Computational Biology and Chemistry 53 (2014) 79–83

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Evidence of a cancer type-specific distribution for consecutive somatic mutation distances

Jose M. Muiño[a,*], Ercan E. Kuruoğlu[b], Peter F. Arndt[a]

[a] Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin,Germany
[b] Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), 56124 Pisa, Italy

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online 23 August 2014

Keywords:
Mutation distribution
Cancer
Power-law
Kataegis

ABSTRACT

Specific molecular mechanisms may affect the pattern of mutation in particular regions, and therefore leaving a footprint or signature in the DNA of their activity. The common approach to identify these signatures is studying the frequency of substitutions. However, such an analysis ignores the important spatial information, which is important with regards to the mutation occurrence statistics. In this work, we propose that the study of the distribution of distances between consecutive mutations along the DNA molecule can provide information about the types of somatic mutational processes. In particular, we have found that specific cancer types show a power-law in interoccurrence distances, instead of the expected exponential distribution dictated with the Poisson assumption commonly made in the literature. Cancer genomes exhibiting power-law interoccurrence distances were enriched in cancer types where the main mutational process is described to be the activity of the APOBEC protein family, which produces a particular pattern of mutations called Kataegis. Therefore, the observation of a power-law in interoccurence distances could be used to identify cancer genomes with Kataegis.

1. Introduction

DNA mutations are one of the main generators of variability and complexity at the DNA sequence level in the genome. Several molecular mechanisms can induce mutations, and interestingly, some of them produce specific patterns in the DNA sequence that can be recognized by computational analysis. For example, the pattern produced by the APOBEC protein-family (Roberts et al., 2013) is characterized by the substitution of cytosines (C) by thymines (T), often associated to the motif tC, i.e., the capitalized mutated C flanked by a T on its 3' side. This mutagenic activity is also affected by the methylation status of the cytosine: methylated cytosines having a higher rate of mutability (see for example (Wijesinghe and Bhagwat, 2012)). Therefore, the characterization and identification of these signatures can help to understand the complexity of the genome at the DNA sequence level, and even how this complexity has been generated.

The characterization and identification of these mutational signatures has been recently facilitated by the release of the sequenced genomes of a large number of patients with different types of cancer, since: (1) in contrast to genomes of healthy donors which contain only few somatic mutations, the genomes of cancer patients often show large number of mutations (see for example (Alexandrov et al., 2013)), (2) most of the cancer types studied can be described by the influence of two main mutational processes (Alexandrov et al., 2013), and (3) there is a large number of cancer genomes sequenced and publically available. In fact, (Alexandrov et al., 2013) has already characterized more than 20 mutational signatures in terms of nucleotide substitution frequencies among more than 7000 cancers. However, it is possible that the mutagenic mechanisms influence not only the frequency of substitutions, but also other properties of the distribution of mutations, as in for example the distribution of the distances among consecutive mutations. Nevertheless, we are not aware of any previous work addressing this.

Assuming that mutations happen randomly in the genome following a Poisson dynamics which is the most common model for arrival or counting processes, it is expected that the distribution of distances between consecutive somatic mutations will follow a discrete exponential distribution independently of the overall frequency of mutations in the genome/region of interest. However, recently, a particular pattern of localized somatic mutations in cancer genomes has been observed. This pattern has been termed Kataegis (Nik-Zainal et al., 2012) and its defining feature are mutations spaced one to several hundred nucleotides apart that are clustered over kilobase-sized regions. In particular, the "rule of

[*] Corresponding author. Tel. : +49 30 8413 1169; fax: +49 30 8413 1152.
E-mail address: muino@molgen.mpg.de (J.M. Muiño).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.012
1476-9271/© 2014 Elsevier Ltd. All rights reserved.

thumb" used to define these regions is as having six or more consecutive mutations with an average distance of less than or equal to 1 kb (e.g., (Taylor et al., 2013)). This pattern has been linked to the activity of particular proteins of the APOBEC family, often acting in the specific motif tCw (where w is adenine or thymine), and associated with the activity of APOBEC near rearrangement break points (Taylor et al., 2013). Clusters of mutations may produce an excess of short distances among consecutive mutations than expected by the exponential distribution, and therefore it will be an excellent starting point to study how mutational processes may affect this distribution. In this manuscript, we describe for the first time that the distribution of distances among consecutive mutations in some cancer types departs from the exponential distribution in the range of short distances (less than 5 kb) and, in fact, it follows a power-law. Among the cancer genomes showing this phenomenon, we see an enrichment on cancer types associated with the mutational process of APOBEC activity.

