Megan Hanna[1,2], Bruce E. Johnson[1], Mark G. Kris[11], Alex Lash[11], Ling Lin[5], Neal Lindeman[3,16], Elaine R. Mardis[5], John D. McPherson[19], John D. Minna[8], Margaret B. Morgan[19], Mark Nadel[1,2], Mark B. Orringer[14], John R. Osborne[5], Brad Ozenberger[20], Alex H. Ramos[1,2], James Robinson[2], Jack A. Roth[21], Valerie Rusch[11], Hidefumi Sasaki[18], Frances Shepherd[25], Carrie Sougnez[2], Margaret R. Spitz[22], Ming-Sound Tsao[25], David Twomey[2], Roel G. W. Verhaak[2], George M. Weinstock[19], David A. Wheeler[19], Wendy Winckler[1,2], Akihiko Yoshizawa[11], Soyoung Yu[1], Maureen F. Zakowski[11], Qunyuan Zhang[6], David G. Beer[14], Ignacio I. Wistuba[23,24], Mark A. Watson[7], Levi A. Garraway[1,2], Marc Ladanyi[11,12], William D. Travis[11], William Pao[11,12], Mark A. Rubin[2,3], Stacey B. Gabriel[2], Richard A. Gibbs[19], Harold E. Varmus[13], Richard K. Wilson[5], Eric S. Lander[2,17,26] & Matthew Meyerson[1,2,16]

Somatic alterations in cellular DNA underlie almost all human cancers[1]. The prospect of targeted therapies[2] and the development of high-resolution, genome-wide approaches[3–8] are now spurring systematic efforts to characterize cancer genomes. Here we report a large-scale project to characterize copy-number alterations in primary lung adenocarcinomas. By analysis of a large collection of tumours ($n = 371$) using dense single nucleotide polymorphism arrays, we identify a total of 57 significantly recurrent events. We find that 26 of 39 autosomal chromosome arms show consistent large-scale copy-number gain or loss, of which only a handful have been linked to a specific gene. We also identify 31 recurrent focal events, including 24 amplifications and 7 homozygous deletions. Only six of these focal events are currently associated with known mutations in lung carcinomas. The most common event, amplification of chromosome 14q13.3, is found in ~12% of samples. On the basis of genomic and functional analyses, we identify *NKX2-1* (NK2 homeobox 1, also called *TITF1*), which lies in the minimal 14q13.3 amplification interval and encodes a lineage-specific transcription factor, as a novel candidate proto-oncogene involved in a significant fraction of lung adenocarcinomas. More generally, our results indicate that many of the genes that are involved in lung adenocarcinoma remain to be discovered.

A collection of 528 snap-frozen lung adenocarcinoma resection specimens, with at least 70% estimated tumour content, was selected by a panel of thoracic pathologists (Supplementary Table 1); samples were anonymized to protect patient privacy. Tumour and normal DNAs were hybridized to Affymetrix 250K Sty single nucleotide polymorphism (SNP) arrays. Genomic copy number for each of over 238,000 probe sets was determined by calculating the intensity ratio between the tumour DNA and the average of a set of normal DNAs[9,10]. Segmented copy numbers for each tumour were inferred with the GLAD (gain and loss analysis of DNA) algorithm[11] and normalized to a median of two copies. Each copy number profile was then subjected to quality control, resulting in 371 high-quality samples used for further analysis, of which 242 had matched normal samples (Methods).

To identify regions of copy-number alteration, we applied GISTIC (genomic identification of significant targets in cancer)[12], a statistical method that calculates a score that is based on both the amplitude and frequency of copy-number changes at each position in the genome, using permutation testing to determine significance (Methods).

GISTIC identified 26 large-scale events and 31 focal events, reported below. Although the overall pattern is broadly consistent with the literature on lung cancer[8,13–15], our sample size and resolution provide more power to accurately identify and localize both large-scale and focal chromosomal alterations. With respect to large-scale events, no single previous study has identified more than 5 of the gains or 11 of the losses[13,14] (Supplementary Table 2). With respect to focal events, three recent studies[8,14,15] report a total of ~200 events, including 23 of the 31 recurrent focal events observed in our study. The overlap among these three studies is limited to only four events (amplification of *EGFR*, *CCNE1*, *MDM2* and 8p11, all seen

[1]Department of Medical Oncology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [2]Cancer Program, Genetic Analysis Platform, and Genome Biology Program, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [4]Institute of Pathology, University of Ulm, Ulm 89081, Germany. [5]Genome Sequencing Center, [6]Division of Statistical Genomics and [7]Department of Pathology and Immunology, Washington University in Saint Louis, Saint Louis, Missouri 63130, USA. [8]University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [9]Max Planck Institute for Neurological Research with Klaus-Joachim-Zülch Laboratories of the Max-Planck Society and the Medical Faculty of the University of Cologne, Cologne 50931, Germany. [10]Center for Integrated Oncology and Department I for Internal Medicine, University of Cologne, Cologne 50931, Germany. [11]Departments of Medicine, Surgery, Pathology, and Computational Biology, [12]Human Oncology and Pathogenesis Program, [13]Cancer Biology and Genetics Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [14]Section of Thoracic Surgery, Department of Surgery and [15]Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109, USA. [16]Department of Pathology and [17]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [18]Department of Surgery, Nagoya City University Medical School, Nagoya 467-8602, Japan. [19]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [20]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [21]Department of Thoracic and Cardiovascular Surgery, [22]Department of Epidemiology, [23]Department of Pathology and [24]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA. [25]University Health Network and Princess Margaret Hospital, Toronto M5G 2C4, Canada. [26]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.
*These authors contributed equally to this work.

here; Supplementary Table 3 and Supplementary Results). A genome-wide view of segmented copy number reveals that most chromosomal arms undergo either amplification or deletion across a large proportion of the samples (Fig. 1a). The distinctive pattern of amplification and loss is also apparent when the median copy number for each chromosome arm is plotted (Supplementary Fig. 1 and Supplementary Table 4). In total, GISTIC identifies 26 large segments (at least half of a chromosome arm), 10 with significant gains and 16 with significant losses (Fig. 1b and Supplementary Table 5).

Visual inspection reveals that similar chromosomal patterns of copy number loss and gain across the genome are found in almost all samples, but that the samples show substantial differences in the amplitude of copy-number variation (Fig. 1a). When the samples are partitioned into tertiles on the basis of overall variation in copy-number amplitude, each shows similar regions of amplification and loss across the genome. The attenuation seen in many samples is consistent with admixture with euploid non-tumour DNA, which we estimate at 50%, 65% and 78% respectively in the three tertiles (Supplementary Results and Supplementary Fig. 2). The significant non-tumour admixture in these tumour samples also makes it difficult to assess genome-wide loss of heterozygosity (LOH). Because normal DNA admixture limits sensitivity, we report LOH only in the top tertile; we see both LOH associated with copy-number loss and copy-neutral LOH (chromosomes 17p and 19p) (Supplementary Results, Supplementary Figs 3 and 4, and Supplementary Table 6).

