Bioinformatics, 31(10), 2015, 1526–1535
doi: 10.1093/bioinformatics/btu858
Advance Access Publication Date: 7 January 2015
Original Paper

Genome analysis

MADGiC: a model-based approach for identifying driver genes in cancer

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

[1]Department of Statistics and [2]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison WI 53706, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 11, 2014; revised on November 19, 2014; accepted on December 23, 2014

Abstract

Motivation: 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 frequency-based criteria, where a gene is identified as having a driver mutation if it is altered in significantly more samples than expected according to a background model. Although useful, frequency-based methods are limited in that all mutations are treated equally. It is well known, however, that some mutations have no functional consequence, while others may have a major deleterious impact. The spatial pattern of mutations within a gene provides further insight into their functional consequence. Properly accounting for these factors improves both the power and accuracy of inference. Also important is an accurate background model.

Results: Here, we develop a Model-based Approach for identifying Driver Genes in Cancer (termed MADGiC) that incorporates both frequency and functional impact criteria and accommodates a number of factors to improve the background model. Simulation studies demonstrate advantages of the approach, including a substantial increase in power over competing methods. Further advantages are illustrated in an analysis of ovarian and lung cancer data from The Cancer Genome Atlas (TCGA) project.

Availability and implementation: R code to implement this method is available at http://www.biostat.wisc.edu/ kendzior/MADGiC/.

Contact: kendzior@biostat.wisc.edu

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

1 Introduction

Cancer is thought to result from the accumulation of causal somatic mutations throughout the lifetime of an individual. These cancer-driving mutations function by altering one of three broad classes of genes: oncogenes, which activate neoplastic activity; tumor-suppressor genes, which decrease a cell's ability to inhibit abnormal cell proliferation and stability genes, which affect a cell's damage repair mechanisms (Kinzler and Vogelstein, 1997; Vogelstein and Kinzler, 2004). A first causal mutation in one of these classes of genes (or a rate-limiting combination thereof) leads to tumorigenesis, and subsequent causal mutational events drive tumor progression by providing a selective advantage to the cancer cells through positive selection (Bozic *et al.*, 2010; Vogelstein *et al.*, 2013; Vogelstein and Kinzler, 2004; Wood *et al.*, 2007).

A major area of cancer research revolves around identifying these causal mutations, as doing so may provide new insights into gene function as well as potential targets for drug development. Methods for distinguishing genes with causal mutations ('driver genes') from those containing only background mutations ('passenger genes') which are irrelevant to cancer growth are also vital in making sense of the vast amounts of information being gathered from tumor sequencing studies such as The Cancer Genome Atlas project (http://cancergenome.nih.gov/) and the Cancer Genome Project (http://www.sanger.ac.uk/research/projects/cancergenome/).

A common approach to this problem is to identify genes that harbor significantly more somatic mutations than expected by chance. Methods using this approach, termed 'frequency-based' methods, rely on an estimate of a background mutation rate which represents the rate of random passenger mutations. Early frequency-based methods assumed a single background rate, constant across the genome and common to all samples (Ding *et al.*, 2008). However, a number of features are known to affect mutation rate: mutation type (transition versus transversion), nucleotide context (which base is at the mutation site), dinucleotide context (which bases are located at neighboring sites to the mutation), replication timing of the region and expression level of the gene. Further details are provided in Section 2.3.

In an effort to get a more accurate estimate of the background mutation rate, subsequent frequency-based methods have been developed that adjust for one or more of these factors. Sjoblom *et al.* (2006) account for nucleotide and dinucleotide context in searching for drivers of breast and colorectal cancer. MuSiC (Dees *et al.*, 2012) accounts for mutation type and allows for sample-specific mutation rates; and in addition to these factors, Lawrence *et al.* (2013) (MutSigCV) also allow for the inclusion of gene-specific factors such as expression level and replication timing.

Although useful, a main limitation of methods based solely on mutation frequency is the inherent assumption that driver genes have relatively high mutation rates. This is often not the case. Indeed, with a few notable exceptions such as TP53 and KRAS, which show consistently high mutation rates in many cancers, most driver genes harbor surprisingly few mutations (Vogelstein *et al.*, 2013; Wood *et al.*, 2007). Consequently, additional criteria need to be incorporated into the search beyond frequency if reliable driver gene identifications are to be made.

Recent developments provide at least two new sources for such information. The first are methods such as Sorting Intolerant From Tolerant (SIFT) [first reported by Ng and Henikoff (2001), later updated by Kumar *et al.* (2009)], Polyphen (Adzhubei *et al.*, 2010) and MutationAssessor (Reva *et al.*, 2011) that incorporate information from sequence context, position and protein characteristics to assess a mutation's functional impact. Recognizing the advantage of prioritizing genes by functional impact information, Gonzalez-Perez and Lopez-Bigas (2012) exploited bias in these scores as evidence of driver activity in their method OncodriveFM.

