Signatures of mutational processes in human cancer

Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T. W. Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N. J. Tutt, Rafael Valdés-Mas, Marit M. van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, Michael R. Stratton

Angaben zur Veröffentlichung / Publication details:

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. "Signatures of mutational processes in human cancer." *Nature* 500 (7463): 415–21. https://doi.org/10.1038/nature12477.

Signatures of mutational processes in human cancer

A list of authors and their affiliations appears at the end of the paper

All cancers are caused by somatic mutations; however, understanding of the biological processes generating these mutations is limited. The catalogue of somatic mutations from a cancer genome bears the signatures of the mutational processes that have been operative. Here we analysed 4,938,362 mutations from 7,042 cancers and extracted more than 20 distinct mutational signatures. Some are present in many cancer types, notably a signature attributed to the APOBEC family of cytidine deaminases, whereas others are confined to a single cancer class. Certain signatures are associated with age of the patient at cancer diagnosis, known mutagenic exposures or defects in DNA maintenance, but many are of cryptic origin. In addition to these genome–wide mutational signatures, hypermutation localized to small genomic regions, 'kataegis', is found in many cancer types. The results reveal the diversity of mutational processes underlying the development of cancer, with potential implications for understanding of cancer aetiology, prevention and therapy.

Somatic mutations found in cancer genomes[1] may be the consequence of the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA, or defective DNA repair. In some cancer types, a substantial proportion of somatic mutations are known to be generated by exposures, for example, tobacco smoking in lung cancers and ultraviolet light in skin cancers[2], or by abnormalities of DNA maintenance, for example, defective DNA mismatch repair in some colorectal cancers[3]. However, our understanding of the mutational processes that cause somatic mutations in most cancer classes is remarkably limited.

Different mutational processes often generate different combinations of mutation types, termed 'signatures'. Until recently, mutational signatures in human cancer have been explored through a small number of frequently mutated cancer genes, notably *TP53* (ref. 4). Although informative, these studies have limitations. To generate a mutational signature, a single mutation from each cancer sample is entered into a mutation set aggregated from several cases of a particular cancer type. A signature that contributes the large majority of somatic mutations in the tumour class is accurately reported. However, if multiple mutational processes are operative, a jumbled composite signature is generated. Furthermore, because such studies are based on 'driver' mutations[1], signatures of selection are superimposed on the signatures of mutational processes.

Recent advances in sequencing technology have overcome past limitations of scale[1]. Thousands of somatic mutations can now be identified in a single cancer sample, offering the possibility of deciphering mutational signatures even when several mutational processes are

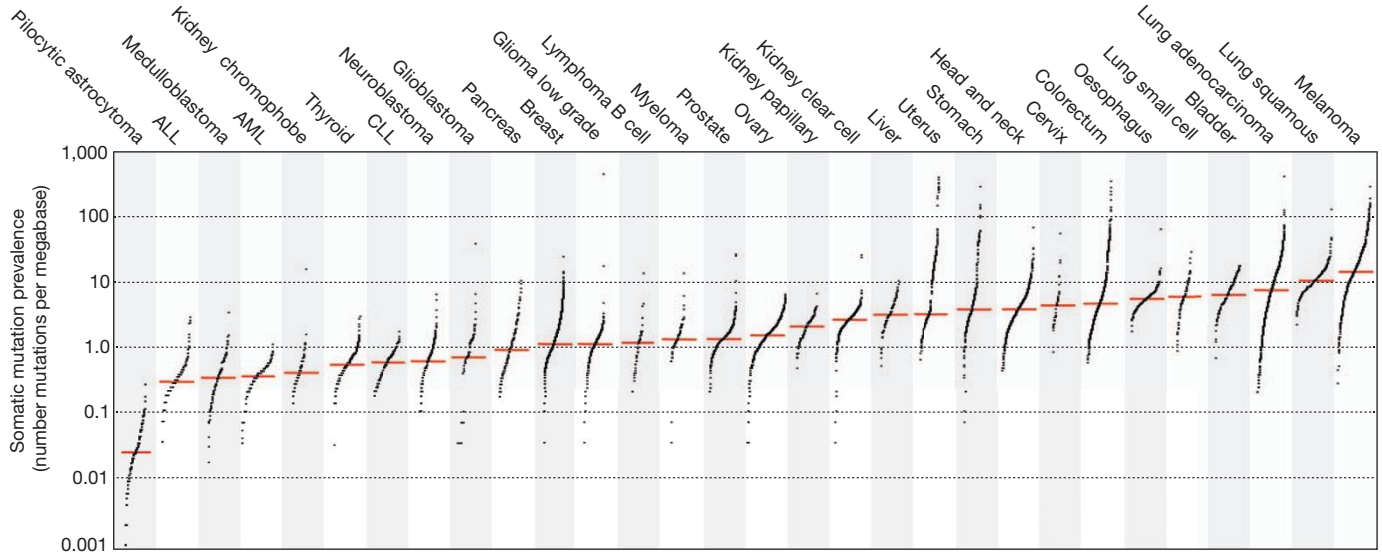

Figure 1 | The prevalence of somatic mutations across human cancer types. Every dot represents a sample whereas the red horizontal lines are the median numbers of mutations in the respective cancer types. The vertical axis (log scaled) shows the number of mutations per megabase whereas the different cancer types are ordered on the horizontal axis based on their median numbers of somatic mutations. We thank G. Getz and colleagues for the design of this figure[26]. ALL, acute lymphoblastic leukaemia; AML, acute myeloid leukaemia; CLL, chronic lymphocytic leukaemia.

