Diverse somatic mutation patterns and pathway alterations in human cancers

Zhengyan Kan1,2, Bijay S. Jaiswal1, Jeremy Stinson1, Vasantharajan Janakiraman1, Deepali Bhatt1, Howard M. Stern3, Peng Yue3, Peter M. Haverty2, Richard Bourgon2, Jianbiao Zheng4, Martin Moorhead1, Subhra Chaudhuri1, Lynn P. Tomsho5, Brock A. Peters1, Kanan Pujara1, Shaun Cordes1, David P. Davis1, Victoria E. H. Carlton4, Wenlin Yuan1, Li Li5, Weiru Wang6, Charles Eigenbrot6, Joshua S. Kaminker7, David A. Eberhard5, Paul Waring3, Stephan C. Schuster5, Zora Modrusan1, Zemin Zhang2, David Stokoe1, Frederic J. de Sauvage1, Malek Faham1 & Somasekar Seshagiri1

The systematic characterization of somatic mutations in cancer genomes is essential for understanding the disease and for developing targeted therapeutics. Here we report the identification of 2,576 somatic mutations across ~1,800 megabases of DNA representing 1,507 coding genes from 441 tumours comprising breast, lung, ovarian and prostate cancer types and subtypes. We found that mutation rates and the sets of mutated genes varied substantially across tumour types and subtypes. Statistical analysis identified 77 significantly mutated genes including protein kinases, G-protein-coupled receptors such as GRM8, BAI3, AGTR1 (also called APLNR) and LPHN3, and other druggable targets. Integrated analysis of somatic mutations and copy number alterations identified another 35 significantly altered genes including GNAS, indicating an expanded role for Ga subunits in multiple cancer types. Furthermore, our experimental analyses demonstrate the functional roles of mutant GNAOI (a Ga subunit) and mutant MAP2K4 (a member of the JNK signalling pathway) in oncogenesis. Our study provides an overview of the mutational spectra across major human cancers and identifies several potential therapeutic targets.

Identification of mutated cancer genes has provided insights into the biological processes underlying tumorigenesis1. Large-scale studies published so far have primarily used Sanger sequencing to identify mutations2–5. These studies were conducted on either a small number of samples across the entire coding genome or a larger number of samples across a small number of coding genes2–5. In this study, we used mismatch repair detection (MRD) technology6–11 (Methods) to identify somatic mutations in 441 human primary tumour samples comprising 183 breast (59 HER2+, 65 hormone receptor (HR; ER/PR)+ and 59 triple negative (ER/PR/HER2 negative)), 134 lung (57 non-small-cell lung cancer (NSCLC) adenocarcinoma (adeno), 63 NSCLC squamous, 5 small cell lung cancers (SCLC) and 9 poorly differentiated tumours designated as ‘others’), 58 ovarian, 58 prostate and 10 pancreatic cancers (Supplementary Fig. 1 and Supplementary Table 1) across the coding exons and flanking splice sites of 1,507 genes (Supplementary Table 2). The 1,507 candidate genes spanned ~4 megabases (Mb) of DNA and comprised known cancer genes and druggable genes that included protein kinases, E3 ligases, deubiquitinating enzymes (DUBs), G-protein-coupled receptors (GPCRs) and other enzymes. A total of ~1,800 Mb of tumour DNA sequences was successfully analysed for the presence of tumour-specific nucleotide changes after MRD sorting10,11, variation prediction and mass spectrometric validation (Methods and Supplementary Fig. 2). Although MRD can detect variations that are as low as ~1% in abundance12, it is not efficient in detecting insertions/deletions (indels)10 that typically are a small proportion of the somatic mutations13–15, and hence this class of mutation was not included in our analysis.

