WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

An integrative approach for the identification of somatic mutations that drive cancer

Keegan Korthauer[1]* and Christina Kendziorski[1,2]

[1]Department of Statistics, UW-Madison, WI; [2]Department of Biostatistics and Medical Informatics, UW-Madison, WI 53706

ABSTRACT

Identifying and prioritizing somatic mutations is an important and challenging area of cancer research that can provide new insights into gene function as well as new targets for drug development.

Most methods for prioritizing mutations rely primarily on recurrence-based criteria, where a gene is identified as having a causal mutation (driver) if it is altered in significantly more patients than expected according to a background model describing random (passenger) mutations. Although useful, the background models considered to date do not accommodate gene-specific features that are known to have a significant effect on mutation rate, such as replication timing. Furthermore, these methods do not incorporate information concerning the likelihood that a given mutation is functional.

Here we develop an integrative approach that uses an improved background mutation model and incorporates both recurrence and functional impact criteria for inferring driver gene status. Applying this model to data from The Cancer Genome Atlas (TCGA) Ovarian project, we identify several genes as drivers that were not identified based on recurrence criteria alone.

INTRODUCTION

Cancer arises from the accumulation of causal somatic mutations (termed drivers) that confer a selective advantage; possible forms of selective advantage are activation of proliferation signals, suppression of apoptosis signals, or suppression of DNA repair mechanisms. Further understanding of these processes and the genes involved will provide valuable insight into tumor biology as well as new therapeutic potentials. Thus, it is of great interest to separate the driver mutations that contribute to tumorigenesis from those random passenger mutations that are irrelevant to the cancer phenotype.

Cancer genome sequencing projects, such as The Cancer Genome Atlas (TCGA) project, have identified somatic mutations in the exomes of hundreds of patients for several cancer types, but statistical methods for prioritizing these mutations are needed to accurately infer driver gene status.

Table 1

Mutation Category	Nucleotide Context	Type
Transition	A:T → G:C	1
	C:G → T:A (non CpG)	2
	C:G → T:A (CpG)	3
Transversion	A:T → C:G or T:A	4
	C:G → A:T or G:C (non CpG)	5
	C:G → A:T or G:C (CpG)	6
Other (Indel)	In Frame	7
	Frameshift	8

Existing methods for driver gene identification often rely on recurrence-based criteria, which infer that a gene is a driver if it is mutated in significantly more patients than expected according to a background mutation model. Background models considered to date include factors such as mutation type and nucleotide context. **Table 1** (above) shows the mutation type and nucleotide context factors adjusted for in the background mutation model of Youn and Simon (2011).

These models do not, however, adjust for region-specific factors such as replication-timing, which is known to be associated with somatic mutation rate. Here, we extend the background mutation model of Youn and Simon (2011) to include adjustment for replication timing.