2. Material and methods

Genomic locations of somatic mutations for 507 cancer individuals comprising 10 cancer types were obtained from (Alexandrov et al., 2013). We considered only whole genome sequencing data, and only single nucleotide substitutions located in autosomal chromosomes (1–22). In the following, the distances between mutations were calculated as distances between consecutive mutations per chromosome. To plot the distribution of absolute frequencies we used the *R* function hist, with parameter breaks = 100; this means that 100 bins of equal size were obtained, and for each bin the number of counts was calculated. Later, if needed, the distribution was plotted in the log–log scale.

Vuong closeness test (Vuong, 1989) was used to determine when the discrete power-law is a closer representation to the data than the discrete exponential distribution. This is a likelihood ratio test for model selection using the Kullback–Leibler criteria. The test statistic, *R*, is the ratio of the log-likelihoods of the data between the two competing models. The sign of *R*, indicates which model is better. The null hypothesis is that both distributions are equally far from the true distribution of the data. To apply this test, we used the *R* package *poweRlaw* (Clauset et al., 2009). We used default parameters except for $x_{\min}$ that was set to 20. This parameter indicates that data points with a lower value than 20 will not be considered to calculate the likelihood. We set this parameter to 20, because as it can be observed in Fig. 1A and D, there is a peak in the distribution for the range 0–10 bp. The probability function used by *powerRlaw* for the discrete power-law distribution was:

$$p(x) = \frac{1}{\xi(\alpha, x_{\min})} \qquad (1)$$

and for the discrete exponential distribution:

$$p(x) = \left(1 - \mathrm{e}^{-\lambda}\right)\mathrm{e}^{\lambda x_{\min}} \qquad (2)$$

where $x_{\min}$ was set to 20 for both cases, as explained above.

