W514–W522 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 4 May 2017*
doi: 10.1093/nar/gkx367

Exploring background mutational processes to decipher cancer genetic heterogeneity

Alexander Goncearenco[1,*], Stephanie L. Rager[1,2], Minghui Li[1], Qing-Xiang Sang[3], Igor B. Rogozin[1] and Anna R. Panchenko[1,*]

[1]National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA, [2]Columbia University, School of Engineering and Applied Science, New York, NY 10027, USA and [3]Department of Chemistry and Biochemistry, Florida State University, Tallahassee, Florida 32306, USA

Received February 28, 2017; Revised April 18, 2017; Editorial Decision April 20, 2017; Accepted April 21, 2017

ABSTRACT

Much remains unknown about the progression and heterogeneity of mutational processes in different cancers and their diagnostic and clinical potential. A growing body of evidence supports mutation rate dependence on the local DNA sequence context for various types of mutations. We propose several tools for the analysis of cancer context-dependent mutations, which are implemented in an online computational framework MutaGene. The framework explores DNA context-dependent mutational patterns and underlying somatic cancer mutagenesis, analyzes mutational profiles of cancer samples, identifies the combinations of underlying mutagenic processes including those related to infidelity of DNA replication and repair machinery, and various other endogenous and exogenous mutagenic factors. As a result, the combination of mutagenic processes can be identified in any query sample with subsequent comparison to mutational profiles derived from malignant and benign samples. In addition, mutagen or cancer-specific mutational background models are applied to calculate expected DNA and protein site mutability to decouple relative contributions of mutagenesis and selection in carcinogenesis, thus elucidating the site-specific driving events in cancer. MutaGene is freely available at https://www.ncbi.nlm.nih.gov/projects/mutagene/.

INTRODUCTION

Cancer genomics studies have revealed high intra- and inter-tumor phenotypic and genetic heterogeneity (1–3). This may be the consequence of various forms of infidelity of DNA replication and repair machinery, differences in tumor micro-environments and various other endogenous and exogenous mutagenic factors. A growing body of evidence supports mutation rate dependence on the local DNA sequence context for various types of mutations (4) and sequence motifs (5). Several DNA context-dependent mutational patterns have been reported that are characteristic for a particular cancer type, tissue (6–11) or mutagen (12–16): UV light, chemical agents, aberrant activity of APOBEC/Activation Induced Deaminase (AID)-family cytidine deaminases and defective mismatch repair or other factors. In addition, sequence dependent structural and thermodynamic properties of the DNA molecule may also affect the DNA repair and replication, and therefore mutagenesis (17,18). Relative contributions of extrinsic and intrinsic factors to DNA mutagenesis have long been debated and the consensus view suggests that many different mutagenic processes have cumulatively shape the observed somatic mutational profiles in cancer.

DNA context-dependent mutational patterns have been analyzed previously (10,17,19–23). Some of these studies examined local DNA sequence contexts around mutated sites for several thousands of cancer genomes and exomes and reported 5 to 30 pervasive mutational signatures (21). It was suggested that these signatures could correspond to the underlying processes of mutagenesis and the etiology of some of them was characterized. Moreover, several computational tools were created to derive signatures from the cancer genomics data (24–28). Nonetheless, there is still a limited understanding of extracted signatures and their clinical potential.

Identification of cancer driver genes and mutations is one of the central problems in cancer research. This calls for further advances in computational techniques to more precisely predict the effects of cancer mutations on protein stability, binding and function (29,30). Statistical models accounting for differential transition and transversion mutation frequencies, kataegis and naïve estimates of the background somatic mutation rates have been used in several driver prediction methods (31). In order to leverage an ex-

*To whom correspondence should be addressed. Tel: +1 301 435 5891; Fax: +1 301 480 4637; Email: panch@ncbi.nlm.nih.gov
Correspondence may also be addressed to Alexander Goncearenco. Tel: +1 301 496 2936; Fax: +1 301 480 4637; Email: gonceare@ncbi.nlm.nih.gov

ponential growth of cancer sequencing data and existing evidence of dependence of mutation rate on cancer type and local DNA sequence contexts (32), it is necessary to explicitly integrate the context-dependent mutations into the cancer specific mutational models to reduce false positive rates in driver gene and driver mutation predictions (33,34).

Here we propose several methods for the analysis of context-dependent mutations, which are implemented in an online computational framework MutaGene (https://www.ncbi.nlm.nih.gov/projects/mutagene/). MutaGene constructs DNA context-dependent mutational profiles and derives signatures from major cancer whole exome and genome sequencing studies available through the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) repositories. Mutational profiles are categorized based on cancer type and primary tumor sites and are normalized by removing the bias from mutational hotspots with recurring mutations. Mutational profiles from the human germline SNPs and benign tissue samples from cancer patients are examined as well. Individual cancer samples are further analyzed in terms of their underlying mutagenesis to explore within and between cancer heterogeneity.