Figure 1

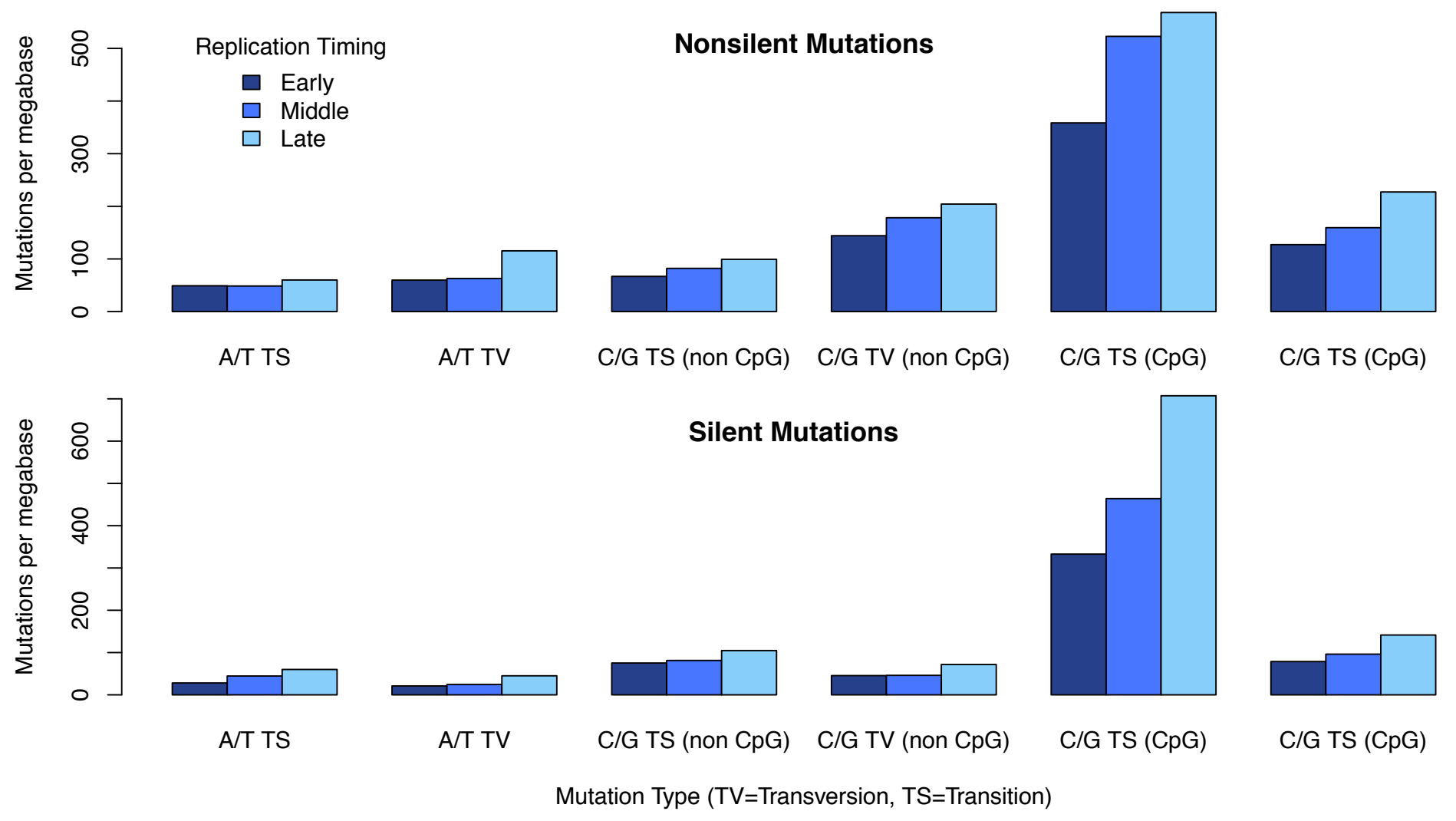
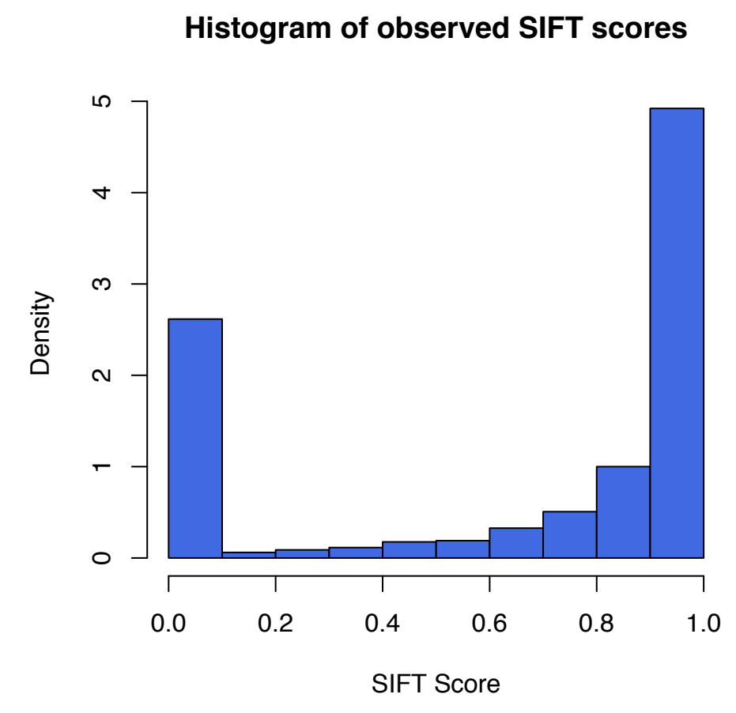


