doi:10.1038/nature12213

Mutational heterogeneity in cancer and the search for new cancer-associated genes

Michael S. Lawrence[1]*, Petar Stojanov[1,2]*, Paz Polak[1,3,4]*, Gregory V. Kryukov[1,3,4], Kristian Cibulskis[1], Andrey Sivachenko[1], Scott L. Carter[1], Chip Stewart[1], Craig H. Mermel[1,5], Steven A. Roberts[6], Adam Kiezun[1], Peter S. Hammerman[1,2], Aaron McKenna[1,7], Yotam Drier[1,3,5,8], Lihua Zou[1], Alex H. Ramos[1], Trevor J. Pugh[1,2,3], Nicolas Stransky[1,9], Elena Helman[1,10], Jaegil Kim[1], Carrie Sougnez[1], Lauren Ambrogio[1], Elizabeth Nickerson[1], Erica Shefler[1], Maria L. Cortés[1], Daniel Auclair[1], Gordon Saksena[1], Douglas Voet[1], Michael Noble[1], Daniel DiCara[1], Pei Lin[1], Lee Lichtenstein[1], David I. Heiman[1], Timothy Fennell[1], Marcin Imielinski[1,5], Bryan Hernandez[1], Eran Hodis[1,2], Sylvan Baca[1,2], Austin M. Dulak[1,2], Jens Lohr[1,2], Dan-Avi Landau[1,2,11], Catherine J. Wu[2,3], Jorge Melendez-Zajgla[12], Alfredo Hidalgo-Miranda[12], Amnon Koren[1,3], Steven A. McCarroll[1,3], Jaume Mora[13], Ryan S. Lee[2,3,14], Brian Crompton[2,14], Robert Onofrio[1], Melissa Parkin[1], Wendy Winckler[1], Kristin Ardlie[1], Stacey B. Gabriel[1], Charles W. M. Roberts[2,3,14], Jaclyn A. Biegel[15], Kimberly Stegmaier[1,2,14], Adam J. Bass[1,2,3], Levi A. Garraway[1,2,3], Matthew Meyerson[1,2,3], Todd R. Golub[1,2,3,8], Dmitry A. Gordenin[6], Shamil Sunyaev[1,3,4], Eric S. Lander[1,3,10] & Gad Getz[1,5]

Major international projects are underway that are aimed at creating a comprehensive catalogue of all the genes responsible for the initiation and progression of cancer[1–9]. These studies involve the sequencing of matched tumour–normal samples followed by mathematical analysis to identify those genes in which mutations occur more frequently than expected by random chance. Here we describe a fundamental problem with cancer genome studies: as the sample size increases, the list of putatively significant genes produced by current analytical methods burgeons into the hundreds. The list includes many implausible genes (such as those encoding olfactory receptors and the muscle protein titin), suggesting extensive false-positive findings that overshadow true driver events. We show that this problem stems largely from mutational heterogeneity and provide a novel analytical methodology, MutSigCV, for resolving the problem. We apply MutSigCV to exome sequences from 3,083 tumour–normal pairs and discover extraordinary variation in mutation frequency and spectrum within cancer types, which sheds light on mutational processes and disease aetiology, and in mutation frequency across the genome, which is strongly correlated with DNA replication timing and also with transcriptional activity. By incorporating mutational heterogeneity into the analyses, MutSigCV is able to eliminate most of the apparent artefactual findings and enable the identification of genes truly associated with cancer.

Recent cancer genome studies have led to the identification of scores of cancer-associated genes in glioblastoma[1], ovarian[2], colorectal[3], lung[4], head and neck[5], multiple myeloma[6], chronic lymphocytic leukaemia[7], diffuse large B-cell lymphoma (DLBCL)[8,9] and many other cancers. Studies are now underway through The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) and the International Cancer Genome Consortium (http://www.icgc.org/) to create a comprehensive catalogue of significantly mutated genes across all major cancer types.