The proposed methods can analyze any given set of mutations, determine the contributions of predefined annotated mutational signatures and identify the cancer type, primary tumor site and cohorts of patients with similar mutagenic processes. This can be interpreted in terms of malfunctions in DNA damage repair mechanisms and exposure to mutagens, for further analyzes with regard to survival, treatment prognosis and drug response. Finally, for any gene or genomic region, MutaGene can apply context-dependent mutational profiles and individual signatures to calculate the background DNA mutability and amino acid substitution rates expected from a given underlying mutagenesis process and not affected by selection. The background mutability can be compared to the observed frequencies of mutations in cancer patients that allows to link cancer genotype with the phenotype to decipher relative roles of mutagenesis and selection in carcinogenesis.

MATERIALS AND METHODS

Data sources for extracting cancer mutations

In order to avoid a bias toward more frequently sequenced genes or mutations identified by genotyping we only considered single base substitutions in protein-coding genomic loci from the whole genome and whole exome-sequenced samples originating from ICGC (35) projects, TCGA (36) and the Pediatric Cancer Genome Project (Supplementary Figure S1). We relied upon annotations from The Catalogue of Somatic Mutations in Cancer (COSMIC) release v75 (37) that curates mutations from these sources and verifies whether mutations are somatic, discarding all mutations with an 'unknown somatic status'. Currently, MutaGene includes 9450 cancer samples from 37 projects with 1,139,534 non-recurring mutations (Figures 1B and S2). Additionally, we included 1,953 somatic mutations identified in 70 benign TCGA samples (38) and common germline variants with no clinical association from the dbSNP database (39).

Construction of context-dependent mutational profiles

We identified the DNA sequences of transcripts affected by mutations using GRCh38 reference human genome assembly. The nucleotide context of a mutation was defined as the neighboring nucleotides in 5′ and 3′ directions from the mutated nucleotide according to the transcript sequence. Altogether, six substitution types (C→A, C→T, C→G, T→A, T→C, T→G) in 16 possible 5′3′ contexts result in 96 context-dependent mutation types: $\tau = \{N[x \in Y \rightarrow N \backslash x] N\}, \; where \; N = \{A, T, C, G\}$. Given the number of observed mutations f_i for each type τ_i, a mutational profile can be represented as a probability mass function of a multinomial distribution for all possible context-dependent mutation types $V(v_1, \ldots, v_{96}): \; v_i = \frac{f_i}{\sum_i f_i} = \Pr(\text{mutation} = \tau_i)$. Recurring mutations observed at the same genomic location in multiple patients (Figures S2 and S3) were counted only once since these mutations and sites might be under selection (hotspots).

Derivation of mutational signatures

Cancer specific context-dependent mutational profile is a manifestation of different mutational processes. These processes may have distinct etiology and may result in distinct sets of mutations in characteristic DNA sequence contexts, so called *context-dependent mutational signatures* (21). Previously, the numeric deconvolution of matrices of mutational profiles of cancer samples (Figure 2A) was obtained using the non-negative matrix factorization (NMF) method (28,40–42). The major goal of this procedure was to obtain the underlying mutational signatures (Figure 2B) and represent cancer samples in terms of exposure to these signatures (Figure 2C). Mutational signatures should ideally represent distinct uncorrelated context-dependent mutational processes that can be further annotated in terms of their etiology. Sparseness of mutational signatures becomes an important aspect when it comes to annotation of signatures. Assuming that a limited number of mutational processes affects each cancer sample, the exposure matrix also has to be sparse. Therefore we applied non-smooth (ns)NMF method (43) with sparse random initialization allowing to obtain sparse solutions for both signature and exposure matrices (Supplementary Figures S7C and S7D) while avoiding high correlation between the signatures (Supplementary Figure S7B).

Given n cancer samples, we represented context dependent mutations ($m = 96$) in these samples as a non-negative matrix $V_{m \times n}$. NMF finds two non-negative matrices $W_{m \times k}$ and $H_{k \times n}$ for a given number of components k, so that $V \approx WH$. Matrix $W_{m \times k}$ contains k mutational signatures and matrix $H_{k \times n}$ describes exposures of n samples to k mutational processes defined by signatures in a matrix W. In case of non-smooth NMF decomposition method (nsNMF) (43), the problem is formulated as $V \approx WSH$, where a square non-negative smoothing matrix $S_{k \times k}$ allows reconstructing into globally sparse solutions. Figure 2 and Supplementary Figures S5 and S6 illustrate the deconvolution for NMF and nsNMF methods.