Figure 1 shows that later-replicating regions are associated with an increased somatic mutation rate in the TCGA ovarian cohort.

In addition to using recurrence criteria as evidence for selection, we would like to prioritize mutations based on their predicted functional impact. Several variant impact predictors have been developed, but here we focus on SIFT (Sorting Intolerant From Tolerant) scores, which represent the predicted likelihood that a mutation is deleterious.

Figure 2



Tolerant ⟷ **Intolerant**

Figure 2 displays the observed distribution of SIFT scores in the TCGA ovarian cohort. We assume that the distributions of both mutational frequencies and SIFT scores differ for driver versus passenger mutations and take advantage of this in order to make inference about driver gene activity. Recurrence and functional impact criteria are combined using a mixture model framework.

METHODS

• Background Mutation Model

The background mutation model of Youn and Simon assumes (1) whether or not a mutation generated from the background model is silent or not is determined by the genetic code, (2) the relative frequencies of mutation types are constant across samples, and (3) the probability of mutation at a given site depends on the possible mutation types and their nucleotide context, the sample, and the timing of replication for that site.

j = sample
k = position
t = mutation type
q_j = mutation rate sample j
n_{kt} = # possible mutations of type t at position k
m_t = relative rate of mutation type t
s_k = relative rate of replication timing region at k
X_{jk} = mutation status of sample j, position k

Assume

$$\Pr(X_{jk}=1)=\sum_t n_{kt}q_j m_t s_k \equiv b_{jk}$$

Replication timing data was obtained from Koren et al. (2012) and used to map each position of the exome to one of three replication timing categories (Early, Middle, and Late), defined by splitting on the tertiles of the observed genome-wide distribution.

To fit the background model, we obtain method of moments estimates for the relative rates of each of the 8 mutation types in **Table 1**, as well as the relative rates of mutation for the three replication timing categories. Empirical Bayes methods are used to estimate the distribution of the sample-specific overall mutation rates. Obtain gene-level estimates of the probability of mutation by summing over all mutation types and positions in the gene, and then integrating over the distribution of sample-specific rates.

• Mixture model framework

Recurrence component:

j : sample
X_{jg} : mutation status gene g, sample j
Z_g : driver status gene g
p_0 : prior probability to be passenger
g : gene
b_{jg} : background mutation rate gene g, sample j
d_{jg} : driver mutation rate gene g, sample j
p_1 : prior probability to be driver

Assume:

$$X_{jg} \mid Z_g = 0 \sim Bern(b_{jg})$$
$$X_{jg} \mid Z_g = 1 \sim Bern(d_{jg})$$

Then $X_{jg} \sim p_0 b_{jg}^{X_{jg}}(1-b_{jg})^{1-X_{jg}} + p_1 d_{jg}^{X_{jg}}(1-d_{jg})^{1-X_{jg}}$

where b_{jg} comes from the background mutation model, summing over all positions in gene g, and $d_{jg} \sim Beta(\alpha,\beta)$, truncated to satisfy constraint $d_{jg} > \frac{1}{J}\sum_j b_{jg} = b_{.g}$ and where α and β are elicited from a set of putative driver genes in Vogelstein et al. (2013) and their mutational frequencies in the COSMIC (Catalogue of Somatic Mutations in Cancer).

