Identification of a signature of evolutionarily conserved stress-induced mutagenesis in cancer

Luis H. Cisneros[1,2,3,*†], **Charles Vaske**[1,4], **Kimberly J. Bussey**[1,2,5†]

***For correspondence:** lhcisner@asu.edu (LHC)

[†]These authors contributed equally to this work

[1]NantOmics, LLC, Santa Cruz, CA; [2]The BEYOND Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ; [3]Current affiliation: Biodesign Center for Biocomputing, Security and Society, Arizona State University, Tempe, AZ; [4]Current affiliation: Claret Bioscience, Santa Cruz, CA; [5]Current affiliation: Midwestern University, Glendale, AZ.

Abstract The clustering of mutations observed in cancer cells is reminiscent of the stress-induced mutagenesis (SIM) response in bacteria. SIM employs error-prone polymerases resulting in mutations concentrated around DNA double strand breaks with an abundance that decays with genomic distance. We performed a quantitative study on single nucleotide variant calls for whole-genome sequencing data from 1950 tumors and non-inherited mutations from 129 normal samples. We introduce statistical methods to identify mutational clusters and quantify their distribution pattern. Our results show that mutations in both normal and cancer samples are indeed clustered and have shapes indicative of SIM. We found the genomic location of groups of close mutations are more likely to be prevalent across normal samples than in cancer suggesting loss of regulation over the mutational process during carcinogenesis.

Introduction

Genomic instability is a well known hallmark of cancer manifested as higher than normal rates of genomic mutations. However these mutations do not typically arise at uniformly random locations across the genome. Rather, they typically follow a non-uniform distribution resulting in mutational clustering ***Drake*** (***2007***); ***Wang et al.*** (***2007***); ***Chen et al.*** (***2009***); ***Ye et al.*** (***2010***); ***Roberts et al.*** (***2012***); ***Nik-Zainal et al.*** (***2012***); ***Alexandrov et al.*** (***2013***); ***Kamburov et al.*** (***2015***); ***Nik-Zainal et al.*** (***2016***). This phenomenon is observed in its extreme form as *kataegis*, consisting of six or more mutations with inter-mutational distances of 1 kb or less ***Alexandrov et al.*** (***2013***); ***Nik-Zainal et al.*** (***2016***).

In particular, large mutational loads in human cancer have been associated with replication repair deficiency ***Campbell et al.*** (***2017***); ***Ma et al.*** (***2018***); ***Campbell et al.*** (***2021***), and thus underlying defects in the DNA repair machinery are thought to lead to biases in the types and locations of passenger mutations and structural events acquired during the progression of cancer. These general ideas justify targeting DNA repair and checkpoint inhibitors in cancer therapies ***Murai*** (***2017***); ***Forment and O'Connor*** (***2018***); ***Ubhi and Brown*** (***2019***); ***Zhu et al.*** (***2020***). But given that most mutations are either neutral or deleterious, the likelihood that randomly distributed mutations would result in gains in fitness is considered to be low ***Ram and Hadany*** (***2014***), whereas concerted patches of mutation, particularly when occurring within specific genes, could lead to neo-functionalization

and increased cellular fitness ***Drake*** (***2007***); ***Ram and Hadany*** (***2014***); ***Cortés-Ciriano et al.*** (***2020***).

Previous work has shown that even though cancer samples typically exhibit a lot more mutations outside of genes, clustered mutations are enriched in genes relative to the intergenic spaces ***Cisneros et al.*** (***2017***); ***Supek and Lehner*** (***2017***). In particular, mutation clustering in non-coding regions have been associated with structural changes that possibly cause elevated mutation rates but by themselves very rarely constitute driver mutations ***Nik-Zainal et al.*** (***2016***); ***Rheinbay et al.*** (***2020***).

Other studies have identified the action of the AID/APOBEC family of cytosine deaminases as well as the action of Pol-η as contributing mechanisms to the phenomenon of mutational clustering ***Lada et al.*** (***2012***); ***Roberts et al.*** (***2013***); ***Taylor et al.*** (***2013***); ***Supek and Lehner*** (***2017***); ***Buisson et al.*** (***2019***); ***Roper et al.*** (***2019***); ***Shi et al.*** (***2020***). However these processes only explain a subset of the mutational clusters observed and thus a more general mechanism remains to be determined.

Stress-induced mutagenesis (SIM) in bacterial occurs when DNA damage happens in the context of additional cellular stress sufficient to initiate the SOS response ***McKenzie et al.*** (***2000***); ***Foster*** (***2007***); ***Janion*** (***2008***); ***Shee et al.*** (***2012***); ***Rosenberg et al.*** (***2012***). SIM has been shown to increase the mutation rates locally around DNA lesions as cells strive to adapt to the challenging environment ***Foster*** (***2007***); ***Rosenberg et al.*** (***2012***); ***Fitzgerald et al.*** (***2017***). In the course of double-strand-break-mediated mutagenesis in bacteria, DNA repair switches from high-fidelity homologous recombination to a repair mechanism that relies on the error-prone DNA polymerase Pol IV, encoded by the gene dinB. The result of this mechanism is a spectrum of both single nucleotide variants (SNV) and copy number amplifications. The molecular signature of this process is a clustering of SNVs around the site of the double strand break (DSB) spanning hundreds of kilobases in size and with a decaying probability of mutation as a function of the distance from the DSB that remains above background for up to a megabase ***Shee et al.*** (***2012***); ***Fitzgerald et al.*** (***2017***). The molecular fingerprint associated with stress-induced mutagenesis manifests as a random probability distribution of SNVs centered on a putative DSB and with a decay distance of about two hundred kilobases ***Rosenberg et al.*** (***2012***).

The evidence of clustering in cancer coupled with the known intra-tumor chromosomal structural heterogeneity that characterizes many cancers ***Roschke et al.*** (***2002***, 2003, 2005) prompted us to inquire into a comparable process to bacterial stress-induced mutagenesis happening during carcinogenesis, an idea that has been previously suggested by Fitzgerald, Rosenberg and colleagues ***Fitzgerald et al.*** (***2017***); ***Xia et al.*** (***2019***). Adaptive mutagenesis has been recently shown in the context of the emergence of drug resistance, with evidence of down-regulation of mismatch repair (MMR) and homologous recombination (HR), and up-regulation of error-prone polymerases in drug-tolerant colorectal tumor cells ***Russo et al.*** (***2019***). Furthermore an mTOR stress signaling has been shown to facilitate SIM in multiple human cancer cell lines exposed to non-genotoxic drug selection ***Cipponi et al.*** (***2020***).

We investigated SNV distributions observed by whole genome sequencing of non-inherited mutations in normal samples and a wide variety of solid tumors. We found clear evidence of mutational clustering as demonstrated by enrichment of closer-than-expected mutations, particularly for samples with low mutational loads. Additionally, by characterizing the distributions of clusters we observed that there is a greater consistency of cluster locations across normal samples than in cancer samples, suggesting a degree of regulation control for mutations in normal tissue that breaks down during carcinogenesis. Finally, we identified the molecular signal of SIM in the SNV distributions of clustered mutations and showed a relationship with clinical outcome.

Results

Variant distribution is not uniform

We analyzed the patterns of mutational density across the genome in non-inherited mutations from 129 normal individuals (CGI data) as well as somatic mutations in 1950 tumors from 14 differ-

ent tissues (PCAWG data) (see section 1 for details), and compared them with simulated patterns of $N_{\text{SNV}} = 1000, 2500, 5000, 10000, 25000$ and 50000 total uniformly distributed mutations (500 replicates each).

First, we measured the distribution of inter-SNV distances x as a function of the total number of mutations. This is, for each sample we find the number of segments with length x inside each 15 kb bin up to 150 kb, and plot against the total mutational load of the sample (Fig.1).

In both normal samples, Fig.1(B), and cancer samples, Fig.1(C), short intervals are more frequently observed than expected from a theoretical null model (see section 1), thus revealing a tendency of mutations to cluster together in genomic space. The effect is considerably stronger for lower values of the mutational load, particularly with $N_{\text{SNV}} < 3000$ where the numbers of short segments ($x \leq 15$ kb) can be over an order of magnitude larger than expected. On the other hand longer interval distances are progressively less over-represented: for $x \sim 75$ kb they have appear at about the expected frequency.

As the total number of mutations increases we observe a drop in the theoretical prediction of numbers of inter-SNV segments. This drop is due to a saturation effect; as the number of uniformly distributed SNVs goes over 100,000, the expected inter-event distance in a 3 billion base genome would be under 30kb, and thus long intervals become more and more unlikely. Interestingly, in cancer samples the over-representation of small segments is prevalent even for large values of mutational loads. This effect is compensated with an under-representation of moderate to large length intervals, yet for very large mutational loads ($N_{\text{SNV}} > 100,000$) long intervals are also more frequent than expected. This suggests that there are regions of the genome somehow avoided by mutations, manifesting in the form of unexpected long conserved, or protected, regions. These features are consistent across all samples and is evidently not associated to number fluctuations, since the dispersion in 500 simulated replicates cannot account for it (Fig.1(A)). From this analysis we conclude mutations in both normal and cancer samples tend to form groups.

We then looked at the number of groups, with "group" defined as a set of contiguous SNVs with inter-SNV distances $x \leq D^\star$ ($D^\star = 15$ kb) as a function of the mutational load. We deemed these groups *tuples*, while a *singleton* (i.e. a mutation that is not grouped) is simply a tuple of size $n = 1$, .