The expectation has been that larger sample sizes will increase the power both to detect true cancer driver genes (sensitivity) and to distinguish them from the background of random mutations (specificity). Alarmingly, recent results seem to show the opposite phenomenon: with large sample sizes, the list of apparently significant cancer-associated genes grows rapidly and implausibly. For example, when we applied current analytical methods to whole-exome sequence data from 178 tumour–normal pairs of lung squamous cell carcinoma[10], a total of 450 genes (Supplementary Table 1 and Supplementary Methods 2) were found to be mutated at a significant frequency (false-discovery rate $q < 0.1$). Although the list contains some genes known to be associated with cancer, many of the genes seem highly suspicious on the basis of their biological function or genomic properties. Almost a quarter (101/450) of the putative significant genes encode olfactory receptors. The list is also highly enriched for genes encoding extremely large proteins, including more than one-fifth of the 83 genes encoding proteins with >4,000 amino acids ($P < 10^{-11}$, Fisher's exact test). These include the two longest human proteins, the muscle protein titin (36,800 amino acids) and the membrane-associated mucin *MUC16* (14,500 amino acids), as well as another mucin (*MUC4*), cardiac ryanodine receptors (*RYR2*, *RYR3*), cytoskeletal dyneins (*DNAH5*, *DNAH11*) and the neuronal synaptic vesicle protein piccolo (*PCLO*). The prominence of these genes is not simply the consequence of their long coding regions, because the statistical tests already account for the larger target size. Furthermore, the list also contains genes with very long introns, including one-sixth of the 73 genes spanning a genomic region of >1 megabase (Mb) ($P < 10^{-6}$), such as those encoding cub- and sushi-domain proteins (*CSMD1*, *CSMD3*), and many neuronal proteins, such as the neurexins *NRXN1*, *NRXN4* (also known as *CNTNAP2*), *CNTNAP4* and *CNTNAP5*, the neural adhesion molecule *CNTN5*, and the Parkinson's disease protein *PARK2*. When we performed similar analyses for several other cancer types with many samples, we similarly obtained large lists including many of the same genes (data not shown).

After recognizing the problem of apparent false-positive findings, we reviewed the published literature and found that some of these potentially spurious genes have already been nominated as cancer-associated genes in recently published cancer genome studies: for example, *LRP1B* in glioblastoma[2] and lung adenocarcinoma[1,4]; *CSMD3* in ovarian cancer[2]; *PCLO* in DLBCL[9]; *MUC16* in lung squamous carcinoma[11], breast cancer[12] and DLBCL[8]; *MUC4* in melanoma[13]; olfactory receptor *OR2L13* in glioblastoma[14]; and *TTN* in breast cancer[12] and other tumour types[15]. We therefore set out to understand the source of the problem.

Analytical approaches in wide use today[1–9,13–16] identify as significantly mutated those genes harbouring more mutations than expected given the average background mutation frequency for the cancer type.

[1]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA. [2]Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [3]Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [5]Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [6]Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, NIH, DHHS, Durham, North Carolina 27709, USA. [7]Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [8]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. [9]Blueprint Medicines, Cambridge, Massachusetts 02142, USA. [10]Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [11]Yale Cancer Center, Department of Hematology, New Haven, Connecticut 06510, USA. [12]Instituto Nacional de Medicina Genómica, Mexico City 14610, Mexico. [13]Department of Pediatric Oncology, Hospital Sant Joan de Déu, Barcelona 08950, Spain. [14]Boston Children's Hospital, Boston, Massachusetts 02115, USA. [15]Children's Hospital, Philadelphia, Pennsylvania 19104, USA.
*These authors contributed equally to this work.

These methods use a handful of parameters: an average overall mutation frequency for a cancer type; and a few parameters about the relative frequencies of different categories of mutations (small insertions/deletions and transitions versus transversions at CpG dinucleotides, other C:G base pairs and A:T base pairs). Average values of these parameters are typically estimated from the samples under study. Various efforts, by us and others, have recently began to incorporate sample-specific mutation rates into the analysis[3,9].

