ARTICLE

doi:10.1038/nature12912

Discovery and saturation analysis of cancer genes across 21 tumour types

Michael S. Lawrence[1], Petar Stojanov[1,2], Craig H. Mermel[1,3], James T. Robinson[1], Levi A. Garraway[1,2,4], Todd R. Golub[1,2,4,5], Matthew Meyerson[1,2,4], Stacey B. Gabriel[1], Eric S. Lander[1,4,6]* & Gad Getz[1,3,4]*

Although a few cancer genes are mutated in a high proportion of tumours of a given type (>20%), most are mutated at intermediate frequencies (2–20%). To explore the feasibility of creating a comprehensive catalogue of cancer genes, we analysed somatic point mutations in exome sequences from 4,742 human cancers and their matched normal-tissue samples across 21 cancer types. We found that large-scale genomic analysis can identify nearly all known cancer genes in these tumour types. Our analysis also identified 33 genes that were not previously known to be significantly mutated in cancer, including genes related to proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis. Down-sampling analysis indicates that larger sample sizes will reveal many more genes mutated at clinically important frequencies. We estimate that near-saturation may be achieved with 600–5,000 samples per tumour type, depending on background mutation frequency. The results may help to guide the next stage of cancer genomics.

Comprehensive knowledge of the genes underlying human cancers is a critical foundation for cancer diagnostics, therapeutics, clinical-trial design and selection of rational combination therapies. It is now possible to use genomic analysis to identify cancer genes in an unbiased fashion, based on the presence of somatic mutations at a rate significantly higher than the expected background level.

Systematic studies have revealed many new cancer genes, as well as new classes of cancer genes[1,2]. They have also made clear that, although some cancer genes are mutated at high frequencies, most cancer genes in most patients occur at intermediate frequencies (2–20%) or lower. Accordingly, a complete catalogue of mutations in this frequency class will be essential for recognizing dysregulated pathways and optimal targets for therapeutic intervention. However, recent work suggests major gaps in our knowledge of cancer genes of intermediate frequency. For example, a study of 183 lung adenocarcinomas[3] found that 15% of patients lacked even a single mutation affecting any of the 10 known hallmarks of cancer, and 38% had 3 or fewer such mutations.

In this paper, we analysed somatic point mutations (substitutions and small insertion and deletions) in nearly 5,000 human cancers and their matched normal-tissue samples ('tumour–normal pairs') across 21 tumour types. The questions that we examine here are: first, whether large-scale genomic analysis across tumour types can reliably identify all known cancer genes; second, whether it will reveal many new candidate cancer genes; and third, how far we are from having a complete catalogue of cancer genes (at least those of intermediate frequency). We used rigorous statistical methods to enumerate candidate cancer genes and then carefully inspected each gene to identify those with strong biological connections to cancer and mutational patterns consistent with the expected function.

The analysis reveals nearly all known cancer genes and revealed 33 novel candidates, including genes related to proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis. Importantly, the data show that the catalogue of cancer genes is still far from complete, with the number of candidate cancer genes still increasing sharply with sample size. These analyses enable us to estimate the sample sizes that will be needed to approach saturation.

Cancer-genome data

We collected and analysed data from 4,742 samples, consisting primarily of whole-exome sequence from tumour–normal pairs. The samples span 21 tumour types, which include 12 from The Cancer Genome Atlas (TCGA) and 14 from non-TCGA projects at the Broad Institute, with some overlapping tumour types (Table 1 and Supplementary Table 1). The number of samples per tumour type varied between 35 and 892.

Data were all analysed through the Broad's stringent filtering and annotation pipeline to obtain a uniform set of mutation calls (Methods). The data set consists of 3,078,483 somatic single nucleotide variations (SSNVs), 77,270 small insertions and deletions (SINDELs) and 29,837 somatic di-, tri- or oligonucleotide variations (DNVs, TNVs and ONVs, respectively), with an average of 672 per tumour–normal pair. The mutations include 540,831 missense, 207,144 synonymous, 46,264 nonsense, 33,637 splice-site, and 2,294,935 non-coding mutations (used to improve our background model). The analysis has sensitivity of >90% based on the sequencing depth and tumour purity and ploidy[4,5].

Mutation frequencies vary over more than five orders of magnitude (from 0.03 per megabase (Mb) to 7,000 per Mb) within and across tumour types, consistent with our recent study of mutational heterogeneity[6] of approximately 3,000 samples (of which 2,502 are included in this data set) (Supplementary Fig. 1). Mutation spectra also vary sharply within and across tumour types[6] (Supplementary Fig. 2).

Cancer-genome analysis

We analysed these data to identify candidate cancer genes, by which we mean genes harbouring somatic point mutations (that is, substitutions

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. [3]Massachusetts General Hospital, Cancer Center and Department of Pathology, 55 Fruit Street, Boston, Massachusetts 02114, USA. [4]Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA. [5]Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815, USA. [6]Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.
*These authors contributed equally to this work.

Table 1 | List of the 21 tumour types analysed

Tumour type	No. of tumour–normal pairs	Median somatic mutation frequency (per Mb)	No. of significantly mutated genes	No. of additional significant genes found under RHT
Acute myeloid leukaemia	196	0.4	26	1
Bladder	99	7.1	24	10
Breast	892	1.2	32	5
Carcinoid	54	0.5	1	0
Chronic lymphocytic leukaemia	159	0.6	7	8
Colorectal	233	3.1	23	12
Diffuse large B-cell lymphoma	58	3.3	16	7
Endometrial	248	2.5	58	15
Oesophageal adenocarcinoma	141	4.0	8	7
Glioblastoma multiforme	291	2.2	22	4
Head and neck	384	3.9	25	9
Kidney clear cell	417	1.9	15	6
Lung adenocarcinoma	405	8.1	22	10
Lung squamous cell carcinoma	178	9.9	11	13
Medulloblastoma	92	0.3	2	1
Melanoma	118	12.9	19	9
Multiple myeloma	207	1.6	11	3
Neuroblastoma	81	0.5	1	0
Ovarian	316	1.7	5	5
Prostate	138	0.7	4	2
Rhabdoid tumour	35	0.1	1	0

The number of significantly mutated genes detected using the MutSig suite when analysing the full set of genes. RHT, restricted hypothesis testing on the set of cancer genes found in all the other tumour types. Supplementary Table 3 lists the cancer genes found in each tumour type and their frequencies (per cent of patients with mutations).

and small insertion or deletions) at a statistically significant rate or pattern in cancer. (Such genes will ultimately need to be verified by biological experiments to be considered validated cancer genes.) In this paper, we do not seek to implicate genes based on other criteria (such as amplification or deletion, translocations or epigenomic modification; however, see ref. 7 for an analysis of copy-number alterations across many tumour types).