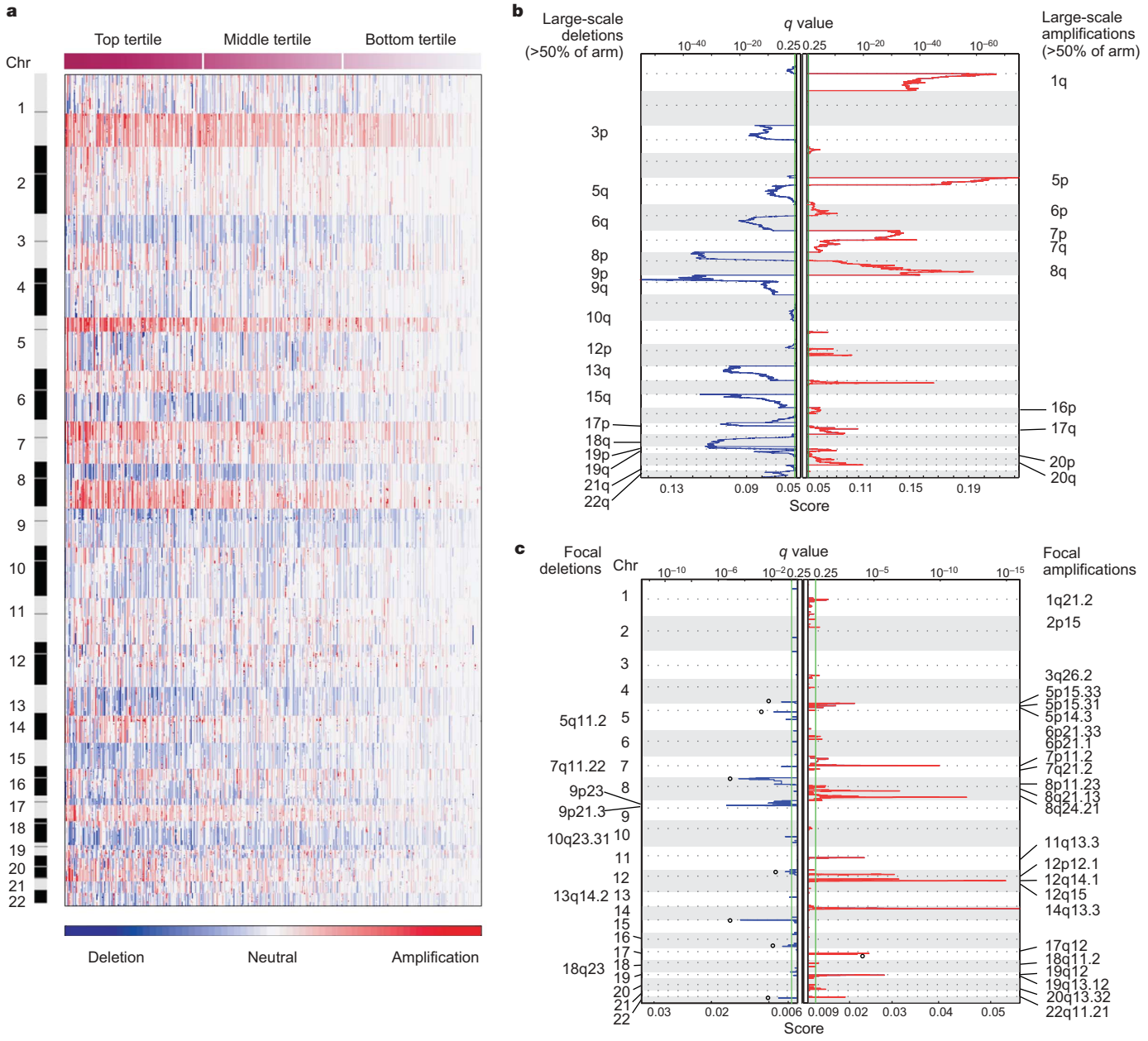

Figure 1 | Large-scale genomic events in lung adenocarcinoma. **a**, Smoothed copy number data for 371 lung adenocarcinoma samples (columns; ordered by degree of interchromosomal variation and divided into top, middle and bottom tertiles) is shown by genomic location (rows). The colour scale ranges from blue (deletion) through white (neutral; two copies in diploid specimens) to red (amplification). **b**, **c**, False-discovery rates (*q* values; green line is 0.25 cut-off for significance) and scores for each alteration (*x* axis) are plotted at each genome position (*y* axis); dotted lines indicate the centromeres. Amplifications (red lines) and deletions (blue lines) are shown for large-scale events (**b**; ≥50% of a chromosome arm; copy number threshold = 2.14 and 1.87) and focal events (**c**; copy number threshold = 3.6 and 1.2). Open circles label known or presumed germline copy-number polymorphisms.

The most common genomic alteration in lung adenocarcinoma is copy-number gain of chromosome 5p, which is found in 60% of total samples and over 80% of the top tertile (Supplementary Table 5). Another 15 large-scale events are seen in at least 33% of all samples and over 40% of the top tertile. Together, the regions of common copy-number gain (~650 megabases (Mb)) and copy-number loss (~1,010 Mb) comprise more than half of the human genome (Supplementary Results and Supplementary Table 5). Despite their high frequency, few of these large-scale events have been clearly related to functional effects on specific genes. Loss of a chromosome arm is likely to act by uncovering an inactivated tumour suppressor gene, yet such mutations have been well-established in lung adenocarcinoma in only three of the sixteen deleted chromosome arms (*CDKN2A* on 9p, *TP53* on 17p and *STK11* on 19p)[16–18]. We tested for correlations between the large-scale lesions and clinical parameters, but none was significant after correction for multiple hypothesis testing (Supplementary Results and Supplementary Table 7).

Focal deletions may help pinpoint tumour suppressor genes, particularly on chromosome arms that show frequent copy-number loss. At a threshold set to detect homozygous deletions in the presence of stromal contamination, GISTIC analysis identified seven focal candidate regions (Fig. 1c and Supplementary Table 8). The most significant focal deletions, detected in 3% of all samples and 6.5% of the top tertile, encompass *CDKN2A*/*CDKN2B*, two well-documented tumour suppressor genes on chromosome 9p21 (Fig. 1c, Table 1 and Supplementary Table 8). The protein products of *CDKN2A* and *CDKN2B* inhibit two cyclin-dependent kinases, Cdk4 and Cdk6, the genes of which both reside in frequently amplified regions (see below). Two other deleted regions also encompass known tumour suppressor genes, *PTEN* and *RB1* (Supplementary Table 8).

Three additional deletion regions each localize to a single gene. Deletions of the 5′ untranslated region of *PTPRD*, encoding a tyrosine phosphatase, occur in 4% of the top tertile. Although *PTPRD* deletions have been reported in lung adenocarcinoma cell lines[8,19,20], this is the first observation in primary human lung adenocarcinomas. Homozygous deletions of *PDE4D* occur in 1.6% of the top tertile and typically remove several hundred kilobases and affect multiple exons (Supplementary Fig. 5). These deletions may be significant for lung biology because *PDE4D* encodes the major phosphodiesterase responsible for degrading cyclic AMP in airway epithelial cells[21]. Another single-gene deletion occurs within *AUTS2*, a gene of unknown function in chromosome 7q11.22 (Table 1 and Supplementary Table 8). We cannot exclude the possibility that some recurrent copy-number losses are due to genomic fragility unrelated to carcinogenesis; the presence of point mutations would provide additional support for a role in cancer.