A total of 2,576 somatic mutation events were identified in 967 of the 1,507 genes studied (Supplementary Table 3 and Supplementary Fig. 3), including 1,833 missense, 538 synonymous, 141 nonsense and 64 canonical splice-site changes. Of the 2,576 somatic mutations identified, 95% (2,449) have not been previously reported in COSMIC13 or in any of the recent large cancer studies3–9. A total of 37 recurrent mutations in 13 genes were found, of which mutations in EPHB1, GSK3B and RUNX1T1 were previously unknown (Supplementary Table 3). Furthermore, computational analysis using five different methods at the individual mutation level predicted 19% of the mutations to have a potential functional effect based on positive predictions from a majority of the methods applied (Supplementary Table 4 and Supplementary Fig. 4). Notably, this subset of mutations predicted to be functionally relevant were also enriched in significantly mutated candidate cancer genes (Supplementary Fig. 5).

Of the 441 tumours studied, 85% (377) carried one or more protein-altering mutation in the 1,507 genes studied. Among tumours with mutations, 91 carried a single mutation and another 225 had between two and nine mutations. The remaining 61 tumours, none of which was mismatch-repair defective (Supplementary Table 1), contained 10 or more mutations, with one lung adenocarcinoma tumour carrying 39 mutations (Supplementary Fig. 1). On average, the tumours analysed contained 1.8 protein-altering mutations per Mb of DNA, with individual cancers and subtypes showing wide deviations from this background rate (Supplementary Fig. 1). Among the 150 tumours types analysed, we found lung adenocarcinomas and squamous carcinomas to have high protein-altering mutation rates of 3.5 and 3.9 per Mb, respectively. In contrast, prostate cancers, where ~75% carried the TMPRSS2–ERG gene fusion16, had a low mutation rate of 0.33 per Mb.

Statistical approaches can be used to identify significantly mutated cancer genes17, although they may not capture rare functionally relevant changes. Applying a Poisson probability-based approach18 we identified 77–19 in lung squamous carcinoma, 18 in lung adenocarcinoma, 5 in small cell lung cancer, 10 in HR+ breast, 6 in HER2+ breast cancers. [...]

1Department of Molecular Biology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA. 2Department of Bioinformatics, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA. 3Department of Pathology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA. 4Affymetrix Inc, 3420 Central Expressway, Santa Clara, California 95051, USA. 5Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, 310 Wartik Lab, University Park, Pennsylvania 16802, USA. 6Department of Protein Engineering, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA.
Prostate cancers (Supplementary Table 5)—as having significant prevalence of protein-altering mutations, with a $q$-score $\geq 1.0$ ($\approx 10\%$ false discovery rate; see Methods). Furthermore, the set of significantly mutated genes varied across each tumour type and subtype, indicating the complexity of the genetic mechanisms underlying carcinogenesis (Fig. 1). We further confirmed the relevance of the significantly mutated cancer genes by analysing 50 additional lung squamous carcinoma samples for somatic mutations in 18 genes using 454-sequencing. We identified and validated 56 additional somatic mutations and found that the frequency of mutations in this prevalence screen was comparable to that seen in the primary screen for a majority of the significantly mutated novel cancer genes including GRM8, BAI3 and KEAP1 (Fig. 2, Supplementary Fig. 6 and Supplementary Table 6).

Among the major class of genes found to be mutated were the kinases where we found 315 mutations in 157 of the 230 protein kinases studied, including mutations in 49 receptor tyrosine kinases (RTKs). In particular, we found 27 distinct somatic mutations in 11 ephrin receptors, 22 of which were from lung cancers (Supplementary Fig. 7).

In addition to kinases, we examined a total of 156 GPCRs and found 87 to be mutated. Of the 87 genes, 13 were significantly mutated, including LPHN3, GRM8, CMKLR1, MASL1, AGTR1 and PTGFR (Fig. 2a and Supplementary Tables 4 and 5). GRM8 is a member of the metabotropic receptor family that is implicated in cancer. GRM8 was mutated in 8% (5 of 63 or 9 of 103) of NSCLC squamous subtype tumours and GRM1 in 7% (4 of 57) of NSCLC adenocarcinomas (Fig. 2a and Supplementary Fig. 8). In addition to GRM, we found mutations in BAI3, a member of the brain-specific angiogenesis inhibitor (BAI) family of GPCR, to be significantly mutated in lung cancers (Supplementary Fig. 7).