The numbers of tuples and singleton variations for simulated, normal and cancer data are shown in Fig. 2(A). The most salient feature is that the frequency of singletons are significantly underrepresented for low mutational loads, while tuples are typically over-represented with respect to the theoretical expectation (Poisson point process model). The number of tuples is particularly high for samples with $N_{\text{SNV}} < 3000$, a mutational load for which a uniformly random process would very rarely lead to any proximal mutations. Namely, at $N_{\text{SNV}} \sim 1000$ SNVs only a handful of tuples are expected yet dozens to hundreds are typically observed in cancer samples. And interesting conclusion of the observation that many groups form is that, though mutations tend to cluster together, they don't do it as a large scale condensate, but rather in many small clusters. In particular, mutations in normal samples seem to mainly aggregate in groups of 2 or 3 while mutations in cancer tend to cluster is groups of a much wider size spectrum (see Appendix section 1 for detailed distributions for different tuple sizes). On the other hand, as the total number of mutations increases the distributions approached the predicted curve, but then departed again. For large mutational loads the relationship between the proportion of tuples and singletons with respect to the expected behavior is inverted, supporting the idea that certain regions in the genome are protected from accumulation of mutations as singletons become very rare.

All together these observations demonstrate that the mutational rate is an heterogeneous property of the genome, and thus likely a regulated or constrained process. We hypothesize two, non-mutually exclusive basic ways to generate these kinds of mutational patterns:

1. The mutational process is modulated, such that it can be modeled as non-uniform Poisson process with a location-dependent rate $\lambda(x)$ at location x. This modulation is an effect of differential DNA repair efficiencies along the genome, conservation or protection of certain

genomic regions due to topological or folding/packing molecular properties and other sequence or location dependent processes.

2. Mutations are inter-dependent events, entailing either nucleation of mutations during subsequent DNA replications (i.e. mutations induce new errors) or a process in which events happen together as a single burst of proximal mutations.

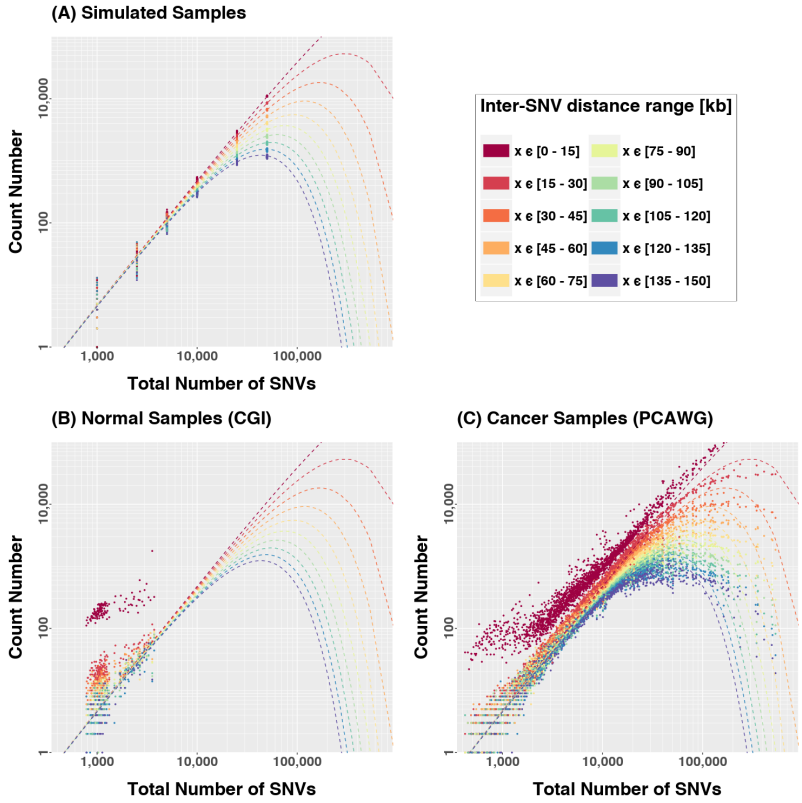


Figure 1. Number of inter-SNV segments of different lengths as a function of the mutational load in (A) Simulated, (B) Normal and (C) Cancer samples. Dashed lines are the theoretical predictions for a Poisson point process. Both normal and cancer cases show significant enrichment of small segments indicating that mutations are typically closer than expected.

Now, if the mutational process is dependent on genomic location, then tuples would tend to occur at the same places across samples. We compared tuple distributions across samples for which tuple enrichment was most obvious (Fig. 2(A)): N_s=129 normal samples and N_s=784 cancer samples have $N_{\text{SNV}} < 5100$. We identified all regions in the genome containing tuples in at least $\sqrt{N_s}$ of the samples (corresponding to 8.8% of the normal samples and 3.5% of the cancer samples), providing confidence that our measurement is above the Poisson-counting statistical noise. For normal samples (Fig. 2(B)) we found 128 regions with an overlap going as high as 30%. These regions were no longer than 30 kb and about a quarter of them were single base locations repeatedly mutated in several samples. Many of these regions were close together rendering more than 50 coarse-grained ranges as shown in Fig. 2(B). In contrast, cancer samples had few overlaps. We observed 19 susceptible regions that coarse-grained to 5 distinct ranges (see Fig. 2(B)): a $\sim$ 117 kb region in chromosome 6, associated to the human leukocyte antigen (HLA) complex, which contained tuples in up to 7% of the samples, two $\sim$ 1 kb regions in chromosomes 2 and 3, a single point mutation in chromosome 1 overlapping in $\sim$ 4% of the samples which is associated with the zinc finger protein ZNF678 and a half kilobase region in chromosome Y with 11% overlap in the 479 male samples. These results suggest that both processes of non-uniform mutation could be at play. The close proximity of mutations in cancer seems to be driven by a process in which events are not necessarily independent from each other, perhaps occurring simultaneously, yet otherwise distributed

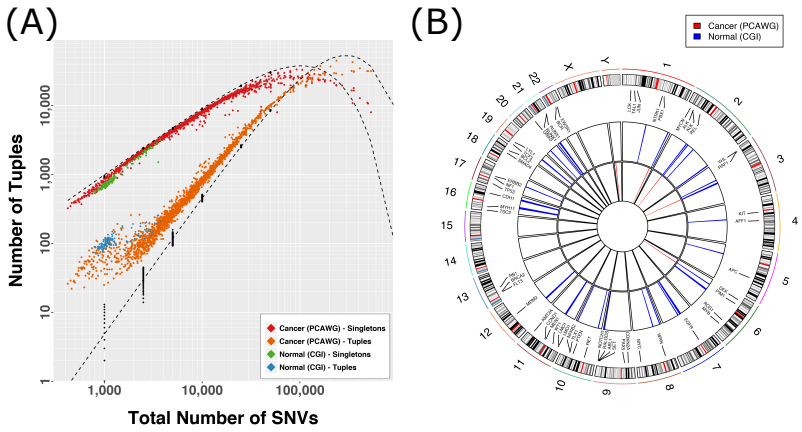


Figure 2. (A) Observed number of tuples and singletons as a function of the total mutational load. A tuple is a set of consecutive mutations with inter-event distance $x \leq 15$ kb. A singleton is a mutation farther than 15 kb from any other mutation (1-tuple). Black dots are simulated data, dashed lines are the expected curves according to Poisson statistics. (B)Susceptible regions for $N_{\text{SNV}} < 5100$. In normal samples (blue) and cancer samples (red) all regions that are part of a tuple in at least 8.8% of the normal samples and 3.5% of the cancer samples, based on the square root of the number of samples. These regions are evidently more common in normal than in cancer samples.

mostly randomly across the genome and therefore not showing large overlap between samples. In contrast, non-inherited mutations in normal tissue appear at least partially driven by a location-specific and/or sequence-specific process, quite possibly sculpted by evolution and regulated across the genome.

Quantification of Cluster Shapes

Previous work demonstrates that SNVs in both normal tissues and cancer cluster together and the sequence context of both the reference and mutant calls can be used to infer mechanism ***Roberts et al.*** (***2012***). The association of APOBEC cytosine deaminases with clusters is well established ***Lada et al.*** (***2012***); ***Burns et al.*** (***2013***); ***Taylor et al.*** (***2013***); ***Roberts et al.*** (***2013***), but it accounts for at most 50% of the clusters observed ***Roberts et al.*** (***2013***). Furthermore, there is nothing about the mechanism of APOBEC that would suggest a characteristic shape of the clusters. In contrast, the stress-induced mutational response of bacteria, mediated by Pol IV and encoded by dinB, leads to a clustering pattern with a characteristic cluster shape where the number of SNVs in the center of the cluster will be greater than those found at the edges ***Shee et al.*** (***2012***). Therefore, we looked at how both the number and shape of clusters, defined as statistically unlikely tuples of size $n \geq 3$ (see 1), varied with total mutational load among non-inherited mutations or somatic mutations in cancer.

Data simulated under the null hypothesis of uniform random mutation showed that as the total number of SNVs increases, we expect to see the number of clusters and the fraction of SNVs in those clusters increase (Fig. 3(A)-(B)). We note that, in agreement with observations presented above, no clusters were observed in simulated data with mutational loads $N_{\text{SNV}} < 2500$, and a mean of only four clusters per genome was detected in samples with 2500 mutations. This indicates that under the null hypothesis at least several thousand mutations are required to observe any measurable clustering. In contrast, both non-inherited mutations in normal tissue and somatic mutations in cancer show extensive clustering when the mutational burden is this low (Fig 3(A)). On the other hand, for very large numbers of mutations we observed a sudden plateau in the number of clusters. Again, saturation is expected as in a genome 3 billion bases long 100,000 mutations yields an average inter-mutation distance of ~ 30 kb, tuples would not be unlikely and the thus number of statistically significant clusters would drop.

