ARTICLES

Patterns of somatic mutation in human cancer genomes

Christopher Greenman[1], Philip Stephens[1], Raffaella Smith[1], Gillian L. Dalgliesh[1], Christopher Hunter[1], Graham Bignell[1], Helen Davies[1], Jon Teague[1], Adam Butler[1], Claire Stevens[1], Sarah Edkins[1], Sarah O'Meara[1], Imre Vastrik[2], Esther E. Schmidt[2], Tim Avis[1], Syd Barthorpe[1], Gurpreet Bhamra[1], Gemma Buck[1], Bhudipa Choudhury[1], Jody Clements[1], Jennifer Cole[1], Ed Dicks[1], Simon Forbes[1], Kris Gray[1], Kelly Halliday[1], Rachel Harrison[1], Katy Hills[1], Jon Hinton[1], Andy Jenkinson[1], David Jones[1], Andy Menzies[1], Tatiana Mironenko[1], Janet Perry[1], Keiran Raine[1], Dave Richardson[1], Rebecca Shepherd[1], Alexandra Small[1], Calli Tofts[1], Jennifer Varian[1], Tony Webb[1], Sofie West[1], Sara Widaa[1], Andy Yates[1], Daniel P. Cahill[3], David N. Louis[3], Peter Goldstraw[4], Andrew G. Nicholson[4], Francis Brasseur[5], Leendert Looijenga[6], Barbara L. Weber[7], Yoke-Eng Chiew[8], Anna deFazio[8], Mel F. Greaves[9], Anthony R. Green[10], Peter Campbell[1], Ewan Birney[2], Douglas F. Easton[11], Georgia Chenevix-Trench[12], Min-Han Tan[13], Sok Kean Khoo[13], Bin Tean Teh[13], Siu Tsan Yuen[14], Suet Yi Leung[14], Richard Wooster[1], P. Andrew Futreal[1] & Michael R. Stratton[1,9]

Cancers arise owing to mutations in a subset of genes that confer growth advantage. The availability of the human genome sequence led us to propose that systematic resequencing of cancer genomes for mutations would lead to the discovery of many additional cancer genes. Here we report more than 1,000 somatic mutations found in 274 megabases (Mb) of DNA corresponding to the coding exons of 518 protein kinase genes in 210 diverse human cancers. There was substantial variation in the number and pattern of mutations in individual cancers reflecting different exposures, DNA repair defects and cellular origins. Most somatic mutations are likely to be 'passengers' that do not contribute to oncogenesis. However, there was evidence for 'driver' mutations contributing to the development of the cancers studied in approximately 120 genes. Systematic sequencing of cancer genomes therefore reveals the evolutionary diversity of cancers and implicates a larger repertoire of cancer genes than previously anticipated.

Cancers are clonal proliferations that arise owing to mutations that confer selective growth advantage on cells. The mutated genes that are causally implicated in cancer development are known as 'cancer genes' and more than 350 have thus far been identified (ref. 1 and http://www.sanger.ac.uk/genetics/CGP/Census/). Cancer genes have been identified by several different physical and genetic mapping strategies, by biological assays and as plausible biological candidates. Each of these approaches has identified a subset of cancer genes, leaving the possibility that others have been overlooked. The provision of the human genome sequence, therefore, led to the proposal that systematic resequencing of cancer genomes could reveal the full compendium of mutations in individual cancers and hence identify many of the remaining cancer genes[2].

Somatic mutations occur in the genomes of all dividing cells, both normal and neoplastic. They may occur as a result of misincorporation during DNA replication or through exposure to exogenous or endogenous mutagens. Cancer genomes carry two biological classes of somatic mutation arising from these various processes. 'Driver' mutations confer growth advantage on the cell in which they occur, are causally implicated in cancer development and have therefore been positively selected. By definition, these mutations are in 'cancer genes'. Conversely, 'passenger' mutations have not been subject to selection. They were present in the cell that was the progenitor of the final clonal expansion of the cancer, are biologically neutral and do not confer growth advantage. A challenge to all systematic mutation screens will, therefore, be to distinguish driver from passenger mutations. However, the prevalence and characteristics of driver and passenger mutations in cancer genomes are not currently well defined. The aim of these studies was to survey the numbers and patterns of somatic point mutations in a diverse set of human cancer genomes and hence to obtain insights into the relative contributions of driver and passenger mutations.

Somatic protein kinase mutations

The protein kinase gene family was selected for these studies because the protein kinase is the domain most commonly found among known cancer genes[1] and because inhibitors of mutated protein kinases have recently shown remarkable efficacy in cancer treatment[3].

