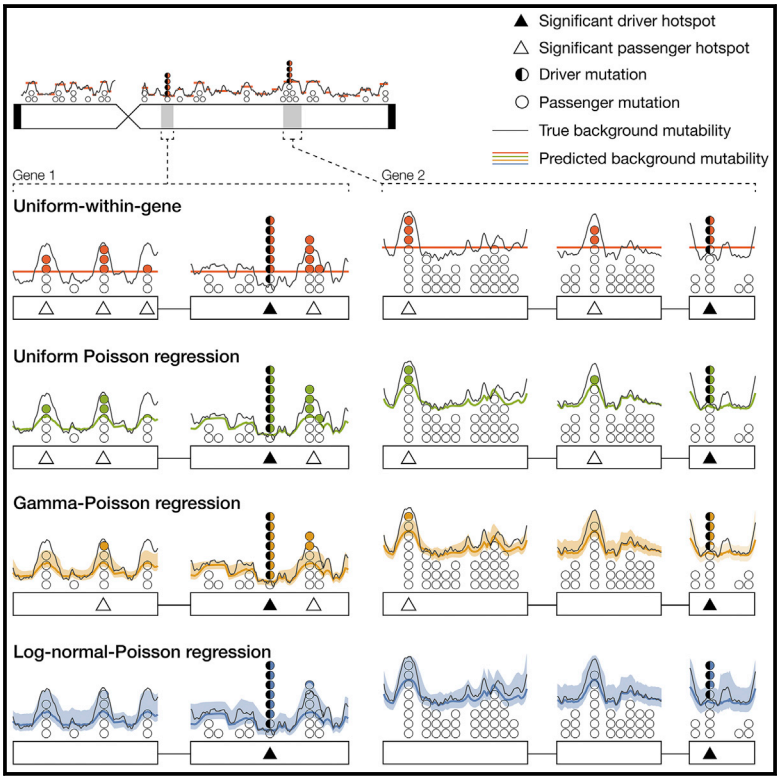
Article

Cancer Cell

Passenger Hotspot Mutations in Cancer

Graphical Abstract



Highlights

- Many cancer hotspots are passengers, recurring at inherently mutable genomic sites
- Known genomic covariates are insufficient to fully predict inherent mutability
- Our LNP model accurately infers latent variability beyond what current covariates predict
- Our LNP model identifies putative driver hotspots with far fewer false-positives

Authors

Julian M. Hess, Andre Bernards, Jaegil Kim, ..., Nicholas J. Haradhvala, Michael S. Lawrence, Gad Getz

Correspondence

mslawrence@mgh.harvard.edu (M.S.L.), gadgetz@broadinstitute.org (G.G.)

In Brief

Somatic hotspot mutations found in tumors are generally considered evidence for selection and are used to nominate tumor drivers. Hess et al. show that many hotspots occur at inherently mutable sites without selection and develop a model that accounts for these passenger hotspots, which can more accurately nominate true driver mutations.

Hess et al., 2019, Cancer Cell *36*, 288–301
September 16, 2019
https://doi.org/10.1016/j.ccell.2019.08.002

CellPress

CellPress

Cancer Cell
Article

Passenger Hotspot Mutations in Cancer

Julian M. Hess,[1] Andre Bernards,[2,3] Jaegil Kim,[1] Mendy Miller,[1] Amaro Taylor-Weiner,[1] Nicholas J. Haradhvala,[1,2] Michael S. Lawrence,[1,2,3,4,*] and Gad Getz[1,2,3,4,5,*]

[1]The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]Center for Cancer Research, Massachusetts General Hospital, Boston, MA 02114, USA
[3]Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115, USA
[4]Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA
[5]Lead Contact
*Correspondence: mslawrence@mgh.harvard.edu (M.S.L.), gadgetz@broadinstitute.org (G.G.)
https://doi.org/10.1016/j.ccell.2019.08.002

SUMMARY

Current statistical models for assessing hotspot significance do not properly account for variation in site-specific mutability, thereby yielding many false-positives. We thus (i) detail a Log-normal-Poisson (LNP) background model that accounts for this variability in a manner consistent with models of mutagenesis; (ii) use it to show that passenger hotspots arise from all common mutational processes; and (iii) apply it to a ~10,000-patient cohort to nominate driver hotspots with far fewer false-positives compared with conventional methods. Overall, we show that many cancer hotspot mutations recurring at the same genomic site across multiple tumors are actually passenger events, recurring at inherently mutable genomic sites under no positive selection.

INTRODUCTION

The genome of a cell lineage continually accrues mutations over time. The vast majority of mutations are either selectively neutral ''passengers'' that leave the lineage phenotypically unaltered, or selectively negative mutations that result in slower growth or cell death. However, occasionally, rare selectively positive mutations (''drivers'') that increase a cell's proliferative fitness can occur. Such a cell may acquire additional driver events that enable it to outcompete its neighbors, eventually transforming into cancer (Cairns, 1975; Stratton et al., 2009).

Selectively positive mutations accumulate in tumor suppressors and oncogenes. These cancer driver genes are recurrently mutated across tumors, and can be identified based on having mutational densities significantly above the background passenger density. This requires accurate estimation of the mutational background, a task complicated by its considerable heterogeneity (Hodgkinson and Eyre-Walker, 2011). Some genomic elements are recurrently mutated not due to positive selection but rather simply due to their higher mutability (Hodgkinson and Eyre-Walker, 2011; Stamatoyannopoulos et al., 2009; Pleasance et al., 2010). Over the past decade, the field has developed increasingly sophisticated statistical models (Lawrence et al., 2013, 2014; Weghorn and Sunyaev, 2017; Martincorena et al., 2017; Dees et al., 2012; Getz et al., 2007) to infer and account for the heterogeneous mutational background. This has increased power and specificity to distinguish true drivers (with an excess of positively selected mutations) from false-positives (with increased mutation density due to high background mutability alone).

Some of the best-known oncogenes are recurrently mutated at the same codon in many tumors, e.g., the V600E mutation in *BRAF*. Inspired by these examples, many methods for detecting driver events use exact positional recurrence as a signal of positive selection. This requires a model of background mutability at the site-specific level. The prevailing assumption has been that all equivalent base pairs within a particular gene (e.g., all sites with the same *k*-mer sequence context) have the

Significance

Hotspot mutations have conventionally been taken as universal evidence of somatic positive selection, unequivocally pinpointing genes driving tumorigenesis. We show here that this convention is falsely premised on an inaccurate statistical model of background mutagenesis. Many hotspots are actually passengers, recurring at inherently mutable sites under no positive selection, which current background models do not account for. We develop a model that accounts for variation in site-specific mutability and use it to nominate driver hotspots with far fewer false-positives compared with conventional methods. As the research community faces critical decisions in prioritizing putative driver mutations for deep experimental characterization to assess therapeutic potential, we offer our findings as a guide to avoid wasting valuable resources on passenger hotspots.

same background probability of being mutated. Thus, it is unlikely to observe by chance many tumors sharing mutations at a particular base pair in a gene, while the other equivalent base pairs of the gene remain unmutated. This leads to the conventional assumption that mutational hotspots *must* reflect true driver events and that no "passenger hotspots" occur. Here, we present evidence to the contrary.

RESULTS

In the same way that certain regions of the genome are more highly mutated than others, and certain fragile sites of the genome are more prone to breakage (Schrock and Huebner, 2015; Mitsui and Tsuji, 2012), certain individual base pairs in the genome appear to be more mutable, simply because they are more vulnerable to damage and/or more refractory to repair. Thus, to identify significantly mutated hotspots, we must first be able to statistically estimate position-specific mutation frequencies. This requires a large cohort of very high-quality somatic mutation calls to have sufficient power to accurately estimate base-level mutation frequencies and to avoid recurrent sequencing artifacts or germline polymorphisms that can severely distort base-level analyses. Here, we examined a cohort of 9,023 quality-controlled whole-exome-sequenced tumor/normal pairs spanning 32 tumor types, generated by the The Cancer Genome Atlas MC3 mutation-calling initiative (Ellrott et al., 2018), a dataset containing 2,288,080 somatic single-nucleotide variations.

Discovering Significant Hotspots Based on a Statistical Model of the Site-Specific Background Mutational Frequency

A common approach to find significantly mutated bases (i.e., "hotspots") is to compare the observed number of mutations across a cohort at a given site with the distribution of the expected number of mutations predicted by a given background model, generating a p value for that site. Sites passing multiple-hypothesis correction (e.g., false discovery rate [FDR] q value $\leq$ 0.1) are then considered significant hotspots. Obviously, what gets called statistically significant is only biologically meaningful if the underlying statistical approximation of the true background mutation frequency is accurate. Although we cannot directly evaluate the underlying model's overall accuracy since we lack ground truth background mutation frequencies for every genomic position, we can assess a given model's specificity via orthogonal criteria for whether mutations deemed significant are indeed under positive selection.

One orthogonal criterion we can use to assess whether a mutation is under positive selection is the distribution of variant protein-coding effects compared with expectation. Variant effects of mutations under no selective pressure will be randomly distributed according to the codon structure and signatures of mutational processes operating throughout the exome. In contrast, positively selected mutations are mostly nonsynonymous or splice altering, because essentially all coding driver mutations alter their corresponding protein. A recent study (Martincorena et al., 2017) estimated the fraction of synonymous driver mutations at approximately 5%, most of which are splice altering. Therefore, an effective proxy for the specificity of a given algorithm is the degree to which the candidate driver hotspots it finds are enriched for nonsynonymous mutations beyond the expected baseline.

We calculate this baseline by using the overall exonic substitution frequencies for each trinucleotide context to generate the expected background frequencies of each coding effect—synonymous, missense, nonsense—in that context. By adding the contributions from all 96 trinucleotide substitutions, weighted by their relative frequencies, we find an overall expected baseline distribution of 28% synonymous, 67% missense, and 5% nonsense mutations (Figure S1). Note that throughout our entire analysis, we avoid splice-altering mutations by excluding genomic positions and mutations within 5 base pairs of a splice site, since synonymous events at those positions frequently disrupt splicing (Supek et al., 2014). We also avoid mutations overlapping transcription factor binding sites by excluding positions overlapping regions harboring strong evidence of transcription factor binding (i.e., sites assayed in more than 90 cell lines with >70% of scores >800), as inferred by ENCODE chromatin immunoprecipitation sequencing experiments (Gerstein et al., 2012).

Hotspots in Non-KCGs Found Significant by Conventional Algorithms Are Mostly Neutral

A simple and still common background model employs the conventional assumption used in many studies up until recently (Van den Eynden et al., 2015; Lohr et al., 2012; Lawrence et al., 2014; Chang et al., 2016; Baeissa et al., 2017; Araya et al., 2016; Miller et al., 2015) that all sites within the same *k*-mer context in a given gene are equally mutable. Our version of this model, which we will refer to as "Uniform-within-gene," assumes that the background mutation frequency at a given position is proportional to the average exome-wide mutation frequency for the base substitution/trinucleotide combination at that position, weighted by a gene-specific mutability factor that captures the overall number of mutations in the gene. Expected counts are Poisson distributed around this background frequency. This null distribution can also be generated by permuting mutations while maintaining their sequence context within each gene (e.g., Lawrence et al., 2014). Applying these assumptions yields a long list of 1,677 significant hotspots within 1,218 genes across our pan-cancer cohort. Although many of these significant hotspots occur in known cancer genes (KCGs), as defined by the Cancer Gene Census (CGC) (v.85) (Futreal et al., 2004)—77/133 KCGs contain significant hotspots (Figure 1A) and recapitulate known driver events (e.g., *BRAF* V600E)—1,264 significant hotspots occur in 1,131 genes with uncertain oncogenic roles (non-KCGs) (Figure 1A).

Significant hotspots in KCGs are nearly devoid of synonymous mutations (0.4% of mutations at hotspots versus the expected 27.9%), providing strong orthogonal evidence that these mutations are indeed under positive selection. If the 1,264 putative driver hotspots in non-KCGs were also drivers, one would expect that they too would be highly depleted of synonymous mutations. Instead, however, we find that the distribution of their coding effects is very close to what we expect by chance (observed silent 19.1% versus expected 24.3%, Figure 1B), indicating that most of these hotspots are equivalent to randomly chosen mutations, and are thus neutral passengers under little-to-no selective pressure.

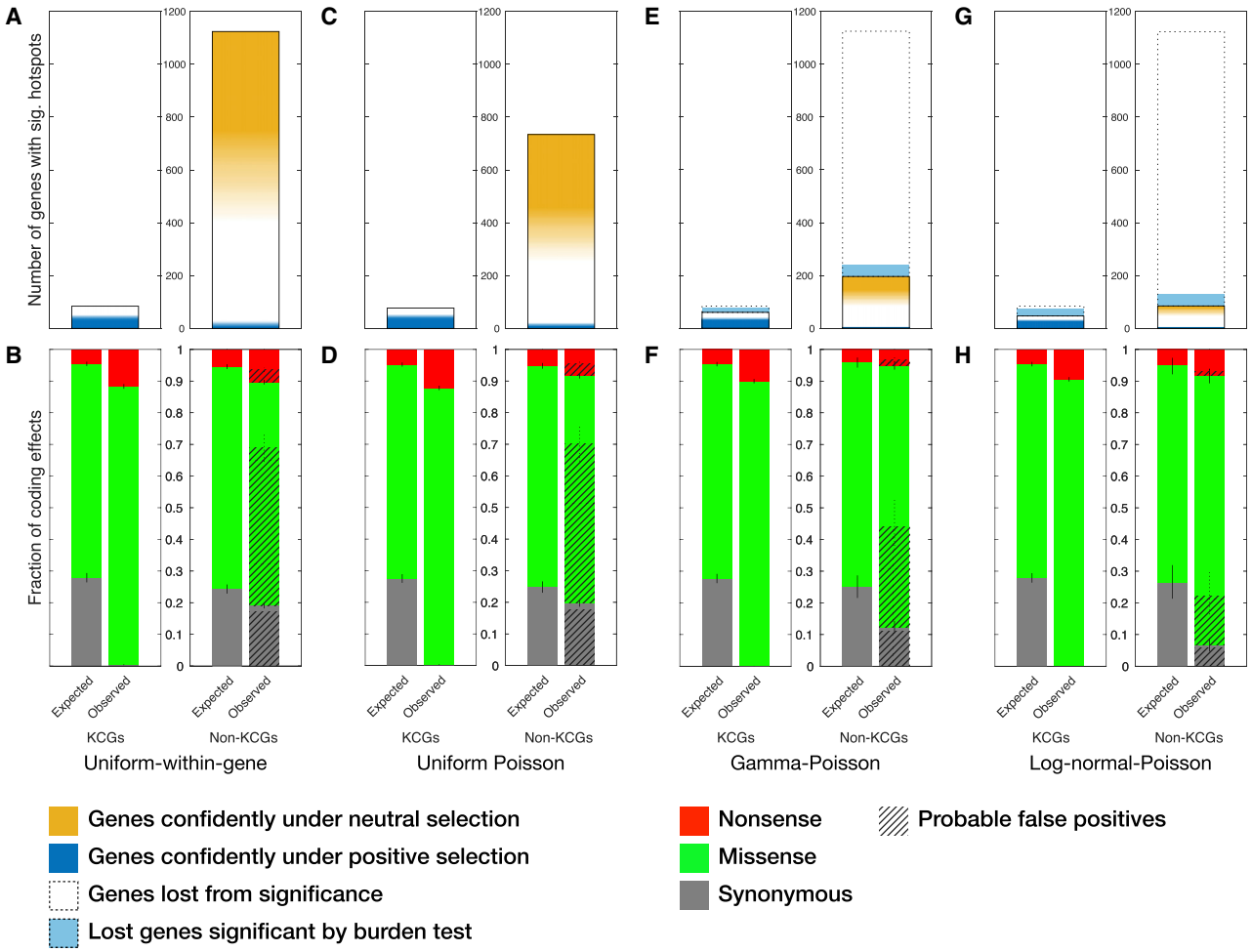


Figure 1. Comparison of Number of Genes Containing Hotspots and Protein-Coding Effects of Hotspots Found Significant by the Four Statistical Models

(A) Number of genes containing significant hotspots ($q \leq 0.1$) according to the Uniform-within-gene statistical model, segregated by whether the gene is a known cancer gene (KCG) (i.e., in the Cancer Gene Census). Genes under neutral selection are conservatively defined to have $\geq$95% probability of their *dN*/*dS* falling between 0.8 and 1.2; orange scale fades to white when this probability falls below 0.5. Genes under positive selection are defined to have $\geq$97.5% probability $dN / dS > 1.2$; blue scale fades to white when probability falls below 0.7. Genes are denoted as lost from significance relative to the Uniform-within-gene method.
(B) Expected and observed distributions of protein-coding effects of hotspot mutations significant ($q \leq 0.1$) by the Uniform-within-gene model, segregated by known cancer gene status. We estimate the overall fraction of false-positive passenger mutations (hatched bars) by assuming 90% of significant synonymous hotspots are false-positives. Thus, the proportion of observed nonsynonymous mutations concordant with the expected ratio between synonymous and nonsynonymous mutations will also be passengers.
(C) Same as (A) but according to the Uniform Poisson model.
(D) Same as (B) but by the Uniform Poisson model.
(E) Same as (A) but by the Gamma-Poisson model.
(F) Same as (B) but by the Gamma-Poisson model.
(G) Same as (A) but by the Log-normal-Poisson model.
(H) Same as (B) but by the Log-normal-Poisson model. Lines between stacked bars denote 95% confidence intervals.
See also Figure S1, Tables S1, S2, and S3.

Assuming that mutational patterns in KCGs reflect mutational patterns of driver genes in general, then the overwhelming majority of all silent hotspots are likely to be passenger events, given their extreme paucity in KCGs. However, this is a conservative assumption—although the fraction of synonymous drivers was estimated at 5% (Martincorena et al., 2017) (of which most are splice altering), it is possible that non-splice-altering synonymous mutations can have a functional effect, e.g., by affecting mRNA folding stability (Katz and Burge, 2003), binding of regulatory factors, or translational efficiency (Quax et al., 2015). Such driver mutations are likely rare, owing to their absence in KCGs, but their exact prevalence is difficult to quantify. Therefore, we liberally assume that 10% of non-splice-altering synonymous variants are in fact functional.

We would expect a proportion of observed nonsynonymous mutations concordant with the expected ratio between (the

90% presumed nonfunctional) synonymous and nonsynonymous mutations to also be passenger events. For example, if the expected ratio of synonymous to nonsynonymous mutations was 1:3, and 10% of putative significant hotspots were synonymous (of which 90% are presumed nonfunctional), then we would expect an additional $3 \times 0.1 \times 0.9 = 27\%$ of the putative nonsynonymous hotspots to also be passengers. By this reasoning, at least 66.4% of putative nonsilent driver hotspots in non-KCGs according to the conventional Uniform-within-gene model are also false-positives, illustrated by the hatched bars in Figure 1A.

In addition to finding that these hotspot mutations at the base pair level are neutral, we also noted that a large fraction of the genes containing these hotspots are themselves neutral. To determine this, we assessed whether a gene was under neutral selection (i.e., not under positive or negative selection) using the molecular evolution criterion of the ratio of a gene's nonsynonymous:synonymous somatic mutation densities (*dN*/*dS*) (Kimura, 1977; Martincorena et al., 2017; Nei and Gojobori, 1986). After normalizing for signature heterogeneity and genes' codon structures (Greenman et al., 2006), we expect parity between these densities (i.e., $dN/dS \approx 1$) in genes under neutral selection. Computing *dN*/*dS* for each gene can only yield a confident estimate in genes with sufficient numbers of mutations, allowing us to conservatively assess whether they were under neutral selection. We identified 194 (of the ~20,000) genes confidently under neutral selection ($\geq$95% probability that the gene's *dN*/*dS* is between 0.8 and 1.2). A total of 22.8% of non-KCGs that contained significant hotspots was either within these 194 genes (n = 70) or only contained silent hotspots (n = 200) (Figure 1A; Table S1), further confirming the poor specificity of the conventional Uniform-within-gene model. Overall, we conclude that analysis methods based on the conventional assumption that equivalent base pairs have the same background mutability produce lists of candidate hotspots with many false-positives.

Thus, the apparent significant recurrence of mutations even at sites/genes under neutral selection is due to a naive background model that fails to account for the underlying variability in the background base-wise mutability. We therefore need to model mutational recurrences with a site-specific background model.