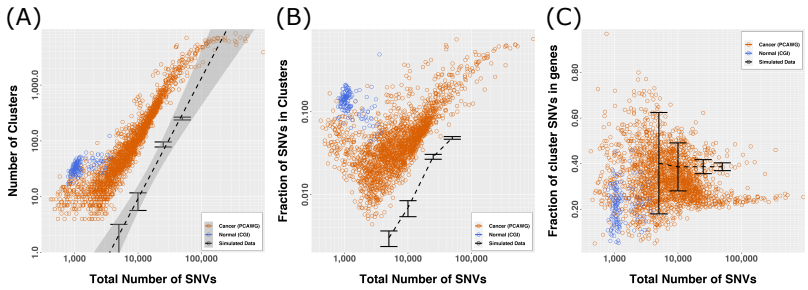


Figure 3. (A) Number of Clusters per sample as a function of the mutational load. Dashed line is the best fit for simulated data. Clustering is clearly larger than expected. The plateau at large N_{SNV} is related to the limit at which the average mutation distance in a background of 3 billion bases approaches 30 kb, which produces many statistically likely tuples and thus less clusters. (B) Fraction of SNVs in clusters versus N_{SNV}. The fraction of SNV in clusters increases with the number of mutations, suggesting that as more mutations are accumulated in the genome they are preferentially placed in clusters. (C) Fraction of SNVs in clusters in genes versus N_{SNV}, showing that cluster do not typically overlap genes for high N_{SNV}.

Figure 3(B) shows that as the number of mutations increase in cancer samples, the fraction of SNV in clusters increased as well. This suggests that as more mutations are accumulated, a larger fraction of those mutations are preferentially placed in clusters. Thus the clustering process itself could be implicated in the mechanism driving cancer mutations in some form of positive feedback loop or nucleation process. Another interesting observation is that the fraction of SNVs in cluster is very high for normal samples as compared to cancer samples with the same mutational load, and a load for which we don't expect any clustering under the null hypothesis. This indicates that the mutational process in normal samples is in fact driven by a mechanism that favors close proximity of variations, and is likely restricted to susceptible genomic regions as suggested by Fig.2(B).

When we looked at whether the SNVs are found in genes or in intergenic regions, the null hypothesis predicts as the number of SNVs increases, the proportion of SNVs located in genes converges to about 40% and remains constant (Fig.3(C)). Private non-inherited (Normal) mutations converged to about 37% (range 31.3%-42.8%) of SNVs localizing within genes, while cancer was defined by a large amount of variability that converged to about 25% being located in genes.

The clustering behavior for the cell-line set is somewhat consistent with cancer samples, but more intense: it shows more clusters, a larger fraction of SNVs in clusters and a rather low fraction of them localized in genes. In all cases these samples are equivalent to the most extremes cases in the cancer set in terms of clustering for the same mutational burden.

In agreement with our previous study ***Cisneros et al.*** (***2017***), SNVs in clusters in cancer are preferentially excluded from genes (Fisher's exact, Odds Ratio (OR) = 0.6002, 95% CI =0.5992-0.6013, p-value $< 2.2\mathrm{x}10^{16}$). When we looked specifically at the position of clusters within genes by counting SNVs that are in genes versus those that are not, we observed a slight enrichment for SNVs in clusters to be in the 3'-end of genes compared to SNVs that are not in clusters (Fisher's exact, OR = 1.024, 95% CI = 1.021-1.027, p-value $=3.067\mathrm{x}10^{-57}$), confirming the observations of Supek and Lehner ***Supek and Lehner*** (***2015***, 2017).

To evaluate the shape of the clusters, we introduce the Stress-Introduced Heterogeneity (SItH) score (see Methods and Fig. 6). The SItH score was computed both on individual clusters (cluster SItH) and over all clusters in a tumor (overall SItH). In simulated data, increasing the number of SNVs led to decreasing overall SItH scores (Fig. 4(A)) and produced a sigmoid shape for the variability in cluster SItH measured by the inner-quartile rage (IQR) (Fig. 4(B)). The overall SItH score was higher in cancer than in normal samples. In comparison to the simulated data, overall SItH scores were larger at the extremes of mutational burden and lower in the mid-range of total SNV count (Fig. 4(A)). Moreover, cancer samples showed a greater diversity of SITH scores than predicted under the assumption of random uniform mutations or compared to normal (Fig. 4(B)). Interestingly, the

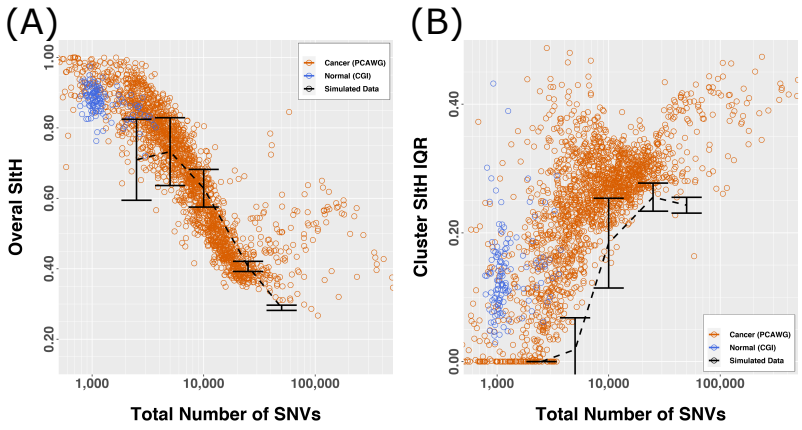


Figure 4. SItH scores by number of SNvs. (A) Overall SItH score as a function to the mutational load.(B) Inner-quartile range (IQR) of Cluster SItH scores as a function of the mutational load.

diversity of cluster shapes reaches a plateau at mutational loads corresponding to higher than expected overall SItH scores. Combined, these results suggest the mutational clustering in cancer is complex and likely driven by multiple mechanisms simultaneously.

Survival Analysis

A key characteristic of SNV clusters that result from stress induced mutation mechanisms is a decay in the frequency of incidental SNVs as a function of distance from the DSB that triggered error-prone repair response ***Shee et al.*** (***2012***). We postulated that a more positive overall SItH score reflects a greater contribution of the adaptive mutation process to the mutational landscape of the tumor. Therefore SItH provides a measure of the evolutionary response, or the adaptive capacity, of a tumor to a source of stress such as chemotherapy. Overall SItH scores ranged from 0.145 to 0.999 (Fig. 4(A)) and varied significantly by organ site and whether the tumor was one of multiple tumors from a single donor (ANOVA, organ, F = 136.70, $p< 2.2x10^{-16}$; multiple tumor, F = 3.07, p = 0.0799; Maximum SITH Score, F = 16.14, $p = 6.098x10^{-5}$).

To determine the relationship between SItH scores and clinical outcome, we conducted Cox Proportional Hazard analysis of both overall SItH score as well as the inter-quartile range of the cluster SItH. The models are specified as follows:

$$\text{Overall Survival} \sim \text{SItH} + \text{multiple.tumor} + \text{is.Max.SItH} + \text{strata(Organ)} \quad (1)$$

where the data analyzed were either primary tumors or the group of metastases and recurrences. For inter-quartile range of the cluster SItH the model is:

$$\text{Overall Survival} \sim \text{SItH IQR} + \text{multiple.tumor} + \text{is.Max.SItH} + \text{strata(Organ)} \quad (2)$$

After controlling for organ site and multiple tumor status, we found overall SItH scores predict patient survival but with different effects depending on whether the sample was a primary tumor or from a metastasis or recurrence. In primary tumors, more positive overall SItH scores predicted better patient survival (Cox Proportional Hazard Regression (CPHR), Hazard Ratio (HR) = 0.4516, 95% CI: 0.2274 -0.8968, p=0.0231, see Supplemental Fig. 5). However, when the recurrences and metastatic tumors were considered as a group, the overall SItH score predicted a worse survival, with a HR of 14.84 (CPHR, 95% CI: 1.934-113.876, p= 0.00947). When we looked at the diversity of SItH scores on a cluster basis, the type of tumor sample was no longer relevant. The inner quartile range (IQR) of cluster-level SItH scores associated with worse survival, with a HR of 5.744 (CPHR, 95% CI: 1.824 -18.09, p= 0.00283). We then asked whether there was a difference in survival between patients with SItH IQRs above or below the median SItH IQR, as clinical translation will

likely require a creating a cut-off value above which one would predict poor prognosis. As is seen in Fig. 5, there is a significant difference in survival, even after accounting for the baseline differences in survival by tissue of origin (CPHR, HR = 1.26, 95% CI: 1.043-1.531, p= 0.0168).

Effects of the maximum inter-SNV distance in the definition of clusters

Our definition of a cluster is a tuple with a probability of less than 1% as measured by a negative binomial test. The test is conducted so that a given tuple might not satisfy the second condition but part of it might (e.g. one with a higher concentration of mutations in one end). In this case only that portion of the tuple is called a cluster. The specific value $D^\star = 15$ kb was chosen because it's a good balance between signal and noise: (a) if $D^\star$ is too small, very few clusters are found unless the total number of mutations is very large. Even though less clusters are found the restrictive condition given by $D^\star$ yields more concentrated clusters with small dispersion as measured by the SItH IQR. (b) On the other hand if $D^\star$ is too large many clusters are found and the noise level is larger. In this case there is more room for different configurations of clusters, producing larger values of the IQR. Low to moderate mutational loads typically have smaller SItH scores since clusters tend to be more uniform (i.e. less peaked), but samples with large mutational loads exhibit saturation effects that limit the number of clusters (i.e if tuples are common then they are not clusters). There is therefore a trade-off between the effects on smaller and larger mutational loads. In order to find a good signal-noise balance we ran our analysis with 8 different values of $D^\star$. Tables 1 and 2 show the correlations between the overall SItH score and SItH IQR in cancer samples. Based on these results we conclude that $D^\star = 15$ kb is a good choice: result values are well correlated with cases in both ends, indicating that this parameter captures well the signal for both small and large mutational loads without too much compromise on the quality.