We therefore sequenced all exons of *AUTS2*, *PDE4D* and *PTPRD*, as each of these genes showed single-gene deletions but no mutations have been reported in primary tumours. Although we did not detect somatic mutations in *AUTS2* or *PDE4D*, we identified validated somatic *PTPRD* mutations in 11 of 188 samples sequenced. Notably, three of the mutations encode predicted inactivating changes in the tyrosine phosphatase domain (Supplementary Table 9 and Supplementary Fig. 6). These results implicate *PTPRD* as a probable cancer-associated gene, although further studies are needed to establish a causative role in cancer via gain or loss of function.

We focused above on homozygous deletions, but note that this approach will miss important genes. Notably, the *TP53* locus is known to be mutated in ~50% of lung adenocarcinomas but shows no homozygous deletions in our data.

We next focused on focal amplification events, for which it may be easier to pinpoint target genes. At a threshold designed to identify high-copy amplification, the GISTIC analysis identified 24 recurrent genomic segments with maximum copy number ranging from about 4- to 16-fold (Fig. 1c, Table 1 and Supplementary Table 10). The amplification events are seen in 1–7% of all samples (1–12% in the top tertile). Each of these events is seen in at least two samples and all but eight are seen in at least five samples. In the 13 most significant amplifications ($q < 0.01$), the regions can be localized to relatively small genomic segments containing 15 or fewer genes. Although 14 of the 24 regions of recurrent amplification contain a known proto-oncogene (Supplementary Table 10), only three of these genes (*EGFR*, *KRAS* and *ERBB2*) have been previously reported to be mutated in lung adenocarcinoma (Supplementary Results). The remaining 11 genes are clear targets for re-sequencing in lung tumours.

Table 1 | Top focal regions of amplification and deletion

Cytoband*	*q* value	Peak region (Mb)*	Max/Min inferred copy no.	Number of genes*†	Known proto-oncogene/ tumour suppressor gene in region*‡	New candidate(s)
Amplifications						
14q13.3	2.26×10^{-29}	35.61–36.09	13.7	2	–	*NKX2-1, MBIP*
12q15	1.78×10^{-15}	67.48–68.02	9.7	3	*MDM2*	–
8q24.21	9.06×10^{-13}	129.18–129.34	10.3	0	*MYC*§	–
7p11.2	9.97×10^{-11}	54.65–55.52	8.7	3	*EGFR*	–
8q21.13	1.13×10^{-7}	80.66–82.55	10.4	8	–	–
12q14.1	1.29×10^{-7}	56.23–56.54	10.4	15	*CDK4*	–
12p12.1	2.83×10^{-7}	24.99–25.78	10.4	6	*KRAS*	–
19q12	1.60×10^{-6}	34.79–35.42	6.7	5	*CCNE1*	–
17q12	2.34×10^{-5}	34.80–35.18	16.1	12	*ERBB2*	–
11q13.3	5.17×10^{-5}	68.52–69.36	6.5	9	*CCND1*	–
5p15.33	0.000279	0.75–1.62	4.2	10	*TERT*	–
22q11.21	0.001461	19.06–20.13	6.6	15	–	–
5p15.31	0.007472	8.88–10.51	5.6	7	–	–
1q21.2	0.028766	143.48–149.41	4.6	86	*ARNT*	–
20q13.32	0.0445	55.52–56.30	4.4	6	–	–
5p14.3	0.064673	19.72–23.09	3.8	2	–	–
6p21.1	0.078061	43.76–44.12	7.7	2	–	*VEGFA*
Deletions						–
9p21.3	3.35×10^{-13}	21.80–22.19	0.7	3	*CDKN2A/ CDKN2B*	–
9p23	0.001149	9.41–10.40	0.4	1	–	*PTPRD*‖
5q11.2	0.005202	58.40–59.06	0.6	1	–	*PDE4D*
7q11.22	0.025552	69.50–69.62	0.7	1	–	*AUTS2*
10q23.31	0.065006	89.67–89.95	0.5	1	*PTEN*	–

* Based on hg17 human genome assembly.
† RefSeq genes only.
‡ Known tumour suppressor genes and proto-oncogenes defined as found in either COSMIC[30], CGP Census[31] or other evidence; if there is more than one known proto-oncogene in the region, only one is listed (priority for listing is, in order: known lung adenocarcinoma mutation; known lung cancer mutation; other known mutation (by COSMIC frequency); listing in CGP Census).
§ *MYC* is near, but not within, the peak region.
‖ Single gene deletions previously seen, this study provides new mutations as well.

Our data localize the amplification peak on chromosome 5p to the telomerase catalytic subunit gene, *TERT*. Although broad amplification of chromosome 5p has been described in non-small-cell lung cancer (NSCLC)[13,22,23], the target of 5p amplification has not been determined. In our data set, eight tumours with amplicons in chromosome 5p15 delineate a region containing ten genes, including *TERT* (Table 1 and Supplementary Table 10), suggesting that *TERT* may be the target of the amplification and thereby contributes to cellular immortalization.

Chromosome 6p21.1 shows focal amplification in four samples in a region containing two genes, one of which (*VEGFA*) encodes vascular endothelial growth factor (Table 1 and Supplementary Table 10). This amplification suggests a possible mechanism for increased angiogenesis and for the reported response to angiogenic inhibitors such as the anti-VEGF antibody bevacizumab in lung adenocarcinoma[24,25]. Similarly, amplification of regions including several cell cycle genes such as *CDK4*, *CDK6* and *CCND1* suggests an important role for these genes (Table 1 and Supplementary Table 10).

Notably, the most common focal amplification does not include any known proto-oncogenes: chromosome 14q13.3 is amplified in 6% of the samples overall and 12% of the samples in the top tertile (Fig. 1c, Table 1 and Supplementary Table 10; $q < 10^{-28}$). Although previous studies have reported amplification of 14q13 in lung cancer cell lines[14] and the region is mentioned in studies of primary lung tumours[8,15], the target gene in this region had not been identified. With our large sample size, we are able to narrow the critical region to a 480-kilobase interval containing only two known genes, *MBIP* and *NKX2-1* (Fig. 2a, b, Table 1 and Supplementary Table 10). Data for a single tumour with a small region of high-level amplification, comprising *MBIP* and *NKX2-1*, exclude the neighbouring gene, *NKX2-8* (Fig. 2c).

We confirmed the amplification of the region by fluorescence *in situ* hybridization (FISH) and quantitative polymerase chain reaction (qPCR; data not shown). FISH analysis was performed with a bacterial artificial chromosome (BAC) probe containing *NKX2-1* and *NKX2-8* (Fig. 2c) on an independent set of 330 lung adenocarcinoma samples from tissue microarrays. High-level amplification of the chromosome 14q13.3 region was seen in 12% (40 out of 330) of these lung tumours. The FISH studies revealed amplification up to an estimated 100-fold (Fig. 2d and Supplementary Fig. 7); the lower amplification estimated on the SNP arrays (up to 14-fold) probably reflects signal saturation, stromal admixture and tumour heterogeneity. No significant difference in patient survival after surgical resection and long-term follow-up was observed between tumours with amplified or non-amplified *NKX2-1* (Supplementary Fig. 8 and Supplementary Table 11). Exon-based sequencing in 384 lung adenocarcinoma DNA samples showed no somatic mutations in either *NKX2-1* or *MBIP* (Supplementary Results), indicating that any oncogenic function might be exerted by the wild-type gene.