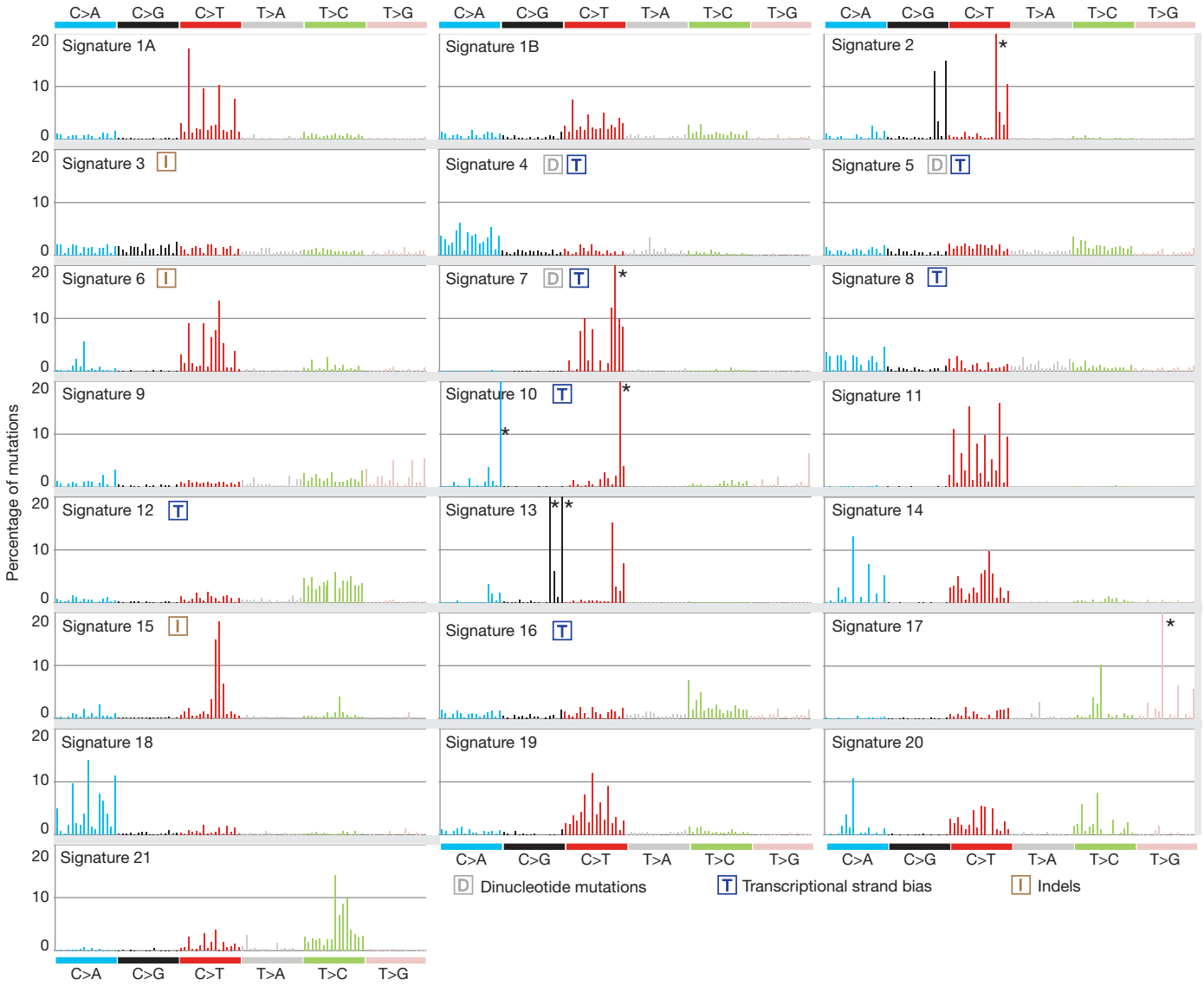

Figure 2 | Validated mutational signatures found in human cancer. Each signature is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 3′ and 5′ to the mutated base. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. All mutational signatures are displayed on the basis of the trinucleotide frequency of the human genome. A higher resolution of each panel is found respectively in Supplementary Figs 2–23. Asterisk indicates mutation type exceeding 20%.

operative. Moreover, because most mutations in cancer genomes are 'passengers'[1] they do not bear strong imprints of selection.

We recently developed an algorithm to extract mutational signatures from catalogues of somatic mutations and applied it to 21 breast cancer whole-genome sequences[5,6]. Novel and known signatures were revealed, with the contribution of each signature to each cancer sample and the timing of its activity estimated[6,7]. Further studies have demonstrated that the approach can also be applied, albeit with less power, to mutational catalogues from sequences of all coding exons (exomes)[5]. Global sequencing initiatives are now yielding catalogues of somatic mutations from thousands of cancers[8]. We have therefore applied this method to survey the repertoire of mutational signatures and processes operating across the spectrum of human neoplasia.

Mutational catalogues

We compiled 4,938,362 somatic substitutions and small insertions/deletions (indels) from the mutational catalogues of 7,042 primary cancers of 30 different classes (507 from whole genome and 6,535 from exome sequences) (Supplementary Fig. 1). In all cases, normal DNA from the same individuals had been sequenced to establish the somatic origin of variants.

The prevalence of somatic mutations was highly variable between and within cancer classes, ranging from about 0.001 per megabase (Mb) to more than 400 per Mb (Fig. 1). Certain childhood cancers carried fewest mutations whereas cancers related to chronic mutagenic exposures such as lung (tobacco smoking) and malignant melanoma (exposure to ultraviolet light) exhibited the highest prevalence. This variation in mutation prevalence is attributable to differences between cancers in the duration of the cellular lineage between the fertilized egg and the sequenced cancer cell and/or to differences in somatic mutation rates during the whole or parts of that cellular lineage[1].

The landscape of mutational signatures

In principle, all classes of mutation (such as substitutions, indels, rearrangements) and any accessory mutation characteristic, for example, the sequence context of the mutation or the transcriptional strand on which it occurs, can be incorporated into the set of features by which a mutational signature is defined. In the first instance, we extracted mutational

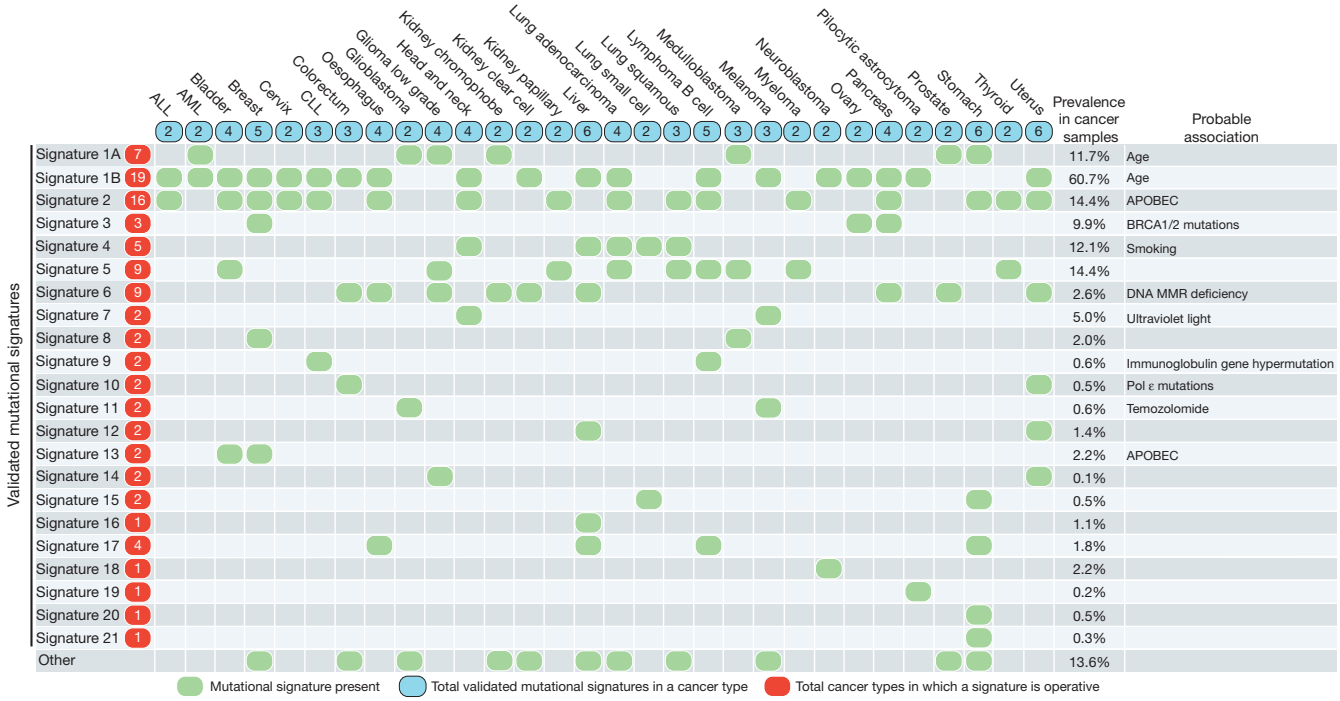

Figure 3 | The presence of mutational signatures across human cancer types. Cancer types are ordered alphabetically as columns whereas mutational signatures are displayed as rows. 'Other' indicates mutational signatures for which we were not able to perform validation or for which validation failed (Supplementary Figs 24–28). Prevalence in cancer samples indicates the percentage of samples from our data set of 7,042 cancers in which the signature contributed significant number of somatic mutations. For most signatures, significant number of mutations in a sample is defined as more than 100 substitutions or more than 25% of all mutations in that sample. MMR, mismatch repair.

signatures using base substitutions and additionally included information on the sequence context of each mutation. Because there are six classes of base substitution—C>A, C>G, C>T, T>A, T>C, T>G (all substitutions are referred to by the pyrimidine of the mutated Watson–Crick base pair)—and as we incorporated information on the bases immediately 5′ and 3′ to each mutated base, there are 96 possible mutations in this classification. This 96 substitution classification is particularly useful for distinguishing mutational signatures that cause the same substitutions but in different sequence contexts.

Applying this approach to the 30 cancer types revealed 21 distinct validated mutational signatures (Supplementary Table 1 and Supplementary Figs 2–28). These show substantial diversity (Fig. 2 and Supplementary Figs 2–23). There are signatures characterized by prominence of only one or two of the 96 possible substitution mutations, indicating remarkable specificity of mutation type and sequence context (signature 10). By contrast, others exhibit a more-or-less equal representation of all 96 mutations (signature 3). There are signatures characterized predominantly by C>T (signatures 1A/B, 6, 7, 11, 15, 19), C>A (4, 8, 18), T>C (5, 12, 16, 21) and T>G mutations (9, 17), with others showing distinctive combinations of mutation classes (2, 13, 14).