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [3]Molecular Pathology Unit, Neurosurgical Service and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [4]Royal Brompton Hospital, London SW3 6NP, UK. [5]Ludwig Institute for Cancer Research, 1200 Brussels, Belgium. [6]Laboratory of Pathology/Experimental Patho-Oncology, Erasmus MC University Medical Center Rotterdam, Daniel den Hoed Cancer Center, Josephine Nefkens Institute, 3000 DR Rotterdam, UCL 745, B-1200, The Netherlands. [7]University of Pennsylvania Cancer Centre, Philadelphia, Pennsylvania 19104-6160, USA. [8]Department of Gynaecological Oncology, Westmead Hospital and Westmead Institute for Cancer Research, University of Sydney at the Westmead Millennium Institute, Westmead NSW 2145, Australia. [9]Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [10]Department of Haematology, Addenbrooke's NHS Trust and University of Cambridge, Cambridge CB2 0QQ, UK. [11]Cancer Research UK Genetic Epidemiology Unit, University of Cambridge, Cambridge CB1 8RN, UK. [12]Queensland Institute of Medical Research, Royal Brisbane Hospital, Herston, Queensland 4029, Australia. [13]Van Andel Research Institute, Grand Rapids, Michigan 49503, USA. [14]Department of Pathology, The University of Hong Kong, Queen Mary Hospital, Pokfulam Road, Hong Kong.

Furthermore, the coding sequences of the protein kinases (Supplementary Table 3) constitute a much larger sample of cancer genome, approximately 1.3 Mb of DNA per case, than has previously been analysed across many cancer types, thus permitting insights into the general patterns of somatic mutation in human cancers.

Human cancers ($n = 210$) including breast, lung, colorectal, gastric, testis, ovarian, renal, melanoma, glioma and acute lymphoblastic leukaemia (Supplementary Table 3) were screened for somatic mutations in the coding exons and splice junctions of the 518 protein kinase genes[4]; a total of 274 Mb of cancer genome. Of the 210 cancers analysed 169 were primary tumours, 2 were early cultures and 39 were immortal cancer cell lines.

One-thousand-and-seven somatic mutations were detected (Supplementary Table 2 and http://www.sanger.ac.uk/genetics/CGP/Studies/). Of these, 921 were single base substitutions, 78 were small insertions or deletions and 8 were complex changes, usually double nucleotide substitutions. Of the single base substitutions, 620 encoded mis-sense changes, 54 caused nonsense changes, 28 were at highly conserved positions of splice junctions and 219 were synonymous (silent) mutations. Approximately one-third of these mutations have previously been reported[5–8].

Prevalence of somatic mutations

Although there is extensive information on the prevalence of somatic rearrangements and copy number changes in human cancer genomes (from studies using cytogenetics and comparative genomic hybridization) there has previously been limited insight into the prevalence of somatic point mutations[5,6,8–10]. The results of the current studies show that the number of somatic point mutations varies widely both within and between classes of cancer (Fig. 1 and Supplementary Fig. 1).

Seventy-three out of the two-hundred-and-ten cancers showed no somatic mutations at all, whereas others showed exceptionally large numbers (Fig. 1 and Supplementary Fig. 1). The highest mutation prevalence (~77 mutations per Mb) was in two gliomas that were recurrences after treatment with the anticancer drug temozolomide, an alkylating agent that is a known mutagen[7,11,12]. Some individual melanomas and lung cancers also showed substantial numbers of mutations that may relate to the extent of past exposure to ultraviolet radiation (UV) and tobacco smoke carcinogens, respectively. Abnormalities in DNA repair also influenced the number of somatic mutations. Five cancers with defective DNA mismatch repair leading to microsatellite instability had a high prevalence of both base substitutions (14–40 per Mb) and small insertions and deletions at polynucleotide tracts (5–12 per Mb). Occasional cancers without known prior treatment, defects in DNA repair or mutagenic exposure also showed very large numbers of mutations.

Excluding individual cancers with known DNA repair defects or previous treatment, there were differences in overall mutation prevalence between different cancer types (Table 1). Among primary cancers, lung carcinomas showed the highest prevalence of somatic mutations (4.21 per Mb), followed by gastric cancers (2.10 per Mb), ovarian cancers (1.85 per Mb), colorectal cancers (1.21 per Mb, a prevalence similar to that previously reported[10]) and renal cancers (0.74 per Mb). Conversely, testis cancers (0.12 per Mb), lung carcinoids (0 per Mb) and most breast cancers (0.19 per Mb) manifested a much lower prevalence of mutations. The cancer types with high mutation prevalence mainly originate from high turnover, surface epithelia that are subject to recurrent exogenous mutagen exposure (for example, colorectal, lung and gastric). However, other less well understood factors may have a role. For example, the prevalence of somatic mutations in ovarian cancer was higher than that of colorectal cancer. Most ovarian cancers are thought to arise from the specialized peritoneal lining overlying the ovary (or ovarian inclusion cysts deriving from it), for which major exogenous exposures are not recognized and, unlike normal colorectal epithelium, is not thought to be rapidly turning over.