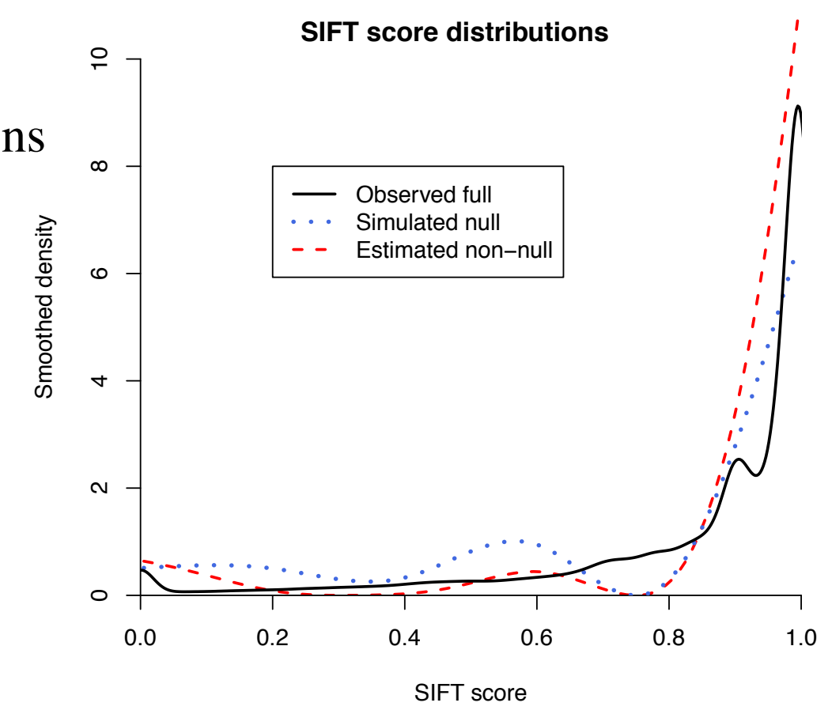
Functional impact component:

S_{jg} : functional impact score (SIFT) sample j, gene g
f^b : distribution of SIFT scores for background mutations
f^d : distribution of SIFT scores for driver mutations

Assume:

$$S_{jg} \mid X_{jg} = 0 \sim \text{ point mass at -1}$$
$$S_{jg} \mid X_{jg} = 1,\ Z_g = 0 \sim f^b$$
$$S_{jg} \mid X_{jg} = 1,\ Z_g = 1 \sim f^d$$

Figure 3



where f^b is estimated from scoring a random sample of mutations simulated from the background mutation model, and f^d is estimated by nonparametric spline regression on the ratio of the simulated null to the observed full distribution f of scores across bins of the score range (see Efron (2001)). These components are illustrated in **Figure 3**.

• Posterior probability of driver activity

For an independent sample of J tumors, the prior predictive distributions for the mutational status and functional impact score vectors of gene g are

$$P(S_g = s,\ X_g = x \mid Z_g = 0) = \prod_j (f^b(s_j)b_{jg})^{x_j}(1-b_{jg})^{1-x_j}$$

$$P(S_g = s,\ X_g = x \mid Z_g = 1) = \frac{[1-F_{(\mathrm{A,B})}(b_{.g})]B(\mathrm{A,B})}{[1-F_{(\alpha,\beta)}(b_{.g})]B(\alpha,\beta)}\prod_j f^d(s_j)^{x_j}$$

where parameters $\mathrm{A} = \sum x_j + \alpha$ and $\mathrm{B} = J - \sum x_j + \beta$, $F_{(\mathrm{A,B})}(y)$ is the cumulative distribution function of the $Beta(\mathrm{A,B})$ distribution, and B is the Beta function The posterior probability of gene g being a driver may be calculated using Bayes rule:

