Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences

Victor Trevino*  

*Corresponding Author: vtrevino@tec.mx  

1 Tecnologico de Monterrey, Escuela de Medicina, Av Morones Prieto No. 3000, Colonia Los Doctores, Monterrey Nuevo León, México, Zip Code 64710.
Abstract

In cancer, recurrently mutated sites in DNA and proteins, called hotspots, are thought to be raised by positive selection and therefore important due to its potential functional impact. Although recent evidence for APOBEC enzymatic activity have shown that specific types of sequences are likely to be false, the identification of putative hotspots is important to confirm either its functional role or its mechanistic bias. In this work, an algorithm and a statistical model is presented to detect hotspots. The model consists of a beta-binomial component plus fixed effects that efficiently fits the distribution of mutated sites. The algorithm employs an optimal step-wise approach to find the model parameters. Simulations show that the proposed algorithmic model is highly accurate for common hotspots. The approach has been applied to TCGA mutational data from 33 cancer types. The results show that well-known cancer hotspots are easily detected. Besides, novel hotspots are also detected. An analysis of the sequence context of detected hotspots show a preference for TCG sites that may be related to APOBEC or other unknown mechanistic biases. The detected hotspots are available online in http://bioinformatica.mty.itesm.mx/HotSpotsAnnotations.

Keywords

Hotspots, Beta-Binomial, Recurrent mutations, Cancer, Algorithm, Simulations
Introduction

It is thought that recurrently mutated amino-acid positions in cancer genes, namely mutation hotspots, are likely to have an important functional impact [1]. Several well-known examples support this view. One of the most frequent hotspots, BRAF V600E mutation, is known to over-activate the RAS pathway [2,3]. BRAF is top mutated in thyroid carcinoma [4], melanoma [5], and hairy-cell leukemia [3], and also frequent in colon and lung cancers [6–8]. Other hotspots are also well-known such R132H in IDH1 for low-grade gliomas [9], G12/G13 in KRAS for lung [10], and Q61 in NRAS for melanoma [11]. Many other genes also show hotspots [12].

Some non-cancer genes seem to show hotspots that become clear when mutations from all cancers are aggregated [1,12,13]. For example, in Chang et al. analysis [13], the RRAS2 showed a hotspot in Q72, which is still not marked as a cancer gene in the Cosmic curated database revision 2019 [14] neither detected for positive selection in Martincorena analysis [15]. This suggests that the identification of putative novel hotspots is important in cancer.

Some methods have been reported regarding the detection of mutation hotspots. Of the seminal approaches, there was a tendency to identify regions [16,17] or domains [1,18] when the available mutations were more limited. Similarly, some approaches focused on the three-dimensional protein structure to identify mutation-rich 3D-regions [19–21]. Then, position-specific models were proposed [12,13,22,23]. These approaches used a binomial or a Poisson distribution to model mutation distribution across genes. Nevertheless, the mutation distribution per gene may depend on cofactors such as sequence context [12], gene length [24], cancer type [25,26], mutational processes [26,27], or relative position along nucleosomes [28]. Modeling all these cofactors together is a very difficult task given its complexity and lack of data to sufficiently estimate embed parameters. To account for these and other unknown factors, an over-dispersion model is preferred [15,24,29]. Thus, other approaches utilize more appropriate models such as the beta-binomial model [24,29], which were applied to non-coding regions.
Although the above methods have been useful, there are some pitfalls. Some approaches use binomial or Poisson models with one or two cofactors [12,13,22] but this may lead to many false positives or negatives. For example, there are 20 genes reported by Chang et al. [13] that show significant hotspot associations backed up by only two mutations such as SESN2. The certainty to pursue experimental validations on SESN2 is very difficult if there is no available information about the parameters fitted and all related information used regarding the gene and mutations. Some methods use randomization of the mutations to estimate significance [1] but this would lead to biased estimations if not all cofactors are considered, which is difficult because there is still uncertainty about possible cofactors. Other methods use well-known cancer genes as positive controls and presumed negatives to estimate sensitivity and specificity [30]. One of the reported problems of this strategy is that it sacrifices sensitivity for specificity [30], which may show difficulties when used as a discovery tool. In this context, simulations may be a good strategy.

One of the strengths of methods that detect regions, domains, and 3D structures is that estimations can be more reliable because many more mutations can be analyzed within regions than within positions. Nevertheless, this is also a weakness because it is known that the sequence context plays a role [12] and these methods lack nucleotide sequence resolution. Another issue is that some methods focus on single nucleotide variants, presumably because of the lack of corrections for small insertions and deletions (INDELS) [13]. Regarding the types of mutations, most referred methods focus mainly on missense mutations. This is sensible because these hotspots mark positions on the protein that may change its function. Besides, missense mutations represent a large proportion of all mutations. Nevertheless, other mutations may be interesting such as those generated by small insertions and deletions that may easily accumulate at repetitive sequences [31]. A deeper analysis of methods is presented elsewhere [32].
In this work, firstly, a comparison of the fitting of the distribution of all types of small mutations by two canonical distributions (Binomial, Geometric) and two more that consider over-dispersion (Beta-Binomial, and Zero-Inflated Beta-Binomial) is presented. The comparison leads to the determination that, overall, the beta-binomial model seems to be the best model. Then, to account for genuine hotspots that do not fit well even considering over-dispersion by the beta-binomial, a mixture model with fixed effects is proposed to better fit the observed mutation distribution per gene without covariates. The need of fixed effects on high frequent mutations suggests the presence of hotspots. Simulations show that the proposed mixture model is accurate. Then, the mixture model has been applied to The Cancer Genome Atlas (TCGA) dataset and the putative hotspots are analyzed. The analysis shows that there is a bias for a sequence context centered at the mutation position and that systematic bias is observed in most co-localized olfactory receptors and other co-localized gene families. More importantly, some detected genes not considered as mutations hotspots show comparable statistics that current well-known cancer genes carrying hotspots. To the author knowledge, this is one of the few methods that use simulations to evaluate the sensitivity and specificity of the proposed method.

Material and Methods

Mutational data

The mutation annotation files (maf) were obtained from the public cancer repository TCGA (http://firebrowse.org/) in January 2018 corresponding to 33 cancer types, 10,182 patients, and 3,175,929 mutations (Supplementary Table 1). Only mutations annotated to an amino acid position within its corresponding transcript were used.

Distribution of mutated positions

For each gene, the mutations were counted per amino acid position depending on their corresponding transcript and protein. Then, the number of amino acid positions having $m_{g,j}$
mutations (from 0 to $M_g$) were aggregated where $g$ is the gene, $i$ is the number of mutations, and $M_g$ is the maximal number of mutations of gene $g$ at any amino acid position.

**Distribution Models**

To find the optimal parameters to fit a distribution model to the histogram of mutational data, a numerical method implemented in the `optim` function from the `stats` package was used minimizing the difference to the observed distribution (method=“L-BFGS-B” for function `optim` in `stats` package in R, [https://cran.r-project.org/](https://cran.r-project.org/)). To estimate the difference between fitted and observed distribution was based on the $G$-test statistic, $G = 2 \sum o_i \log(o_i/e_i)$, which is equivalent to the Kullback–Leibler divergency metric used to compare distributions. The geometric and binomial distributions were fitted using the `stats` package in R. The Zero-inflated beta-binomial (ZIBB) was fitted using the `gamlss` package in R. The beta-binomial was fitted using the `emdbook` package in R.

**Beta-binomial model with fixed effects**

Conceptually, the problem is schematized in Figure 1A while the algorithm is shown in Figure 1B. The model, $M = \text{BetaBin}(\alpha, \beta) + F$, assumes a fixed effect on positions with an excess of mutations presumably due to hotspots where $M_k$ is the number of positions carrying $k$ mutations, $F$ is the fixed hotspot effect vector, and $\text{BetaBin}$ is the beta-binomial density function scaled conveniently to sum the total number of mutations minus the sum of $F$. A stepwise algorithm was devised to fit this model. The algorithm starts setting $F_k=0$ and fitting the beta-binomial model using an optimization algorithm as described in previous section. Then, a matrix of improvements is estimated where each cell represents an independent possible fixed effect in a mutation number $k$ (in columns) and at a fraction of the total number of sites (in rows). The value of the cell is the ratio of the $G$ statistic before applying the fixed effect divided by the $G$ statistic after applying the representing fixed effect. The largest ratio represents an improvement if larger than 1 and therefore it is taken. The corresponding level (positions) and number of mutations $f_{i,k}$ are aggregated to the $F$ vector of fixed effects. The algorithm continues until the largest ratio is not greater than 1 (no improvement), when the number of steps is
larger than 3 times the maximum number of mutations, or when the $G$ statistic is lower than 1 to avoid over-fitting. The 0 positions ($k=0$), the zero mutations ($m_{g,i}=0$), or fractions that do not achieve at least 1 mutation in any $k$ mutations, are not explored. The output of the algorithm is the fixed effect vector $F$ representing the mutations and the magnitude (number of positions), $F_k$, that cannot be explained by the beta-binomial model with updated $\alpha$, and $\beta$ parameters. The algorithm was implemented in R and is available upon request.

