# Estimating a gene's mutation burden by the number of observed synonymous base substitutions

Perry Evans and Michael Krauthammer
*Department of Pathology, Yale University School of Medicine*
*New Haven, CT, USA*
*Email: perry.evans@yale.edu*

***Abstract*—A common goal of tumor sequencing projects is the identification of genes whose mutations are selected for during tumor development. This is accomplished by finding genes that have more nonsynonymous mutations than expected by an estimated background mutation frequency. While this frequency is unknown, it can be estimated using both the observed synonymous mutation frequency, and the nonsynonymous to synonymous mutation ratio. The synonymous mutation frequency can be determined across all genes, or in a gene-specific manner. This choice introduces an interesting trade-off. A gene-specific frequency is difficult to estimate given small or missing synonymous mutation counts, but adjusts for an underlying mutation load bias. Using a genome-wide synonymous frequency is more robust, but is less suited for adjusting for the same bias. Studying three evaluation criteria for identifying genes with high nonsynonymous mutation burden (preferential selection of expressed genes, genes with mutations in conserved bases, and genes that show loss of heterozygosity), we find that the gene-specific synonymous frequency is superior in the gene expression and conservation tests, while both frequencies perform similarly for the loss of heterozygosity test. In conclusion, we believe that the use of the gene-specific synonymous mutation frequency is well suited for estimating a gene's nonsynonymous mutation burden.**

***Keywords*-cancer; sequencing**

## I. Introduction

Today's genome-wide and exome-wide cancer sequencing projects annotate mutations across all genes in tumors from a large cohort of patients. Genes with more nonsynonymous mutations than expected by chance are hypothesized to be important for tumors, and therefore important for the understanding of tumor biology and the development of anti-cancer drugs. The process of finding these genes with high mutation burden is confounded by mutation biases that cause genes to have more or less mutations than other genes, while not affecting a gene's importance to tumors. These biases include genomic sequence context, gene expression, distance from the transcription start site, and the timing of gene replication [1], [2]. The factors mentioned here explain less than 40% of the mutation variation observed across genes [1], indicating that there are still other factors to be discovered.

Given this dilemma, it is impossible to model the expected frequency of nonsynonymous mutations using the biases themselves. However, synonymous mutations are affected by the same biases that influence nonsynonymous mutations, making it possible to use synonymous mutations as a proxy for bias influence. Given the nonsynonymous to synonymous mutation (NS:SN) ratio, the expected nonsynonymous mutations can be estimated by counting the synonymous ones. This approach was used to evaluate the mutation burden for 623 genes from 188 lung adenocarcinomas [3]. In the analysis, the synonymous mutation frequency was determined in two ways. First, in a global approach, the frequency was found across all genes, and every gene's expected nonsynonymous mutation count was derived from this constant frequency. Second, in a gene-specific approach, the synonymous frequency was found separately for each gene, allowing each gene's expected nonsynonymous mutation count to be estimated using the gene's synonymous mutations. The global approach helps estimate frequencies for genes with under-sampled synonymous mutations, but it fails to capture individual gene biases. The gene-specific approach does account for these biases, but becomes problematic when a gene has no synonymous mutations. These two measures of nonsynonymous frequencies, global and gene-specific, have never been compared to see which yields better lists of genes with significantly high nonsynonymous mutations.

In this paper, we make two contributions to the discussion of gene mutation bias and burden by assessing somatic mutations from exome-wide sequencing of tumors from a cohort of 99 patients with melanoma. First, we expand upon the mutation biases associated with melanoma. It has been shown that mutations in a single melanoma sample were influenced by expression and replication timing [1]. We confirm these results using our much larger panel of somatic mutations. We show that unexpressed genes accumulate more mutations than expressed genes, and that late replicating genes have more mutations than early replicating genes. Second, we address the global and gene-specific use of synonymous mutations as a proxy for mutation bias that helps find genes that drive cancer. We use three criteria to evaluate genes deemed to have more nonsynonymous mutations than expected by the two frequencies. We consider a gene a good cancer driver candidate if it is expressed

978-1-4673-2560-8/12/$31.00 

in melanoma, has mutations at conserved positions, and shows loss of heterozygosity (LOH). Based on our criteria, we conclude that the gene-specific synonymous mutation frequency yields the best list of candidate driver genes.