To gain a more complete understanding of the mutations relevant to cancer, we performed an integrated analysis by combining comparative genome hybridization array (CGH)-derived copy number alteration profiles obtained for a majority of the tumours screened for somatic mutations. Analysis of this data in combination with the mutation data using alteration $q$-score, a statistical significance measure (Methods), identified 35 additional cancer genes which included GNAS, STK11 and EPHB1 (Supplementary Table 5). In particular, GNAS—the G-protein α subunit, known to be associated with multiple human diseases, including some cancers—besides being mutated at R201, was also found to be amplified in 12% (6 of 49) of ovarian cancers, 20% (10 of 50) of HER2 breast cancers and 13% (7 of 53) of HR breast cancers (Fig. 3b, c). Furthermore, we found several significantly mutated genes, and genes like RUNX1T1, SPOP and GRM8 that carry mutations predicted to have functional effects, to be amplified, indicating that these genes may function as oncogenes (Fig. 3a and Supplementary Figs 9 and 10).

In addition to mutations in GNAS, we found the Gα subunit GNAO1 to be mutated at residue R243, a position analogous to GNAS residue R265, which is known to be mutated in Albright hereditary osteodystrophy. Furthermore, structural analysis predicted that this mutation is also likely to impair the GTPase activity of GNAO1 (Fig. 3b), as does mutation at R201 in GNAS. Consistent with this, when overexpressed in human mammary epithelial cells, we found that the R243H mutation of GNAO1 promoted anchorage-independent growth (Fig. 3d–f), possibly through STAT3 signalling. Recently, activating mutations in GNAQ, another Gα subunit, in uveal and blue naevi were reported. Together, these data indicate that the Gα subunit alterations potentially have a more significant role in cancers than previously thought. In contrast to amplifications, genes showing recurrent copy number losses typically have a role in tumour suppression. We found known tumour suppressors such as NF1, FBXW7, PTEN and CDKN2A to be frequently mutated or deleted in multiple cancers (Fig. 3a), expanding the role of these tumour suppressors in additional tumour types.

Understanding mutations in genes involved in signalling cascades can aid the identification of pathways critical for tumorigenesis and provide strategies for therapeutic intervention in specific tumour types. Statistical evaluation of the combined prevalence of somatic mutations and copy number alterations at the pathway level (Methods) found

Figure 1 | Diverse patterns of significantly mutated genes across cancer subtypes. Each solid circle represents a gene and the size of the circle is proportional to the mutation count for that gene. The genes are represented in alphabetical order from left to right on the x axis. Selected genes with significant $q$-scores are labelled.
RTKs signalling together with the RAS/MAPK pathway (RTK/RAS) to be one of the most significantly altered pathways across all cancer types, except for prostate (Fig. 4a, b, Supplementary Fig. 11 and Supplementary Table 7). In addition to the RTK/RAS pathway we found the GPCR pathway genes also to be significantly mutated in lung adenocarcinoma and lung squamous carcinoma, and the JNK pathway to be significantly mutated in HR+ breast cancer (Supplementary Table 7).

Because the JNK signalling pathway was significantly altered, we studied the role of MAP2K4, a JNK pathway kinase, in oncogenic signalling. Whereas the presence of homozygous MAP2K4 deletions...
in multiple cancers has indicated a tumour suppressor role for MAP2K4, results from a few other studies have implicated an oncogenic role for this kinase. To understand the relevance of MAP2K4, results from a few other studies have implicated an oncogenic role for this kinase. To understand the relevance of MAP2K4, we have interrogated 22 of the 26 published significant lung adenocarcinoma genes in our lung adenocarcinoma cancer genes. The diverse spectrum of mutational and genomic changes from largescale sequencing studies, including this, show that each tumour is unique even within a given type and subtype. Understanding the set of changes at the individual patient level will enable personalized treatment. Furthermore, patient selection based on tumour mutational profile and genomic alterations for clinical drug testing will be critical for successful development of new treatments.