Currently Known Covariates Cannot Account for all Base-Wise Mutational Heterogeneity

It is possible that the unaccounted variability might be completely explained by previously reported covariates that affect mutation frequencies on both coarse and fine scales. Covariates such as replication timing (Stamatoyannopoulos et al., 2009), chromatin state (Polak et al., 2015), and gene expression levels (Pleasance et al., 2010) have been reported to influence background mutability on a broad scale (~100 kbp–1 Mbp). More recent studies have discovered that other covariates like nucleosome positions or transcription factor binding activity influence mutability on a much smaller scale (10s of base pairs) (Poulos et al., 2016; Sabarinathan et al., 2016; Mao et al., 2018). We tested a fixed regression model in which the base-wise mutation frequency is entirely determined by these covariates, which we refer to as the "Uniform Poisson" model. However, we find that the covariates alone are incapable of explaining all of the variability, because 25.6% ($n(dN/dS \approx 1) = 50$, n(only silent hotspots) = 149) of the 737 non-KCGs containing significant hotspots are confidently under neutral selection (Figure 1C). Moreover, the distribution of protein-coding effects of hotspots in non-KCGs found to be significant by this method is essentially identical to that of the Uniform-within-gene model (observed silent fraction of 19.8% versus expected 24.8%; Figure 1D). Although additional yet-undiscovered covariates may be able to better explain this variability in the future, we currently are forced to update the model to explicitly allow for variability in base-wise mutation frequencies beyond what can be fully modeled with currently known covariates.

Introducing Uncertainty in Site-Specific Mutability Improves Specificity of Finding Driver Hotspots with Minimal Loss of Sensitivity

In contrast to the Uniform Poisson model, another way to account for the base-wise variability is by allowing the mutation frequency at each site to be drawn from a probability distribution reflecting additional variability in the background mutability. The choice and parameterization of the underlying distribution can make a large difference in model performance. Recent methods have employed a Gamma-Poisson model, both at the gene level (Weghorn and Sunyaev, 2017; Martincorena et al., 2017; Imielinski et al., 2017) and on the base level (Smith et al., 2016), in which mutation counts are still Poisson distributed, but the Poisson rates vary according to a fitted gamma distribution, adding an additional parameter to represent the overdispersion. There is no inherent biological rationale for using the gamma distribution to represent the uncertainty of the Poisson rates—it is merely mathematically convenient, because there is an easy closed-form expression for a Poisson distribution whose rate is gamma-distributed: the negative binomial/Gamma-Poisson distribution. We found that by applying the Gamma-Poisson regression model indeed explained a large amount of the additional variability, but it still did not capture all of it; while the set of significant hotspots in non-KCGs is depleted of synonymous events (observed 12.3% versus expected 25.1%), we still see that 15.7% (31/197) of non-KCGs containing significant hotspots are neutral (Figures 1E and 1F), suggesting that all the variability is not yet accounted for in this model and there is additional specificity to be gained.

Next, we tried to improve on the Gamma-Poisson model by replacing the gamma distribution with a log-normal distribution. Unlike the gamma distribution, arbitrarily chosen for its mathematical convenience, the log-normal distribution is based on the idea that mutant base pairs do not instantaneously arise but are the net result of many independent consecutive events (e.g., damage and repair processes), each with an independent probability of occurring. For instance, a mutagen has a certain probability of initially damaging a nucleotide, which in turn has a certain probability of being missed by all repair mechanisms before S phase. Should all of these events occur, the DNA polymerase must also fail to recognize the lesion and incorporate the wrong complementary base, which must survive another cell cycle without being recognized to become a mutated base pair in the genome. By the geometric central limit theorem, this product of probabilities approaches a log-normal distribution (Sutton, 1997).

Applying this Log-normal-Poisson (LNP) model to our cohort identified a list of candidate driver hotspots in non-KCGs with the lowest fraction of non-KCGs under neutral selection with significant hotspots (7% [6/86] of genes; Figure 1G) and was most depleted of synonymous events relative to expectation (observed 6.6% versus expected 26.5%; Figure 1H). These results suggest that the LNP model has the highest specificity among the four tested models.

Although the LNP model increases specificity of finding true oncogenic hotspots, this potential advantage may come at the expense of decreased sensitivity. Indeed, 31 KCGs containing significant hotspots according to the Uniform-within-gene model are lost from significance by the LNP model; 18 KCGs are lost from significance by the Gamma-Poisson model (Figures 1E and 1G); and 28/15 genes are lost from significance relative to the Uniform Poisson model by the LNP/Gamma-Poisson models, respectively. Although both the Gamma-Poisson and LNP models find fewer KCGs containing hotspots than both conventional models, not all cancer genes driven by point mutations must have strong mutational hotspots. For example, many tumor suppressors can be inactivated via truncating mutations anywhere in their open reading frame; thus, while the gene as a whole is recurrently mutated, its mutations do not need to recur at the same specific genomic position to incur the same functional effect. A position with many truncating mutations would not have any additional fitness advantage over any other position with fewer truncating mutations, since any truncating mutations far enough upstream of the C terminus will either induce nonsense-mediated degradation of the mRNA or produce a nonfunctional partial protein product. Thus, recurrent mutations in cancer genes not driven by hotspots should not be considered false-negatives for an algorithm that solely evaluates mutations at the single-site level.

Although such genes may not be driven by hotspots, they will still display an overall excess of nonsilent mutations and should therefore be significant by a gene-level burden test. Of the 31 genes lost from significance under the LNP model, 23 are still significant by a conservative burden test (Prob[$dN/dS \geq 1.2$] $\geq$0.975). The additional eight genes have too few mutations to confidently establish an excess of nonsilent events by our conservative criterion but are all deemed significant by more sophisticated methods (e.g., MutSigCV [Lawrence et al., 2013], which incorporates genomic covariates to estimate gene-level background mutation frequencies).

Finally, we assess sensitivity on the individual allele level. When searching for totally uncharacterized driver events (i.e., in non-KCGs), the goal is to identify both hotspots and the genes that harbor them. However, many researchers are looking to prioritize yet-uncharacterized potentially actionable mutations in already characterized genes. To this end, to increase our statistical power, we restricted our hypothesis testing to only sites in KCGs. This yielded 775 hotspots (at the codon substitution level; Table S2) significant by any of the four methods ($q_{RHT} < 0.1$), 201 of which are not significant by the LNP model. However, most of these are loss-of-function events in tumor suppressors (determined analytically by enrichment in truncating mutations, Prob [$dT/dS \geq 1.2$] $\geq$ 0.975, Table S3); by the above reasoning, these should not be considered false-negatives in a site-specific significance analysis. Excluding nonsense mutations and tumor suppressors, the LNP model misses 70 hotspots, corresponding to a 9% loss in sensitivity under the very liberal assumption that all 775 hotspots identified by any method are indeed true driver hotspots. The Gamma-Poisson, Uniform-within-gene, and Uniform Poisson methods miss 8.9%, 7.5%, and 1% of the 775, respectively.

Thus, both the LNP and Gamma-Poisson models retain sensitivity on both the gene and allele levels if we specifically differentiate cancer genes driven by hotspots from genes driven by an overall excess of mutations.

Comparative Analyses Confirm Improved Performance of the LNP Model

To more rigorously assess the sensitivity and specificity tradeoffs of the different methods, we used receiver operating characteristic (ROC) curve analysis employing as ground truth sets: (i) a false-positive set of 194 genes confidently under neutral selection, as defined before (Prob[$0.8 \leq dN/dS \leq 1.2$] $\geq$95%); and (ii) a true-positive set of 42 genes, defined as KCGs with a high concentration of nonsilent mutations at recurrently mutated positions. We identified these 42 true-positive genes by requiring that the dN/dS of each gene dropped by more than 5% when we removed all sites that were near significance ($q \leq 0.25$) by the least conservative Uniform-within-gene model (Table S1). This ensures that it will attain 100% sensitivity by definition, setting a relevant point of comparison for the other models, because the Uniform-within-gene model has been widely employed (Van den Eynden et al., 2015; Lohr et al., 2012; Lawrence et al., 2014; Chang et al., 2016; Baeissa et al., 2017; Araya et al., 2016; Miller et al., 2015). However, we were aware that this true-positive set depends on a specific method, and thus assembled an alternate one comprising all KCGs with five or more recurrent mutations in our cohort, with evidence in the literature supporting an oncogenic role for the recurrent event. We only included genes whose hotspots' phenotypic effect was deemed consistent with primary tumor development, thus excluding mutations found only to confer resistance to treatment, or mutations found only to promote metastasis. We also only included genes tested in cell lines corresponding to the tissues in which the hotspot is observed *in vivo*. This alternate true-positive set comprised 53 KCGs (Table S1).

We then plotted ROC curves based on these ground truth sets for the four methods, using each method's minimum q value across all sites in the gene as the discrimination threshold (Figures 2A and S2A). We observed that the LNP model had the highest area under the curve (AUC), followed by the Gamma-Poisson, Uniform Poisson, and Uniform-within-gene methods. The differences in AUC were almost entirely due to differences in specificity, since all methods had comparable sensitivities. On each curve, we denote the positions on the ROC curves corresponding to the standard significance threshold of $q \leq 0.1$, at which the LNP model identified 494 significant hotspots in 134 genes. Of these 134 genes, none of the genes belong to the negative truth set; 39 genes, containing 169 hotspots, belong to the positive truth set, corresponding to a sensitivity of 89% ($CI_{95\%}$[75%, 95%]). At the same threshold, the Gamma-Poisson model achieved an insignificantly higher sensitivity of 95% ($CI_{95\%}$[85%, 99%]) but a false-positive rate of 4.1% ($CI_{95\%}$[1.8%, 7.29%]), corresponding to an FDR of

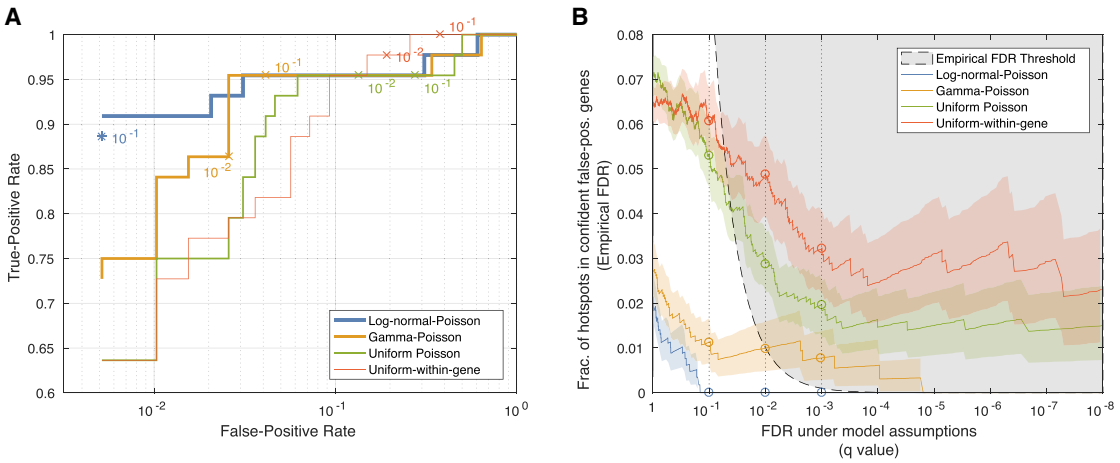


Figure 2. Performance of the Four Methods Quantified by ROC and FDR Analysis

(A) ROC curves for each method evaluated at the gene level. Truth set used for estimating false-positive rate comprises genes confidently under neutral selection; truth set used for estimating true-positive rate comprises KCGs with a high proportion of mutations at recurrently mutated positions. A gene is considered a hit for the ROC analysis if it has at least one hotspot more significant than a specific q value cutoff. q value cutoffs of 0.1 and 0.01 are marked on each curve. Because the Log-normal-Poisson model has a false-positive rate of 0 even at $q < 0.1$, we indicate this position in ROC space with an "*" marker at the lowest possible non-zero false-positive rate.

(B) Fraction of loci falling in false-positive truth set genes (empirical FDR) as a function of q value. The gray area indicates the region for which the empirical FDR exceeds the q value; methods whose curves lie in this region yield more false-positives than expected by the q value cutoff. Colored regions indicate 95% beta distribution confidence intervals on the fractions. q value thresholds of 0.1, 0.01, and 0.001 are shown as vertical dotted lines, with circles denoting where they intersect the curves.

See also Figure S2 and Table S1.

~16% (8 false-positive genes out of 50). The specificity losses for the conventional models are even higher, with false-positive rates of 25%/38% and FDRs of 56%/62% for the Uniform Poisson/Uniform-within-gene models, respectively. Re-running the benchmarks using the alternate true-positive set did not significantly affect the results (Figure S2B). This ROC analysis suggests that the LNP model performs the best among the four models, having the highest specificity without a significant loss in sensitivity for cancer genes driven by hotspots.

Another way to quantify the inflation of significant results of different methods is to examine their quantile-quantile (QQ) plots. Since we expect that most sites in the genome do not harbor driver events, we expect their p values to be uniformly distributed. Indeed, when comparing the four different methods, we observe that the QQ plots of the conventional methods are inflated, demonstrating deviation from the uniform distribution toward more significant p values in a large fraction of genomic loci (Figure S2C). The inflation of the models also affects the resulting q values and produces lists of significant hotspots (and genes) that may contain more false-positives than expected by the q value cutoff. In the case of a well-calibrated model (and hence a well-behaved QQ plot), setting a specific q value threshold (e.g., $q \leq 0.1$) would result in a list of significant hits that contain (on average) at most the desired fraction (10%) of false-positives. It is therefore important to test whether this is indeed the case.

We used our set of false-positives to measure the empirical FDR. Since this is a conservative list, we expect to have an even lower FDR than the chosen q value cutoff. We compare the empirical FDR as a function of q value among the four models (Figure 2B). The LNP model is the only model for which the empirical FDR did not exceed the desired FDR (Table 1). By contrast, even at the extreme q value threshold of 10^{-8}, approximately 3% of hotspots significant by the Uniform-within-gene or Uniform Poisson models are in false-positive genes. Thus, these data confirm that only the LNP model's q value cutoff properly bounds the FDR.

Finally, in addition to the ROC/QQ analyses, we present a nonparametric benchmark: the fraction of false-positives nominated by each method in the top *N* significant hotspots in non-KCGs. This reflects the approach of a researcher looking to prioritize putative drivers for experimental characterization, whose finite resources only allow for investigation of a fixed number of hotspots (e.g., 100). The method returning the fewest hotspots in the false-positive truth set would likely contain the most true-positives, since a set of hotspots containing fewer confident false-positives will contain more potential drivers.

Out of its top 100 hotspots in non-KCGs, the LNP model nominates zero occurring in a false-positive gene. The other three methods nominate between 5 and 9 false-positive hotspots. If we picked a gene at random out of the 19,315 genes we analyzed, there is a ~1% chance of it being in the false-positive truth set (194 genes). Thus, we see that the other methods' top 100 hotspots are significantly enriched for false-positives, and thus further depleted of potential true-positives.

The LNP Model Produced the Most Accurate Estimates of Neutral Mutation Frequency

In addition to evaluating model performance by examining the protein-coding effects of hotspots found significant by each model, we can also assess how well each model predicts the expected number of mutated patients at each genomic position.

Table 1. Fraction of Hotspots in Confident False-Positive Genes (Empirical FDR) at the Indicated q Value Cutoff

Method	$q \leq 0.1$	$q \leq 0.01$	$q \leq 0.001$
Uniform-within-gene	0.06 (102/1,678)	0.05 (45/920)	0.03 (19/590)
Uniform Poisson	0.05 (82/1,547)	0.03 (29/1,006)	0.02 (14/713)
Gamma-Poisson	0.01 (8/710)	0.01 (5/504)	0.01 (3/387)
Log-normal-Poisson	0.0 (0/511)	0.0 (0/290)	0.0 (0/191)

A well-calibrated model accurately infers the background mutation frequency at each position, so the expected mutation frequencies at positions under neutral selection will be concordant with the observed frequencies. On the other hand, a poorly calibrated model that inaccurately models the background frequency will predict mutational frequencies that significantly deviate from the observed frequencies.

To compare the accuracy of the different background models, we calculated the observed distribution of recurrent synonymously mutated sites (i.e., the fraction of sites with a specific number of mutations out of all sites that can harbor synonymous mutations) (Figure 3A). We then compared with the expected distributions predicted by each of the four models. While the conventional models underestimate the fraction of sites mutated in three or more patients—synonymous sites mutated in exactly three patients are 3.5× more likely than expected by the Uniform-within-gene model; overall, synonymous mutations recurring in ≥3 patients are 12× more likely than expected by the model—both overdispersed models (Gamma-Poisson and LNP) are more accurate, with the LNP model most correctly recapitulating the observed distribution even for highly recurrent events, with synonymous mutations recurring in ≥3 patients 1.84× more likely than expected by the Gamma-Poisson model, but only 1.05× more likely by the LNP model.

Although the two overdispersed models performed much better in predicting mutation frequencies than the non-overdispersed models, they are not identical. We illustrate a specific instance of this by looking at a sequence context containing recurrent mutations in likely passenger genes that only the LNP model avoids calling significant (Figure 3B). Three of the most recurrently mutated positions in the sequence context A(A→C)G occur in Spectrin alpha (*SPTA1*; ten missense mutations: six in stomach adenocarcinoma and one each in bladder, cervical, colon, and lung squamous tumors), Titin (*TTN*; seven missense mutations: three stomach, two colorectal, one each in lung squamous and esophageal tumors), and regulating synaptic exocytosis protein 2 (*RIMS2*; ten synonymous mutations: five stomach, three colon, one each in esophageal and liver tumors). All three of these mutations were found to be significant by the Gamma-Poisson model ($q = 5\times10^{-4}$, $q = 0.002$, and $q = 0.002$, respectively) and by both conventional models (q values $< 10^{-13}$ for all genes and models) but not by the LNP model ($q = 0.97$, $q = 0.72$, and $q = 1.0$, respectively). Because all of these genes are only expressed in specific tissue types (*SPTA1* in red blood cell progenitors, *TTN* in muscle cells, and *RIMS2* in neurons), it is highly likely that these represent passenger hotspots. Additional evidence that these genes are passengers is provided by the *dN*/*dS* analysis, which assigns the genes tight confidence intervals around 1 (*SPTA1*: mean *dN*/*dS* 1.05, $CI_{95\%}$ [0.92,1.19]; *TTN*: mean 0.98, $CI_{95\%}$ [0.94,1.02]; *RIMS2* mean 1.02, $CI_{95\%}$ [0.86,1.22]), and by the fact that the hotspot in *RIMS2* is synonymous. By contrast, three of the other most recurrently mutated positions in the context fall in known drivers (*EGFR*, *TP53*, and *KRAS*, mutated in 8, 13, and 14 patients, respectively), and are significant by all four methods. Thus, taken together, these data suggest that the overdispersed LNP model most accurately predicts actual mutation frequencies and makes the fewest false-positive calls.

Significant Hotspots Identified by the LNP Model

The main scientific interest of any significance analysis method is to discover promising but not yet experimentally validated driver candidates for subsequent experimental follow-up. Since the LNP model performed well in excluding many more false-positive passenger mutations than the other three tested models, we can have more confidence that the genes in the resulting list of still-significant hotspots are enriched for true drivers. The LNP model yielded 494 significant hotspots ($q \leq 0.1$) in 134 genes (49 KCGs, containing 405 hotspots, and 85 non-KCGs, containing 89 hotspots) (Table S2). The KCGs contain 29 of the conservative true-positive genes (including *KRAS*, *BRAF*, and *PIK3CA*) and also 20 other genes not in the truth set but still with significant hotspots, including *PTEN*, *SMAD4*, and *CDKN2A*.

Since we use an FDR threshold of $q \leq 0.1$, approximately 49 out of the 494 significant hotspots should be false-positives. If we assume that none of the 405 hotspots in KCGs are false-positives, then we expect that 40 (= 89 − 49) out of the 89 hotspots in 85 non-KCGs are true-positives.

Within the 85 non-KCGs (Figure 4), 26 have been previously experimentally implicated in cancer but are not yet well-known enough to be included in the CGC, or are in the CGC but not due to somatic point mutations (e.g., implicated by germline risk alleles or copy-number/structural alterations). We discuss these genes in depth in Table S4 and Figure S3.