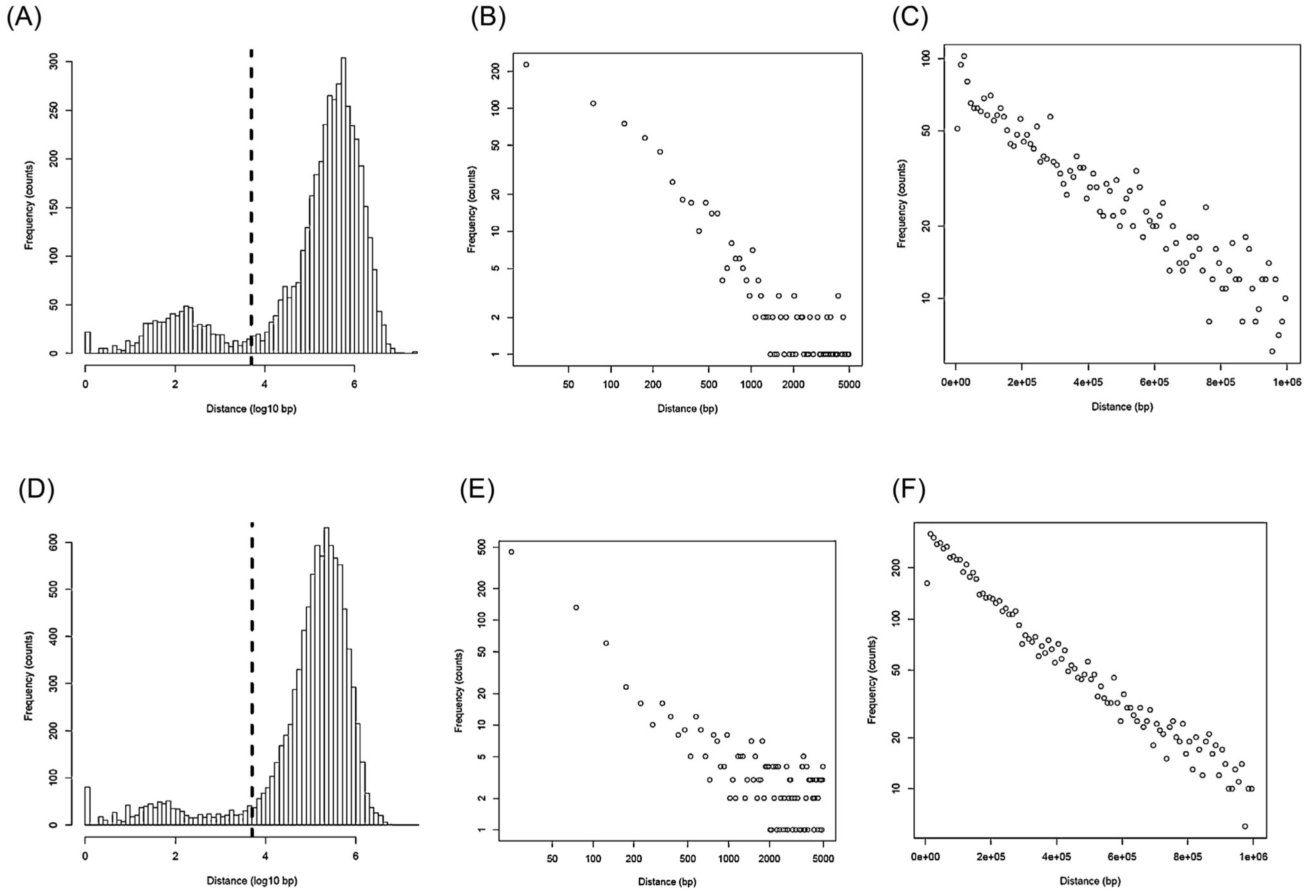

Fig. 1. Histogram of distances (log 10 bp) among consecutive mutations in autosomal chromosomes from breast cancer genomes of the individual *PD4103a* (A) and *PD4107a* (D) is shown. Vertical dash line indicates the distance 5 kb that we have used to separate the population in two: one with short distances (0–5 kb) and other with long distances (5 kb–1 Mb). A log–log plot of the frequency distribution of the subpopulation with short distances for the individual *PD4103a* (B) and *PD4107a* (E) is shown. A log-linear plot of the frequency distribution of the subpopulation with long distances (5 kb–1 Mb) for the individual *PD4103a* (E) and *PD4107a* (F) is shown.

3. Results

3.1. Distribution of mutation distances in two breast cancer genomes with Kataegis

The phenomenon of Kataegis is illustrated for first time in (Nik-Zainal et al., 2012) with the analysis of the genome of two breast cancer patients (ids: *PD4103a* and *PD4107a*). Fig. 1A and D shows their distribution of distances among consecutive mutations (on a log scale). From these distributions, it seems that there are two clear subpopulations, one comprising distances less than 5 kb and other for distances larger than 5 kb; we will define this second region as the one comprising distances from 5 kb to 1 Mb throughout this manuscript to guarantee a good estimation of the density. Interestingly, when we study their distribution, the first subpopulation have a linear behavior in the log–log scale (Fig. 1B and E), meanwhile the second subpopulation shows a linear behavior in the log-linear scale (Fig. 1C and F). This is an indication that the distribution of short distances (ranging from 0 to 5 kb) is well described by a power-law, meanwhile the distribution for the subpopulation ranging from 5 kb to 1 Mb seems to be visually better described by an exponential behavior. Indeed, when we used Vuong's closeness test to determine if the subpopulation of shorter distances follows a discrete power-law compared to a discrete exponential distribution, we obtained that the data is represented significantly better by the discrete power-law than the discrete exponential ($\text{pv} \leqq 0.0012$ for *PD4103a* and $\text{pv} \leqq 9 \times 10^{-19}$ for *PD4107a*), meanwhile the subpopulation ranging from 5 kb to 1 Mb is represented better by a discrete exponential distribution than the discrete power-law ($\text{pv} \leqq 10^{-60}$ in both cases). The maximum likelihood estimation for the exponent of the power-law distribution was 1.44 for *PD4103a* and 1.41 for *PD4107a*.