SItH	1kb	2kb	5kb	10kb	15kb	20kb	25kb	50kb
1kb	1	0.988498379	0.953254764	0.913062667	0.885251282	0.872055507	0.86602581	0.890076618
2kb	0.988498379	1	0.979158317	0.944003381	0.914886845	0.896847418	0.884576285	0.885092543
5kb	0.953254764	0.979158317	1	0.981798123	0.95732091	0.937928221	0.92095692	0.885284312
10kb	0.913062667	0.944003381	0.981798123	1	0.988106173	0.973637963	0.957521878	0.900404232
15kb	0.885251282	0.914886845	0.95732091	0.988106173	1	0.992613883	0.980773228	0.919694508
20kb	0.872055507	0.896847418	0.937928221	0.973637963	0.992613883	1	0.993667498	0.939087871
25kb	0.86602581	0.884576285	0.92095692	0.957521878	0.980773228	0.993667498	1	0.957075149
50kb	0.890076618	0.885092543	0.885284312	0.900404232	0.919694508	0.939087871	0.957075149	1

Table 1. Correlation of overall SItH scores for different $D^\star$ values with PCAWG data.

SItH IQR	1kb	2kb	5kb	10kb	15kb	20kb	25kb	50kb
1kb	1	0.938197026	0.838362333	0.78391552	0.737391598	0.688942755	0.65370155	0.445513598
2kb	0.938197026	1	0.906115983	0.830270283	0.774522805	0.721045756	0.680923471	0.454410934
5kb	0.838362333	0.906115983	1	0.921681646	0.848406024	0.784003812	0.736097232	0.493501676
10kb	0.78391552	0.830270283	0.921681646	1	0.933918906	0.868093155	0.815392267	0.584146019
15kb	0.737391598	0.774522805	0.848406024	0.933918906	1	0.941611092	0.882449786	0.664941136
20kb	0.688942755	0.721045756	0.784003812	0.868093155	0.941611092	1	0.944940353	0.736770549
25kb	0.65370155	0.680923471	0.736097232	0.815392267	0.882449786	0.944940353	1	0.789022886
50kb	0.445513598	0.454410934	0.493501676	0.584146019	0.664941136	0.736770549	0.789022886	1

Table 2. Correlation of SItH IQR scores for different $D^\star$ values with PCAWG data.

Discussion

Our study provides evidence that a signature of stress-induced mutagenesis, characterized by clusters of SNVs with a defined geometry, is widespread across multiple cancer types. Furthermore,

(A)

(B)

Figure 5. Survival difference based on SItH score IQR being above or below the median SItH score IQR. a) Kaplan-Meir curves for tumors with cluster-level SItH IQR above and below the median SItH IQR for 1895 tumors. b) Results from the Cox Proportional Hazard analysis. Survival data from 1950 tumors, of which 1201 samples had SItH IQR scores in 14 different cancer types were used. Hazard ratio for IQR group was controlled for by multiple tumors, maximum IQR value and tissue of origin.

the association of both overall cluster shape and increased cluster shape variability with patient survival suggests stress-induced mutation has a clinical impact. Both the overall SItH value and the heterogeneity represented by the SItH IQR are likely derived from a combination of the strength of stress-induced mutation as a mutational process within a tumor and the clonal diversity of the tumor, both of which would be expected to impact disease outcome ***Andor et al.*** (***2016***). The relationship between overall SItH and SItH IQR with respect to survival suggests cluster heterogeneity predominantly represents a combination of the amount of time stress-induced mutagenesis has been active during carcinogenesis and clonal heterogeneity, while overall SItH represents the ratio of stress-induced mutagenesis relative to other mutational processes. Our work showed an increase in mutational load leads to both increasing cluster sizes as well the percentage of SNVs involved in clusters, but only up to a point. In tumors with high mutational burdens, the number of clusters, the genomic distance covered by cluster, and the number of SNVs contained within a cluster plateau. This implies that under high mutational burden the variations in mutation density across the genome flatten out, likely due to alterations in DNA repair pathways such as a loss of mismatch repair ***Supek and Lehner*** (***2017***); ***Campbell et al.*** (***2017***), and obscure the detection of clusters.

The influence of intra-tumor diversity on clinical outcome is an area of active investigation. Evidence from measures of clonal diversity and copy number diversity are associated with worse outcome and therapeutic response ***Andor et al.*** (***2016***); ***Davoli et al.*** (***2017***); ***Roh et al.*** (***2017***); ***Dagogo-Jack and Shaw*** (***2017***); ***Turajlic et al.*** (***2019***); ***Ben-David and Amon*** (***2019***). However, cancer must balance the introduction of genomic rearrangements that contribute to cellular diversity with a sufficient level of genome stability to avoid a genomic error catastrophe. Our results are consistent with this notion in that very large positive overall SItH scores associates with better patient survival. The SItH IQR represents a measure of mutational heterogeneity that ties intra-tumor diversity to a mutational process underlying an evolutionarily conserved response to cellular stress. The diversity measured by the SItH IQR is a measure of the heterogeneity of adaptive strategies within a patient. This diversity manifests as a broader ensemble of mutational cluster shapes within a tumor driven by the heterogeneity in mutational processes to generate genomic diversification. This in turn increases the substrates available for broad phenotypic plasticity, including transcriptional responses. Such responses have been shown to be important in the rapid acquisition of resistance to doxorubicin ***Wu et al.*** (***2015***). In this case high diversity becomes a direct survival advantage for the tumor, allowing it to respond to a wider range of stresses and leading to a poor outcome for patients.

Others have found clustered mutations and proposed mechanisms for them ***Roberts et al.*** (***2012***); ***Lada et al.*** (***2012***); ***Burns et al.*** (***2013***); ***Taylor et al.*** (***2013***); ***Roberts et al.*** (***2013***); ***Supek and Lehner*** (***2017***). Our definition of mutational clusters spanning over kilobase distances***Shee et al.*** (***2012***); ***Fitzgerald et al.*** (***2017***) is broader than that of Supek and Lehner who showed Pol-η, a TLS polymerase closely related to Pol IV, is involved in the generation of clustered mutations that preferentially locate to the 3'-end of active genes***Supek and Lehner*** (***2015***). However, we were able to confirm that key finding with our cluster definition.

An open question that remains is whether the clusters we and others detect arise from single events reflective of bursts of mutational activity or are accumulated over time, therefore marking regions of the genome prone to mutation. Allele fraction has been suggested as one way to address this question. However, the precision of most allele fraction measurements prevents the accurate discrimination of varying degrees of heterogeneity across a tumor. For example, the 95/95 binomial tolerance interval for a true allele fraction of 0.5 at a read depth of 60x ranges from 0.25 to 0.75 (see Appendix). This interval represents the bounds in which we are 95% confident that 95% of the measurements of a true allele fraction of 0.5 will lie. If we have a cluster where the allele fractions of the SNVs all fall within this range, we cannot rule out they actually represent a true allele fraction of 0.5 and therefore all come from the same event. Experimental evidence in mammalian systems leading to cluster formation is necessary to answer this question. This is an

important study to pursue as the strategies one might propose for influencing mutational patterns with impact on clinical outcomes will depend on whether the target is the mutational process itself or the regions of the genome being acted upon by the mutational process.

Conclusions

Cancer is notorious for outsmarting the physician. To make progress we need to factor in how cancer cells evolve and adapt in the face of treatment challenges. Understanding the mechanisms of mutation and adaptation in cancer is therefore an essential pre-requisite for improving patient outcomes. Stress-induced mutagenesis, an ancient and evolutionarily conserved adaptive mutation mechanism well-characterized in E. coli, underlies in part the genomic instability seen in cancer and contributes to the ability of the tumor to evolve resistance to therapy ***Fitzgerald et al.*** (***2017***). We have described a way to quantify this antagonism and shown that SIM has a strong association with poor prognosis. Further investigations into the process of SIM in cancer should lead to better patient outcomes by giving clinicians a measure allowing them to tailor treatment that checks tumor progression while minimizing the risk of triggering an aggressive evolutionary response.

Methods

Null model - Uniform random mutations

A uniform distribution of point mutations can be modeled as a Bernoulli process. Because the total number of mutations is typically much smaller than the number of nucleobases in the genome the expected inter-event distance can be approximated as an exponential function. This distribution is equivalent to the expression for the waiting time distribution in a Poisson process ***Cinlar*** (***1975***) or the survival density function in a constant hazard process ***Moore*** (***2016***).

The probability of observing an inter-event distance x is given by the density function:

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \tag{3}$$

Where $\lambda = \mathrm{N_{SNV}}/L$ is the mutational rate, $\mathrm{N_{SNV}}$ the total number of SNV mutations and L the length of the genome. The probability of $x \leq D$, or *cumulative waiting time function*, is

$$F(D) = \int_0^D f(x)\mathrm{d}x = 1 - \mathrm{e}^{-\lambda D} \tag{4}$$

and the probability of $x > D$, or *survival function*, is $S(D) = 1 - F(D) = \mathrm{e}^{-\lambda D}$.

The probability associated with the range of interval lengths $[d_i, d_i + D_{\mathrm{bin}}]$ is

$$P_i = F(d_i + D_{\mathrm{bin}}) - F(d_i) = \mathrm{e}^{-\lambda d_i} \cdot \left(1 - \mathrm{e}^{-\lambda D_{\mathrm{bin}}}\right) = S(d_i) \cdot F(D_{\mathrm{bin}}) \tag{5}$$

And the expected number of intervals with length in this range is $E_i = \mathrm{N_{SNV}} \cdot P_i$.

We define a n-tuple as a set of n consecutive mutations that are closer than $D^\star = 15$ kb. By definition all tuples are separated by intervals $x > D^\star$ from each other. In particular 1-tuples, or *singletons*, are those SNVs that are farther than $D^\star$ bases from its closest neighbors.

From eq. (5) the probability of $x \leq D^\star$ is $F^\star = 1 - \mathrm{e}^{-\lambda D^\star}$ and the probability of $x > D^\star$ is $S^\star = \mathrm{e}^{-\lambda D^\star}$.