We analyzed the reproducibility of deconvolution results by calculating a cophenetic correlation coefficient of con-

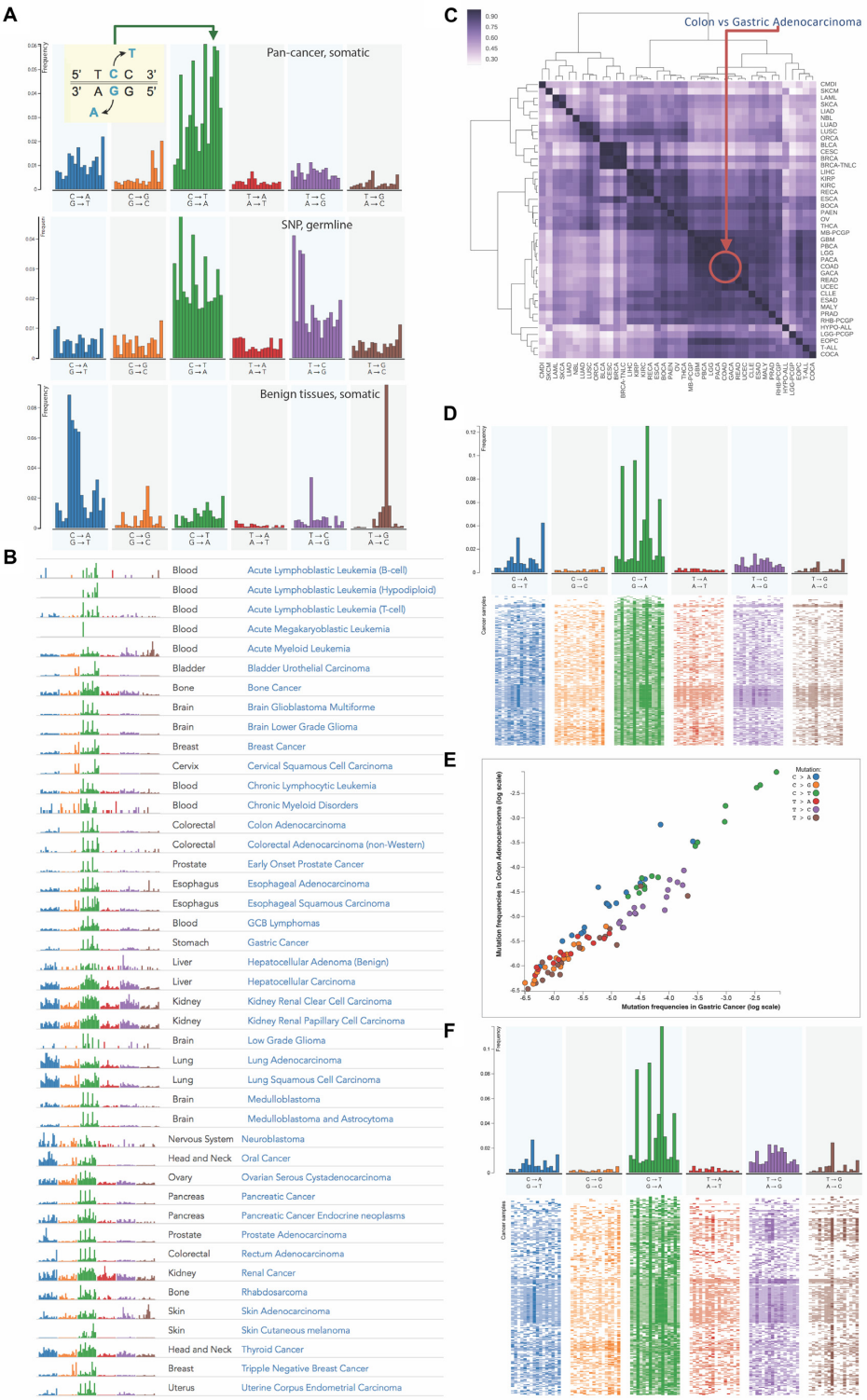

Figure 1. Exploring and comparing context-dependent mutational profiles in various cancer types. (**A**) Mutational profiles of pan-cancer somatic mutations, germline mutations (single nucleotide polymorphisms) and somatic mutations found in benign tissues in cancer patients. (**B**) The list of fingerprints of mutational profiles of 37 cancer types defined based on large scale whole-genome and whole-exome projects. (**C**) Similarity matrix calculated by different distance measures between cancer-specific mutational profiles and clustering of cancer types according to similarities of their profiles using cosine metric. (**D**–**F**) 2D profiles of colon and gastric cancers illustrate within-cancer heterogeneity, where each line corresponds to the tumor sample and each column to the type of context-dependent mutation. (E) Comparison of relative frequencies of mutation types between colon and gastric cancers on log scale.