We note that the three other models (and other conventional methods) also find all of our significant hotspots, but are diluted by many other false-positives. Overall, we demonstrate that the LNP model produces a highly accurate list of driver hotspots that provide clear biological hypotheses, some of which have already been supported by experimental data. Future experiments will be needed to validate the functional role of the other hotspots.

Hotspot-Generating Mutational Processes Have Similar Base-Wise Heterogeneity Despite Having Vastly Different Mutation Frequencies

Our results provide evidence of pervasive variability in base-wise mutation frequency across cancer, irrespective of mutagen or underlying mutational process. Next, we tested whether different mutational processes have different levels of variability. Each mutational process has specificity for particular genomic contexts and features, defining a set of "bases-at-risk" for that process. We might expect differences in the degree to which different mutational processes diverge from uniformly targeting their bases-at-risk and thus different levels of variability. Because there is considerable heterogeneity in

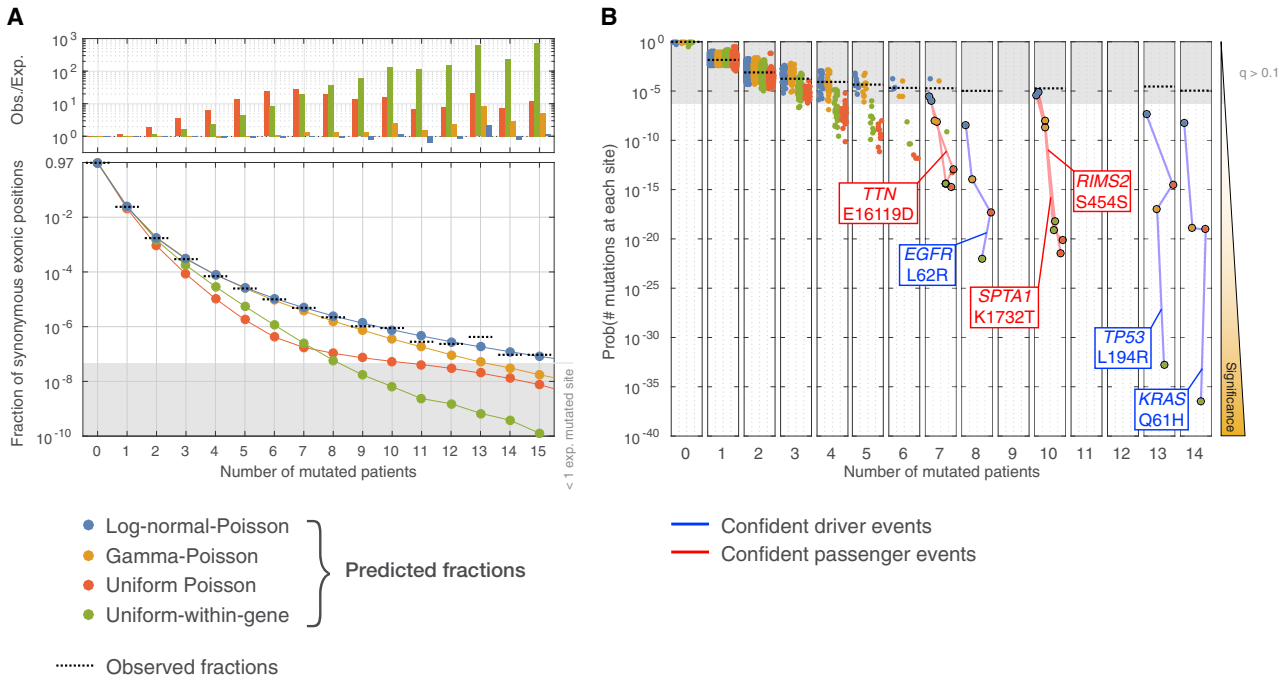


Figure 3. Predicted Frequencies of Recurrent Mutations According to the Four Methods

(A) Fraction of synonymous hotspot mutations observed in multiple patients (0–15; total cohort size 9,023). Colored lines represent the expected fraction of sites as predicted by each of the four models; dashed black lines represent the observed fractions. The ratios of observed:expected fractions for each model are plotted above each recurrence level. Log-normal-Poisson model best matches the observed fractions. The gray region indicates a fraction corresponding to <1 base pair of synonymous exonic territory (21.4 million possible base pair substitutions that could yield a synonymous codon change).

(B) Probabilities of observing every mutation in sequence context A(A→C)G as predicted by the four models; each dot corresponds to an observed mutation. Lines connect the different models' predictions for the same mutation; blue lines highlight confident driver mutations, while red lines highlight confident passenger mutations. The gray region indicates model predictions with a q value > 0.1 (i.e., non-significant recurrence), indicating that, while driver hotspot mutations are significant by all models, only the Log-normal-Poisson model correctly infers passenger hotspot mutations as non-significant.

the overall mutation frequencies of different processes (ranging from $\sim 10\,\text{muts}/\text{million bases} - \text{at} - \text{risk}$ for methylated-CpG deamination to $>5,000\ \text{muts}/\text{million bases} - \text{at} - \text{risk}$ for *POLE* hypermutation), we might predict a correspondingly wide range in the variability across bases-at-risk of different processes. Tellingly, many published analyses have been forced to exclude hypermutated tumors from significance analyses because they yield too many significantly mutated genes/hotspots, largely due to the inaccurate background models in wide use (Bailey et al., 2018; Martincorena et al., 2017). One might surmise from this that high-mutation-frequency processes are more prone to generating hotspots, reflecting greater variability in their base-wise mutation frequencies.

The LNP framework provides a natural way of quantifying both the mutation frequency and variability of different mutational processes. We fit the model to all bases-at-risk for each process; the log-normal parameters e^{μ} and e^{σ} are equal to the geometric mean and geometric standard deviation, respectively, of the base-wise mutation frequency. The geometric mean is simply equivalent to the median of the log-normal distribution. The geometric standard deviation e^{σ} is a dimensionless scale factor indicating the average multiplicative distance from the mean; for example, $e^{\sigma} = 2$ signifies that at 1 standard deviation above the mean, bases will be twice as mutable as the average. At its minimum value $e^{\sigma} = 1$, there is no variability. Hence, the same value of e^{σ} indicates an equivalent amount of base-wise variability, irrespective of mutation frequency. Our fully Bayesian model finds not only the optimal values of e^{μ} and e^{σ}, but also the posterior distributions of these parameters. In general, mutational processes that generate few mutations will have high posterior uncertainty, while processes that generate many mutations will have lower uncertainty.

We selected the following eight mutational processes to examine using the LNP model because they are mostly of known causes and very specific to certain patients and sequence contexts, making assignments of patients to these signatures unambiguous: APOBEC (3A+3B, COSMIC Signatures 2 and 13), aging (spontaneous methylated-CpG deamination, COSMIC Sig. 1), esophageal (Dulak et al., 2013) (COSMIC Sig. 17), MSI (COSMIC Sig. 6), POLE (COSMIC Sig. 10), POLE + MSI (Haradhvala et al., 2018) (COSMIC Sig. 14), smoking (COSMIC Sig. 4), and UV (UV-A only, COSMIC Sig. 7). We used SignatureAnalyzer (Kim et al., 2016), which is based on a Bayesian implementation of non-negative matrix factorization, to infer the probabilities of each mutation being assigned to each process. We then identified 8 subcohorts of patients (each comprising between 44 and 4,739 patients) in which each of these processes dominated ($\geq$75% assignment probability) at their relevant sequence contexts (e.g., C→T mutations at CpG sites for aging). The mutational spectra of these subcohorts are shown in Figure S4. For each

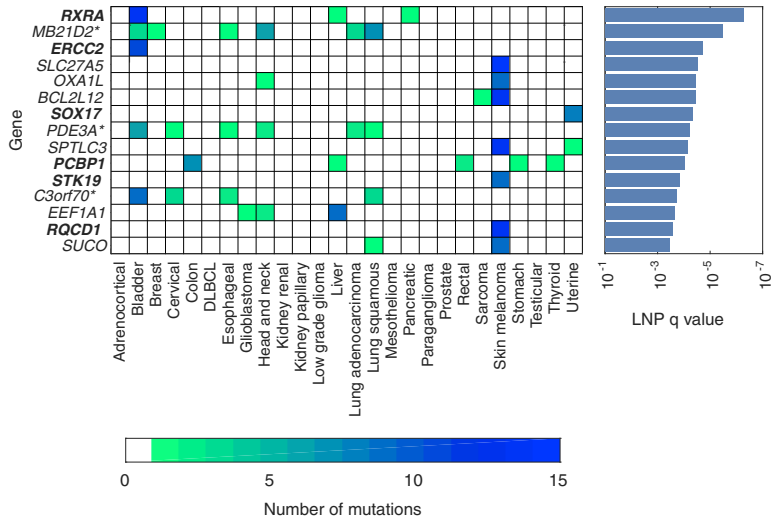


Figure 4. The Top 15 Non-KCGs with Significant Hotspots by the LNP Model (q < 0.1) Segregated by Tumor Type

For each non-KCG-containing significant hotspots, we indicate the total number of mutated patients within each tumor type at all significant hotspots. Genes in bold have been previously experimentally characterized as cancer genes but are not in the CGC. Genes marked with a star have hotspots that occur at optimal APOBEC3A hairpin substrate sequences and are likely passengers. See also Tables S2 and S4, and Figure S3.

of these "process-centric" subcohorts, we fit the LNP model to the relevant contexts.

We plotted the posterior distributions of e^{μ} and e^{σ} for each pentamer context belonging to the eight mutational processes to test whether processes with higher mutation frequencies also have higher base-wise variability. To our surprise, our results show that, despite extreme heterogeneity in overall mutation frequency (spanning nearly 5 orders of magnitude), most mutational processes show a similar amount of non-zero base-wise variability (Figure 5A), with e^{σ} approximately between 2 and 3.

Notably, the only exception was the esophageal mutational process, which showed the highest variability despite having one of the lowest mutation frequencies. This may indicate that additional yet-undiscovered factors may correspond to elevated mutability at specific bases-at-risk for this process. It was recently reported that the esophageal signature disproportionately mutates positions within CTCF binding sites (Katainen et al., 2015), possibly due to bound CTCF transcription factors occluding damaged bases from repair processes.

To test whether variability was independent of mutation rate even within individual processes, we partitioned the high mutation frequency processes between hypermutated samples (top decile of mutation frequency) and non-hypermutants, and compared their e^{σ} values. We observed similar levels of base-wise variability between the two partitions (Figure S5A), indicating that there is nothing unusual about the distribution of mutations in hypermutants, and nothing that should warrant their exclusion in significance analyses.

Explanatory Power of Genomic Covariates Differs among Mutational Processes

Although we previously showed in the Uniform Poisson regression model analysis that genomic covariates cannot fully explain all base-wise mutational variability, their explanatory power is non-zero. The LNP model provides a natural way to quantify the contribution of each covariate toward explaining this variability: the amount that e^{σ} decreases as we incorporate additional covariates corresponds directly to the variance explained by the added covariates. The value of e^{σ} after all covariates have been incorporated is the unexplained variance, which would approach $e^{\sigma} = 1$ if the covariates completely explained the base-wise variability, because there would be no extra variance beyond what the covariates predict.

We grouped the covariates by genomic scale: replication timing and expression (Lawrence et al., 2013) influence mutation frequencies on a coarse scale (~100 kbp–1 Mbp), while nucleosome positions and DNase hypersensitivity influence mutation frequencies on a fine scale (~10 bp). We also quantified the effect of accounting for pentamer context specificity on mutation frequency (i.e., considering the flanking ±2 upstream/downstream positions in addition to the immediate 5′/3′ positions).

In Figure 5B, we plot the amount of total variance explained by the aforementioned covariate sets and the amount of remaining unexplained variance for each of the trinucleotide contexts associated with each mutational signature. From this analysis, the following universal patterns stand out: (1) unexplained variance is higher than explained variance; (2) pentamer contexts are almost always important (with the notable exception of the aging signature); and (3) there is no correlation between the amount of unexplained variance and the amount of explained variance, and processes with more total variance do not necessarily have more variance explained by covariates.

There are likely yet-undiscovered properties of the genome that affect mutability. For example, we examined the effect of adding XR-seq coverage (Hu et al., 2015)—an extrinsic measurement of nucleotide excision repair (NER) activity, used to repair pyrimidine dimers resulting from UV damage—as a covariate in our model. We found that this covariate explained an average additional 10% of variance in the UV subcohort (Figure S5B), indicating that many passenger UV hotspots occur at loci predictably refractory to NER. However, the intrinsic genomic properties that determine amenability to NER are unknown.

For certain processes, standard genomic covariates are completely non-explanatory. Although certain APOBEC trinucleotides display considerable variability among pentamer contexts (Chan et al., 2015), neither coarse nor fine covariates explain any additional variability. However, Buisson et al. (2019) recently showed that the APOBEC3A (A3A) cytidine deaminase preferentially targets DNA hairpin loops. Cytosines residing at optimal hairpin loop positions/motifs can show a >200× increase

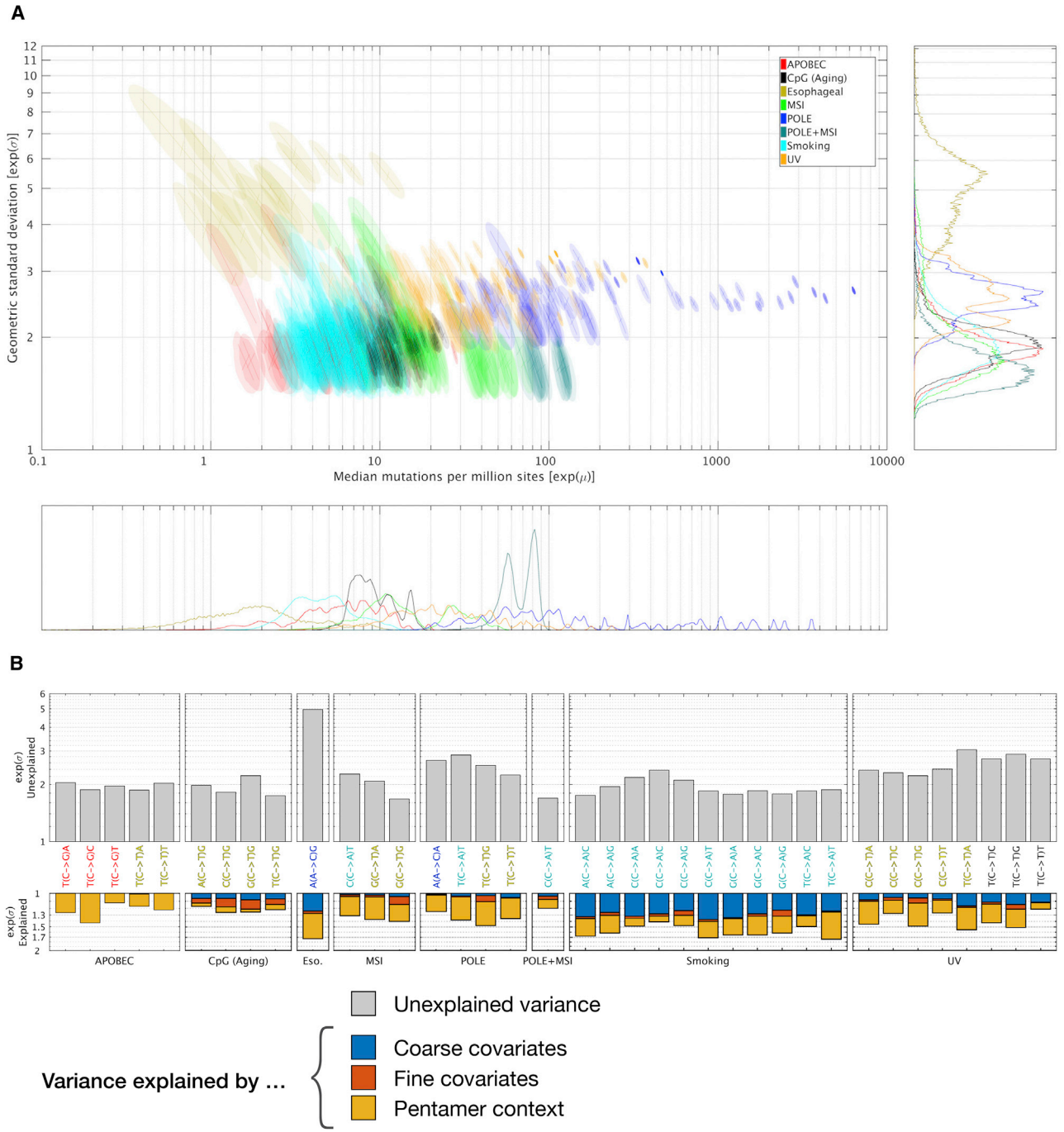


Figure 5. Comparison of the Heterogeneity of Base-Wise Mutability for Different Mutational Processes as Inferred by the Log-Normal-Poisson Model

(A) Log-normal-Poisson posterior distributions of base-wise mutation frequency (median mutations per million sites e^{μ}) and mutations' deviation from being uniformly Poisson distributed (geometric standard deviation e^{σ}) for different mutational processes. Each colored area represents the posterior 95% confidence region for a pentamer context associated with a given mutational process (Figure S4). $e^{\sigma} = 1$ corresponds to uniformly Poisson distributed mutations. Since $e^{\sigma}>1$ for all processes, we see that base-wise variability is universal and pervasive. Marginal distributions of e^{μ} and e^{σ} are shown below and to the right of the plot, respectively. Lines within each region are principal axes of the posterior density (i.e., eigenvectors of the covariance matrix of the estimated posterior density). (B) Amount of base-wise mutation rate variability e^{σ} explained by model covariates (colored), and remaining unexplained variance after all covariates have been incorporated (gray), for each relevant trinucleotide context in each mutational process.

See also Figures S4 and S5.

in A3A-induced mutation rate. Incorporating A3A substrate optimality, as defined in Buisson et al. (2019), as a covariate into our model removes several A3A mutational hotspots from significance (Figure S5C, Table S2).

Conversely, coarse covariates—namely gene expression—explain a substantial amount of smoking mutational variability. This is expected since the C→A mutations comprising the smoking signature are caused by benzo[α]pyrene guanine adducts (Denissenko et al., 1996), which are often corrected by transcription-coupled repair (Fousteri and Mullenders, 2008), which occurs more frequently in highly expressed genes (Pleasance et al., 2010).

DISCUSSION

Discovering cancer drivers from sequencing data requires overcoming the problem of detecting signal (mutational recurrence) above a background of non-random noise (variable intrinsic mutability). We detect drivers by statistically modeling this background and looking for recurrence that significantly exceeds it. This approach is only fruitful if the underlying model is accurate. As the corpus of cancer genomes has exponentially grown over the last decade, we have become statistically powered to observe background mutational variability at increasingly fine genomic scales, which we must accordingly account for in our background models. Initially, around 2007, cohorts were so small (~10 patients) that we lacked the power to observe any variability at all, although we suggested that this might indeed be the case (Getz et al., 2007). Thus, models assumed that every genomic region was equally mutable, and that all recurrent mutation stemmed from positive selection (Sjöblom et al., 2006). As cohorts grew to ~1,000 patients, around 2013, we became powered to observe heterogeneity on the scale of a gene, and infer that most recurrently mutated genes were actually passengers with high intrinsic mutability. However, cohorts were still too small to estimate mutability on levels smaller than a gene, so models did not account for it. This led to the assumption that any base pair recurrently mutated beyond the overall background level of its gene had to be a driver (Van den Eynden et al., 2015; Lohr et al., 2012; Lawrence et al., 2014; Chang et al., 2016; Araya et al., 2016; Baeissa et al., 2017).

We are able to challenge this assumption in the current era of cohorts comprising ~10,000 patients, which powers us to estimate background mutability on the smallest possible genomic scale: that of the individual base pair. Here, we demonstrate that base-wise mutational variability is so extreme that a large proportion of recurrently mutated base pairs in non-KCGs are in fact passengers. Unlike most hotspots found significant by our LNP model, most significant hotspots under the two toy models we compare it against (Uniform-within-gene and Uniform Poisson) are under no positive selection by orthogonal criteria.