The expected total number of tuples can be estimated as:

$$T^\star = \mathrm{N_{SNV}} \cdot S^\star = \mathrm{N_{SNV}} \cdot \mathrm{e}^{-\lambda D^\star} \tag{6}$$

The total number of mutation events in tuples can be estimated as $E^\star \sim \mathrm{N_{SNV}} \cdot F^\star$, where the identity is not exact because of edge effects; for instance, location differences are calculated on each chromosome (except Y) independently and therefore the total number of inter-mutation intervals is $\mathrm{N_{SNV}} - 25$.

Following these definitions, the probability of observing a n-tuple can be written as the combination of probabilities:

$$P^\star(n) = (S^\star)^2 (F^\star)^{n-1} = \mathrm{e}^{-2\lambda D^\star} \cdot \left(1 - \mathrm{e}^{-\lambda D^\star}\right)^{n-1} \tag{7}$$

Thus the expected number of n-tuples is $N^\star(n) = \mathrm{N_{SNV}} \cdot P^\star(n)$ and the probability mass function of n-tuples is

$$P_n = \frac{N^\star(n)}{T^\star} = S^\star(F^\star)^{n-1} = \mathrm{e}^{-\lambda D^\star}\left(1 - \mathrm{e}^{-\lambda D^\star}\right)^{n-1}, \tag{8}$$

which is equivalent to the binomial mass function of the first order $P_r(1, s, \lambda)$.

Data

We obtained variant calls for normal and cancer from public repositories where all cases had been called by a standard pipeline. For non-inherited mutations in normal tissue, we used WGS data from the Complete Genomics Indices database in the 1000 Genome Project ***The 1000 Genomes Project Consortium*** (***2015***)(release 20130502, see Supplementary Materials Table 7 in Cisneros, et al. ***Cisneros et al.*** (***2017***) for a list of donors). This data has average genome coverage of $47X$. The VCFs of 129 trios were analyzed using the `vcf_contrast` function from the `VCFTools` analysis toolbox to compare each child with the two corresponding parents. The resulting potential novel variants were then filtered such that the child and both parents must be flagged as `PASS` (the variant passed all filters in the calling algorithm); the child must have a read depth of at least 20; and the alternative (aka novel) allele frequency was ≥ 0.35. For cancer, we analyzed the simple somatic mutations and corresponding clinical data from the PCAWG coordinated WGS calls for 1950 tumor samples from 1830 donors representing 14 different primary sites ***Campbell et al.*** (***2017***). Somatic variants for all data sets were classified as previously published ***Cisneros et al.*** (***2017***).

We generated 500 sample replicates for eight groups of simulated data defined by their total mutational load ($N_{\mathrm{SNV}} = 500; 1000; 2500; 5000; 10,000; 25,000; 50,000; 100,000$). We modeled a uniform, random distribution of SNVs across the genome as a one-dimensional Bernoulli Process, corresponding to our null hypothesis. The number of events in a region of size X is a random variable with a probability mass function that can be approximated as a Poisson distribution:

$$P(n) = \frac{(X \cdot \lambda)^n}{n!}\mathrm{e}^{-(X \cdot \lambda)} \tag{9}$$

with $\lambda = (N_{\mathrm{SNV}}/L)$ the total mutational rate and L the genome length.

In order to characterize the clustering of genomic mutations we defined a tuple as a set of consecutive SNVs such that the inter-event distance $x < D^\star = 15$ kb for all event pairs in it. According with Poisson statistics (per equation 8) the expected number of n-tuples in a sample with N_{SNV} mutation is given as

$$N_{\mathrm{T}}(n) = N_{\mathrm{SNV}}\left(1 - \mathrm{e}^{-\lambda D^\star}\right)^{n-1}\mathrm{e}^{-\lambda D^\star}. \tag{10}$$

Detection of mutation clusters

A group of SNVs is deemed a "cluster" if it is a tuple of at least 3 variations and the probability of finding it by chance is less than 1% according to the negative binomial regression given by the total rate of observed mutations in the genome. In other words, the particular group of variations is statistically unlikely to happen in the background given by the mutational load of the sample. For each WGS sample in our database, all possible clusters were identified and the "center of mass" (genomic location of cluster centroid) in each case is calculated along with other properties like start and end locations, length and size (number of variations) ***Cisneros et al.*** (***2017***).

Detection of cluster shape

We treated cluster centroids as likely locations of the DSBs that induced the accumulation of variations. Therefore, the expected signature for stress-induced mutagenesis should be evident as a concentration of mutations around these centroids that decays with distance. Thus, for each cluster i we computed the cumulative distribution of SNV events $F_i(X)$, as a function of the distance X from the cluster centroid up to 250 kb and in both the $3'$ and the $5'$ directions. By aggregating together all cumulative distributions observed in each sample we generated a representative overall curve $F(X) = \sum F_i(X)$ that conveys the probability of finding a mutation at a given distance from

a cluster center. If the distribution of SNV events were uniformly random (and therefore do not typically decay) then $F(X)$ is expected to increase proportionally with X. This assumption gives us a background of mutations against to which we can compare the observed distribution pattern. It is important to note that this definition is itself independent of the definition of clusters. By construction, if the background distribution is uniform as assumed, then we should not observe clusters at all since they are statistically unlikely by random chance. In order to define a useful score, we normalize X by 250 kb and F by the number of events closer than 250 kb, thus mapping all cluster-associated cumulative distribution curves to a unit box:

$$\frac{X}{250\text{kb}} \to x \quad ; \quad x \in [0,1]$$

$$\frac{F(X)}{F(250\text{kb})} \to f(x) \quad ; \quad f(x) \in [0,1]$$

If the null hypothesis were correct for these events, $f(x) = x$. We define a measure of the degree of deviation from the null hypothesis by integrating the difference between the normalized cumulative distribution $f(x)$ and the expected value x as follows:

$$S\left(f(s)\right) = 2 \cdot \sum_{x=0}^{x=1} \left(f(x) - x\right) \tag{11}$$

The value of S is a signed statistic with range $S \in [-1, 1]$ (see Figure 6). As S approaches one, smaller windows close to the origin (cluster center) contain more events than expected from a random uniform distribution, indicating that SNV events concentrate near the center of the clusters and sharply decay with the distance. A negative value indicates that the events are typically depleted from the center and concentrated on the edges of the cluster, and values of S close to zero indicate that the concentration of events is mostly uniform across the 250 kb interval length, supporting the null hypothesis. We call this the **Overall Stress Introduced Heterogeneity** score, or **SItH** score, of the distribution of somatic SNVs and use it to address the typical cluster geometry in a sample. Following the same definition on individual clusters we can estimate a **Cluster SItH** score using the function $F_i(X)$ instead of $F(X)$, thus leading to $S_i = S(f_i(x))$. This definition is statistically less robust than the overall measure but allows us to assess the diversity of behaviors in clusters within a sample. We do this by estimating the quartile statistics on the ensemble of S_i values for each sample.

Code Availability

- Code to compute SItH scores is available upon request: Charles Vaske at Charlie.Vaske@nantomics.com.
- Code for all other analysis, including data sets with computed SItH scores, is available at: https://github.com/kjbussey/SItH.

Data Availability

All data in this study are publicly available for analysis:

- Cancer data: https://dcc.icgc.org/pcawg.
- Normal tissue variant data from the Complete Genomics Indices database in the 1000 Genome Project (release 20130502): ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/cgi_variant_calls/

Acknowledgments

We thank Paul Davies, Adam Orr, Charles Lineweaver, Susan Rosenberg, Julia A. Bos and Robert Austin for insightful discussions into the role of genomic instability and DNA repair in cancer. We also acknowledge the work of the clinical collaborators, data analysis teams, and funders generating the WGS data in the Pan-Cancer Analysis Working Group of the International Cancer Genome Consortium and the Complete Genomics Indices database in the 1000 Genome Project database.

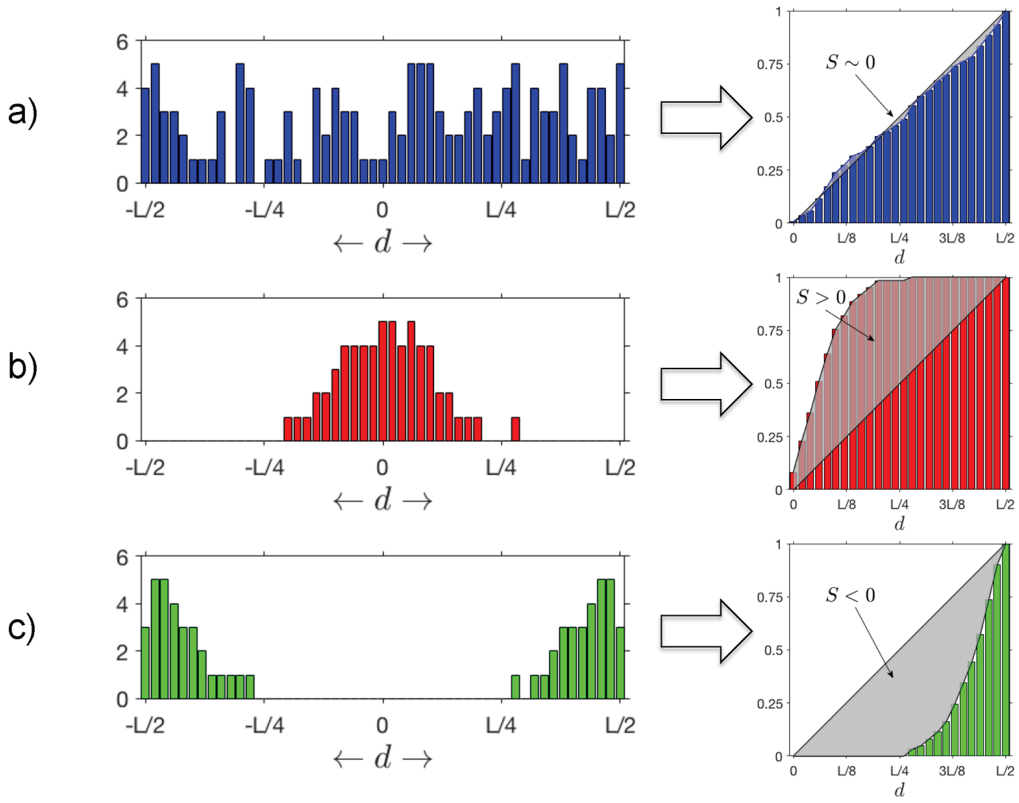


Figure 6. SItH Scores. Different outcomes for distributions of mutations as a function to the distance d to cluster centroids. a) Uniformly distributed mutations yield a linear cumulative distribution and $S \sim 0$, while b) $S > 0$ signifies a bell-shaped distribution of mutations around the centroid, and c) $S < 0$ signifies a distribution of mutations that increase with the distance to the centroid.