To account for both frequency and function, Youn and Simon (2011) (referred to hereinafter as YS) model mutation type, account for sample-specific mutation rates and incorporate BLOSUM80 (BLOcks Substitution Matrix) alignment scores (Henikoff and Henikoff, 1992) into their approach. BLOSUM80 alignment scores reflect empirical probabilities associated with amino acid substitutions; and YS use these scores as a measure of functional impact. The idea is that if an amino acid substitution is rarely seen, it is likely detrimental. Although useful, power and specificity is gained by using methods such as those mentioned above that directly assess functional impact specific to the gene and mutation of interest (Ng and Henikoff, 2001).

In addition to advances regarding our ability to assess functional impact at the single nucleotide level, major advances have also been made with respect to our understanding of the spatial pattern of mutations within driver genes. Indeed, Vogelstein *et al.* (2013) recently noted that the best way to identify driver genes is not through their mutation frequency as has often been done in the past, but rather through their spatial patterns of mutation. Vogelstein's claim is based on the recognition that oncogenes are often mutated recurrently at the same amino acid positions while tumor suppressor genes tend to have an over-abundance of truncating mutations (frameshift indels, non-sense mutations or mutations at the normal stop codon). These characteristic patterns were not known just a few years ago, since they only become apparent with very large sample sizes. For example, even when looking at a dataset with close to 500 samples such as the TCGA ovarian, spatial patterning of mutations is not obvious (see Fig. 1, left panel).

Recognition of such spatial patterns has been facilitated largely by a project to catalogue somatic mutations in cancer (Forbes *et al.*, 2011). The so-called COSMIC (Catalogue Of Somatic Mutations In Cancer) project was initiated by the NIH in 2004 and is ongoing, with new datasets being added several times per year. The database currently contains mutation information for close to one million samples in over 40 tissue types, including data from several thousand whole exomes. Recent results from an integrative analysis across multiple cancers in COSMIC identified 'highly characteristic and non-random' patterns of mutation that were not apparent when studying cancers by type in isolation (Vogelstein *et al.*, 2013). In particular, results demonstrated that many known oncogenes consistently harbor mutations at relatively few specific amino acid positions, suggesting that oncogenic activity does not result from random mutation(s) in an oncogene, but rather requires a mutation in one of a few locations. OncodriveCLUST (Tamborero *et al.*, 2013) was designed to exploit such evidence of positional clustering to identify oncogenes (but like OncodriveFM does not utilize other sources of information such as frequency of mutation and functional impact). Non-random mutational patterns are also observed in known tumor suppressor genes, which tend to exhibit an over-abundance of protein-truncating alterations. The right panel of Figure 1 provides a few examples. As we demonstrate, accounting for these non-random spatial patterns and abundance of truncating mutations improves both the sensitivity and specificity with which driver genes may be identified.

In addition to these recently characterized patterns of mutation, it is well known that alteration of DNA repair genes such as BRCA1 or BRCA2 leads to an increased accumulation of mutations (Birkbak *et al.*, 2013). For those samples with mutations in known DNA repair genes, any global increase in mutation rate will be accommodated by our model's sample-specific background rate estimation (see Section 2.3.2).

In summary, the most important sources of information to consider when identifying driver genes include: mutation frequency, mutation type, gene-specific features such as replication timing and expression level that are known to affect background rates of mutation, mutation-specific scores that assess functional impact and the spatial patterning of mutations that only becomes apparent when thousands of samples are considered. Previously developed methods incorporate many of these features (see Table 1 for an overview), but not all at once. In this paper, we provide a unified empirical Bayesian Model-based Approach for identifying Driver Genes in Cancer (MADGiC) that utilizes each of these features. The Bayesian framework provides a natural way to leverage the mutational patterns observed in COSMIC as prior information and provides gene-specific posterior probabilities of driver gene activity. The posterior probabilities are informed by mutation frequency relative to a background model that incorporates mutation type and the gene-specific features mentioned above as well as position specific functional impact scores. Results from a simulation study in Section 3.1 suggest improved performance over currently available methods. Further advantages are demonstrated in an analysis of data from the Cancer Genome Atlas (TCGA) project (Section 3.2).