Signatures 1A and 1B were observed in 25 out of 30 cancer classes (Fig. 3). Both are characterized by prominence of C>T substitutions at NpCpG trinucleotides. Because they are almost mutually exclusive among tumour types they probably represent the same underlying process, with signature 1B representing less efficient separation from other signatures in some cancer types. Signature 1A/B is probably related to the relatively elevated rate of spontaneous deamination of 5-methyl-cytosine which results in C>T transitions and which predominantly occurs at NpCpG trinucleotides[9]. This mutational process operates in the germ line, where it has resulted in substantial depletion of NpCpG sequences, and in normal somatic cells[10].

Signature 2 is characterized primarily by C>T and C>G mutations at TpCpN trinucleotides and was found in 16 out of 30 cancer types (Fig. 3). On the basis of similarities in mutation type and sequence context we previously proposed that signature 2 is due to over activity of members of the APOBEC family of cytidine deaminases, which convert cytidine to uracil, coupled to activity of the base excision repair and DNA replication machineries[6,11].

In most cancer classes at least two mutational signatures were observed, with a maximum of six in cancers of the liver, uterus and stomach. Although these differences may, in part, be attributable to differences in the power to extract signatures, it seems likely that some cancers have a more complex repertoire of mutational processes than others.

Most individual cancer genomes exhibit more than one mutational signature and many different combinations of signatures were observed (Fig. 4 and Supplementary Figs 29–88). The patterns of contribution to individual cancer samples vary markedly between signatures. Signature 1A/B contributes relatively similar numbers of mutations to most cancer cases whereas other signatures contribute overwhelming numbers of mutations to some cancer samples but very few to others of the same cancer class, for example, signatures 2, 3, 4, 6, 7, 9, 10, 11, 13 (Fig. 4).

Mutational signatures and age of cancer diagnosis

We examined each cancer type for correlations between age of diagnosis and the number of mutations attributable to each signature in each sample. Signature 1A/B exhibited strong positive correlations with age in the majority of cancer types of childhood and adulthood (Supplementary Table 2). No other mutational signature showed a consistent correlation with age of diagnosis.

The mutations in a cancer genome may be acquired at any stage in the cellular lineage from the fertilized egg to the sequenced cancer cell. The correlation with age of diagnosis is consistent with the hypothesis that a substantial proportion of signature 1A/B mutations in cancer genomes have been acquired over the lifetime of the cancer patient, at a relatively constant rate that is similar in different people, probably in normal somatic tissues. The absence of consistent correlation of all

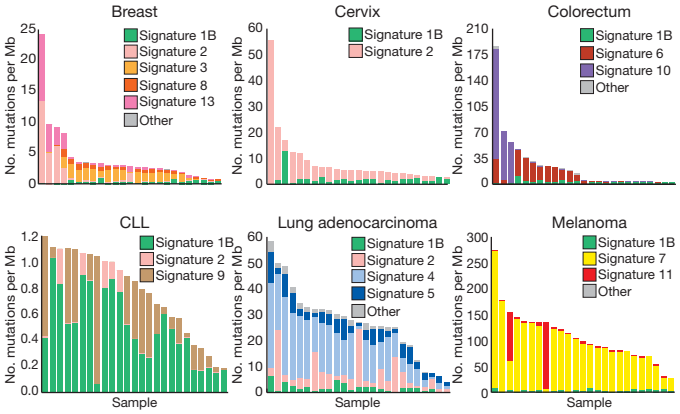

Figure 4 | The contributions of mutational signatures to individual cancers of selected cancer types. Each bar represents a typical selected sample from the respective cancer type and the vertical axis denotes the number of mutations per megabase. Contributions across all cancer samples could be found in Supplementary Figs 29–58. Summary of the total contributions for all operative mutational processes in a cancer type can be found in Supplementary Figs 59–88. 'Other' indicates mutational signatures for which we were not able to perform validation or for which validation failed (Supplementary Figs 24–28).

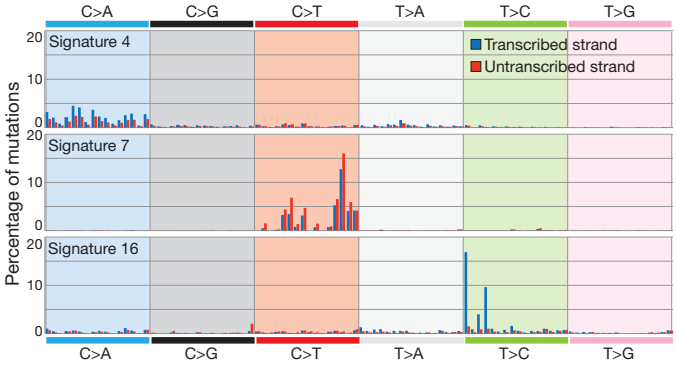

Figure 5 | Selected mutational signatures with strong transcriptional strand bias. Mutations are shown according to the 192 mutation classification incorporating the substitution type, the sequence context immediately 5′ and 3′ to the mutated base and whether the mutated pyrimidine is on the transcribed or untranscribed strand. The mutation types are displayed on the horizontal axis, whereas the vertical axis depicts the percentage of mutations attributed to a specific mutation type. A higher resolution version of all mutational signatures with strong transcriptional strand bias is found respectively in Supplementary Figs 89–95.

other signatures with age suggests that mutations associated with these have been generated at different rates in different people, possibly as a consequence of differing carcinogen exposures or after neoplastic change has been initiated.

Mutational signatures with transcriptional strand bias

The efficiency of DNA damage and DNA maintenance processes can differ between the transcribed and untranscribed strands of genes. The most well known cause of this phenomenon is transcription-coupled nucleotide excision repair (NER) that operates predominantly on the transcribed strand of genes and is recruited by RNA polymerase II when it encounters bulky DNA helix-distorting lesions[12].

We re-extracted substitution mutational signatures incorporating the transcriptional strand on which each mutation has taken place. Because a mutation in a transcribed genomic region may be either on the transcribed or the untranscribed strand, this generates a classification with 192 mutation subclasses.

Several signatures showed substantial differences in mutation prevalence between transcribed and untranscribed strands (known as transcriptional strand bias) (Fig. 5 and Supplementary Figs 89–95). For example, signature 4 shows transcriptional strand bias for C>A mutations (Fig. 5). Signature 4 is observed in lung adeno, squamous and small cell carcinomas, head and neck squamous, and liver cancers (Fig. 3), most of which are known to be caused by tobacco smoking. Therefore, signature 4 is probably an imprint of the bulky DNA adducts generated by polycyclic hydrocarbons found in tobacco smoke and their removal by transcription-coupled NER[13]. The higher prevalence of C>A mutations on transcribed compared to untranscribed strands is consistent with the propensity of many tobacco carcinogens to form adducts on guanine.

Similarly, signature 7, mainly found in malignant melanoma, shows a higher prevalence of C>T mutations on the untranscribed compared to the transcribed strands consistent with the formation, through ultraviolet exposure, of pyrimidine dimers and other lesions which are known to be repaired by transcription-coupled NER[14].

Beyond these known examples of DNA damage processed by transcription-coupled NER, other signatures show strong transcriptional strand bias (5, 8, 10, 12, 16). Notably, signature 16, which is characterized by T>C mutations at ApTpA, ApTpG and ApTpT trinucleotides and is observed in hepatocellular carcinomas, shows the strongest transcriptional strand bias of any signature, with T>C mutations occurring almost exclusively on the transcribed strand (Fig. 5). Similarly, signature 12, which features T>C mutations at NpTpN trinucleotides, also found in hepatocellular carcinomas, shows strong transcriptional strand bias with more T>C mutations on the transcribed than untranscribed strands (Supplementary Fig. 94). On the assumption that the transcriptional strand biases in signatures 12 and 16 are introduced by transcription-coupled NER, these currently unexplained signatures may be the result of bulky DNA helix-distorting adducts on adenine. However, there is no previous basis for invoking transcription-coupled NER in the genesis of these signatures and other causes of transcriptional strand bias may exist.

Mutational signatures with insertions and deletions

We re-extracted the mutational signatures including, in addition to the 96 substitution types, two further classes of mutation: indels at short nucleotide repeats and indels with overlapping microhomology at breakpoint junctions. Three of the 21 base substitution signatures associated with large numbers of indels. Signature 6, which is characterized predominantly by C>T at NpCpG mutations, but is distinct from signature 1A/B, contributes very large numbers of substitutions and small indels (mostly of 1 bp) at nucleotide repeats to subsets of colorectal, uterine, liver, kidney, prostate, oesophageal and pancreatic cancers. This pattern of indels, often termed 'microsatellite instability', is characteristic of cancers with defective DNA mismatch repair[15]. Consistent with this explanation, the presence of signature 6 was strongly associated with the inactivation of DNA mismatch repair genes in colorectal cancer ($P = 3.3 \times 10^{-5}$).