3.2. Mutation distances distribution as a specific signature of cancer genomes with Kataegis

The previous section demonstrates the presence of two subpopulations in the distribution of mutation distances in two breast cancer patients. Unexpectedly, the subpopulation of short distances (0–5 kb) follows a power-law, instead of the expected exponential distribution. Next, we wanted to find out if this pattern is also present in other cancer types.

We downloaded the location of somatic mutations of 507 cancer individuals comprising 10 cancer types from (Alexandrov et al., 2013). We selected only mutations (single nucleotide substitutions) in autosomal chromosomes and calculated the consecutive distance between mutations. For distances in the range 0–5 kb, we applied Vuong's closeness test (Vuong, 1989) to determine which distribution, discrete power-law or discrete exponential, can represent the data better.

Table 1 shows in the second column the number of cancer genomes/individuals where the distribution of short distances (0–5 kb) can be represented significantly more closely ($\text{pv} < 0.05$) by a discrete power-law compared to the discrete exponential distribution. Breast cancer and B-cell lymphoma are among the cancer types with relatively more individuals with a significant power-law distribution. It has been proposed that there is a link between Kataegis and APOBEC activity (e.g., (Alexandrov et al., 2013; Nik-Zainal et al., 2012; Taylor et al., 2013)). Among cancer genomes/individuals with a power-law distribution, we detected a significant enrichment (hypergeometric test; $\text{pv} < 7 \times 10^{-4}$) of cancer genomes associated with the APOBEC mutational signal. We extracted the information about the association of the mutational pattern of APOBEC with different cancer types from (Alexandrov et al., 2013).

3.3. Numerical simulation of clustered mutations recapitulated the observed power law distribution in some cancer genomes

In order to be able to understand the distribution observed in the distances of mutations, we have run simulations. We hypothesize that the nucleotide substitutions with short distances (0–5 kb) is mainly due to a mutational process that generates mutations in a localized region (e.g., by the APOBEC activity), and that the mutations in the distribution of the long distances (5 kb–1 Mb) can be explained by distances of mutations belonging to two clusters of mutations generated by APOBEC activity. In this simulation, we assume that number of mutations generated by the APOBEC activity is much larger than those generated by other processes and, therefore, we ignore the second ones. We try to model a scenario where we have APOBEC binding to c different genomic positions (p_i, where $i = 1, \ldots, c$). Next, m mutations are generated in each genomic region bound by APOBEC. We assume that these mutations will occur with higher probability near the APOBEC binding position (p_i) and this probability will decay as we move away from the APOBEC binding site. For illustrative reasons, we assumed a genome of only one chromosome of length 10 Mb. We further assumed $c = 1500$ APOBEC binding regions, but similar results are obtained for other values. The position of the i-th APOBEC binding (p_i) is drawn from a uniform distribution between 0 and 10 Mb to simulate the random creation of these clusters in the genome. The position of each mutation is drawn from a normal distribution with mean p_i, and standard deviation 100 (after taking the integer part). We choose the value of 100 for the standard deviation arbitrarily, and other values for this parameter gives similar results provided that is large enough compared with the

Table 1
Number of individuals with a power-law by cancer type.