Simulations
For simulations, sampling with the same protein size and mutations from the beta-binomial model with parameters $\alpha$ and $\beta$ was used. Then the $F$ vector was added depending on the simulation. For no hotspots, $F_k=0$, otherwise some $F_k > 0$. In any case, after running the proposed algorithm, a hotspot was recognized if the fitted value of $F_k$ is larger than 50% of the mutations at $k$. From the 2,000 genes taken for simulations, only 1,973 genes generated successful distributions.

Hotspots from cancer data
For cancer data, a hotspot or biased position was recognized if the fitted value of $F_k$ is larger than 50% of the mutations, whose mutations were 4 or more, and whose $q$-value (corrected p-value) was $<= 0.01$. These criteria were used to avoid calling hotspots in low positions that helped to improve model fitting but unlikely to represent hotspots (see Supplementary Figure 1).

Sequence context
The context sequence of a mutation was annotated using the R package BSgenome.Hsapiens.NCBI.GRCh38.
Results

Comparisons of competing distributions

To determine the best canonical distribution matching the observed mutations distributions in cancer, a comparison was performed between binomial, geometric, beta-binomial, and zero-inflated beta-binomial (ZIBB) [33]. For this, the Kullback-Leibler divergence metric was used to determine which distribution provides the best fit to the observed distribution. The ZIBB was included due to the observation that sites at zero mutations seem to be exacerbated. Under randomness, the binomial is the expected result. Nevertheless, the results show that the beta-binomial and the geometric functions capture the largest number of genes (Supplementary Figure 2A). The former is expected because the beta-binomial can capture over-dispersion commonly present in binomial data [34]. However, the geometric distribution performed surprisingly high. Then, to assess whether there is a preference of a density function for cancer genes, the same process was performed for cancer genes according to Cosmic [14] or Martincorena [15] and, on the other hand, for olfactory factors, which are believed to be mostly negative for cancer genes [35]. The results demonstrate that the beta-binomial and the geometric distributions dominates the best fit (Supplementary Figure 2A). If only the beta-binomial and the geometric distributions were compared, 63% of the genes were best fitted using the beta-binomial (Supplementary Figure 2B). Moreover, for those genes best fitted with the geometric distribution, 98% would best fit the beta-binomial if the geometric were not considered whereas the genes best fitted with beta-binomial would not prefer the geometric (Supplementary Figure 2C). These results suggest that, overall, the best distribution tested is the beta-binomial.

Hotspot detection algorithm

As shown above, the beta-binomial distribution seems to be a good model for most of the genes and it has been used to estimate recurrent alterations [36–38]. The use of a distribution is interesting because it provides the probability of observing k mutations allowing the possibility of assigning a p-value to biased amino acid positions (putative hotspots). Although a distribution could be a good model, the presence of hotspot mutations or biased sites would
artificially increase the mutations counts at specific positions generating longer tails. This will generate deviations in the parameter values modifying the corresponding p-values and therefore falsely calling or not calling hotspots at uncertain conditions. To handle this, a mixed model is proposed having two components as $M = \text{BetaBin}(\alpha, \beta) + F$ where $M_k$ is the count of amino acid positions mutated $k$ times. Without hotspots or deviated sites, the $F$ vector is zero (all $F_k=0$) and the number of amino acid sites mutated are explained entirely by the beta-binomial component. This would generate very low differences between the observed and fitted distribution, which is measured by the Kullback-Leibler (KL) divergency (or $G$-test, see Methods). In the presence of hotspots or sequence biases, the KL divergency will be higher. Nevertheless, within the model, $F$ can absorb the excess of amino acid positions at $k$ mutations ($F_k > 0$), providing a better fit for the beta-binomial and lowering the KL divergency. Therefore, the problem is to finding the optimal $\alpha$, $\beta$, and $F$. For this, the devised step-wise algorithm, schematized in Figure 1B, first sets $F_k=0$, then finds the most deviated amino acid positions at $k$ mutations looking for lower values of cell scores. This is achieved exploring the possible combinations of $k$ mutations and fractions of amino acid positions. In the example shown in Figure 1B, the first iteration finds $F_7 = 1$ while the second iteration finds $F_6 = 1$. The process ends because there is no sufficient improvement at the third iteration. In this way, the fitted beta-binomial, conditioned to the fitted $F$, is more representative of most sites and mutations providing an unbiased estimation of the probability of $k$ mutations at updated parameters $\alpha$ and $\beta$, which can be very different that without using the fixed effect at the start of the algorithm.

Assessing the performance of the proposed algorithm

To objectively evaluate the performance of the proposed algorithm, simulations were used. The first simulation was performed assuming no hotspots. To simulate realistic scenarios, all genes were first fit to the beta-binomial without a fixed effect. Then, the observed $\alpha_g$ and $\beta_g$ values for 2,000 random genes $g$ were used to generate positions distributions at the same number of the observed mutations. Finally, the proposed algorithm was run with this artificial data. The
results show that the proposed algorithm has a specificity of 84.3% recognizing 0 hotspots when there are none (Figure 2A).

The second simulation was performed assuming one or more hotspots (or biased amino acid positions). Note that the number of amino acid positions or the number of mutations is important because it could deviate far from the overall distribution or can be masked within dense regions of the distribution. For example, in Figure 1A, there is one hotspot carrying 6 mutations and another carrying 7 mutations, which are at +3 and +4 mutations farther than the last mutated ‘random’ mutation at 4. Similarly, in Figure 2B, two examples are shown. First, two hotspots are added having 3 mutations (relative to the maximum 4, these are at \( r_{Mut}=-1 \)). Then one hotspot is added at 5 mutations (\( r_{Mut}=+1 \)). To generalize for any gene, for the simulations, the number of amino acid positions injected were \( n_{Hot}={1, 3, 5} \) whereas the number of mutations tested was \( r_{Mut}={-3, -2, -1, 0, 1, 3, 5} \) relative to the maximum number of observed mutations. In this way, injected hotspots at \( r_{Mut} \leq 0 \) are harder to detect because are mixed with the overall distribution. Contrary, high values of \( r_{Mut} \) or larger \( n_{Hot} \) are easier to detect because the alteration has a deeper impact on the distribution. For these simulations, the same 2,000 genes employed in the first simulations were used. The results show that the proposed algorithm only fails to detect at least one hotspot in 15% of the simulations (Figure 2C). Thus, the algorithm has an overall sensitivity of 85%. Nevertheless, in more than 10% of the simulations, more than one hotspot was detected. To study the conditions of this behavior deeply, the performance of the algorithm for different \( r_{Mut} \) values was analyzed as shown in Figure 2D. The ideal well-known hotspots should contain more than the maximum random mutations, which corresponds to \( r_{Mut} > 0 \). The performance in these ideal hotspots was \( \geq 99\% \) for 1, 3, and 5 injected hotspots. If the hotspots are precisely the ones at the maximum number of mutations (\( r_{Mut}=0 \)), the performance is 76% if there is only one hotspot, or close to 100% if there are 3 or more. If a hotspot is present but in the observed data is still below the maximum number of mutations (corresponding to \( r_{Mut} < 0 \)), the performance decreases with both \( n_{Hot} \), and \( r_{Mut} \) (Figure 2D). This scenario seems counterintuitive but because data in cancer has not been uniformly nor comprehensive acquired in all cancer types, it may be still
useful if detected. In these cases, if only one hotspot is present, the overall performance decreases to 81%, 50%, or 21% corresponding to -1, -2, and -3 relative mutations and more than 15% of the times another false ‘hotspot’ is detected (see rows at $r\text{Mut}=-1,-2,-3$ and columns 2, 4, and 6). When three or five hotspots are present below the maximum mutations, the performance is in general higher but also increases the number of false ‘hotspots’ detected.

In summary, the proposed algorithm has an ideal performance (> 99% sensitivity and specificity) when the hotspots are those at the maximum number of mutations and the performance decreases with the number of hotspots or the relative position to the maximum number of mutations.