We used RNA interference (RNAi) to test the roles of both *MBIP* and *NKX2-1* with respect to cell survival and oncogenic properties. Expression of two different short hairpin RNAs (shRNAs) targeting *NKX2-1* significantly reduced the levels of NKX2-1 protein in NCI-H2009 cells (Fig. 3a) and NCI-H661 cells (data not shown)—NSCLC lines that carry 14q amplifications[14]. No NKX2-1 protein was detected in A549 cells that lack 14q amplification (Fig. 3a).

RNAi-mediated inhibition of *NKX2-1* expression substantially decreased the ability of NCI-H2009 cells to grow in an anchorage-independent manner as measured by colony formation in soft agar (Fig. 3b), which may be due, in part, to a loss of cell viability. NCI-H661 cell viability was also impaired by *NKX2-1* RNAi (Supplementary Fig. 9). *NKX2-1* knockdown leads to a decrease in colony formation in lung adenocarcinoma lines (NCI-H1975 and HCC1171) that lack chromosome 14q13 amplification but express NKX2-1 (Supplementary Fig. 10), but has no effect on either soft agar colony formation or cell viability in A549 cells, which express little or no NKX2-1 protein (Fig. 3a, c). In contrast to the results for *NKX2-1*, RNAi-based *MBIP* knockdown neither decreased colony formation in NCI-H2009 cells (Fig. 3d, e) or in NCI-H661 cells (Supplementary Fig. 11a, b), nor reduced cell viability (Supplementary Fig. 11c, d). It thus seems that *NKX2-1*, but not *MBIP*, is essential for the survival and tumorigenic properties of lung adenocarcinoma cell lines that express *NKX2-1*.

Systematic understanding of the molecular basis of a particular type of cancer will require at least three steps: comprehensive characterization of recurrent genomic aberrations (including copy-number changes, nucleotide sequence changes, chromosomal rearrangements and epigenetic alterations); elucidation of their biological role in cancer pathogenesis; and evaluation of their utility for diagnostics, prognostics and therapeutics. This study represents a step towards comprehensive genomic characterization of one of the

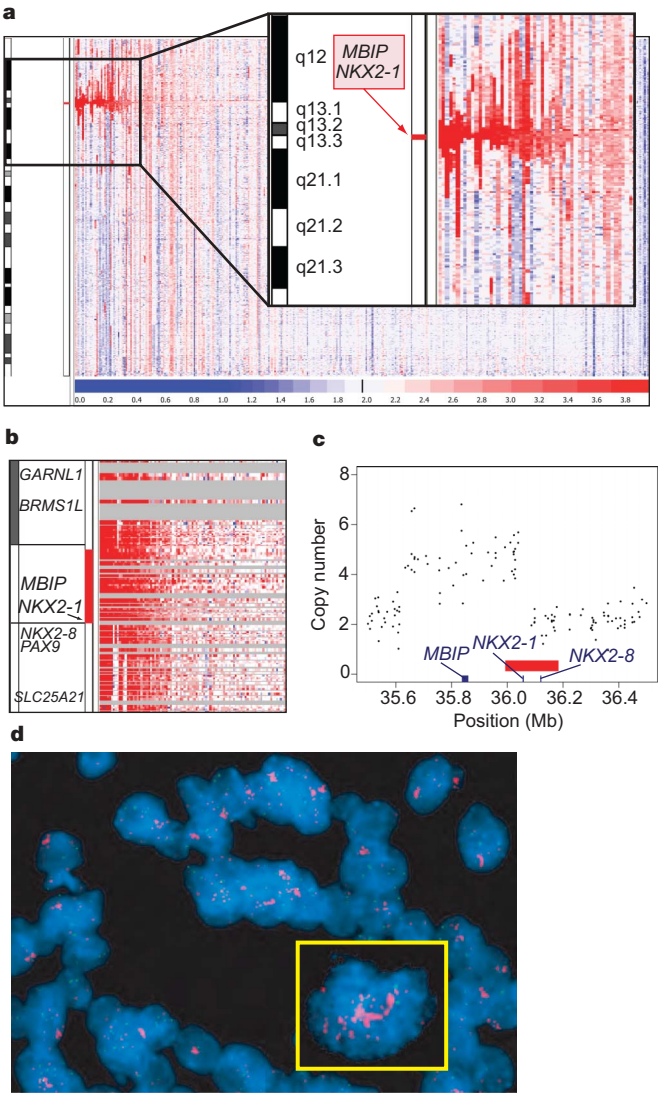

Figure 2 | High-prevalence amplification of the *MBIP/NKX2-1* locus on chromosome 14q. **a**, Copy number on chromosome 14q is shown for 371 lung adenocarcinomas (columns; ordered by amplification) from centromere (top) to telomere (bottom). Colour scale as in Fig. 1. **b**, Magnified view of the amplified region from **a**; grey bars represent the absence of SNPs on the array. **c**, Raw copy number data (*y* axis) for one sample defining the minimally amplified region are plotted according to chromosome 14 position (*x* axis; scale in megabases). Genomic positions of *MBIP*, *NKX2-1*, *NKX2-8* and the BAC used for FISH (red bar) are shown along the *x* axis. **d**, FISH for *NKX2-1* (red) and a chromosome 14 reference probe (green) on a lung adenocarcinoma specimen with high-level amplification of the *NKX2-1* probe. Nuclei are stained with 4,6-diamidino-2-phenylindole (DAPI; blue). The yellow box shows a single nucleus.