Cancer type	Number of individuals with a power-law	Number of individuals considered[a]	Total number individuals	APOBEC mutational signature[b]
Acute lymphocytic leukemia	0	1	1	Yes
Breast cancer	7	87	119	Yes
Chronic lymphocytic leukemia	1	14	28	Yes
Lung adenocarcinoma	1	24	24	Yes
B-cell lymphoma	5	19	24	Yes
Pancreas cancer	1	14	15	Yes
Acute myeloid leukemia	0	0	7	No
Liver cancer	0	84	88	No
Medulloblastoma	1	2	100	No
Pilocytic astrocytoma	0	0	101	No

[a] Only individuals with more than 50 somatic mutations in the subpopulation with distances 0–5 kb were considered.
[b] Obtained from (Alexandrov et al., 2013).

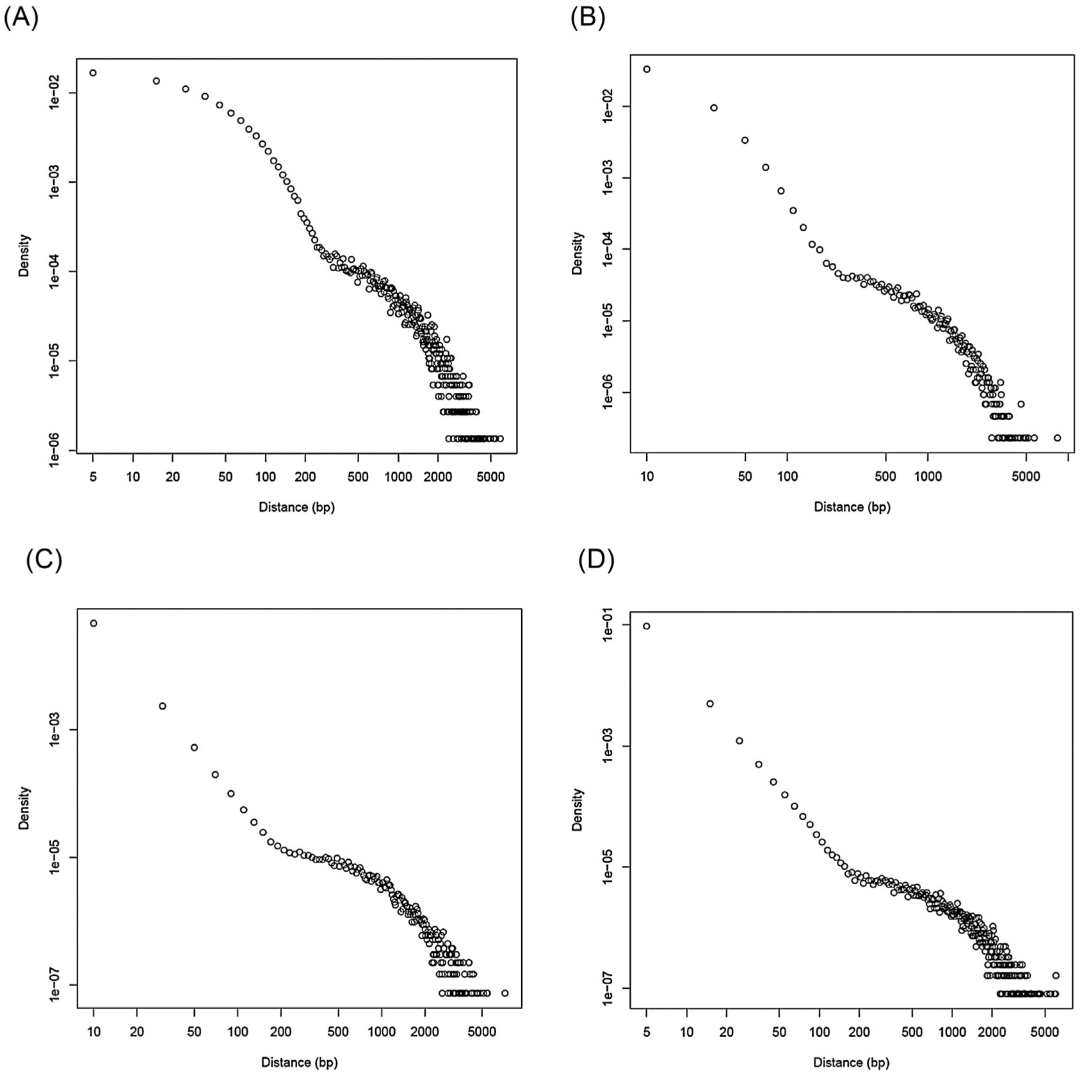

Fig. 2. Distribution of distances among consecutive mutations simulated as described in the text. We used several values for the parameter m (number of mutations per cluster): 2 (A), 5 (B), 25 (C), 50 (D).