**Detecting hotspots in cancer data**

From the proposed algorithm, the fixed effects $F$ absorbs those positions that cannot be explained by the beta-binomial model alone. Thus, the fixed effect vector $F$ mark hotspots while the fitted beta-binomial is able to, less biasedly, estimate its probability. The p-value was then corrected by a false discovery rate (FDR) approach [39]. Because potential hotspots are only those with a sensible number of recurrent positions, the FDR correction was estimated for sites whose recurrence were 4 or more. Only positions having FDR $\leq 0.01$ were considered as hotspots. This was applied to TCGA mutational data, which includes 3,175,929 mutations from 10,182 patients across 33 cancer types (Supplementary Table 1). As a correction, hotspots were also called if the number of mutations were 9 or greater which includes many amino acid positions in $TP53$, $PIK3CA$, and $PTEN$, which result presumably to the overwhelming number of hotspots in these genes (Supplementary Figure 3). The detected hotspots are part of a database, Hotspots Annotations, available online (http://bioinformatica.mty.itesm.mx/HotSpotsAnnotations). Some representative examples of the hotspot detection are shown in Figure 3. For a well-known cancer gene, $EGFR$, 4 hotspots are clearly recognized carrying from 11 to 27 mutations. In addition, there were 4 AA positions carrying 5 mutations, 1 of 6 mutations, and 2 of 7 mutations that were effectively recognized by the algorithm but that were not significant under the above criteria after FDR correction.
Similarly, for NBPF12 and GK2, not recognized as cancer genes in COSMIC, there were 1 hotspot accumulating 12 mutations in NBPF12, and 4 hotspots showing 5 to 6 mutations in GK2. In total, 3,860 hotspots were detected in 3,115 genes where 2,639 genes had only 1 hotspot, 378 genes contain 2 hotspots, and 98 genes showed 3 or more hotspots (Figure 4A). These hotspots cover 39,815 mutations representing 1.25% of the total mutations and 0.19% of the mutated sites. Common cancer genes showed many hotspots such as TP53, PIK3CA, APC, PTEN, CDKN2A, ARID1A, FBXW7, NFE2L2, and 6 or more were estimated in ERBB2, CTNNB1, BRAF, CIC, KMT2D, and DNAH5. TTN showed 7 ‘hotspots’ but has been marked repeatedly as a ‘false positive’ gene due to its size. The Table 1 shows the 96 genes showing 3 or more hotspots. This list is highly enriched in cancer genes, it contains 38% (n=37, \( p < 10^{-53} \)) and 39% (n=38, \( p < 10^{-31} \)) cancer genes from Cosmic [14] and Martincorena [15] respectively. Hotspots containing many mutations are commonly well-known because they have been spotted time ago such as IDH1 in gliomas, BRAF in thyroid, melanoma, and other cancer types. Nevertheless, an analysis of the distribution of mutations show high density corresponding to mutations between 5 and 9 reaching ~70% of detected hotspots (Figure 4B). This suggest that many hotspots are needed to be analyzed and experimentally studied.

**Table 1. Genes showing 3 or more recognized hotspots.**

| Gene      | HotSpots | Mutations Min, Max | Gene      | HotSpots | Mutations Min, Max |
|-----------|----------|--------------------|-----------|----------|--------------------|
| BRAF      | 6        | 11, 594            | KIAA2026  | 3        | 6, 11              |
| KRAS      | 4        | 24, 564            | MDN1      | 3        | 7, 11              |
| PIK3CA    | 22       | 9, 290             | PBRM1     | 3        | 9, 11              |
| TP53      | 63       | 16, 251            | PPM1D     | 3        | 6, 11              |
| NRAS      | 3        | 15, 203            | ZNF442    | 3        | 8, 11              |
| PTEN      | 15       | 10, 112            | DNAH5     | 5        | 9, 10              |
| FBXW7     | 9        | 9, 69              | ATRX      | 3        | 7, 10              |
| JAK1      | 3        | 14, 60             | C5orf42   | 3        | 7, 10              |
| CTNNB1    | 6        | 30, 50             | CSMD3     | 3        | 9, 10              |
| HRAS      | 4        | 7, 50              | MECOM     | 3        | 9, 10              |
| CDKN2A    | 12       | 10, 41             | UVRAG     | 3        | 5, 10              |
| APC       | 18       | 9, 40              | TTN       | 7        | 8, 9               |
| ERBB2     | 6        | 9, 40              | MGA       | 4        | 7, 9               |
| Gene    | Chr | Start | End   | Gene    | Chr | Start | End   |
|---------|-----|-------|-------|---------|-----|-------|-------|
| PPP2R1A | 3   | 13,33 | OR51S1 | 4     | 8,9 |
| NFE2L2  | 9   | 7,32  | ZNF292 | 4     | 6,9 |
| ARID1A  | 9   | 9,29  | ADAMTS3 | 3    | 7,9 |
| EGFR    | 4   | 11,27 | ADNP   | 3     | 6,9 |
| KMT2D   | 5   | 9,26  | ALB    | 3     | 7,9 |
| FGFR2   | 3   | 8,25  | CHD1   | 3     | 7,9 |
| SCAF4   | 3   | 10,24 | DNAAH7 | 3     | 9,9 |
| SF3B1   | 4   | 7,21  | RASA1  | 3     | 8,9 |
| SPOP    | 4   | 9,19  | SAMD9  | 3     | 8,9 |
| PIK3R1  | 4   | 10,17 | TRIM23 | 3     | 6,9 |
| KMT2B   | 3   | 11,17 | ZNF14  | 3     | 7,9 |
| MBD6    | 3   | 12,17 | ZNF732 | 3     | 7,9 |
| MFRP    | 3   | 6,17  | KIF20B | 4     | 7,8 |
| TPTE    | 4   | 9,16  | SLCO1B7 | 4   | 5,8 |
| CD93    | 3   | 12,16 | CWF19L2 | 3   | 7,8 |
| KANSL1  | 4   | 8,15  | FAM193A | 3    | 6,8 |
| CHD4    | 3   | 11,15 | OR2T2  | 3     | 7,8 |
| CTCF    | 3   | 11,14 | UNC79  | 3     | 7,8 |
| ZBTB7C  | 3   | 9,14  | ZNF502 | 3     | 6,8 |
| CIC     | 5   | 7,13  | MSH6   | 4     | 6,7 |
| NF1     | 4   | 8,13  | VPS13C | 4     | 6,7 |
| ARHGAP5 | 3   | 7,13  | BTBD7  | 3     | 6,7 |
| PRKDC   | 3   | 8,13  | CCDC27 | 3     | 7,7 |
| SMAD2   | 3   | 7,13  | CFAP61 | 3     | 7,7 |
| YLPM1   | 3   | 7,13  | GTF3C4 | 3     | 5,7 |
| CASP8   | 4   | 9,12  | PTPN11 | 3     | 5,7 |
| ANK3    | 3   | 9,12  | TDRD6  | 3     | 6,7 |
| CNOT1   | 3   | 7,12  | CCDC168 | 4    | 6,6 |
| CNTNAP2 | 3   | 9,12  | CSGALNACT1 | 3   | 5,6 |
| MYOCRD  | 3   | 10,12 | GK2    | 3     | 6,6 |
| THSD7B  | 3   | 10,12 | PTPN13 | 3     | 5,6 |
| ZFHX4   | 3   | 9,12  | RALGAPA1 | 3   | 6,6 |
| ALG13   | 3   | 8,11  | RPS6KA5 | 3    | 5,6 |
| C6      | 3   | 8,11  | MAN2A1 | 4     | 5,5 |
| DDX17   | 3   | 5,11  | B3GAT2 | 3     | 5,5 |
| HCN1    | 3   | 9,11  | CLCA4  | 3     | 5,5 |
Variant types and sequence context in hotspots

Most hotspots methods focus on missense and nonsense mutations, which cover around 75% of all mutations. This has the advantage of focusing on clear biological effects but has the disadvantage of ignoring possible sequence biases that may help to recognize mechanistic effects. In addition, the proposed algorithm is inspired in estimating biases in the distribution of mutations along protein coding regions, which will be affected by selecting types of mutations. Therefore, all small mutations types were used. The disadvantage, however, is that not all variant types may show an interesting biological effect. In addition, it is known that hotspots may be focalized in specific sequence contexts [40]. Accordingly, a comparison of variant types and sequence contexts were performed between hotspots and the overall data in unique positions. To clearly expose the differences, only hotspots carrying 10 or more mutations were compared as shown in Figure 5 while the complete analysis is shown in Supplementary Figure 5.

From the input data, the most frequent variant types are missense, silent, and nonsense accumulating 1.44, 0.564 and 0.116 million mutations. In hotspots, although the most frequent mutations are missense (n=750) surprisingly, frame shift deletions counts are very similar (n=742) even that frame shift deletions are more than 20 times less frequent in the overall data. Frame shift insertions were also high (n=327).

The Figure 5 clearly show that while the sequence context TCN dominates the overall mutated positions mainly in the TCT sequence context (where the C marks the site of mutation), the TCG is by far the most recurrent context for hotspots while TCT, TCA, and TCC generally decrease. This pattern seems to be clearly present in missense and nonsense and partially also in silence mutations suggesting that there is some type of preference or selection for the TCG context in these types of variants. Similarly, for hotspots carrying 5 to 9 mutations, the TCG increase is also observed (Supplementary Figure 4). However, in these hotspots, an increase in GCC, then CCG and AC, were also present suggesting that the overall preference for 5 to 9 mutations seems to be xNCG. All these results concur with the pattern of mutations from APOBEC [40].
For frame shift deletions the observed differences are not so strong, suggesting that, overall, selection pressure is absent or low. The highest increases in differences (+5 relative %) were in ACC, CTT, and TTA. For other types of variants, the changes or the number of occurrences in hotspots are low.