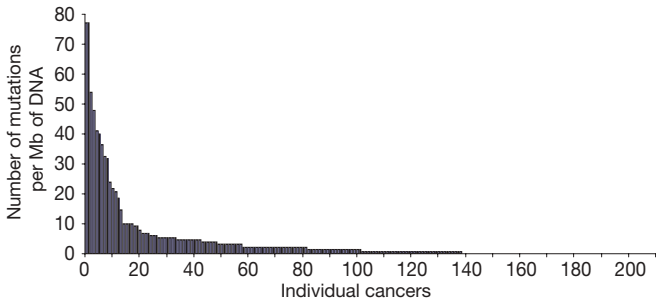

Figure 1 | The prevalence of somatic mutations in human cancer genomes. The number of somatic mutations (base substitutions, insertions/deletions and complex mutations) per Mb of DNA in 210 individual human cancers.

Signatures of somatic mutation

The large numbers of somatic mutations found in this screen also allow comparison of the mutational signatures of cancers. These signatures can carry the specific imprint of previous mutagenic exposures or DNA repair defects and hence provide insights into cancer aetiology. Signatures derived in the past from driver mutations in known cancer genes, notably *TP53* (see http://www-p53.iarc.fr/index.html), have been informative but are inevitably influenced by biological selection, which distorts the patterns generated by the underlying mutational processes. In contrast, in systematic mutation screens most somatic mutations turn out to be passengers (see below) and are therefore not affected by selection.

Mutational signatures differed between cancer types (Fig. 2). In the lung cancers, melanomas and glioblastomas studied they may reflect previous exposure to tobacco carcinogens, UV light and mutagenic alkylating chemotherapy, respectively[6,7]. However, the pathogenesis of other mutational signatures is not understood. For example, we previously showed that a subset of breast cancers has an unusual mutational signature characterized by a high prevalence of C:G>G:C transversions (Fig. 2) that occur in a specific sequence context, at TpC/GpA dinucleotides[5]. We now demonstrate that C:G>G:C changes in lung, ovarian and other cancers are also strongly enriched at TpC/GpA dinucleotides (Table 2), indicating that the underlying mutational process may be more widespread than previously appreciated. In contrast, the TpC/GpA sequence context was not observed in germline C:G>G:C polymorphisms in the protein kinases, suggesting that the process is restricted to cancer cells (Supplementary Table 4). The biological basis of this mutational signature remains

Table 1 | Somatic mutation prevalence by cancer type.

Cancer type	Mutations per Mb of DNA	Number of samples	Number of mutations
ALL	0.57	8	2
Breast	2.70 (†0.19)	16	56
Colorectal	1.21	28	44
Gastric	2.10	18	49
Glioma	22.37 (‡0.32)	9	69
Lung carcinoma	4.21	20	109
Lung carcinoid	0.00	6	0
Ovarian	1.85	25	60
Renal	0.74	23	22
Testis	0.12	13	2
MMR-deficient	32.29	5	209
Melanoma*	18.54	6	144
Other cell lines*	5.64	33	241
All tissues	3.93	210	1,007

ALL, acute lymphoblastic leukaemia; MMR-deficient, mismatch-repair-deficient cancers (two colorectal, two gastric and one ovarian).
* All samples except those indicated are primary cancers or early cultures.
† Removing the single breast cancer PD0119 decreases the breast mutation prevalence to 0.19 per Mb.
‡ Removing temozolomide-exposed PD1487 and PD1489 reduces the glioma mutation prevalence to 0.32 per Mb.