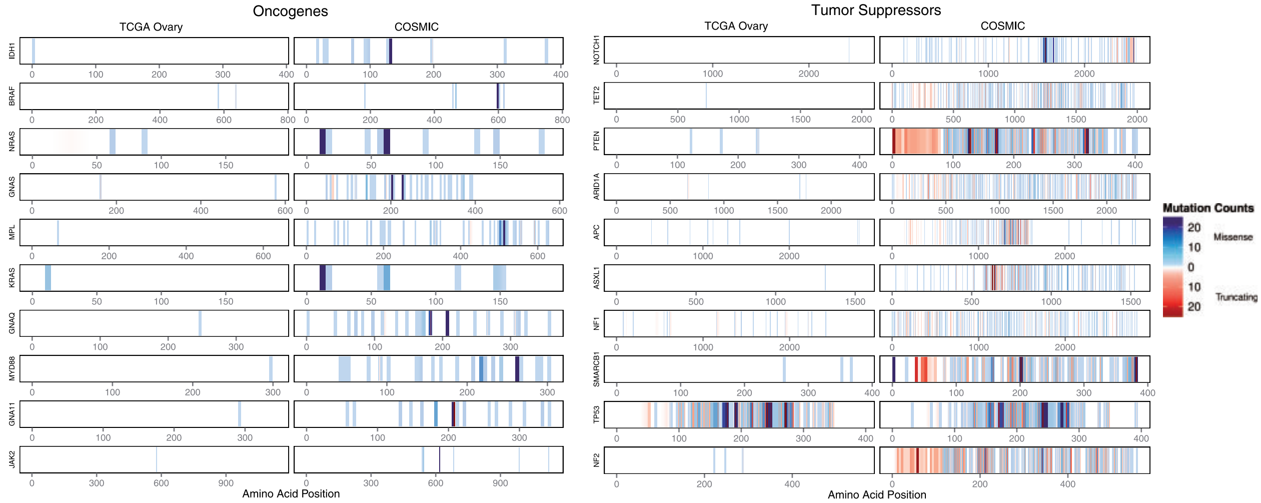

Fig. 1. Counts of samples with mutation by position and type for TCGA ovarian and COSMIC (Catalogue Of Somatic Mutations In Cancer) datasets. The left panel displays the ten genes with the lowest entropy in COSMIC (putative oncogenes) that have at least one mutation in TCGA ovarian. The right panel displays the ten genes with the highest proportion of truncating mutations (putative tumor-suppressor genes) that have at least one mutation in TCGA ovarian. Blue represents missense mutations and red represents a location with at least one truncating mutation. Each vertical bar spans five amino acids and darker colors correspond to more mutations. For genes with more than 500 mutations, a random sample of 500 was plotted, and positions with more than 25 mutations are given the same color intensity as those with 25 mutations

Table 1. Summary of features of methods to identify driver genes

Methods	Mutation Type	Frequency	Gene-specific Background	Functional Impact	Spatial Patterning
MADGiC	✓	✓	✓	✓	✓
MuSiC	✓	✓			
YS	✓	✓		✓	
MutSigCV	✓	✓	✓		
OncodriveFM	✓			✓	
OncodriveCLUST	✓				✓

2 Methods

2.1 TCGA somatic mutation data

The TCGA somatic mutation datasets consist of exome somatic mutation calls between tumor tissue samples and normal samples (from either matched tissue or blood) of cancer patients and are freely available for download from the TCGA data download portal (available at https://tcga-data.nci.nih.gov/tcga/). Each somatic mutation is annotated for the sample(s) in which it occurs, its chromosome and position, the gene in which it is located, the allele found in the reference genome, the specific nucleotide change(s) and the type of mutation (silent, missense, non-sense, frameshift indel, in frame indel). The analysis presented here includes all available ovarian and squamous cell lung cancer samples as of October 1, 2013.

2.1.1 Ovarian cancer

In the collection of 463 ovarian cancer samples, there are 5849 silent mutations (mutations that do not alter the amino acid sequence) located in 4369 genes and 21 800 non-silent mutations (mutations that cause a change in the amino acid sequence) located in 10 164 genes. The median (range) total number of mutations per sample is 60 (1–209). For silent mutations, the median (range) is 11 (0–51) and 41 (0–175) for non-silent mutations. There is very little positional overlap of mutations across samples and only 62 genes have a nonsilent mutation in more than 10 samples.

2.1.2 Squamous cell lung carcinoma

In the collection of 178 squamous cell lung cancer samples, there are 15 883 silent mutations located in 8191 genes and 49 418 non-silent mutations located in 13 238 genes. The median (range) total number of mutations per sample is 299.5 (4–3922). For silent mutations, the median (range) is 71.5 (0–1374) and 229 (3–2548) for non-silent mutations. There is very little positional overlap of mutations across samples, but 649 genes have a non-silent mutation in more than 10 samples.

The vast majority of squamous cell lung cancer cases are attributed to cigarette smoking (Kenfield *et al.*, 2008). Since cigarette smoking is a known mutagen that results in an increased mutation rate as well as characteristic mutation signatures (Pleasance *et al.*, 2009), it is plausible that the driver genes may differ between smokers and non-smokers because they are subject to different mutational processes. This is problematic since most methods assume there exists a common set of driver genes. To minimize the possibility of including non-smoking-related cancer cases in the analysis, samples with a mutation rate below the 5th percentile that were also recorded as current or lifelong non-smokers at the time of data collection were excluded. This resulted in the removal of 10 samples.