Hotspots across cancer types

It is known that cancer types differ in the frequency of mutations per gene [35]. It has also been proposed that driver mutations may accumulate from 1 to 10 depending on the cancer type [15]. Therefore, a comparison of hotspots across cancer types were performed. First, it was noted that the percentage of samples not carrying any hotspot mutation formed three to four clusters of cancer types (Figure 6A), which also correlated with the overall mutation rate. The clusters include more than 60% of samples (TGCT, KIRP, KIRC, MESO, PCPG, PRAD, KICH, and ACC), then between 25% and 60% of samples (THCA, THYM, OV, GBM, BRCA, CESC, LAML, DLBC, LIHC, SARC, CHOL), those between 10% and 25% (PAAD, LGG, LUSC, HNSC, ESCA, LUAD, BLCA, STAD), and those below 10% (SKCM, COAD, UCEC). UVM, UCS, and READ show also low percentage of samples not carrying hotspots but its distribution is more similar to one of the first three clusters. STAD, SKCM, COAD, and UCEC show around 20% or more samples carrying 10 or more hotspots, which is also consistent with the high rate of mutations of these cancer types. It is well known that TP53, PIK3CA, and RAS gene family show recurrence in many cancer types but others genes are more specific. For example, IDH1/2 in gliomas, AKT1 and GATA3 in BRCA, SPOP in PRAD, and BRAF in THCA. Therefore, three approaches were performed to highlight cancer-specific hotspots. First, the top 10 most frequent hotspots per cancer type were estimated as shown in Table 2. Beside the above cancer-specific genes, other high frequent hotspot can be noted such as GTFI2 in THYM, GNAQ in UVM, CTNNB1 in LIHC, VHL in KIRC, CDKN2A in HNSC, and NFE2L2 in LUSC. Second, an analysis of the number of cancer types per hotspot shows that most hotspots (91%) are formed by mutations from 2 to 6 cancer types (Figure 6B). Thus, only 95 hotspots (2.46%) are strictly cancer type-specific (Figure 6C). For example, VHL p.158 in KIRC, APC p.935 in COAD, and CDH1 p.23 in BRCA. Third, because of these results, for each hotspot the major cancer type was calculated. Then, if its contribution to
the total number of mutations were higher than 50% or if it were higher than 25% and the number of mutations were higher than 10, it was selected as ‘cancer-enriched’. Thus, the number of hotspots per cancer type was very high for UCEC, STAD, SKCM, and COAD as shown in Figure 6C, presumably due to high mutations rates. The Table 3 shows the hotspots for the rest of cancer types and the complete list is shown in Supplementary Table 1. This is interesting because it highlights genes not well studied such as NBPF12 in BRCA, LPAR6 or ASXL2 in BLCA, and FGGY in LUSC, which is being studied recently [41].

Table 2. Top 10 most frequent hotspots per cancer type (# patients GENE position).

| TYPE  | TOP 1 | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      | 10     |
|-------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| ACC   | 6 TMEM24 | 5 CTNNB1 | 3 CTNNB1 | 2 MUC4 | 3515  | 2 OR4K2 | 207    | 2 RPL22 | 15     | 2 TP53 | 125    |
|       | 7 128  | 45      | 34      |        |        |        |        |        |        |        | 394    |
| BLCA  | 35 PIK3CA | 30 FGFR3 | 24 TP53 | 248    | 22 ERBB2 | 310    | 15 TP53 | 280    | 14 RXRA | 427    |
|       | 545    | 249     | 248     | 18     | PIK3CA  | 542    |        |        |        |        |        |
| BRCA  | 133 PIK3CA | 69 PIK3CA | 41 PIK3CA | 25 AKT1 | 17   | GATA3  | 308    | 20 TP53 | 273    | 19 TP53 | 175    |
|       | 1047   | 545     | 542     | 21     |        |        |        |        |        |        | 345    |
|       |        |         |         |        |        |        |        |        |        |        | 407    |
|       |        |         |         |        |        |        |        |        |        |        | 464    |
|       |        |         |         |        |        |        |        |        |        |        | 419    |
|       |        |         |         |        |        |        |        |        |        |        | 28     |
| CESC  | 37 PIK3CA | 23 PIK3CA | 10 MAPK1 | 7 FBXW7 | 505  | 7 KRAS | 12     | 6 ERBB2 | 310    | 6 FBXW7 | 465    |
|       | 545    | 542     | 322     | 10     |        |        |        |        |        |        |        |
|       |        |         |         |        |        |        |        |        |        |        |        |
| CHOL  | 5 IDH1 | 5 IDH2 | 2 ERBB2 | 2 IDH2 | 755   | 172    |        |        |        |        |        |
| COAD  | 102 KRAS | 49 BRAF | 35 PIK3CA | 545    | 32 KRAS | 13     | 27      | SETD1B | 5 175  | 23 APC | 1450   |
|       | 12     | 600     | 13      | 27     | 175    |        |        |        |        |        | 526    |
|       |        |         |         |        |        |        |        |        |        |        | 692    |
| DLBC  | 2 B2M 1 |        |        |        |        |        |        |        |        |        |        |
| ESCA  | 15 TP53 | 11 TP53 | 9 TP53  | 237    | 135    | 6 TP53  | 135    | 5 TP53  | 220    | 5 TP53  | 282    |
|       | 248    | 175     | 273     |        |        |        |        |        |        |        | 282    |
|       |        |         |         |        |        |        |        |        |        |        | 79     |
|       |        |         |         |        |        |        |        |        |        |        | 1047   |
|       |        |         |         |        |        |        |        |        |        |        | 187    |
| GBM   | 22 IDH1 | 21 EGFR | 14 EGFR | 598    | 13 TP53 | 248    | 7 PTEN | 130    | 7 TP53  | 175    | 7 TP53  | 273    |
|       | 132    | 289     | 248     | 13     | 175    |        |        |        |        |        | 273    |
|       |        |         |         |        |        |        |        |        |        |        | 376    |
|       |        |         |         |        |        |        |        |        |        |        | 282    |
|       |        |         |         |        |        |        |        |        |        |        | 173    |
| HNSC  | 24 PIK3CA | 20 CDKN2A | 18 PIK3CA | 13 TP53 | 248  | 13 TP53 | 273    | 12 TP53 | 175    | 11 CDKN2A | 5 175  |
|       | 545    | 80      | 542     | 1047   | 238    | 273    | 175    |        |        |        | 12     |
|       |        |         |         |        |        |        |        |        |        |        | 13     |
| KICH  | (none)  |        |        |        |        |        |        |        |        |        |        |
| KIRC  | 9 VHL 155 | 9 VHL 158 | 2 PBRM1 | 710    | 2 PWWP2A | 270    |        |        |        |        |        |
| KIRP  | 5 KRAS 12 | 4 ERBB2 | 3 PIK3CA | 542    | 2 BRAF | 600    | 2 FGFR3 | 373    | 2 NFE2L2 | 82     | 2 OR13G1 | 383    |
|       | 755     | 285     | 542     | 1047   | 12      |        |        |        |        |        | 54     |
| TUMOR | Genes | Count |
|-------|-------|-------|
| LAML  | 12 NPM1 10 DNM3A 10 FLI3 9 IDH2 7 IDH1 5 KIT 816 4 NRAS 4 RIMS4 3 KRAS 3 NRAS | 287 |
| LGG   | 390 IDH1 59 TP53 20 IDH2 14 TP53 12 CIC 10 TP53 9 TP53 8 TP53 7 TP53 6 ATRX | 132 |
| LIHC  | 17 CTNNB1 17 CTNNB1 12 CTNNB1 10 TP53 9 TP53 8 TP53 7 TP53 5 TP53 | 32 |
| LUAD  | 136 KRAS 23 EGFR 10 TP53 10 TP53 9 TP53 8 BRAF 8 KRAS 8 TP53 7 TP53 | 858 |
| LUSC  | 18 TP53 17 TP53 17 TP53 16 NFE2L2 14 TP53 14 TP53 12 TP53 11 TP53 | 125 |
| MESO  | 2 PTEN 2 TP53 | 246 |
| OV    | 163 TP53 20 TP53 16 TP53 12 TP53 10 TP53 10 TP53 9 TP53 8 TP53 7 TP53 | 248 |
| PAAD  | 313 KRAS 10 TP53 8 NGS 6 TP53 | 12 |
| PCPG  | 16 HRAS 2 FGFR1 2 HRAS | |
| PRAD  | 19 SPOP 14 SPOP 10 SPOP 8 TP53 5 TP53 4 IDH1 3 PIK3CA | 133 |
| READ  | 41 KRAS 13 TP53 11 TP53 10 APC 9 TP53 7 TP53 6 TP53 | 12 |
| SARC  | 5 TP53 4 TP53 4 TP53 3 KRATAP1 3 NUP93 | 175 |
| SKCM  | 243 BRAF 110 NRAS 22 RAC1 21 SLC27A5 17 MAP2K1 16 IDH1 15 BCL2L12 | 600 |
| STAD  | 31 XYL2 29 ACVR2A 26 SLC3A2 25 RPL22 21 UBR5 17 LARP4B | 526 |
| TGCT  | 11 KIT 7 KRAS 12 KRAS 3 KRAS 3 NRAS 2 KRAS 2 NRAS 2 PIK3CA | 816 |
| THCA  | 281 BRAF 39 NRAS 17 HRAS 5 INTS2 5 INTS2 3 AKT1 3 NUP93 2 BRAF | 600 |
| THYM  | 62 GTF2I 4 HDAC4 4 HRAS 3 NRAS 2 SRP18 179 | 424 |
| UCEC  | 78 PTEN 67 KRAS 49 SETD1B 47 RPL22 41 JAK1 40 RNF43 32 DOCK3 | 130 |
| UCS   | 7 FBXW7 7 KRAS 7 TP53 5 PIK3CA 5 PIK3CA | 465 |
| UVM   | 37 GNAQ 34 GNA11 14 SF3B1 | 209 |
### Table 3. Cancer enriched hotspots.