Although our toy models are merely illustrative, many recently published algorithms intended for actual driver discovery (both at the gene and site level) do not account for base-wise variability in their background models. Using these naive models, various groups reported lists of significant genes or hotspots with many likely false-positives. Our own MutSigCL method (Lawrence et al., 2014), which analyzes mutational clustering on the gene level, reports a long list of significant genes at FDR < 10% when run on the cohort analyzed here, some of which fall in our false-positive truth set. Likewise, the statistical model used in another study (Chang et al., 2016) yielded 1,202 significant hotspots at FDR < 1%, which required additional analyses and manual review to yield a final report of 470 hotspots. Since FDRs are only meaningful if their underlying statistics are well calibrated, rigorous validation of the statistical models should be performed whenever the significance levels are overly confident. Failure to do so has potentially severe consequences in situations like identifying genes for deep experimental follow-up or as drug targets. Owing to the substantial cost of such follow-up experiments, properly accounting for base-wise variability is essential when selecting these candidates to avoid wasting valuable scientific resources on passenger hotspots.

In addition to providing an improved model for assessing mutational significance, our LNP model also sheds light on the fundamental nature of mutagenesis. It has long been known that background mutability correlates with coarse-scale genomic features, and more recent studies have shown that specific fine-scale genomic features undergo localized increased mutability (Mao et al., 2018; Poulos et al., 2016; Sabarinathan et al., 2016; Katainen et al., 2015). We show here that neither of these factors can explain the amount of observed base-wise variability, suggesting that there are yet-undiscovered properties of the genome that affect mutability. These may include: (1) other secondary/tertiary DNA structural motifs (Harteis and Schneider, 2014; Georgakopoulos-Soares et al., 2018); (2) yet-unknown proteins bound to the DNA; (3) a combination of the two (e.g., binding of ETS transcription factors has been shown to rotate adjacent pyrimidines into a more favorable conformation to form a cyclobutane dimer [Mao et al., 2018]); (4) local chromatin structure (which could affect gene expression or accessibility to mutagens or repair enzymes); or (5) sequence-specific polymerase error modes that our datasets are not yet powered to detect.

Accounting for all potential covariates, however, may never be able to fully explain the observed variability since the mutations we observe in a tumor genome are merely a snapshot of the aggregated effect of many fluctuating mutational processes that have been active over the course of the tumor's life history. These processes' activity levels and bases-at-risk vary as a function of continuously changing factors, such as the tumor's microenvironment, mutagen exposure, or epigenetic state, which are complicated to model with static covariates. We probabilistically represent the unexplained variability with a log-normal distribution because it represents the net product of consecutive molecular events. The log-normal distribution has theoretical justification and currently best fits the observed data. But, as datasets grow in size, increasing our power to distinguish low-frequency drivers from background, we also become more powered to infer the distribution of background mutability, which may require us to further refine our probabilistic models.

Regardless of the underlying causes of the dramatic heterogeneity we observe in base-wise mutability, its existence has implications in the fields of molecular evolution and population genetics. We find that base-wise heterogeneity is pervasive across all mutational processes, including methylated-CpG deamination, which is overwhelmingly responsible for *de novo* germline mutations (Ehrlich and Wang, 1981; Hodgkinson and

Eyre-Walker, 2011). This suggests that the infinite-sites model underpinning many assumptions in population genetics may be incorrect—for instance, the probability of identical alleles in multiple unrelated individuals originating from different common ancestors may be much higher than a naive coalescent theory would predict. Many somatic methods also rely on the infinite-sites model; for example, tools for clonal structure inference often impose a hard constraint that the same mutation can never arise in multiple subclones independently.

In addition, large-scale genomic organization has been thought to reflect coarse variability in background mutability (Chuang and Li, 2004), wherein genes more tolerant of mutation are thought to reside in more highly mutable regions of the genome. Variability at the base pair level may equivalently mold genomic architecture on fine scales. For example, it has long been speculated that the sequence composition of immunoglobulin variable chains is specifically biased to induce AID hypermutation hotspots at positions that accelerate the process of antigen selection (Jolly et al., 1996). Genome-wide selective pressure may be analogously guided or constrained due to variable site-specific mutability.

In conclusion, further growth of cancer sequencing datasets will allow us to survey the landscape of drivers with even greater precision, reveal the intricacies of mutational processes, and even elucidate how these mutational processes shape evolutionary selection. But, as datasets continually grow and evolve, so must our methods and the conclusions we draw from them.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Mutation Calling and QC
  - Definition of Known Cancer Genes
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Calculating Expected Coding Effect Fractions
  - Gene dN/dS Calculation
  - Description of Significance Methods
  - MCMC Implementation
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.ccell.2019.08.002.

ACKNOWLEDGMENTS

We thank Daniel Rosebrock and Dimitri Livitz for providing *MYC* absolute copy-number calls, Yosef E. Maruvka for providing microsatellite indel calls with respect to *UPF2* mutant status, Esther Rheinbay for providing TFBS activity and chromatin state tracks, as well as Chip Stewart and Barry Taylor for helpful comments and fruitful discussion on the manuscript. J.M.H. and G.G. were partially funded by NCI GDAC grants (U24CA143845, U24CA210999). M.S.L. was partially funded by G.G. funds at the Broad Institute and M.S.L. startup funds at Massachusetts General Hospital. G.G. was partially funded by Paul C. Zamecnik Chair in Oncology, MGH Cancer Center.

AUTHOR CONTRIBUTIONS

Conceptualization, J.M.H., M.S.L., and G.G.; Methodology, J.M.H. and G.G.; Software, J.M.H.; Formal Analysis and Investigation, J.M.H.; SignatureAnalyzer Results, J.K.; Data Curation, J.M.H., M.S.L., and A.B.; Visualization, J.M.H.; Writing – Original Draft, J.M.H., M.S.L., and G.G.; Writing – Review & Editing, G.G., M.S.L., M.M., A.T.-W., and N.J.H.; Supervision, G.G. and M.S.L.; Funding Acquisition, G.G.

DECLARATION OF INTERESTS

G.G. receives research funds from IBM and Pharmacyclics. G.G. is an inventor on patent applications related to MuTect, MutSig, ABSOLUTE, and POLYSOLVER. M.S.L. is an inventor on patent applications related to MuTect and MutSig.

Received: November 21, 2018
Revised: May 15, 2019
Accepted: August 6, 2019
Published: September 16, 2019

REFERENCES

Araya, C.L., Cenik, C., Reuter, J.A., Kiss, G., Pande, V.S., Snyder, M.P., and Greenleaf, W.J. (2016). Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nat. Genet. *48*, 117–125.

Baeissa, H., Benstead-Hume, G., Richardson, C.J., and Pearl, F.M.G. (2017). Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. Oncotarget *8*, 21290–21304.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18.

Buisson, R., Langenbucher, A., Bowen, D., Kwan, E.E., Benes, C.H., Zou, L., and Lawrence, M.S. (2019). Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science *364*, https://doi.org/10.1126/science.aaw2872.

Cairns, J. (1975). Mutation selection and the natural history of cancer. Nature *255*, 197–200.

Chan, K., Roberts, S.A., Klimczak, L.J., Sterling, J.F., Saini, N., Malc, E.P., Kim, J., Kwiatkowski, D.J., Fargo, D.C., Mieczkowski, P.A., et al. (2015). An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat. Genet. *47*, 1067–1072.

Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J.J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. *34*, 155–163.

Chuang, J.H., and Li, H. (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS Biol. *2*, 253–263.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. *41*, e67.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

Denissenko, M.F., Pao, A., Tang, M., and Pfeifer, G.P. (1996). Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. Science *274*, 430–432.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast computation and applications of genome mappability. PLoS One *7*, e30377.

Dulak, A.M., Stojanov, P., Peng, S., Lawrence, M.S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S.E., Shefler, E., et al. (2013). Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat. Genet. *45*, 478–486.

Ehrlich, M., and Wang, R.Y. (1981). 5-Methylcytosine in eukaryotic DNA. Science *212*, 1350–1357.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. *6*, 271–281.e7.

Fousteri, M., and Mullenders, L.H.F. (2008). Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. Cell Res. *18*, 73–84.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M., and Nik-Zainal, S. (2018). Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res. *28*, 1264–1271.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91–100.

Getz, G., Höfling, H., Mesirov, J.P., Golub, T.R., Meyerson, M., Tibshirani, R., and Lander, E.S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers". Science *317*, 1500.

Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R., and Easton, D.F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics *173*, 2187–2198.

Haradhvala, N.J., Kim, J., Maruvka, Y.E., Polak, P., Rosebrock, D., Livitz, D., Hess, J.M., Leshchiner, I., Kamburov, A., Mouw, K.W., et al. (2018). Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat. Commun. *9*, 1746.

Harteis, S., and Schneider, S. (2014). Making the bend: DNA tertiary structure and protein-DNA interactions. Int. J. Mol. Sci. *15*, 12335–12363.

Hodgkinson, A., and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. *12*, 756–766.

Hu, J., Adar, S., Selby, C.P., Lieb, J.D., and Sancar, A. (2015). Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev. *29*, 948–960.

Imielinski, M., Guo, G., and Meyerson, M. (2017). Insertions and deletions target lineage-defining genes in human cancers. Cell *168*, 460–472.e14.

Jolly, C.J., Wagner, S.D., Rada, C., Klix, N., Milstein, C., and Neuberger, M.S. (1996). The targeting of somatic hypermutation. Semin. Immunol. *8*, 159–168.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821.

Katz, L., and Burge, C.B. (2003). Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. *13*, 2042–2051.

Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Tiao, G., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A.D., et al. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. *48*, 600–606.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature *267*, 275–276.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lohr, J.G., Stojanov, P., Lawrence, M.S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y.W., Slager, S.L., et al. (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc. Natl. Acad. Sci. U S A *109*, 3879–3884.

Mao, P., Brown, A.J., Esaki, S., Lockwood, S., Poon, G.M.K., Smerdon, M.J., Roberts, S.A., and Wyrick, J.J. (2018). ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. Nat. Commun. *9*, 2626.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal patterns of selection in cancer and somatic tissues. Cell *171*, 1029–1041.e21.

Miller, M.L., Reznik, E., Gauthier, N.P., Aksoy, B.A., Korkut, A., Gao, J., Ciriello, G., Schultz, N., and Sander, C. (2015). Pan-cancer analysis of mutation hotspots in protein domains. Cell Syst. *1*, 197–209.

Mitsui, J., and Tsuji, S. (2012). Common chromosomal fragile sites: breakages and rearrangements in somatic and germline cells. Atlas Genet. Cytogenet. Oncol. Haematol. https://doi.org/10.4267/2042/46078.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418–426.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordóñez, G.R., Bignell, G.R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M.S., Reynolds, A., Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J.A., et al. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature *518*, 360–364.

Poulos, R.C., Thoms, J.A.I., Guan, Y.F., Unnikrishnan, A., Pimanda, J.E., and Wong, J.W.H. (2016). Functional mutations form at CTCF-cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif. Cell Rep. *17*, 2865–2872.

Quax, T.E.F., Claassens, N.J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. Mol. Cell *59*, 149–161.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature *532*, 264–267.

Schrock, M.S., and Huebner, K. (2015). WWOX: a fragile tumor suppressor. Exp. Biol. Med. *240*, 296–304.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268–274.

Smith, T.C.A., Carr, A.M., and Eyre-Walker, A.C. (2016). Are sites with multiple single nucleotide variants in cancer genomes a consequence of drivers, hypermutable sites or sequencing errors? PeerJ *4*, e2391.

Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nat. Genet. *41*, 393–395.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.

Sutton, J. (1997). Gibrat's legacy. J. Econ. Lit. *35*, 40–59.

Van den Eynden, J., Fierro, A.C., Verbeke, L.P.C., and Marchal, K. (2015). SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinformatics *16*, 1–12.

Weghorn, D., and Sunyaev, S. (2017). Bayesian inference of negative and positive selection in human cancers. Nat. Genet. *49*, 1785–1788.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Log-normal-Poisson regression algorithm	This paper	https://github.com/broadinstitute/getzlab-LNP
Code to reproduce this paper's figures and analyses	This paper	https://github.com/broadinstitute/getzlab-PHS
Other		
TCGA MC3 somatic mutation callset (protected)	TCGA	https://www.synapse.org/#!Synapse:syn5917256
TCGA MC3 somatic mutation callset (unprotected)	TCGA	https://www.synapse.org/#!Synapse:syn7824274
TCGA hg19 BAM files	TCGA	https://portal.gdc.cancer.gov/legacy-archive/

LEAD CONTACT AND MATERIALS AVAILABILITY

Inquiries on materials availability should be directed to corresponding author Gad Getz (gadgetz@broadinstitute.org).

Both the exact TCGA MC3 mutation calls used in this paper and the BAM files used to generate/QC them are protected data, requiring appropriate dbGaP authorization to access. We have also included a link to an unprotected version of the MC3 callset that will yield essentially identical results.

METHOD DETAILS

Mutation Calling and QC

We obtained the MC3 TCGA mutation calls from Synapse, a cohort comprising 10,510 unique patients (3,850,525 somatic single-nucleotide variations [sSNVs] called by two or more mutation callers) before filtering. MC3 is a curated set of high-quality mutation calls run through numerous QC filters with failing samples/mutations annotated in the mutation annotation file (MAF), allowing the end user to exclude them.

We excluded all patients failing MC3 sample-level filters (i) excessive cross-sample contamination (MAF filter `contest`); (ii) whole genome amplified libraries (MAF filters `wga_only`, `native_wga_mix`); and (iii) bad sequencing runs due to mid-run sequencer failure (MAF filter `badseq`). This left us with a cohort of 9,023 patients, from which we further excluded mutations failing the following MC3 site-level filters:

(i) Panel-of-normals (PoN): for each genomic position, the PoN encodes the distribution of alternate read fractions (AFs) across ≈ 8,000 TCGA normals. We removed candidate variant calls that occurred at sites recurrently harboring alternate reads across the PoN whose AFs were consistent with the candidate variant. For a full description of the panel-of-normals filter, see the Supplemental Information of (Ellrott et al., 2018). (MAF filter `broad_PoN_v2`).
(ii) ExAC: site appeared in ExAC (Lek et al., 2016) (≈ 60,000 normal samples) with allele count ≥ 50. This gave us greater power than the PoN to filter germline polymorphisms somehow missing from the matched normal. (MAF filter `common_in_exac`).
(iii) Low normal coverage: sequencing coverage of the matched normal at the position of the somatic variant was fewer than eight reads, making it difficult to distinguish the somatic call from a rare germline polymorphism. (MAF filter `ndp`)
(iv) OxoG: variant was likely a sequencing artifact caused by oxidative damage during shearing in library preparation. For a full description of this artifact mode, see (Costello et al., 2013). (MAF filter `oxog`).

In total, 861,664 failing mutations were excluded, leaving us with 2,988,861 passing sSNVs.
In addition to removing MC3 flagged calls, we applied two additional filters not employed in MC3:

(i) We performed another round of PoN filtering using a panel comprising exomes captured using Illumina's ICE protocol, which was used only for TCGA samples sequenced late in the project. The ICE protocol introduces recurrent artifacts not observed in exomes captured using Agilent's capture kit, which was used for the majority of TCGA and comprises the entirety of the PoN used by MC3.
(ii) We found that many recurrent mutations were in fact false positives caused by read misalignments, verifying this by BLATing reads supporting the variant calls and seeing that they mapped with fewer mismatches elsewhere in the reference genome. To

mitigate this problem, we excluded calls in 75mer windows (chosen to match the typical read length of TCGA exomes) that were not completely unique (Derrien et al., 2012) in the reference.

This left us with 2,288,080 analysis-ready sSNVs.

Finally, we annotated all mutations' gene names and protein-coding effects according to our own transcript definitions, namely all GENCODE v19 protein coding transcripts with unambiguous translation start/stop sites and splice sites.

We re-annotated because accurate calculation of dN/dS (STAR Methods, Gene dN/dS Calculation) requires knowing the precise number of sequence contexts in each gene that can give rise to each coding effect (e.g., number of T(C→T)G sites that yield synonymous mutations), weighted by sequencing coverage across the cohort. Such exact transcript definitions were unavailable in MC3, whose calls were annotated using a pipeline that does not expose them. We weighted sequence context/protein coding effect territory by coverage because some context/effect combinations, although nominally possible based on a gene's open rading frame (ORF) sequence, occur at positions that consistently lack sequencing coverage, thereby deflating mutation frequency estimates. For each coding position, we computed the fraction of sufficiently covered tumor/normal pairs across 7,732 TCGA whole exome pairs. We used the binary criterion of sufficient coverage employed by MuTect (Cibulskis et al., 2013), i.e. 14 high-quality reads/bases in the tumor, 8 high-quality reads/bases in the matched normal. This allowed us to compute coverage-weighted context/effect territories by weighting each ORF position yielding a given context/effect by the fraction of sufficiently covered samples at that position.

Furthermore, these exact transcript definitions are required for most conservatively estimating overall protein-coding effect distributions (STAR Methods, calculating expected coding effect fractions), which requires accounting for loci in overlapping transcripts that yield different protein-coding effects. We annotated mutations with the most deleterious effect across all transcripts at each position (synonymous < missense < nonsense) to avoid erroneously misclassifying mutations that are synonymous on one transcript but protein-altering on a different overlapping transcript.

Definition of Known Cancer Genes

We assembled our list of Known Cancer Genes from the Cancer Gene Census (CGC) (v85) (Futreal et al., 2004), which comprises 719 genes. However, our KCG list is a subset of 133 genes, since not all genes listed in the CGC are relevant to this study, which only considers somatic single nucleotide variants (sSNVs). The CGC indexes many genes driven by a variety of mechanisms outside the scope of this study, such as genes only driven by structural variants (e.g., *MYC*), or genes that are only germline risk factors (e.g., *APOBEC3B*). In addition, many genes in the CGC are only reported as drivers in tumor types not included in TCGA (e.g., many blood cancers). Finally, some CGC genes are only drivers in non-primary tumors (e.g., relapse or post-treatment resistance tumors). We thus only considered genes explicitly listed in the CGC as drivers of primary tumors in the 32 tumor types in our cohort, excluding genes driven only by structural variants.

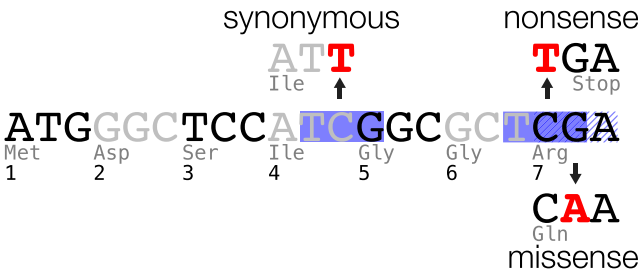
QUANTIFICATION AND STATISTICAL ANALYSIS

Calculating Expected Coding Effect Fractions

Given a set of genes and their observed coding mutations, we wish to calculate the distribution of the expected fractions of protein-coding effects (i.e., synonymous, missense, nonsense), assuming the mutations were distributed uniformly throughout the coding sequences of these genes, conserving sequence context (in order to normalize for mutational processes).

For example, in the cartoon coding sequence, the three effects of all possible strand-collapsed T(C→T)G mutations (highlighted in blue) are: (i) a synonymous substitution at codon 4 (Ile.4→Ile); (ii) a missense substitution at codon 7 (Arg.7→Gln, note mutation is on the reverse complement strand, shown as hatching); and (iii) a nonsense substitution also at codon 7 (Arg.7→Stop).

Assuming every TCG in the cartoon sequence is equally mutable, we would therefore expect random T(C→T)G mutations to result in an even distribution of coding effects — 1/3 synonymous, 1/3 missense, and 1/3 nonsense substitutions.



We can calculate the expected distribution of coding effects across a set of genes, if we assume that the variability of mutation rates across genes is independent from the gene-specific tendency to generate synonymous, missense, or nonsense mutations. Given a large enough set of genes, this assumption is reasonable.

For each trinucleotide context t, we calculate $F_t(e_j|c_i)$, the fraction of genomic positions that match t across the whole exome (or a subset of genes, e.g., known cancer genes only) that would yield coding effect e_j, given base change c_i, where $j \in \{1,2,3\}$ corresponds to the three coding effects (synonymous/missense/nonsense), and $i \in \{1,2,3\}$ corresponds to the three possible base substitutions. We compute these fractions by enumerating the codon changes for every possible substitution at every coding position in our gene set.