Signature 15 also contributes very large numbers of substitutions and small indels at nucleotide repeats but, compared to signature 6, exhibits greater prominence of C>T at GpCpN trinucleotides. Signature 15 was found in several samples of lung and stomach cancer and its origin is currently unknown.

By contrast, substantial numbers of larger deletions (up to 50 bp) with overlapping microhomology at breakpoint junctions were found in breast, ovarian and pancreatic cancer cases with major contributions from signature 3. A subset of cancer cases of these three classes is known to be due to inactivating mutations in *BRCA1* and *BRCA2*, and the presence of signature 3 was strongly associated with *BRCA1* and *BRCA2* mutations within the individual cancer types ($P = 1.6 \times 10^{-8}$ for breast cancer and $P = 0.02$ for pancreatic cancer)[6]. Indeed, almost all cases with *BRCA1* and *BRCA2* mutations showed a large contribution from signature 3. However, some cases with a substantial contribution from signature 3 did not have *BRCA1* and *BRCA2* mutations,

indicating that other mechanisms of *BRCA1* and *BRCA2* inactivation or abnormalities of other genes may also generate it.

BRCA1 and BRCA2 are implicated in homologous-recombination-based DNA double-strand break repair[16]. Abrogation of their functions results in non-homologous end-joining mechanisms, which can use microhomology at rearrangement junctions to rejoin double-strand breaks, taking over DNA double-strand break repair. The results show that, in addition to the genomic structural instability conferred by defective double-strand break repair, a base substitution mutational signature is associated with BRCA1 and BRCA2 deficiency.

Associating cancer aetiology and mutational signatures

Each mutational signature is the imprint left on the cancer genome by a mutational process that may include one or more DNA damage and/or DNA maintenance mechanisms, with the latter either functioning normally or abnormally. Here we consider likely mechanisms or underlying causes by comparing signatures with mutation patterns of known causation in the scientific literature or by associating them with epidemiological and biological features of particular cancer types.

Signature 1A/B is probably due to the endogenous mutational process present in most normal and neoplastic cells that is initiated by deamination of 5-methyl-cytosine[9]. Other signatures are probably attributable to exogenous mutagenic exposures. Signature 7 is observed in malignant melanoma and squamous carcinoma of the head and neck and has the known features of ultraviolet-light-induced mutations. Signature 4 is found in cancers associated with tobacco smoking (Fig. 3) and has the mutational features associated with tobacco carcinogens[13]. The causal relationship between tobacco smoking and signature 4 is supported by a strong positive association between smoking history and the contributions of signature 4 to individual cancers ($P = 1.1 \times 10^{-7}$, Supplementary Figs 44–46, 74–76 and 96).

Cigarette smoke contains over 60 carcinogens[13] and it is possible that this complex mixture may initiate other mutational processes. Signatures 1A/B, 2 and 5 were also found in lung adenocarcinoma. Signature 5, but not signatures 1A/B and 2, also showed a positive correlation between smoking history and mutation contribution ($P = 8.0 \times 10^{-3}$, Supplementary Fig. 96). Thus, in lung cancer, signature 5, which is characterized predominantly by C>T and T>C mutations, may also be due to tobacco carcinogens. However, it is also present in nine other cancer types, most of which are not strongly associated with tobacco consumption, and therefore its aetiology overall is unclear (Fig. 3).

Some anticancer drugs are mutagens[17]. Signature 11 is found in malignant melanomas and glioblastoma multiforme pretreated with the alkylating agent temozolomide ($P = 4.0 \times 10^{-3}$) and has mutational features very similar to those previously reported in experimental studies of alkylating agents[18].

Abnormalities in DNA maintenance may also be responsible for mutational signatures, and the roles of defective DNA mismatch repair (signature 6) and defective homologous-recombination-based DNA double-strand break repair (signature 3) have been discussed above. Other signatures may result from abnormal activity of enzymes that modify DNA or of error-prone polymerases. Signatures 2 and 13 have been attributed to the AID/APOBEC family of cytidine deaminases[6]. On the basis of similarities in the sequence context of cytosine mutations caused by APOBEC enzymes in experimental systems, a role for APOBEC1, APOBEC3A and/or APOBEC3B in human cancer seems more likely than for other members of the family[19–21]. However, the reason for the extreme activation of this mutational process in some cancers is unknown. Because APOBEC activation constitutes part of the innate immune response to viruses and retrotransposons[22] it may be that these mutational signatures represent collateral damage on the human genome from a response originally directed at retrotransposing DNA elements or exogenous viruses. Confirmation of this hypothesis would establish an important new mechanism for initiation of human carcinogenesis.

Signature 9, observed in chronic lymphocytic leukaemia and malignant B-cell lymphomas, is characterized by T>G transversions at ApTpN and TpTpN trinucleotides, and is restricted to cancers that have undergone somatic immunoglobulin gene hypermutation (IGHV-mutated) associated with AID ($P = 2.5 \times 10^{-4}$ in chronic lymphoid leukaemia (CLL)). Signature 9 does not, however, have the known mutational features of AID[20], and has been proposed to be due to polymerase η, an error-prone polymerase involved in processing AID-induced cytidine deamination[11,23]. Similarly, signature 10, which generates huge numbers of mutations in subsets of colorectal and uterine cancer, has been previously associated with altered activity of the error-prone polymerase Pol ε consequent on mutations in the gene[24,25].

Many mutational signatures do not, however, have an established or proposed underlying mutational process or aetiology. Some, for example signatures 8, 12 and 16, show strong transcriptional strand bias (Fig. 5) and possibly reflect the involvement of transcription-coupled nucleotide excision repair acting on bulky DNA adducts due to exogenous carcinogens. Others, for example signatures 14, 15 and 21, show overwhelming activity in a small number of cancer cases (Supplementary Figs 38, 45 and 56, respectively) and are perhaps more likely to be due to currently uncharacterized defects in DNA maintenance.

Localized hypermutation

Foci of localized substitution hypermutation, termed kataegis after the Greek for thunderstorm, were recently described in breast cancer[6]. Kataegis is characterized by clusters of C>T and/or C>G mutations which are substantially enriched at TpCpN trinucleotides and on the same DNA strand. Foci of kataegis include from a few to several thousand mutations and are often found in the vicinity of genomic rearrangements. The genomic regions affected are different in different cancers. On the basis of the substitution types and sequence context of kataegis substitutions, an underlying role for APOBEC family enzymes was proposed for kataegis as well as for signatures 2 and 13 (ref. 6).

The 507 whole-cancer genome mutation catalogues were searched for clusters of mutations. Cancers of breast (67 of 119), pancreas (11 of 15), lung (20 of 24), liver (15 of 88), medulloblastomas (2 of 100), CLL (15 of 28), B-cell lymphomas (21 of 24) and acute lymphoblastic leukaemia (1 of 1) showed occasional (<10), small (<20 mutations) foci of kataegis, whereas acute myeloid leukaemia (0 of 7) and pilocytic astrocytoma (0 of 101) did not. Subsets of breast (7), lung (6) and haematological cancers (3) showed numerous (>10) kataegic foci and two breast and one pancreatic cancer showed major foci of kataegis (>50 mutations) (Fig. 6 and Supplementary Figs 97 and 98).

Kataegic foci are often associated with genomic rearrangements (Supplementary Fig. 98). In yeast, introduction of a DNA double-strand break greatly increases the likelihood of kataegis in its vicinity, indicating a role for such breaks in initiating the process[20]. However, even in cancer cases with kataegis, most rearrangements do not exhibit nearby kataegis, indicating that a double-strand break is not sufficient.

In neoplasms of B-lymphocyte origin, including CLL and many lymphomas, mutation clusters recurrently occurred at immunoglobulin loci. In these cancers the mutation characteristics were different (Supplementary Fig. 98), bearing the hallmarks of somatic hypermutation associated with AID, which is operative during the generation of immunological diversity[20].