## II. Results and Discussion

For this study, we used novel somatic mutations found by comparing exome-wide sequencing data from paired melanoma and normal skin or blood samples taken from 99 patients [4]. We searched for novel somatic mutations by ignoring variants catalogued in normal samples and the variant databases dbSNPv135 [5] and 1000 Genomes [6]. For each sample, we surveyed roughly 22 megabases covering about 15000 genes. Somatic nonsynonymous mutations in each sample ranged from zero to nearly two thousand. Somatic synonymous mutations in each sample ranged from zero to just over one thousand.

### A. Mutation bias

Here we consider the mutation bias contributors gene expression and replication timing. We show how both factors contribute to mutation bias by showing their correlation with synonymous mutation counts.

*1) Gene expression bias:* Expressed and unexpressed genes were determined using RNA-Seq data from two normal melanocyte samples. Of the genes with synonymous mutations, 3,281 genes were found to be expressed in normal melanocytes, while 3,224 were not. As seen in Figure 1, expressed genes were found to have less synonymous mutations than unexpressed genes (t-test p-value $<$ 5e-11). This mutation bias is attributed to transcription coupled repair, where mutations are corrected as genes are transcribed.

*2) Replication timing bias:* During S phase of the cell cycle, genes are replicated at different stages. Replication timing is believed to have an affect on mutation rate [7], with more variation appearing in late replicating genes than early ones. We gathered replication timing data from twelve cell types [8], and found genes that were consistently replicated late or early across all cell types. In total we had 2,081 early replicating genes and 258 late replicating genes with synonymous mutations. We found that late replicating genes had more synonymous mutations than early replicating genes (Wilcoxon test p-value $<$ 3e-16, Figure 2)

### B. Using synonymous mutations as a proxy for mutation bias

Having demonstrated that synonymous mutations are influenced by mutation biases, we now survey different methods for incorporating synonymous mutation counts into the gene burden mutation model to address these biases. We limit our analysis to 61 sun-exposed tumors because sun-shielded tumors have few mutations, and might have a different tumor biology.

In the assessment of nonsynonymous mutation burden for a gene, a binomial test is used to gauge the significance of

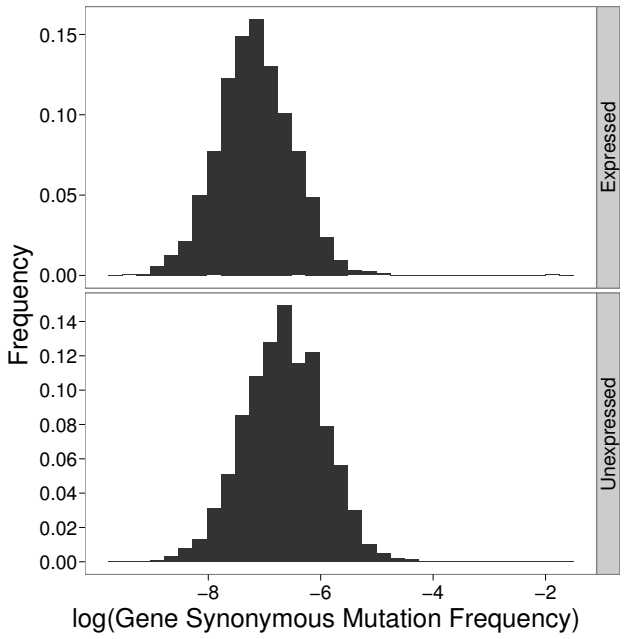


Figure 1. Gene expression mutation bias: Unexpressed genes have significantly more synonymous mutations than expressed genes. Genes with somatic synonymous mutations were classified into expressed and unexpressed gene sets of roughly 3,000 genes. For each gene, the somatic synonymous mutation frequency was determined by dividing a gene's total somatic synonymous mutations by the gene's coding length. Histograms of the log of this mutation frequency were plotted for expressed and unexpressed genes to clearly show the relation between the gene sets. Note that a more negative log mutation frequency indicates a lower somatic synonymous mutation frequency.

a gene's nonsynonymous mutations given the gene's length, and expected the number of nonsynonymous mutations seen across the exome. The expected number of nonsynonymous mutations is found using the nonsynonymous:synonymous (NS:SN) mutation ratio, and a frequency of synonymous mutations. The NS:SN ratio is found by simulating mutations across the genome, and recording the NS:SN ratio observed in total, and for individual genes. The expected nonsynonymous mutations for a gene are then estimated using the frequency of synonymous mutations accumulated across all genes, or the frequency of synonymous mutations for the gene in question. We address the question of which frequency is better by introducing four alternative methods below, and evaluating their performance based on tumor gene expression, genes found in loss of heterozygosity (LOH) regions, and the conservation of amino acids involved in mutations.