For a given set of observed coding mutations at context t within this gene set, we denote the fraction that are of the ith base change as f_{t,c_i}. For example, if 95% of observed mutations at t = TCG are c_i = C→T, then $f_{\mathrm{TCG,C\to T}} = 0.95$.

For each of the three coding effects e_j, the joint probability of seeing $\mathbf{n}_{t,e} \equiv \{n_{t,e1}, n_{t,e2}, n_{t,e3}\}$ mutations given N_t total mutations within context t is multinomial distributed

$$\mathbf{n}_{t,e}|\mathbf{F}_t, \mathbf{f}_t, N_t \sim \mathrm{mult}\left(\underbrace{\sum_i F_t(e_1|c_i) f_{t,c_i}}_{p(e_1|\mathbf{f}_t)}, \sum_i F_t(e_2|c_i) f_{t,c_i}, \sum_i F_t(e_3|c_i) f_{t,c_i}, N_t\right).$$

However, because some trinucleotide contexts have low mutation counts for a given base change, there can be considerable uncertainty on f_{t,c_i}. We quantify this uncertainty by modeling the f's not as single values, but drawn from a distribution. The ideal distribution for the f's is a Dirichlet distribution, which quantifies the uncertainty on multinomial probabilities derived from counts, defining the hierarchical model

$$\begin{aligned} \mathbf{n}_{t,e}|\mathbf{F}_t, \mathbf{f}_t, N_t &\sim \mathrm{mult}(p(e_1|\mathbf{f}_t), p(e_2|\mathbf{f}_t), p(e_3|\mathbf{f}_t), N_t) \\ \mathbf{f}_t|\mathbf{n}_{t,c} &\sim \mathrm{dir}(n_{t,c_1}+1, n_{t,c_2}+1, n_{t,c_3}+1). \end{aligned}$$

We obtain the probability distribution on $\mathbf{n}_{t,e}$ by marginalizing over $\mathbf{f}_t$, i.e.,

$$\begin{aligned} p(\mathbf{n}_{t,e}|\mathbf{n}_{t,c}, N_t) &= \int d\mathbf{f}_t \mathrm{mult}(n_{t,e_1}, n_{t,e_2}, n_{t,e_3}|p(e_1|\mathbf{f}_t), p(e_2|\mathbf{f}_t), p(e_3|\mathbf{f}_t), N_t) \\ &\quad \times \mathrm{dir}(f_{t,c_1}, f_{t,c_2}, f_{t,c_3}|n_{t,c_1}+1, n_{t,c_2}+1, n_{t,c_3}+1) \\ &\propto \int d\mathbf{f}_t \prod_{j=1}^{3} \frac{p(e_j|\mathbf{f}_t)^{n_{t,e_j}}}{n_{t,e_j}!} \prod_{j=1}^{3} f_{t,c_j}^{n_{t,c_j}} \end{aligned} \qquad \text{(Equation 1.1)}$$

Since there is no closed form for this integral over the 2D Dirichlet simplex, we compute it by sampling the full joint distribution $p(\mathbf{n}_{t,e}, \mathbf{f}_t|\mathbf{n}_{t,c}, N_t)$ via Markov Chain Monte Carlo (MCMC), and only considering the values drawn for $\mathbf{n}_{t,e}$. Our full conditionals are

$$\begin{aligned} p(\mathbf{n}_{t,e}|-) &\propto \prod_{j=1}^{3} \frac{p(e_j|\mathbf{f}_t)^{n_{e_j}}}{n_{e_j}!} \\ &\propto \mathrm{mult}(n_{t,e_1}, n_{t,e_2}, n_{t,e_3}|p(e_1|\mathbf{f}_t), p(e_2|\mathbf{f}_t), p(e_3|\mathbf{f}_t), N_t) \\ p(\mathbf{f}_t|-) &\propto \prod_{j=1}^{3} p(e_j|\mathbf{f}_t)^{n_{t,e_j}} \prod_{j=1}^{3} f_{t,c_j}^{n_{t,c_j}} \end{aligned}$$

the first of which we can sample from directly, the second of which cannot be analytically sampled from and requires Metropolis-Hastings sampling.

To compute the overall expected fraction of coding effects for the set of observed mutations, we convolve over the count distributions for the 32 strand-collapsed trinucleotides and 3 substitutions from (Equation 1.1),

$$n_{\mathrm{tot},e_1}, n_{\mathrm{tot},e_2}, n_{\mathrm{tot},e_3} \sim \overset{32}{\underset{t=1}{*}} \overset{3}{\underset{c_i=1}{*}} p(\mathbf{n}_{t,e}|\mathbf{n}_{t,c}, N_t) \qquad \text{(Equation 1.2)}$$

and then consider the fractions of each effect type (i.e., after dividing the counts by $N = \sum_t N_t$) as the Dirichlet distribution

$$f_{\mathrm{tot},e_1}, f_{\mathrm{tot},e_2}, f_{\mathrm{tot},e_3} \sim \mathrm{dir}(n_{\mathrm{tot},e_1}, n_{\mathrm{tot},e_2}, n_{\mathrm{tot},e_3}).$$

Gene dN/dS Calculation

In the previous section, we computed the expected *fractions* of protein coding effects, given a set of mutations and their contexts/substitutions. A related but distinct problem is computing whether the nonsynonymous mutation *density* significantly exceeds the

synonymous mutation density for a given gene. This ratio-of-densities, referred to as *dN / dS* in molecular evolution, is a good proxy for inferring whether a gene as a whole is under any kind of selective pressure:

- A gene under **neutral selection** will be mutated randomly and will have no bias towards generating nonsynonymous over synonymous events, after normalizing for the gene sequence and codon structure. Therefore, the *dN / dS* of genes under neutral selection will be very close to 1.
- A gene under **positive selection** will have an excess of nonsynonymous events, relative to synonymous events, since phenotypes that change fitness (positive or negative) arise overwhelmingly more often from protein altering mutations than from synonymous mutations. Therefore, the *dN / dS* of genes under positive selection will exceed 1.
- A gene under **negative selection** will have a dearth of nonsynonymous events, since protein altering substitutions would be selected out, keeping relatively more synonymous substitutions. Therefore, *dN / dS* of genes under negative selection will be less than 1.

We are interested in identifying genes that are confidently drivers (to define a true positive set) or genes that are confidently passengers (for a true negative set). As in the previous section, we are not simply interested in a point estimate of *dN / dS* for a given gene, but rather its distribution. It is crucial to estimate the *uncertainty* on *dN / dS* to confidently conclude whether a gene is indeed under positive or neutral selection. Uncertainty in *dN / dS* decreases as the number of mutations in the gene increases.

As mentioned previously, we must normalize for heterogeneous mutation frequencies at different genomic contexts and the codon structure of the gene. For each of the 96 trinucleotide channels (16 contexts × 6 base changes), the expected number of mutations with coding effect *e* in gene *g* is simply

$$\lambda_{g,e,c} \equiv \langle n_{g,e,c} \rangle = r_c N_{g,e,c} \ ,$$

where r_c is the exome-wide mutation frequency for channel *c* and $N_{g,e,c}$ is the number of positions (corrected for sequencing coverage) in the ORF of *g* within channel *c* that would yield a codon substitution of a particular effect *e* (i.e., synonymous, missense, nonsense).

If mutation frequencies were constant across all genes, then the observed number of mutations for channel *c* and effect *e* in any gene would be Poisson distributed around this average, i.e.,

$$n_{g,e,c} \sim \text{pois}(\lambda_{g,e,c}). \qquad \text{(Equation 2.1)}$$

We know this is not the case, as different genes have different background mutabilities, so the overall number of mutations could be higher or lower than this expectation. Furthermore, driver genes are under positive selection, so their number of nonsynonymous mutations could be higher than this expectation. We correct for this heterogeneity with a gene-specific scale factor we infer from the data.

We assume background mutability/selective pressure is channel-independent — for example, we do not expect a more highly mutable gene to be more mutable only with respect to T(C→T)G mutations, or a driver gene to be enriched for nonsynonymous events preferentially for A(A→T)G mutations. Because of this channel-independence assumption, we scale $\lambda_{g,e,c}$ by a channel-agnostic correction factor $\beta_{g,e}$, so the Poisson distribution in (Equation 2.1) becomes

$$n_{g,e,c} \sim \text{pois}(\beta_{g,e}\lambda_{g,e,c}). \qquad \text{(Equation 2.2)}$$

$\beta_{g,e}$ reflects the overall excess/dearth of mutations with coding effect *e* in gene *g*. For example, $\beta_{g,\text{nonsyn}}$ close to 1 means that the number of nonsynonymous mutations is close to what we expect given average mutation frequencies; values significantly higher than 1 indicate an excess of nonsynonymous events; values close to 0 indicate a paucity. Of course, $\beta_{g,\text{nonsyn}} \gg 1$ is not necessarily indicative of positive selection — a highly mutable gene under neutral evolution would have a comparable excess of both nonsynonymous *and* synonymous events. To infer selective pressure on gene *g*, we are interested in the *ratio* of these scale factors,

$$dN / dS \equiv \frac{\beta_{g,\text{nonsyn}}}{\beta_{g,\text{syn}}}.$$

We calculate $\beta_{g,e}$ from its likelihood, which is the product of (Equation 2.2) over the 96 channels,

$$\begin{aligned} \mathcal{L}\left(\beta_{g,e} \middle| \{n_{g,e,c_i}, \lambda_{g,e,c_i}\}_{i \in \{1\ldots96\}}\right) &= \prod_{i=1}^{96} \text{pois}(n_{g,e,c_i}; \beta_{g,e}\lambda_{g,e,c_i}) \\ &\propto \prod_{i=1}^{96} (\beta_{g,e}\lambda_{g,e,c_i})^{n_{g,e,c_i}} \exp(-\beta_{g,e}\lambda_{g,e,c_i}) \\ &\propto \beta_{g,e}^{\sum_i n_{g,e,c_i}} \exp\left(-\beta_{g,e}\sum_i \lambda_{g,e,c_i}\right). \end{aligned}$$

For genes with many mutations, we could simply find the maximum likelihood estimate of $\beta_{g,e}$, as the likelihood would be sharply peaked. However, since many genes have only a few mutations and have considerable uncertainty around the maximum likelihood estimate (MLE), we use the full posterior on $\beta_{g,e}$. Luckily, the Poisson likelihood is proportional to the gamma distribution, allowing us to explicitly define this posterior:

$$\begin{aligned} p(\beta_{g,e}; \mathbf{n}_{g,e}, \boldsymbol{\lambda}_{g,e}) &\propto \mathcal{L}(\beta_{g,e} | \mathbf{n}_{g,e}, \boldsymbol{\lambda}_{g,e}) \\ &\propto \beta_{g,e}^{\sum_i n_{g,e,c_i}} \exp\left(-\beta_{g,e} \sum_i \lambda_{g,e,c_i} \right) \\ &\propto \text{gamma}\left(\beta_{g,e} \middle| 1 + \sum_i n_{g,e,c_i}, \sum_i \lambda_{g,e,c_i} \right). \end{aligned}$$

We can then treat both $\beta_{g,\text{nonsyn}}$ and $\beta_{g,\text{syn}}$ as gamma random variables, with the ratio dN/dS their quotient,

$$\frac{dN}{dS} \sim \frac{\beta_{g,\text{nonsyn}}}{\beta_{g,\text{syn}}} \qquad \begin{aligned} \beta_{g,\text{nonsyn}} &\sim \text{gamma}\left(1 + \sum_i n_{g,\text{nonsyn},c_i}, \sum_i \lambda_{g,\text{nonsyn},c_i}\right) \\ \beta_{g,\text{syn}} &\sim \text{gamma}\left(1 + \sum_i n_{g,\text{syn},c_i}, \sum_i \lambda_{g,\text{syn},c_i}\right). \end{aligned}$$

Finally, we obtain the distribution of dN / dS via a Monte Carlo simulation.

We can generalize this approach to infer the distribution of arbitrary protein-coding effect ratios — for example, whether genes are enriched for *truncating* mutations relative to synonymous mutations, thereby increasing sensitivity to detect tumor suppressors, which often harbor truncating events.

We list such genes in Table S3 (Prob[dT / dS > 2.5] > 0.975). As expected, the list almost entirely comprises well-known tumor suppressors as per the Cancer Gene Census (CGC), or genes not yet indexed by the CGC but with evidence in the literature supporting their status as tumor suppressors, whose citations we have included in the table. While the list contains few totally uncharacterized potential drivers, it is nonetheless confirmation that the method we use to compute d^* / dS works well.

Description of Significance Methods

In this section, we detail the four models presented in the main text: the two conventional methods, **Uniform-within-gene** and **Uniform Poisson regression**, which both assume that base-wise mutation frequencies have no latent variability beyond what can be predicted by sequence context and genomic covariates, and the two overdispersed methods **Gamma-Poisson regression** and **Log-normal-Poisson regression**, which allow mutation frequencies to have additional underlying variability beyond what the covariates predict (modeled as gamma or log-normal distributions, respectively).

Uniform-within-Gene Model

The Uniform-within-gene model assumes that all sites of the same *k*-mer context in the same gene are equally mutable. There are many methods that employ this conventional assumption, with varying implementations (Miller et al., 2015; Van den Eynden et al., 2015; Lohr et al., 2012; Lawrence et al., 2014; Chang et al., 2016; Baeissa et al., 2017; Araya et al., 2016).

One particular method described by (Chang et al., 2016) assumes that the background mutation rate at a given position with trinucleotide context t in gene g is proportional to the average exome-wide mutation rate across all contexts t, weighted by a gene- and codon-specific mutability factor. In other words, it assumes that every context/codon combination t,c within gene g is equally mutable. It also does not account for differing substitution rates, e.g., it considers all mutations at TCG contexts equally likely, irrespective of base change.

Because we consider mutations solely on the positional level, not on the codon level, we re-implemented a method similar to that of Chang et al., which we call **Uniform-within-gene**, that does not include codon-specific factors, but does account for different substitution rates.

Let $n_{s,c}$ equal the total number of mutations with base substitution s at k-mer context c, and N_c equal the total number of contexts c in the exome. The overall mutation rate for this context/substitution is

$$r_{s,c} = n_{s,c} / N_c \ . \qquad \text{(Equation 3.1)}$$

If all genes were equally mutable, the expected number of mutations $x_{s,c}$ at any genomic position of context c with base substitution s would simply be Poisson distributed (assuming binomial convergence) with rate parameter $r_{s,c}$,

$$x_{s,c} \sim \text{pois}(r_{s,c}). \qquad \text{(Equation 3.2)}$$

The Uniform-within-gene model accounts for the fact that different genes have different intrinsic mutabilities by scaling $r_{s,c}$ by a gene-specific mutability factor F_g, which explicitly assumes that all contexts/substitutions within the same gene are equally scaled. Suppose the coding sequence of gene g contains $N_{g,c}$ positions of context c. Assuming no gene-specific scaling, the expected total number of mutations in g would be $\lambda_{g,s,c} = r_{s,c} N_{g,c}$, with $r_{s,c}$ from (Equation 3.1). The observed numbers of mutations $x_{g,s,c}$ for

this gene/context/substitution is again Poisson distributed around this average value, but this time scaled by an unknown parameter F_g:

$$x_{g,s,c} \sim \text{pois}(\lambda_{g,s,c} F_g).$$

To calculate F_g, we first compute its likelihood over all $4^k/2$ strand-collapsed contexts×3 substitutions (equals to 96 for $k = 3$),

$$\mathcal{L}(F_g | \{x_{i,j}\}) = \prod_{i=1}^{4^k/2} \prod_{j=1}^{3} \text{pois}(x_{i,j} | \lambda_{g,i,j} F_g).$$

Next, we find the maximum likelihood estimate for F_g,

$$\widehat{F}_g = \frac{\sum_{i,j} x_{g,i,j}}{\sum_{i,j} \lambda_{g,i,j}}. \qquad \text{(Equation 3.3)}$$

Scaling the rate parameter of (Equations C.3.2) by (C.3.3), we can compute a p value (i.e., the probability of seeing at least the observed number of mutations by chance) for a particular gene, site, and substitution type by plugging in the observed number of mutations, $\tilde{x}_{g,c,s}$, into the upper cumulative distribution function (CDF) of the Poisson distribution

$$P(\tilde{x}_{g,c,s}) \equiv \Pr(x \geq x_{g,c,s}) = \sum_{x=\tilde{x}_{g,c,s}}^{\infty} \text{pois}(x | r_{s,c} \widehat{F}_g).$$

Uniform Poisson Regression Model

Rather than estimate the background mutability of a gene from its own mutation burden as in the Uniform-within-gene model, we can estimate it by regressing against covariates known to influence background mutability. Because covariates can be associated with genomic regions of different scales (e.g., from the single base-pair to entire chromatin regions), this approach is not limited to learning mutability solely on the gene scale, as is the case with the Uniform-within-gene model, but rather allows pooling information from genomic regions of arbitrary scales.

We implement this model via standard Poisson regression, i.e., a fixed effect general linear model (GLM) with a log link function. For a given genomic position i mutated $x_{i,k}$ times across the cohort with context and substitution k and covariate vector $\vec{c}_i$, we assume the observed mutation count is Poisson distributed, with the log of its rate parameter linearly defined by the covariates, i.e.,

$$x_{i,k} | \vec{c}_i \sim \text{pois}(\lambda_{i,k}) \qquad \log \lambda_{i,k} = \beta_{k0} + \vec{\beta}_k \cdot \vec{c}_i. \qquad \text{(Equation 3.4)}$$

To find the intercept β_{k0} and slope $\vec{\beta}_k$ parameters, we maximize their likelihood function with respect to all exonic positions of context and substitution k, whose set we denote $\mathcal{E}_k$:

$$\mathcal{L}(\beta_{k0}, \vec{\beta}_k | \{x_{i,k}\}, \vec{\mathbf{c}}) = \prod_{i \in \mathcal{E}_k} \text{pois}(x_{i,k} | \exp(\beta_{k0} + \vec{\beta}_k \cdot \vec{c}_i)).$$

This likelihood function is log-convex so its maximum $\widehat{\beta}_{k0}$, $\widehat{\vec{\beta}}_k$ can be found via simple optimization methods, e.g., gradient descent. In practice, we use Newton-Raphson because the number of parameters is low enough that computing their full Hessian ($O(n^2)$ in number of parameters) is feasible.

Once we have found our model parameters for each k, we can assign a p value for any position with context and substitution k with counts $\tilde{x}_{i,k}$ and specific covariates $\widetilde{\vec{c}}_i$ from the upper CDF of the Poisson distribution, as before:

$$P(\tilde{x}_{i,k} | \widetilde{\vec{c}}_i) \equiv \Pr(x \geq \tilde{x}_{i,k} | \widetilde{\vec{c}}_i) = \sum_{x=\tilde{x}_{i,k}}^{\infty} \text{pois}(x | \exp(\widehat{\beta}_{k0} + \widehat{\vec{\beta}}_k \cdot \widetilde{\vec{c}}_i)).$$

Gamma-Poisson Regression Model

The previous two models both assume that the base-wise mutability (i.e., the Poisson distribution's rate parameter) is fixed once the model parameters are found. However, this assumption does not allow for additional uncertainty in the base-wise mutability. To allow for additional uncertainty, we can probabilistically model the base-wise mutability using an arbitrary distribution $p(\boldsymbol{\theta})$. This forms a hierarchical model for observed mutation counts x,

$$x \sim \text{pois}(\lambda) \qquad \lambda \sim p(\boldsymbol{\theta})$$

whose probability density function (PDF) is the compound distribution

$$p(x | \boldsymbol{\theta}) = \int_0^{\infty} d\lambda \, \text{pois}(x | \lambda) p(\lambda | \boldsymbol{\theta}), \qquad \text{(Equation 3.5)}$$

with the parameters $\boldsymbol{\theta}$ of the latent distribution of the base-wise mutability learned from the data, e.g., using maximum likelihood.