In principle, candidate cancer genes can be discovered by sequencing enough tumour–normal pairs; based on the presence of an excess of somatic mutations compared to expectation. However, careful analysis is required to assess statistical significance. The mere presence of somatic mutations is insufficient to implicate a gene in cancer, inasmuch as 93% of genes carried mutations in at least five samples.

We showed recently[6] that heterogeneity of mutation rates and patterns in cancer can give rise to false positives and described methods to overcome this problem. We applied these methods to identify candidate cancer genes. We used the most recent version of the MutSig suite of tools (Supplementary Fig. 3a and Methods), which looks for three independent signals: high mutational burden relative to background expectation, accounting for heterogeneity; clustering of mutations within the gene[8]; and enrichment of mutations in evolutionarily conserved sites[8]. We combined the significance levels (*P* values) from each test to obtain a single significance level per gene (Methods).

We analysed each tumour type separately, as well as the entire cohort ('combined' set), using the same methodology to ensure that the results can be compared across types. We verified that each analysis accurately calculates the significance level of genes, based on the fact that the vast majority of genes fit the null hypothesis and lie on the diagonal in a Q–Q plot (Supplementary Fig. 3b). For each analysis, genes with false discovery rate (FDR) $q \leq 0.1$ were declared to be candidate cancer genes (Methods). Using an FDR of $q \leq 0.1$ ensures that the expected fraction of false positives in each analysis does not exceed 10%. This well-established statistical procedure results in an increase in statistical power to detect true positives, while controlling the proportion of false positives. We also analysed the merged set of gene × tumour-type pairs identified from the 22 individual analyses (here we include the combined set as one of the 'tumour types'), using methods discussed below.

Data and results are posted at http://www.tumorportal.org/. The site includes graphical displays of the mutations in each of the 18,388 genes studied; see examples in Fig. 1 and Supplementary Fig. 4. The site also includes tables of mutational data for each significant gene and Q–Q plots for each statistical test.

Candidate cancer genes across 21 tumour types

A total of 334 gene × tumour-type pairs were found by our analysis to be significantly mutated. These 334 pairs involve 224 distinct genes. The number of genes detected per tumour type varied considerably (range of 1–58), with 7 types having fewer than 10 genes and 2 (breast and endometrial) having more than 30 (Fig. 2, Supplementary Fig. 5 and Table 1). The specific genes differed substantially across tumour types, although some pairs of tumour types showed clear similarity, such as lung squamous cancer and head and neck squamous cancer (Methods and Supplementary Fig. 6).

Notably, only 22 genes were found to be significant in three or more tumour types. The well-established cancer genes *TP53, PIK3CA, PTEN, RB1, KRAS, NRAS, BRAF, CDKN2A, FBXW7, ARID1A* and *MLL2*, as well as *STAG2*, were significant in four or more tumour types. An additional 10 genes (*ATM, CASP8, CTCF, ERBB3, HLA-A, HRAS, IDH1, NF1, NFE2L2* and *PIK3R1*) were significant in three tumour types.

Although the power to detect cancer genes varied across tumour types (based on sample size and background mutation frequency), the marked differences across tumour types do not simply reflect differences in detection power. For example, tumour types with low mutation frequency or many samples often show fewer cancer genes despite having greater statistical power to detect them (Table 1). Moreover, many genes that are highly enriched in one (for example, *VHL, WT1*) or a few (for example, *HRAS, FBXW7*) tumour types fail to show detectable enrichment across the entire data set (Supplementary Table 2). Notably, most of the significant gene × tumour-type pairs involve only a small fraction of patients (with one half of the significant pairs involving ≤6.1% of patients, and one quarter involving ≤3.1%).

We then analysed the combined set, which produced 114 genes (Fig. 3 and Supplementary Table 2). Although 84 of these genes were already identified from analysis of individual tumour types, the remaining 30 achieved significance based only on the frequency of mutations across tumour types, underscoring the value of cross-tumour-type analysis. Conversely, 140 of the 224 genes found in analysis of individual tumour types did not reach significance when analysing the combined set (Fig. 3, bottom-right quadrant), consistent with the observation that many genes show strong enrichment in only one or a few tumour types.

By merging the 22 lists above, we obtained a Cancer5000 set containing 254 genes. Although the expected proportion of false positive genes in each list does not exceed 10%, the expected proportion in the merged list is actually higher (because true positives will tend to occur across several tumour types, whereas false positives will tend to be random