We proposed that the problem might be due to heterogeneity in the mutational processes in cancer. Whereas it is obvious that assuming an average mutation frequency that is too low will lead to spuriously significant findings, it is less well appreciated that using the correct average rate but failing to account for heterogeneity in the mutational process can also lead to incorrect results. To illustrate this point, we compared two simple scenarios both sharing the same average mutation frequency: (1) a constant frequency of 10 mutations per Mb (10/Mb) across all genes, versus (2) frequencies of 4/Mb, 8/Mb and 20/Mb in 25%, 50% and 25% of genes, respectively (Supplementary Fig. 1). If the second case is analysed under the erroneous assumption of a constant rate, many of the highly mutable genes will falsely be declared to be associated with cancer. Notably, the problem grows with sample size: because the threshold for statistical significance decreases with sample size, modest deviations due to an erroneous model are declared significant. For the same reason, the problem is also more pronounced in tumour types with higher mutation rates. Heterogeneity in mutation frequencies across patients can also lead to inaccurate results, including the potential to produce both false-positive, as described earlier, and false-negative results if the baseline frequency is overestimated.

We therefore set out to study heterogeneity in mutation rates, using a data set of 3,083 tumour–normal pairs across 27 tumour types, for which the whole-exome sequence was available for 2,957 and the whole-genome sequence was available for 126 (Supplementary Table 2). Approximately 92% of the samples were sequenced at the Broad Institute and thus were processed using a uniform experimental and analytical pipeline (see Methods). In this data set, an average of 30 Mb of coding sequence per sample was covered to adequate depth for mutation detection, yielding a total of 373,909 non-silent coding mutations or an average of 4.0/Mb per sample (median of 44 non-silent coding mutations per sample, or 1.5/Mb).

We analysed three types of heterogeneity, with the aim of achieving more accurate detection of cancer-associated genes. First, we analysed heterogeneity across patients with a given cancer type. Analysis of the 27 cancer types revealed that the median frequency of non-synonymous mutations varied by more than 1,000-fold across cancer types (Fig. 1). About half of the variation in mutation frequencies (measured on a logarithmic scale) can be explained by tissue type of origin. Paediatric cancers showed frequencies as low as 0.1/Mb (approximately one change across the entire exome), whereas at the opposite extreme, melanoma and lung cancer exceeded 100/Mb. The highest mutation frequencies are in some cases attributable to extensive exposure to well known carcinogens, such as ultraviolet radiation in the case of melanoma and tobacco smoke in the case of lung cancers.

More surprisingly, mutation frequencies varied markedly across patients within a cancer type. In melanoma and lung cancer, the frequency ranged across 0.1–100/Mb. Despite the low median frequency in acute myeloid leukaemia (AML; 0.37/Mb), the patient-specific frequencies similarly spanned three orders of magnitude, from 0.01 to 10/Mb. Variation may in some cases be due to key biological factors, such as melanomas not attributed to ultraviolet exposure or on unexposed skin, colon cancers with or without mismatch repair defects[3], or head and neck tumours with viral or non-viral origin[5] (Supplementary Fig. 2).

Second, after analysing total mutation frequency, we analysed heterogeneity in the mutational spectrum of the tumours. Starting with all 96 possible mutations (12 mutations at a base times 16 possible flanking bases, then collapsed by strand symmetry), we used non-negative matrix factorization (NMF) to reduce the dimensionality, with each spectrum represented as a linear combination of six basic spectra (Methods). We represented the mutational spectrum of each tumour on a circular plot, with distance from the origin representing total mutation rate and angle representing the relative contribution of the