The four approaches used for finding synonymous mutation frequencies include one gene-specific method, two variations on the global gene-wide frequencies, and a mixture of gene-specific and global frequencies. The *Gene* frequency uses a different synonymous mutation frequency for each gene. This frequency is calculated directly from the gene's

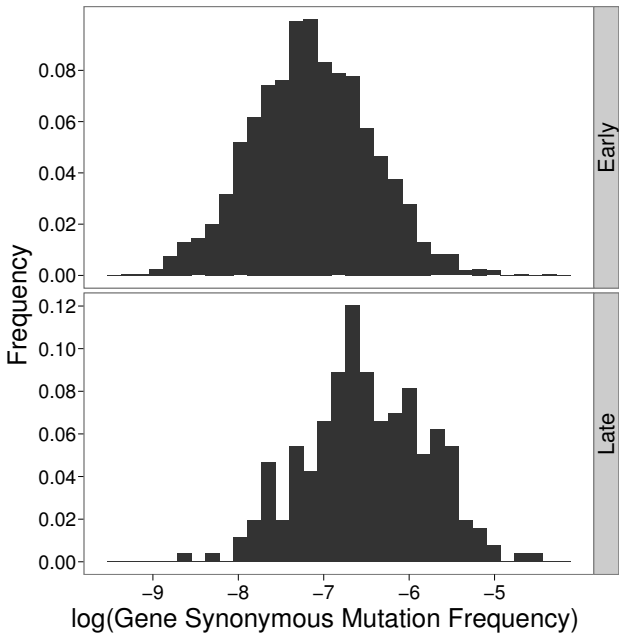


Figure 2. Replication timing mutation bias: Genes with consistent late replication across tissues have significantly more synonymous mutations than genes with consistent early replication. Genes with somatic synonymous mutations were classified into sets of 2,081 early replicating genes and 258 late replicating genes. For both of these groups we determined the frequency of somatic synonymous mutations found per gene. As in Figure 1, we plot histograms of the log of mutation frequency to better illustrate the synonymous mutation load difference between early and late replicating genes.

synonymous mutations. The *Global* frequency uses the same synonymous frequency for all genes. While developing the *Global* frequency, we noticed that some genes have more synonymous mutations than expected from the exome wide synonymous mutation frequency. These genes also tended to have higher nonsynonymous counts, which made them stand out as significant when using the global synonymous mutation frequency. To fix this problem, we divided the *Global* frequency into two frequency measures: *Global_All* and *Global_Exclude*. *Global_All* uses the synonymous mutation frequency across all genes, while *Global_Exclude* uses the synonymous frequency from genes whose synonymous mutation counts fall within the distribution from *Global_All*. The *Mix* frequency is a combination of the *Gene* and *Global* frequencies, introduced to fix the issue that some genes have more synonymous mutations than expected. In this method, a gene's synonymous mutation count is first compared to the overall distribution of synonymous mutations. If a gene has more synonymous mutations than expected, the synonymous correction comes from the gene's synonymous mutation count, as in the *Gene* frequency. Otherwise, the synonymous correction comes the *Global_Exclude* frequency. Table I lists the number of genes found to have significant somatic nonsynonymous mutation burden for each of the four methods.

Table I
THE NUMBER OF GENES WITH SIGNIFICANT SOMATIC NONSYNONYMOUS MUTATION BURDEN UNDER EACH METHOD.

| Method | Significant Genes |
|---|---|
| *Gene* | 17 |
| *Global_All* | 35 |
| *Global_Exclude* | 95 |
| *Mix* | 38 |

In the remainder of this section, we demonstrate that the *Gene* frequency performs best, according to three criteria. We find that using this frequency to determine genes with more nonsynonymous mutations than expected produces a gene list that has most genes expressed in tumors, indicating that the mutations in these genes are relevant to tumor development. This frequency also produces genes whose mutations are relatively more conserved. The genes in this list have a higher fraction of genes in LOH compared to the gene list produced by the *Global_Exclude* frequency, but the *Global_All* and *Mix* frequencies produce gene lists that have a higher fraction of genes in LOH.

