BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 2 2011, pages 175–181
doi:10.1093/bioinformatics/btq630

Sequence analysis

Identifying cancer driver genes in tumor genome sequencing studies

Ahrim Youn and Richard Simon*

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda MD 20892-7434, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Major tumor sequencing projects have been conducted in the past few years to identify genes that contain 'driver' somatic mutations in tumor samples. These genes have been defined as those for which the non-silent mutation rate is significantly greater than a background mutation rate estimated from silent mutations. Several methods have been used for estimating the background mutation rate.

Results: We propose a new method for identifying cancer driver genes, which we believe provides improved accuracy. The new method accounts for the functional impact of mutations on proteins, variation in background mutation rate among tumors and the redundancy of the genetic code. We reanalyzed sequence data for 623 candidate genes in 188 non-small cell lung tumors using the new method. We found several important genes like PTEN, which were not deemed significant by the previous method. At the same time, we determined that some genes previously reported as drivers were not significant by the new analysis because mutations in these genes occurred mainly in tumors with large background mutation rates.

Availability: The software is available at: http://linus.nci.nih.gov/Data/YounA/software.zip

Contact: rsimon@mail.nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 22, 2010; revised on September 23, 2010; accepted on November 7, 2010

1 INTRODUCTION

Major tumor sequencing projects (Ding *et al.*, 2008; Greenman *et al.*, 2007; Sjoblom *et al.*, 2006; Wood *et al.*, 2007) have been conducted and initiated in the past few years to identify genes that are frequently mutated and thereby are expected to have primary roles in the development of tumor. One of the challenges in interpreting this data is distinguishing driver mutations, which have a role in oncogenesis or in the cancer phenotype from passenger mutations that accumulate through DNA replication but are irrelevant to tumor development. To find these driver genes, each gene is tested for whether its mutation rate is significantly higher than the background (or passenger) mutation rate. The background mutation rate is estimated based on silent mutations which do not change amino acid encoding and which are therefore considered to be passenger mutations.

All current methods for estimating the background mutation rate are based on a common approach in which background non-silent mutation rate ρ_N is estimated as a product $\rho_S R$, where the background silent mutation rate ρ_S is obtained by dividing the observed number of silent mutations by the number of base pairs sequenced and R is the average ratio of the number of potential non-silent mutations to the number of potential silent mutations. Having estimated the background non-silent mutation rate ρ_N, each gene can be tested whether the number of mutations is significantly greater than that expected under the background mutation rate using a binomial test.

The methods used for calculating R vary. Ding *et al.* (2008) calculated R in the following way. They mutate each nucleotide of each codon *in silico* to determine whether it results in a non-silent or silent mutation. They then calculate the average of each hypothetical non-silent or silent mutation by weighting it according to its relative frequency.

Some investigators (Sjoblom *et al.*, 2006) further divide mutations into several types according to the nucleotide and the neighboring nucleotides of the mutations. They estimate a separate background mutation rate for each mutation type by multiplying relative frequencies of each mutation type by the background rate ρ_N. They then test each gene by using a likelihood ratio test to assess whether the number of mutations occurring in the gene is unlikely under the background mutation rates.

There are three shortcomings in the approaches previously developed for identifying driver genes. First, previous approaches ignore the fact that different types of mutations can have different impact on proteins. Non-silent mutations include missense mutations which change an amino acid to another amino acid, nonsense mutations which change an amino acid to a stop codon, mutations in splice sites and insertions or deletions (indels). The indels can also be divided into two types namely, inframe indels and frameshift indels according to whether it changes the reading frame.

Since frameshift indels and nonsense mutations change all the amino acids that come after the amino acid where the mutation occurred, they have the greatest impact on the protein function. Mutations in splice sites also have strong impact since they disrupt splicing. Also different types of missense mutations may have different impact based on how similar are the chemical properties of the original and new amino acids.

Several studies also have shown that the selection pressures vary by mutation type and sequence location in cancer mutation datasets.

*To whom correspondence should be addressed.

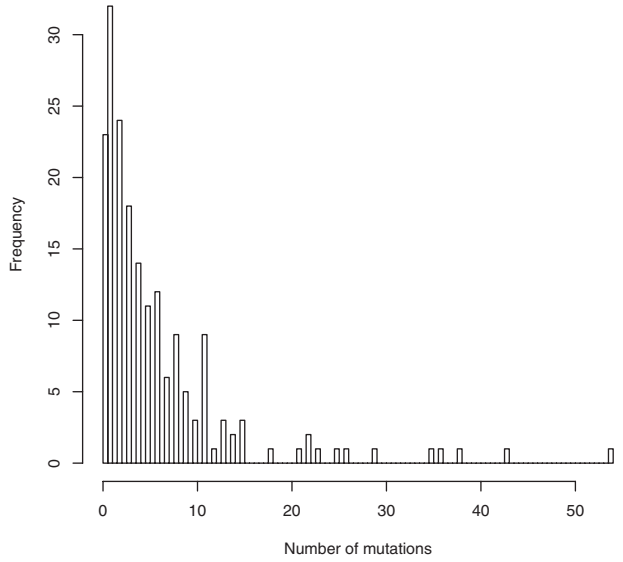

Fig. 1. Histogram of the number of mutations per sample. The data are from Ding *et al.* (2008) who sequenced 623 genes in 188 tumor samples.