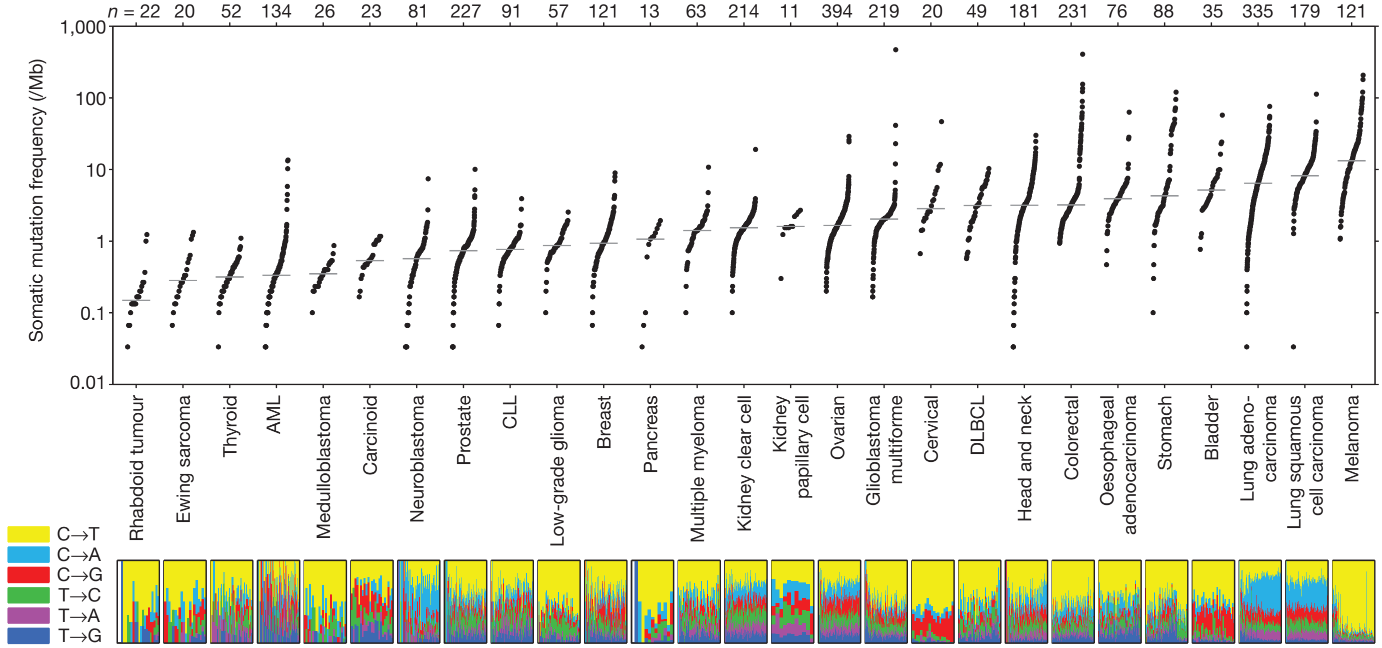

Figure 1 | Somatic mutation frequencies observed in exomes from 3,083 tumour–normal pairs. Each dot corresponds to a tumour–normal pair, with vertical position indicating the total frequency of somatic mutations in the exome. Tumour types are ordered by their median somatic mutation frequency, with the lowest frequencies (left) found in haematological and paediatric tumours, and the highest (right) in tumours induced by carcinogens such as tobacco smoke and ultraviolet light. Mutation frequencies vary more than 1,000-fold between lowest and highest across different cancers and also within several tumour types. The bottom panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend on the left. See also Supplementary Table 2.

six basic spectra (Fig. 2). This representation reveals natural groupings with respect to mutational spectrum.

Lung cancers (Fig. 2, red cluster at 2 o'clock position), for example, share a mutational spectrum dominated by C→A mutations, consistent with their exposure to the polycyclic aromatic hydrocarbons in tobacco smoke[17]. Melanoma (Fig. 2, black cluster at 12 o'clock) shows a distinct pattern reflecting the frequent C→T mutations caused by misrepair of ultraviolet-induced covalent bonds between adjacent pyrimidines[18]. Gastrointenstinal tumours (oesophageal, colororectal and gastric; Fig. 2, green cluster at 8 o'clock) show extremely high frequencies of transition mutations at CpG dinucleotides, which may reflect higher methylation levels in these tumour types[3].