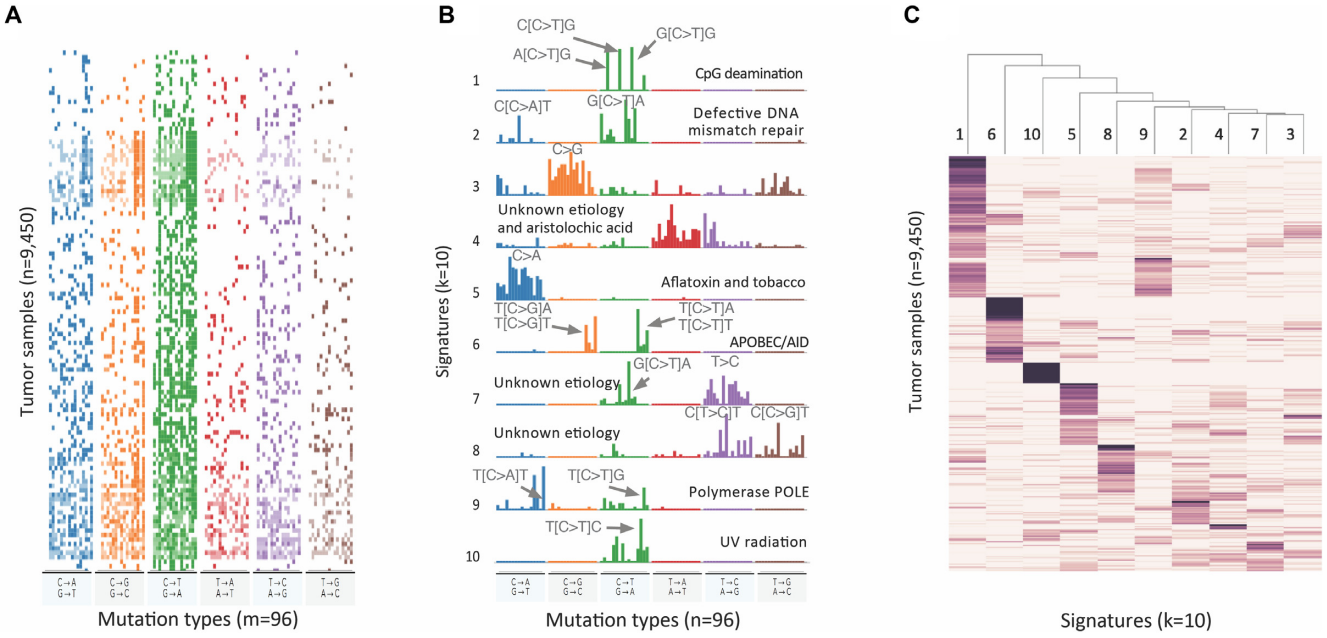

Figure 2. Decomposition of mutational profiles of pan-cancer samples into mutational signatures. (**A**) The matrix of mutational profiles of pan-cancer samples V; (**B**) Ten mutational signatures MutaGene10 in matrix W annotated by the corresponding mutagenic processes and (**C**) exposure matrix H representing relative contributions of mutational processes (represented by ten signatures) in each tumor sample. The matrices are transposed for visualization purpose. Matrices V and H are heat maps where pixel intensity indicates the frequency, whereas in matrix W the values of row vectors as shown as bar plots.

sensus matrix $\bar{C}$ (see Supplementary Data). Previously it was used to determine an optimal number of components in NMF deconvolution as a point where this coefficient begins to decrease (42). A consensus matrix $\bar{C}_{n \times n}$ is an average of connectivity matrices produced as a result of multiple runs of NMF or nsNMF algorithms. A connectivity matrix $C_{n \times n}$ is calculated based on the exposure matrix H and shows if samples are exposed to the same dominating mutagenic process. Supplementary Figure S7A shows that cophenetic correlation coefficient decreases after five and ten components when applying NMF and nsNMF respectively. Moreover, NMF deconvolution with more than ten signatures/components resulted in highly correlated signatures with correlation coefficients larger than 0.6 (Supplementary Figures S7B and S13). Therefore, we used nsNMF decompositions with five and ten components and obtained two *MutaGene5* and *MutaGene10* signature sets listed and annotated on the MutaGene website.

Identification of mutational profiles, signatures and mutagenic processes for a query set of mutations

MutaGene tools determine the mutational DNA context according to the reference human genome assembly and construct a query mutational profile for any given set of mutations. The query profile can be compared to the collection of profiles in the MutaGene database using the 'Identify' tool, which ranks the query mutational profile by its distance to the MutaGene profiles using the k-nearest neighbors classifier and distance measures listed in Supplementary Data. MutaGene also provides Naïve Bayes, random forest and linear support vector machines (SVM) classifiers pre-trained with 4-fold cross-validation, treating different cancer types and primary tumor sites as classification labels. Performance evaluation results are shown in Supplementary Figures S8–11. In addition, contributions (exposures) h of pre-annotated signature sets W (*MutaGene5*, *MutaGene10* and *COSMIC30*) for a sample query profile v are calculated by solving $W\boldsymbol{h} = \boldsymbol{v}$ with a non-negative constrained least squares method.

Analysis of mutability

A mutational background model represented by the mutational profile or mutational signature can be applied to any protein-coding gene sequence or any other genomic region in order to calculate the DNA mutability expected for each particular site. DNA mutability of a base b_i in a trinucleotide context $t = b_{i-1}\, b_i b_{i+1}$ is calculated using the total number of mutations F_t for a given trinucleotide t according to the context-dependent mutational profile F. Given the number of samples included in the mutational profile n_F and the number of trinucleotides t in a diploid human exome x_t according to the reference genome assembly GRCh38, one can define mutability as $\omega_t = \frac{F_t}{n_F\, x_t}$. The number of trinucleotides in the human exome is calculated for each nucleotide base considering its trinucleotide context. In case when mutability is calculated using signatures, mutation rate $\frac{F_t}{n_F}$ is set to 120, which corresponds to the pan-cancer average of the number of mutations per exome (Supplementary Figure S4). MutaGene also allows to specify an arbitrary mutation rate for mutability calculations. Mutability is estimated per Megabase per sample (Figure 3). Mutability of a codon Ω_j is calculated as the