| Cancer | HotSpot | N  | %  | Cancer | HotSpot | N  | %  | Cancer | HotSpot | N  | %  |
|--------|---------|----|----|--------|---------|----|----|--------|---------|----|----|
| BLCA   | FGFR3 249 | 30 | 83 | BRCA   | PAK3CA 1047 | 133 | 47 | LGG    | IDH1 132 | 390 | 85 |
|        | ERBB2 310 | 22 | 55 |        | AKT1 17 | 25 | 47 |        | IDH2 172 | 20 | 77 |
|        | TP53 280 | 15 | 33 |        | GATA3 308 | 21 | 95 |        | CIC 215 | 12 | 92 |
|        | RXRA 427 | 14 | 82 |        | PAK3CA 345 | 16 | 40 |        | ATRX 1426 | 6 | 60 |
|        | TP53 285 | 10 | 34 |        | GATA3 407 | 12 | 75 |        | CIC 1512 | 6 | 55 |
|        | C3orf70 6 | 9  | 45 |        | PAK3CA 726 | 10 | 33 |        | EGFR 252 | 5 | 45 |
|        | ERCC2 238 | 9  | 100|        | CDH1 23 | 9 | 100|        | EGFR 858 | 23 | 100|
|        | AHR 383  | 8  | 73 |        | HIST1H2A 128 | 9 | 43 |        | BRAF 469 | 8 | 36 |
|        | FGFR3 373 | 8  | 80 |        | SF3B1 700 | 8 | 57 |        | BRAF 466 | 6 | 46 |
|        | KDM6A 555 | 8  | 100|        | ERBB2 755 | 7 | 39 |        | STK11 51 | 6 | 75 |
|        | LPAR6 316 | 7  | 88 |        | NBP1F12 125 | 7 | 58 |        | OR2T2 14 | 4 | 50 |
|        | SF3B1 902 | 7  | 100|        | PAK3CA 453 | 7 | 28 |        | SNPD3 96 | 4 | 67 |
|        | RARS2 6  | 6  | 43 |        | RTF1 235 | 5 | 42 |        | GBM 289 | 21 | 78 |
|        | TP53 271 | 6  | 32 |        | ERBB2 777 | 4 | 31 |        | EGFR 598 | 14 | 74 |
|        | CELSR3 356 | 5  | 83 |        | FOXA1 226 | 4 | 57 |        | PIK3R1 376 | 6 | 46 |
|        | MROH2B 1109 | 5  | 56 |        | CDKN2A 80 | 20 | 49 |        | KRTAP4-6 62 | 4 | 57 |
|        | PDE3A 275 | 5  | 42 |        | CDKN2A 58 | 11 | 52 |        | PTEN 132 | 4 | 36 |
|        | TFPI2 222 | 5  | 50 |        | HRAS 12 | 11 | 58 |        | LIHC 33 | 17 | 40 |
|        | ACTB 158 | 4  | 57 |        | HRAS 13 | 11 | 35 |        | CTNNB1 45 | 17 | 46 |
|        | ASXL2 330 | 4  | 67 |        | CDKN2A 153 | 10 | 56 |        | EEF1A1 432 | 8 | 73 |
|        | C12orf43 | 4  | 50 |        | CDKN2A 110 | 9 | 43 |        | MUC4 3515 | 6 | 40 |
|        | FOXO1 135 | 4  | 80 |        | RAC1 159 | 6 | 75 |        | ADRA1D 554 | 4 | 80 |
|        | HIST2H2BE 71 | 4  | 44 |        | TP53 298 | 6 | 33 |        | THCA 281 | 47 |
|        | RB1 405 | 4  | 44 |        | CDKN2A 88 | 5 | 38 |        | HRAS 61 | 17 | 34 |
|        | TMCO4 13 | 4  | 50 |        | EP300 1399 | 5 | 33 |        | INTS2 577 | 5 | 62 |
| LUSC   | TP53 125 | 18 | 26 |        | KRT6A 487 | 5 | 71 |        | INTS2 578 | 5 | 71 |
|        | TP53 158 | 17 | 33 |        | CDKN2A 51 | 4 | 27 |        | PRAD 133 | 19 | 100|
|        | NFE2L2 34 | 16 | 50 |        | TP53 195 | 12 | 30 |        | SPOP 131 | 14 | 88 |
|        | NFE2L2 29 | 14 | 48 |        | RIF1 1718 | 6 | 55 |        | SPOP 102 | 8 | 89 |
|        | TP53 157 | 14 | 36 |        | ZNF12 417 | 6 | 86 |        | GNAQ 209 | 37 | 92 |
|        | NFE2L2 79 | 10 | 34 |        | FAH 153 | 5 | 62 |        | GNA1 209 | 34 | 89 |
|        | TP53 234 | 9  | 38 |        | BRAP 577 | 4 | 67 |        | SF3B1 625 | 14 | 67 |
|        | CDKN2A 84 | 7  | 41 |        | DDR2 85 | 4 | 67 |        | CESC 53 | 10 | 53 |
|        | MB12D 311 | 7  | 28 |        | SLC9A4 353 | 4 | 36 |        | KLF5 419 | 5 | 38 |
|        | CDKN2A 108 | 6  | 40 |        | LAM1 287 | 12 | 100|        | KIRC 155 | 9 | 90 |
|        | KRT5 492 | 6  | 55 |        | DNMT3A 882 | 10 | 83 |        | VHL 158 | 9 | 100|
|        | NFE2L2 31 | 6  | 55 |        | FLT3 835 | 10 | 100|        | THYM 424 | 62 | 97 |
|        | TP53 105 | 5  | 31 |        | IDH2 140 | 9 | 82 |        | HDAC4 746 | 4 | 33 |
|        | FGGY 484 | 4  | 44 |        | NRAS 13 | 4 | 27 |        | ACCM247 128 | 6 | 30 |
|        | NFE2L2 30 | 4  | 40 |        | RIMS4 85 | 4 | 40 |        | READ 537 | 5 | 50 |
|        | PTEN 245 | 4  | 40 |        | STAD (444 see Suppl) | (444 see Suppl) | 40 | 40 |        | KIT 816 | 11 | 65 |
| UCEC   | (1076 see Suppl) | 7  | 28 |        | SKCM (252 see Suppl) | (252 see Suppl) | 40 | 40 |        | COAD (249 see Suppl) | (249 see Suppl) | 40 | 40 |
Model parameters correlates with background mutation rates

The estimation of background mutation rates is important for mutation detection methods because it helps to determine deviations [42]. Instead of the expected number of mutations, the fitted $\text{beta-binomial}$ model can be used to provide estimations of the probability of $k$ mutations along chromosomes. By definition, contiguous genes should show similar probabilities even that the fitting was independent. Small deviations of an overall probability should highlight important genes and systematic deviations should show artifactual genes or regions. To validate this, the estimated p-values were compared between genes along chromosomes. The Figure 7 shows a representative example of the estimations for the chromosome 1 (Supplementary Figure 4 shows all chromosomes) for the p-value of 0 and 1 mutations (shown in black and red respectively). It is clear that the smoothed mean show some peaks that colocalize with olfactory receptors (vertical gray lines), which has been shown to be highly correlated to late replication timing, low expression, and higher mutation rates [35]. Other gene clusters can be identified, for example, late cornified envelope (LCE) gene cluster in Chr1 (Figure 7), regenerating family member (REG) in Chr2, protocadherin beta gene cluster (PCDHB) in Chr5, the histone 1 cluster in Chr6, among others (Supplementary Figure 5). Specific deviations such as CDKN2A in Chr9, PTEN in Chr10, TP53 in Chr17 among other are also visible (Supplementary Figure 5). These results show that the proposed algorithm provides consistent estimations.

Discussion

This manuscript shows an algorithm to identify highly recurrent mutations at specific amino acid positions in cancer. The algorithm fits the distribution of amino acid positions along number of mutations using a mixed model that includes a beta-binomial model plus a fixed effect (Figure 1).

The comparisons of different distributions lead to select the $\text{beta-binomial}$ model. This makes sense because, in principle, the mutation can be seen as a binomial process during replication
and/or repair. Then, instead of fixing $p$ along the gene in the binomial process, $p$ is random drawn from a \textit{beta} distribution, which absorbs uncertainty due to patient, different positions, and sequence contexts resulting in allowing more uncertainty, covering observed overdispersion, and fitting the data better. Other statistical models could be tested but the justification, the interpretation, and the adequacy of the model may be difficult.

One of the problems when proposing a predicting or discovery algorithm is how assessing the accuracy. Although other algorithms and models have been proposed, most of them use lists of positive and/or negative curated genes as benchmarking. Instead, simulations were used here showing that overall the sensitivity and specificity was $\sim 85\%$. More importantly, in conditions common for hotspots such as at highest number of mutations, the algorithm shows accuracies around 99\%.

Few genes such as TP53, PIK3CA, and PTEN, showed a different tendency in fitting than most genes (Supplementary Figure 3). This is presumably due to the high number of hotspots and mutations backed up by the observation that closer genes such as CDKN2A, GATA3, and APC also show high hotspots. This is not a problem because these are well-known cancer genes. Nevertheless, it would be interesting to observe other genes once more mutation data is aggregated in the coming years.