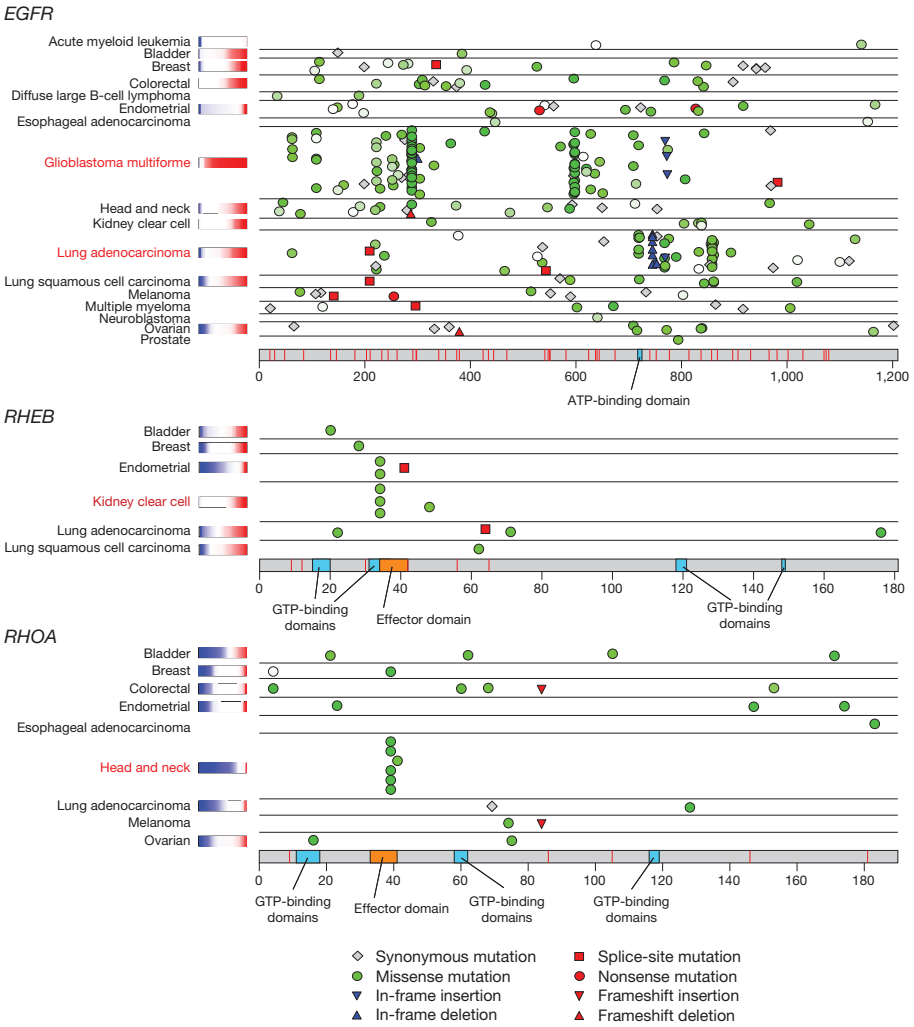

Figure 1 | Mutation patterns for one known and two novel cancer genes. *EGFR* shows distinctive tumour-type-specific concentrations of mutations in different regions of the gene. *RHEB*, which encodes a small GTPase in the Ras superfamily, shows a mutational hotspot in the effector domain. *RHOA*, another a member of the Ras superfamily, also shows a mutational hotspot in the effector domain. Coloured bars after tumour-type names are copy-ratio distributions for the gene, when available (red, amplified; blue, deleted). Missense mutations are represented by green circles of varying saturation indicating degree of evolutionary conservation of the mutated base pair, from highly conserved (dark green), to medium conservation (light green), to totally unconserved (white). Tumour types with names shown in red were significantly mutated in this gene, in dark red were nearly significantly mutated, or in black were not significantly mutated. Thin red strokes in the protein ideogram represent splice sites (see also Supplementary Fig. 4; similar diagrams for all genes are available at http://www.tumorportal.org).

singletons). A rigorous solution is to analyse the gene × tumour-type pairs as approximately 400,000 distinct hypotheses (approximately 18,400 genes × 22 types) and apply an FDR of $q \leq 0.1$. This analysis produces 403 significant pairs, which involve 219 distinct genes. We refer to this set as the Cancer5000-S (for 'stringent') genes. (All but six of the genes are contained in the Cancer5000 set.) Of the 403 significant pairs, 10% (approximately 40) at most are expected to be false positives. Assuming conservatively that the 40 pairs affect 40 distinct genes, we expect 179 of the 219 genes to be true cancer genes. Below, we discuss genes from both the Cancer5000 and Cancer5000-S sets.

Coverage of known cancer genes

We first asked whether all cancer genes that have been discovered and validated to date can be identified by hypothesis-free genomic analysis. As a reference set, we used the Cancer Gene Census (CGC), which is a manually curated catalogue of cancer genes. The current version (v65) contains 130 cancer genes driven by somatic point mutations (as well as additional genes mutated by other mechanisms), of which 82 are associated with 1 or more of the 21 tumour types studied here.

Of these 82 genes, 60 were identified in our Cancer5000 set. Of the remaining 22 genes, 8 fell just below significance in our data set, 6 appear in the CGC based on focused studies of the gene in very large samples (typically >1,000), and 8 genes harboured few mutations and seem to lack adequate evidence to justify association with any of the tumour types we studied. The first two categories would clearly be captured with larger sample sizes.

Analysis of novel candidate cancer genes

Of the 219 genes in the Cancer5000-S set, 81 are neither listed in the CGC as affected by point mutations in these tumour types (v65) nor discussed in papers published so far (Supplementary Table 4). (The list includes three genes that appear in tables in published papers based on mutations in a handful of samples, but were not noted or interpreted in the text.) Of the 41 additional genes in the Cancer5000 (but not Cancer5000-S) set, none are in the CGC but 3 are reported in recent publications (Supplementary Table 4).

We closely analysed these 81 'novel' genes to look for connections with cancer biology, together with a mutational pattern consistent with the biology. Where loss-of-function would be expected, we looked for an excess of disruptive changes, such as nonsense and frameshift mutations. In cases in which gain-of-function would be expected, we examined whether the overall collection of mutations included hotspots that resulted in recurrent changes at identical or nearby amino acids (often causing precisely the same change). Conversely, where we observed distinctive mutation patterns, we examined whether they were consistent with known biology.

As discussed above, the Cancer5000-S set is expected by design to contain approximately 40 false positives. Assuming conservatively that these false positives fall exclusively in the novel set, we expect approximately 41 of the 81 novel genes to be true positives.

In fact, we identified strong and consistent connections to cancer for at least 21 of the novel genes in the Cancer5000-S set. Among the 38 additional novel genes in the larger Cancer5000 set, we found 12 additional strong candidates. (References supporting the biological

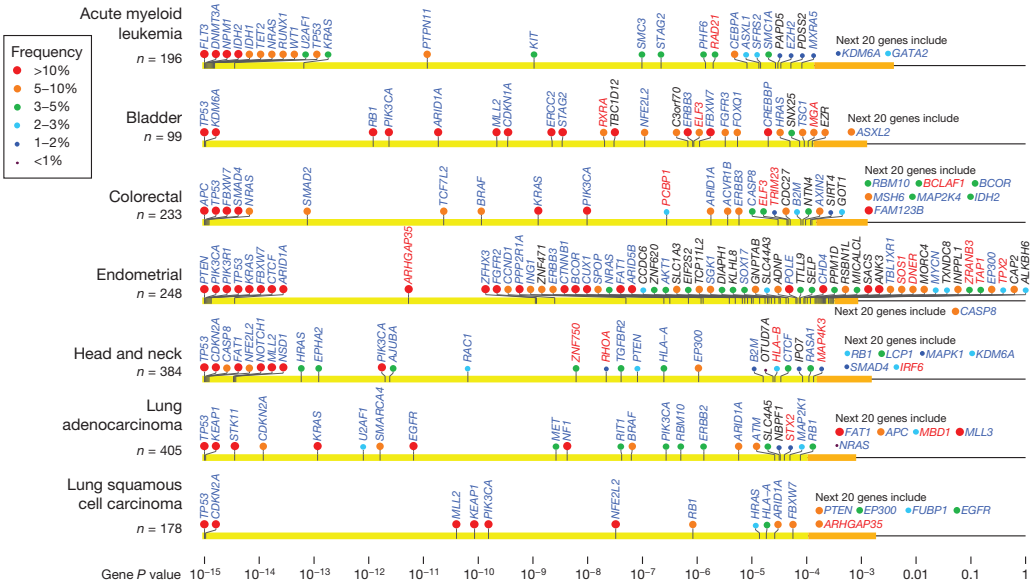

Figure 2 | Cancer genes in selected tumour types. Genes are arranged on the horizontal line according to *P* value (combined value for the three tests in MutSig). Yellow region contains genes that achieve FDR $q \leq 0.1$. Orange interval contains *P* values for the next 20 genes. Gene-name colour indicates whether the gene is a known cancer gene (blue), a novel gene with clear connection to cancer (red, discussed in text), or an additional novel gene (black). Circle colour indicates the frequency (percentage of patients carrying non-silent somatic mutations) in that tumour type; see also Supplementary Fig. 5.

roles of the genes are provided in Supplementary Table 5.) We briefly describe below these 33 genes not previously reported as significantly mutated in cancer.