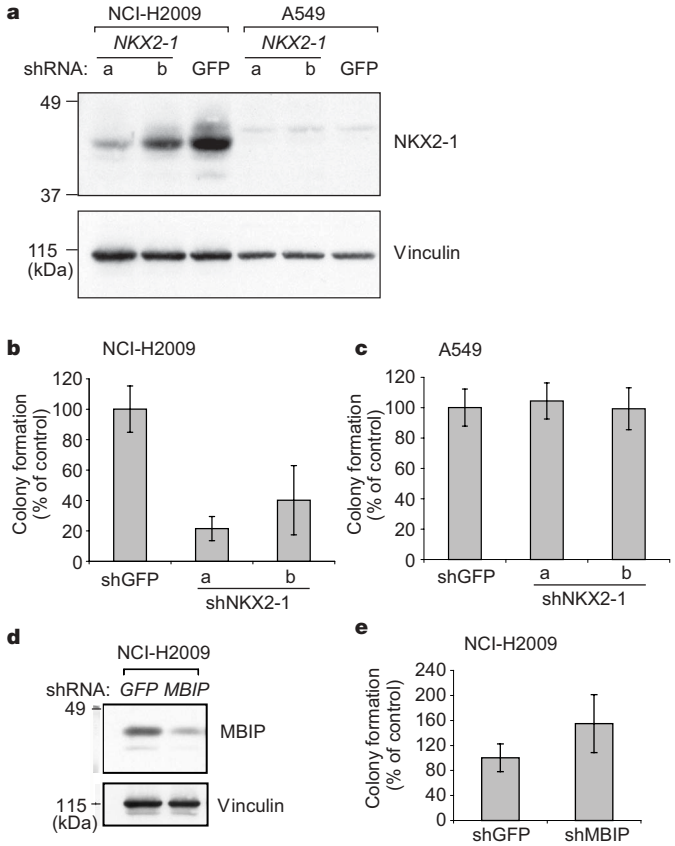

Figure 3 | *NKX2-1* RNAi leads to reduced anchorage-independent growth and viability of NCI-H2009 cells but not A549 cells. a, Anti-NKX2-1 and control anti-vinculin immunoblots of lysates from NCI-H2009 and A549 cells expressing shRNA against *NKX2-1* (shNKX2-1a and shNKX2-1b) or *GFP* (shGFP) as control. **b**, Soft agar colony formation by NCI-H2009 cells is shown relative to the shGFP control as a mean percentage (± standard deviation in triplicate samples; $P = 5.8 \times 10^{-6}$ when comparing shGFP to shNKX2-1a and $P = 5.1 \times 10^{-4}$ when comparing shGFP to shNKX2-1b). **c**, Colony formation assays as in **b** for A549 cells ($P > 0.5$). **d**, Anti-MBIP and anti-vinculin immunoblots of lysates from shRNA-expressing NCI-H2009 cells. **e**, Colony formation of shMBIP NCI-H2009 cells relative to that of control shGFP cells ($P = 0.0344$).

most common cancers, lung adenocarcinoma. We define two main types of recurrent events in this disease: frequent, large-scale events and rare, focal events. Further efforts to identify the target genes of the frequent, large-scale events will probably involve systematic screens to produce orthogonal data sets (mutational, epigenetic, expression and loss-of-function phenotypes).

Strikingly, the single most common focal event in lung adenocarcinoma (amplification of 14q13.3) was not previously associated with a specific gene. We show here that the target gene is *NKX2-1*, a transcription factor that has an essential role in the formation of type II pneumocytes, the cell type that lines the alveoli of the lung[26,27]. *Nkx2-1* knockout mice fail to develop normal type II pneumocytes or alveoli and die of respiratory insufficiency at birth[28], which highlights the importance of *NKX2-1* in lung development. *NKX2-1* shows hallmarks of a novel lineage-survival oncogene, similar to the *MITF* gene in melanoma[7]. The lineage-restricted amplification of such genes contrasts with the more ubiquitous amplifications seen for genes in cell cycle (for example, *CDK4*, *CDK6*, *CCND1*, *CCNE1*) and signal transduction (for example, *EGFR*, *ERBB2*, *KRAS*) pathways.

More generally, our results, together with other recent studies[29], illustrate the power of systematic copy-number analysis with SNP arrays. They make clear that many important cancer-related genes remain to be discovered and can be revealed by systematic genomic study.

METHODS SUMMARY

DNA specimens were labelled and hybridized to the Affymetrix 250K Sty I array to obtain signal intensities and genotype calls. Loci identified by GISTIC analysis were further characterized by sequencing, genotype validation, tissue microarray FISH and functional studies. RNAi was performed by stable expression of shRNA lentiviral vectors targeting *NKX2-1*, *MBIP* or *GFP* in lung cancer cell lines, which were then used in soft agar and cell proliferation assays. Raw data and related files are available at http://www.broad.mit.edu/tsp.

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

Received 12 April; accepted 10 October 2007.
Published online 4 November 2007.

1. Weir, B., Zhao, X. & Meyerson, M. Somatic alterations in the human cancer genome. *Cancer Cell* **6,** 433–438 (2004).
2. Sawyers, C. Targeted cancer therapy. *Nature* **432,** 294–297 (2004).
3. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20,** 207–211 (1998).
4. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23,** 41–46 (1999).
5. Bignell, G. R. *et al.* High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14,** 287–295 (2004).
6. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64,** 3060–3071 (2004).
7. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436,** 117–122 (2005).
8. Zhao, X. *et al.* Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res.* **65,** 5561–5570 (2005).
9. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* **98,** 31–36 (2001).
10. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* **2,** RESEARCH0032 (2001).
11. Hupe, P., Stransky, N., Thiery, J. P., Radvanyi, F. & Barillot, E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20,** 3413–3422 (2004).
12. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl Acad Sci. USA*. (in the press).
13. Balsara, B. R. & Testa, J. R. Chromosomal imbalances in human lung cancer. *Oncogene* **21,** 6877–6883 (2002).
14. Garnis, C. *et al.* High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Int. J. Cancer* **118,** 1556–1564 (2006).
15. Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proc. Natl Acad. Sci. USA* **102,** 9625–9630 (2005).
16. Hayashi, N., Sugimoto, Y., Tsuchiya, E., Ogawa, M. & Nakamura, Y. Somatic mutations of the *MTS* (multiple tumor suppressor) 1/*CDK4I* (cyclin-dependent kinase-4 inhibitor) gene in human primary non-small cell lung carcinomas. *Biochem. Biophys. Res. Commun.* **202,** 1426–1430 (1994).
17. Sanchez-Cespedes, M. *et al.* Inactivation of *LKB1/STK11* is a common event in adenocarcinomas of the lung. *Cancer Res.* **62,** 3659–3662 (2002).
18. Takahashi, T. *et al.* p53: a frequent target for genetic abnormalities in lung cancer. *Science* **246,** 491–494 (1989).
19. Sato, M. *et al.* Identification of chromosome arm 9p as the most frequent target of homozygous deletions in lung cancer. *Genes Chromosomes Cancer* **44,** 405–414 (2005).
20. Cox, C. *et al.* A survey of homozygous deletions in human cancer genomes. *Proc. Natl Acad. Sci. USA* **102,** 4542–4547 (2005).
21. Barnes, A. P. *et al.* Phosphodiesterase 4D forms a cAMP diffusion barrier at the apical membrane of the airway epithelium. *J. Biol. Chem.* **280,** 7997–8003 (2005).
22. Zhu, C. Q. *et al.* Amplification of telomerase (*hTERT*) gene is a poor prognostic marker in non-small-cell lung cancer. *Br. J. Cancer* **94,** 1452–1459 (2006).
23. Zhang, A. *et al.* Frequent amplification of the telomerase reverse transcriptase gene in human tumors. *Cancer Res.* **60,** 6230–6235 (2000).
24. Johnson, D. H. *et al.* Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer. *J. Clin. Oncol.* **22,** 2184–2191 (2004).
25. Sandler, A. *et al.* Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N. Engl. J. Med.* **355,** 2542–2550 (2006).
26. Bingle, C. D. Thyroid transcription factor-1. *Int. J. Biochem. Cell Biol.* **29,** 1471–1473 (1997).
27. Ikeda, K. *et al.* Gene structure and expression of human thyroid transcription factor-1 in respiratory epithelial cells. *J. Biol. Chem.* **270,** 8108–8114 (1995).

28. Yuan, B. *et al.* Inhibition of distal lung morphogenesis in *Nkx2.1*$^{-/-}$ embryos. *Dev. Dyn.* **217,** 180–190 (2000).
29. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
30. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* **91,** 355–358 (2004).
31. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).