It is assumed that a hotspot have functional impact in cancer [1]. Nevertheless, recent advances have shown that many hotspots arise by artifacts in local sequences such as hairpins susceptible for APOBEC enzymatic activity [40], including the detected gene MB21D2. Therefore, it is difficult to confirm in advance which hotspots will be functional. However, the first step is to detect those that under a certain model seems to be potential hotspots. These hotspots are provided here. Thus, how hotspots must be selected for functional validation? First, those that are well-known cancer genes whose hotspot have not been experimentally tested. Second, the genes showing many hotspots or high number of mutations at the hotspot. These would provide further certainty that any of its hotspots are indeed functional.
Nevertheless, in the analysis of cancer data, most genes only show 1 hotspot and most hotspots were found supported by less than 10 mutations (Figure 4). Third, check that the gene has not been listed for APOBEC activity [40]. In this context, the database HotSpotsAnnotations has been created (http://bioinformatica.mty.itesm.mx:8080/HotSpotsAnnotations) which has been annotated for APOBEC, the ratio of non-synonymous by synonymous mutations, and can be manually annotated by the research community [43]. Fourth, further verification is needed if the gene is super-sized or within artifactual regions such as those around olfactory receptors. Fifth, check the criteria of the ratio of non-synonymous to synonymous mutations [15]. Finally, frame shifts deletions and insertions have not been well studied in the hotspot context and in statistical models. Around one third of the detected hotspots included these mutation types.

The observation that TC\textsubscript{G} is more prone to form hotspots does not seem to be due to the lack of covariates in the model used. This is based on the fact that sequence context in hotspots were analyzed after normalization by percentage comparing the observed mutational spectra and the hotspots. That is, if all mutational contexts would have similar probability of being established as a hotspot, similar percentages would be observed in hotspots. Instead, more than two-fold was observed in TC\textsubscript{G} for single nucleotide variants.

Most hotspots carry between 5 to 9 mutations (70%) and also are formed by mutations of different cancer types (91%). Therefore, many hotspots were only detected when mutation from all cancer types were aggregated highlighting the importance of integrating databases. Consequently, as more mutation data is accumulated, more precise detections can be done. One issue is that all datasets must be processed in compatible pipelines, genome annotations, and transcripts to avoid inconsistencies. In this context, other databases such as those from the International Cancer Genome Consortium (ICGC) should improve and confirm the results.
Conclusion

Simulations of the proposed algorithm that fit a mixed model of beta-binomial plus a fixed effect demonstrated excellent performance for hotspots at highest mutations (around 99% accuracy) and acceptable overall performance (85%). The algorithm was applied to TCGA cancer data detecting more than 3,860 hotspots after FDR correction that account for around 1.25% of the total number of mutations and 0.19% of the mutated amino acid sites.

Acknowledgements

I thank Dr. Jose Tamez, Dr. Emmanuel Martinez, and all participants in the Bioinformatics seminar for their comments and recommendations.

Conflict of Interest

The author declares none conflict of interest.

Figure Captions

**Figure 1. Hotspot concept and proposed algorithm.** Panel A Cartoon conceptualization of random mutations along a protein (top) and similar number of mutations resulting in two hotspots (bottom). Histogram of sites per number of mutations is shown. Count are missing for clarity. Panel B shown the proposed algorithm to find the optimal parameters of the mixed model that better fit the observed mutation distribution.

**Figure 2. Performance of the proposed algorithm on simulated data.** Panel A shows the distribution of the number of detected hotspots in simulation 1, which does not contain hotspots. Panel B shows how hotspots were injected. The left histogram shows original data. The middle and right histogram show the result of adding hotspots having 3 or 5 mutations respectively. The number of mutations is relative to the maximum number of observed mutations. In the example at left, the maximum is 4. Thus, $r_{Mut}=-1$ refers to 1 mutation lower than 4 while $r_{Mut}=+1$ refers to 1 mutation higher than 4. Finally, the $nPos$ is the number of amino acid positions added to the specified number of mutations. Panel C shows the overall results of all simulations having hotspots. Here ‘detected hotspot’ stand for the sum of values > 0 from fitted F vectors. Panel D shows the performance of the algorithm depending on $r_{Mut}$.
Each combination shows the percentage of simulated genes that showed the corresponding hotspots at the relative number of mutations.

Figure 3. Examples of hotspots detections. Three examples of hotspots detections from TCGA data. The left drawings show the mutations along the protein sequence of three genes. Colors correspond to different cancer types. Symbols correspond to different types of mutations. The right histograms show the corresponding amino acid positions (vertical, in logarithmic scale) per number of mutations (horizontal). The beta-binomial component is represented in light bar colors and dotted line. The fixed effect is represented by darker bar colors. Significant hotspots are marked. Non-Significant fixed effects are also shown. Figures taken from http://bioinformatica.mty.itesm.mx/HotSpotsAnnotations developed in our research group.

Figure 4. Distribution of hotspots per gene and mutations.

Figure 5. Comparison of mutated context sequences in hotspots. The left heatmap show the relative percentage of mutated positions per mutation type found in the whole dataset of TCGA data used. Only mutations types found in hotspots of 10 or more mutations are shown. Only distinct sites are considered. Total positions (N), are shown in thousands (k=1000). The heatmap at the middle shows equivalent percentages found at hotspots positions carrying 10 or more mutations. To facilitate interpretation, the heatmap at the right show the difference of the percentages. Fold-changes may vary. The Supplementary Figure 4 show details of other mutations types and for hotspots of 5 to 9 mutations.

Figure 6. Distribution of hotspots across cancer types. Panel A shows the percentage of samples containing k hotspots. Panel B shows the different types of cancer per hotspot. Panel C shows the 95 hotspots found at only one cancer type. Supplementary Table 2 shows the genes that are strict cancer type-specific. Panel D shows the hotspots that are majorly represented by one cancer type. Supplementary Table 3 shows the genes that are enriched by cancer type.

Figure 7. Model estimations along chromosome 1. The figure shows the density estimations of 0 mutations (dots in black) and 1 mutation (dots in red). The red line in top and black line in bottom show the smoothed estimation (window=5). The mean value, 0.818 for the former and 0.151 for the last, is shown at right and represented by a horizontal gray line. Vertical gray lines represent genomic positions for annotated olfactory receptors. Some genes farther than 3 standard deviations are annotated. Supplementary Figure 5 shows equivalent information for all chromosomes.

Supplementary Material

Supplementary Figure 1. Fitting characteristics in TCGA data. The figure shows the distribution of the fitted fixed effect F (horizontal) relative to the total number of mutations (vertical). A
value of 1 represent 100% of the mutations. It is clear that most fittings absorb 100% of mutations (highest vertical bar at top). Nevertheless, few fittings are fractions of the total number of mutations, mainly at low number of mutations (5 or lower). Therefore, these fittings capture biases in the distribution relative to the beta-binomial model.

**Supplementary Figure 2. Comparisons of best fitting among four distributions.** Panel A shows a comparison of the best distribution for different sets of genes. Panel B compares the best two distributions. Panel C shows results if one distribution is not considered.

**Supplementary Figure 3. FDR estimation for amino acid positions carrying 4 or more mutations.** Two tendencies are observed, a major below ~20 mutations and a minor above ~20. The last includes TP53, PIK3CA, and PTEN hotspots mainly as labelled. The vertical nMut was jittered for clarity in density estimation. FDR q-value was limited to 10^{-10} (at left) for clarity.

**Supplementary Figure 4. Comparison of mutated context sequences in hotspots.** The first two heatmaps show the relative percentage of mutated positions per mutation type found in the whole dataset of TCGA data used. The first show the types of mutations not found in hotspots of 10 mutations or more. The second show the types of mutations found in hotspots of 10 or more mutations. Only distinct sites are considered. Total positions (N), are shown in thousands (k=1000). The third heatmap shows equivalent percentages found at hotspots positions carrying 10 or more mutations. The last heatmap at right show equivalent percentages for hotspots carrying 5 to 9 mutations.

**Supplementary Figure 5. Model estimations along chromosomes.** Each panel shows the density estimations of 0 mutations (dots in black) and 1 mutation (dots in red) for a chromosome (as labeled at right). The red line in top and black line in bottom show the smoothed estimation (window=5). The mean value is shown at right and represented by a horizontal gray line. Vertical gray lines represent genomic positions for annotated olfactory receptors. Some genes farther than 3 standard deviations are annotated.

**Supplementary Table 1. Number of samples mutations per cancer type used in this study.**

**Supplementary Table 2. Genes strictly cancer type-specific.**

**Supplementary Table 3. Genes enriched by cancer type-specific.**
References

[1] Miller ML, Reznik E, Gauthier NP, Ciriello G, Schultz N, Miller ML, et al. Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. Cell Syst 2015;1:197–209. doi:10.1016/j.cels.2015.08.014.

[2] Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature 2002;417:949–54. doi:10.1038/nature00766.

[3] Tiacci E, Trifonov V, Schiavoni G, Holmes A, Kern W, Martelli MP, et al. BRAF Mutations in Hairy-Cell Leukemia. N Engl J Med 2011;364:2305–15. doi:10.1056/NEJMoa1014209.