Interestingly, there is a multifarious cluster at the 10 o'clock position in Fig. 2 corresponding to cervical, bladder and some head and neck tumours, all sharing frequent mutations at Cs in the TpC context (that is, Cs with a T on their 5′ side) that change the C to either T or G or (less often) A. This pattern is characteristic of mutations caused by the APOBEC family of cytidine deaminases, innate immunity enzymes restricting the propagation of retroviruses and retrotransposons[19,20]. Some APOBECs can be induced by certain classes of viruses[21]. Cervical cancer is known to be caused in over 90% of cases by the human papillomavirus (HPV)[22]. Recent studies have also implicated HPV in head and neck cancers[5]. The similar mutational spectrum in bladder cancer may indicate a viral aetiology in a significant subset of this tumour type; a potential role of HPV in bladder cancer is a subject of active investigation[23]. This cluster also contains sporadic examples of breast tumours (consistent with a recent report[12]), as well as some tumours from lung and other tissues. Recent work[19,20] has shown that the TpC mutations tend to occur in proximity to one another, consistent with the activity of APOBEC enzymes in damaged long single-strand DNA regions. One last minor cluster (Fig. 2, 4 o'clock position) consists of samples dominated by A→T mutations in the TpA context. This cluster contains mostly leukaemia samples (AML and chronic lymphocytic leukaemia (CLL)), as well as one breast cancer sample and one neuroblastoma sample.

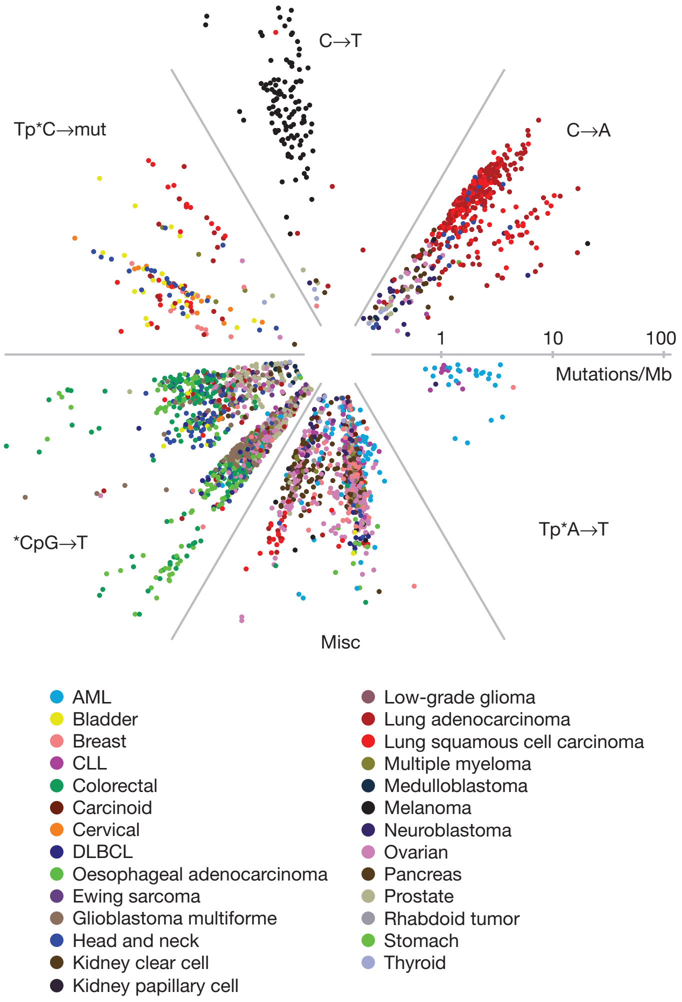

Figure 2 | Radial spectrum plot of the 2,892 tumour samples with at least 10 coding mutations. The angular space is compartmentalized into the six different factors discovered by NMF (see Methods). The distance from the centre represents the total mutation frequency. Different tumour types segregate into different compartments based on their mutation spectra. Notable examples are: lung adenocarcinoma and lung squamous carcinoma (red; 2 o'clock position); melanoma (black; 12 o'clock position); stomach, oesophageal and colorectal cancer (various shades of green; 8 o'clock position); samples harbouring mutations of the HPV or APOBEC signature (bladder, cervical and head and neck cancer, marked in yellow, orange and blue, respectively; 10 o'clock position); and AML and CLL samples sharing the Tp*A→T signature, 4 o'clock position. Misc, miscellaneous. See also Supplementary Table 3.

The rich variation in mutational spectrum across tumours underscores the problems with using an overly simplistic model of the average mutational process for a tumour type and failing to account for heterogeneity within a tumour type.