*1) Gene expression evaluation:* One way to judge the quality of a candidate cancer driver gene list is to look at the fraction of genes that are expressed in tumors. Genes that are not expressed in tumors are unlikely to have mutations under selection, so gene lists with a higher fraction of expressed genes are more correct. We determined gene expression status based on microarray data from a panel of fifteen melanomas. The order of expression performance for the different synonymous mutation frequencies, from worst to best, is *Mix*, *Global_All*, *Global_Exclude*, *Gene*. Figure 3 compares the driver gene lists from the four frequencies using tumor gene expression. The best frequency, *Gene* uses each gene's synonymous mutation counts, and produces a list with 64% of the genes expressed across tumors. Both *Global* frequencies perform better than the *Mix* frequency. Of these *Global* frequencies, *Global_Exclude*, which uses the modified global synonymous mutation count, rather than the total used by *Global_All* improves the fraction of expressed genes (57% vs. 45%). However, *Global_Exclude* also produces a larger gene list, which may not be desired when experimental followups are performed. In last place, the *Mix* frequency only has 39% of the genes expressed.

*2) Loss of heterozygosity evaluation:* Our second comparison of candidate driver gene lists derived from the different frequencies uses LOH. For each candidate driver gene, we determined the fraction of nonsynonymous mutations where the mutation is homozygous, and the gene is in an LOH region for the sample. When most of a gene's mutations are in LOH, it is likely that the gene plays a role in cancer. Figure 4 compares the distributions of LOH mutation fractions for the results obtained using the four synonymous frequencies. While no frequency stands out as superior, the *Global_Exclude* frequency yields genes whose mutations are

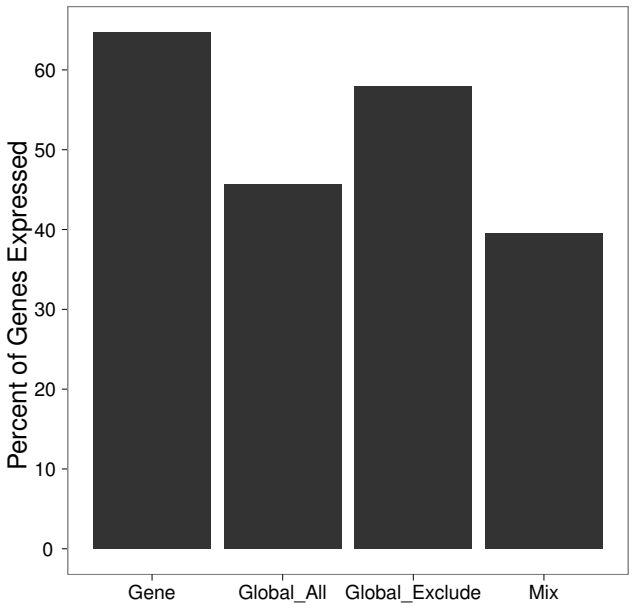


Figure 3. Expression evaluation. We found candidate cancer driver genes with more nonsynonymous mutations than expected according to each of the four synonymous mutation frequencies. The resulting gene lists were compared based on the fraction of genes that were expressed in melanomas. Gene lists with higher fractions of expressed genes were considered more correct.