[4] Cancer T, Atlas G, Agrawal N, Akbani R, Aksoy BA, Ally A, et al. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. Cell 2014;159:676–90. doi:10.1016/j.cell.2014.09.050.

[5] Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat J-P, et al. A Landscape of Driver Mutations in Melanoma. Cell 2012;150:251–63. doi:10.1016/j.cell.2012.06.024.

[6] Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7. doi:10.1038/nature11252.

[7] Cancer T, Atlas G, (Network TCGAR, institution.) (Participants are arranged by area of contribution and then by, Institute G data analysis centres: B sequencing centres: B, Hammerman PS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519–25. doi:10.1038/nature11404.

[8] Salimian KJ, Fazeli R, Zheng G, Ettinger D, Maleki Z. V600E BRAF versus Non-V600E BRAF Mutated Lung Adenocarcinomas: Cytomorphology, Histology, Coexistence of Other Driver Mutations and Patient Characteristics. Acta Cytol 2018;62:79–84. doi:10.1159/000485497.

[9] Gliomas L. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med 2015;2481–98. doi:10.1056/NEJMoa1402121.

[10] Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543–50. doi:10.1038/nature13385.

[11] Akbani R, Akdemir KC, Aksoy BA, Albert M, Ally A, Amin SB, et al. Genomic Classification of Cutaneous Melanoma. Cell 2015;161:1681–96. doi:10.1016/j.cell.2015.05.044.

[12] Chang MT, Bhattarai TS, Schram AM, Bielski CM, Donoghue TA, Jonsson P, et al. Accelerating discovery of functional mutant alleles in cancer. Cancer Discov 2017;8:174–83. doi:10.1158/2159-8290.CD-17-0321.

[13] Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 2015;34:155–63. doi:10.1038/nbt.3391.

[14] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 2019;47:D941–7. doi:10.1093/nar/gky1015.
[15] Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 2017;171:1029-1041.e21. doi:10.1016/j.cell.2017.09.042.

[16] Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 2013;29:2238–44. doi:10.1093/bioinformatics/btt395.

[17] Jia P, Wang Q, Chen Q, Hutchinson KE, Pao W, Zhao Z. MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. Genome Biol 2014;15:489. doi:10.1186/s13059-014-0489-9.

[18] Baeissa H, Benstead-hume G, Richardson CJ, Pearl MG. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors 2017;8:21290–304.

[19] Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure 2016:3719–32. doi:10.1158/0008-5472.CAN-15-3190.

[20] Gao J, Chang MT, Johnsen HC, Gao SP, Sylvestre BE, Sumer SO, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med 2017;9:4. doi:10.1186/s13073-016-0393-x.

[21] Chen T, Wang Z, Zhou W, Chong Z, Meric-bernstam F, Mills GB, et al. Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types. BMC Genomics 2016;17. doi:10.1186/s12864-016-2727-x.

[22] Munro D, Ghersi D, Singh M. Two critical positions in zinc finger domains are heavily mutated in three human cancer types 2018:1–17.

[23] Juul M, Bertl J, Guo Q, Nielsen MM, Świtnicki M, Hornshøj H, et al. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. Elife 2017;6. doi:10.7554/eLife.21778.

[24] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S a JR, Behjati S, Biankin A V, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21. doi:10.1038/nature12477.

[25] Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: An integrated framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res 2015;43:8123–34. doi:10.1093/nar/gkv803.

[26] Hess JM, Bernards A, Kim J, Haradhvala NJ, Lawrence MS, Getz G, et al. Passenger Hotspot Mutations in Cancer. Cancer Cell 2019:288–301.
[31] Kucab JE, Zou X, Morganella S, Arlt VM, Phillips DH, Nik-zainal S, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell 2019:1–16. doi:10.1016/j.cell.2019.03.001.

[32] Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci 2016;113:14330–5. doi:10.1073/pnas.1616440113.

[33] Hu T, Gallins P, Zhou Y-H. A zero-inflated beta-binomial model for microbiome data analysis. Stat 2018;7:e185. doi:10.1002/sta4.185.

[34] Hinde J, Demtrio CGB. Overdispersion: Models and estimation 1998;27:151–70.

[35] Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8. doi:10.1038/nature12213.

[36] Vandin F. Computational methods for characterizing cancer mutational heterogeneity. Front Genet 2017;8:1–12. doi:10.3389/fgene.2017.00083.

[37] Rheinbay E, Parasuraman P, Grimsby J, Tiao G, Engreitz JM, Kim J, et al. Recurrent and functional regulatory mutations in breast cancer. Nature 2017. doi:10.1038/nature22992.

[38] Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 2012;28:2747–54. doi:10.1093/bioinformatics/bts526.

[39] Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9:811–8.

[40] Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science 2019;364:eaaw2872. doi:10.1126/science.aaw2872.

[41] Zhang R, Zhang F, Sun Z, Liu P, Zhang X, Ye Y, et al. LINE-1 Retrotransposition Promotes the Development and Progression of Lung Squamous Cell Carcinoma by Disrupting the Tumor-Suppressor Gene FGGY. Cancer Res 2019;79:4453–65. doi:10.1158/0008-5472.CAN-19-0076.

[42] Jiang L, Zheng J, Kwan JSH, Dai S, Li C, Li MJ, et al. WITER: a powerful method for estimation of cancer-driver genes using a weighted iterative regression modelling background mutation counts. Nucleic Acids Res 2019;47. doi:10.1093/nar/gkz566.

[43] Trevino V. HotSpotAnnotations - A database for hotspot mutations and annotations in cancer. Database 2019:(In revision).
**Figure 1**

\[ M \approx (BB+F) \]

**Input:** Positions per mutation \((M_k)\)

**Estimation of** \(f_i\), the best possible fixed effect given \(F_{i-1}\)

**Aggregate & re-iterate**

**Output:**
- Aggregated \(F\)
- Updated \(\alpha, \beta\)
A

Simulation 1 (no hotspots)

1663
84.29%

1163
84.29%

Detected Hotspots

B

Performance of Simulation 2 (>=1 hotspot)

0.3% 100% 0.1% 100% 0.1% 100%

0.7% 99% 0.1% 100% 0.5% 100%

1.4% 99% 0.1% 100% 0.8% 99%

25% 76% 0.3% 100% 13% 87% 0.1%

3% 81% 16% 78% 23% 82% 16%

32% 50% 19% 83% 21% 73% 22% 0.1%

62% 21% 25% 67% 30% 64% 29% 2.1% 0.2%

Recognized Hotspots

(%) Relative to 1973 simulations, ideally, at columns 1, 3, and 5)

C

Simulation 2 (>=1 hotspot)

29743
71.82%

6331
15.29%

Detected Hotspots

D

Injected AA Positions (nHot)

- 1

- 3

- 5
**Figure 3**

- **EGFR**
- **NBPF12**
- **GK2**

Mutations and AA Positions are plotted against each other for each protein. The graphs show the distribution of mutations at different AA positions, with notable hotspots highlighted.

- **EGFR** has mutations at positions 7, 11, 19, 23, and 27.
- **NBPF12** has a mutation at position 12.
- **GK2** has mutations at positions 6, 6, 6, 5, and 5.

Additionally, there is a bar chart indicating the significance of these hotspots after FDR correction for multiple comparisons.
Figure 4

A

Hotspots per Gene

2639 378 64 19 KMT2D, CIC, DNAH5
3 3 ERBB2, CTNNB1, BRAF
1 3 ARID1A, FBXW7, NFE2L2
1 1 CDKN2A, PTEN
1 1 APC
1 1 PIK3CA
1 1 TP53

B

Hotspots per number of mutations

471 683 700 480 383 268 153 100 96 57 49 49 29 27 26 218

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20+
### % All Mutated Positions

| TCN | (N in k) | 565 | 1444 | 116 | 7 | 29 | 2 | 6 | 37 | 36 |
|-----|----------|-----|------|-----|---|----|---|---|----|----|
| Silent | 0.9 | 2.5 | 0.9 | 0.5 | 2.9 | 1.5 | 0 | 5.2 | 2.6 | 0.3 |
| Missense | 1.2 | 0.9 | 1.6 | 6.2 | 2.8 | 1.1 | 0 | 3.8 | 2.5 | 0.2 |
| Nonsense | 0.8 | 1.4 | 1.1 | 7.7 | 2.1 | 1.1 | 0 | 1.7 | 0.8 | 0.2 |
| Frame Shift Del | 0.8 | 0.6 | 1.7 | 4.5 | 2.1 | 5.0 | 0 | 2.6 | 2.9 | 0.2 |
| In Frame Del | 1.3 | 7.9 | 40.3 | 15.4 | 7.6 | 6.5 | 4.6 | 0 | 4.6 | 67.1 |
| Splice Region | 1.7 | 7.2 | 6.4 | 9.8 | 1.5 | 3.5 | 0 | 0.9 | 1.2 | 0 |
| Trans Start | 7.2 | 7.7 | 7.5 | 7.4 | 5.4 | 3.4 | 0 | 3.7 | 2.5 | 0.2 |
| In Frame Ins | 8.6 | 5.8 | 19.1 | 8.5 | 4.5 | 17 | 23.6 | 3.9 | 5.2 | 0.2 |
| Frame Shift Ins | 12.1 | 1.6 | 2.2 | 0.4 | 7.4 | 1.7 | 4.5 | 2.6 | 1.1 |
| Splice Site | 2.9 | 1.7 | 2.8 | 0.7 | 5.1 | 0.2 | 1.7 | 2.6 | 4.1 |