Of all the kinds of heterogeneity in mutational processes, the most important turns out to be the third kind we analysed: regional heterogeneity across the genome. By examining the whole-genome sequence from 126 tumour–normal pairs across ten tumour types, we found marked variation in mutation frequency across the genome, with differences exceeding fivefold (Fig. 3a, b); the profile of the genomic variation was similar across and within tumour types (Supplementary Fig. 3). Recent studies have noted regional variation in cancer mutation rates and begun to explore correlations with genomic features[6,17,18,24].

We focused on two factors that were especially powerful in explaining mutational heterogeneity. The first factor is gene expression level. It is known that the germline mutation rate is somewhat lower in genes that are highly expressed in the germ line[18], owing to a process termed transcription-coupled repair[25]. With the whole-genome and whole-exome data analysed here, we found a strong correlation between somatic mutation frequency in cancers and gene expression level (averaged across many cell lines, with similar results for expression in matched normal tissue) (Fig. 3a, b and Supplementary Fig. 3 and Supplementary Tables 4, 5). The average mutation rate is ~2.9-fold higher in the bottom expression level percentile than in the top one. Although statistically highly significant, this effect is insufficient to explain regional variation in mutation levels fully.

The second important factor is the replication time of a DNA region during the cell cycle. Recent studies have reported that germline mutation rates are correlated with DNA replication time[26–28]: late-replicating regions have much higher mutation rates, possibly due to depletion of the pool of free nucleotides[26]. With the whole-genome and whole-exome data here, we see a marked correlation between somatic mutation frequency in cancers and DNA replication timing (as measured in HeLa cells[27]) (Fig. 3a, b), with similar results for blood cell lines[28] (Supplementary Fig. 3). The average mutation rate is ~2.9-fold higher in the latest- versus earliest-replicating percentile, and there is a ~2.1-fold difference in mutation rate between the latest- and earliest-replicating decile.

These two features explain most of the suspicious entries on the putative cancer-associated gene lists. Olfactory receptor genes, for example, have low expression ($P < 10^{-172}$, Kolmogorov–Smirnoff test; Fig. 3e), are uniformly late in replication timing ($P < 10^{-109}$; Fig. 3f) and have a high regional non-coding mutation rate ($P < 10^{-81}$), which accounts for the high frequency of somatic mutations in their coding regions. Large genes have similarly low expression and are late replicating (Fig. 3e, f), including the genes cited in the lung cancer example earlier, such as titin and the ryanodine receptors. Importantly, these results undermine the evidence supporting several recent reports, such as the suggestion that *CSMD3* is associated with ovarian cancer[2]. As