Discussion

The diversity and complexity of somatic mutational processes underlying carcinogenesis in human beings is now being revealed through mutational patterns buried within cancer genomes. It is likely that more mutational signatures will be extracted, together with more precise definition of their features, as the number of whole-genome sequenced cancers increases and analytical methods are further refined.

The mechanistic basis of some signatures is, at least partially, understood but for many it remains speculative or unknown. Elucidating the

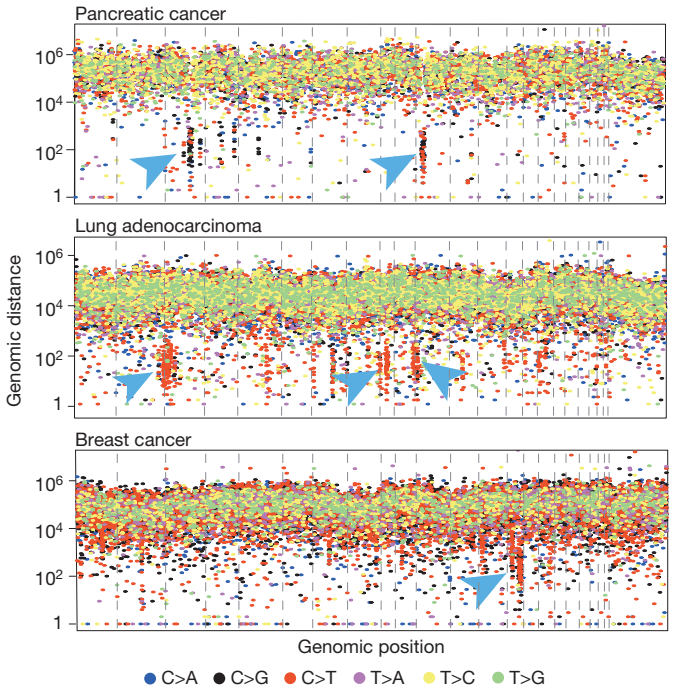

Figure 6 | Kataegis in three cancers. Each of these 'rainfall' plots represents an individual cancer sample in which each dot represents a single somatic mutation ordered on the horizontal axis according to its position in the human genome. The vertical axis denotes the genomic distance of each mutation from the previous mutation. Arrowheads indicate clusters of mutations in kataegis.

underlying mutational processes will depend upon two major streams of investigation. First, compilation of mutational signatures from model systems exposed to known mutagens or perturbations of the DNA maintenance machinery and comparison with those found in human cancers. Second, correlation of the contributions of mutational signatures with other biological characteristics of each cancer through diverse approaches ranging from molecular profiling to epidemiology. Collectively, these studies will advance our understanding of cancer aetiology with potential implications for prevention and treatment.

METHODS SUMMARY

Mutational catalogues were stringently filtered and our previously developed computational framework[5,6] was used to extract mutational signatures from them. The computational framework for deciphering mutational signatures and all mutational catalogues are freely available for download from http://www.mathworks.com/matlabcentral/fileexchange/38724, whereas the complete set of somatic mutations is available from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl. All presented mutational signatures were validated. Kataegis was detected using an algorithm based on piecewise constant fitting.

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
2. Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med.* **2,** 54 (2010).
3. Peña-Diaz, J. *et al.* Noncanonical mismatch repair as a source of genomic instability in human cells. *Mol. Cell* **47,** 669–680 (2012).
4. Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* **2,** a001008 (2010).
5. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3,** 246–259 (2013).
6. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979–993 (2012).
7. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
8. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
9. Pfeifer, G. P. Mutagenesis at methylated CpG sequences. *Curr. Top. Microbiol. Immunol.* **301,** 259–281 (2006).
10. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).
11. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76,** 1–22 (2007).
12. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nature Rev. Mol. Cell Biol.* **9,** 958–970 (2008).
13. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21,** 7435–7451 (2002).
14. Pfeifer, G. P., You, Y. H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutat. Res.* **571,** 19–31 (2005).
15. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138,** 2073–2087 (2010).
16. Thompson, L. H. Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: the molecular choreography. *Mutat. Res.* **751,** 158–246 (2012).
17. 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).
18. Tomita-Mitchell, A. *et al.* Mismatch repair deficient human cells: spontaneous and MNNG-induced mutational spectra in the HPRT gene. *Mutat. Res.* **450,** 125–138 (2000).
19. Taylor, B. J. M. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* e00534 (2013).
20. Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494,** 366–370 (2013).
21. Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* **10,** 1247–1253 (2002).
22. Koito, A. & Ikeda, T. Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. *Front. Microbiol.* **4,** 28 (2013).
23. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
24. The Cancer Genome Atlas Network.. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
25. Cancer Genome Atlas Research. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
26. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

Acknowledgements We would like to thank the Wellcome Trust for support (grant reference 098051) together with many other funding bodies and individuals (Supplementary Note 1).

Author Contributions L.B.A., S.N.-Z. and M.R.S. conceptualized the study and analysed the mutational signatures and kataegis data. L.B.A. performed data curation, data filtering and mutational signature extraction. S.N.-Z. and D.C.W. performed kataegis identification. S.N.-Z. performed visual validation. A.P.B., K.R., J.W.T. and D.J. provided bioinformatics support for mutational signature and kataegis analysis. S.A.J.R.A., S.Be., A.V.B., G.R.B., N.B., A.B., A.-L.B.-D., S.Bo., B.B., C.C., H.R.D., C.D., R.E., J.E.E., J.A.F., M.G., F.H., B.H., T.I., S.I., M.I., N.J., D.T.W.J., S.K., M.K., S.R.L., C.L.-O., S.M., N.C.M., H.N., P.A.N., M.P., E.P., A.P., J.V.P., X.S.P., M.R., A.L.R., J.R., P.R., M.S., T.N.S., P.N.S., Y.T., A.N.J.T., R.V.-M., M.M.v.B., L.V.V., A.V.-S., N.W., L.R.Y., J.Z.-R., P.A.F., U.M., P.L., M.M., S.M.G., R.S., E.C., T.S., S.M.P. and P.J.C. contributed samples, clinical data and scientific advice. M.R.S. and L.B.A. wrote the manuscript. M.R.S. directed the overall research.

Author Information The computational framework for deciphering mutational signatures and all mutational catalogues are freely available for download from http://www.mathworks.com/matlabcentral/fileexchange/38724, whereas the complete set of somatic mutations is available from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R.S. (mrs@sanger.ac.uk).

Ludmil B. Alexandrov[1], Serena Nik-Zainal[1,2], David C. Wedge[1], Samuel A. J. R. Aparicio[3,4,5], Sam Behjati[1,6], Andrew V. Biankin[7,8,9,10,11], Graham R. Bignell[1], Niccolò Bolli[1,12,13], Ake Borg[14], Anne-Lise Børresen-Dale[15,16], Sandrine Boyault[17], Birgit Burkhardt[18,19], Adam P. Butler[1], Carlos Caldas[20], Helen R. Davies[1], Christine Desmedt[21], Roland Eils[22], Jórunn Erla Eyfjörd[23], John A. Foekens[24], Mel Greaves[25], Fumie Hosoda[26], Barbara Hutter[22], Tomislav Ilicic[1], Sandrine Imbeaud[27,28], Marcin Imielinski[29], Natalie Jäger[22], David T. W. Jones[30], David Jones[1], Stian Knappskog[31,32], Marcel Kool[30], Sunil R. Lakhani[33], Carlos López-Otín[34], Sancha Martin[1], Nikhil C. Munshi[35,36], Hiromi Nakamura[26], Paul A. Northcott[30], Marina Pajic[7], Elli Papaemmanuil[1], Angelo Paradiso[37], John V. Pearson[38], Xose S. Puente[34], Keiran Raine[1], Manasa Ramakrishna[1], Andrea L. Richardson[39,40,41], Julia Richter[42], Philip Rosenstiel[43], Matthias Schlesner[22], Ton N. Schumacher[44], Paul N. Span[45], Jon W.