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

Acknowledgements This work was supported by grants from the US National Cancer Institute (B.A.W., M.S.W., M.R.S., M.M., I.I.W., A.F.G., J.A.R., M.S., J.D.M.), the US National Human Genome Research Institute (R.A.G, R.K.W., E.S.L.), the Canadian Cancer Society/National Cancer Institute (M.S.T.), the American Lung Association (M.M.), Joan's Legacy Foundation (M.M.), the American Cancer Society (M.M.), the International Association for the Study of Lung Cancer (R.K.T.), the US Department of Defense (R.B., I.I.W., J.D.M.) and the Carmel Hill Fund (W.P., M.G.K., H.E.V.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.M. (matthew_meyerson@dfci.harvard.edu).

METHODS

Primary lung specimens. A total of 575 DNA specimens were obtained from primary lung tumours (all of them with the original diagnosis of lung adenocarcinoma, 528 of which were confirmed to be lung adenocarcinomas), 439 matched normal samples and 53 additional normal specimens. These DNAs were labelled and hybridized to SNP arrays (see below) without previous whole-genome amplification. Each of the selected tumour samples was determined to have greater than 70% tumour percentage by pathology review.

Of the 575 selected tumours, 384 anonymous lung tumour and matched normal DNAs for the Tumour Sequencing Project (TSP) were collected from five sites: Memorial-Sloan Kettering Cancer Center (102 tumours and paired normal samples), University of Michigan (101 tumours and paired normal samples), MD Anderson Cancer Center (29 tumours and paired normal samples), Washington University (84 tumours and paired normal samples) and Dana-Farber Cancer Institute/The Broad Institute (68 tumours and paired normal samples). Additional anonymous lung adenocarcinoma samples or DNAs were collected from the Brigham and Women's Hospital tissue bank (19 tumours and 18 paired normal samples), H. Sasaki at the Nagoya City University Medical School (112 tumours and 37 paired normal samples) and from the University Health Network in Toronto (60 tumour samples). In addition to the matched normal samples, 53 unmatched normal tissue or blood samples were used for SNP array normalization purposes (sources include J. Llovet, S. Pomeroy, S. Singer, the Genomics Collaborative, Inc., Massachusetts General Hospital and R. Beroukhim). All tumour samples were surgically dissected and frozen at −80 °C until use.

SNP array experiments. For each sample, SNPs were genotyped with the Sty I chip of the 500K Human Mapping Array set (Affymetrix Inc.). Array experiments were performed according to manufacturer's directions. In brief, for each sample, 250 ng of genomic DNA was digested with the StyI restriction enzyme (New England Biolabs). The digested DNA was then ligated to an adaptor with T4 ligase (New England Biolabs) and PCR-amplified using an Applied Biosystems 9700 Thermal Cycler I and Titanium Taq (Clontech) to achieve a size range of 200–1,100 bp. Amplified DNA was then pooled, concentrated and put through a clean-up set. The product was then fragmented using DNaseI (Affymetrix Inc.) and subsequently labelled, denatured and hybridized to arrays. Hybridized arrays were scanned using the GeneChip Scanner 3000 7G (Affymetrix Inc.). Batches of 96 samples were processed as a single plate using a Biomek FX robot with dual 96 and span-8 heads (Beckman Coulter) and a GeneChip Fluidics Station FS450 (Affymetrix Inc.). Samples and plates were tracked using ABGene 2D barcode rack and single tube readers (ABGene). Tumour and paired normal sample (where applicable) were always placed in adjacent wells on the same plate to minimize experimental differences. Raw data (.CEL and .txt files) are available at http://www.broad.mit.edu/tsp.

Primary SNP array data analysis. SNP arrays were processed as a plate of 96 samples using the GenePattern software package[32], with modules based on dChipSNP algorithms[9,10]. GenePattern modules are available at http://www.broad.mit.edu/cancer/software/genepattern/. Intensity (.CEL) files were normalized and modelled using the PM-MM difference modelling method[9] with the SNPfileCreator module. Array normalization, similar to quantile normalization, was performed[33]; 6,000 matching quantiles from the probe density distributions of two arrays were used to fit a running median normalization curve for normalization of each array to a common baseline array[10].

Array quality control analysis. Further analysis was performed on arrays that met certain quality control criteria. As a first step, non-adenocarcinoma samples ($n = 47$) from the TSP set of 384 tumours were removed from further analysis (leaving 528 adenocarcinomas). Technical failure criteria (removing 33 tumours) included a requirement for correct tumour/normal matching, genotyping call rates (% of SNPs that a genotype call can be inferred for) greater than 85% and a score measuring copy-number variation between neighbouring SNPs of less than 0.5. The measure of local SNP copy number variation is calculated by the formula: variation score = mean$[(\log(\mathrm{RC}_i) - \log(\mathrm{RC}_{i+1}))^2 + (\log(\mathrm{RC}_i) - \log(\mathrm{RC}_{i-1}))^2]$, where RC_i is the raw copy number at SNP i and the mean is taken over all SNPs. Criteria also included a requirement that after taking the log2 ratio and performing segmentation by GLAD[11], the number of times the smoothed copy number crossed ±0.1 on the log scale in the genome of tumour samples was < 100 (removing 73 tumours). The same test was used to exclude normal samples, with the number of times the smoothed copy number crossed ±0.1 decreased to <45 (removing 50 normal samples). A histogram quality control step, as part of the GISTIC procedure, then removed tumours ($n = 51$) with high degrees of non-tumour DNA contamination by looking for samples with only one peak of copy number across its whole genome. This histogram quality control step also removed normals ($n = 20$) with tumour DNA contamination by looking for samples with greater than one peak of copy number across its whole genome.

GISTIC analysis. GISTIC analysis[12] was performed on arrays that met certain quality control criteria. Raw intensity value files from the GenePattern SNPfileCreator module were used as input into the GISTIC algorithm. In brief, batch correction, data normalization, copy-number determination using either the paired normal sample or the average of the five closest normal samples and copy number segmentation was performed. Data-set-specific copy number polymorphisms were identified by running GISTIC on the set of normal samples alone; the regions identified from this analysis were then also removed from the subsequent analysis of tumours. GISTIC then assigns G^{AMP} and G^{DEL} scores to each locus, respectively representing the frequency of amplifications (deletions) seen at that locus, multiplied by the average increase (decrease) in the log2 ratio in the amplified (deleted) samples. The score (G) is based on the average amplitude (a) of the lesion type (amplification or deletion) and its frequency (f) in the data set according to the formula: $G_i^{(\text{lesion type})} = f_i^{(\text{lesion type})}\, a_i^{(\text{lesion type})}$. The significance of each score is determined by comparison to similar scores obtained after permuting the data within each sample. The resulting q-value is an upper bound for the expected fraction of false positives among all regions with a particular q-value or less. GISTIC also implements a peel-off step, which identifies additional secondary peaks within a region.

GISTIC analysis was performed essentially the same as is described in a future publication[12], with the following exceptions. Copy number determination was performed for each tumour using its matched normal sample when available and of good quality ($n = 242$). For all others, the average of the five closest normal samples was used ($n = 129$). Copy number segmentation was performed using the GLAD algorithm with parameter $d = 10$. GLAD segments less than eight SNPs in length were also removed.

Regions identified by GISTIC were also compared to known copy-number polymorphisms[34] and were manually reviewed for the presence of the alteration in the paired normal sample. Focal deletion regions with events that occurred in tumour samples that did not have paired normal samples were considered presumed polymorphisms and also removed from the list. Secondary peaks and known and presumed germline copy number polymorphisms are listed in Supplementary Tables 12 and 13.