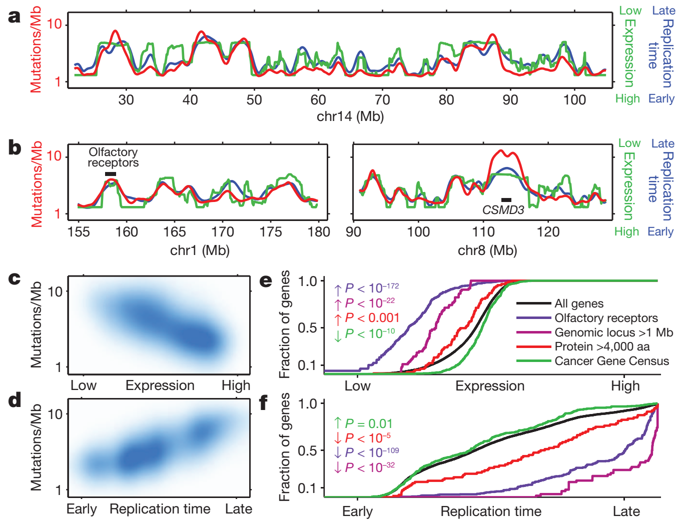

Figure 3 | Mutation rate varies widely across the genome and correlates with DNA replication time and expression level. a, **b**, Mutation rate, replication time and expression level plotted across selected regions of the genome. Red shows total non-coding mutation rate calculated from whole-genome sequences of 126 samples (excluding exons). Blue shows replication time[27]. Green shows average expression level across 91 cell lines in the Cancer Cell Line Encyclopedia determined by RNA sequencing. Note that low expression is at the top of the scale and high expression at the bottom, in order to emphasize the mutual correlations with the other variables. Panels show entire chromosome 14 (**a**) and portions of chromosomes 1 and 8 (**b**), with the locations of two specific loci: a cluster of 16 olfactory receptors on chromosome (chr)1 and the gene *CSMD3* on chromosome 8. These two loci have very high mutation rates, late replication times and low expression levels. The local mutation rate at *CSMD3* is even higher than predicted from replication time and expression, suggesting contributions from additional factors, perhaps locally increased DNA breakage—the locus is a known fragile site. **c**, **d**, Correlation of mutation rate with expression level and replication time for all 100 Kb windows across the genome. **e**, **f**, Cumulative distribution of various gene families as a function of expression level and replication time. Olfactory receptor genes, genes encoding long proteins (>4,000 amino acids (aa)) and genes spanning large genomic loci (>1 Mb) are significantly enriched towards lower expression and later replication. By contrast, known cancer-associated genes (as listed in the Cancer Gene Census) trend towards slightly higher expression and earlier replication. See also Supplementary Fig. 9 and Supplementary Tables 4, 5 and 6.

an independent test, we confirmed that these two genomic features correlated strongly with the overall frequency of silent substitutions in coding regions and mutations in introns (Fig. 3c, d and Supplementary Table 6). However, we note that silent substitutions alone provide inadequate data to correct mutation frequencies on a gene-by-gene basis in most tumour types and for most genes, owing to the sparsity of the data and the resulting uncertainty in estimated rates.

Using the observations above, we developed a new integrated approach to identify significantly mutated genes in cancer. The method (MutSigCV) corrects for variation by using patient-specific mutation frequency and spectrum, and gene-specific background mutation rates incorporating expression level and replication time (Supplementary Methods 3). MutSigCV is freely available for non-commercial use (http://www.broadinstitute.org/cancer/cga/mutsig).

When we applied MutSigCV to the lung cancer example earlier, the list of significantly mutated genes shrank from 450 to 11 genes. Most of the genes in this shorter list have been previously reported to be mutated in squamous cell lung cancer (*TP53*, *KEAP1*, *NFE2L2*, *CDKN2A*, *PIK3CA*, *PTEN*, *RB1*; refs 11, 16) or in other tumour types (*MLL2* (also known as *KMT2D*), *NOTCH1*, *FBXW7*). An additional novel gene in the list, *HLA-A*, suggests that mutations in immune-related genes may help tumours evade immune surveillance, a finding that requires follow-up experimental work. These significantly mutated genes are discussed in the TCGA lung squamous publication[10], in which we applied our novel methodology.

With the ability to eliminate many obviously suspicious genes, it is now feasible to start analysing large cancer collections, including combined data sets across many cancer types.

We note that other forms of heterogeneity in tumours merit further investigation. These include the co-occurrence of many mutations in proximity to each other ('kataegis'[19] or 'clustered mutations'[20]) (see Supplementary Fig. 10) and transcription-coupled repair (see Supplementary Fig. 11). In addition, it will be crucial to have a full understanding of heterogeneity across cancer cells within a tumour, reflecting the evolutionary process of a tumour[29].

Our results make clear that the accurate identification of new cancer-associated genes will require accurate accounting of mutational processes. Although MutSigCV resolves the most serious current problems, the ultimate solution will probably involve using empirically observed local mutation rates obtained from massive amounts of whole-genome sequencing.