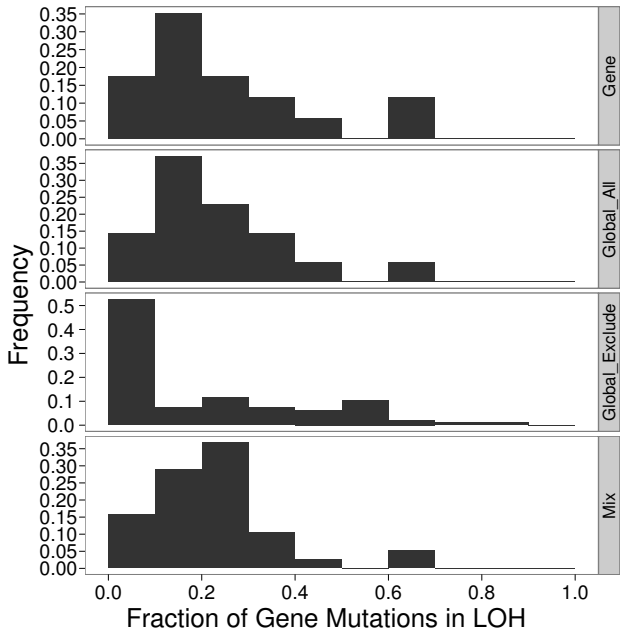


Figure 4. LOH evaluation. We found candidate cancer driver genes with more nonsynonymous mutations than expected according to each of the four synonymous mutation frequencies. The resulting gene lists were compared based on the fraction of a gene's nonsynonymous mutations that were in LOH across the samples. Genes lists with higher LOH fractions of were considered more correct. Here we show the frequencies of LOH fractions.

mostly not in LOH. The other three frequencies perform similarly, with the *Mix* frequency having a slight lead.

*3) Conservation evaluation:* For a final comparison of the synonymous mutation frequencies, we examined the conservation of the residues at each driver gene's nonsynonymous mutation positions. We assumed that a higher conservation of a residue indicated more importance to the protein, and took a gene with nonsynonymous mutations at mostly conserved residues to be more relevant to driving tumor development than a gene with little conservation as its nonsynonymous mutant residues. To evaluate mutant conservation for each gene, we compared the mean nonsynonymous mutation phyloP score to the mean phyloP score for the whole gene. Figure 5 shows the distribution of the log ratio of the mean mutation score to the mean score across the gene. The *Gene* frequency stands out as the best, with the most genes with positive ratios. As in the LOH evaluation, the *Global_Exclude* frequency performs the worst.

## III. Conclusion

To accurately find genes important to tumor development, we need robust models of mutation to find genes with more mutations than expected. Developing these models is difficult due to the mutation biases that are both tumor type specific and common across all cancers. Here we demonstrated that mutation bias can be captured by synonymous mutation

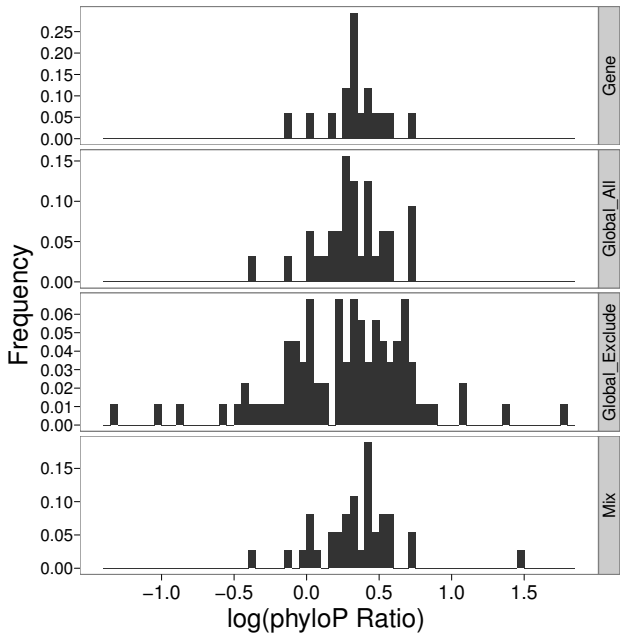


Figure 5. Conservation evaluation. We found candidate cancer driver genes with more nonsynonymous mutations than expected according to each of the four synonymous mutation frequencies. For each driver gene, we compared the mean of phyloP scores at nonsynonymous mutations to the mean phyloP score across the gene using the log ratio of the foreground nonsynonymous mutation mean to the gene background mean. Positive ratios indicate that the gene is more likely to important to cancer. Here we show the frequency of the phyloP log ratios.

frequencies, relieving the modeler from accounting for all types of mutation bias. We showed that finding the expected number of nonsynonymous mutations using synonymous mutation counts can be accomplished by using a gene-specific frequency, the same global frequency for all genes, or a mixture of both frequencies. We showed that the gene-specific method produces the best gene list by delivering more expressed genes whose nonsynonymous mutations are more likely to be conserved. We advocate the use of this gene-specific method when each gene has many synonymous mutations, but acknowledge that small sequencing projects will need to fall back on the global method when most genes are without synonymous mutations.