Greenman *et al.* (2006) developed tests to examine the significance of selection toward missense, nonsense and splice site mutations in somatic cancer mutation datasets. They found that the selection pressures for nonsense and splice mutations are much higher than those for missense mutations. Also Radivojac *et al.* (2008) found that somatic cancer mutation datasets have a significant enrichment for mutations disrupting phosphorylation sites.

If two genes A and B have a similar number of mutations, but all mutations in gene A are expected to affect its protein function significantly while those in gene B are not, then the gene A is more likely to be a driver gene than gene B. However, the current methods are not able to differentiate genes A and B since they ignore the information of mutation types.

A second limitation of previous approaches is that they ignore the fact that different samples have different background mutation rates. Tumors differ substantially with regard to the number of somatic mutations accumulated. Samples are exposed to different levels of mutagens (for example, smoking) and some samples have mutations in genes that repair mutations. Therefore, some samples have much higher background mutation rate than others. This can be seen in Figure 1 for the data from Ding *et al.* (2008) who sequenced 623 genes in 188 tumor samples to identify 1013 non-silent mutations and 108 silent mutations. Figure 1 shows the distribution of the number of mutations that occurred in each of the 188 samples. It shows that the number of mutations per sample ranges from 0 to 54.

If a gene has mutations only in the samples with high background mutation rate, then those mutations are more likely byproducts of the high background mutation rate rather than the cause of a cancer. In contrast, if a gene has mutations only in the samples with low background mutation rate, then the gene is more likely to be a driver gene even if the number of mutations is small. If we assume the same background mutation rate across samples, the analysis will be biased toward falsely identifying as drivers those genes that have mutations in highly mutated samples and falsely missing those genes with a small number of mutations in samples with low mutation rates.

Third, previous approaches ignore the fact that a different number of non-silent mutations can occur at each base pair according to the genetic code. For example, consider a codon TGG which encodes the amino acid Tryptophan. Since this is the only codon encoding Tryptophan, any mutation at any nucleotide of the codon would change the amino acid. Therefore any mutation results in a non-silent mutation. In contrast, six codons encode the same amino acid arginine: AGA, AGG, CGA, CGG, CGC and CGT. Therefore within the codon CGA, no non-silent mutations can occur at the third position of the codon and only two non-silent mutations can occur at the first position. If a protein A consists mostly of tryptophan, and a protein B consists mostly of arginine, the gene A encoding the protein A is susceptible to more non-silent mutations than the gene B encoding the protein B. Thus, methods ignoring this difference will tend to misclassify as drivers genes rich in codons with limited redundancy and misclassify as non-drivers genes rich in codons with substantial redundancy.

In this article, we propose and evaluate a new method for identifying driver genes. In Section 2, we will define *P*-values for testing whether a gene is a driver gene. In Section 3, we will evaluate the new method using lung tumor genome sequences.

2 METHODS

2.1 Definition of *P*-values for identifying driver genes

For each gene, we test if the number of samples with 'driver-like' non-silent mutations is higher than that expected by the background mutation model M_0.

Let

$$Y_{ij} = \begin{cases} 0 & \text{if no non-silent mutation occurred in sample } j \text{ for gene } i \\ 1 & \text{if any non-silent mutation occurred in sample } j \text{ for gene } i \end{cases}$$

then Y_{ij} is a Bernoulli random variable.

Define

$$s_{ij} = P(Y_{ij} = 0 | M_0).$$

Since we assume a different mutation rate for each sample, the probability s_{ij} varies across samples $j = 1, \ldots, J$. It is calculated from the background mutation model which will be described in the Section 2.2.

We assign a score to every possible non-silent mutation according to its expected impact on the protein function: higher score for mutations with stronger impact. As will be shown, the order between scores rather than the actual scores determines the test statistics. Therefore, one can assign any score to each non-silent mutation to reflect the order of its impact on the protein function. We assign scores so that they comply with the following order: missense < inframe indel < mutation in splice sites < frameshift indel = nonsense. We also assign different scores to different types of missense mutations based on BLOSUM80 matrix, which is a matrix of scores for each of the 190 possible substitutions of the 20 standard amino acids.

Let T_{ij} be the maximum score of the non-silent mutations that occurred in sample j for gene i. If no mutation occurred, let $T_{ij} = 0$. Define $F_{ij}(x) = P(T_{ij} < x | Y_{ij} = 1, M_0)$. We can obtain the distribution F_{ij} from the background mutation model described in the Section 2.2.