Figure 3. Analysis of mutability in DNA and protein sequences. Two examples showing the results of mutability calculations using the pan-cancer mutational profile: (**A**) a sequence fragment of *TP53* gene and protein with Arg273 interacting with the DNA (red arrow) and Arg283 not in direct contact with the DNA (blue arrow) and (**B**) mutability of site Leu858 (red arrow) in epidermal growth factor receptor *(EGFR)* gene and protein using a pan-cancer mutational profile. Expected DNA mutability depending on the local sequence context, depicted as a green line, is scaled per Megabase DNA per cancer sample. The heatmap shows expected mutabilities for each DNA position, where the color encodes silent, missense and nonsense nucleotide substitutions and color intensity represents the scale of mutability values. Expected mutability is translated onto the protein level (shown in orange line for each codon) and the heatmap below shows the mutability values for each amino acid substitution. Yellow circles indicate mutations observed in cancer patients, with height of the pins showing relative numbers of observed mutations in log scale. Note that in sites pointed by arrows expected mutability and observed mutation frequencies show the opposite trends. Such maps are generated by MutaGene 'Analyze mutability' tool and are designed for interactive use, where clicking on a circle shows a distribution of observed mutations over cancer types.

sum of mutabilities of the three nucleotides comprising the codon $\Omega_j = \sum_{i=1,3j-2}^{3j} \omega_i$. Amino acid substitutions corresponding to each mutation are calculated by translating the mutated codon using a standard codon table. The relative propensities of all types of mutations in a codon j including missense, silent and nonsense mutations sum up to Ω_j.

RESULTS

Exploring the diversity of mutational profiles in human cancer

A context-dependent mutational profile is a results of contributions of different DNA context-dependent processes characteristic for a given cancer sample. Importantly, mutational profile is calculated solely based on the types of nucleotide substitutions and their context, regardless of the chromosome location and gene type. Mutational profiles are represented as probability mass functions in order to emphasize the relative mutational preferences, because an absolute value of mutation rate (Supplementary Figure S2) is not necessarily directly associated with the underlying mutagenic processes and may be determined by the number of replications, cell divisions and other factors (44). Moreover, some sites (hotspots) may harbor several hundreds of mutations from different samples (45) and mainly represent mutations offering selective advantage to the clone. In fact, we found that about eight percent of all cancer mutations were recurring. Proportions of recurring mutations varied depending on cancer type reaching up to 30% of recurring mutations in some cases (Figures S3). To avoid biases caused by selection acting upon particular genomic sites, mutational profiles have been derived by counting mutations only once and excluding recurring mutations (Figure S14).

A collection of mutational profiles in MutaGene allows to explore the diversity of mutagenic processes in different cancers and tissues (Figure 1B). In Figure 1A, a pan-cancer somatic mutational profile is shown along with the profiles of germline mutations obtained from human SNPs and somatic mutations found in benign human tissue samples. Pan-cancer and most cancer specific mutational profiles (Figure 1B) contain a dominating C→T mutations in the NpCpG context, which is also characteristic for germline mutational profile, pointing to striking similarities between the accumulation of mutations in tumors and in germline cells. The mutation rate at nucleotide C in the CpG dinucleotide context associated with methylation was previously found to be much larger than that of other sites (46). The mutational profile of benign pan-tissue somatic mutations allows distinguishing cancer-specific somatic patterns and highlights the differences between benign and malignant tissues, particularly in T→G and C→A transversions.

Figure 1C shows a heat map representing the pairwise comparisons of all cancer-specific mutational profiles. This comparison reveals inter-cancer similarities in terms of their mutational profiles. Namely, the similarities have been detected between lung and oral carcinomas; breast, bladder and cervical carcinomas; liver and kidney carcinomas; and a large group of blood, brain, gastric and colorectal cancers. However, while mutational profiles of different cancer types can be very similar, as in the case of colon and gastric adenocarcinomas (cosine similarity of 0.98) (Figure 1E), mutational profiles of individual samples within the same cancer type may reveal large heterogeneity. In Figure 1D and 1F individual cancer samples are ordered based on the distances between their mutational profiles, shown as bands on the heat maps, indicating that within cancer types these differences may be more pronounced than between cancer types. Understanding of cancer genetic heterogeneity is deeply rooted in our understanding of the underlying mutagenic processes and will be explored in the following sections.