Teague[1], Yasushi Totoki[26], Andrew N. J. Tutt[46], Rafael Valdés-Mas[34], Marit M. van Buuren[44], Laura van 't Veer[47], Anne Vincent-Salomon[48], Nicola Waddell[38], Lucy R. Yates[1], Australian Pancreatic Cancer Genome Initiative*, ICGC Breast Cancer Consortium*, ICGC MMML-Seq Consortium*, ICGC PedBrain*, Jessica Zucman-Rossi[27,28], P. Andrew Futreal[1], Ultan McDermott[1], Peter Lichter[49], Matthew Meyerson[29,39,40], Sean M. Grimmond[38], Reiner Siebert[42], Elías Campo[50], Tatsuhiro Shibata[26], Stefan M. Pfister[30,51], Peter J. Campbell[1,12,13] & Michael R. Stratton[1]

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [2]Department of Medical Genetics, Box 134, Addenbrooke's Hospital NHS Trust, Hills Road, Cambridge CB2 0QQ, UK. [3]Molecular Oncology, Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver V5Z 1L3, Canada. [4]Centre for Translational and Applied Genomics, Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver V5Z 1L3, Canada. [5]Department of Pathology, University of British Columbia, G227-2211 Wesbrook Mall, British Columbia, Vancouver V6T 2B5, Canada. [6]Department of Paediatrics, University of Cambridge, Hills Road, Cambridge CB2 2XY, UK. [7]Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD, UK. [8]West of Scotland Pancreatic Unit, Glasgow Royal Infirmary, Glasgow G4 0SF, UK. [9]The Kinghorn Cancer Centre, 370 Victoria Street, Darlinghurst, and the Cancer Research Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, New South Wales 2010, Australia. [10]Department of Surgery, Bankstown Hospital, Eldridge Road, Bankstown, Sydney, New South Wales 2200, Australia. [11]South Western Sydney Clinical School, Faculty of Medicine, University of New South Wales, Liverpool, New South Wales 2170, Australia. [12]Department of Haematology, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [13]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK. [14]Department of Oncology, Lund University, SE-221 85 Lund, Sweden. [15]Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway. [16]The K.G. Jebsen Center for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, N-0310 Oslo, Norway. [17]Plateforme de Bioinformatique Synergie Lyon Cancer, Centre Léon Bérard, 28 rue Laennec, 69373 Lyon Cedex 08, France. [18]NHL-BFM Study Center and Department of Pediatric Hematology and Oncology, University Children's Hospital, 48149 Münster, Germany. [19]NHL-BFM Study Center and Department of Pediatric Hematology and Oncology, University Children's Hospital, 35392 Giessen, Germany. [20]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, UK. [21]Breast Cancer Translational Res Lab -BCTL, Université Libre de Bruxelles—Institut Jules Bordet, Boulevard de Waterloo, 125, B-1000 Brussels, Belgium. [22]Department of Theoretical Bioinformatics (B080), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. [23]Cancer Research Laboratory, Faculty of Medicine, Biomedical Centre, University of Iceland, 101 Reykjavik, Iceland. [24]Department of Medical Oncology, Erasmus MC Cancer Institute, 3015 CE Rotterdam, The Netherlands. [25]Department of Haemato-oncology, Institute of Cancer Research, London SM2 5NG, UK. [26]Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Tokyo 104-0045, Japan. [27]INSERM, UMR-674, Génomique Fonctionnelle des Tumeurs Solides, Institut Universitaire d'Hematologie (IUH), 75475 Paris, France. [28]Université Paris Descartes, Labex Immuno-oncology, Sorbonne Paris Cité, Faculté de Médecine, 75006 Paris, France. [29]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA. [30]Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [31]Section of Oncology, Department of Clinical Science, University of Bergen, 5020 Bergen, Norway. [32]Department of Oncology, Haukeland University Hospital, 5021 Bergen, Norway. [33]The University of Queensland Centre for Clinical Research, School of Medicine and Pathology Queensland, The Royal Brisbane & Women's Hospital, Herston 4029, Brisbane, Queensland, Australia. [34]Departamento Bioquímica y Biología Molecular, IUOPA-Universidad de Oviedo, 33006 Oviedo, Spain. [35]Jerome Lipper Multiple Myeloma Disease Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02215, USA. [36]Boston Veterans Administration Healthcare System, West Roxbury, Massachusetts 02132, USA. [37]Clinical Experimental Oncology Laboratory, National Cancer Institute, Via Amendola, 209, 70126 Bari, Italy. [38]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, Brisbane, Queensland 4072, Australia. [39]Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. [40]Harvard Medical School, Boston, Massachusetts 02115, USA. [41]Department of Pathology, Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. [42]Institute of Human Genetics, Christian-Albrechts-University, 24118 Kiel, Germany. [43]Institute of Clinical Molecular Biology, Christian-Albrechts-University, 24118 Kiel, Germany. [44]Division of Immunology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [45]Department of Radiation Oncology and department of Laboratory Medicine, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500HB Nijmegen, The Netherlands. [46]Breakthrough Breast Cancer Research Unit, King's College London School of Medicine, London SW3 6JB, UK. [47]The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands. [48]Institut Curie, Departement de Pathologie, INSERM U830, 26 rue d'Ulm, 75248 Paris Cedex 05, France. [49]Division of Molecular Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [50]Unidad de Hematopatología, Servicio de Anatomía Patológica, Hospital Clínic, Universitat de Barcelona, IDIBAPS, 08036 Barcelona, Spain. [51]Department of Pediatric Hematology and Oncology, 69120 Heidelberg, Germany.

*A list of authors and affiliations appears in the Supplementary Information.

METHODS

Validating mutational signatures. Validating a mutational signature requires ensuring that a large set of somatic mutations attributed to this signature is genuine in at least one sample. Validation is complicated as multiple mutational processes are usually operative in most cancer samples, and thus every individual somatic mutation can be probabilistically assigned to several mutational signatures. To overcome this limitation, we examined our data set for samples that are predominantly generated by one mutational signature (that is, more than 50% of the somatic mutations in the sample belong to an individual mutational signature) and/or for samples in which all operative mutational processes have mutually exclusive patterns of mutations (for example, a sample with mutations only from signature 1B, which is predominantly C>T substitutions, and signature 18, which is predominantly C>A substitutions). We identified the optimal available sample for every mutational signature and attempted to validate the subset of somatic mutations attributed to this signature using one of three methods (Supplementary Fig. 99): (1) validation through re-sequencing with an orthogonal sequencing technology; (2) validation through re-sequencing with the same sequencing technology (including RNA-seq, bisulphite sequencing, etc.); (3) validation through visual examination of somatic mutations by an experienced curator using a genomic browser and BAM files for both the tumour and its matched normal.

For some of the previously published samples, we used the already reported validation data. When possible, somatic mutations were validated by either re-sequencing with orthogonal technology or re-sequencing using the same sequencing technology. We resorted to visual validation only when there was no other possibility for validating a mutational signature. 22 out of the 27 originally identified mutational signatures were validated (Supplementary Table 1 and Supplementary Fig. 99). Three mutational signatures failed validation: signatures R1 to R3 (Supplementary Figs 24 to 26). We were unable to validate two mutational signatures: signatures U1 and U2 (Supplementary Figs 27 and 28), due to lack of available biological samples and access to BAM files for the samples with sufficient number of somatic mutations generated by these two mutational signatures.

Samples and curation of freely available cancer data. Informed consent was obtained from all subjects. Collection and use of patient samples were approved by the appropriate Internal Review Board of each institution. In addition to newly generated data, we curated freely available somatic mutations from three other sources: (1) the data portal of The Cancer Genome Atlas (TCGA); (2) the data portal of the International Cancer Genome Consortium (ICGC); (3) previously published data in peer-review journals, see additional references[6,23,27–59].

Filtering, estimating mutation prevalence and generating mutational catalogues. In all examined samples, normal DNA from the same individuals had been sequenced to establish the somatic origin of variants. Extensive filtering was performed to remove any residual germline mutations and technology-specific sequencing artefacts before analysing the data. Germline mutations were filtered out from the lists of reported mutations using the complete list of germline mutations from dbSNP[60], 1000 genomes project[61], NHLBI GO Exome Sequencing Project[62], and 69 Complete Genomics panel (http://www.completegenomics.com/public-data/69-Genomes/). Technology-specific sequencing artefacts were filtered out by using panels of BAM files of (unmatched) normal tissues containing more than 120 normal genomes and 500 normal exomes. Any somatic mutation present in at least three well-mapping reads in at least two normal BAM files was discarded. The remaining somatic mutations were used for generating a mutational catalogue for every sample.