number of mutations (m) to guarantee that the region considered is not completely saturated with mutations. Mutations located in the same position are only considered once. Other distributions than normal give us similar results (data not shown). Fig. 2 shows distributions of distances between consecutive mutations for different values of $m = 2$, 5, 25, and 50. We observed clearly two subpopulations of distances and with increasing m we observed that the distribution of short distances seems to follow a power-law.

4. Discussion

The specific mutational signatures originated by particular molecular process can help us to understand better the complexity and variation of pattern of nucleotides in genomes, and also to understand how they originated since we could identify the mutational processes that generated them. This is of particular importance in cancer, since mutations are one of the main drivers of development of cancer. To identify particular molecular processes that cause the pattern of mutations observed in specific cancer types is important in order to elaborate potential treatments that target these mutagenic processes and therefore to affect cancer development. Until now, most of the work to identify mutational signatures is based in discovering patterns of frequencies of nucleotide substitutions.

In this work, we have investigated the distribution of distances among consecutive mutations in cancer genomes, and to our surprise we have found that different cancer types show different distributions. This opens the door to use this information in the future, potentially together with frequencies of substitutions, to find mutational signatures on the DNA.

In particular, a power-law distribution has been associated in this work with the cluster of mutations (Kataegis) observed in some cancers genomes. In fact, there is not yet a quantitative definition for Kataegis, but a "rule of thumb" is used to define these regions as having six or more consecutive mutations with an average distance of less than or equal to 1 kb (e.g., (Taylor et al., 2013)). The aim of these conditions is to control the number of regions identified as Kataegis under the null hypothesis that the mutations are scattered randomly in the genome. Since the probability to find one mutation followed by five other mutations within a distance of X bp is given by $p = \text{P}(\text{Pois}(X\,n/G) \geq 5)$, where G is the length of the genome and n the number of mutations considered which may vary from about 0.001 per megabase (Mb) to more than 400 per Mb, the error of this procedure can be controlled. However, the assumption that the mutations are scattered randomly in the genome may not hold always, since different genomic regions may present very different frequency of mutations independent of any cancer process (e.g., (Arnheim and Calabrese, 2009)). An alternative method to define Kataegis independently of the overall frequency of mutation is to use the property identified in this work and to test statistically if the distances of mutations follow a power-law distribution (cancer with Kataegis) versus an exponential distribution. Future studies are needed to reveal the real biological importance of this discovery to identify Kataegis in cancer genomes, and whether other properties of this distribution may help to identify new mutational signatures.

Acknowledgements

This work was supported by the research project of the German Federal Ministry of Education and Research "CancerEpiSys". EE Kuruoglu acknowledges funding from Alexander von Humboldt Foundation in the form of Experienced Research Fellowship.

References

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.

Arnheim, N., Calabrese, P., 2009. Understanding what determines the frequency and pattern of human germline mutations. Nat. Rev. Genet. 10, 478–488.

Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-law distributions in empirical data. SIAM Rev. 51, 661–703.

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.

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.

Taylor, B.J., Nik-Zainal, S., Wu, Y.L., Stebbings, L.A., Raine, K., Campbell, P.J., Rada, C., Stratton, M.R., Neuberger, M.S., 2013. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. eLife 2, e00534.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.

Wijesinghe, P., Bhagwat, A.S., 2012. Efficient deamination of 5-methylcytosines in DNA by human APOBEC3A, but not by AID or APOBEC3G. Nucleic Acids Res. 40, 9206–9217.