### % Hotspots Positions

| TCN | (N) | 143 | 750 | 171 | 742 | 29 | 7 | 2 | 327 | 59 |
|-----|-----|-----|-----|-----|-----|----|---|---|-----|-----|
| Silent | 0.9 | 2.1 | 2.7 | 0.3 | 6.9 | 0 | 0 | 0 | 0.3 | 0 |
| Missense | 0.7 | 0.3 | 0.6 | 9.3 | 3.4 | 0 | 0 | 0 | 0.0 | 0 |
| Nonsense | 4.9 | 3.6 | 0 | 8.4 | 6.9 | 0 | 0 | 0 | 0.0 | 0 |
| Frame Shift Del | 1.4 | 0.8 | 4.1 | 9.2 | 6.9 | 0 | 0 | 0 | 0.0 | 0 |
| In Frame Del | 1.4 | 0.8 | 4.1 | 9.2 | 6.9 | 0 | 0 | 0 | 0.0 | 0 |
| Splice Region | 0 | 4.2 | 9.5 | 0 | 4.2 | 9.5 | 0 | 0 | 0.0 | 0 |
| Trans Start | 23.8 | 19.5 | 40.9 | 0.9 | 6.9 | 0 | 0 | 0 | 3.4 | 0 |
| In Frame Ins | 12.6 | 6.8 | 4.7 | 6.6 | 0 | 0 | 0 | 5.2 | 2.1 | 0 |
| Frame Shift Ins | 2.8 | 1.9 | 4.7 | 0.3 | 6.9 | 14.3 | 0 | 5.2 | 0.3 | 4.7 |
| Splice Site | 1.7 | 2.8 | 0.7 | 5.0 | 0 | 14.3 | 0 | 0 | 0 | 12.0 |

### Difference in %

| Difference in % | TCN | TTT | TTN | TTA | TCT | TCG | TCC | TCA |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Silent | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Missense | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Nonsense | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Frame Shift Del | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| In Frame Del | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Splice Region | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Trans Start | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| In Frame Ins | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Frame Shift Ins | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |
| Splice Site | 0 | 0 | -1 | 0 | -1 | 0 | 0 | 0 |

*Note: The table shows the percentage of mutated positions for different classes (Silent, Missense, Nonsense, Frame Shift Del, In Frame Del, Splice Region, Trans Start, In Frame Ins, Frame Shift Ins, Splice Site) across different points (N in k). The difference in percentage shows the variation between Low and High values.*
Figure 6

A) Samples

0% to 80% of samples with hotspot mutations

B) Types of cancer per hotspot

- UVM
- THCA
- THYM
- OV
- GBM
- BRCA
- CESC
- LAML
- DLBC
- LIHC
- HCC
- ERCC2, KDM6A, SF3B1
- CDH1
- MXD3, HMX3, APC
- VHL
- NPM1, FLT3
- EGFR
- SPOP
- ITGB1

C) ‘Strict’ Cancer Type Specific

D) Hotspots ‘enriched’ by cancer type

- UCEC
- STAD
- SKCM
- COAD
- BLCA
- LUSC
- BRCA
- HNSC
- OV
- LAML
- LG
- LLAD
- GBM
- LIHC
- THCA
- PRAD
- UVM
- CESC
- KIRC
- KIRP
- PCPG
- KICH
- ACC
Supplementary Figure 3
| Nonstop | Silent | Missense | Nonsense | Frame Shift Del | In Frame Del | Splice Region | Trans Start | In Frame Ins | Frame Shift Ins | Splice Site | Nonstop | Silent | Missense | Nonsense | Frame Shift Del | In Frame Del | Splice Region | Trans Start | In Frame Ins | Frame Shift Ins | Splice Site |
|---------|--------|----------|----------|---------------|-------------|--------------|-------------|-------------|----------------|-------------|---------|--------|----------|----------|---------------|-------------|--------------|-------------|-------------|----------------|------------|
|         |        |          |          |               |             |              |             |             |                |             |         |        |          |          |               |             |              |             |             |                |           |
| 1.7     | 15     | 22       | 63       | 178           | 139         | 192          | 109         | 307         | 33             | 90          | 60      | 52     | 6        | 1        |               |             |              |             |             |                |           |
| 2.1     | 18     | 94       | 80       | 359           | 356         | 528          | 257         | 777         | 522           | 234          | 183     | 170    | 136      | 106      |               |             |              |             |             |                |           |
| 2.7     | 41     | 73       | 35       | 707           | 704         | 1031         | 527         | 1562        | 431           | 212          | 143     | 128    | 106      | 76       |               |             |              |             |             |                |           |
| 5.1     | 60     | 64       | 26       | 755           | 752         | 1159         | 562         | 1692        | 372           | 176          | 122     | 109    | 91       | 66       |               |             |              |             |             |                |           |
| 28.7    | 47     | 40       | 10       | 767           | 764         | 1181         | 564         | 1723        | 463           | 220          | 152     | 138    | 113      | 87       |               |             |              |             |             |                |           |
| 1.7     | 18     | 94       | 80       | 359           | 356         | 528          | 257         | 777         | 522           | 234          | 183     | 170    | 136      | 106      |               |             |              |             |             |                |           |
| 2.1     | 18     | 94       | 80       | 359           | 356         | 528          | 257         | 777         | 522           | 234          | 183     | 170    | 136      | 106      |               |             |              |             |             |                |           |
| 2.7     | 41     | 73       | 35       | 707           | 704         | 1031         | 527         | 1562        | 431           | 212          | 143     | 128    | 106      | 76       |               |             |              |             |             |                |           |
| 5.1     | 60     | 64       | 26       | 755           | 752         | 1159         | 562         | 1692        | 372           | 176          | 122     | 109    | 91       | 66       |               |             |              |             |             |                |           |
| 28.7    | 47     | 40       | 10       | 767           | 764         | 1181         | 564         | 1723        | 463           | 220          | 152     | 138    | 113      | 87       |               |             |              |             |             |                |           |

**Supplementary Figure 4**

| (N in k) | 17 | 19 | 23 | 24 | 25 | 26 | 30 |
|----------|----|----|----|----|----|----|----|
| 1.7      | 15 | 22 | 63 | 178| 139| 192| 109|
| 2.1      | 18 | 94 | 80 | 359| 356| 528| 257|
| 2.7      | 41 | 73 | 35 | 707| 704| 1031| 527|
| 5.1      | 60 | 64 | 26 | 755| 752| 1159| 562|
| 28.7     | 47 | 40 | 10 | 767| 764| 1181| 564|

**% All Mutated Positions**

| (N in k) | 14 | 25 | 28 | 33 | 41 | 49 | 57 | 61 | 66 | 72 | 74 |
|----------|----|----|----|----|----|----|----|----|----|----|----|
| 1.7      | 15 | 22 | 63 | 178| 139| 192| 109| 307| 33 | 90 | 60 |
| 2.1      | 18 | 94 | 80 | 359| 356| 528| 257| 777| 522| 234| 183|
| 2.7      | 41 | 73 | 35 | 707| 704| 1031| 527| 1562| 431| 212| 143|
| 5.1      | 60 | 64 | 26 | 755| 752| 1159| 562| 1692| 372| 176| 122|
| 28.7     | 47 | 40 | 10 | 767| 764| 1181| 564| 1723| 463| 220| 152|

**% Hotspots (nMut >= 10)**

| (N in k) | 17 | 17 | 23 | 24 | 26 | 30 | 57 | 61 | 66 | 72 | 74 |
|----------|----|----|----|----|----|----|----|----|----|----|----|
| 1.7      | 15 | 22 | 63 | 178| 139| 192| 109| 307| 33 | 90 | 60 |
| 2.1      | 18 | 94 | 80 | 359| 356| 528| 257| 777| 522| 234| 183|
| 2.7      | 41 | 73 | 35 | 707| 704| 1031| 527| 1562| 431| 212| 143|
| 5.1      | 60 | 64 | 26 | 755| 752| 1159| 562| 1692| 372| 176| 122|
| 28.7     | 47 | 40 | 10 | 767| 764| 1181| 564| 1723| 463| 220| 152|

**% Hotspots (nMut 5 ~ 9)**

| (N in k) | 17 | 17 | 23 | 24 | 26 | 30 | 57 | 61 | 66 | 72 | 74 |
|----------|----|----|----|----|----|----|----|----|----|----|----|
| 1.7      | 15 | 22 | 63 | 178| 139| 192| 109| 307| 33 | 90 | 60 |
| 2.1      | 18 | 94 | 80 | 359| 356| 528| 257| 777| 522| 234| 183|
| 2.7      | 41 | 73 | 35 | 707| 704| 1031| 527| 1562| 431| 212| 143|
| 5.1      | 60 | 64 | 26 | 755| 752| 1159| 562| 1692| 372| 176| 122|
| 28.7     | 47 | 40 | 10 | 767| 764| 1181| 564| 1723| 463| 220| 152|

*Note: All values are percentages.*