## IV. Methods and Data

### A. Melanoma somatic mutation calling

We used exome sequence data from 99 matched tumor/germline pairs for the automated calling of somatic mutations. For a full description of methods and data, see our previously published work [4]. Novel mutations found in tumors were classified as somatic and inherited variations as follows: we first called the tumor mutation, and then used sequencing data in a matched germline DNA sample to determine the presence or absence of variant reads at the same position. The mutation was called somatic in the absence of variant reads in the germline DNA samples, tolerating one mutant read in normal, and expecting a sufficient variant to total read ratio in tumor and normal as assessed by the Fishers exact test (p-value threshold of 0.001). Tumor mutations not classified as somatic, and called in normal were classified as inherited.

### B. Loss of heterozygosity

LOH regions were determined for each tumor/normal paired samples individually. First, heterozygous genomic positions in the normal sample were identified using the mutant allele frequency. A position whose mutant allele frequency was not significantly different from 0.5 according the the binomial test was considered heterozygous. For heterozygous normal positions, the corresponding tumor mutant allele frequency was tested for homozygosity using the binomial test. This resulted in a set of genomic locations with binary states: heterozygous in normal and tumor, and heterozygous in normal, but homozygous in tumor. The R Bioconductor DNAcopy package [9] was used to split these states into continuous segments representing tumor regions with and without LOH.

LOH regions were further filtered by treating each LOH region as one genomic location, and accumulating the mutant allele frequency for the LOH region. This combined mutant allele frequency was tested using a binomial test to ensure that it was significantly different from the expected 0.5 heterozygous mutant allele frequency.

### C. Microarray gene expression

Whole genome gene expression was derived from hybridization to NimbleGen human whole genome expression microarrays. Array analysis was performed on 15 melanomas [10]. Data from the array analysis were used to identify expressed genes in melanomas. Genes with a median expression value of 550 and above were called expressed. These expression data were used to assess the performance of the four gene burden methods presented here.

### D. RNA-Seq gene expression

RNA-Seq was performed on two independent cultures of two normal human melanocytes cultures derived from newborn foreskins and adult skin. Total RNA was extracted using Trizol (Invitrogen) followed by DNase digestion and Qiagen RNeasy (Qiagen, Valencia, CA) column purification following the manufacturer's protocol. The RNA integrity was verified using an Agilent Bioanalyzer 2100 (Agilent, Palo Alto, CA). One microgram of high-quality RNA was processed using an Illumina RNA-Seq sample prep kit following the manufacturer's instructions (Illumina, San Diego, CA). Final RNA-Seq libraries were sequenced at 75 bp/sequence using an GAIIx Illumina sequencer. Reads were processed with bwa and SAMtools. Mapping was performed against the reference genome. Reads were counted in bins of 100 basepairs, and normalized with regard to the median. To calculate the expression value for a particular RefSeq transcript, we determined the transcript exon boundaries, and summed up all bin read values for bins within the boundaries. The transcript length-normalized, and log-transformed value was used as the measure of gene expression. A two component Gaussian mixture model was fit to the data, and a lower bound for expressed genes was chosen as two standard deviations away from the higher distribution mean. The RNA-Seq data were used to identify expressed genes in normal melanocytes for the gene burden analysis.

### E. Replication timing

Replication timing data were taken from the ENCODE Repli-seq tracks provided by the University of Washington [8]. We used the percent signal files of both replicants from the following cells: Bj, Bg02es, Gm06990, Gm12801, Gm12878, Helas3, Hepg2, Huvec, Imr90, Mcf7, Nhek, Sknsh. For early replication values, we used cell cycle stages G1b and S1. For late replication values, we used S4 and G2. For each gene and cell type, we found the ratio of the average replication activity for early and late stages. We considered a gene to replicate early if the ratio was positive for all cells and all replicants. Similarly, a late replicating gene had a ratio that was consistently negative across all cells and replicants.