References

Alexandrov LB, Nik-Zainal S, Wedge DC, Sajr A, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen SG, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb; 578(7793):94–101. http://www.nature.com/articles/s41586-020-1943-3, doi: 10.1038/s41586-020-1943-3.

Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Current Opinion in Genetics & Development. 2014 Feb; 24:52–60. https://linkinghub.elsevier.com/retrieve/pii/S0959437X13001639, doi: 10.1016/j.gde.2013.11.014.

and T. Pan-cancer analysis of whole genomes. Nature. 2020 Feb; 578(7793):82–93. http://www.nature.com/articles/s41586-020-1969-6, doi: 10.1038/s41586-020-1969-6.

Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat Med. 2016; 22:105–13.

Baez-Ortega A, Gori K. Computational approaches for discovery of mutational signatures in cancer. Briefings in Bioinformatics. 2019 Jan; 20(1):77–88. https://academic.oup.com/bib/article/20/1/77/4056408, doi: 10.1093/bib/bbx082.

Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nature Reviews Genetics. 2019 sep; 21(1):44–62. https://doi.org/10.1038%2Fs41576-019-0171-x, doi: 10.1038/s41576-019-0171-x.

Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, Lawrence MS. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science. 2019 Jun; 364(6447):eaaw2872. https://www.sciencemag.org/lookup/doi/10.1126/science.aaw2872, doi: 10.1126/science.aaw2872.

Burns MB, Temiz NA, Rs H. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013; 45:977–83.

Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, Davidson S, Edwards M, Elvin JA, Hodel KP, Zahurancik WJ, Suo Z, Lipman T, Wimmer K, Kratz CP, Bowers DC, Laetsch TW, Dunn GP, Johanns TM, Grimmer MR, et al. Comprehensive Analysis of Hypermutation in Human Cancer. Cell. 2017 Nov; 171(5):1042–1056.e10. https://linkinghub.elsevier.com/retrieve/pii/S009286741731142X, doi: 10.1016/j.cell.2017.09.048.

Campbell B, Galati M, Stone S, Riemenschneider A, Edwards M, Sudhaman S, Siddaway R, Komosa M, Nunes N, Nobre L, Morrissy AS, Zatzman M, Zapotocky M, Joksimovic L, Kalimuthu S, Samuel D, Mason G, Bouffet E, Morgenstern D, Aronson M, et al. Mutations in the RAS/MAPK pathway drive replication repair deficient hypermutated tumors and confer sensitivity to MEK inhibition. Cancer Discov; Cancer Discovery. 2021; .

Chen JM, Férec C, Cooper DN. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. Human Mutation. 2009; 30:1435–48.

Chen Z, Wen W, Bao J, Kuhs KL, Cai Q, Long J, Shu Xo, Zheng W, Guo X. Integrative genomic analyses of APOBEC-mutational signature, expression and germline deletion of APOBEC3 genes, and immunogenicity in multiple cancer types. BMC Medical Genomics. 2019 Dec; 12(1):131. https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0579-3, doi: 10.1186/s12920-019-0579-3.

Cinlar E. Introduction to Stochastic Processes. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.; 1975.

Cipponi A, Goode DL, Bedo J, McCabe MJ, Pajic M, Croucher DR, Rajal AG, Junankar SR, Saunders DN, Lobachevsky P, Papenfuss AT, Nessem D, Nobis M, Warren SC, Timpson P, Cowley M, Vargas AC, Qiu MR, Generali DG, Keerthikumar S, et al. MTOR signaling orchestrates stress-induced mutagenesis, facilitating adaptive evolution in cancer. Science. 2020 Jun; 368(6495):1127–1131. https://www.sciencemag.org/lookup/doi/10.1126/science.aau8768, doi: 10.1126/science.aau8768.

Cisneros L, Bussey KJ, Orr AJ, Miočević M, Lineweaver CH, Davies PC. Ancient genes establish stress-induced mutation as a hallmark of cancer. PLOS ONE. 2017; 12:e0176258.

Cortés-Ciriano I, Lee JJK, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang CZ, Pellman DS, Park PJ. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nature Genetics. 2020 Mar; 52(3):331–341. http://www.nature.com/articles/s41588-019-0576-7, doi: 10.1038/s41588-019-0576-7.

505 **Dagogo-Jack I**, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clin-
506 ical Oncology. 2017 nov; 15(2):81–94. https://doi.org/10.1038%2Fnrclinonc.2017.166, doi: 10.1038/nrcli-
507 nonc.2017.166.

508 **Davoli T**, Uno H, Wooten EC, Sj E. Tumor aneuploidy correlates with markers of immune evasion and with
509 reduced response to immunotherapy. Science. 2017; 355:eaaf8399.

510 **Dodgshun AJ**, Fukuoka K, Edwards M, Bianchi VJ, Das A, Sexton-Oates A, Larouche V, Vanan MI, Lindhorst
511 S, Yalon M, Mason G, Crooks B, Constantini S, Massimino M, Chiaravalli S, Ramdas J, Mason W, Ashraf S,
512 Farah R, Damme AV, et al. Germline-driven replication repair-deficient high-grade gliomas exhibit unique
513 hypomethylation patterns. Acta Neuropathologica. 2020 sep; 140(5):765–776. https://doi.org/10.1007%
514 2Fs00401-020-02209-8, doi: 10.1007/s00401-020-02209-8.

515 **Drake JW**. Mutations in clusters and showers. Proc Natl Acad Sci. 2007; 104:8203–4.

516 **Fitzgerald DM**, Hastings PJ, Sm R. Stress-Induced Mutagenesis: Implications in Cancer and Drug Resistance.
517 Annual Review of Cancer Biology. 2017; 1:119–140.

518 **Forment JV**, O'Connor MJ. Targeting the replication stress response in cancer. Pharmacology and Ther-
519 apeutics. 2018; 188:155–167. https://www.sciencedirect.com/science/article/pii/S0163725818300536, doi:
520 https://doi.org/10.1016/j.pharmthera.2018.03.005.

521 **Foster P**. Stress-Induced Mutagenesis in Bacteria. Crit Rev Biochem Mol Biol. 2007; 42:373–397.

522 **Goldmann JM**, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM, Hoischen A,
523 Roach JC, Vockley JG, Veltman JA, Solomon BD, Gilissen C, Niederhuber JE. Parent-of-origin-specific signatures
524 of de novo mutations. Nature Genetics. 2016 Aug; 48(8):935–939. http://www.nature.com/articles/ng.3597,
525 doi: 10.1038/ng.3597.

526 **Goldmann JM**, Veltman JA, Gilissen C. De Novo Mutations Reflect Development and Aging of the Human
527 Germline. Trends in Genetics. 2019 Nov; 35(11):828–839. https://linkinghub.elsevier.com/retrieve/pii/
528 S0168952519301787, doi: 10.1016/j.tig.2019.08.005.

529 **Haradhvala NJ**, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braun-
530 stein LZ, Kamburov A, Hanawalt PC, Wheeler DA, Koren A, Lawrence MS, Getz G. Mutational Strand Asym-
531 metries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016 Jan; 164(3):538–549.
532 http://www.sciencedirect.com/science/article/pii/S0092867415017146, doi: 10.1016/j.cell.2015.12.050.

533 **Harris K**. The randomness that shapes our DNA. eLife. 2018 Oct; 7:e41491. https://elifesciences.org/articles/
534 41491, doi: 10.7554/eLife.41491.

535 **Harris K**, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome
536 Research. 2014 Sep; 24(9):1445–1454. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4158752/, doi:
537 10.1101/gr.170696.113.

538 **Helleday T**, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nature
539 Reviews Genetics. 2014 Sep; 15(9):585–598. http://www.nature.com/articles/nrg3729, doi: 10.1038/nrg3729.

540 **Hu X**, Xu Z, De S. Characteristics of mutational signatures of unknown etiology. NAR Cancer. 2020
541 Sep; 2(3):zcaa026. https://academic.oup.com/narcancer/article/doi/10.1093/narcan/zcaa026/5911782, doi:
542 10.1093/narcan/zcaa026.

543 **Janion C**. Inducible SOS Response System of DNA Repair and Mutagenesis in Escherichia coli. Int J Biol Sci.
544 2008; 4:338–344.

545 **Jia P**, Pao W, Zhao Z. Patterns and processes of somatic mutations in nine major cancers. BMC Medical Ge-
546 nomics. 2014 Dec; 7(1):11. http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-7-11, doi:
547 10.1186/1755-8794-7-11.

548 **Jin ZB**, Li Z, Liu Z, Jiang Y, Cai XB, Wu J. Identification of *de novo* germline mutations and causal genes for sporadic
549 diseases using trio-based whole-exome/genome sequencing: Trio-based de novo mutations detection. Bio-
550 logical Reviews. 2018 May; 93(2):1014–1031. http://doi.wiley.com/10.1111/brv.12383, doi: 10.1111/brv.12383.

551 **Jónsson H**, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson
552 HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, Jonasdottir A, Jonasdottir
553 A, Rafnar T, Frigge M, Stacey SN, et al. Parental influence on human germline de novo mutations in 1,548
554 trios from Iceland. Nature. 2017 Sep; 549(7673):519–522. http://www.nature.com/articles/nature24018, doi:
555 10.1038/nature24018.