Then,

$$\log P(Y_{ij} = y_{ij}, T_{ij} \geq t_{ij}, j = 1, \ldots, J | M_0)$$
$$= \log \prod_{j=1}^{J} P(T_{ij} \geq t_{ij} | Y_{ij} = y_{ij}, M_0) P(Y_{ij} = y_{ij} | M_0)$$

$$= \log \prod_{j=1}^{J} s_{ij}^{1-y_{ij}} ((1-s_{ij})(1-F_{ij}(t_{ij})))^{y_{ij}}$$

$$= \sum_{j=1}^{J} y_{ij} \left(\log\left(\frac{1-s_{ij}}{s_{ij}} \right) + \log(1-F_{ij}(t_{ij})) \right) + \sum_{j=1}^{J} \log s_{ij}$$

We use $Z_i = \sum_{j=1}^{J} Y_{ij}(\log(\frac{1-s_{ij}}{s_{ij}}) + \log(1-F_{ij}(T_{ij})))$ as our test statistic and define P-values by $P(Z_i < z_i | M_0)$, where z_i is the observed value of Z_i. Z_i can be interpreted as a sum of mutated sample indicators weighted by $\log(\frac{1-s_{ij}}{s_{ij}}) + \log(1-F_{ij}(T_{ij}))$. Since s_{ij} is larger than 0.5, $\frac{1-s_{ij}}{s_{ij}}$ is less than one and thus, the weights are negative. The larger the s_{ij} and $F_{ij}(T_{ij})$, the smaller the weight. Therefore, mutations with higher scores (stronger impact on the protein function) occurring in samples with low mutation rates (samples with large s_{ij}) contribute more in decreasing Z_i and thus, P-values.

We can generate the distribution of Z_i under the background mutation model by simulating Y_{ij} from $Bernoulli(s_{ij})$ and T_{ij} from F_{ij} for $j = 1, \ldots, J$ and then calculating the sum of $Y_{ij}(\log(\frac{1-s_{ij}}{s_{ij}}) + \log(1-F_{ij}(T_{ij})))$. Then we can approximate the P-value $P(Z_i < z_i | M_0)$ by computing the tail area of this background distribution beyond the observed value z_i for each gene i.

2.2 Background mutation model

The most distinguishing features of our background mutation model are that it does not assume separate mutation rates for non-silent and silent mutations and that it assumes separate mutation rates for different samples. We assume that each passenger mutation is generated from one background mutation rate process and that whether the mutation is non-silent or silent depends on the genetic code.

There are six types of mutations:

$$A{:}T \to G{:}C, \quad A{:}T \to C{:}G, \quad A{:}T \to T{:}A$$
$$G{:}C \to A{:}T, \quad G{:}C \to T{:}A, \quad G{:}C \to C{:}G.$$

The transitions $A{:}T \to G{:}C$ and $G{:}C \to A{:}T$ change a purine to another purine or a pyrimidine to another pyrimidine. The transversions change a purine to a pyrimidine or vice versa. Because transitions occur more frequently than transversions, we assume separate mutation rates for transitions and transversions. Also it is generally observed that a mutation occurs more often at $C{:}G$ than $A{:}T$ and that a $C{:}G$ appearing in *CpG* dinucleotides has a higher mutation rate than a $C{:}G$ appearing in non-*CpG* dinucleotides. Therefore, we assume a separate background mutation rate for each combination of base pair types and *CpG* dinucleotides context.

We also assume that different tumor samples have different mutation rates. To keep the number of parameters manageable, we assume that relative frequencies of different types of mutations are same for each sample. Thus, the mutation rate in sample j for mutation type m is defined as the product of p_m, the ratio of mutation rate of the type m relative to the type 1 ($A{:}T \to G{:}C$) and q_j, the mutation rate of the sample j for the mutation type 1 (Table 1).

To estimate the parameters in the background mutation model, we could fit the model in Table 1 to the sequences for which silent mutations were identified. (Most previous projects have evaluated silent mutations for only a subset of the genes.) To estimate the background mutation rate for insertions and deletions (indels), which are non-silent, however, we included in our estimation genes which have at most one non-silent mutation across all tumor samples; these genes are not likely to be related to tumorigenesis and thus the non-silent mutations in these genes are likely to be passenger mutations. However, since we selected these genes based on the total number of mutations occurring in each gene, the estimated background rates for these genes may be biased. Since the selection was based on the total number of mutations, it is unlikely that the relative frequencies of different types of mutations are subject to the bias, but the sample-specific mutation rates may be. Let q_j' be the mutation rate of the sample j for mutation type 1 in the selected genes. Then we assume $q_j' = r \cdot q_j$, where r is the selection bias and q_j is the unbiased sample-specific mutation rate.

Table 1. Background mutation rates

Mutation type	Mutation type ID	Mutation rate
$A{:}T \to G{:}C$	1	$q_j p_1$
$A{:}T \to C{:}G$	2	$q_j p_2$
$A{:}T \to T{:}A$	2	$q_j p_2$
$C{:}G \to T{:}A$ at non *CpG*	3	$q_j p_3$
$C{:}G \to A{:}T$ at non *CpG*	4	$q_j p_4$
$C{:}G \to G{:}C$ at non *CpG*	4	$q_j p_4$
$C{:}G \to T{:}A$ at *CpG*	5	$q_j p_5$
$C{:}G \to A{:}T$ at *CpG*	6	$q_j p_6$
$C{:}G \to G{:}C$ at *CpG*	6	$q_j p_6$
Inframe indels	7	$q_j p_7$
Frameshift indels	8	$q_j p_8$

*j is sample index.