Analysis of query mutational profiles

MutaGene can analyze any set of mutations from one or several cancer samples to identify cancers of unknown primary tumor site, to detect the most likely mutagenic process and to distinguish tumorigenic from benign mutation sets (Supplementary Figure S12). First, a mutational profile is calculated for a query sample of interest. Next, MutaGene compares the query profile to the collection of profiles and signatures in the MutaGene database and calculates the contributions of different annotated mutagenic processes to the mutational profile of interest (exposures). We thoroughly assessed the performance of cancer type and primary tumor site identification with a cross-validation benchmarks, in particular the dependence of its accuracy on the number of mutations in the query sample. We found that the average accuracy of primary site identification ranges from 38 to 85% (Supplementary Figure S9D). According to our benchmarks, random forest classification method outperformed multinomial Naïve Bayes and SVM classifier with a linear kernel (Supplementary Figures S9, S10), therefore random forest is used in MutaGene by default. For cancer types that have similar mutational profiles (Figure 1C), it could be sufficient to identify a correct cancer or tissue type within the top two or three matches. Using this more relaxed criterion, the average accuracy of cancer type prediction using random forest classifier increases from 66 to 90%, if we consider top three matches (Supplementary Figure S10B). For primary site prediction, the same approach would show a boost in accuracy from 72 to 92% (Supplementary Figure S10E).

Per-class performance analysis shows (Supplementary Figure S11A and Table ST1) that some cancer types, such as pancreatic cancer , breast cancer and renal adenocarcinoma are not predicted correctly as a top-matching hit. Due to within-cancer heterogeneity, some samples belonging to lung squamous cell carcinoma were attributed to lung adenocarcinoma (LUAD), however almost all LUAD samples were correctly identified. Regarding the primary tumor sites, esophagus and pancreas are the most problematic sites, where the classifier incorrectly attributes most of the samples to colon and stomach because mutational profiles of stomach and colorectal samples are practically indistinguishable. However, despite high heterogeneity, MutaGene correctly identifies the primary site for almost all liver, lung and colorectal samples (Supplementary Figure S11B and Table ST2). Therefore, we show that it is possible to identify cancer types and primary sites for a given cancer sample with sufficient accuracy using only the information about

its mutational profile. This analysis uncovers and illustrates diagnostic potential of context-dependent mutational profiles, however in practical diagnostic applications it may be necessary to combine mutational profile with other types of data such as presence/absence of mutations in certain genes, copy number variations, gene expression and DNA methylation.

Estimating the background DNA and protein site mutability

MutaGene provides background mutational models in the form of cancer-specific mutational profiles or mutagen-specific signatures that can be used to calculate the number of mutations expected as a result of underlying mutagenesis, not affected by selection pressure in somatic cells. The site mutability (see 'Materials and Methods' for definition) can be estimated for each genomic site thus allowing to compare relative mutabilities of different sites between each other and simultaneously relate them to the frequencies of mutations observed in certain sites in cancer patients. For protein-coding sequences MutaGene calculates the rates of expected amino acid substitutions for each codon thus taking into account the local DNA context, the nucleotides surrounding the codons of each amino acid. Figure 3A shows the DNA and protein mutability for a fragment of gene *TP53* calculated using pan-cancer mutational profile. This figure shows two sites: R273 that is involved in DNA binding (red arrow) and R283 site (blue arrow) that is not directly involved in binding of DNA. There is experimental evidence (47) that any missense mutations in codons of DNA-binding arginine result in a loss of function, whereas many amino acid substitutions of another arginine can be tolerated (48). These two arginines have different mutability values since mutability depends on both codons and nucleotides surrounding these codons. Particularly interesting is that the key position involved in interactions with DNA (R273) has the highest numbers of observed mutations in cancer patients, however its mutability is much lower than that of other arginine (R283) that is not involved in DNA interactions. Consistent with this, Figure 3B shows a very low expected mutability of an oncogene epidermal growth factor receptor (*EGFR*) L858 site (red arrow) although it is frequently mutated in cancer. Respectively, expected mutabilities of adjacent codons, that are supposedly not under selection in cancer, are high. In general, by comparing observed frequencies to expected mutabilities one can potentially get important clues about the potential cancer driving events.

DISCUSSION

Cancers are notorious for their intra- and inter-tumor functional and genetic heterogeneity, which imposes difficulties in terms of cancer type classification and targeted drug therapy. Exploring the heterogeneity of cancer in terms of mutagenic processes is not trivial. First, mutational profiles of cancer samples with only a few mutations could be too sparse and not well defined. Second, mutational profiles of samples represent a combination of different mutational signatures and processes, many of which remain uncharacterized (21,22,49). Third, mutational processes may act independently, but their signatures may be overlapping, for instance the signature of somatic hypermutation enzyme, *AID,* as we identified recently, overlaps with the CpG methylation site (50). Finally, mutational profiles and signatures are intended to represent the context-dependent propensities determined by the underlying background mutagenic processes rather than selection and a signal coming from selection processes is hard to eliminate. The evolution of cancer is largely driven by somatic mutations and clonal selection of these mutations (51); it is therefore important to decouple mutagenesis from selection in order to characterize driving events in tumor evolution. Mutagenesis can be affected by the local DNA sequence context around the mutated site and therefore sequence context should be accounted for in estimating the mutational probability and mutation rate at any given site. Context-dependent mutational models allow MutaGene to calculate the expected background mutability of nucleotide and protein sites, thereby linking processes operating at the DNA level to the protein phenotype. The choice of mutational model is crucial and the expected mutability may largely depend on the background model. Additionally, considering mutational hotspots and excluding recurring mutations that may be subject to selection is important for calculating the accurate background mutational model.