Unfortunately, the integral in (Equation 3.5) is only analytically tractable for a few choices of $p(\lambda|\boldsymbol{\theta})$. One common choice is the gamma distribution; as mentioned in the main text, there is no intrinsic biological motivation for using this distribution — it is merely mathematically convenient. If $p(\lambda \mid \boldsymbol{\theta}) \equiv \text{gamma}(\lambda \mid a,b)$, (Equation 3.5) becomes

$$\begin{aligned} p(x|a,b) &= \int_0^\infty d\lambda\ \text{pois}(x|\lambda)\text{gamma}(\lambda|a,b) \\ &= \frac{1}{x!}\frac{b^{-a}}{\Gamma(a)}\int_0^\infty d\lambda\ \exp(-\lambda)\lambda^x \times \exp(-\lambda/b)\lambda^{a-1}. \end{aligned} \qquad \text{(Equation 3.6)}$$

This integral is of the form $\int_0^\infty dx\ \exp(-ax)x^{b-1} = a^{-b}\Gamma(b)$, so (Equation 3.6) becomes

$$\begin{aligned} p(x|a,b) &= \frac{1}{x!}\frac{b^{-a}}{\Gamma(a)}\left(\frac{b}{1+b}\right)^{x+a}\left(\frac{b}{1+b}\right)^{a}\Gamma(a+x) \\ &\Downarrow \\ p(x|a,p) &= \frac{\Gamma(a+x)}{\Gamma(x+1)\Gamma(a)}p^x(1-p)^a, \end{aligned}$$

the PDF of the negative binomial distribution for $p = b/(1 + b)$. Unlike the Poisson distribution, which is immediately amenable to parameterization as a GLM whose log mean can be linearly parameterized by covariates (Equation 3.4), the negative binomial must be re-written in terms of its mean, $\mu = pa/(1-p)$, yielding the GLM

$$p\left(x_{i,k}\middle|a,\beta_0,\vec{\beta},\vec{c}_i\right) \propto \left(\frac{\mu_i}{a+\mu_i}\right)^{x_{i,k}}\left(\frac{a}{a+\mu_i}\right)^{a} \qquad \log\mu_{i,k} = \beta_{k0} + \vec{\beta}_k \cdot \vec{c}_i.$$

As before, optimal values for model parameters (intercept β_{k0}, slope vector $\vec{\beta}_k$, and dispersion parameter a) can be found via maximum likelihood estimation. Finally, we assign p values as with the other methods by computing the upper CDF of the negative binomial distribution parameterized by its MLE values.

Log-Normal-Poisson Regression Model

A more principled choice for the latent distribution for the base-wise mutability is the log-normal distribution. Mutations are the product of independent consecutive events, each with an independent probability of occurring; by the geometric central limit theorem, this product of probabilities approaches a log-normal distribution (Sutton, 1997).

As in the Uniform Poisson regression model (Equation 3.4), the covariates linearly determine the log of the Poisson rate, but this time we include an additional linear factor $\epsilon_{i,k}$, which is a random variable that represents additional latent mutation rate variability at position i (with context and substitution k) not explained by the covariates. For genomic position i mutated x_i times with covariate vector $\vec{c}_i$ in channel k, our model is

$$x_{i,k}\left|\vec{c}_i \sim \text{pois}(\lambda_{i,k}) \qquad \log\lambda_{i,k} = \vec{\beta}_k \cdot \vec{c}_i + \mu_k + \sigma_k\epsilon_{i,k} \qquad \epsilon_{i,k} \sim \mathcal{N}(0,1).\right.$$

The PDF of $x_{i,k}$ is then

$$p(x_{i,k}|\mu_k,\sigma_k,\vec{\beta}_k,\vec{c}_i) = \int_0^\infty d\epsilon_{i,k}\ \text{pois}(x_{i,k}|\exp(\vec{\beta}_k \cdot \vec{c}_i + \mu_k + \sigma_k\epsilon_{i,k}))\mathcal{N}(\epsilon_{i,k}|0,1). \qquad \text{(Equation 3.7)}$$

Since $\epsilon_{i,k}$ is a standard normal random variable, $\exp(\vec{\beta}_k \cdot \vec{c}_i + \mu_k + \sigma_k\epsilon_{i,k})$ is log-normally distributed with geometric mean $\vec{\beta}_k \cdot \vec{c}_i + \mu_k$ and geometric standard deviation σ_k. These parameters are interpretable: μ_k represents the overall geometric mean mutation frequency for context/substitution k, σ_k the overall geometric standard deviation around this mean (i.e., the average multiplicative distance from the mean), and $\vec{\beta}_k$ the geometric covariate slope vector associated with k (i.e., the multiplicative effect each covariate has on the mutation rate).

In addition to being better statistically calibrated than the other methods (see Results), the log-normal-Poisson model carries other advantages:

1. Because σ_k is multiplicative, it quantifies overdispersion irrespective of mutation frequency, allowing for comparisons between different mutational processes. For example, $\exp(\sigma_k) = 2$ indicates that bases in context-substitution pair k that are one geometric standard deviation above the mean will be approximately *twice* as mutable as sites at the mean, while bases falling one g. standard deviation below the mean will be approximately *half* as mutable, regardless of the mean mutation frequency.
2. The log-normal-Poisson model accommodates nested variance (variance is additive). Suppose that we are analyzing mutations within trinucleotide context c. We would fit the model to all exonic sites of that sequence context. Now suppose that we would like to account for additional variance explained by the 16 flanking pentanucleotide contexts. The set of base-pairs belonging to the trinucleotide context c is a superset of the base-pairs belonging to a pentamer context flanking that

trinucleotide. Thus, we can express the total variance for base-pair i at context/substitution k as the sum of the overall trinucleotide variance and the additional variance explained by the pentamer context of i:

$$\log\lambda_{i,k} = \vec{\beta}_k \cdot \vec{c}_i + \mu_{k_{\text{pent}}} + \sigma_{k_{\text{tri}}}\epsilon_{i_{\text{tri}}} + \sigma_{k_{\text{pent}}}\epsilon_{i_{\text{pent}}}, \quad \text{(Equation 3.8)}$$

with $\epsilon_{i\text{tri}}$ and $\epsilon_{i\text{pent}}$ representing independent standard normal random variables. The total variance for the pentamer context is $\sigma^2_{k_{\text{tri}}} + \sigma^2_{k_{\text{pent}}}$, since the variance of the sum of independent normal random variables is the sum of their variances. In general linear mixed model parlance, the pentamers are *categorical random effects*.

3. The natural parameterization in terms of mean and variance allows us to naturally quantify the amount of variance explained by model factors (e.g., covariates). The amount that σ^2 drops after adding a covariate is equivalent to the variance explained by that covariate, given the other covariates already present in the model. To obtain the average linear contribution of each covariate, we fit the model for all possible sets of covariates, and then perform an ordinary least squares fit to find each marginal linear contribution. For example, if there are three covariates, we fit a model for each of the $2^3 = 8$ possible combinations, obtaining $\sigma^2_0 \ldots \sigma^2_7$, the observed values of model parameter σ^2 for each covariate set (0: no covariates, 7: all covariates). We then fit the following linear model,

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \end{bmatrix} \vec{\sigma}^2_{\text{lin}} = \begin{bmatrix} \sigma^2_0 \\ \sigma^2_1 \\ \sigma^2_2 \\ \sigma^2_3 \\ \vdots \\ \sigma^2_7 \end{bmatrix},$$

with a 1 in the data matrix representing the presence of each covariate. The first column is the intercept and thus entirely comprises ones. $\vec{\sigma}^2_{\text{lin}}$ is then the marginal linear contribution of each covariate; its first element (i.e., the intercept) is the remaining amount of unexplained variance after all covariates have been incorporated; the other three elements are the variance explained by their respective covariate.

To fit the log-normal-Poisson model's parameters, we employ a fully Bayesian approach, sampling from the parameters' posterior distribution $p(\mu_k, \sigma_k, \vec{\beta}_k | \{x_{i,k}\}, \{\vec{c}\})$ via MCMC. We use MCMC instead of a standard likelihood (or quasi-likelihood) approach because:

1. The model parameters provide meaningful summary statistics of the base-wise mutation rate; as such, we are not merely interested in the model's predictions (i.e., significance values), but also its parameters. It is therefore useful to not only obtain their optimal best-fit values, but their overall posterior distribution, to quantify our confidence in the model.
2. The integral in (Equation 3.7) is analytically intractable, but it is amenable to approximation via Gibbs sampling, given the hierarchical nature of the model.

We describe the MCMC in detail in STAR Methods, MCMC implementation.

To compute p values for the previous three models, we first find a point estimate $\hat{\boldsymbol{\theta}}$ of optimal model parameters (e.g., via maximum likelihood estimation), and then use the upper CDF parameterized by $\hat{\boldsymbol{\theta}}$, for an observed number of mutations $\tilde{x}_i$ with covariates $\vec{\tilde{c}}_i$ at base-pair i,

$$P(\tilde{x}_i | \vec{\tilde{c}}_i) = \sum_{x=\tilde{x}_i}^{\infty} p(x | \hat{\boldsymbol{\theta}}, \vec{\tilde{c}}_i).$$

Because the log-normal-Poisson model is fully Bayesian, we no longer use a point estimate for $\hat{\boldsymbol{\theta}}$, but rather integrate the CDF over the full domain $\boldsymbol{\Theta}$ of the posterior on $\boldsymbol{\theta}$, yielding posterior predictive p values

$$P\left(\tilde{x}_i \middle| \vec{\tilde{c}}_i\right) = \sum_{x=\tilde{x}_i}^{\infty} \int_{\boldsymbol{\Theta}} d\boldsymbol{\theta}\, p\left(x \middle| \boldsymbol{\theta}, \vec{\tilde{c}}_i\right) p(\boldsymbol{\theta} | \{x\}, \{\vec{c}\}) \quad \text{(Equation 3.9)}$$

MCMC Implementation

In this section, we detail the implementation of the Markov Chain Monte Carlo (MCMC) sampler used to fit the log-normal-Poisson regression model. As with most non-conjugate hierarchical models, the log-normal-Poisson probability density function (Equation 3.7) is analytically intractable — its most general form is the compound distribution

$$p(x | \mu, \sigma) = \int_0^{\infty} d\epsilon\, \text{pois}(x | \epsilon) \log\mathcal{N}(\epsilon | \mu, \sigma), \quad \text{(Equation 4.1)}$$

which has no closed form, so analytically sampling any distribution derived from the log-normal-Poisson PDF (e.g., the posterior on model parameters) will also lack closed form.

Model Statement

Recall the model hierarchy of the general linear mixed model defined in Equations 3.7 and 3.8 in Log-Normal-Poisson Regression Model.

For observed number of mutated patients x_i at the ith base-pair whose covariate vector is $\vec{c}_i$, we have

$$x_i|\vec{c}_i \sim \text{pois}(\lambda_i) \qquad \log\lambda_i = \vec{\beta}\cdot\vec{c}_i + \mu + \sigma\epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0,1).$$

For a set $\{x\}$ of n base-pairs with covariates $\{\vec{c}\}$, the full posterior distribution across all parameters — i.e., including the latent ϵ in (Equation 4.1) or (Equation 3.7) — is

$$p(\mu,\sigma,\vec{\beta},\epsilon_1,\ldots,\epsilon_n|\{x\},\{\vec{c}\},\mathcal{H}) \propto \prod_{i=1}^{n}[p(x_i|\exp(\vec{\beta}\cdot\vec{c}_i+\mu+\sigma\epsilon_i))p(\epsilon_i)]p(\mu,\sigma,\vec{\beta}|\mathcal{H}) \qquad \text{(Equation 4.2)}$$

for set of hyperparameters $\mathcal{H}$. In the case of nested variability (i.e., when adding m total categorical random effects), the model hierarchy becomes

$$x_i|\vec{c}_i, 1_{i\cdot} \sim \text{pois}(\lambda_i) \qquad \log\lambda_i = \vec{\beta}\cdot\vec{c}_i + \sum_{j=1}^{m}(\mu_j+\sigma_j\epsilon_{ij})1_{ij} + \sigma_0\epsilon_{i0} \qquad \{\epsilon_{ij}\} \sim \mathcal{N}(0,1), \qquad \text{(Equation 4.3)}$$

where 1_{ij} is an indicator function for whether the ith base is of category j, and the set $\{\epsilon_{ij}\}$ comprises *independent* standard normal random variables. The full posterior (Equation 4.2) becomes

$$\begin{aligned} &p(\mu_1,\ldots,\mu_m,\sigma_0,\sigma_1,\ldots,\sigma_m,\vec{\beta},\{\epsilon\}|\{x\},\{\vec{c}\},\mathcal{H}) \\ &\propto \prod_{i=1}^{n}\left[p\left(x_i\middle|\exp\left(\vec{\beta}\cdot\vec{c}_i+\textstyle\sum_{j=1}^{m}(\mu_j+\sigma_j\epsilon_{ij})1_{ij}+\sigma_0\epsilon_{i0}\right)\right)p(\epsilon_{i0})\prod_{j=1}^{m}p(\epsilon_{ij})^{1_{ij}}\right]p(\{\mu\},\{\sigma\},\vec{\beta}|\mathcal{H}) \end{aligned} \qquad \text{(Equation 4.4)}$$

In our analyses, we only use the categorical random effects to represent flanking pentamer contexts within each trimer. Suppose that we fit the model to the set of all base-pair substitutions $\{x_k\}$ at trinucleotide substitution k, e.g. TCT→TAT. Our categories within $\{x_k\}$ are the $j = 1 \ldots 16$ pentamer contexts flanking k (e.g., **A**TCT**G**), which are unique for each base-pair. Then (Equations C.4.3) becomes (C.3.8); the sum in (C.4.3) disappears, since $1_{ij} = 0$ for all but a single value of j — each base-pair can only have a single pentamer context. We denote the pentamer context associated with the ith base-pair as $j(i)$.

Explicitly writing out the full posterior distribution for the single random effect model for n base-pairs in trinucleotide context k, we have

$$\begin{aligned} &p(\mu_1,\ldots,\mu_{16},\sigma_0,\sigma_1,\ldots,\sigma_{16},\vec{\beta},\{\epsilon\}|\{x_k\},\{\vec{c}\},\mathcal{H}) \\ &\propto \prod_{i=1}^{n}\left[p\left(x_i\middle|\exp\left(\underbrace{\vec{\beta}\cdot\vec{c}_i+\mu_{j(i)}+\sigma_0\epsilon_{i0}+\sigma_{j(i)}\epsilon_{ij(i)}}_{\lambda_i}\right)\right)p(\epsilon_{i0})p(\epsilon_{ij(i)})\right]p(\{\mu\},\{\sigma\},\vec{\beta}|\mathcal{H}) \end{aligned} \qquad \text{(Equation 4.5)}$$

By sampling from the full posterior in (Equations 4.5) and ignoring the samples for latent parameters ϵ, we simultaneously sample the distribution on the parameters of interest ($\{\mu\}$, $\{\sigma\}$, $\vec{\beta}$) while marginalizing out $\{\epsilon\}$, i.e., performing the integral in (Equations 4.1). Because these full posteriors can have millions of parameters (each $\epsilon_{ij}\in\{\epsilon\}$ corresponds to a single base-pair, and we fit the model to all base-pairs within a trinucleotide context), they are impractical to sample in a single draw. We therefore use a Metropolis-within-Gibbs scheme: we partition the joint posterior's parameters into uncorrelated blocks, and independently sample from the full conditional of each block. In the next sections, we detail these parameter blocks and their full conditionals. We will focus exclusively on the nested model in (Equations 4.5), since the non-nested model is a simplified subset of it.

Metropolis-within-Gibbs Description

The Gibbs sampler allows us to draw from a high dimensional probability distribution by iteratively sampling lower dimensional subsets of its parameters conditioned on all the other parameters. For example, if sampling in three dimensions from the distribution $p(x,y,z)$ is difficult, iterative univariate samples from full conditional distributions $p(x|y,z)$, $p(y|x,z)$, and $p(z|x,y)$ will converge to multivariate samples (x,y,z) from the full joint. We will denote full conditional distributions as $p(x|-)$: x conditioned on all other parameters in the joint, i.e., (y, z).

For notational convenience, let us first expand (Equations C.4.5) in terms of the PDFs of the Poisson and standard normal distributions corresponding to $p(x_i|\lambda_i)$ and $p(\epsilon_{ij})$, respectively. Recall that λ_i is a function of $\mu_{j(i)}$, σ_0, $\sigma_{j(i)}$, $\vec{\beta}$, $\{\epsilon\}$, and $\vec{c}_i$; we omit the arguments for brevity.

$$p(\{\mu\},\{\sigma\},\vec{\beta},\{\epsilon\}|\{x_k\},\{\vec{c}\},\mathcal{H}) \propto \prod_{i=1}^{n}\left[\exp(\lambda_i x_i - \exp(\lambda_i))\exp\left(-\frac{\epsilon_{i0}^2}{2}\right)\exp\left(-\frac{\epsilon_{ij(i)}^2}{2}\right)\right] \times p(\{\mu\},\{\sigma\},\vec{\beta}|\mathcal{H}) \quad \text{(Equation 4.6)}$$

We see that the ϵ's are independent, so each ϵ_{ij} can be sampled independently as its own 1-dimensional block. To allow for correlation between each μ_j and each σ_j, we sample (μ_j, σ_j) as a 2-D block. Likewise, if there is correlation between covariates, we will expect correlation between elements of $\vec{\beta} \in \mathbb{R}^b$, and will therefore sample $\vec{\beta}$ as a b-dimensional block.

Because none of the full conditionals can be sampled from analytically, we will sample from them using the Metropolis-Hastings algorithm: given some target distribution $p(x)$ that is impossible to analytically sample from, we randomly sample x^* from some easy-to-sample proposal distribution $q(x^*|x)$ that approximates $p(x)$, and accept the new value x^* with probability

$$\Pr(x \to x^*) = \min\left\{1, \frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}\right\}.$$

Iterating this procedure will yield random samples from target $p(x)$.

Of course, performance of this method critically depends on choosing a proposal distribution that well-approximates the distribution being sampled. All of the full conditionals are log-concave; furthermore, terms above second-order in their series expansions around their maxima are negligible, so we will use proposal distributions that quadratically approximate the full conditionals centered at their maxima. We find each full conditional's global maximum, and propose using a (multivariate) t distribution whose mean is the global maximum and whose (co)variance is the curvature at the maximum,

$$\widehat{\vec{\mu}} = \langle \hat{x}, \hat{y} \rangle = \underset{x,y}{\operatorname{argmax}}(\log p(x,y|-)) \qquad \hat{\mathbf{\Sigma}} = -\mathbf{H}^{-1}[\log p(\hat{x},\hat{y}|-)] \qquad q(x^*,y^*|x,y) \sim t_\nu\left(\widehat{\vec{\mu}}, \hat{\mathbf{\Sigma}}\right),$$

where $\mathbf{H}^{-1}$ is the inverse Hessian. ν is the degrees-of-freedom of the t distribution, which can be tuned to achieve better approximation of the target distribution.

To find the maxima, we use Newton-Raphson iterations (since we are already calculating the Hessian), augmented with backtracking linesearch to avoid potentially overshooting.

Metropolis-within-Gibbs Sampling Blocks

Here, we detail the sampling procedure used for each full conditional block.