2.1.3 Simulated data

To facilitate comparisons with existing methods, two types of simulations were considered. For SIM I simulations, 100 sets of random passenger mutations were obtained by shuffling the observed mutations from the TCGA ovarian dataset while preserving nucleotide context and mutation type, but ignoring gene-specific factors that affect mutation rate such as replication timing and expression level (see Fig. 2 and Section 2.3 for details of these gene-specific factors). Each mutation in a given sample was randomly assigned a new position, drawn from all possible positions with the same reference nucleotide and mutation type. Next, 100 sets of 30 driver genes were randomly selected from the Cancer Gene Census [a set of nearly 500 genes that have been implicated in some form of cancer, manually curated by Futreal *et al.* (2004)] and non-silent mutations were randomly added at three levels: either 3, 5 or 10 mutations (total across all samples; 10 genes at each level). The choice of 30 driver genes was made to be on the order of the median number of genes identified as drivers in the case studies. This resulted in a total of 100 unique simulated datasets. One hundred sets of random passenger mutations were obtained for SIM II in a similar way, but accommodating the dependence of mutation rate on replication timing and expression level. Specifically, in this case each mutation in a given sample was randomly assigned a new position, drawn from all possible positions with the same reference nucleotide and mutation type in the same replication timing and gene expression categories. As in SIM I, 100 sets of 30 driver genes were randomly selected from the Cancer Gene Census and non-silent mutations were randomly added at three levels.

This same process was repeated to generate 100 SIM I and 100 SIM II datasets using the TCGA lung data since it was suspected that some sample characteristics may influence the ability to detect driver genes. In particular, the lung dataset differs from the ovarian in that it has less than half the number of samples but more than twice the number of somatic mutations. In addition, the sample-specific mutation rates are much more heterogeneous in the lung dataset compared to the ovarian. This can be seen in the ranges of detected mutations per sample reported above. Note that the absolute number of mutations in the true driver genes is the same for both simulation sets, and consequently the relative mutation rate for driver genes in the simulated ovarian data is higher than that in the simulated lung data.

2.2 Driver gene model framework

Our primary aim is to prioritize genes that have been somatically mutated in cancer based on the likelihood that they are driver genes. A driver gene is defined as a gene harboring a mutation that provides a selective advantage to the cancer cell. The empirical Bayesian hierarchical mixture model framework we develop considers three sources of evidence for driver activity: (i) increased frequency of