Four genes encode anti-proliferative proteins, in which loss-of-function mutations would be expected to contribute to oncogenesis. A notable example is *ARHGAP35* (previously called *GRLF1*), which encodes a Rho-GTPase-activating protein, for which only a single tumour type reaches statistical significance on its own, but which gives a strong signal ($q = 2 \times 10^{-12}$) in the combined set of 4,742 tumours (83 missense, 38 nonsense, 16 frameshift and 2 splice site). Notably, the gene resides in a small genomic region that is focally deleted in many tumours. Other examples are *MGA*, whose product competes with Myc for binding to Max and which resides in small focal deletions (containing ≤4 genes) in ovarian and various epithelial cancers; the interferon regulatory factor *IRF6*, which is known to have tumour suppressive roles in keratinocytes and is mutated in head and neck squamous cancer; and the delta/notch-like EGF-repeat gene *DNER*.

Six additional genes encode proteins that are clearly involved in cell proliferation: *RHEB*, *RHOA*, *SOS1*, *ELF3*, *SGK1* and *MYOCD*. Notably, *RHEB* and *RHOA* encode small GTPases, in which recurrent mutations affect the 9-amino-acid effector domain. For *RHEB*, five tumours (two endometrial and three kidney clear cell cancer) carry Tyr35Asn mutations, which alter the first amino acid of the effector domain. For *RHOA*, seven tumours (six head and neck, one breast) carry mutations affecting the effector domain: these include six Glu40Gln mutations and a single Tyr42Ile mutation, which alter the seventh and ninth amino acids of the effector domain, respectively. *SOS1* encodes a guanine nucleotide exchange factor that promotes activation of Ras proteins, in which gain-of-function mutations might contribute to oncogenesis. Consistent with this notion, *SOS1* carries Asn233Tyr mutations in six tumours (four endometrial and two lung adenocarcinoma) and Arg 552 alterations in three tumours (two endometrial and one AML). Notably, the same Arg 522 alterations in *SOS1* have been found to be germline mutations causing Noonan syndrome and to cause gain of function, resulting in Ras activation. *ELF3* encodes an ETS-domain transcription factor that functions in cell differentiation; it carries many truncating mutations in bladder and colon cancer. Myocardin (*MYOCD*), which encodes a transcriptional regulator involved in differentiation and cell migration, has a cluster of 9 mutations at amino acids 750–770 (7 in melanoma, 1 head and neck, 1 lung adenocarcinoma) with a hotspot of four at Ser 763. The retinoid X receptor alpha *RXRA*, which forms a heterodimer with retinoic acid receptors to regulate cell growth and survival, shows a clear hotspot of recurrent mutations at Ser 427 in bladder cancer and nearby mutations in lung, head and neck, and oesophageal cancers.

Five genes encode pro-apoptotic factors, in which loss-of-function mutations would be expected to promote oncogenesis. These genes encode alpha-kinase 2 (*ALPK2*); Bcl2-associated factor 1 (*BCLAF1*); a MAP kinase (*MAP4K3*) reported to post-transcriptionally regulate the apoptotic proteins PUMA (also known as BBC3), BAD and BIM (also known as BCL2L11); a zinc-finger protein (*ZNF750*, which harbours many early frameshift and nonsense mutations in head and neck cancer and is the only known gene residing in a small current focal deletion in head and neck and lung squamous cancers); and tumour necrosis factor (*TNF*, which harbours mutations in five diffuse large B-cell lymphomas that are tightly clustered in the region encoding the membrane and cytoplasmic domain, rather than the soluble TNF protein).

Six genes encode proteins related to genome stability. These include *CEP76* (encoding a centrosomal protein, whose depletion drives aberrant amplification of centrioles), which harbours early nonsense mutations in many tumour types and resides in a focal deletion peak in acute myeloid leukaemia; *RAD21* (encoding a protein crucial for chromosome segregation and double-strand break repair), which is mutated at significant rates in acute myeloid leukaemia and also harbours mutations in other tumour types; the p53-binding protein *TP53BP1* (encoding a check-point protein that binds to double-strand breaks), which does not reach significance in any single tumour type, but is significant in the combined data set owing to truncating mutations in many tumour types; *TPX2* (encoding a protein involved in mitotic spindle formation, whose depletion leads to aneuploidy); and *ZRANB3* (encoding a translocase that helps to rescue stalled replication forks). In addition, *STX2* encodes a protein required for cytokinesis, whose disruption may promote aneuploidy; *STX2* harbours recurrent mutations at Arg 107 in lung and endometrial tumours.

Five genes are associated with chromatin regulation. *SETDB1* encodes a H3K9 histone methyltransferase (*SETD2*, which encodes a H3K36 histone methyltransferase, has been shown previously to be mutated in cancer). *MBD1* encodes a protein that binds methylated-CpG and is required for SETDB1 activity; it contains five mutations in endometrial cancer in the amino-terminal methyl-binding domain. *EZH1* encodes a H3K27 histone methyltransferase; it does not reach significance in any individual tumour type, but is significant in the combined set owing to truncating mutations in multiple tumour types. *EZH1* shows a similar pattern of mutations as seen in the well-established cancer gene *EZH2*, with truncating mutations along the gene and a hotspot of

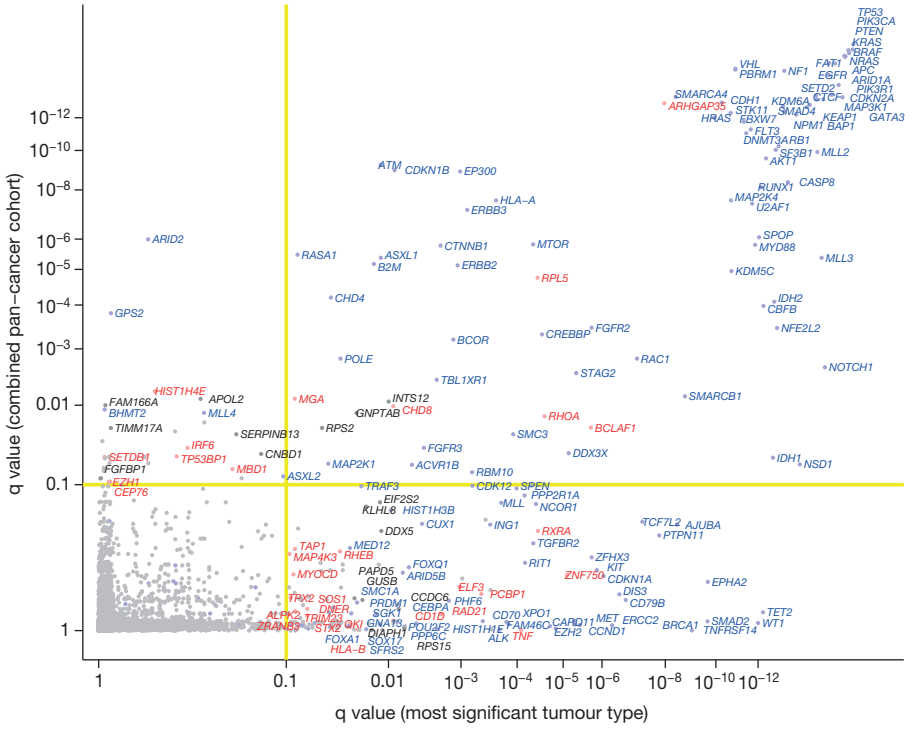

Figure 3 | Cancer genes identified from a data set of 4,742 tumours. Genes are plotted by the q value (FDR) in the most significant of the 21 tumour types (*x* axis) and the q value when the 4,742 tumours are analysed as a combined ('pan-cancer') cohort (*y* axis). Genes in the top-left quadrant reached significance only in the combined analysis. Genes in the bottom-right quadrant reached significance only in one or more single-type analyses. Genes in the top-right quadrant were significant in both the combined set and in individual tumour types. Colour of gene names is as in Fig. 2.