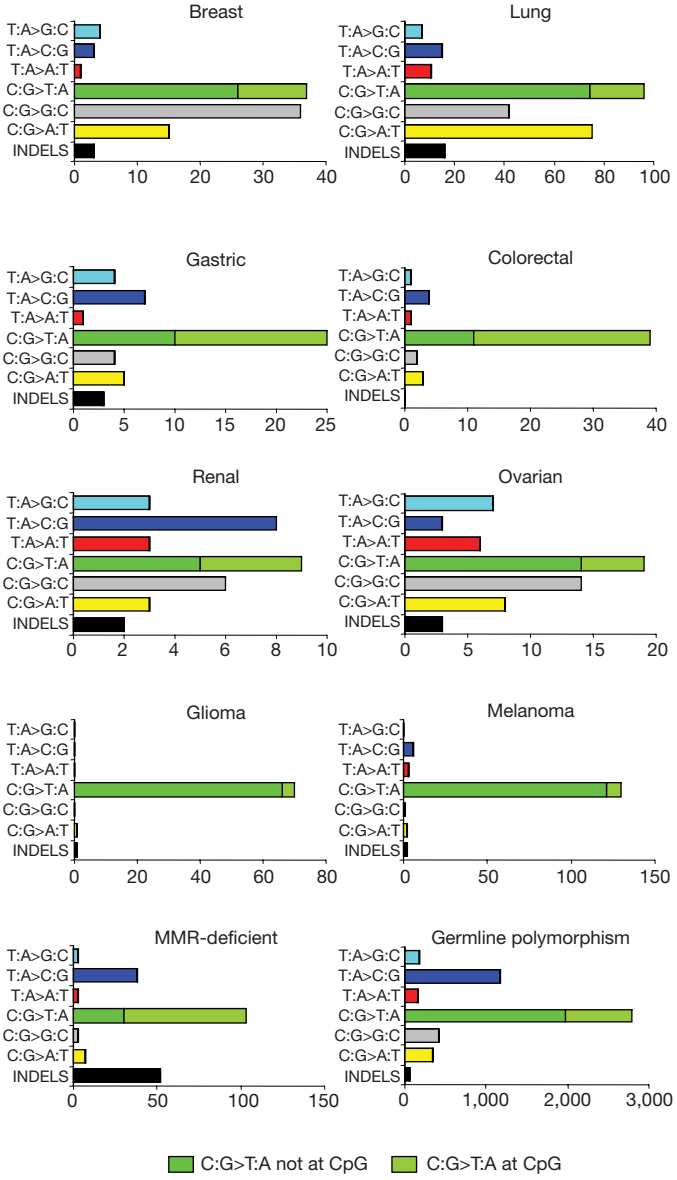

Figure 2 | Mutation spectra of human cancers by tumour type. The numbers of each of the six classes of base substitution and insertion/deletions are shown. C:G>T:A substitutions have been divided into those at CpG dinucleotides and those not at CpG dinucleotides. The data for germline polymorphisms were generated from the protein kinase screen. The data from the two colorectal, two gastric and ovarian cancers that were mismatch-repair-deficient have been shown separately (MMR-deficient).

unknown and may be due to a defect in DNA repair or a shared mutagenic exposure.

Prevalence of driver and passenger mutations

Sequencing the coding exons of the 518 kinases yielded 921 base substitution somatic mutations. These were annotated as non-synonymous (changing an amino acid) or synonymous (not changing an amino acid). To investigate the numbers of driver and passenger mutations we examined the observed ratio of non-synonymous: synonymous mutations compared with that expected by chance alone[13,14] (see Supplementary Methods for details). The underlying assumption of the analysis is that biological selection is exerted mainly on non-synonymous mutations because these may alter the structure and function of proteins. Conversely, synonymous mutations are generally biologically silent and hence cannot be selected. Therefore, a higher ratio of non-synonymous:synonymous mutations compared with that expected by chance indicates positive selection overall (selection pressure > 1) and is indicative of the presence of driver mutations. A lower non-synonymous:synonymous ratio compared with that expected by chance indicates negative selection overall (selection pressure < 1). This approach has been widely used in studies of selection during evolution[15]. In these analyses we have corrected for several other factors that might influence the non-synonymous:synonymous ratio (see Methods). We are, therefore, interpreting deviation from the expected ratio as owing to selection. However, we cannot completely exclude the existence of other, currently cryptic, factors that might influence the non-synonymous: synonymous ratio and hence imitate the effects of selection.

The selection pressure of all 921 base substitution mutations was 1.29 (95% confidence interval, 1.10–1.51; $P = 0.0013$), demonstrating an excess of non-synonymous mutations compared with that expected and thus providing evidence for the existence of driver mutations within the set. Eleven out of the nine-hundred-and-twenty-one mutations (eight in *BRAF* and three in *STK11*) would have been clearly implicated, on the basis of prior knowledge, in the development of the cancers analysed[16,17]. Removing these mutations, however, only marginally reduces the selection pressure to 1.28 ($P = 0.0025$), indicating that most driver mutations detected were not previously known to be involved in oncogenesis.

To evaluate further the significance of this observation, genes carrying non-synonymous somatic mutations in each cancer type were examined in additional series of each cancer. An additional 454 cancers were examined in this follow-up screen and 91 additional somatic mutations were identified (see Supplementary Information). The selection pressure among this set of mutations was 1.66, indicating that the gene set examined in the follow-up screen was enriched in cancer genes compared with the main screen (selection pressure 1.29, see above), supporting the notion that a proportion of protein kinases harbour oncogenic, driver mutations.