556 **Kamburov A**, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES, Getz G. Comprehensive assess-
557 ment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci. 2015; 112:E5486–
558 E5495.

559 **Khurana E**, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J,
560 Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, et al.
561 Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science. 2013 Oct;
562 342(6154):1235587. http://www.sciencemag.org/content/342/6154/1235587, doi: 10.1126/science.1235587.

563 **Lada AG**, Dhar A, Boissy RJ, Hirano M, Rubel AA, Rogozin IB, et al. AID/APOBEC cytosine deaminase induces
564 genome-wide kataegis. Biol Direct. 2012; 7.

565 **Li Y**, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE,
566 Imielinski M, Weischenfeldt J, Beroukhim R, Campbell PJ. Patterns of somatic structural variation in human
567 cancer genomes. Nature. 2020 Feb; 578(7793):112–121. http://www.nature.com/articles/s41586-019-1913-9,
568 doi: 10.1038/s41586-019-1913-9.

569 **Ma J**, Setton J, Lee NY, Riaz N, Powell SN. The therapeutic significance of mutational signatures from DNA repair
570 deficiency in cancer. Nature Communications. 2018 aug; 9(1). https://doi.org/10.1038%2Fs41467-018-05228-y,
571 doi: 10.1038/s41467-018-05228-y.

572 **McKenzie GJ**, Harris RS, Lee PL, Sm R. The SOS response regulates adaptive mutation. Proc Natl Acad Sci. 2000;
573 97:6646–51.

574 **Moore DF**. Applied Survival Analysis Using R. Switzerland: Springer International Publishing; 2016.

575 **Murai J**. Targeting DNA repair and replication stress in the treatment of ovarian cancer. International
576 Journal of Clinical Oncology. 2017 jun; 22(4):619–628. https://doi.org/10.1007%2Fs10147-017-1145-7, doi:
577 10.1007/s10147-017-1145-7.

578 **Nik-Zainal S**, Alexandrov LB, Wedge DC, Loo PV, Greenman CD, Raine K, et al. Mutational Processes Molding
579 the Genomes of 21 Breast Cancers. Cell. 2012; 149:979–93.

580 **Nik-Zainal S**, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560
581 breast cancer whole-genome sequences. Nature. 2016; 534:47–54.

582 **Noorani A**, Bornschein J, Lynch AG, Secrier M, Achilleos A, Eldridge M, Bower L, Weaver JMJ, Crawte J, Ong
583 CA, Shannon N, MacRae S, Grehan N, Nutzinger B, O'Donovan M, Hardwick R, Tavaré S, Fitzgerald RC,
584 Consortium obotOCCaMSO, Elliott RF, et al. A comparative analysis of whole genome sequencing of
585 esophageal adenocarcinoma pre- and post-chemotherapy. Genome Research. 2017 Jun; 27(6):902–912.
586 http://genome.cshlp.org/content/27/6/902, doi: 10.1101/gr.214296.116.

587 **Patch AM**, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S, Nones K, Cowin P, Alsop K, Bailey
588 PJ, Kassahn KS, Newell F, Quinn MCJ, Kazakoff S, Quek K, Wilhelm-Benartzi C, Curry E, Leong HS, Hamilton
589 A, Mileshkin L, et al. Whole–genome characterization of chemoresistant ovarian cancer. Nature. 2015 May;
590 521(7553):489–494. http://www.nature.com/articles/nature14410, doi: 10.1038/nature14410.

591 **Petljak M**, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, Dawson KJ, Ju YS, Iorio F, Tubio JMC,
592 Koh CC, Georgakopoulos-Soares I, Rodríguez-Martín B, Otlu B, O'Meara S, Butler AP, Menzies A, Bhosle SG,
593 Raine K, Jones DR, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic
594 APOBEC Mutagenesis. Cell. 2019 Mar; 176(6):1282–1294.e20. https://linkinghub.elsevier.com/retrieve/pii/
595 S0092867419301618, doi: 10.1016/j.cell.2019.02.012.

596 **Phillips DH**. Mutational spectra and mutational signatures: Insights into cancer aetiology and mechanisms
597 of DNA damage and repair. DNA Repair. 2018 Nov; 71:6–11. https://linkinghub.elsevier.com/retrieve/pii/
598 S156878641830168X, doi: 10.1016/j.dnarep.2018.08.003.

599 **Poulos RC**, Wong YT, Ryan R, Pang H, Wong JWH. Analysis of 7,815 cancer exomes reveals associations between
600 mutational processes and somatic driver mutations. PLOS Genetics. 2018 Nov; 14(11):e1007779. https:
601 //dx.plos.org/10.1371/journal.pgen.1007779, doi: 10.1371/journal.pgen.1007779.

602 **Pouyet F**, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more
603 than 95% of the human genome and bias demographic inferences. eLife. 2018 Aug; 7:e36317. https://
604 elifesciences.org/articles/36317, doi: 10.7554/eLife.36317.

605 **Ram Y**, Hadany L. Stress-induced mutagenesis and complex adaptation. Proc R Soc Lond B Biol Sci. 2014;
606 281:20141025.

Ram Y, Hadany L. Evolution of Stress-Induced Mutagenesis in the Presence of Horizontal Gene Transfer. The American Naturalist. 2019 Jul; 194(1):73–89. https://www.journals.uchicago.edu/doi/10.1086/703457, doi: 10.1086/703457.

Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, Hornshøj H, Hess JM, Juul RI, Lin Z, Feuerbach L, Sabarinathan R, Madsen T, Kim J, Mularoni L, Shuai S, Lanzós A, Herrmann C, Maruvka YE, Shen C, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020 Feb; 578(7793):102–111. http://www.nature.com/articles/s41586-020-1965-x, doi: 10.1038/s41586-020-1965-x.

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45:970–6.

Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA. Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. Molecular Cell. 2012 May; 46(4):424–435. https://linkinghub.elsevier.com/retrieve/pii/S1097276512002997, doi: 10.1016/j.molcel.2012.03.030.

Roh W, P-l C, Reuben A, Spencer CN, Prieto PA, Miller JP, et al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. Sci Transl Med. 2017; 9:eaah3560.

Roper N, Gao S, Maity TK, Banday AR, Zhang X, Venugopalan A, Cultraro CM, Patidar R, Sindiri S, Brown AL, Goncearenco A, Panchenko AR, Biswas R, Thomas A, Rajan A, Carter CA, Kleiner DE, Hewitt SM, Khan J, Prokunina-Olsson L, et al. APOBEC Mutagenesis and Copy-Number Alterations Are Drivers of Proteogenomic Tumor Evolution and Heterogeneity in Metastatic Thoracic Tumors. Cell Reports. 2019 Mar; 26(10):2651–2666.e6. https://linkinghub.elsevier.com/retrieve/pii/S2211124719301998, doi: 10.1016/j.celrep.2019.02.028.

Roschke AV, Lababidi S, Tonon G, Gehlhaus KS, Bussey K, Weinstein JN, Ir K. Karyotypic "state" as a potential determinant for anticancer drug discovery. Proc Natl Acad Sci U S A. 2005; 102:2964–2969.

Roschke AV, Stover K, Tonon G, Sch"affer AA, Ir K. Stable Karyotypes in Epithelial Cancer Cell Lines Despite High Rates of Ongoing Structural and Numerical Chromosomal Instability. Neoplasia. 2002; 4:19–31.

Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, et al. Karyotypic complexity of the NCI-60 drug-screening panel. Cancer Res. 2003; 63:8634–8647.

Rosenberg SM, Shee C, Frisch RL, Pj H. Stress-induced mutation via DNA breaks in Escherichia coli: A molecular mechanism with implications for evolution and medicine. BioEssays. 2012; 34:885–92.

Russo M, Crisafulli G, Sogari A, Reilly NM, Arena S, Lamba S, Bartolini A, Amodio V, Magrì A, Novara L, Sarotto I, Nagel ZD, Piett CG, Amatu A, Sartore-Bianchi A, Siena S, Bertotti A, Trusolino L, Corigliano M, Gherardi M, et al. Adaptive mutability of colorectal cancers in response to targeted therapies. Science. 2019 Dec; 366(6472):1473–1480. https://www.sciencemag.org/lookup/doi/10.1126/science.aav4474, doi: 10.1126/science.aav4474.

Saini N, Gordenin DA. Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells: Human Somatic Mutation Load and Spectra. Environmental and Molecular Mutagenesis. 2018 Oct; 59(8):672–686. http://doi.wiley.com/10.1002/em.22215, doi: 10.1002/em.22215.

Sakofsky CJ, Ayyar S, Deem AK, W-h C, Ira G, Malkova A. Translesion Polymerases Drive Microhomology-Mediated Break-Induced Replication Leading to Complex Chromosomal Rearrangements. Mol Cell. 2015; 60:860–72.

Scarpa A, , Chang DK, Nones K, Corbo V, Patch AM, Bailey P, Lawlor RT, Johns AL, Miller DK, Mafficini A, Rusev B, Scardoni M, Antonello D, Barbi S, Sikora KO, Cingarlini S, Vicentini C, McKay S, Quinn MCJ, et al. Whole-genome landscape of pancreatic neuroendocrine tumours. Nature. 2017 feb; 543(7643):65–71. https://doi.org/10.1038%2Fnature21063, doi: 10.1038/nature21063.

Shee C, Gibson JL, Sm R. Two Mechanisms Produce Mutation Hotspots at DNA Breaks in Escherichia coli. Cell Rep. 2012; 2:714–21.

Shi MJ, Meng XY, Fontugne J, Chen CL, Radvanyi F, Bernard-Pierrot I. Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer. Genome Medicine. 2020 Dec; 12(1):85. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00781-y, doi: 10.1186/s13073-020-00781-y.

Srivastava AK, Han C, Zhao R, Cui T, Dai Y, Mao C, et al. Enhanced expression of DNA polymerase eta contributes to cisplatin resistance of ovarian cancer stem cells. Proc Natl Acad Sci. 2015; 112:4411–6.

Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015; 521:81–4.

Supek F, Lehner B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell. 2017 Jul; 170(3):534–547.e23. https://linkinghub.elsevier.com/retrieve/pii/S0092867417307742, doi: 10.1016/j.cell.2017.07.003.

Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. eLife. 2013; 2:e00534.

Telis N, Aguilar R, Harris K. Selection against archaic hominin genetic variation in regulatory regions. Nature Ecology & Evolution. 2020 Nov; 4(11):1558–1566. http://www.nature.com/articles/s41559-020-01284-0, doi: 10.1038/s41559-020-01284-0.

Temprine K, Campbell NR, Huang R, Langdon EM, Simon-Vermot T, Mehta K, Clapp A, Chipman M, White RM. Regulation of the error-prone DNA polymerase Polκ by oncogenic signaling and its contribution to drug resistance. Science Signaling. 2020 Apr; 13(629):eaau1453. https://stke.sciencemag.org/lookup/doi/10.1126/scisignal.aau1453, doi: 10.1126/scisignal.aau1453.

Teng K, Qiu M, Li Z, Luo H, Zeng Z, Luo R, et al. DNA polymerase ? protein expression predicts treatment response and survival of metastatic gastric adenocarcinoma patients treated with oxaliplatin-based chemotherapy. J Transl Med. 2010; 8:126.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. doi: 10.1038/nature15393.

Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. Nature Reviews Genetics. 2019 mar; 20(7):404–416. https://doi.org/10.1038%2Fs41576-019-0114-6, doi: 10.1038/s41576-019-0114-6.

Ubhi T, Brown G. Exploiting DNA Replication Stress for Cancer Treatment. Cancer Res; Cancer Research. 2019; 79(8):1730–1739.

Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K, Quinn MCJ, Robertson AJ, Fadlullah MZH, Bruxner TJC, Christ AN, Harliwong I, Idrisoglu S, Manning S, Nourse C, Nourbakhsh E, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature. 2015 Feb; 518(7540):495–501. http://www.nature.com/articles/nature14169, doi: 10.1038/nature14169.

Wang J, Gonzalez KD, Scaringe WA, Tsai K, Liu N, Gu D, Li W, Hill KA, Ss S. Evidence for mutation showers. Proc Natl Acad Sci. 2007; 104:8403–8408.

Waters LS, Minesinger BK, Wiltrout ME, D'Souza S, Woodruff RV, C WG. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. Microbiol Mol Biol Rev MMBR. 2009; 73:134–54.

Wu A, Zhang Q, Lambert G, Khin Z, Gatenby RA, Kim JH, et al. Ancient hot and cold genes and chemotherapy resistance emergence. Proc Natl Acad Sci. 2015; 112:10467–10472.

Xia J, Chiu LY, Nehring RB, Bravo Núñez MA, Mei Q, Perez M, Zhai Y, Fitzgerald DM, Pribis JP, Wang Y, Hu CW, Powell RT, LaBonte SA, Jalali A, Matadamas Guzmán ML, Lentzsch AM, Szafran AT, Joshi MC, Richters M, Gibson JL, et al. Bacteria-to-Human Protein Networks Reveal Origins of Endogenous DNA Damage. Cell. 2019 Jan; 176(1-2):127–143.e24. https://linkinghub.elsevier.com/retrieve/pii/S0092867418316222, doi: 10.1016/j.cell.2018.12.008.

Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH. Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics. 2010 Dec; 11(1):11. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-11, doi: 10.1186/1471-2105-11-11.

Zhang L, Vijg J. Somatic Mutagenesis in Mammals and Its Implications for Human Disease and Aging. Annual Review of Genetics. 2018 Nov; 52(1):397–419. https://www.annualreviews.org/doi/10.1146/annurev-genet-120417-031501, doi: 10.1146/annurev-genet-120417-031501.

705 **Zhou W**, Chen Y, Liu X, Chu P, Loria S, Wang Y, et al. Expression of DNA translesion synthesis polymerase eta in
706 head and neck squamous cell cancer predicts resistance to gemcitabine and cisplatin-based chemotherapy.
707 PloS One. 2013; 8:e83978.

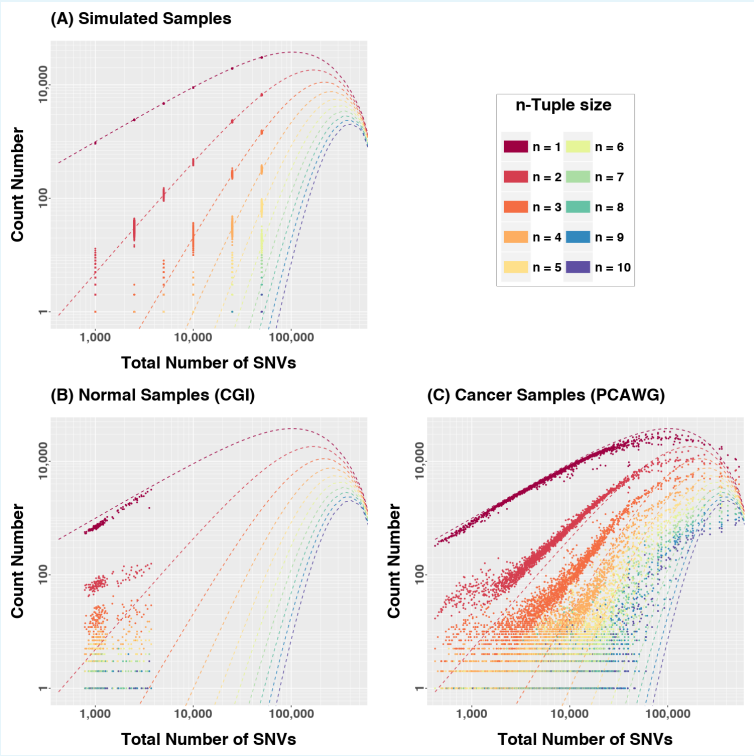
708 **Zhu H**, Swami U, Preet R, Zhang J. Harnessing DNA Replication Stress for Novel Cancer Therapy. Genes. 2020
709 aug; 11(9):990. https://doi.org/10.3390%2Fgenes11090990, doi: 10.3390/genes11090990.

Appendix 1

Detailed tuple distributions

We consider a *n-tuple* as a set of n contiguous SNVs in genomic space with inter-SNV distances $x \leq D^\star$ and $D^\star = 15$ kb. Then different values of n we observed numbers of n-tuples in simulated, normal and cancer data, shown in Figure 1. Singletons (1-tuples) are significantly underrepresented for low mutational loads, while tuples of size two or more are typically over-represented with respect to a Poisson point process model.

And interesting observation, in normal samples 2-tuples are the most extremely overrepresented, and the overrepresentation decays quickly with he tuple size. Tuples of size $n = 4-5$ are basically as frequently observed as expected. But in cancer samples tuples of all sizes are over-represented. In fact, perhaps large tuples sizes are even more extreme than small ones.

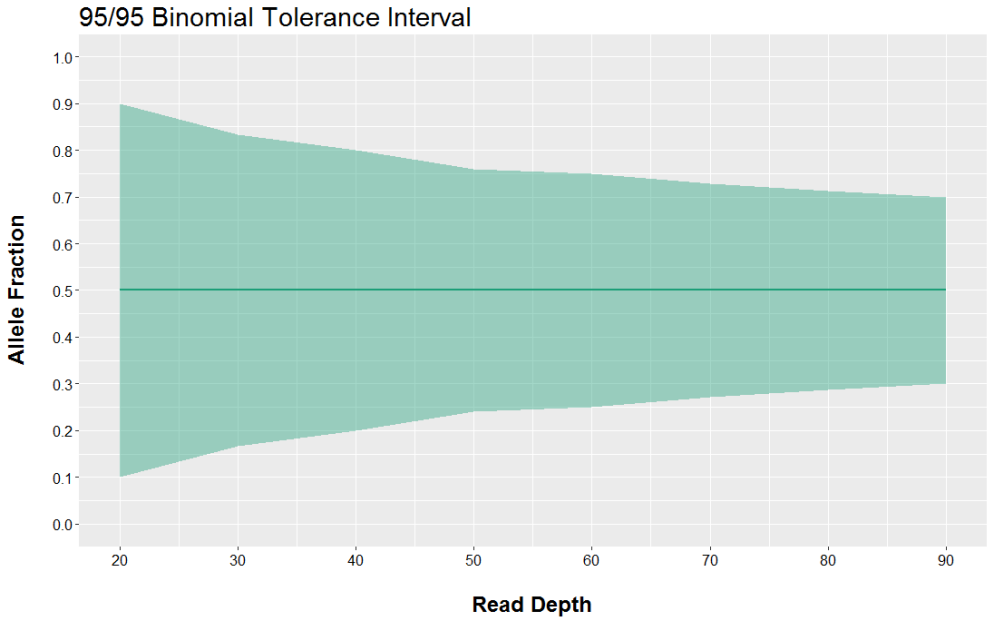


Appendix 1 Figure 1. Number of n-tuples in (A) simulated, (B) normal and (C) cancer samples. Singletons (1-tuples) are less frequent than expected in both normal and cancer cases, while larger n-tuples are more frequent than expected. This relation is inverted for large mutational loads.

Appendix 2

Precision in allele fraction estimations

The 95/95 binomial tolerance interval for a true allele fraction of 0.5 at a read depth as high as 60x ranges from 0.25 to 0.75 (Figure 1), meaning that random fluctuations in allele fraction estimations anywhere in that range cannot be ruled out. According to this much larger read depths are necessary to have the precision power to use allele fractions as a methods to estimate mutation lineages and discriminate varying degrees of heterogeneity across a tumor.



Appendix 2 Figure 1. The shaded interval represents the bounds in which we are 95% confident that 95% of the measurements of a true allele fraction of 0.5 will lie as a function of the real read depth.