Table 2. Definition of probabilities of X_{jk}

$$P(X_{jk}) = \begin{cases} q_j c_k p_{t_k} I(k \in K) & \text{for } X_{jk} = \text{sts} \\ q_j d_k p_{v_k} I(k \in K) & \text{for } X_{jk} = \text{stv} \\ q_j r e_k p_{t_k} I(k \in L) & \text{for } X_{jk} = \text{nts} \\ q_j r f_k p_{v_k} I(k \in L) & \text{for } X_{jk} = \text{ntv} \\ q_j r p_7 I(k \in L) & \text{for } X_{jk} = \text{iid} \\ q_j r p_8 I(k \in L) & \text{for } X_{jk} = \text{fid} \\ 1 - q_j a_k & \text{for } X_{jk} = \text{non} \end{cases}$$

$a_k = (c_k p_{t_k} + d_k p_{v_k}) I(k \in K) + r(e_k p_{t_k} + f_k p_{v_k} + p_7 + p_8) I(k \in L)$;
$I(x)$, indicator function, 1 if x is true and 0 otherwise;
c_k, number of silent transitions possible at position k (0 or 1); d_k, number of silent transversions possible at position k (0, 1 or 2); e_k, number of non-silent transitions possible at position k (0 or 1); f_k, number of non-silent transversions possible at position k (0, 1 or 2); t_k, mutation type ID for the transition at position k (1, 3 or 5); v_k, mutation type ID for the transversion at position k (2, 4 or 6); non, no mutation; sts, silent transition; stv, silent transversion; nts, non-silent transition; ntv, non-silent transversion; iid, inframe indel; fid, frameshift indel.

To estimate the parameters r, q_j and $p_1, \ldots, p_8$, we first define the position of base pairs across all the sequenced genes. Since we assume that background mutation rates are independent of genes, we do not need to differentiate genes. Therefore, we concatenate all the sequenced genes and determine the position of each base pair from 1 to N, the total number of base pairs that are sequenced. Let K denote the subset of positions of the base pairs belonging to the genes used for silent mutation detection and let L denote the subset of positions of the base pairs belonging to the genes which have at most one non-silent mutation across all samples.

For position k in genes for which silent mutations have been evaluated, the probability that a silent transition of type i (second column in Table 1) occurs in sample j equals $q_j c_k p_i$ where c_k is 1 if a transition at position k results in a silent mutation, otherwise c_k is 0. The probability that a silent transversion of type i occurs at that position equals $q_j d_k p_i$, where d_k is the number of silent transversions possible at position k (0, 1 or 2). The full set of probabilities definitions are shown in Table 2 based on the indicators X_{jk}, which indicate the type of mutation occurring at position k in sample j. Since any mutation is either a silent mutation or a non-silent mutation, $c_k + e_k = 1$ (number of possible transition mutations) and $d_k + f_k = 2$ (number of possible transversion mutations). When a mutation occurs within splice sites, it is considered to be non-silent, therefore $c_k = d_k = 0, e_k = 1, f_k = 2$ if k belongs to splice sites.

All of the constants shown in Table 2 can be determined from the gene sequence and genetic code. However, the values are ambiguous in cases where genes have several alternative transcripts, and where some base pairs

belong to different codons in alternative transcripts. We describe how to determine the values of the constants in such cases in the Supplementary Material.

2.3 Estimation of parameters

We use the method of moments to estimate r and p_m. The process of obtaining the method of moments estimates $\hat{r}$ and $\hat{p}_m$ is described in the Supplementary Material.

The estimation of q_j is more complex because the number of base pairs sequenced per sample is not sufficient to estimate the extremely small mutation rate q_j accurately. For example, no mutations were found for many samples in the data from Ding *et al.* (2008). Therefore, the maximum likelihood estimate of q_j would be zero for those samples, which are problematic point estimates. To improve the accuracy of the estimates, we use empirical Bayes methods to estimate the distribution of q_j. Empirical Bayes methods borrow information from all the samples for estimating each q_j, therefore give more robust estimates of q_j. (Casella, 1985)

We assume the prior distribution f of q_j is uniform on (α, β). As estimates of α, β, we use the values maximizing the marginal likelihood given the estimates, $\hat{r}$ and $\hat{\vec{p}}=(\hat{p}_1,\hat{p}_2,\hat{p}_3,\hat{p}_4,\hat{p}_5,\hat{p}_6,\hat{p}_7,\hat{p}_8)$.

The posterior distribution of q_j given the data X_{jk} for $k \in K \cup L, j=1,\ldots,J$ and the estimated parameters $\hat{\alpha}, \hat{\beta}, \hat{r}, \hat{\vec{p}}$ is

$$h_j(q_j) = \lambda f(q_j|\hat{\alpha},\hat{\beta}) \prod_{k \in K \cup L} P(X_{jk}|q_j,\hat{r},\hat{\vec{p}})$$

where the product is over all positions k, the probabilities $P(X_{jk}|q_j,\hat{r},\hat{\vec{p}})$ are computed from the formulas in Table 2 and λ is the normalizing constant.