METHODS SUMMARY

Frozen human tumour tissue with patient-matched frozen normal tissue was procured from commercial vendors with appropriate institutional approval. Tumour tissues with ≥50% viable neoplastic cell content were used for nucleic acid isolation. Tumour DNA was scanned for mutations using MRD and the predicted somatic mutations were further validated by nucleic acid mass spectrometry (Supplementary Fig. 2). Tumour samples were also analysed for copy number variations using the Agilent 244K CGH array.
Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 July 2009; accepted 27 May 2010.
Published online 28 July 2010.


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

Acknowledgements We acknowledge Genentech DNA Sequencing, Oligo and Microarray, and FACS labs for their help with the project. We thank the Genentech Bioinformatics group for informatics infrastructure support and the Pathology Core Labs for providing histology, immunohistochemistry and tissue management support. We also thank U. Vitt and W. Forrest for their help during the course of this project, and L. Phillips and D. Shames for their help with preparation of this manuscript.

METHODS
Genes and samples. The 1,507 genes studied comprised known oncogenes, tumour suppressors, as well as ‘druggable targets’ that included protein kinases, E3 ligases, DUBs, GPCRs and other genes with enzymatic activities (Supplementary Table 2). Patient-matched frozen tumour and normal tissue were obtained from multiple commercial vendors. Tumour and normal match was confirmed using a panel of 15 single nucleotide polymorphisms (SNPs)11,12. Tumour samples with ten or more somatic mutations were assessed for microsatellite instability using an MSI detection kit (Promega).

The samples used in the study were selected from 730 breast, 750 lung (662 NSCLC and 88 SCLC), 218 ovarian, 150 prostate and 96 pancreatic patient tumour samples based on histological diagnosis and/or immunohistochemistry (IHC) analysis of tumour, assessment of tumour content, absence of tumour cells in normal, availability of matched adjacent normal tissue and the yield of high-quality nucleic acid (DNA and RNA).

Each tissue was subclass and evaluated histologically for both quality and diagnosis by a board-certified pathologist. Classification of tumours was primarily performed based on histology. In some cases, IHC was used to help classify poorly differentiated tumours, and included chromogranin A and synaptophysin for SCLC, and TTF1 and cytokeratin 5 for lung adenocarcinoma (TTF1+, CK5–) and squamous cell carcinoma (TTF1–, CK5+)13. IHC using the HercepTest (Dako) and/or HER2 FISH (Abbott PathVision kit) was used to identify HER2+ breast samples. Samples that were negative for HER2 by IHC (0 or 1+) were subjected to ER and PR IHC. Genomic DNA from prostate samples was tested for TMPRSS2-ERG fusion status using a TaqMan assay (5p primer, TAGGCCACCTTGAAAGCATG; 3p primer, GTTACACTCTAAACACCGG; and probe, CTCACCTGACACTGTATAGGCTTGCC) following recommended protocols (Applied Biosystems).

Somatic mutation and genomic alteration detection. We used Escherichia coli-based mismatch repair detection (MRD) to scan individual tumour samples for mutations. Matched normal sample corresponding to each of the tumour, in pools of five, was also screened for variations against the reference DNA to help identify germline changes that may not be represented in dbSNP. The E. coli served as a sorter for mutant alleles present in a DNA sample relative to the pools of five, was also screened for variations against the reference DNA to help identify somatic mutations relative to the background using Poisson probabilities. The background rates of protein-altering change were extrapolated from the empirical rates of synonymous mutations using estimated ratios of protein-altering (NS) versus synonymous (S) mutations.