***μ_j, σ_j* block** The full conditional of each (μ_j,σ_j) block is

$$\log p(\mu_j,\sigma_j|-) \propto \sum_{i \in \mathbf{J}}\left[(\mu_j + \sigma_j\epsilon_{ij})x_i - \exp\left(\vec{\beta}\cdot\vec{c}_i + \mu_j + \sigma_0\epsilon_{i0} + \sigma_j\epsilon_{ij}\right)\right] + \log p(\mu_j,\sigma_j|\mathcal{H}).$$

Until this point, we have not specified the form of our priors. Because

$$X \sim \mu_j + \sigma_j\mathcal{N}(0,1) \equiv X \sim \mathcal{N}\left(\mu_j,\sigma_j^2\right),$$

it makes sense to specify a joint prior $p(\mu_j,\sigma_j|\mathcal{H})$ on (μ_j,σ_j) conjugate to the normal distribution, to make interpretation of hyperparameters $\mathcal{H}$ easy. In this case, this is the normal-inverse-gamma distribution:

$$\mu_j,\sigma_j^2 \sim \mathcal{NG}^{-1}(A,B,M,S). \quad \text{(Equation 4.7)}$$

where A and B are the shape/scale parameters of the gamma distribution, and M and S are the mean/variance parameters of the normal distribution, respectively. These parameters signify that σ_j^2 was estimated from $2A$ observations with sample mean M and sum of sample squared deviations $2B$; μ_j was estimated from S observations with sample mean M. To actually implement this prior, we make change-of-variable $\tau_j = 1/\sigma_j^2$, turning the normal-inverse-gamma prior into a normal-gamma prior

$$\mu_j,\tau_j \sim \mathcal{NG}(A,B,M,S).$$

With our prior in hand, we need to compute the gradient and Hessian of the full conditional in order to perform our Newton-Raphson iterations. Our gradient is

$$\vec{\nabla}\log p(\mu_j,\tau_j|-) = \left\langle \frac{\partial}{\partial\mu_j}\log p(\mu_j,\tau_j|-), \frac{\partial}{\partial\tau_j}\log p(\mu_j,\tau_j|-)\right\rangle$$

$$\frac{\partial}{\partial\mu_j}\log p(\mu_j,\tau_j|-) = \sum_{i\in\mathbf{J}} x_i - \exp(\mu_j)\sum_{i\in\mathbf{J}}\exp\left(\vec{\beta}\cdot\vec{c}_i + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij}}{\sqrt{\tau_j}}\right) \underbrace{-S\tau_j(\mu_j - M)}_{\frac{\partial}{\partial\mu_j}\log\mathcal{NG}(\mu_j,\tau_j|A,B,M,S)}$$

$$\frac{\partial}{\partial\tau_j}\log p(\mu_j,\tau_j|-) = -\frac{1}{2\tau_j^{3/2}}\sum_{i\in\mathbf{J}}\epsilon_{ij}x_i + \exp(\mu_j)\sum_{i\in\mathbf{J}}\exp\left(\vec{\beta}\cdot\vec{c}_i + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij}}{\sqrt{\tau_j}}\right)\frac{\epsilon_{ij}}{2\tau_j^{3/2}} + \underbrace{\frac{A-1}{\tau_j} - \frac{1}{B} - \frac{S}{2}(\mu_j - M)^2 + \frac{1}{2\tau_j}}_{\frac{\partial}{\partial\tau_j}\log\mathcal{NG}},$$

and our Hessian components are

$$H_{\mu,\mu} = -\exp(\mu_j)\sum_{i\in\mathbf{J}}\exp\left(\vec{\beta}\cdot\vec{c}_i + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij}}{\sqrt{\tau_j}}\right) \underbrace{-S\tau}_{\frac{\partial^2}{\partial\mu_j^2}\log\mathcal{NG}}$$

$$H_{\tau,\tau} = \frac{3}{4\tau_j^{5/2}}\sum_{i\in\mathbf{J}}\epsilon_{ij}x_i + \exp(\mu_j)\sum_{i\in\mathbf{J}}\exp\left(\vec{\beta}\cdot\vec{c}_i + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij}}{\sqrt{\tau_j}}\right)\epsilon_{ij}\left[-\frac{3}{4\tau_j^{5/2}} - \frac{\epsilon_{ij}}{4\tau_j^3}\right] \underbrace{-\frac{2A-1}{2\tau_j^2}}_{\frac{\partial^2}{\partial\tau_j^2}\log\mathcal{NG}}$$

$$H_{\mu,\tau} = \exp(\mu_j)\sum_{i\in\mathbf{J}}\exp\left(\vec{\beta}\cdot\vec{c}_i + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij}}{\sqrt{\tau_j}}\right)\frac{\epsilon_{ij}}{2\tau_j^{3/2}} \underbrace{-T(\mu_j - M)}_{\frac{\partial^2}{\partial\mu_j\partial\tau_j}\log\mathcal{NG}}$$

To actually use this gradient/Hessian to perform Newton-Raphson iterations, and since Newton-Raphson updates are only stable on functions whose domain is $\mathbb{R}$, we need to make change-of-variable $g^{-1}: \mathbb{R}^+ \to \mathbb{R},\ \tau_j \mapsto \log(\tau_j)$, because $\tau_j \in (0,\infty)$. Recall that when transforming parameters of a probability distribution with some function g^{-1}, we need to multiply by the Jacobian of the inverse transform g to preserve the volume element. Thus, for $g^{-1}(\tau_j) = \log(\tau_j)$,

$$p(\mu_j, g(\tau_j)|-)\frac{dg}{d\tau_j} = p(\mu_j,\exp(\tau_j)|-)\exp(\tau_j).$$

As a result, our log-space gradient/Hessian are actually

$$\vec{\nabla}[\log p(\mu_j,\exp(\tau_j)|-) + \tau_j] \qquad \mathbf{H}[\log p(\mu_j,\exp(\tau_j)|-) + \tau_j]$$

Performing this change-of-variables requires invoking the chain rule. For the gradient, this is straightforward:

$$\vec{\nabla} f(x,g(y)) = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial g}\frac{\partial g}{\partial y}\right\rangle,$$

so for $g(y) = \exp(y)$, we simply multiply the untransformed τ component of the gradient $(\partial_g f)$ by $\exp(\tau_j)$, and then add 1 to account for the Jacobian. For the Hessian, we need to invoke the chain rule twice, to account for the second derivative. H_{xx} is unchanged. The other components transform as follows:

$$H_{yy} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial g}\frac{\partial g}{\partial y}\right] = \frac{\partial f}{\partial g}\frac{\partial^2 g}{\partial y^2} + \left(\frac{\partial g}{\partial y}\right)^2 \frac{\partial^2 f}{\partial g^2} \qquad H_{xy} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right] = \frac{\partial g}{\partial y}\frac{\partial^2 f}{\partial g \partial x}$$

$\partial_g f$ is the untransformed τ component of the gradient; $\partial_y^2 g = \exp(\tau_j)$; $(\partial_y g)^2 = \exp(2\tau_j)$; $\partial_g^2 f$ is the untransformed (τ, τ) component of the Hessian; and $\partial_{gx}^2 f$ is the untransformed μ, τ component of the Hessian. Because the Jacobian term in the gradient is a constant 1, it disappears entirely in the Hessian.

σ_0 block In addition to each (μ_j,σ_j) for each pentamer context, we also have an overall variance term σ_0 shared between the pentamers. Its full conditional is

$$\log p(\sigma_0|-) \propto \sum_{i=1}^{n}\left[\sigma_0 \epsilon_{i0} x_i - \exp\left(\vec{\beta}\cdot\vec{c}_i + \mu_{j(i)} + \sigma_0\epsilon_{i0} + \sigma_{j(i)}\epsilon_{ij(i)}\right)\right] + \log p(\sigma_0|\mathcal{H}).$$

We place an inverse gamma prior on $\sigma_0^2 \sim \mathcal{G}^{-1}(A,B)$, whose parameters have the same conjugate interpretation as the normal-inverse-gamma prior for (μ_j,σ_j), (Equation 4.7). As before, we perform change-of-variable $\tau_0 = 1/\sigma_0^2$ to actually implement this model structure. Our gradient and Hessian are simply the first and second derivatives of $p(\sigma_0|-)$:

$$\frac{\partial}{\partial \tau_0}\log p(\tau_0|-) = -\frac{1}{2\tau_0^{3/2}}\sum_{i=1}^{n}\epsilon_{i0}x_i + \sum_{i=1}^{n}\exp\left(\vec{\beta}\cdot\vec{c}_i + \mu_{j(i)} + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij(i)}}{\sqrt{\tau_{j(i)}}}\right)\frac{\epsilon_{i0}}{2\tau_0^{3/2}} + \underbrace{\frac{A-1}{\tau_0} - \frac{1}{B}}_{\frac{\partial}{\partial\tau_0}\log\mathcal{G}(A,B)}$$

and

$$\frac{\partial^2}{\partial \tau_0^2}\log p(\tau_0|-) = \frac{3}{4\tau_0^{5/2}}\sum_{i=1}^{n}\epsilon_{i0}x_i + \sum_{i=1}^{n}\exp\left(\vec{\beta}\cdot\vec{c}_i + \mu_{j(i)} + \frac{\epsilon_{i0}}{\sqrt{\tau_0}} + \frac{\epsilon_{ij(i)}}{\sqrt{\tau_{j(i)}}}\right)\epsilon_{i0}\left[-\frac{3}{4\tau_0^{5/2}} - \frac{\epsilon_{i0}}{4\tau_0^3}\right] \underbrace{-\frac{A-1}{\tau_0^2}}_{\frac{\partial^2}{\partial\tau_0^2}\log\mathcal{G}(A,B)}.$$

As before, we actually perform the Newton-Raphson iterations on $\log p(\exp(\tau_0)|-)$, transforming the PDF and first/second derivatives with the appropriate Jacobian and chain rule factors, respectively.

$\vec{\beta}$ block Our full conditional is

$$\log p(\vec{\beta}|-) \propto \sum_{i=1}^{n}\left[\vec{\beta}\cdot\vec{c}_i x_i + \exp\left(\vec{\beta}\cdot\vec{c}_i + \mu_{j(i)} + \sigma_0\epsilon_{i0} + \sigma_{j(i)}\epsilon_{ij(i)}\right)\right] + \log p(\vec{\beta}|\mathcal{H}),$$

with $p(\vec{\beta}|\mathcal{H})$ equal to multivariate normal prior $\vec{\beta} \sim \mathcal{N}_m(\vec{\mu}_\beta, \mathbf{\Sigma}_\beta)$. For both notational convenience and computational efficiency, we express the sums of dot products (i.e., $\sum_i \vec{\beta}\cdot\vec{c}_i$) as matrix multiplications over covariate matrix $\mathbf{C} = \begin{bmatrix}\vec{c}_1 & \cdots & \vec{c}_n\end{bmatrix}^\top$ and data vector $\vec{x} = \langle x_1, \ldots x_n\rangle^\top$. We similarly vectorize our sets of parameters and latent random variables, with parameter vectors $\vec{\sigma}_{\mathbf{J}} = \langle\sigma_{j(1)}, \sigma_{j(2)}, \ldots, \sigma_{j(n)}\rangle^\top$ and $\vec{\mu}_{\mathbf{J}} = \langle\mu_{j(1)}, \ldots, \mu_{j(n)}\rangle^\top$, and latent random variable vectors $\vec{\epsilon}_0 = \langle\epsilon_{10}, \epsilon_{20}, \ldots, \epsilon_{n0}\rangle^\top$ and $\vec{\epsilon}_{\mathbf{J}} = \langle\epsilon_{1j(1)}, \ldots, \epsilon_{nj(n)}\rangle^\top$. Our gradient is

$$\vec{\nabla} p(\vec{\beta}|-) = \vec{x}^\top\mathbf{C} - \underbrace{\exp\odot(\mathbf{C}\vec{\beta} + \vec{\mu}_{\mathbf{J}} + \sigma_0\vec{\epsilon}_0 + \vec{\sigma}_{\mathbf{J}}\cdot\vec{\epsilon}_{\mathbf{J}})^\top}_{\vec{q}}\mathbf{C} - \underbrace{\mathbf{\Sigma}_\beta^{-1}(\vec{\beta} - \vec{\mu}_\beta)}_{\vec{\nabla}_{p(\vec{\beta}|\vec{\mu}_\beta,\mathbf{\Sigma}_\beta)}},$$

where $\exp\odot$ is the elementwise exponential. Our Hessian is simply

$$\mathbf{H} = -\mathbf{C}^\top\mathrm{diag}(\vec{q})\mathbf{C} - \mathbf{\Sigma}_\beta^{-1}.$$

Because $\mathbf{C}$ can have millions of rows, we speed up calculation of the Hessian by noting that any matrix multiplication of the form $\mathbf{C}^\top\mathbf{D}\mathbf{C}$ with diagonal $\mathbf{D}$ can be expressed as

$$\mathbf{C}^\top \mathbf{D}\mathbf{C} = \mathbf{C}^\top \left(\sum_{j=1}^{n} d_j \mathbf{E}_{jj}\right)\mathbf{C} = \sum_{j=1}^{n} d_j \mathbf{C}^\top \mathbf{E}_{jj}\mathbf{C} = \sum_{j=1}^{n} d_j \vec{c}_j \otimes \vec{c}_j$$

where d_j is the jth diagonal element of $\mathbf{D}$, $\mathbf{E}_{jj}$ the corresponding standard basis matrix, and $\vec{c}_j$ the jth row of $\mathbf{C}$. Because the covariate matrix $\mathbf{C}$ is constant, we only need to calculate each outer product $\vec{c}_j \otimes \vec{c}_j$ once ahead of time, greatly saving time when computing the Hessian, which only requires computing a linear combination of the outer product matrices.

ϵ **blocks** For each shared latent random variable $\epsilon_{i0} \in \epsilon_0$, our full conditional is

$$\log p(\epsilon_{i0}|-) \propto \sigma_0 \epsilon_{i0} x_i - \exp\left(\vec{\beta} \cdot \vec{c}_i + \mu_{j(i)} + \sigma_0 \epsilon_{i0} + \sigma_{j(i)} \epsilon_{ij(i)}\right) - \frac{\epsilon_{i0}^2}{2}.$$

Likewise, each nested latent random variable's full conditional is

$$\log p\left(\epsilon_{ij(i)}\middle|-\right) \propto \sigma_j \epsilon_{ij(i)} x_i - \exp\left(\vec{\beta} \cdot \vec{c}_i + \mu_{j(i)} + \sigma_0 \epsilon_{i0} + \sigma_{j(i)} \epsilon_{ij(i)}\right) - \frac{\epsilon_{ij(j)}^2}{2}.$$

Each of these full conditionals is univariate, so their gradients/Hessians will simply be the first/second derivatives, which are of the form

$$\begin{aligned}
\frac{\partial}{\partial \epsilon_{i0}} \log p(\epsilon_{i0}|-) &= \sigma_0 x_i - \sigma_0 \exp\left(\vec{\beta} \cdot \vec{c}_i + \mu_{j(i)} + \sigma_0 \epsilon_{i0} + \sigma_{j(i)} \epsilon_{ij(i)}\right) - \epsilon_{i0} \\
\frac{\partial^2}{\partial \epsilon_{i0}^2} \log p(\epsilon_{i0}|-) &= -\sigma_0^2 \exp\left(\vec{\beta} \cdot \vec{c}_i + \mu_{j(i)} + \sigma_0 \epsilon_{i0} + \sigma_{j(i)} \epsilon_{ij(i)}\right) - 1,
\end{aligned}$$

substituting the appropriate σ/ϵ as necessary.

Posterior Predictive Calculation

To actually infer whether an observed level of recurrent mutation significantly exceeds the expected background, we compute a posterior predictive p value as stated in Equation 3.9. This entails integrating the parameters of the log-normal-Poisson distribution over the posterior probabilities of the model parameters, whose domain we denote $\boldsymbol{\Theta}$. Given some observed number of mutations $\tilde{x}$ with covariates $\widetilde{\vec{c}}$ and category (i.e., pentamer context) $\tilde{\jmath}$, the posterior predictive is

$$\begin{aligned}
p(\tilde{x}|\{x\}, \{\vec{c}\}, \tilde{\vec{c}}) = \int_{\boldsymbol{\Theta}} \overbrace{\left[\int_{-\infty}^{\infty} d\epsilon_0 \, d\epsilon_{\tilde{\jmath}} \, \mathrm{pois}(\tilde{x}|\exp(\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \sigma_0 \epsilon_0 + \sigma_{\tilde{\jmath}} \epsilon_{\tilde{\jmath}})) \mathcal{N}(\epsilon_0|0,1) \mathcal{N}(\epsilon_{\tilde{\jmath}}|0,1)\right]}^{I(\tilde{x}, \tilde{c}, \mu_{\tilde{\jmath}}, \sigma_0, \sigma_{\tilde{\jmath}}, \vec{\beta})} \\
\times p(\mu_{\tilde{\jmath}}, \sigma_0, \sigma_{\tilde{\jmath}}, \vec{\beta}|\{x\}, \{\vec{c}\}, \mathcal{H}) \, d\mu_{\tilde{\jmath}} \, d\sigma_{\tilde{\jmath}} \, d\sigma_0 \, d\vec{\beta}.
\end{aligned}$$

To compute the outer integral over model parameters, we simply average draws from the MCMC:

$$\int_{\boldsymbol{\Theta}} d\mu_{\tilde{\jmath}} \, d\sigma_{\tilde{\jmath}} \, d\sigma_0 \, d\vec{\beta} \, I(\tilde{x}, \tilde{c}, \mu_{\tilde{\jmath}}, \sigma_0, \sigma_{\tilde{\jmath}}, \vec{\beta}) \approx N^{-1} \sum_{i=1}^{N} I(\tilde{x}, \tilde{c}, \mu_{\tilde{\jmath}}^{(i)}, \sigma_0^{(i)}, \sigma_{\tilde{\jmath}}^{(i)}, \vec{\beta}^{(i)}),$$

where (i) is the ith MCMC draw.

To compute the inner integrals over latent parameters, we use Hermite quadrature. Hermite polynomials provide a basis for approximating integrals of the form $\int_{-\infty}^{\infty} dx \, f(x) \exp(-x^2)$. In our case, the inner integrand is

$$\begin{aligned}
g(\epsilon_0, \epsilon_{\tilde{\jmath}}) &= \mathrm{pois}(\tilde{x}|\exp(\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \sigma_0 \epsilon_0 + \sigma_{\tilde{\jmath}} \epsilon_{\tilde{\jmath}})) \mathcal{N}(\epsilon_0|0,1) \mathcal{N}(\epsilon_{\tilde{\jmath}}|0,1) \\
&\propto \exp\left[(\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \sigma_0 \epsilon_0 + \sigma_{\tilde{\jmath}} \epsilon_{\tilde{\jmath}}) \tilde{x} - \exp(\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \sigma_0 \epsilon_0 + \sigma_{\tilde{\jmath}} \epsilon_{\tilde{\jmath}})\right] \\
&\quad \times \exp(-\epsilon_0^2/2) \exp(-\epsilon_{\tilde{\jmath}}^2/2),
\end{aligned}$$

omitting normalizing constants for brevity. We can transform this into a function of a single variable by noting that the sum of normal random variables ϵ_0 and is equivalent to the single normal random variable $\mu_{\tilde{\jmath}} + \sqrt{\sigma_0^2 + \sigma_{\tilde{\jmath}}^2}\mathcal{N}(0,1)$. Our integrand becomes

$$\begin{aligned}
g(\epsilon) &= \mathrm{pois}\left(\tilde{x}|\exp\left[\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \epsilon\sqrt{\sigma_0^2 + \sigma_{\tilde{\jmath}}^2}\right]\right) \mathcal{N}(\epsilon|0,1) \\
&\propto \underbrace{\exp\left[\left(\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \epsilon\sqrt{\sigma_0^2 + \sigma_{\tilde{\jmath}}^2}\right)\tilde{x} - \exp\left(\vec{\beta} \cdot \tilde{\vec{c}} + \mu_{\tilde{\jmath}} + \epsilon\sqrt{\sigma_0^2 + \sigma_{\tilde{\jmath}}^2}\right)\right]}_{f(\epsilon)} \exp(-\epsilon^2/2),
\end{aligned}$$

which is of the form $f(x) \exp(-x^2)$ after change-of-variable $\xi = \epsilon/\sqrt{2}$, and therefore amenable to approximation via Hermite quadrature,

$$\int_{-\infty}^{\infty} d\epsilon\, f(\epsilon)\exp\left(-\epsilon^2/2\right) = \sqrt{2}\int_{-\infty}^{\infty} d\xi\, f(\xi)\exp\left(-\xi^2\right) \approx \sqrt{2}\sum_{i=1}^{n} w_i f(x_i),$$

where x_i is the ith root of Hermite polynomial H_n, and w_i is the ith Hermite quadrature weight. Note the additional factor of $\sqrt{2}$ from transforming the volume element $d\epsilon$.

DATA AND CODE AVAILABILITY

Code for running the MCMC to sample the Log-normal-Poisson model posterior and compute significance values is available at https://github.com/broadinstitute/getzlab-LNP. Code for reproducing the manuscript figures and analyses is available at https://github.com/broadinstitute/getzlab-PHS.