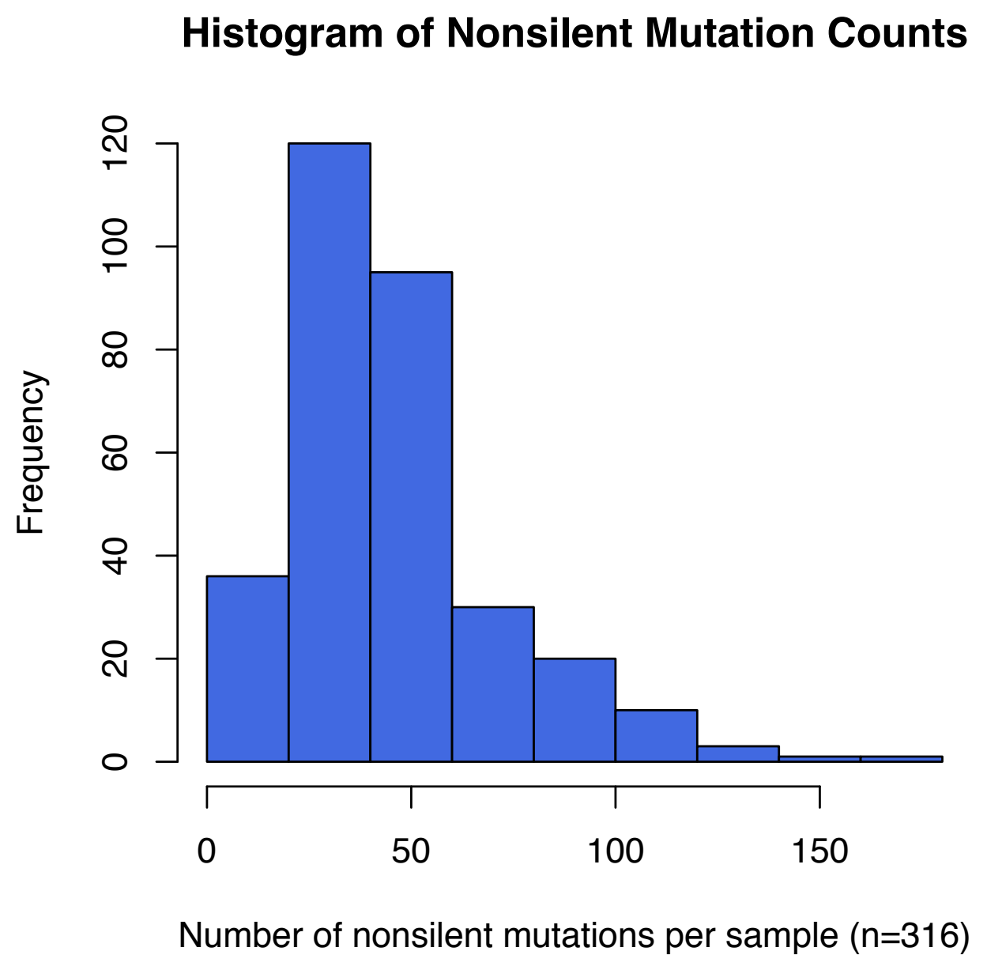
$$P(Z_g = 1 \mid S_g = s,\ X_g = x\ ) = \frac{p_1[1-F_{(\mathrm{A,B})}(b_{.g})]B(\mathrm{A,B})\prod_j f^d(s_j)^{x_j}}{p_1[1-F_{(\mathrm{A,B})}(b_{.g})]B(\mathrm{A,B})\prod_j f^d(s_j)^{x_j} + p_0[1-F_{(\alpha,\beta)}(b_{.g})]B(\alpha,\beta)\prod_j (f^b(s_j)b_{jg})^{x_j}(1-b_{jg})^{1-x_j}}$$

CASE STUDY RESULTS

We considered the somatic mutation data from the TCGA ovarian cohort, publicly available from http://cancergenome.nih.gov/. This dataset consists of somatic mutation calls for the exomes of 316 patients between normal blood and tumor tissue samples. Each somatic mutation is annotated for the sample(s) in which it occurs, its chromosome and base pair position, the gene in which it is located, the allele found in the reference genome, the specific nucleotide change, and the type of mutation (silent, missense, frameshift indel, inframe indel).