We use the posterior distribution of q_j in calculating $s_{ij}=P(Y_{ij}=0|M_0)$ rather than using the point estimates of q_j to take into account the uncertainty in the point estimate. Therefore,

$$s_{ij} = P(Y_{ij}=0|M_0) = \int h_j(q_j) \prod_{k \in G_i} (1-q_j b_k) dq_j$$

where G_i is the subset of positions of the base pairs belonging to the gene i and $b_k = e_k \hat{p}_{t_k} + f_k \hat{p}_{v_k} + \hat{p}_7 + \hat{p}_8$. The integration with regard to the posterior distribution of q_j is performed numerically. The resulting values of s_{ij} are used as described in Section 2.1 for computing statistical significance.

The distributions $F_{ij}(x)=P(T_{ij}<x|Y_{ij}=1,M_0)$ are also needed for the significance tests used to identify driver genes and are computed from:

$$F_{ij}(x) = \frac{\sum_{k \in G_i} P(T'_{jk} < x | X_{jk} = \text{nts, ntv, iid, or fid}) b_k}{\sum_{k \in G_i} b_k}$$

where T'_{jk} is the score of the mutation occurring in position k and sample j. The distribution $P(T'_{jk}<x|X_{jk}=\text{nts, ntv, iid , or fid })$ can be easily calculated from the genetic code and background mutation model. The process of the derivation of $F_{ij}(x)$ is explained in the Supplementary Material.

3 RESULTS

We applied our method to the data of Ding *et al.* (2008). They sequenced coding exons and splice donor/acceptor sites (dinucleotides in the $5'/3'$ ends of introns) of 623 genes in 188 samples from patients with lung adenocarcinoma to identify 1013 non-silent mutations. They selected a subset of 250 genes to identify 108 silent mutations for measuring a background mutation rate. The table describing all the identified mutations is available in the paper of Ding *et al.* (2008), but the patient-specific gene sequences are not. Thus, we used the reference sequence of coding exons and splice donor/acceptor sites (dinucleotides in the $5'/3'$ ends of introns) of the 623 genes from Ensembl release 46.

3.1 Simulation study

We first performed a simulation study to evaluate our method. For the comparison with the method of Ding *et al.* (2008), we did not include the mutation score T_{ij} in the test statistics, that is, we use the test statistic $Z_i=\sum_{j=1}^{J} Y_{ij}(\log(\frac{1-s_{ij}}{s_{ij}}))$ instead of $Z_i=\sum_{j=1}^{J} Y_{ij}(\log(\frac{1-s_{ij}}{s_{ij}})+\log(1-F_{ij}(T_{ij})))$.

We generate simulated data based on the data of Ding *et al.* (2008). We first generate passenger mutations by shuffling the locations of all observed non-silent and silent mutations across the genes sequenced. There are 1013 non-silent mutations observed in 623 genes and 108 silent mutations observed in 250 genes. For these mutations, we change the base pair positions in which the mutation occurred as follows: we randomly sample the base pair positions from the base pair positions within the sequence of all genes, which correspond to the same base pair types as the mutations. If a mutation occurred in the base pair A, we sample its new base pair position from all the base pair positions within the sequence of all genes corresponding to a base pair A. If the base pair is G or C, we also restrict the sampling by the CpG dinucletoide context. We then determine which of these mutations are non-silent or silent according to the genetic code. Since we randomly sample the base pair positions of all the mutations, they become evenly spread across all genes.

To see the effect of variation of mutation rates across samples, we change the sample ID in which mutations occurred by sampling a new sample ID under two different distributions namely, moderate sample variation and high sample variation.

The first distribution, moderate sample variation, is estimated from the background mutations of the data from Ding *et al.* (2008). We sample each sample ID with the probability proportional to the number of passenger mutations (silent mutations and non-silent mutations observed in genes with at most one non-silent mutations) that occurred in the sample. For the second distribution, high sample variation, we increase the mutation rates of the 10 samples with highest mutation rate by a factor of 10.

Finally, we make 20 driver genes by adding five non-silent mutations to 20 selected genes.

In our simulations, we have used the true expected ratio of non-silent to silent mutations (R) in applying the method of Ding *et al.* (2008) because we did not have their software for estimating R. This may somewhat overestimate the accuracy of their method.

Each simulation was repeated for 200 replications. The average number of true and false positive driver genes claimed based on P-value cutoffs of 0.005 and 0.01, respectively, are shown in Table 3. Our method finds more true positives and fewer false positives than the method of Ding *et al.* (2008). We did Wilcoxon signed rank test of the null hypothesis that the distribution of number of true or false positives from both methods are same and presented one-sided P-values in the last column of Table 3. For moderate sample variation, the P-values for false positives are 0.0001 and 0.0008, and the P-values for true positives are less than 10^{-16}. For high sample variation, all the P-values are less than 10^{-16}. This shows that the difference in the number of true positives or false positives between

Table 3. Result for simulated data

Sample variation	Cutoff	Average number	Our method	Ding's method	*P*-value
Moderate	0.005	TP	12.9	9.9	<1e-16
		FP	1.3	1.7	1e-04
	0.01	TP	14.9	11.7	<1e-16
		FP	3	3.4	8e-04
High	0.005	TP	13.4	9.9	<1e-16
		FP	0.2	2.0	<1e-16
	0.01	TP	15.1	11.7	<1e-16
		FP	0.6	3.9	<1e-16

TP, true positives; FP, false positives.