In addition to histological characterization of a cancer sample, methods of molecular diagnostics are aimed toward correct and timely diagnosis and the optimal choice of personalized treatment for a cancer patient. Currently these methods are mostly relying on biomarkers related to differential gene expression, methylation, copy number variation and by the presence or absence of mutations in certain genes. However, much remains unknown about the mutational processes operating at the level of DNA in any given cancer patient or sample. Identification of the underlying mutational processes can improve molecular subtype classification, particularly in cancers with high heterogeneity. Identification of cancer type and primary site is also important for free-floating DNA blood samples and metastatic cancer samples, where the original tumor site may be unknown. Additionally, it may help to identify the actual source of tumor in case of a metastatic sample. Mutational studies in cell culture, viral and animal models may also require a comparison to the reference human datasets using MutaGene. Coupled to the analysis of clinical features, such as drug response, resistance and survival for different cohorts of patients with similar mutational profiles mutational analysis with MutaGene server provides an additional factor to consider in explaining cancer heterogeneity.

AVAILABILITY

MutaGene is freely available at https://www.ncbi.nlm.nih.gov/projects/mutagene/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Michael Lynch, Lyudmil Alexandrov, Nir Ben-Tal, Gerhard Manning and Marie Evangelista for helpful discussions and Janet Coleman for proofreading.

Author Contributions: A.G. and A.R.P. designed the analysis and wrote the paper. A.G. developed the framework. S.L.R, M.L., Q.X.S. and I.B.R. applied the framework to the analysis of cancer genomes. All authors approved the manuscript.

FUNDING

Intramural Research Programs of the National Library of Medicine; National Institutes of Health. Funding for open access charge: National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
2. Swanton,C. (2012) Intratumor heterogeneity: evolution through space and time. *Cancer Res.*, **72**, 4875–4882.
3. Andor,N., Graham,T.A., Jansen,M., Xia,L.C., Aktipis,C.A., Petritsch,C., Ji,H.P. and Maley,C.C. (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, **22**, 105–113.
4. Francioli,L.C., Polak,P.P., Koren,A., Menelaou,A., Chun,S., Renkens,I. and Genome of the Netherlands, C.Genome of the Netherlands, C., van Duijn,C.M., Swertz,M., Wijmenga,C., van Ommen,G. *et al.* (2015) Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.*, **47**, 822–826.
5. Sharp,N.P. and Agrawal,A.F. (2016) Low genetic quality alters key dimensions of the mutational spectrum. *PLoS Biol.*, **14**, e1002419.
6. Pfeifer,G.P. and Hainaut,P. (2003) On the origin of G→T transversions in lung cancer. *Mutat. Res.*, **526**, 39–43.
7. Boutros,P.C., Fraser,M., Harding,N.J., de Borja,R., Trudel,D., Lalonde,E., Meng,A., Hennings-Yeomans,P.H., McPherson,A., Sabelnykova,V.Y. *et al.* (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.*, **47**, 736–745.
8. Alexandrov,L.B., Nik-Zainal,S., Siu,H.C., Leung,S.Y. and Stratton,M.R. (2015) A mutational signature in gastric cancer suggests therapeutic strategies. *Nat. Commun.*, **6**, 8683.
9. Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.
10. Nik-Zainal,S., Alexandrov,L.B., Wedge,D.C., Van Loo,P., Greenman,C.D., Raine,K., Jones,D., Hinton,J., Marshall,J., Stebbings,L.A. *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979–993.
11. Schulze,K., Imbeaud,S., Letouze,E., Alexandrov,L.B., Calderaro,J., Rebouissou,S., Couchy,G., Meiller,C., Shinde,J., Soysouvanh,F. *et al.* (2015) Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.*, **47**, 505–511.
12. Brash,D.E. (2015) UV signature mutations. *Photochem. Photobiol.*, **91**, 15–26.
13. Poon,S.L., Huang,M.N., Choo,Y., McPherson,J.R., Yu,W., Heng,H.L., Gan,A., Myint,S.S., Siew,E.Y., Ler,L.D. *et al.* (2015) Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med.*, **7**, 38.
14. Langie,S.A., Koppen,G., Desaulniers,D., Al-Mulla,F., Al-Temaimi,R., Amedei,A., Azqueta,A., Bisson,W.H., Brown,D.G., Brunborg,G. *et al.* (2015) Causes of genome instability: the effect of low dose chemical exposures in modern society. *Carcinogenesis*, **36**, S61–S88.
15. Roberts,S.A., Lawrence,M.S., Klimczak,L.J., Grimm,S.A., Fargo,D., Stojanov,P., Kiezun,A., Kryukov,G.V., Carter,S.L., Saksena,G. *et al.* (2013) An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.*, **45**, 970–976.
16. Sammalkorpi,H., Alhopuro,P., Lehtonen,R., Tuimala,J., Mecklin,J.P., Jarvinen,H.J., Jiricny,J., Karhu,A. and Aaltonen,L.A. (2007) Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer Res.*, **67**, 5691–5698.
17. Sung,W., Ackerman,M.S., Gout,J.F., Miller,S.F., Williams,E., Foster,P.L. and Lynch,M. (2015) Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Mol. Biol. Evol.*, **32**, 1672–1683.
18. Bauer,N.C., Corbett,A.H. and Doetsch,P.W. (2015) The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res.*, **43**, 10083–10101.
19. Siepel,A. and Haussler,D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, **21**, 468–488.
20. Pfeifer,G.P. and Besaratinia,A. (2009) Mutational spectra of human cancer. *Hum. Genet.*, **125**, 493–506.
21. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Borresen-Dale,A.L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
22. Helleday,T., Eshtad,S. and Nik-Zainal,S. (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, **15**, 585–598.
23. Rogozin,I.B., Babenko,V.N., Milanesi,L. and Pavlov,Y.I. (2003) Computational analysis of mutation spectra. *Brief Bioinform.*, **4**, 210–227.
24. Petljak,M. and Alexandrov,L.B. (2016) Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*, **37**, 531–540.
25. Hollstein,M., Alexandrov,L.B., Wild,C.P., Ardin,M. and Zavadil,J. (2016) Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene*, **36**, 158–167.
26. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
27. Ardin,M., Cahais,V., Castells,X., Bouaoun,L., Byrnes,G., Herceg,Z., Zavadil,J. and Olivier,M. (2016) MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics*, **17**, 170.
28. Gehring,J.S., Fischer,B., Lawrence,M. and Huber,W. (2015) SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, **31**, 3673–3675.
29. Li,M., Kales,S.C., Ma,K., Shoemaker,B.A., Crespo-Barreto,J., Cangelosi,A.L., Lipkowitz,S. and Panchenko,A.R. (2016) Balancing protein stability and activity in cancer: a new approach for identifying driver mutations affecting CBL ubiquitin ligase activation. *Cancer Res.*, **76**, 561–571.
30. Li,M., Goncearenco,A. and Panchenko,A.R. (2017) Annotating mutational effects on proteins and protein interactions: designing novel and revisiting existing protocols. *Methods Mol. Biol.*, **1550**, 235–260.
31. McFarland,C.D., Korolev,K.S., Kryukov,G.V., Sunyaev,S.R. and Mirny,L.A. (2013) Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 2910–2915.
32. Lynch,M. (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 961–968.
33. Chang,M.T., Asthana,S., Gao,S.P., Lee,B.H., Chapman,J.S., Kandoth,C., Gao,J., Socci,N.D., Solit,D.B., Olshen,A.B. *et al.* (2016) Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.*, **34**, 155–163.
34. Carter,H., Chen,S., Isik,L., Tyekucheva,S., Velculescu,V.E., Kinzler,K.W., Vogelstein,B. and Karchin,R. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, **69**, 6660–6667.
35. Gonzalez-Perez,A., Mustonen,V., Reva,B., Ritchie,G.R., Creixell,P., Karchin,R., Vazquez,M., Fink,J.L., Kassahn,K.S., International Cancer Genome Consortium Mutation, P. and Consequences Subgroup of the Bioinformatics Analyses Working, G. *et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, **10**, 723–729.

36. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
37. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
38. Yadav,V.K., DeGregori,J. and De,S. (2016) The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Res.*, **44**, 2075–2084.
39. Coordinators,N.R. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
40. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Campbell,P.J. and Stratton,M.R. (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, **3**, 246–259.
41. Rosenthal,R., McGranahan,N., Herrero,J., Taylor,B.S. and Swanton,C. (2016) DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, **17**, 31.
42. Brunet,J.P., Tamayo,P., Golub,T.R. and Mesirov,J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 4164–4169.
43. Pascual-Montano,A., Carazo,J.M., Kochi,K., Lehmann,D. and Pascual-Marqui,R.D. (2006) Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans Pattern Anal. Mach. Intell.*, **28**, 403–415.
44. Tomasetti,C. and Vogelstein,B. (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**, 78–81.
45. Rogozin,I.B. and Pavlov,Y.I. (2003) Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat. Res.* **544**, 65–85.
46. Zhao,Z. and Jiang,C. (2007) Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. *Mol. Biol. Evol.*, **24**, 23–25.
47. Petitjean,A., Mathe,E., Kato,S., Ishioka,C., Tavtigian,S.V., Hainaut,P. and Olivier,M. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum. Mutat.*, **28**, 622–629.
48. Petitjean,A., Achatz,M.I., Borresen-Dale,A.L., Hainaut,P. and Olivier,M. (2007) TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene*, **26**, 2157–2165.
49. Alexandrov,L.B. and Stratton,M.R. (2014) Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.*, **24**, 52–60.
50. Rogozin,I.B., Lada,A.G., Goncearenco,A., Green,M.R., De,S., Nudelman,G., Panchenko,A.R., Koonin,E.V. and Pavlov,Y.I. (2016) Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. *Sci. Rep.*, **6**, 38133.
51. Greaves,M. (2015) Evolutionary determinants of cancer. *Cancer Discov.*, **5**, 806–820.