METHODS SUMMARY

All samples were obtained under Institutional Review Board approval and with documented informed consent. A complete list of samples is given in Supplementary Table 2. Whole-exome capture libraries were constructed and sequenced on Illumina HiSeq flowcells to an average coverage of 118×. Whole-genome sequencing was done with the Illumina GA-II or Illumina HiSeq sequencer, achieving an average of ~30× coverage depth. Reads were aligned to the reference human genome build hg19 using an implementation of the Burrows-Wheeler Aligner, and a BAM file was produced for each tumour and normal sample using the Picard pipeline[6]. The Firehose pipeline was used to manage input and output files and submit analyses for execution. The MuTect[30] and Indelocator (A. Sivachenko *et al.*, manuscript in preparation) algorithms were used to identify somatic single-nucleotide variants and short somatic insertions and deletions, respectively. Mutation spectra were analysed using NMF. Significantly mutated genes were identified using MutSigCV, which estimates the background mutation rate for each gene–patient–category combination based on the observed silent mutations in the gene and non-coding mutations in the surrounding regions. Because in most cases these data are too sparse to obtain accurate estimates, we increased accuracy by pooling data from other genes with similar properties (for example, replication time, expression level). Significance levels (*P* values) were determined by testing whether the observed mutations in a gene significantly exceeded the expected counts based on the background model. False-discovery rates (*q* values) were then calculated, and genes with $q \leq 0.1$ were reported as significantly mutated. Full details on methods used are listed in Supplementary Information.

Received 14 June 2012; accepted 22 April 2013.
Published online 16 June 2013.

1. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
3. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
4. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
5. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333,** 1157–1160 (2011).
6. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
7. Wang, L. *et al. SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* **365,** 2497–2506 (2011).
8. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
9. 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).
10. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
11. Shibata, T. *et al.* Cancer related mutations in *NRF2* impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. *Proc. Natl Acad. Sci. USA* **105,** 13568–13573 (2008).
12. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).

13. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485,** 502–506 (2012).
14. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).
15. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
16. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466,** 869–873 (2010).
17. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).
18. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
19. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–993 (2012).
20. Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46,** 424–435 (2012).
21. Vartanian, J. P., Guetard, D., Henry, M. & Wain-Hobson, S. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science* **320,** 230–233 (2008).
22. Walboomers, J. M. *et al.* Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J. Pathol.* **189,** 12–19 (1999).
23. Jimenez-Pacheco, A., Exposito-Ruiz, M., Arrabal-Polo, M. A. & Lopez-Luque, A. J. Meta-analysis of studies analyzing the role of human papillomavirus in the development of bladder carcinoma. *Korean J. Urol.* **53,** 240–247 (2012).
24. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Rev. Genet.* **12,** 756–766 (2011).
25. Fousteri, M. & Mullenders, L. H. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res.* **18,** 73–84 (2008).
26. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nature Genet.* **41,** 393–395 (2009).
27. Chen, C. L. *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* **20,** 447–457 (2010).
28. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91,** 1033–1040 (2012).
29. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152,** 714–726 (2013).
30. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31,** 213–219 (2013).

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 National Human Genome Research Institute. This work was conducted as part of the Slim Initiative for Genomic Medicine, a joint US–Mexico project founded by the Carlos Slim Health Institute. Support to D.A.G. and S.A.R. was through the Intramural Research Program of the National Institute of Environmental Health Sciences (National Institutes of Health, United States Department of Health and Human Services) project ES065073 (principal investigator M. Resnick).

Author Contributions G.G., E.S.L., S.S., D.A.G., T.R.G., M.M., L.A.G., A.J.B., K.S., J.A.B., C.W.M.R., S.B.G., C.J.W., S.A.M., J.M.-Z. and A.H.-M. conceived the project and provided leadership. C.So., L.A., E.N., E.S., M.L.C., D.A., W.W. and K.A. provided project management. W.W., K.A., T.F., R.O. and M.P. planned and carried out DNA sequencing and genetic analysis. T.F., D.V., G.S., M.N., D.D., P.L., L.L. and D.I.H. developed and engineered software to support the project. M.S.L., P.S., P.P., G.V.K., K.C., A.S., S.L.C., C.St., C.H.M., S.A.R., A.Ki., P.S.H., A.M., Y.D., L.Z., A.H.R., T.J.P., N.S., E.H., J.K., M.I., B.H., E.H., S.B., A.M.D., J.L., D.-A.L., C.J.W., J.M.-Z., A.H.-M., A.Ko., S.A.M., R.S.L., J.M., B.C., A.J.B. and D.A.G. analysed the data and contributed to scientific discussions. M.S.L., P.S., P.P., E.S.L. and G.G. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. 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).