mutations within the SET domain. *CHD8* encodes a chromatin helicase DNA binding protein (like the known cancer gene *CHD4*) that suppresses the β-catenin–Wnt signalling pathway. The histone protein *HIST1H4E* is mutated in multiple tumour types; two other histone genes, *HIST1H1E* and *HIST1H3B*, have previously been reported as significantly mutated in CLL and DLBCL, respectively.

Three genes encode proteins whose loss is expected to help tumours evade immune attack; they all recurrently subject to truncating mutations across several tumour types. These include the major histocompatibility protein *HLA-B* (loss of the *HLA-A* gene has been implicated in lung cancer), *TAP1* (which processes intracellular peptides for presentation to the immune system) and *CD1D* (which presents lipid antigens to natural killer cells), the last of which shows a cluster of truncating mutations at the internalization domain that are likely to abolish antigen-presentation function.

Three genes are associated with RNA processing and metabolism. *PCBP1*, whose protein product blocks translation of certain mRNAs by binding to poly(C) regions of messenger RNAs, carries two mutations in each of two nearby leucines (Leu 100 and Leu 102) that mediate dimerization of the protein's K-homology domains. We speculate that disruption of *PCBP1* leads to increased translation of one or more pro-oncogenic mRNAs. *QKI* encodes an RNA-binding protein that regulates pre-mRNA splicing, including the known cancer gene *CDKN1B*; the gene harbours caboxy-terminal truncating mutations in several tumour types that are likely to remove the nuclear localization signal; and the gene resides in a recurrent deletion peak in glioblastoma and ovarian cancer. Finally, the ribosomal protein gene *RPL5* contains early truncating mutations in glioblastoma and other tumour types and resides in a focally deleted region in glioblastoma; heterozygous loss of certain ribosomal proteins has been reported to contribute to cancer.

One gene, *TRIM23*, is involved in protein homeostasis. It encodes an ubiquitin E3 ligase and harbours recurrent mutations at Asn 93 (four tumours) and Asp 289 (three tumours). Mutations in this gene may promote cancer by altering the substrate specificity of the E3 ligase in a manner that leads to accumulation of an oncogenic protein.

Beyond these 33 genes, the set of 81 novel genes is likely to contain additional true cancer genes. For example, we omitted genes with connections to cancer (such as *HSP90AB1*, *PPM1D* and *ITGB7*) in situations in which we could not easily reconcile the function in cancer with the observed pattern of mutations. In addition, we may have overlooked additional candidate cancer genes because we did not identify clear connections with cancer, owing to gaps in the literature or in our knowledge.

Saturation analysis

We next explored whether the discovery of candidate cancer genes is approaching saturation or whether many more genes are likely to be found. An effective test is to perform 'down-sampling'; that is, to study how the number of discoveries increases with sample size, by repeating the analysis on random subsets of samples of various smaller sizes.

For each tumour type (omitting those with five or fewer candidate cancer genes), the number of genes increases roughly linearly with sample size (examples in Fig. 4a; see also Supplementary Fig. 7), indicating that the inventory for each of the tumour types is far from complete. The number of genes also increases linearly with the number of tumour types studied (Fig. 4b), suggesting that it is valuable to increase both the sample size per tumour type and the number of tumour types.

We also studied how the total number of candidate cancer genes varies with sample size when applying the 'stringent' methodology used to create the Cancer5000-S set. Here too, the total number of genes increases steadily with sample size (Fig. 4c). Notably, the saturation analysis varies considerably with the mutation frequency (Fig. 4d). Genes mutated in >20% of tumours are approaching saturation; those mutated at frequencies of 10–20% are still rising rapidly, but at a decreasing rate; those at 5–10% are increasing linearly; and those at <5% are increasingly at an accelerating rate.

We next sought to infer the nature of the genes awaiting discovery in each tumour type. One possibility is that some of these genes are already contained in the Cancer5000 set (by virtue of their contribution to other tumour types) but have not yet reached statistical significance in the given tumour type due to insufficient sample size. To test this idea, we performed restricted hypothesis testing (RHT): for each tumour type *T*, we omitted that tumour type, determined the set of genes (G_T) that are significant based on the remaining tumour

types, and determined which genes in G_T reached significance in the omitted tumour type when correcting for multiple-hypothesis testing based on only the number of genes in G_T rather than all (approximately 18,400) genes in the genome.

The RHT analysis implicated many additional Cancer5000 genes in the individual tumour types (median 6 per tumour type, range 0–15). The number of significant gene × tumour-type pairs increased from 334 to 461 across the 21 tumour types. The RHT analysis indicates that, with somewhat larger sample size, these genes are likely to reach significance in an unrestricted test (Table 1 and Supplementary Table 3). For some tumour types, the number of implicated genes more than doubled: lung squamous cell carcinoma increased from 11 to 24; CLL from 7 to 15; and ovarian from 5 to 10. Notably, three genes now became significant in four tumour types each (*ARID2*, *ERBB2*, *ARHGAP35*) and seven genes in three types each (*CTNNB1*, *FGFR3*, *KRAS*, *PTEN*, *SMAD4*, *MLL3*). Although nine of these genes are well known cancer genes, one (*ARHGAP35*) is absent from the current CGC list. Notably, *ARHGAP35* appears in the Cancer5000 set because it is significantly mutated in endometrial cancer (although not discussed in the recent TCGA publication[9]), but our RHT analysis also finds it to be significant in lung adenocarcinoma, lung squamous cell carcinoma, kidney clear cell, and head and neck cancer. The genes found to be significant in additional tumour types in the RHT analysis are mutated at a median frequency of 3.4%.

However, the data also clearly show that many new candidate cancer genes remain to be discovered beyond those in the current Cancer5000 set. First, in addition to the Cancer5000 genes being shown by RHT to be significant in additional tumour types, the down-sampling analysis shows that the number of novel genes being identified is increasing sharply (using the stringent analysis used to create the Cancer5000-S set). Second, adding additional tumour types typically adds novel 'tumour-type-specific' genes, which are unique to (or at vastly higher frequency in) the tumour type.