GISTIC analysis of large-scale regions. Significant broad regions of amplification and deletion were identified by applying GISTIC with the default thresholds of 2.14/1.87 (log2 ratio of ±0.1). Regions identified by GISTIC that were greater than 50% of a chromosome arm were considered large-scale. Region frequencies were calculated by determining the number of samples that had a median log2 ratio greater/less than the threshold (±0.1), for those SNPs within the region.

GISTIC analysis of focal regions. Significant focal regions of amplification and deletion were identified by applying GISTIC with a threshold of 3.6/1.2 (log2 ratio of 0.848/−0.737).

Data visualization. Normalized raw copy number from GISTIC analysis was used as input for visualization in the GenePattern SNPviewer (http://www.broad.mit.edu/cancer/software/genepattern/)[32]. Mapping information for SNP, Refgene and cytoband locations are based on Affymetrix annotations and the hg17 build of the human genome sequence from the University of California, Santa Cruz (http://genome.ucsc.edu).

Chromosome arm analysis. After segmentation by GLAD, the median of each chromosome arm for each sample was calculated. Amplification or deletion of an arm across the data set was tested for significance by a two-sided binomial test, after removing log2 copy number ratios between ±0.1. P values were false-discovery rate (FDR) corrected to give a FDR q value; significance is set to a q value of 0.01. The standard deviation of the median copy number of significant arms was then used to sort samples into three groups. Higher standard deviation implies higher interchromosomal variation, which correlates with less stromal contamination. Frequencies were then calculated for the total set and for only the top one-third least stromally contaminated samples to give a better idea of true frequencies in the context of attenuated signal owing to stromal contamination.

Comparison between tertiles. A similar chromosome arm analysis was performed independently on the three sample groups, separated according to the standard deviations of their median arm log2 copy number ratios. Amplification or deletion of an arm across the data set was tested for significance by a two-sided binomial test, after removing values between ±0.0125. P values were FDR corrected to give a FDR q value, significance is set to a q value of 0.01.

Estimation of stromal contamination. To attempt to estimate stromal contamination, we calculated the allele-specific copy numbers by taking all informative SNPs in each of the 237 tumours that have a paired normal (removing five bad pairs) and dividing the allele-specific signal from the tumour by that of the normal. Then for each SNP we found M, the minimum between the copy numbers of the A and B alleles. In regions in which one allele has zero copies (for example, one copy loss in diploid cells) M represents the stromal contamination level (as the stroma has one copy of each allele). We calculated the median value

of *M* across each of the chromosome arms and then estimated the stromal contamination by taking their minimum.

LOH analysis. Inferred LOH calls using an HMM algorithm for 242 tumour/normal sample pairs were generated using dChipSNP[35]. Default parameters were used, except the genotyping error rate was set to 0.2. Five bad-quality sample pairs were removed before visualization and GISTIC analysis. GISTIC analysis of LOH calls and copy loss for 237 samples were performed as described[12].

Correlation analysis. Associations were tested between each large-scale alteration identified by GISTIC and certain clinical parameters. A Fisher's exact test was used to determine association of large-scale copy-number lesions with the binary clinical parameters (gender and smoking status). A chi-squared test was used to determine whether each large-scale copy number alteration was independent of each non-binary clinical parameter (age range, differentiation, tumour stage or patient's reported ancestry). *P* values were FDR corrected to give a FDR *q* value, significance is set to a *q* value of 0.05.

Correlation of clinical features and *NKX2-1* amplification. The analysis included 123 consecutive patients with lung adenocarcinoma treated at Brigham and Women's Hospital between January 1997 and December 1999. Fifty-two of these cases had a FISH amplification status that was not assessable (6 cases showed no tumour on the tissue cores and 46 cases had insufficient hybridization). Of the remaining 71 cases, 10 cases had *NKX2-1* amplification, 1 had a *NKX2-1* deletion, and 60 cases showed no *NKX2-1* alteration. All cases for which the *NKX2-1* amplification status was not assessable and the one case that showed a *NKX2-1* deletion were excluded, bringing the final number of cases included in the analysis to 70.

All cases were histologically confirmed as lung adenocarcinomas. For cases that showed a pure solid growth pattern, mucicarmine and immunohistochemical stains were performed to confirm that the tumour was an adenocarcinoma. Well-differentiated tumours were defined as tumours with a purely bronchioloalveolar growth pattern or mixed tumours with an acinar component with cytologic atypia equivalent to that seen with bronchioloalveolar carcinoma. Poorly differentiated tumours were defined as tumours that showed any amount of solid growth. All other tumours were classified as moderately differentiated. Patient demographics, smoking status, tumour location, type of surgical resection, tumour stage (according to the 6th edition of the American Joint Committee on Cancer system for lung carcinoma) and nodal status were recorded.

Overall survival of patients with *NKX2-1* amplification. We excluded from the survival analysis three cases with *NKX2-1* amplification and 11 cases that had no *NKX2-1* alterations. Exclusion criteria included: cancer was a recurrence; patients received neoadjuvant treatment; patients died within the first 30 days after surgery; and patients had another cancer diagnosed in the 5 years before the diagnosis of lung adenocarcinoma. Survival was plotted by Kaplan–Meier method using the date of resection and date of death or last follow-up.

Sequencing. *NKX2-1*, *MBIP* and *AUTS2* were sequenced in all 384 TSP lung adenocarcinomas. Primers were designed in an automated fashion using Primer 3 (ref. 36) and characterized by amplification in genomic DNA from three Coriell cell lines. Primers that show an agarose gel band for at least two of the three DNAs were then used for production PCR. Passing primers were arrayed into 384-well PCR plates along with samples and PCR master mix. A total of 5 ng of whole-genome-amplified sample DNA was PCR amplified over 35 cycles in Thermo-Hybaid units, followed by a SAP/Exo clean-up step. *NKX2-1* PCR reactions for sequencing contained an addition of 5% DMSO. The resulting purified template is then diluted and transferred to new plates for the sequencing reaction. After cycling (also performed on Thermo-Hybaids), the plates are cleaned up with an ethanol precipitation, re-hydrated and detected on an ABI 3730xl DNA analyser (Applied Biosystems). Output from the detectors is transferred back to the directed sequencing platform's informatics pipeline. SNPs and/or mutations are then identified using three mutation-detecting algorithms in parallel: PolyPhred[37] and PolyDHAN (D. Richter *et al.*, manuscript in preparation), which are bundled into the in-house software package SNP Compare, and the commercially available Mutation Surveyor (SoftGenetics, LLC.). Candidates were filtered to remove silent variants, intronic variants (with the exception of potential splice site mutations) and validated SNPs registered in dbSNP or confirmed as SNPs in our previous experiments.