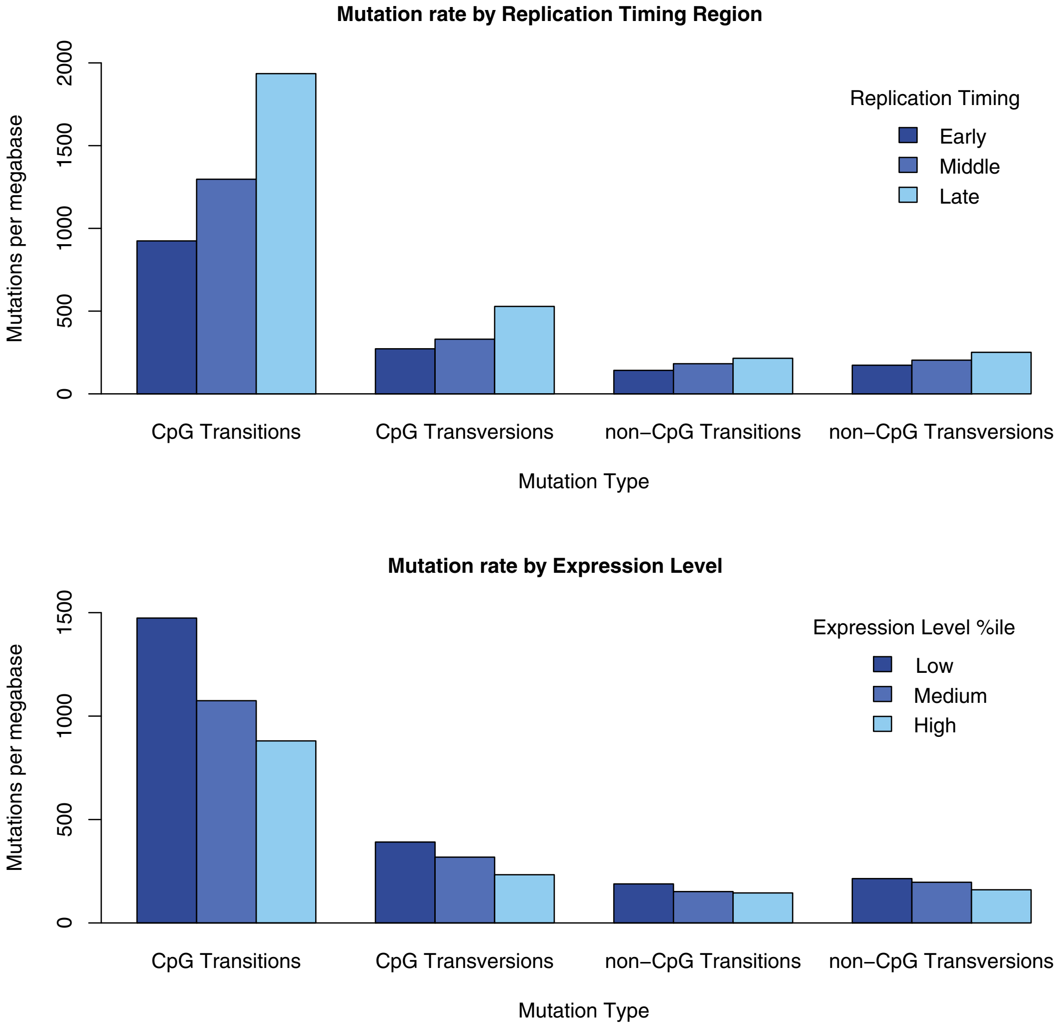

Fig. 2. Mutation rate is shown to depend significantly on replication timing region and expression level. Specifically, mutation rate is shown for three replication timing regions (top) and for three levels of expression (bottom) for four types of mutations in TCGA ovarian data. Within each mutation type, Chi-Square tests of mutation counts stratified by replication timing or expression level categories were found to be significant ($P < 0.05$)

mutation compared to a gene-specific background mutation model, (ii) evidence of functional impact and (iii) a non-random spatial pattern of mutations. We detail the generative model framework in Section 2.2.1 and the calculation of the posterior probabilities in Section 2.2.2. Parameter estimation is discussed in Section 2.2.3 and the use of spatial pattern data to inform the prior probability of oncogenic activity is described in Section 2.2.4.

2.2.1 Generative model

Consider a single gene indexed by g, from a total of G genes having at least one non-silent somatic mutation. Note that non-silent mutations include missense mutations, frameshift indels and in frame indels. Further consider an independent sample of size J, indexed by j, each with at least one non-silent somatic mutation in one or more of the G genes. Let $\vec{X}_g = X_{1g}, \ldots, X_{Jg}$ be the vector of observed non-silent mutation states of gene g for all samples (where $X_{jg} \in \{0, 1\}$ and $1 =$ one or more non-silent mutations anywhere in the gene; $0 =$ no mutations in the gene). Next, let $\vec{S}_g = S_{1g}, \ldots, S_{Jg}$ be the vector of functional impact scores for mutations in gene g for all samples. Finally let $Z_g \in \{0, 1\}$ be the indicator that gene g exhibits driver activity.

We are interested in the posterior probability that gene g is a driver gene given the mutation status and impact score for that gene across J independent samples:

$$P(Z_g = 1|\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}) = 1$$

$$= \frac{P(Z_g = 1)\prod_{j=1}^{J} P(S_{jg} = s_j, X_{jg} = x_j|Z_g = 1)}{\sum_{k\in\{0,1\}} P(Z_g = k)\prod_{j=1}^{J} P(S_{jg} = s_j, X_{jg} = x_j|Z_g = k)} \qquad (1)$$

We assume that the presence of mutations in gene g and sample j depends on driver status. Specifically, $X_{jg}|\,Z_g = z \sim Bern((1 - z)b_{jg} + zd_g)$ where $b_{jg} \in (0, 1)$ is the background (passenger) mutation probability for sample j, gene g and $d_g \in (0, 1)$ is the driver mutation probability for gene g. To enforce that the driver mutation probability is at least as high as the average passenger mutation probability (i.e. that $d_g > \overline{b}_{.g}$), we let $d_g \sim Beta(\alpha, \beta)$ truncated below at $\overline{b}_{.g}$.

Likewise, we assume that the distribution of functional impact scores across all genes and all samples depends on driver status. Specifically, $S_{jg}|X_{jg} = 1, Z_g = z \sim (1 - z)f^p + zf^d$, where f^p is the distribution of functional impact scores for passenger genes and f^d is the distribution of functional impact scores for driver genes. Note that we are assuming a common functional impact score profile for all driver mutations, and another for all passenger mutations, independent of mutation frequency.