Table 4. Driver genes by new method

Gene name	*P*-value
EGFR	0
CDKN2A	0
KRAS	0
STK11	0
TP53	0
EPHA3	2e-06
NF1	2e-06
ATM	3e-06
RB1	4e-06
APC	1.3e-05
INHBA	6.8e-05
ERBB4	0.000109
PTPRD	0.000145
FGFR4	0.000146
PTEN	0.000210
EPHA5	0.000237
NTRK3	0.000298
NTRK1	0.000298
KDR	0.000319
LRP1B	0.000518
PAK3	0.000750
NRAS	0.000848
LTK	0.000876
ZMYND10	0.001091
EPHA7	0.001116
MYO3B	0.001151
NTRK2	0.001322
TFDP1	0.001404

two methods is significant and it gets more significant as the sample variation grows larger.

3.2 Results for the data of Ding *et al.* (2008)

We identified 28 genes as driver genes with the false discovery rate (FDR) controlled at 5% using the Benjamini and Hochberg method. These include EGFR, CDKN2A, KRAS, STK11, TP53, EPHA3, NF1, ATM, RB1, APC, INHBA, ERBB4, PTPRD, FGFR4, PTEN, EPHA5, NTRK3, NTRK1, KDR, LRP1B, PAK3, NRAS, LTK, ZMYND10, EPHA7, MYO3B, NTRK2 and TFDP1. The *P*-values of the selected genes are given in Table 4.

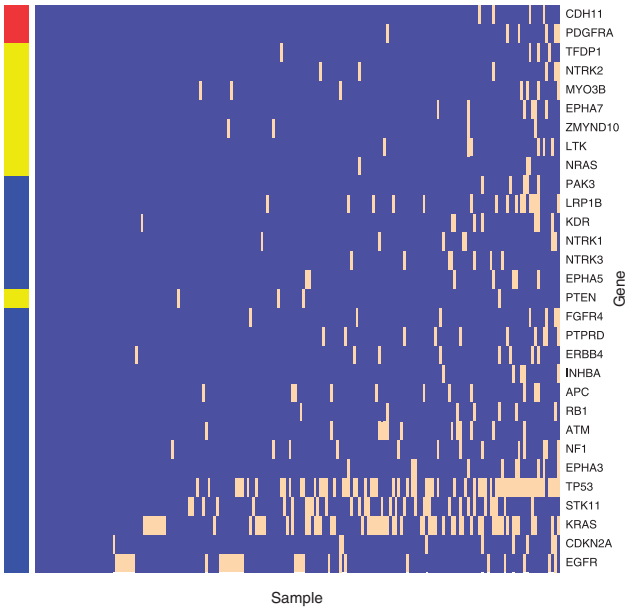

Fig. 2. Map of the 30 selected genes versus tumor samples. Tumor samples with/without mutations in genes are labeled yellow/blue. The rows (genes) are ordered according to the *P*-value obtained by our method. The columns (samples) are ordered according to the total number of genes with non-silent mutations (among all 623 genes) in the corresponding sample. The red/blue/yellow banner across the left side of the map shows the difference between selected genes by the two methods: our method and the method of Ding *et al.* (2008). The genes covered by the red bar are the additional genes found by the method of Ding *et al.* (2008) and those covered by the yellow bar are the additional genes found by our method. The genes covered by the blue bar are those which both methods find significant.

By the method of Ding *et al.* (2008), 22 genes were found to be significant at the 5% FDR level. These genes include two genes that we do not find significant. Our method finds eight genes, which they do not find significant. We drew a map of the genes selected by each method versus tumor samples in Figure 2. The genes are ordered by the *P*-value obtained by our method and samples are ordered according to the total number of genes with non-silent mutation (among all 623 genes). The genes indicated by the red bar are those which Ding *et al.* (2008) find significant but we do not. The genes indicated by the yellow bar are those which we find significant but they do not. The genes indicated by the blue bar are those which both methods find significant.

For most of the well known oncogene and tumor suppressor genes in the dataset we analyzed (EGFR, CDKN2A, KRAS, STK11, TP53, NF1, RB1 PTEN, NRAS), the mutations occurring in them have very high mutation scores. For example, in STK11, most mutations are frameshift indels, nonsense or mutations in splice sites. Even the missense mutations represent poorly conserved amino acid changes. In RB1, all seven mutations that occurred in the gene are either frameshift indel, nonsense or mutations in splice sites. This is consistent with our scoring system that mutations in driver genes will tend to have strong impact on protein functions. By incorporating mutation scores in calculating *P*-values, driver genes have smaller *P*-values and thus are better identified. For the well- known driver genes EGFR, CDKN2A, KRAS, STK11, TP53 which already have computed *P*-values of zero due to frequent mutations, the effect

of scoring makes little difference. But for genes like NF1, RB1, APC, INHBA, ERBB4, FGFR4, PTEN and NRAS, their *P*-values are about one-third on average of the *P*-values calculated without incorporating the mutation scores. Most of those genes are well-known cancer driver genes, and incorporating mutation scores helps in their identification.