All nucleotide positions in the protein-coding region of a gene (g) and validated somatic mutations were classified into one of six nucleotide categories (i) defined by the base of interest and the preceding base (Supplementary Table 10). For each nucleotide category and cancer type (d), we collected statistics on the numbers of protein-coding bases (n), protein-altering somatic mutations (x) and synonymous mutations (y) for individual genes and in aggregate. The NS/S ratios were determined by a simulation where we mutated all residues to every possible allele (a) in protein-coding regions and canonical splice sites of the transcript sequences based on observed somatic mutation probabilities F and determined whether a particular change would result in a synonymous or protein-altering change. We then calculated NS/S ratios (r) as the aggregate number of simulated protein-altering and synonymous changes in each nucleotide category. The background mutation rate fbg is estimated as the number of observed synonymous somatic mutations in each cancer type and nucleotide category multiplied by the NS/S ratios and averaged over all protein-coding sequences analysed:

\[ F_{bg} = \frac{m_{ga}}{\sum_{a=1}^{m_{ga}}} \]

where \( m_{ga} \) is the number of somatic changes from nucleotide type i to allele a

\[ f_{bg} = \frac{s_{i,d} r_{i,d} n_{j,d}}{x_{g,d}} \]

Statistical analysis of combined cancer alterations. In addition to somatic mutation, copy number gains and losses can promote carcinogenesis. The scope of cancer genes can be best delineated by taking into account such somatic genomic alterations and mutations. Here we describe a statistical score called alteration q-score for assessing whether a gene is likely to be involved in cancer based on the combined prevalence of somatic mutation, and copy number gains and losses. In addition to p-values based on mutation prevalence, we calculated for each gene p-values based on the prevalence of copy number alterations using binomial probabilities. Copy number alteration statuses were identified for the 1,507 genes studied across 331 tumour samples using CGH arrays. In each cancer type, we calculated the frequency of copy number alteration as the number of samples with a given copy number value of probes within that segment. For each gene and tumour sample, one of three copy number alteration statuses was determined based on GLAD copy number values: gain (\( g \geq 3 \)), loss (\( l \leq 1.25 \)) and wild type. Copy number alteration and somatic mutation statuses for each gene and tumour sample were integrated to derive the molecular alteration status at the tumour sample based on histological diagnosis and/or immunohistochemistry (IHC) analysis of tumour, assessment of tumour content, absence of tumour cells in normal, availability of matched adjacent normal tissue and the yield of high-quality nucleic acid (DNA and RNA). Usually exhibit a higher-than-expected mutation count due to enrichment of driver mutations. A statistical score called mutation q-score was calculated for each gene based on the probability that protein-altering mutations observed in that gene consist of only passenger mutations occurring at background rates. A higher q-score indicates that the observed mutation count in a gene is more likely not attributable to chance, indicating a functional role in tumour growth and development. This method is similar to the CaMP (cancer mutation prevalence) score14 and incorporated subsequent refinements15,16. For each gene in a cancer type, the q-score is calculated by comparing the observed prevalence of protein-altering somatic mutations relative to the background using Poisson probabilities. The background rates of protein-altering mutations were extrapolated from the empirical rates of synonymous mutations using estimated ratios of protein-altering (NS) versus synonymous (S) mutations.

\[ F_{bg} = \frac{x_{g,d} p_{g,d}}{\sum_{k=0}^{\infty} \frac{x_{g,d}^k e^{-x_{g,d}}}{k!}} \]

Statistical analysis of combined cancer alterations. In addition to somatic mutation, copy number gains and losses can promote carcinogenesis. The scope of cancer genes can be best delineated by taking into account such somatic genomic alterations and mutations. Here we describe a statistical score called alteration q-score for assessing whether a gene is likely to be involved in cancer based on the combined prevalence of somatic mutation, and copy number gains and losses. In addition to p-values based on mutation prevalence, we calculated for each gene p-values based on the prevalence of copy number alterations using binomial probabilities. Copy number alteration statuses were identified for the 1,507 genes studied across 331 tumour samples using CGH arrays. In each cancer type, we calculated the frequency of copy number alteration as the number of samples harbouroing copy number alterations for a gene (\( x_{g,d} \)) divided by the number of samples analysed (\( n_{j,d} \)). The background frequencies (\( f_{bg} \)) were empirically estimated as the mean frequency of alteration across all genes analysed in a cancer type (Supplementary Table 10).
Plasmids, reagents, cells and assays. N-terminally Flag-tagged MAP2K4 and GNAO1 wild-type cDNAs (Origene) or mutants generated by site-directed mutagenesis (Stratagene) were cloned into pRetro-IRESGFP (Clontech) vector for further study. Anti-Flag M2 antibody and EZview Flag-M2-antibody-coupled affinity gel (Sigma Life Science), HRP-conjugated secondary antibodies and chemiluminescence Super signal West Dura substrate (Thermo Fisher Scientific), Cell lysis buffer and Kinase buffer (Cell Signalling Technology) were used in this study.