The numbers of passenger and driver mutations present can be estimated from these results (see Supplementary Methods). Of the 921 base substitutions in the primary screen, 763 (95% confidence interval, 675–858) are estimated to be passenger mutations. Therefore, the large majority of mutations found through sequencing cancer genomes are not implicated in cancer development, even when the search has been targeted to the coding regions of a gene family of high candidature. However, there are an estimated 158 driver mutations (95% confidence interval, 63–246), accounting for the observed positive selection pressure. These are estimated to be distributed in 119 genes (95% confidence interval, 52–149). The number of samples containing a driver mutation is estimated to be 66 (95% confidence interval, 36–77). The results, therefore, provide statistical evidence for a large set of mutated protein kinase genes implicated in the development of about one-third of the cancers studied.

Table 2 | Sequence context of C:G>G:C mutations.

5′ base	Breast	Lung	Others	Germ line	Expected
A	1	6	9	90	20%
C	0	4	4	102	28%
G	0	3	4	114	25%
T	35	29	16	99	26%

Base counts immediately 5′ to cytosine at C:G>G:C somatic mutations and germline variants. The expected percentages were derived from all screened C:G base pairs in the coding sequences of the protein kinases.

Characteristics of driver mutations

To gain further insights into the nature of the driver mutations in protein kinases, we examined how the selection pressure varied among different subsets of mutations. There was no significant difference in selection pressure between mis-sense (1.27), nonsense (1.58) and splice site mutations (1.23) ($P = 0.3363$) or between histological classes of cancer. However, the selection pressure was lower in cancers with defective DNA mismatch repair (MMR) (selection pressure 1.08; $P = 0.72$) compared with MMR-proficient cancers

Figure 3 | P-loop and activation segment mutations. ClustalW multi-sequence alignment of P-loop and activation segments with all positions of mis-sense mutations highlighted with underline/yellow. Positions of BRAF mutations are shown, with previously identified mutations highlighted in blue and mutations from the current study with underline/yellow. The gene name is indicated on the left. Mutations identified in the study are given to the right of the sequence.

(selection pressure 1.35; $P = 0.00089$). As reported above, MMR-deficient cancers have a higher prevalence of base substitutions than MMR-proficient cancers, presumably due to an increased mutation rate. The lower selection pressure in MMR-deficient cancers is therefore compatible with a model in which driver mutations are overwhelmed by passenger mutations.

Many previously described activating mutations in protein kinase genes that contribute to cancer development are in the kinase domain (see http://www.sanger.ac.uk/genetics/CGP/cosmic/). However, the selection pressure was only slightly higher (1.40) among mutations within kinase domains compared with mutations outside (1.23; $P = 0.08$). Mutations within the P loops and activation segments of kinase domains, in which activating mutations in cancer are often located (Fig. 3), showed a selection pressure of 1.75. Overall, the analysis suggests that, although there may be greater selection pressure for kinase domain mutations, many driver mutations are not in the kinase domains.

There were differences in selection pressure between the ten subclasses[4] of protein kinase ($P = 0.04$) with the highest in calmodulin-dependent protein kinases (1.59), atypical/other kinases (1.32) and tyrosine kinase like kinases (1.33). Many previously reported protein kinase cancer genes have been members of the tyrosine kinase or serine/threonine kinase subclasses. These analyses suggest that other subclasses are also contributing to cancer development.

Potential protein kinase cancer genes

To define further which protein kinases are likely to be carrying driver mutations, the 518 genes have been ranked according to the probability that each is carrying at least one driver mutation, conditional on the selection pressure estimate for each gene (Table 3; Supplementary Table 5; and see Methods). *BRAF* and *STK11* are second and sixteenth in this ranking, providing validation of this indicator. Remarkably, the gene at the top of this statistical ranking is Titin (*TTN*), which carries 63 non-synonymous and 13 synonymous mutations. The selection pressure associated with *TTN* is only 2.04 compared with 8.36 and 7.16 for *BRAF* and *STK11* respectively and approximately half of the non-synonymous mutations in *TTN* are likely to be passengers. *TTN* is the largest polypeptide encoded by the human genome[18] and has been extensively studied as a component of the muscle contractile machinery. However, it is expressed in many cell types and has other functions that are compatible with a

Table 3 | Protein kinase genes ranked by probability of carrying at least one driver mutation, conditional on the gene-specific selection pressures.

Gene	Ranking (95% confidence interval)	Selection pressure	Number of non-synonymous mutations
TTN	1 (1–3)	2.036	63
BRAF	2 (1–67)	8.362	8
ATM	3 (2–150)	2.920	10
TAF1L	4 (2–145)	3.588	8
ERN1	5 (2–151)	4.538	6
MAP2K4	6 (2–156)	8.665	4
CHUK	7 (2–205)	5.392	5
FGFR2	8 (2–210)	5.096	5
NTRK3	9 (2–518)	4.808	5
MGC42105	10 (2–170)	7.097	4
TGFBR2	11 (2–187)	5.877	4
EPHA6	12 (3–518)	3.949	5
FLJ23074	13 (3–193)	5.403	4
ITK	14 (3–203)	4.887	4
DCAMKL3	15 (3–204)	4.714	4
STK11	16 (3–518)	7.160	3
PAK7	17 (3–518)	4.215	4
STK6	18 (3–518)	6.018	3
BRD2	19 (4–518)	3.773	4
RPS6KA2	20 (4–518)	3.722	4

The top 20 protein kinase genes are shown. See Supplementary Information for the ranking and selection pressures for all 518 genes.

role in oncogenesis[19–21]. The role of *TTN* as a cancer gene is currently a mathematically based prediction and will require direct biological evaluation.