### F. Mutation burden analysis

To calculate a list of significantly mutated genes, i.e., genes with more mutations than expected by the background mutation frequency, we modified a recently established protocol [3]. We used the nonsynonymous:synonymous (NS:SN) mutation ratio to estimate the nonsynonymous background mutation frequency. This estimate is then used to determine whether some observed number of nonsynonymous mutations in a gene is above the expected count. We also used insights into melanoma-specific mutation patterns to calculate mutation frequencies based on sequence contexts, and on expression of the gene locus. We measured an increase in mutation frequency when studying unexpressed versus expressed genes, and observed that most mutations occur at cytosines in the dipyrimidine context, as described before [2]. This led us to calculate the nonsynonymous background mutation frequencies separately for expressed and unexpressed genes, and separately for the three following sequence contexts: mutating Cs at dipyrimidines, 2) mutating Cs at non-dipyrimidines, and 3) mutating Ts, which stand for, respectively, mutations in cytosines with a flanking pyrimidine, mutations in cytosines without a flanking pyrimidine, and mutations in thymines with no restriction on the flanking bases.

We found expected context-specific nonsynonymous mutation frequencies using context specific NS:SN ratios and synonymous counts, and performed, for each gene, and for each context, a binomial test for whether the observed nonsynonymous mutations in a gene are explained by the expected estimate, receiving three distinct and independent p-values for each context. We then use the Fishers combined probability test to generate an overall p-value measuring whether the number of nonsynonymous mutations in a gene is more than expected.

We estimated gene-specific NS:SN ratios in each of the three contexts. We proceeded as follows: we first identified all bases in a particular gene that are positioned in the context C under consideration. We then performed an in-silico experiment where we mutated each base and recorded whether the change resulted in a nonsynonymous change or not. The resulting ratios between nonsynonymous and synonymous changes were weighted according to the observed frequencies for a particular base change. The frequencies for each base change, in each context, were calculated from the frequencies of the observed synonymous and nonsynonymous base changes, with the exception of nonsynonymous changes in the top 100 mutated genes, which may be enriched for driver mutations. The top 100 genes were determined by dividing the number of observed somatic mutations by the gene length, and ranking of the resulting ratios. We determined an overall NS:SN ratio, across the three contexts, and across all genes, of 1.93 in sun-exposed melanomas.

The final gene burden ranks were matched against similar ranks that were generated by excluding the top 5% of mutated samples, in order to ensure robustness of the results. Only genes that were ranked high in both lists were retained.

## References


[1] A. Hodgkinson, Y. Chen, and A. Eyre-Walker, "The large-scale distribution of somatic mutations in cancer genomes," *Human Mutation*, vol. 33, no. 1, pp. 136–143, 2012.

[2] E. Pleasance, R. Cheetham, P. Stephens, D. McBride, S. Humphray, C. Greenman, I. Varela, M. Lin, G. Ordóñez, G. Bignell *et al.*, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, no. 7278, pp. 191–196, 2009.

[3] L. Ding, G. Getz, D. Wheeler, E. Mardis, M. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. Muzny, M. Morgan *et al.*, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, 2008.

[4] M. Krauthammer, Y. Kong, B. Ha, P. Evans, A. Bacchiocchi, J. McCusker, E. Cheng, M. Davis, G. Goh, M. Choi *et al.*, "Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma," *Nature Genetics*, 2012.

[5] S. Sherry, M. Ward, M. Kholodov, J. Baker, L. Phan, E. Smigielski, and K. Sirotkin, "dbsnp: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[6] N. Siva, "1000 genomes project," *Nature Biotechnology*, vol. 26, no. 3, pp. 256–256, 2008.

[7] J. Stamatoyannopoulos, I. Adzhubei, R. Thurman, G. Kryukov, S. Mirkin, and S. Sunyaev, "Human mutation rate associated with DNA replication timing," *Nature Genetics*, vol. 41, no. 4, pp. 393–395, 2009.

[8] R. Hansen, S. Thomas, R. Sandstrom, T. Canfield, R. Thurman, M. Weaver, M. Dorschner, S. Gartler, and J. Stamatoyannopoulos, "Sequencing newly replicated DNA reveals widespread plasticity in human replication timing," *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 139–144, 2010.

[9] V. Seshan and A. Olshen, "Dnacopy: A package for analyzing DNA copy data," 2010.

[10] R. Halaban, M. Krauthammer, M. Pelizzola, E. Cheng, D. Kovacs, M. Sznol, S. Ariyan, D. Narayan, A. Bacchiocchi, A. Molinaro *et al.*, "Integrative analysis of epigenetic modulation in melanoma cell response to decitabine: clinical implications," *PLoS One*, vol. 4, no. 2, p. e4563, 2009.