We observe that in this sample there are a total of 3,960 silent mutations located in 3,192 genes and 14,710 nonsilent mutations located in 8,181 genes. The median (range) total number of mutations per sample is 59 (8-193). The median (range) number of silent mutations per sample is 13 (0-41) and the median (range) number of nonsilent mutations per sample is 41 (6-161). In **Figure 4** we see that the distribution of nonsilent mutations per sample is right-skewed, with a few individuals containing over a hundred nonsilent mutations.

Figure 4



The set of silent mutations and all sequences containing at most one nonsilent mutation were used to estimate parameters in the background mutation model. The background mutation model was then used to compute the sample- and gene-specific background mutation probabilities p_{jg}, and then posterior probabilities of driver activity were computed for each gene.

Table 2 shows the 29 genes with posterior probability greater than 0.95 of being a driver, along with the number of samples in the TCGA cohort with at least one nonsilent mutation in that gene, and whether or not the gene was found significant by recurrence criteria alone using the background mutation model in Youn and Simon (2011).

Table 2

Gene	Posterior Probability	# Samples Mutated	Significant by Youn & Simon?
TP53	1.000	304	Y
BRCA1	1.000	11	Y
CDK12	1.000	9	Y
NF1	1.000	13	Y
CSMD3	1.000	19	Y
MYO3A	0.997	7	N
MAP3K19	0.997	7	N
RB1	0.995	6	N
EFEMP1	0.995	5	Y
LATS1	0.994	6	N
MYH1	0.994	9	N
PROKR2	0.993	4	N
CCAR1	0.993	6	N
CREBBP	0.984	7	N
PPP1R3A	0.983	8	Y
TTN	0.981	67	Y
KRT72	0.976	4	N
DUSP19	0.975	4	Y
TBX5	0.974	5	Y
STK10	0.974	5	N
OR11G2	0.971	3	N
PGAP1	0.971	5	N
EPHA7	0.971	7	Y
ATG3	0.969	3	N
PAX3	0.963	5	N
ADAMTS14	0.962	7	N
VSIG2	0.961	4	Y
MYH11	0.959	7	N
MC2R	0.955	3	N

CONCLUSIONS

- Existing methods for identifying driver genes that rely primarily on recurrence criteria suffer from (1) inadequate background mutation models and (2) ignoring information about the potential functional effect of a mutation by treating all mutations of a given type equally.
- We have developed a mixture model framework to account for both evidence of recurrence under an improved background model and functional impact criteria in assessing driver activity of genes using somatic mutation data. Simulation studies are underway to evaluate power and type I error.

REFERENCES

- Efron, B. et al. Empirical Bayes analysis of a microarray experiment. *J Am Statist Assoc,* 96(456):1151-1160, 2001.
- Forbes, S. A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res,* 39(suppl 1):D945-D950, 2011.
- Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J of Human Genet,* 91:1033-1040, 2012.
- Kumar, P. et al. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4(8): 1073-1082, 2009.
- Vogelstein, B. et al. Cancer genome landscapes. *Science,* 339(6127): 1546-1558, 2013.
- Youn, A. and Simon, R. Identifying cancer driver genes in tumor sequencing studies. *Bioinformatics*, 27(2):175-181, 2011.

Contact Information:

Keegan Korthauer
kdkorthauer@wisc.edu
Christina Kendziorski
kendzior@biostat.wisc.edu

Acknowledgements: NIGMS Biostatistics Training Grant