Several genes that are high in the statistical ranking have previously been associated with cancer development. Some of these genes may be activated by their somatic mutations and function as dominant cancer genes, for example *NTRK3* and *ITK*, which are activated by rearrangement in secretory breast cancer and T-cell lymphoma respectively (see http://www.sanger.ac.uk/genetics/CGP/Census/). Others are more likely to be inactivated and operate as recessive cancer genes including *ATM*, in which germline mutations predispose to ataxia telangiectasia[22] and breast cancer[23], *TGFBR2,* in which frameshift somatic mutations are frequently found in mismatch repair deficient cancers[24], and *BMPR1A,* in which germline inactivating mutations cause juvenile polyposis[25]. Each of these three genes has at least one somatic nonsense mutation in the screen. However, most of the genes with probable driver mutations have not previously been associated with cancer development.

Several mutations identified in conserved, functional domains are plausible candidate driver mutations. For example, mutations were found in the glycine residues of the ATP-binding P-loop GxGxxG motif of several protein kinases (Fig. 3). Similar mutations in *BRAF* induce cellular transformation and activate downstream MEK signalling[26]. Mutations were also identified within the activation segment (Fig. 3), a domain frequently harbouring oncogenic mutations in known cancer genes such as *EGFR*, *FLT3*, *KIT* and *BRAF* (see http://www.sanger.ac.uk/genetics/CGP/cosmic/). In particular, the highly conserved DFG motif at the amino-terminal end of the activation segment was mutated in eight protein kinases including three closely related members of the SRC family, *HCK*, *LYN* and *FYN*. Similarly, a Y589H mutation was identified in the juxtamembrane domain of *PDGFRB* in a gastric cancer. *PDGFRB* is activated by translocation in leukaemias (http://www.sanger.ac.uk/genetics/CGP/Census/), and activating mutations in the juxtamembrane domain of the *PDGFRB* paralogue, *PDGFRA*, are found in gastrointestinal stromal tumours (http://www.sanger.ac.uk/genetics/CGP/cosmic/). Tyrosine 589 is highly conserved and mutation of this residue increases the baseline kinase activity of PDGFRB, conferring IL3 independence on BaF3 cells[27].

Clustering of mutations in multiple genes implicates the JNK pathway in cancer development. We and others have identified truncating and mis-sense mutations of *MAP2K4* in lung, colorectal and other cancers[6,28–30]. Downstream signalling from MAP2K4 is mediated, in part, through phosphorylation of MAP2K7 (MKK7) and subsequent activation of JNK1 (MAPK8) and JNK2 (MAPK9)[31,32]. We found two different *MAP2K7* mis-sense mutations of codon 162 (p.R162C and p.R162H) within the kinase domain in colorectal cancers. Moreover, we identified activation segment mutations in *MAPK8* (*JNK1*) and a kinase domain mutation in *MAPK9* (*JNK2*). Taken together, these data indicate that mutations in the JNK pathway are likely to be involved in cancer development.

To investigate formally the distribution of mutated genes with respect to biological pathways, we compared the set of genes with a high probability of having at least one driver mutation to a combined data set of human pathway information that is based on Reactome[33], Panther[34] and INOH[35] data sets. Five-hundred-and-thirty-seven non-redundant pathways containing different combinations of protein kinases were examined. The FGF signalling pathway (Panther Accession P00021 http://www.pantherdb.org/) showed the highest enrichment for kinases containing non-synonymous mutations (corrected *P*-value of 0.011). Among genes in this pathway, previous biological and genetic information suggest that the fibroblast growth factor receptors show several plausible driver mutations. Activating germline mutations of *FGFR3* are known to cause dwarfism[36]. Previous studies have shown that the same amino acids in *FGFR3* that are mutant in the germ line, causing thanatophoric dwarfism, are mutated somatically in bladder cancer[37]. We observed the same pattern of coincident germline mutations causing skeletal dysplasia and somatic mutations in cancer for *FGFR1* (p.P252T) and *FGFR2* (p.W290C), both in lung cancers[6]. Other mutated genes in the FGF signalling pathway included several MAP kinases such as *MAP2K4*, *MAP2K7*, *MAPK8* (*JNK1*) and *MAPK9* (*JNK2*). Interestingly, pathways involved in apoptosis and cell cycle checkpoints were not enriched in this analysis, although the relative paucity of kinase-domain-containing genes in these pathways limits the power to draw definitive conclusions. Finally, comparison of our results with previously published screens of protein kinases in colorectal cancer[9,30,38] identifies several genes mutated in both colorectal cancer series including *BRAF*, *MAP2K4*, *ERBB4*, *PRKCZ* and *RET*.