Cancer Cell, Volume *36*

Supplemental Information

Passenger Hotspot Mutations in Cancer

Julian M. Hess, Andre Bernards, Jaegil Kim, Mendy Miller, Amaro Taylor-Weiner, Nicholas J. Haradhvala, Michael S. Lawrence, and Gad Getz

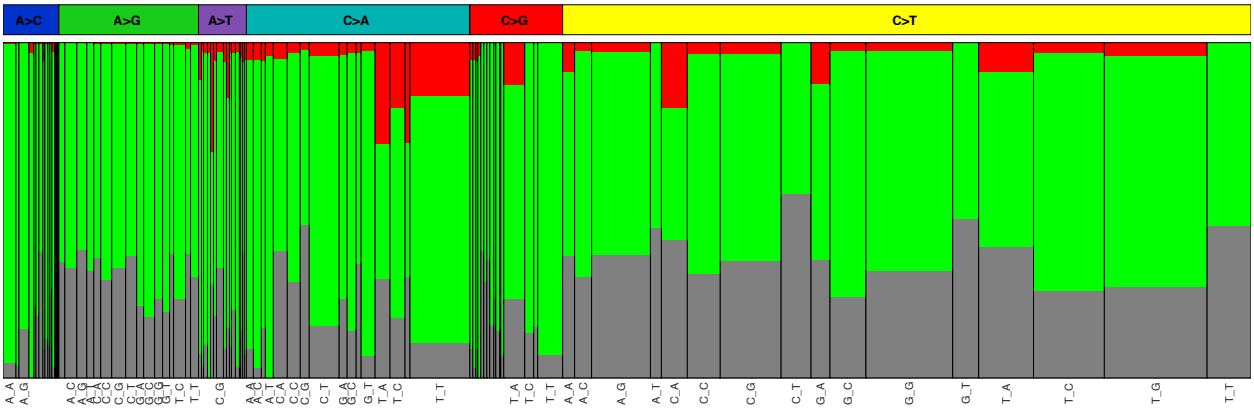


Figure S1: Expected protein-coding effects for each of the 96 trinucleotide substitutions; related to Figure 1. Width of each bar is proportional to observed mutation burden in our cohort for that substitution. Note that A(A→[CT])T and A(C→[AG])T substitutions can never generate synonymous (gray) mutations, and [CG](C→G)[ACG]/T(C→G)[CG]/[ACGT](C→T)T substitutions can never generate nonsense (red) mutations.

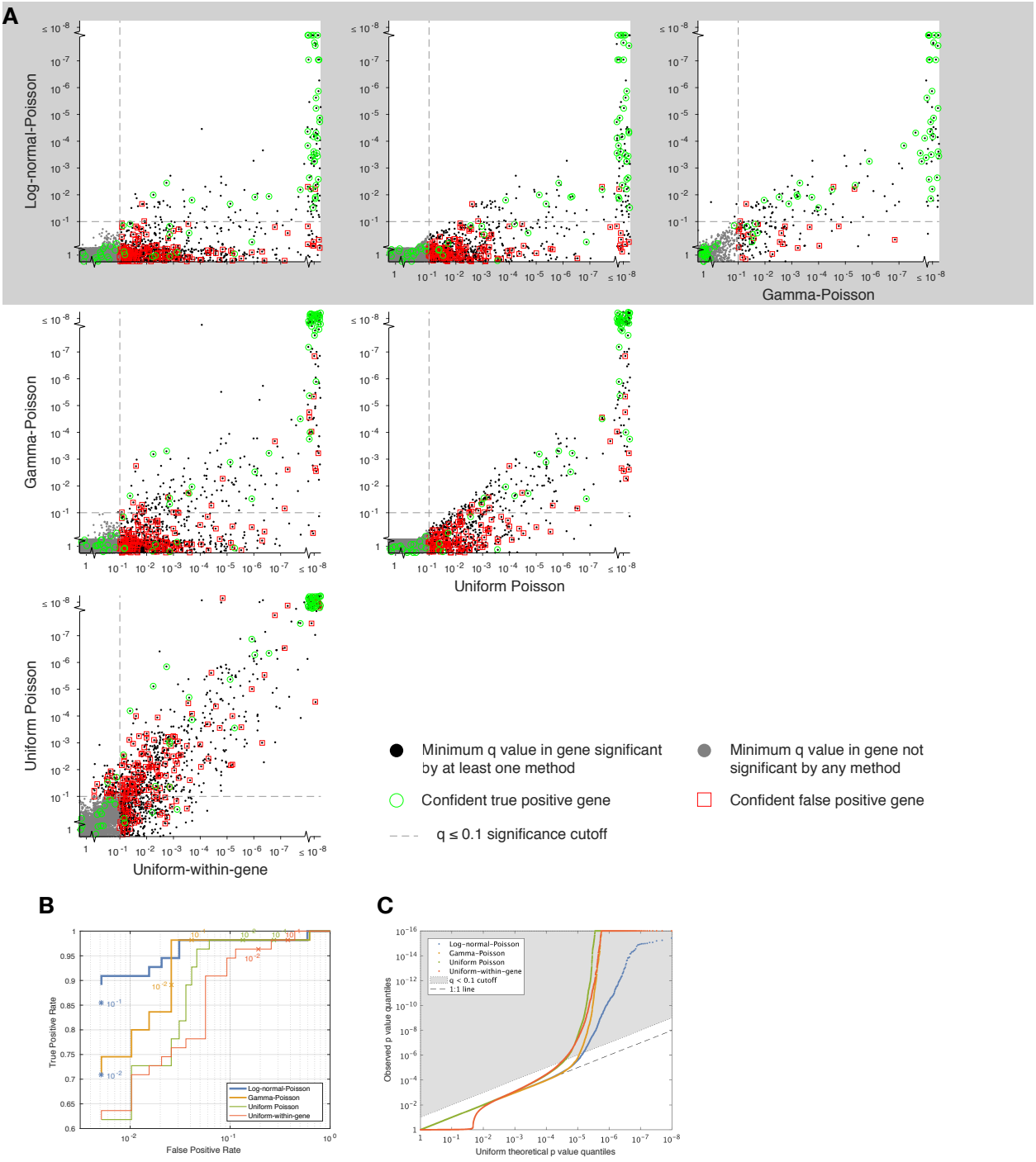


Figure S2: Comparison of methods' statistical calibration; related to Figure 2. (A) Cross-method comparison of minimum q values across all mutations in each gene. Genes in false-positive/true-positive truth sets are circled in red and green, respectively. q value cutoff of 0.1 shown as dashed lines. For clarity, points at q = 1 or q $< 10^{-8}$ are randomly jittered. Comparison of Log-normal-Poisson method against other methods highlighted in gray. (B) Alternate version of Figure 2A showing ROC curves with positive truth set comprising genes with recurrent mutations whose oncogenic roles are supported by the literature. Negative truth set comprises genes confidently under neutral selection, as before. (C) Quantile-Quantile (QQ) plots of each method's p values. Dashed 1:1 line corresponds to uniformly distributed p values; gray area represents region of Benjamini-Hochberg q value < 0.1. For legibility purposes, observed p values are capped at a minimum value of 10^{-16}.

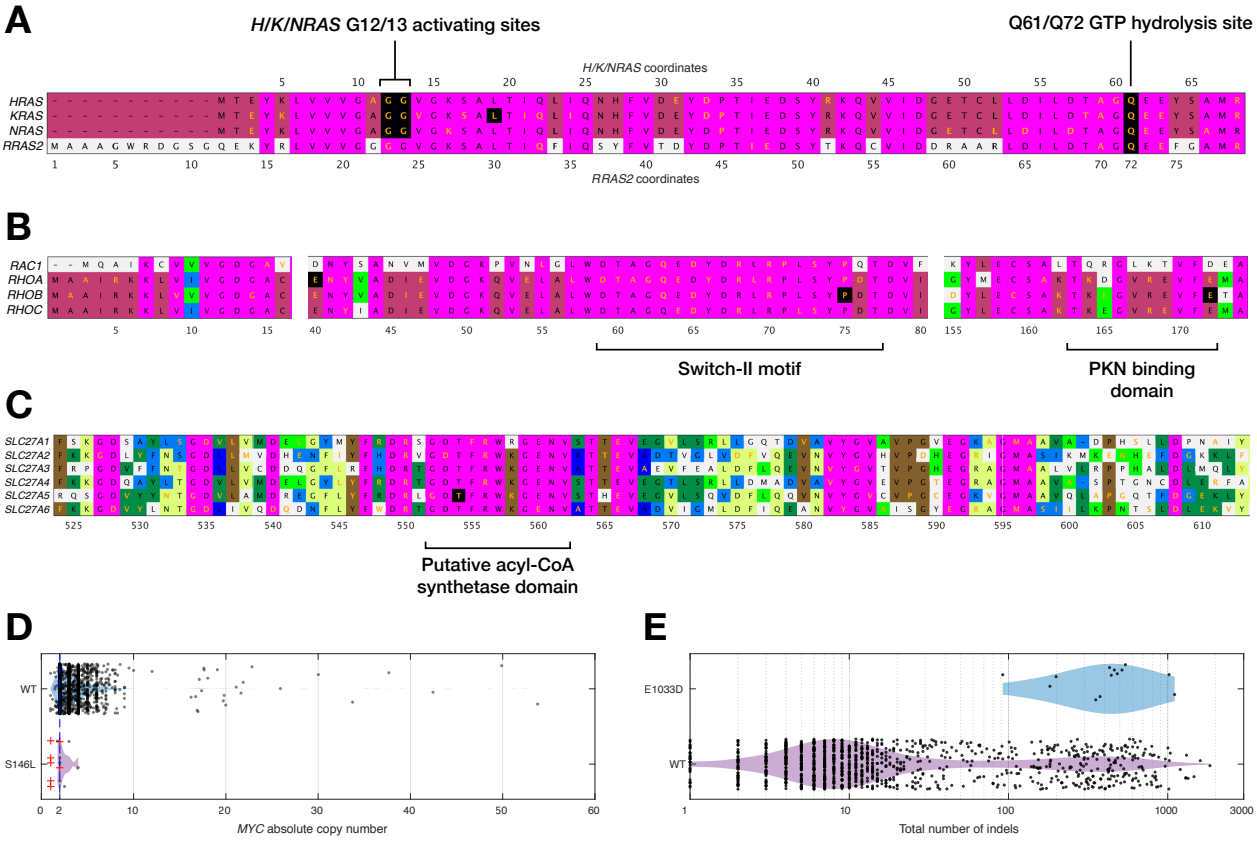


Figure S3: Functional exposition of putative driver hotspots; related to Figure 4. In all sequence alignment plots, amino acids with orange letters are mutated; amino acids with black backgrounds are significant hotspots (by the Log-normal-Poisson model). Other colors represent the number of matching amino acids across the alignment. (A) Sequence alignment of first 68 amino acids of canonical Ras subfamily members *H/K/NRAS* and first 79 amino acids of *RRAS2*, with both *H/K/NRAS* and *RRAS2* coordinates shown to emphasize homology between hydrolysis site Q61 (and well-known driver locus) of canonical subfamily members with Q72 in *RRAS2*. (B) Sequence alignments of three subsections of Rho GTPase subfamily members *RAC1* and *RHOA/B/C*. The first subsection comprises amino acids 1-16, in order to showcase similarity of N termini. The next subsection contains the highly conserved Switch-II motif, which contains a significant hotspot in *RHOB* and has high mutation burden across the four Rho GTPase subfamily members. The third subsection contains the PKN-binding domain, which also contains a significant hotspot in *RHOB*. (C) Sequence alignment of solute carrier 27 family members, illustrating significant hotspot in *SLC27A5* occurring in highly conserved motif (putative acyl-CoA synthetase domain) within a larger nonconserved context. (D) Absolute copy number at *MYC* of all wildtype patients versus S146L mutants. Red crosses show allelic copy number of somatic mutations. Ranksum of mutant vs. wildtype absolute copy number p value = 0.02. (E) Total somatic indel burden of *UPF2* E1033D mutants versus wildtype patients in the same cohort as the mutants. Ranksum of indel burden p value = 2.3×10^{-7}

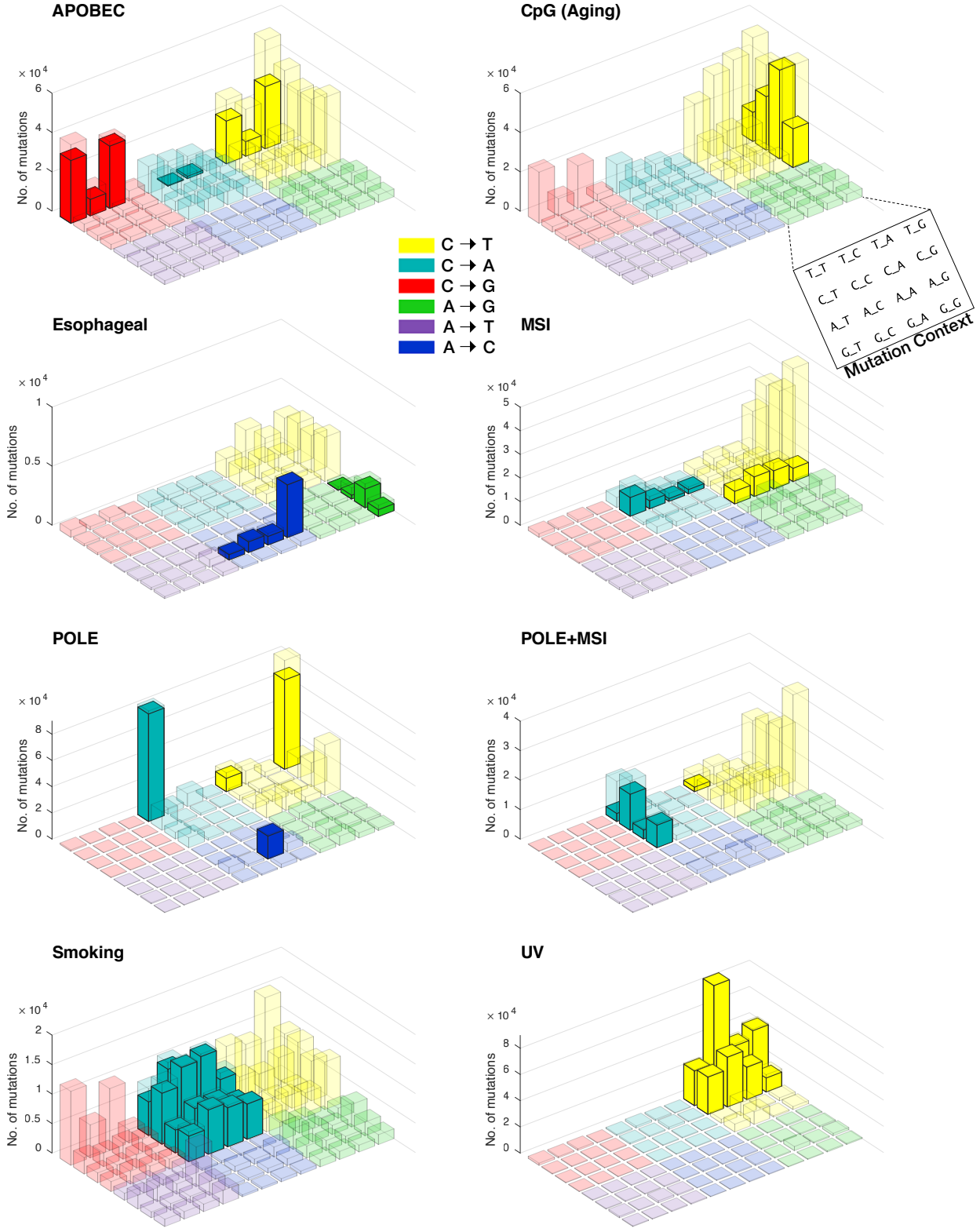


Figure S4: Total mutation counts of the 96 trinucleotide substitutions in each mutational process-centric subcohort; related to Figure 5. Opaque bars show counts of mutations definitively attributed to the relevant mutational process (Bayesian NMF factor assignment probability $\geq$0.75); transparent bars show counts of all other mutations in the process-centric subcohort not attributed to the relevant mutational process.

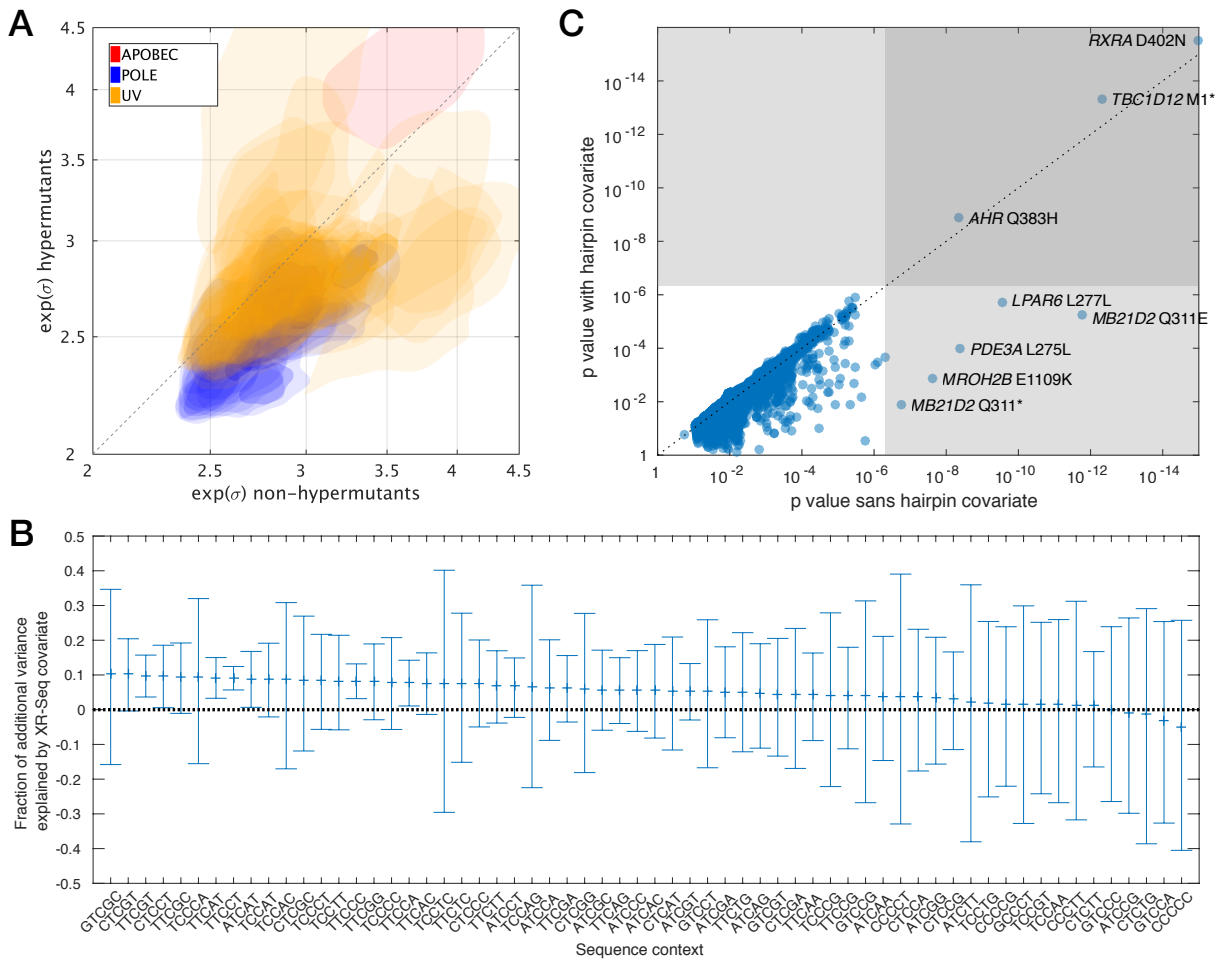


Figure S5: Properties of mutational processes; related to Figure 5. (A) Joint posterior distributions of geometric standard deviation e^{σ} for non-hypermutants vs. hypermutants. Each colored area is the 95% confidence region for a pentamer context associated with a given mutational process. Only mutational processes with sufficiently tight posterior densities in both the hypermutant and non-hypermutant partitions are shown. (B) Variance explained by including XR-seq coverage as a covariate in the UV process-centric subcohort, stratified by pentamer contexts around the 8 pyrimidine dimer trinucleotide contexts most affected by UV mutagenesis. Variance explained is presented as the fold-change in the log-normal variance after including XR-seq coverage as a model covariate, relative to log-normal variance of the model run with only the standard covariates. Bars represent 90% posterior confidence intervals; central dashes represent posterior means. (C) Comparison of Log-normal-Poisson model p values for APOBEC-induced hotspots in non-KCGs, with/without APOBEC hairpin substrate optimality covariate. Grey regions indicate p values falling below 10% FDR threshold. Note simultaneous presence of missense and nonsense events at the same nucleotide in *MB21D2*, and synonymous hotspots in *PDE3A* and *LPAR6* — evidence that these events are passenger hotspots under no positive selection.