2.2.2 Likelihood and posterior calculations

For J independent samples with observed mutation states $\vec{x}$ and scores $\vec{s}$, the data likelihood for gene g given driver status Z_g, driver mutation probability d_g and estimates $\hat{b}_{jg}$, $\hat{f}^p$, $\hat{f}^d$ is

$$P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = z, d_g = \delta)$$

$$= \prod_{j=1}^{J} P(S_{jg} = s_j|X_{jg} = x_j, Z_g = z)P(X_{jg} = x_j|Z_g = z, d_g = \delta)$$

$$= \delta^{z\sum_{j=1}^{J} x_j}(1 - \delta)^{z(J - \sum_{j=1}^{J} x_j)}$$

$$\times \prod_{j=1}^{J} \hat{f}^d(s_j)^{x_j z}(\hat{b}_{jg}\hat{f}^p(s_j))^{x_j(1-z)}(1 - \hat{b}_{jg})^{(1-z)(1-x_j)}$$

Note that this probability depends on d_g, which is unknown. Thus, we calculate the prior predictive distributions $P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = 1)$ and $P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = 0)$ by averaging over the prior distribution of d_g. Then,

$$P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = 1)$$

$$= \frac{B(\alpha^*, \beta^*)[1 - F_{(\alpha^*,\beta^*)}(\overline{b}_{.g})]}{B(\alpha, \beta)[1 - F_{(\alpha,\beta)}(\overline{b}_{.g})]}\prod_{j=1}^{J}\hat{f}^d(s_j)^{x_j}$$

$$P(\vec{S}_g = \vec{s}, \vec{X}_g = \vec{x}|Z_g = 0) = \prod_{j=1}^{J}(\hat{f}^p(s_j)\hat{b}_{jg})^{x_j}(1 - \hat{b}_{jg})^{1-x_j}$$

where $F_{(\alpha,\beta)}$ is the cumulative distribution function of the beta distribution with shape parameters (α, β); B is the Beta function; $\alpha^* = \sum_{j=1}^{J} x_j + \alpha$; and $\beta^* = J - \sum_{j=1}^{J} x_j + \beta$. Then the final form of the posterior probability is easily obtained from Equation (1).

2.2.3 Parameter estimation

We use the background mutation model that will be described in Section 2.3 to get an empirical Bayes estimate of b_{jg}. Recall that the global hyperparameters α and β govern the prior probability that a driver gene is mutated and consequently they are estimated using the method of moments from tissue-specific mutation data of known cancer genes [from the Cancer Gene Census (Futreal *et al.*, 2004)] in COSMIC. To avoid overfitting the model, any samples included in a dataset of interest should be removed prior to estimation of the hyperparameters. Here, e.g. TCGA ovarian and lung samples were removed; see Supplementary Section S2.2.4 for details.

To assess functional impact, we use SIFT scores from Liu *et al.* (2011), which range from zero to one, transformed such that scores closer to one represent high impact (Kumar *et al.*, 2009). If there is more than one non-silent mutation in gene g for sample j, we let S_{jg} take the value of the maximum functional impact score for all mutations in the gene. If there are no non-silent mutations in gene g for sample j, we let $S_{jg} = -1$. To estimate $f^d(\cdot)$, the distribution of functional impact scores for driver genes, we first obtain SIFT scores for a random sample of non-silent mutations, generated by shuffling the observed mutations subject to the constraints of the background mutation model. We then estimate $f^d(\cdot)$ using non-parametric spline regression on the ratio of the simulated null to the observed full distribution $f(\cdot)$ of scores across bins of the score range, a technique used by Efron *et al.* (2001) to estimate the non-null distribution of z-scores in the analysis of gene expression microarray experiments. Specifically, 50 equally spaced bins and a natural spline with 5 degrees of freedom were used. Though our functional impact score of choice here is SIFT, this non-parametric approach accommodates other available functional impact scoring schemes.

2.2.4 Quantifying gene-specific mutation patterns

Motivated by the fact that genes showing a random pattern of mutations across cancers in COSMIC are less likely to be drivers than those showing concentrated mutations (more likely to be oncogenes) or those showing an overabundance of protein-truncating mutations (more likely to be tumor suppressors) (Vogelstein *et al.*, 2013), for every gene g we calculate a prior probability of driver activity ($P(Z_g = 1)$) using all mutations observed for that gene in COSMIC (excluding TCGA ovarian and lung cancer cases and only including data from whole-gene screens). Specifically, to quantify evidence of concentrated mutations, for each gene we calculate its positional entropy compared to a random distribution of mutations across all amino acids. Genes with low entropy are ones with highly concentrated mutations. Similarly, we test each gene to see if it has a significantly higher proportion of truncating mutations than the proportion of truncating mutations observed over all genes. Genes with significantly low entropy or a significantly high proportion of truncating mutations ($P < 0.05$) are assigned a higher prior probability of oncogenic activity ($P(Z_g = 1) = 0.5$); otherwise $P(Z_g = 1) = 0.01$. Examples are shown in Figure 1; see Supplementary Section S2.2.5 for further details. The specific values of 0.5 and 0.01 are arbitrary, but empirical sensitivity analyses (see Supplementary Section S3) demonstrated little variability in results for values between 0.25 and 0.75 and between 0.005 and 0.05, respectively.