Mutation validation by genotyping. Homogeneous mass extension (hME) genotyping for validation of sequencing candidates was performed in 96-well plates with up to 7-plex reactions. PCR was performed with final concentrations of 0.83 mM dNTPs, 1.56× of 10× buffer, 3.38 mM MgCl, 0.03 U μl^{-1} HotStar Taq (Qiagen), 0.10 μM PCR primers. Thermocycling was performed at 92 °C for 15 min, followed by 45 cycles of 92 °C for 20 s, 56 °C for 30 s and 72 °C for 1 min, with an additional extension at 72 °C for 3 min. Shrimp alkaline phosphatase (SAP) clean-up was performed using a master mix made up of 0.5× buffer and SAP. Reactions were performed at 34 °C for 20 min, 85 °C for 5 min and then held at 4 °C. After the SAP clean-up, hME reaction was performed using thermosequenase and final concentrations of 0.06 mM sequenom termination mix (specific to the pool being used), and 0.64 μM extension primer. Reactions were cycled at 94 °C for 2 min, followed by 55 cycles of 94 °C for 5 s, 52 °C for 5 s and 72 °C for 5 s. Samples were then put through a resin clean-up step, then the purified primer extension reaction was loaded onto a matrix pad (3-hydroxypicoloinic acid) of a SpectroCHIP (Sequenom) and detected by a Bruker Biflex III MALDI-TOF mass spectrometer (SpectroREADER, Sequenom).

***PTPRD* mutation discovery and validation.** The *PTPRD* gene was sequenced in 188 lung adenocarcinoma samples. Sequence traces (reads) were aligned to human reference sequence using cross-match. PolyPhred[37] and PolyScan were used to predict SNPs and insertions/deletions. Identified SNPs were validated using the Illumina Goldengate assay. ENST00000356435 is the transcript used for annotating the mutations. Both synonymous and non-synonymous candidates were identified, but only non-synonymous mutations were validated.

Tissue microarray FISH (TMA–FISH). A Biotin-14-dCTP-labelled BAC clone RP11-1083E2 (conjugated to produce a red signal) was used for the *NKX2-1* probe and a Digoxin-dUTP labelled BAC clone RP11-72J8 (conjugated to produce a green signal) was used for the reference probe. Tissue hybridization, washing and colour detection were performed as described previously[7,38]. *NKX2-1* amplification by FISH was assessed using a total of 935 samples (represented by 2,818 tissue microarray cores).

The BAC clones were obtained from the BACPAC Resource Center, Children's Hospital Oakland Research Institute (CHORI, Oakland, California, USA). Before tissue analysis, the integrity and purity of all probes were verified by hybridization to metaphase spreads of normal peripheral lymphocytes. The samples were analysed under a ×60 oil immersion objective using an Olympus BX-51 fluorescence microscope equipped with appropriate filters, a CCD (charge-coupled device) camera and the CytoVision FISH imaging and capturing software (Applied Imaging). Semi-quantitative evaluation of the tests was independently performed by two evaluators (S.P. and L.A.J.); at least 100 nuclei for each case were analysed when possible. Cases with significant differences between the two independent evaluations were refereed by a third person (M.A.R.). The statistical analysis was performed using SPSS 13.0 for Windows (SPSS Inc.) with a significance level of 0.05.

Cell lines and cell culture conditions. NCI-H2009 (ref. 39), NCI-H661 (ref. 40), NCI-H1975 (ref. 39) and HCC1171 (ref. 8) have been previously described. A549 cells were purchased from American Type Culture Collection. NSCLC cells were maintained in RPMI growth media consisting of RPMI 1640 plus 2 mM L-glutamine (Mediatech) supplemented with 10% fetal bovine serum (Gemini Bio-Products), 1 mM sodium pyruvate, and penicillin/streptomycin (Mediatech).

RNAi knockdown. shRNA vectors targeted against *NKX2-1*, *MBIP* and *GFP* were provided by TRC (The RNAi Consortium). The sequences targeted by the *NKX2-1* shRNAs are as follows: shNKX2-1a (TRCN0000020449), 5′-CGCTTGTAAATACCAGGATTT-3′, and shNKX2-1b (TRCN0000020453), 5′-TCCGTTCTCAGTGTCTGACAT-3′. The sequences targeted by the *MBIP* shRNA and *GFP* shRNA are 5′-CCACCGGAAGGAAGATTTATT-3′ (TRCN0000003069) and 5′-GCAAGCTGACCCTGAAGTTCAT-3′, respectively. Lentiviruses were made by transfection of 293T packaging cells with a three plasmid system[41,42]. Target cells were incubated with lentiviruses for 4.5 h in the presence of 8 $\mu g\,ml^{-1}$ polybrene. After the incubation, the lentiviruses were removed and cells were fed fresh medium. Two days after infection, puromycin (0.75 $\mu g\,ml^{-1}$ for NCI-H1975, 1.0 $\mu g\,ml^{-1}$ for NCI-H661, 1.5 $\mu g\,ml^{-1}$ for NCI-H2009, 1.0 $\mu g\,ml^{-1}$ for NCI-H661 and 2.0 $\mu g\,ml^{-1}$ for A549 and HCC1171) was added. Cells were grown in the presence of puromycin for 3 days or until all of the non-infected cells died. Twenty-five micrograms of total cell lysates prepared from the puro-selected cell lines was analysed by western blotting using anti-NKX2-1 polyclonal antibody (Santa Cruz Biotechnology), anti-MBIP polyclonal antibody (Proteintech Group, Inc.) and anti-vinculin monoclonal antibody (Sigma).

Soft agar anchorage-independent growth assay. NCI-H2009 (1×10^4), NCI-H661 (2.5×10^4), A549 (3.3×10^3), NCI-H1975 (5×10^4) or HCC1171 (1×10^4) cells expressing shRNAs targeting *NKX2-1*, *MBIP* or *GFP* were suspended in a top layer of RPMI growth media and 0.4% Noble agar (Invitrogen) and plated on a bottom layer of growth media and 0.5% Noble agar in 35-mm wells. Soft agar colonies were counted 3–4 weeks after plating. The data are derived from two independent experiments unless otherwise noted and are graphed as the percentage of colonies formed relative to the shGFP control cells (set to 100%) ±1 standard deviation of the triplicate samples. *P* values between shGFP and shNKX2-1 or shMBIP samples were calculated using a *t*-test.

Cell proliferation assays. NCI-H2009 (500 cells per well), A549 (400 cells per well) and NCI-H661 (600 cells per well) cells expressing shRNAs targeting *NKX2-1*, *MBIP* or *GFP* were seeded in 6 wells in a 96-well plate. Cell viability

was determined at 24-h time points for a total of 4 days using the WST-1-based colorimetric assay (Roche Applied Science). The percentage of cell viability is plotted for each cell line ± 1 standard deviation of the reading from six wells, relative to day 0 readings. Experiments were performed two or more times and a representative experiment is shown.

32. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38,** 500–501 (2006).
33. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high-density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).
34. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36,** 949–951 (2004).
35. Lin, M. *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20,** 1233–1240 (2004).
36. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132,** 365–386 (2000).
37. Nickerson, D. A., Tobe, V. O. & Taylor, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25,** 2745–2751 (1997).
38. Rubin, M. A. *et al.* Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res.* **64,** 3814–3822 (2004).
39. Phelps, R. M. *et al.* NCI-Navy Medical Oncology Branch cell line data base. *J. Cell. Biochem.*, Suppl. 24, 32–91 (1996).
40. Banks-Schlegel, S. P., Gazdar, A. F. & Harris, C. C. Intermediate filament and cross-linked envelope expression in human lung tumor cell lines. *Cancer Res.* **45,** 1187–1197 (1985).
41. Naldini, L. *et al. In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272,** 263–267 (1996).
42. Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery *in vivo*. *Nature Biotechnol.* **15,** 871–875 (1997).