Human mammary epithelial cells (HMECs; Invitrogen) engineered to express hTERT, and large and small T antigen were maintained in HuMEC medium (Invitrogen) supplemented with HuMEC supplement, bovine pituitary extract, 100 U ml\(^{-1}\) penicillin and 100 mg ml\(^{-1}\) streptomycin. The NIH3T3 (ATCC) and Phoenix cells (OriBeng) were maintained in DMEM supplemented with 10% (v/v) fetal bovine serum, 2 mm l-glutamine, 100 U ml\(^{-1}\) penicillin and 100 mg ml\(^{-1}\) streptomycin. HMEC and NIH3T3 stably expressing MAP2K4 or GNAO1 wild type and mutants were generated as previously described.

Expression of the Flag-tagged proteins was confirmed by western blot analysis using cell lysates. Briefly, cell lysates were resolved on a 4–20% SDS–PAGE (Invitrogen) and transferred onto a nitrocellulose membrane and detected using appropriate primary antibody, HRP-conjugated secondary antibody and chemiluminescence Super signal West Dura chemiluminescence detection substrate. Anchor-independent colony formation assay, HMEC (3 \times 10^6) or NIH3T3 (1 \times 10^6) cells stably expressing either empty vector, wild-type or indicated MAP2K4 and GNAO1 mutants were suspended in full media containing 0.35% agar and plated on a lower layer of 0.5% agar in 6-well plates. After 3 weeks, the presence of colonies was assessed using Gel count imager (Oxford Optronics Ltd). The number of colonies in each plate was quantified using Gel count software (Oxford Optronix Ltd). A Student’s t-test (two-tailed) was used for statistical analyses to compare colonies formed by wild-type and mutant expressing cells using GraphPad Prism 5.00 (GraphPad Software). A P-value < 0.05 was considered statistically significant (*P < 0.05 and **P < 0.01).

Kinase assay. NIH3T3 cells were serum starved for 24 h, washed with PBS and lysed with Cell lysis buffer (20 mm Tris-HCl pH 7.5, 150 mm NaCl, 1 mm EDTA, 1% Triton X-100, 2.5 mm sodium pyrophosphate, 1 mm β-glycerophosphate, 1 mm Na3VO4, 1 μg ml\(^{-1}\) leupeptin and 1 mm PMSF). The clarified lysates were incubated overnight at 4°C with anti-Flag-M2 antibody coupled beads. The beads were then spun down at 10,000g for 30 s and washed three times using the Cell lysis buffer. Flag-M2-immunoprecipitated protein beads containing MAP2K4 wild type or mutants, adjusted for protein concentration, were incubated for 30 min at 30°C either with Myelin basic protein (MBP; Millipore Inc., MA) or inactive MAP2K7/NK2 (Invitrogen) and 2 μCi [γ\(^{32}\)P]ATP (Perkin Elmer) in 25 μl of kinase reaction buffer (25 mm Tris-HCl (pH 7.5), 5 mm β-glycerophosphate, 2 mm dithiothreitol, 0.1 mm Na3VO4, 10 mm MgCl\(_2\) and 10 μM ATP). The kinase reaction was stopped by adding SDS sample buffer. The samples were then resolved by SDS–PAGE. The gel was fixed, dried, exposed to phosphor screen and scanned on a Typhoon phosphorimager (GE Healthcare).