2.3 Background mutation model

We build on the YS background model and extend it to incorporate external information that has been shown to affect mutation rates, namely replication timing and expression level. Substantial variation in somatic mutation rates, up to 33% in normal and 60% in cancer cells, has been attributed to variation in the replication timing of DNA (Koren *et al.*, 2012; Woo and Li, 2012). In short, regions that replicate later have higher mutation rates due to the decreased amount of time the replication machinery has to perform repairs compared to earlier replicating regions (Pleasance *et al.*, 2009).

Figure 2 (top) shows this effect in the TCGA ovarian dataset. As shown, the pattern persists when looking only at a specific mutation type (transitions versus transversions) and nucleotide context (CpG versus non-CpG dinucleotide). Similar patterns are observed in the lung dataset (see Supplementary Fig. S1). Thus, it is not likely that the pattern can be explained by differences in rates of specific types of mutations across the regions. If this factor was to be ignored, then the background rate for late-replication regions would likely be underestimated, whereas the background rate for early-replicating regions would be overestimated.

Variation in mutation rate has also been observed with gene expression level. Specifically, Chapman *et al.* (2011) discovered that there are fewer mutations observed in genes that are expressed at a higher level on average in cancer cells. It is thought that transcription-coupled repair mechanisms are responsible for this effect. As in the case of replication timing, the differences in mutation rate by expression level remain largely consistent within mutation type and nucleotide context. This can be seen in Figure 2 (bottom), where the mutation rate is plotted for three gene expression level categories. As with replication timing, if this factor is ignored, the background rate for lowly expressed genes will be underestimated.

These two gene-specific factors explain additional variation in mutation rate beyond that contributed by the position-specific factors of mutation type and nucleotide context. However, some genes still have an inexplicably high mutation rate even after accounting for all the previously mentioned factors. Notably, the class of genes known as olfactory receptors has a near 2-fold increase in mutation rate compared to genes with similar replication timing and expression levels in the two TCGA datasets examined (see Supplementary Fig. S2). Here we classify genes as olfactory receptors using gene symbols to obtain a set of size 323 genes (see Supplementary Section S2.1 for details). While it is unclear why these genes have elevated rates of somatic mutation, they are known to exhibit substantial genetic diversity in terms of both single nucleotide polymorphisms and copy number variation (Hasin *et al.*, 2008). Consequently, the background model adjusts for the expected increase in the number of background mutations for this class of genes.

2.3.1 Adjusting for gene-specific factors

In order to incorporate the gene-specific factors of replication timing and expression level into the background mutation model, external estimates of replication timing were first obtained from Chen *et al.* (2010), who sequenced the DNA from HeLa cell lines at various stages of the synthesis phase of the cell cycle and provided timing estimates over 100 kb windows across the entire genome. As a robust proxy for replication timing, we divided the genome into three equal parts: (i) Early, (ii) Middle and (iii) Late replicating regions by splitting on the tertiles of the observed distribution. This is desirable since replication timing is not perfectly correlated across cell types, and we do not have ovarian cell line data. However, we note that the implementation of MADGiC is flexible enough to accommodate other sources of replication timing data.

Next, average expression levels of each gene were obtained from the 91 cell lines in the Cancer Cell Line Encyclopedia (CCLE) database with RNA-seq data (Barretina *et al.*, 2012), and genes were divided into tertiles of expression. Averages across many tissue types in the CCLE were used rather than matched expression measurements from TCGA since the pattern of decreased mutations with increased expression was more stable within mutation type, and because this same set of expression data could be used in studies of a different cancer or in studies where expression data was not available. However, as with replication timing data, if other sources of expression data are available, they may be specified in the MADGiC package.

Let λ_n, $n \in \{1, 2, 3\}$ be the relative rates of mutation for a position in replication timing category n, and let ε_h, $h \in \{1, 2, 3\}$ be the relative rates of mutation for a position in expression level category h. In addition, let δ be the relative rate of mutation for olfactory genes compared to all others. These parameters are incorporated into the background model of YS as additional multiplicative factors. They are in addition to the mutation-type and nucleotide context-specific rate parameters p_m, $m = 1, \ldots, 8$ defined in the YS model.

2.3.2 Parameter estimation

Recall that the relative rate parameter estimates $\hat{p}_m$, $\hat{\lambda}_n$, $\hat{\epsilon}_h$ and $\hat{\delta}$ determine background mutation rate probabilities and thus, ideally, they should be obtained by fitting the model only to genes containing silent mutations. Using all genes would mean that driver genes are included, which would violate our assumption that driver genes do not follow the background mutation model. However, because indels are non-silent, we also include genes that have at most one non-silent mutation. This introduces potential selection bias in the sample-specific mutation rates q_j so we follow YS and introduce another parameter r to account for the bias.