Power analysis

As the cancer-gene catalogue remains far from complete, we explored what sample sizes are needed to approach saturation. The power to detect a gene as significantly mutated depends on the properties of the tumour type, namely the average background somatic mutation frequency along the genome for the tumour type ('noise'), and the target frequency (across patients)—above the background rate—that one wishes to detect ('signal'). It also depends on the properties of the gene, namely its background mutation frequency relative to other genes (which depends on length and local mutation rate). We set a target of having 90% power to detect 90% of all genes. In addition, we allow for a false negative rate of 10% in detecting mutations, which increases the sample size by slightly more than 10%.

Figure 5 shows that the current collection lacks the desired power to detect genes mutated at 5% above the background rate for 17 of the 21 tumour types and even at 10% for 7 of the tumour types. These results are consistent with the down-sampling analysis showing that candidate cancer genes with frequency ≥20% are approaching saturation, whereas the number of candidate cancer genes at lower frequencies is continuing to grow rapidly with sample size.

Creating a reasonably comprehensive catalogue of candidate cancer genes mutated in ≥2% of patients will require between approximately 650 samples (for tumours with ~0.5 mutations per Mb, such as neuroblastoma) to approximately 5,300 samples (for melanoma, with 12.9 mutations per Mb).

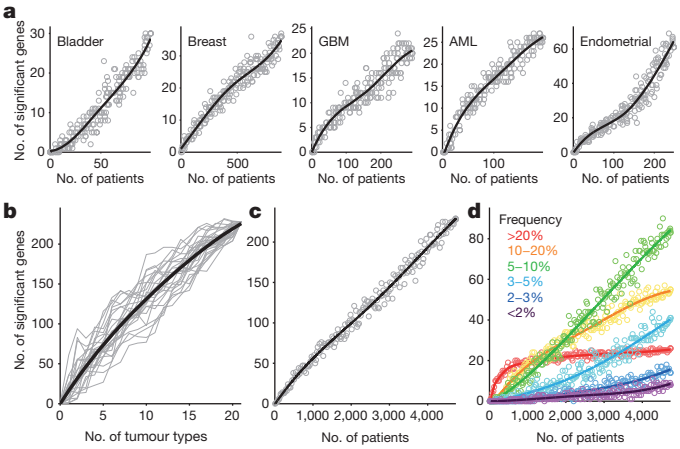

Figure 4 | Down-sampling analysis shows that gene discovery is continuing as samples and tumour types are added. **a**, Analysis within tumour types. Each point represents a random subset of patients. Line is a smoothed fit. **b**, Analysis by adding tumour types. Each grey line represents a random ordering of the 21 tumour types. **c**, Analysis by adding samples. Each point is a random subset of the 4,742 patients. **d**, Analysis in **c** broken down by mutation frequency. Genes mutated at frequencies ≥20% are nearing saturation, and intermediate frequencies show steep growth; see also Supplementary Figs 7 and 8.

Discussion

Precision medicine for cancer will ultimately require a comprehensive catalogue of cancer genes to enable physicians to select the best combination therapy for each patient based on the cellular pathways disrupted in their tumour and the specific nature of the disruptions. Such a catalogue will also guide therapeutic development by identifying druggable targets. In addition, the catalogue and its underlying data will facilitate the interpretation of cell lines, animal models and clinical observations and will reveal patterns of co-occurrence, mutual exclusivity and lineage restriction, which may provide mechanistic insights with profound therapeutic implications.

Although a handful of cancer genes are mutated at high frequency, most cancer genes mutated in most patients occur at intermediate frequencies (2–20%). To provide therapeutic options for most patients, it will therefore be critical to identify and understand the pathway-level implications of all genes mutated at intermediate frequencies (2–20%).

With growing data sets across many tumour types, pan-cancer analyses are becoming of great interest[10,11]. In this paper, we studied somatic point mutations in a collection of nearly 5,000 tumour-normal pairs across 21 cancer types. We identified a Cancer5000 set containing 254 genes, based on merging results from each tumour type and the combined set, and a stringent Cancer5000-S set containing 219 genes, accounting for multiple-hypothesis testing across the types. Nearly all previously known cancer genes in these tumour types are contained within these sets or just below statistical significance.

After eliminating genes reported in the CGC or recent papers and accounting for the expected number of false positives, the stringent Cancer5000-S set is expected to contain approximately 41 novel candidate cancer genes, with additional candidate cancer genes expected in the larger Cancer5000 set. After close inspection, we found 33 genes (21 in the stringent set and 12 more in the larger set) with strong functional connections to cancer and mutation patterns consistent with the presumed function. These genes fall within known 'hallmarks' of cancer[3,12], including cell proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis. Follow-up studies will be required to confirm and understand the functional impact of the mutations in these genes.

Beyond identifying new candidate cancer genes, our study demonstrates that we are far from having a complete catalogue of cancer genes, with many genes at clinically important frequencies within individual tumour types and across cancer as a whole still awaiting identification. The number of such genes is still increasing steeply with the number of samples and the number of tumour types studied. Importantly, these new candidate cancer genes are not rare. Substantial ongoing increases are seen in each of the 10–20%, 5–10% and 2–5% ranges (Fig. 4d).

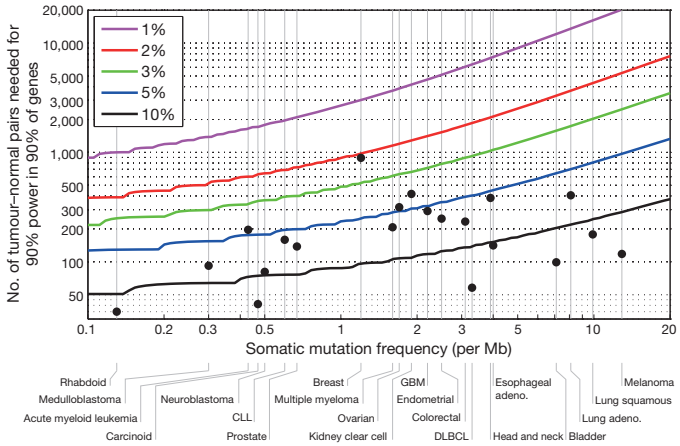

Figure 5 | Number of samples needed to detect significantly mutated genes, as a function of a tumour type's median background mutation frequency and a cancer gene's mutation rate above background. The number of samples needed to achieve 90% power for 90% of genes (y axis). Grey vertical lines indicate tumour type median background mutation frequencies (x axis). Black dots indicate sample sizes in the current study. For most tumour types, the current sample size is inadequate to reliably detect genes mutated at 5% or less above background; see also Supplementary Fig. 9. Adeno., adenocarcinoma.

Notably, of the 33 novel genes above, 5 are mutated at frequencies greater than 10% and fifteen at frequencies greater than 5%.