Not having mutations with high scores does not preclude a gene being a potential driver gene, but these cases tend to be infrequently mutated genes that occur in samples with large mutation rates as was the case for CDH11 and PDGFRA. These two genes are selected to be significant by Ding's method, but not by our method. Figure 2 shows that the mutations in these genes are clustered in the highly mutated samples except one in PDGFRA. However, most mutations in PDGFRA are missense mutations with low mutation scores, offsetting the effect of low mutation rate to the test statistics of PDGFRA. Also, some of the mutations in both genes occur in the same sample, therefore, our method which is based on the number of samples with mutations rather than the total number of mutations assigned larger *P*-values to them than the method of Ding *et al.* (2008).

There are eight genes beside the yellow bar. The gene with the smallest *P*-value is PTEN, a well-known tumor suppressor gene. Ding *et al.* (2008) did not find it significant because the total number of mutations is so small (four). However, since each of the four mutations occurred in different samples with low mutation rates and the score of each mutation is high (one nonsense mutation and three missense mutation with high score), our method could find it significant.

The gene with the second smallest *P*-value is NRAS, a well-known oncogene. Although there were only three total mutations in this gene, all of them are the same missense mutation changing glutamine to leucine, which has a high score. Also, one of the mutation occurred in a sample with low mutation rate, thus we could find it significant.

The gene ZMYND10 is a candidate tumor suppressor gene whose association with carcinomas is suggested by Agathanggelou *et al.* (2003); Cho (2007); Lerman and Minna (2000); Marsit *et al.* (2005); Qiu *et al.* (2004).

The gene EPHA7 is a member of the ephrin receptor family and is known to be related to oncogenesis (Kiyokawa *et al.*, 1994). The other two members EPHA3 and EPHA5 are also selected to be significant by both methods, implying that EPHA7 is potentially involved in oncogenesis.

NTRK2 is a member of the neurotrophic tyrosine receptor kinase (NTRK) family, which phosphorylates members of the MAPK pathway. It is known to be potentially implicated in oncogenesis (Marchetti *et al.*, 2008) and also the other two members of the NTRK receptor family, NTRK1 and NTRK3 are selected to be significant by both methods, supporting the implication of NTRK2 in oncogenesis.

TFDP1 is a transcription factor and its overexpression or amplification is known to be associated with carcinomas (Melchor *et al.*, 2009; Yasui *et al.*, 2003). The role of LTK and MYO3B in oncogenesis is not well known.

4 DISCUSSION

We have developed a new method for identifying driver genes that has several methodological advantages compared with the previously used methods.

First, we assign scores to non-silent mutations according to their expected impacts on the protein function so that the genes with more 'driver-like' mutations will get smaller *P*-values.

Second, we permit each sample to have a different background mutation rate. This has the effect of reducing the false positives and increasing true positives, which was confirmed by the simulation study.

Third, instead of assuming separate background mutation rates for non-silent and silent mutations, we assume that each passenger mutation is generated from one background mutation rate process and that whether the mutation is non-silent or silent depends on the genetic code. Thus, our model accounts for the variable number of possible non-silent mutations that can occur at each base pair according to the genetic code. This takes into account the difference in the number of possible non-silent mutations between genes according to the codon usage within genes.

Fourth, we take into account uncertainties in the background mutation rate by using empirical Bayes methods.

These methodological advances contributed to identifying a different set of driver genes when compared with those identified by Ding *et al.* (2008). First, we did not find the genes CDH11 and PDGFRA which Ding *et al.* (2008) found significant. These genes are not selected by our method because they are mainly mutated in the highly mutated samples and the scores of the mutations are not high. Second, we found PTEN, NRAS, LTK, ZMYND10, EPHA7, MYO3B, NTRK2 and TFDP1, which Ding *et al.* (2008) did not find significant. It shows that our method is more sensitive in finding genes whose total number of mutations is small.

Although we believe that our method provides an improvement over the previous methods, there is room for improvement by extending our approach. First, we measure the functional impact of mutations by the significance of the change to amino acids caused by the mutation. However, the functional impact is also dependent on the position in which a mutation occurs. For example, all three mutations in NRAS occurred in the exact same base pair position, which implies that the mutation in the specific position is crucial to the function of the protein. If a score for each position can be estimated that measures the significance of the position in protein function, it can be used in our test statistics in the same way as the mutation score T_{ij}.

Second, the current scoring system which assigns mutation scores in the order: missense mutation < inframe indel < mutation in splice sites < frameshift indel = nonsense mutation may be biased toward identifying tumor suppressor genes over oncogenes. Loss-of-function mutations such as frameshift indels or nonsense mutations occur more frequently in tumor suppressor genes than in oncogenes. Our use of the BLOSUM80 matrix to refine the scoring of missense mutations helps in the identification of new oncogenes. Alternative scoring systems can be used, however, to increase sensitivity for identifying oncogenes. For example, we can assign the same scores as the current method to the missense mutations, but reduced scores to indels, mutation in splice sites or nonsense mutations.