Prevalence of somatic mutations was estimated on the basis of a haploid human genome after all filtering. Prevalence of somatic mutations in exomes was calculated based on the identified mutations in protein-coding genes and assuming that an average exome has 30 Mb in protein-coding genes with sufficient coverage. Prevalence of somatic mutations in whole genomes was calculated based on all identified mutations and assuming that an average whole genome has 2.8 gigabases with sufficient coverage.

The immediate 5′ and 3′ sequence context was extracted using the ENSEMBL Core programing interfaces for human genome build GRCh37. Curated somatic mutations that originally mapped to an older version of the human genome were re-mapped using UCSC's freely available lift genome annotations tool (any somatic mutations with ambiguous or missing mappings were discarded). Dinucleotide substitutions were identified when two substitutions were present in consecutive bases on the same chromosome (sequence context was ignored). The immediate 5′ and 3′ sequence content of all indels was examined and the ones present at mono/polynucleotide repeats or microhomologies were included in the analysed mutational catalogues as their respective types. Strand bias catalogues were derived for each sample using only substitutions identified in the transcribed regions of well-annotated protein-coding genes. Genomic regions of bidirectional transcription were excluded from the strand bias analysis.

Deciphering signatures of mutational processes. Mutational signatures were deciphered independently for each of the 30 cancer types using our previously developed computational framework[5]. The algorithm deciphers the minimal set of mutational signatures that optimally explains the proportion of each mutation type found in each catalogue and then estimates the contribution of each signature to each catalogue. Mutational signatures were also extracted separately for genomes and exomes. Mutational signatures extracted from exomes were normalized using the observed trinucleotide frequency in the human exome to the one of the human genome. All mutational signatures were clustered using unsupervised agglomerative hierarchical clustering and a threshold was selected to identify the set of consensus mutational signatures. Mis-clustering was avoided by manual examination (and whenever necessary re-assignment) of all signatures in all clusters. 27 consensus mutational signatures were identified across the 30 cancer types. The computational framework for deciphering mutational signatures as well as the data used in this study are freely available and can be downloaded from http://www.mathworks.com/matlabcentral/fileexchange/38724, whereas the complete set of somatic mutations is available from ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl.

Factors that influence extraction of mutational signatures. Recently, using simulated and real data, we described in detail the factors that influence the extraction of mutational signatures[5]. These included the number of available samples, the mutation prevalence in samples, the number of mutations contributed by different mutational signatures, the similarity between the signatures of mutational processes operative in cancer samples, as well as the limitations of our computational approach. Here, we examined data sets with varying sizes from 30 different cancer types and we have taken great care to report only validated mutational signatures. However, our approach identified two similar patterns most likely representing the same biological process; that is, signature 1A and 1B. The reasons for this is, for some cancer types we have sufficient numbers of samples and/or mutations (that is, statistical power) to decipher the cleaner version (that is, signature 1A), whereas for other cancer types we do not have sufficient data and our approach extracts a version of the signature which is more contaminated by other signatures present in that cancer type (that is, signature 1B). Nevertheless, the two signatures are very similar; hence we call them 1A and 1B. Being almost mutually exclusive among cancer types (that is, finding either signature 1A or 1B in each cancer type but not usually both) is supportive of the notion that they represent the same underlying process as is the fact that signatures 1A and 1B both correlate with age and have the same overall pattern of contributions to individual cancer genomes. Indeed, in our view it is likely that if we had sufficient data, signature 1B would disappear and the algorithm would extract only signature 1A.

Displaying mutational signatures. Mutational signatures are displayed using a 96 substitution classification defined by the substitution class and the sequence context immediately 3′ and 5′ to the mutated base. Mutational signatures are displayed in the main text of the report and in Supplementary Information on the basis of the observed trinucleotide frequency of the human genome; that is, representing the relative proportions of mutations generated in each signature based on the actual trinucleotide frequencies of the reference human genome. However, in Supplementary Information we also provide a visualization of mutational signatures based on an equal frequency of each trinucleotide (Supplementary Figs 2–28). The equal trinucleotide frequency representation results, in all mutational signatures, in a greater degree of prominence of C>T substitutions at NpCpG trinucleotides as major features compared to the plots based on the observed trinucleotides. This difference may in some cases reflect the biological reality, that is, a propensity of the particular mutational process to be more active at NpCpG trinucleotides. However, note that it may also in some cases be due to incomplete extraction by the algorithm of the signature in question from signature 1A/B, which is characterized by prominent features at NpCpG trinucleotides. This is likely to happen because (1) signature 1A/B is ubiquitous and (2) because even a small probability of mutations at NpCpG trinucleotides will generate a prominent feature because of the severe depletion of NpCpG trinucleotides in the reference genome. In future, with larger numbers of sequences and large numbers of whole-genome sequences it is anticipated that the latter effect will be reduced.

Approaches for associating cancer aetiology and exposures of validated mutational signatures. Generalized linear models (GLMs) were used to fit signature exposures (that is, number of mutations assigned to a signature) and age of cancer diagnoses. For each cancer type, all mutational signatures operative in it were evaluated using GLMs and the *P* values were corrected for multiple hypothesis testing using the Benjamini–Hochberg false discovery rate procedure. The resulting *P* values indicate that age strongly correlates with signature 1A/B across 15 cancer types (Supplementary Table 2). Exposure to signature 4 also correlates with age of diagnosis in kidney papillary and thyroid cancers. However, in both cancer types, we were not able to detect/extract signature 1A/B due to a low number of mutations in their samples and it is likely that signature 1A/B is

currently mixed within signature 4. Further studies involving whole-genome sequences will be needed to validate this hypothesis. Notably, in melanoma, age of diagnosis also correlates with exposure to signature 7, which we have associated with exposure to ultraviolet light.

Associations between all other aetiologies and signature exposures were performed using two-sample Kolmogorov–Smirnov tests between two sets of samples. The first set contains the signature exposures of the samples with the 'desired feature' (for example, samples that contain a hypermutation in the immunoglobulin gene) and the second set is the signature exposures of the samples without the 'desired feature' (for example, samples that do not contain a hypermutation in the immunoglobulin gene). Samples with unknown feature status (for example, not knowing the status of the immunoglobulin gene) were ignored. Kolmogorov–Smirnov tests were performed for all signatures and all examined 'features' in a cancer type. *P* values were corrected for multiple hypothesis testing using the Benjamini–Hochberg false discovery rate procedure and based on the performed tests in a particular cancer class.

A piecewise-constant-fitting-based algorithm for the detection of kataegis. Foci of localized hypermutation, termed kataegis, were sought in 507 whole-genome sequenced cancers. High-quality variant calls that had been previously subjected to filtering for mutational signature analysis were investigated using an algorithm developed to identify foci of kataegis.

For each sample, all mutations were ordered by chromosomal position and the intermutation distance, defined as the number of base pairs from each mutation to the next one, was calculated. Intermutation distances were then segmented using the piecewise constant fitting (PCF) method[63] to find regions of constant intermutation distance. Parameters used for PCF were $\gamma = 25$ and $k_{\min} = 2$ and were trained on the set of kataegis foci that had been manually identified, curated and validated using orthogonal sequencing platforms[6]. Putative regions of kataegis were identified as those segments containing six or more consecutive mutations with an average intermutation distance of less than or equal to 1,000 bp.

Variation in number of foci of kataegis and relationship with genome-wide mutation burden. To examine the likelihood of kataegis occurring for different mutation burdens, the expected number of kataegis events that would be observed by chance was calculated for a range of total number of mutations per cancer, *n*, between 1,000 and 2,000,000. The probability that any one mutation will be followed by five other mutations within a distance of 5,000 bp, thereby triggering the identification of kataegis, is given by $p = P(\mathrm{Pois}(5{,}000n/g) \geq 5)$, where g is the length of the genome, in base pairs.

Supplementary Fig. 97 shows the expected number of kataegis events identified in genomes with between 100,000 and 500,000 mutations. For cancers with up to 200,000 mutations, the expected number of kataegis events is extremely small (0.16 for a total mutation load of 200,000), making the detection of kataegic foci highly significant for each sample. Supplementary Table 3 presents all the samples in which kataegic foci were identified, the total mutation burden for each sample, the observed number of kataegic foci, and the expected number of foci.