Creating a comprehensive catalogue of genes in which somatic point mutations propel cancer at both high (>20%) and intermediate (2–20%) frequency will require analysing an average of approximately 2,000 tumours for each of at least 50 tumour types, corresponding to approximately 100,000 tumours. (Currently defined tumour types may be divided, based on genomic information, into distinct subtypes, each of which should be analysed on its own. The ultimate number of tumour types will thus be defined iteratively by molecular analysis.)

Analysis should include both point mutations (as studied here), as well as other types of functional variation[7]. Genomic studies of such large numbers of samples is no longer prohibitive, in light of the one-million-fold decrease in the cost of DNA sequencing over the past decade. Given the devastating toll of cancer, with nearly 8 million deaths annually worldwide[13], completing the genomic analysis of this disease should be a biomedical imperative.

METHODS SUMMARY

For TCGA tumour types, mutation data were downloaded from the Synapse website. For non-TCGA tumour types, sequencing data was downloaded from dbGaP and processed through Firehose, the Broad Institute's analysis pipeline. Liftover was used to convert hg18 data. Each mutation in the combined MAF file was filtered against a panel of normal samples. Three significance metrics were calculated for each gene, using the previously described methods MutSigCV, MutSigCL, and MutSigFN. These measure the significance, respectively, of mutation burden, clustering, and functional impact. The three MutSig tests were combined into a single final P value for each gene, and q values were calculated using the method of Benjamini and Hochberg, and genes with $q \leq 0.1$ were declared to be candidate cancer genes. Down-sampling was performed within each tumour type and for the combined data set. MutSig analysis was repeated for a set of many smaller random subsets of patients. Genes were stratified by their maximal frequency across tumour types (Fig. 4d). Power analysis was performed using a binomial power model. We first calculated the probability, p_0, that a patient will have at least one non-silent mutation in a particular gene from the background model. We then calculated the signal we want to detect, $p_1 = p_0 + r(1-m)$, where r is the frequency of non-silent mutations in the population (above background) that a gene is mutated and m is the mis-detection rate of the mutation (we took $m = 0.1$). The power was then calculated using a binomial model, with $p = p_0$ representing the null hypothesis, and $p = p_1$ representing the alternative hypothesis (Methods). To obtain Fig. 5 and Supplementary Fig. 9 we found the number of tumour–normal pairs that yielded 90% power for 90% of genes as a function of background mutation frequency and r.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 September; accepted 27 November 2013.
Published online 5 January 2014.

1. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153,** 17–37 (2013).
2. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
3. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).
4. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnol.* **30,** 413–421 (2012).
5. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31,** 213–219 (2013).
6. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
7. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature Genet.* **45,** 1134–1140 (2013).
8. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl Acad. Sci. USA* **109,** 3879–3884 (2012).
9. Cancer Genome Atlas Research. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
10. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
11. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3,** 2650 (2013).
12. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
13. Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer* **127,** 2893–2917 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was conducted as part of TCGA, a project of the National Cancer Institute and the National Human Genome Research Institute. We are grateful to T. I. Zack, S. E. Schumacher, and R. Beroukhim for sharing their copy-number analyses before publication.

Author Contributions G.G., E.S.L., T.R.G., M.M., L.A.G. and S.B.G. conceived the project and provided leadership. M.S.L., G.G., E.S.L., P.S. and C.H.M. analysed the data and contributed to scientific discussions. M.S.L., E.S.L. and G.G. wrote the paper. J.T.R., M.S.L., E.S.L. and G.G. created the website for visualizing this data set.

Author Information The data analysed in this manuscript have been deposited in Synapse (http://www.synapse.org), accession number syn1729383, and in dbGaP (http://www.ncbi.nlm.nih.gov/gap), accession numbers phs000330.v1.p1, phs000348.v1.p1, phs000369.v1.p1, phs000370.v1.p1, phs000374.v1.p1, phs000435.v2.p1, phs000447.v1.p1, phs000450.v1.p1, phs000452.v1.p1, phs000467.v6.p1, phs000488.v1.p1, phs000504.v1.p1, phs000508.v1.p1, phs000579.v1.p1, phs000598.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.S.L. (lander@broadinstitute.org) and G.G. (gadgetz@broadinstitute.org).

METHODS

Mutation data and preprocessing. Mutation data were obtained as follows. For TCGA tumour types, mutation data were downloaded from the Synapse website (http://www.synapse.org), accession no. syn1729383. For non-TCGA tumour types, sequencing data was downloaded from dbGaP and processed through Firehose, the Broad Institute's analysis platform (http://www.broadinstitute.org/cancer/cga/Firehose). For tumour types that were originally aligned to build hg18, LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) was used to convert the coordinates of each mutation to build hg19. All mutation data were then combined into a single MAF file. Duplicate patients and duplicate mutations were removed. To standardize the definition of a 'splice-site' mutation, any mutation affecting the two bases before or after a splice junction, was labelled as a splice-site mutation. Filtering was performed as follows. To remove common sequencing artefacts or residual germline variation, each mutation in the combined MAF file was subjected to a 'Panel of Normals' filtering process using a panel of over 4000 BAM files from normal samples. For each mutation, the position of the mutation was examined in each normal BAM file. Mutations observed in the panel of normals were removed from the MAF. The final MAF is available at http://www.tumorportal.org/.

MutSig significance calculations. Three significance metrics were calculated for each gene, using the previously described methods MutSigCV, MutSigCL, and MutSigFN. These measure the significance of mutation burden, clustering, and functional impact, respectively (Supplementary Fig. 3). MutSigCV was described previously[6]. MutSigCV determines the P value for observing the given quantity of non-silent mutations in the gene, given the background model determined by silent (and noncoding) mutations in the same gene and the neighbouring genes of covariate space that form its 'bagel'. MutSigCL and MutSigFN were used previously[8] but were not given names in that work. Here we name the methods to reflect the type of evidence of positive selection that they are designed to detect. MutSigCL and MutSigFN measure the significance of the positional clustering of the mutations observed, as well as the significance of the tendency for mutations to occur at positions that are highly evolutionarily conserved (using conservation as a proxy for probably functional impact). MutSigCL and MutSigFN are permutation-based methods and their P values are calculated as follows: The observed nonsilent coding mutations in the gene are permuted T times (to simulate the null hypothesis, $T = 10^8$ for the most significant genes), randomly reassigning their positions, but preserving their mutational 'category', as determined by local sequence context. We used the following context categories: transitions at CpG dinucleotides, transitions at other C–G base pairs, transversions at C–G base pairs, mutations at A–T base pairs, and indels. Indels are unconstrained in terms of where they can move to in the permutations. For each of the random permutations, two scores are calculated: S_{CL} and S_{FN}, measuring the amount of clustering and function impact (measured by conservation) respectively. S_{CL} is defined to be the fraction of mutations occurring in hotspots. A hotspot is defined as a 3-base-pair region of the gene containing many mutations: at least 2, and at least 2% of the total mutations. S_{FN} is defined to be the mean of the base-pair-level conservation values for the position of each non-silent mutation, as obtained from an alignment of 45 vertebrate genomes to the human genome, the UCSC 'phyloP46way' track, which can be downloaded from (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/README.txt). To determine a P_{CL}, the P value for the observed degree of positional clustering, the observed value of S_{CL} (computed for the mutations actually observed), was compared to the distribution of S_{CL} obtained from the random permutations, and the P value was defined to be the fraction of random permutations in which S_{CL} was at least as large as the observed S_{CL}. The P value for the conservation of the mutated positions, P_{FN}, was computed analogously. Finally, we noted that the gene *AJUBA* was referred to in some analyses by the alternative name *JUB*; after reconciling this naming difference, the gene was significant and added to the list of significant genes.