Third, we may refine our background mutation model in Table 1 so that all six types of mutations, $A{:}T \rightarrow G{:}C$, $A{:}T \rightarrow C{:}G$, $A{:}T \rightarrow T{:}A$, $G{:}C \rightarrow A{:}T$, $G{:}C \rightarrow T{:}A$, $G{:}C \rightarrow C{:}G$ have separate mutation rates. We separate the rates of mutations according to the mutation types (transition or transversion), base pair types ($A{:}T$ or $G{:}C$) and their context (CpG dinucleotide contexts). Therefore, we did not separate the rates of the two types of mutation for each

transversion: $A{:}T \rightarrow C{:}G$, $A{:}T \rightarrow T{:}A$ for the transversion at $A{:}T$ and $C{:}G \rightarrow A{:}T$, $C{:}G \rightarrow G{:}C$ for the transversion at $C{:}G$ in non-*CpG* or in *CpG*. However, if the two types of mutations for each transversion have quite different mutation rates, it may induce bias. Therefore, we evaluated a model in which each of them has a separate mutation rate using the simulated data generated described in Section 3.1. For 200 repeated simulations, we calculated the average number of true and false positives for this method. When compared with the original method, the new model increased true positives as well as false positives. Supplementary Table S1 shows these results.

For larger datasets, one could refine our background mutation model to differentiate coding and non-coding strands. Currently, we assume that the mutation rate at the base pair $A{:}T$ for example is same whether A is in the coding strand or T is in the coding strand. Using separate mutation rates according to the coding and non-coding strand, however, will increase the number of parameters by almost 2-fold, and therefore will be feasible only for the large datasets.

Fifth, we did not take into account correlations among mutations in identifying driver genes. Indeed, none of the existing methods for identifying driver mutations that we are aware of utilize estimates of synergism or antagonism for pairs of mutations. However, strong positive or negative correlations between mutations in several pairs of genes have been observed. Therefore, one could attempt to utilize the correlation structure among mutations in identifying driver genes.

Finally, one might combine both copy number variation and sequencing data to identify driver genes. In this article, we used only genomic sequence changes to identify driver genes. However, change of protein functions related to oncogenesis are frequently caused by copy number variation. Therefore, it is desirable to integrate both copy number variation and sequence changes to identify driver genes if both data are available. Our method can be extended to include copy number variation in the test statistics; we can test for each gene if the number of samples with 'driver-like' non-silent mutation or copy number variation is higher than that expected by the background mutation model.

The analysis of tumor sequencing data is of key importance for understanding oncogenesis, identifying molecular targets and personalizing therapy. Learning to read the tumor genome is complex, however, and new methods of analysis are needed. We believe that methods such as those we have described that account for functional impact of mutations, sample variation in mutation rates and the redundancy of the genetic code will be useful for the identification of genes that drive the pathogenesis of cancer.

ACKNOWLEDGEMENTS

We thank Dr Li Ding for kindly sharing the unpublished table of silent mutations and answering questions.

Conflict of Interest: none declared.

REFERENCES

Agathanggelou,A. *et al.* (2003) Epigenetic inactivation of the candidate 3p21.3 suppressor gene blu in human cancers. *Oncogene*, **22**, 1580–1588.

Casella,G. (1985) An introduction to empirical Bayes data analysis. *Am. Stat.*, **39**, 83–87.

Cho,W.C.-S. (2007) Nasopharyngeal carcinoma: molecular biomarker discovery and progress. *Mol. Cancer*, **6**, 1.

Ding,L. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.

Greenman,C. *et al.* (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, **173**, 2187–2198.

Greenman,C. *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.

Kiyokawa,E. *et al.* (1994) Overexpression of ERK, an EPH family receptor protein tyrosine kinase, in various human tumors. *Cancer Res.*, **54**, 3645–3650.

Lerman,M.I. and Minna,J.D. (2000) The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. the international lung cancer chromosome 3p21.3 tumor suppressor gene consortium. *Cancer Res.*, **60**, 6116–6133.

Marchetti,A. *et al.* (2008) Frequent mutations in the neurotrophic tyrosine receptor kinase gene family in large cell neuroendocrine carcinoma of the lung . *Hum. Mutat.*, **29**, 609–616.

Marsit,C.J. *et al.* (2005) Hypermethylation of rassf1a and blu tumor suppressor genes in non-small cell lung cancer: implications for tobacco smoking during adolescence. *Int. J. Cancer*, **114**, 219–223.

Melchor,L. *et al.* (2009) Comprehensive characterization of the dna amplification at 13q34 in human breast cancer reveals tfdp1 and cul4a as likely candidate target genes. *Breast Cancer Res.*, **11**, R86+.

Qiu,G. *et al.* (2004) The candidate tumor suppressor gene blu, located at the commonly deleted region 3p21.3, is an e2f-regulated, stress-responsive gene and inactivated by both epigenetic and genetic mechanisms in nasopharyngeal carcinoma. *Oncogene*, **23**, 4793–4806.

Radivojac,P. *et al.* (2008) Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, **24**, i241–i247.

Sjoblom,T. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.

Wood,L.D. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.

Yasui,K. *et al.* (2003) Association of over-expressed tfdp1 with progression of hepatocellular carcinomas. *J. Hum. Genet*., **48**, 609–613.