Specificity of variants in kataegis foci. Clusters of variant calls can easily occur in regions of low sequence complexity. These are not true substitution mutations but represent systematic sequencing artefacts or mis-mapping of short reads. The quality of variant calls depends on the quality of mutation-calling by individual institutions. Additional filtering was applied to remove likely false-positive calls and then putative kataegic foci were individually curated.

1,436 kataegis foci were called by PCF, with 873 finalized as putative kataegis foci (Supplementary Table 4) involving 9,219 substitution variants. Where possible, BAM files were retrieved, inspected and substitution variants involved in kataegis foci were manually curated to remove likely false-positive calls. Where BAM files were not available to us, substitution variants were strictly excluded if called in: (1) genomic features that generate mapping errors, for example, regions of excessively high coverage due to collapsed repeat sequences in the reference genome[64]; (2) highly repetitive regions with reads consistently demonstrating low mapping qualities in 20 unrelated normal samples; (3) locations with known germline insertions/deletions within the sequencing reads reporting the mutated base.

Several features were seen in the finalized putative kataegis foci, which reinforced the conviction in the validity of these calls. Although clusters of mutations identified by the PCF method were sought in an approach unbiased by mutation type and based exclusively on intermutation distances, we find that the 873 putative foci demonstrate: first, a preponderance to C>T and C>G mutations (Supplementary Fig. 97b); second, the enrichment for a TpC sequence context as previously described[6] (Supplementary Fig. 97b); third, processivity (where consecutive mutations within a cluster were on the same strand; that is, 6 C>T mutations in a row or 6 G>A mutations in a row; Fig. 6c); and fourth, visual curation of reads carrying these processive variants showed that the variants were usually in *cis* (that is, mutations were on the same read (Supplementary Fig. 97c) or on the read mate of other affected alleles within the insert size) with respect to each other, indicating that they had arisen on the same allele. Finally, where data were available, we found that clusters of substitution mutations within the same kataegis foci shared approximately the same variant allele fraction, indicating that they had probably arisen during a single cell cycle event.

BAM files from some samples were not accessible and therefore a proportion of substitution variants involved in kataegis foci were not visually curated. The application of the strict criteria described above and the subsequent finding of the consistency of the mutation-type, sequence context, processive nature of the mutations, with the majority in *cis* on individual sequencing reads, indicates that the vast majority of these foci are probably genuine. However, the possibility that some of the foci are not truly kataegis, particularly for the cancers which have not been validated or visually curated, remains.

Sensitivity of kataegis detection. It is acknowledged that the likelihood of detection of kataegis foci rests on the sensitivity of mutation detection. It is possible for foci to be missed because the mutations were not detected by mutation callers of the various institutions, before our analysis. This is particularly relevant for subclonal mutations bearing a low variant allele fraction or for mutations that occur on a single copy of a multi-copy locus. This is because the likelihood of mutation detection is reduced when uncorrected for copy number and for aberrant cell fraction of the tumour sample. Furthermore, our stringent post-processing criteria, particularly of samples that have not been visually curated, make it more likely that kataegis is under-represented in this analysis.

Relationship between kataegis and large-scale genomic changes. Reinforcing our previous findings[6], we found that some kataegic foci were very closely associated with rearrangements. For example, a breast cancer sample with 1,534 point mutations had only one focus of kataegis which contained 32 point mutations. The same breast cancer sample also had 25 large-scale genomic structural variations scattered throughout the genome. However, one tandem duplication coincided with this single locus of kataegis in this cancer. Notably, no other mutations or structural variations were seen for 2 Mb flanking this extraordinary event (Supplementary Fig. 97b). Another breast cancer (Fig. 6) that contained 22,454 mutations and had 292 rearrangements altogether, had nine regions of kataegis, five of which coincided with large-scale structural variations, underscoring the co-localization of kataegis foci with structural variations. This also highlights that not all foci of kataegis co-localized with structural variations and not all structural variations were associated with kataegis.

Sites of amplification represent a potential source of false variant calls. If the amplification occurred early in the evolution of a cancer, then there is an increased likelihood of substitutions accumulating randomly within the amplified genomic region. When mapped back to the reference genome, these will appear as clustered variants.

A number of features allow us to distinguish such events from 'true' kataegis. These mutations would not be expected to have features associated with kataegis, such as the mutation type, predilection for a TpC sequence context and the processivity. Furthermore, if they have accumulated as random events in a multi-copy locus, then they would be less likely to occur in *cis* (on the same sequencing read) with respect to each other. In contrast, mutations which have occurred at the same time, during one moment of transient hypermutability in a single cell cycle event, would be expected to cluster on one copy of a multi-copy locus, to be in *cis* and to demonstrate approximately the same variant allele fraction. Finally, to achieve the level of hypermutation required to be called as a focus of kataegis (average intermutation distance of less than 1,000 bp for six consecutive mutations equivalent to ~1,000 substitutions per Mb), the degree of copy number amplification would have to be considerable.

To examine this likelihood of false calls in regions of amplification, simulations were performed assuming background mutation rates of 10 per Mb, 40 per Mb and 100 per Mb for different copy number states and for different sizes of focal amplification. The expected number of kataegic foci for these different states are provided in Supplementary Table 5. For most of the samples in which kataegis was detected (all but twenty), a 10 Mb region of amplification would require a copy number state of 36 or above to generate 1 cluster of 6 mutations with an average intermutation distance of less than 1,000 bp. For 19 of the remaining 20 samples, a 10 Mb region of amplification would require a copy number state of 10 or above. For the single cancer with a mutation rate exceeding 40 per Mb, a copy number state of 4 is required to generate a cluster of mutations. As mentioned previously, these clusters would have to be processive, be in *cis* and have roughly the same variant allele fraction to be called as a focus of kataegis.

Definition of kataegis. Kataegis has been identified via a PCF-based method as 6 or more consecutive mutations with an average intermutation distance of less than or equal to 1,000 bp. Other salient features include a preponderance for C>T and C>G mutations, a predilection for a TpC mutation context, processivity, evidence of having arisen on the same parental allele (being in *cis*) on sequencing reads and additionally (but not necessarily) co-localization with large-scale genomic structural variation.

27. Holmfeldt, L. *et al.* The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nature Genet.* **45,** 242–252 (2013).
28. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481,** 157–163 (2012).
29. De Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nature Genet.* **45,** 186–190 (2013).
30. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481,** 506–510 (2012).
31. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).
32. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genet.* **44,** 47–52 (2012).
33. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488,** 660–664 (2012).
34. Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genet.* **45,** 478–486 (2013).
35. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333,** 1154–1157 (2011).
36. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333,** 1157–1160 (2011).
37. Guo, G. *et al.* Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nature Genet.* **44,** 17–19 (2012).
38. Peña-Llopis, S. *et al.* BAP1 loss defines a new class of renal cell carcinoma. *Nature Genet.* **44,** 751–759 (2012).
39. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
40. Seo, J. S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.* **22,** 2109–2119 (2012).
41. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).
42. Love, C. *et al.* The genetic landscape of mutations in Burkitt lymphoma. *Nature Genet.* **44,** 1321–1325 (2012).
43. Zhang, J. *et al.* Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nature Genet.* **45,** 602–612 (2013).
44. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
45. Jiao, Y. *et al.* DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **331,** 1199–1203 (2011).
46. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nature Genet.* **45,** 279–284 (2013).
47. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330,** 228–231 (2010).
48. Wu, J. *et al.* Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. *Proc. Natl Acad. Sci. USA* **108,** 21188–21193 (2011).
49. Sausen, M. *et al.* Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nature Genet.* **45,** 12–17 (2013).
50. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).
51. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487,** 239–243 (2012).
52. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature Genet.* **44,** 685–689 (2012).
53. Rudin, C. M. *et al.* Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nature Genet.* **44,** 1111–1116 (2012).
54. Peifer, M. *et al.* Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nature Genet.* **44,** 1104–1110 (2012).
55. Stark, M. S. *et al.* Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. *Nature Genet.* **44,** 165–169 (2012).
56. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485,** 502–506 (2012).
57. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150,** 251–263 (2012).
58. Zang, Z. J. *et al.* Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nature Genet.* **44,** 570–574 (2012).
59. Wang, K. *et al.* Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nature Genet.* **43,** 1219–1223 (2011).
60. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
61. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
62. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).
63. Baumbusch, L. O. *et al.* Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* **9,** 379 (2008).
64. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27,** 2144–2146 (2011).