Combining MutSig statistics. The three MutSig tests described above (MutSigCV, MutSigCL and MutSigFN) were combined into a single final P value as follows. First, a joint P value (CL + FN) for the observed clustering and conservation was calculated from the joint probability distribution of the random permutations. Next, this was combined with the MutSigCV P value using two methods: the Fisher method of combining P values from independent tests (http://en.wikipedia.org/wiki/Fisher's_method); the truncated product method (TPM) for combining P values, which rewards highly significant P values in any one of the tests The combined P values for both methods were extremely similar. We examined the performance of each of the three metrics separately and each pairwise combination of two metrics. The results of these analyses are presented in Supplementary Table 1 (last tab) and summarized in Supplementary Table 5.

Multiple hypothesis corrections. In the analysis of each tumour type, a total of 18,388 genes were analysed. To correct for these multiple hypotheses, the final MutSig P values were converted to FDR (q values) using the method of Benjamini and Hochberg, and genes with $q \leq 0.1$ were declared to be significantly mutated. This was also done for the analysis of the combined cohort. Genes with $q \leq 0.1$ in any tumour-type analysis or in the combined-cohort analysis were declared to be a member of the Cancer5000 list of candidate cancer genes.

To correct for the 22 analyses thus combined (corresponding to 22 chances for each gene to become significant), a further level of multiple hypothesis correction was applied. A list was made of the 18,388 genes × 22 analyses = 404,536 hypotheses. The Benjamini–Hochberg method was applied to this full set, yielding new FDR q values. Any gene involved in these gene × tumour-type pairs was declared to be a member of the stringently corrected Cancer5000-S list of genes.

Down-sampling analyses. To analyse the dependence of the number of significantly mutated cancer-associated genes upon the size of the data set being analysed, down-sampling was performed. Three different down-sampling analyses are described: first, down-sampling within each tumour type; second, down-sampling of the number of different tumour types; and third, down-sampling of the full Cancer5000-S procedure.

Down-sampling within each tumour type (Fig. 4a and Supplementary Fig. 7): for each tumour type, the MutSig analysis was repeated for a set of many smaller subsets of patients from that tumour type. The sizes of the subsets were chosen to sample regularly the interval from zero patients to the final total number of patients that were in the full analysis. For each of the random subsets thus defined, we repeated the full MutSig calculation (MutSigCV + MutSigCL + MutSigFN) and combined the results of the three tests as described above. This enabled us to determine which genes remained significant when analysing each smaller subset. We counted how many of the genes remained significant at each smaller set size, and plotted this number as a smoothed function of set size. This allowed us to demonstrate that the number of significantly mutated genes detected is continuing to rise steeply in each tumour type. We also repeated this same analysis for the full combined data set (4,742 patients), with similar results.

Down-sampling of the number of different tumour types (Fig. 4b). To examine the effect of adding whole tumour types, we performed the following analysis. We constructed 25 random orderings of the 21 tumour types, and for each ordering we constructed 20 subsets by sequentially adding whole tumour types according to that ordering. Then we repeated the whole MutSig analysis for each of these subsets. This produced a set of curves showing how the number of significantly mutated genes increased as a function of the number of tumour types included in the analysis. The curve depended on the exact ordering of the tumour types as they were added, but all curves showed a steady increase in the number of genes, even at the highest numbers of tumour types. This demonstrated the importance of continuing to sample additional tumour types. We also repeated the analysis with subtraction of the expected number of false positives (Supplementary Fig. 8a); the results were qualitatively unchanged.

Down-sampling of the full Cancer5000-S procedure (Fig. 4c): we repeated our procedure of constructing the Cancer5000-S list by applying the stringent procedure of correction for the approximately 400,000 hypothesis (18,388 genes × 22 analyses), and computed how many genes remained significant at each smaller set size. We plotted the number of significantly mutated genes detected as a function of set size. This produced a curve similar to down-sampling within each tumour type, with the number of significant genes continuing to rise steeply even at the largest set sizes. We also repeated the analysis with subtraction of the expected number of false positives (Supplementary Fig. 8b); the results were qualitatively unchanged. Furthermore, we stratified the genes according to their frequency (calculated as the maximal frequency across tumour types), and plotted separate curves for each of the following frequency categories: 20% and above, 10–20%, 5–10%, 3–5%, 2–3%, and below 2%. This clearly demonstrated that the 20% and above genes have largely been discovered. In contrast, genes at lower frequencies are continuing to be discovered (Fig. 4d). Note that rerunning the analysis produces slightly different results in every run since the calculation of P values has a stochastic component. The genes at the edge of significance (that is, ones with q value close to 0.1) may be declared as significant or insignificant with respect to the cut-off of $q = 0.1$ in different analyses. This slight fluctuation is standard for permutation-based methods.

Power calculations. Power analysis was performed using a binomial power model. We first calculated the probability, p_0, that a patient will have at least one non-silent mutation in a particular gene from the background model. The calculation is based on the length of the gene, L (in coding bases), the background mutation frequency, μ (in mutations per base), the gene-specific mutation rate factor, f_g, (calculated by MutSigCV), the 3:1 typical ratio of non-silent to silent mutations; $p_0 = 1 - (1 - \mu f_g)^{(3L/4)}$. We used $L = 1{,}500$, and $f_g = 3.9$ (representing the 90th percentile of $f_g\, L_g/1{,}500$ across the approximately 18,000 genes and $f_g = 1$ for the 50th percentile gene). We then calculated the signal we want to detect, $p_1 = p_0 + r(1 - m)$, where r is the frequency of non-silent mutations in the population (above background) that a gene is mutated and m is the mis-detection rate of the mutation (we took $m = 0.1$). The power was then calculated by: first,

using a binomial of N trials (that is, N patients) and $p = p_{0,}$ finding the minimal number of patients with mutations that reach a genome-wide significance level ($P \leq 5\times10^{-6}$); and second, calculating the power as the probability of observing at least this many patients with mutations when using a binomial with $p = p_1$. To obtain Fig. 5 and Supplementary Fig. 9 we found the values of N that yielded 90% power as a function of μ and r.