Discussion

These large-scale sequencing studies have shown that the prevalence and signature of somatic mutations in human cancers are highly variable. It is likely that the full range of somatic mutation patterns will not be apparent until thousands of cancer samples have been sequenced, each one yielding several dozen mutations each. For some cancers this may require sequencing of hundreds of megabases. This information, however, will ultimately provide major insights into the mutagenic processes underlying neoplastic change.

Our results demonstrate that most somatic mutations in cancer cells are likely to be passenger mutations; however, they have also revealed surprising insights into the number of cancer genes operative in human cancer. Approximately 120 of the 518 genes screened are estimated to carry a driver mutation and therefore function as cancer genes, a larger number than previously anticipated. Interestingly, however, similar conclusions have recently been reached by others. A recent paper reported a mutational analysis of 13,023 genes in 11 colorectal and 11 breast cancers, covering ~1.7 times as much cancer genome as this study[38]. As in this study, they interpret an excess of observed non-synonymous mutations compared with that expected by chance as evidence for the presence of driver mutations. Their design did not include the examination of synonymous changes and hence the analysis of selection pressure undertaken here. Instead, they estimated the expected number of non-synonymous passenger mutations on the basis of prior published data and identified 189 genes that were mutated at significantly higher frequency. Their conclusion was broadly similar, that a large number of cancer-causing mutations and cancer genes are operative in human cancers.

By studying a gene family with a strong track record of involvement in oncogenesis, it is conceivable that we have improved our chances of detecting new cancer genes and that other gene sets may yield a more meagre harvest. Nevertheless, given that we have studied only 518 genes and limited numbers of each cancer type, it seems likely that the repertoire of mutated human cancer genes is larger than previously envisaged. The work presented here suggests that systematic sequencing studies of larger numbers of tumours from a wide variety of cancer types will yield further insights into the development of human cancer, providing new opportunities for molecular diagnosis and therapeutics.

METHODS

DNA was extracted from primary tumours, cancer cell lines and normal tissue samples. Collection and use of tissue samples were approved by the IRB of each institution. Samples estimated to contain more than 80% tumour cells were used. All samples were analysed using Affymetrix 10K SNP arrays to demonstrate that they were from the same individual and to confirm the presence of copy number changes. Microsatellite instability was assessed using the NCI consensus marker panel[39]. PCR primers were designed to amplify all coding exons of the 518 protein kinases[4] annotated in the human genome (available at http://www.sanger.ac.uk/genetics/CGP/). Approximately 10,000 fragments of 500 base pairs were amplified and directly sequenced in both directions from each cancer. Sequence traces were initially evaluated computationally and subsequently manually reviewed. The existence of the variant was then assessed in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) and, if not present, was directly evaluated in normal DNA from the same individual by PCR sequencing using the appropriate

amplimer. Cancer samples showing putative somatic sequence alterations were then re-amplified and re-sequenced along with the appropriate, matched, non-cancer DNA to confirm the somatic nature of the mutation and to eliminate sequencing artefacts. Statistical analyses are outlined in more detail in Supplementary Methods. Deviation of the ratio of non-synonymous:synonymous mutations from that expected by chance was used to indicate the presence of selection on non-synonymous mutations. To assess the significance of this ratio, an exact Monte Carlo test was developed which was applied to the entire set and to subsets of mutations. Additional methods were developed to determine the number of driver mutations, analyse differences in selection between mismatch-repair-deficient and -proficient cancers and to assess the likelihood of a gene being a cancer gene. A combined pathway database was generated by merging Reactome, Panther and INOH to test for the presence of mutated pathways.

Received 7 September 2006; accepted 18 January 2007.

1. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
2. Futreal, P. A. *et al.* Cancer and genomics. *Nature* **409,** 850–852 (2001).
3. Sawyers, C. Targeted cancer therapy. *Nature* **432,** 294–297 (2004).
4. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934 (2002).
5. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature Genet.* **37,** 590–592 (2005).
6. Davies, H. *et al.* Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res.* **65,** 7591–7595 (2005).
7. Hunter, C. *et al.* A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66,** 3987–3991 (2006).
8. Bignell, G. *et al.* Sequence analysis of the protein kinase gene family in human testicular germ-cell tumours of adolescents and adults. *Genes Chromosom. Cancer* **45,** 42–46 (2006).
9. Bardelli, A. *et al.* Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300,** 949 (2003).
10. Wang, T.-L. *et al.* Prevalence of somatic alterations in the colorectal cancer cell genome. *Proc. Natl Acad. Sci. USA* **99,** 3076–3080 (2002).
11. Lonardi, S., Tosoni, A. & Brandes, A. A. Adjuvant chemotherapy in the treatment of high grade gliomas. *Cancer Treat. Rev.* **31,** 79–89 (2005).
12. Karran, P., Offman, J. & Bignami, M. Human mismatch repair, drug-induced DNA damage, and secondary cancer. *Biochimie* **85,** 1149–1160 (2003).
13. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173,** 2187–2198 (2006).
14. Yang, Z., Ro, S. & Rannala, B. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics* **165,** 695–705 (2003).
15. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–736 (1994).
16. Sanchez-Cespedes, M. *et al.* Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. *Cancer Res.* **62,** 3659–3662 (2002).
17. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417,** 949–954 (2002).
18. Granzier, H. L. & Labeit, S. Titin and its associated proteins: the third myofilament system of the sarcomere. *Adv. Protein Chem.* **71,** 89–119 (2005).
19. Machado, C. & Andrew, D. J. D-Titin: a giant protein with dual roles in chromosomes and muscles. *J. Cell Biol.* **151,** 639–652 (2000).
20. Machado, C., Sunkel, C. E. & Andrew, D. J. Human autoantibodies reveal Titin as a chromosomal protein. *J. Cell Biol.* **141,** 321–333 (1998).
21. Zastrow, M. S., Flaherty, D. B., Benian, G. M. & Wilson, K. L. Nuclear Titin interacts with A- and B-type lamins *in vitro* and *in vivo*. *J. Cell Sci.* **119,** 239–249 (2006).
22. Shiloh, Y. ATM and related protein kinases: safeguarding genome integrity. *Nature Rev. Cancer* **3,** 155–168 (2003).
23. Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature Genet.* **38,** 873–875 (2006).
24. Markowitz, S. *et al.* Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268,** 1336–1338 (1995).
25. Howe, J. R. *et al.* Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nature Genet.* **28,** 184–187 (2001).
26. Wan, P. T. C. *et al.* Mechanism of activation of the RAF–ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116,** 855–867 (2004).
27. Irusta, P. M. *et al.* Definition of an inhibitory juxtamembrane WW-like domain in the platelet-derived growth factor beta receptor. *J. Biol. Chem.* **277,** 38627–38634 (2002).
28. Teng, D. *et al.* Human mitogen-activated protein kinase kinase 4 as a candidate tumor suppressor. *Cancer Res.* **57,** 4177–4182 (1997).
29. Su, G. *et al.* Alterations in pancreatic, biliary, and breast carcinomas support MKK4 as a genetically targeted tumor suppressor gene. *Cancer Res.* **58,** 2339–2342 (1998).
30. Parsons, D. W. *et al.* Colorectal cancer Mutations in a signalling pathway. *Nature* **436,** 792 (2005).
31. Bogoyevitch, M. A., Boehm, I., Oakley, A., Ketterman, A. J. & Barr, R. K. Targeting the JNK MAPK cascade for inhibition: basic science and therapeutic potential. *Biochim. Biophys. Acta* **1697,** 89–101 (2004).
32. Kyriakis, J. M. & Avruch, J. Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation. *Physiol. Rev.* **81,** 807–869 (2001).
33. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33,** D428–D432 (2005).
34. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33,** D284–D288 (2005).
35. Kushida, T., Takagi, T. & Fukuda, K. Event ontology: a pathway-centric ontology for biological processes. *Pac. Symp. Biocomput.* **11,** 152–163 (2006).
36. Wilkie, A., Patey, S., Kan, S., van den Ouweland, A. & Hamel, B. FGFs, their receptors, and human limb malformations: clinical and molecular correlations. *Am. J. Med. Genet.* **112,** 266–278 (2002).
37. Cappellen, D. *et al.* Frequent activating mutations of FGFR3 in human bladder and cervix carcinomas. *Nature Genet.* **23,** 18–20 (1999).
38. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006).
39. Brose, M. S. *et al. BRAF* and *RAS* mutations in human lung cancer and melanoma. *Cancer Res.* **62,** 6997–7000 (2002).

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

Acknowledgements We would like to thank J. Leary and the ABN-Oncology group (funded by the National Health and Medical Research Council of Australia), the Hauenstein Foundation and the Cooperative Human Tissue Network for providing samples for analysis, G. Wu and L. Stein for the development of the joint Reactome, Panther, INOH database, and C. Marshall and N. Rahman for comments. The studies were funded by the NIH and the Wellcome Trust.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.A.F. (paf@sanger.ac.uk) or M.R.S. (mrs@sanger.ac.uk).