As in YS, we use the method of moments to estimate the relative rate parameters for mutation type p_m and selection bias r, as well as the additional parameters λ_n, ε_h and δ. We obtain empirical Bayes estimates of the sample-specific overall mutation rates q_j (by assigning the prior distribution of q_j to be Uniform(a, b) and estimating the posterior mean). The hyperparameters $(\hat{a}, \hat{b})$ are found via maximum likelihood estimation given relative rate parameters ($\hat{r}, \hat{\vec{p}}, \hat{\vec{\lambda}}, \hat{\vec{\epsilon}}$ and $\hat{\delta}$). In this way, the posterior distribution of q_j depends on the observed mutations in sample j, as well as the data-wide parameter estimates of the relative rates of the different types of mutations. Finally, the background probability b_{jg} that a gene g is mutated in sample j under the background model (i.e. given g is a passenger) is approximated by summing the probability of a background mutation across all base pairs in the gene. Note that this procedure is different from YS, who calculate b_{jg} as the expectation with respect to the posterior distribution of q_j; the resulting estimates of b_{jg} using YS are also empirical Bayes estimates and are very similar to the estimates obtained by our procedure, except that the former requires $J * G$ numerical integrations and the latter only J which provides considerable improvement in computation time.

2.4 Implementation and evaluation

In order to evaluate the utility of incorporating functional impact scores in the model, as well as assess what could be gained with a score that was better able to distinguish between passenger and driver mutations, MADGiC was evaluated under three different functional impact profiles: (i) ignoring functional impact, (ii) realistic impact—SIFT score profiles (see Supplementary Section S4 for details) and (iii) high impact—passenger scores drawn from Beta(1,1.5) and driver scores set equal to one. The last represents some idealistic functional impact (FI) scoring system in which the distributions of driver and passenger scores are well-separated (i.e. passenger mutations tend to have low functional impact and driver mutations always have high functional impact) and is designed to assess the upper bound for the amount of improvement than can be achieved by incorporating FI. The background model was fit as described in Section 2.3 and the posterior probabilities of each gene being a driver were computed as described in Section 2.2. Genes with posterior probability greater than 0.95 were classified as drivers.

For comparison, the frequency-based methods YS (original version) and MutSigCV (version 1.3) were also evaluated (YS evaluated for only 50 simulations due to computation time). Genes were classified as drivers by YS or MutSigCV if the Benjamini-Hochberg adjusted P value was less than 0.05. MuSiC was not evaluated since it requires a post-processing step to filter the output, for which general guidelines are not provided by Dees *et al.* (2012); and OncodriveFM and OncodriveCLUST (using Intogen suite version 2.4.1) were only evaluated for the case study data since it is not possible to specify simulated SIFT scores for these approaches. For the Oncodrive methods, we considered genes with q-values less than 0.05.

While we can comment on the characteristic differences among the driver genes identified in the case studies, it should be noted that we do not have a list of 'true positive' driver genes for the ovarian or lung cancer data. As a proxy, we use the list of 125 genes identified as drivers by Vogelstein *et al.* (2013). Note that although some hyperparameters in MADGIC were estimated using COSMIC data (see Section 2.2.3), the TCGA ovarian and lung datasets were removed prior to estimation, and no information regarding the list of drivers in Vogelstein *et al.* (2013) was used. Further, a sensitivity analysis was conducted to examine the effect of the weight placed on COSMIC in assigning prior probabilities that a gene is a driver (see Supplementary Section S3).

3 Results

3.1 Application to simulated data

To facilitate comparisons with existing methods, the simulation study considers two types of simulations: SIM I simulations that ignore the dependence of mutation rate on replication timing and expression level and SIM II simulations that do not. Within each simulation setup, we evaluate the ability of MADGiC and competing methods to identify true driver genes in a scenario that mimics TCGA ovarian (with a relatively large sample size and average number of mutations) as well as one that mimics TCGA lung (relatively small sample size and large number of mutations). In addition, MADGiC was evaluated under three different functional impact settings in order to assess to what degree the inclusion of an FI score may result in increased power. As expected, performance depends on each of these characteristics.

As shown in Table 2, false discovery rate (FDR) is well controlled for all methods when mutation rate is assumed constant across replication timing region and expression level. In the more realistic SIM II setting, FDR increases for methods that do not accommodate this dependence. For the simulated lung data, FDRs are generally higher and power is generally lower for all methods. This is likely due in part to the higher passenger mutation rate relative to the true driver mutation rate as well as greater heterogeneity in sample-specific mutation rates.

When functional impact scores are able to separate driver mutations from passengers (the ideal FI case), MADGiC is very well powered to detect true driver genes and has a well-controlled FDR. In contrast, when no FI information is used, the power of MADGiC is decreased but is still highest among approaches using the ovarian-based simulations, with only moderate increases in FDR. In the lung-based simulations, YS has higher power than MADGiC, but the FDR is considerably inflated. Under the more intermediate FI setting that is based on observed SIFT score profiles, MADGiC has more power than when FI information is ignored, with comparable FDR. Thus, MADGiC performs best when FI scores are set to be near ideal, however, it still shows favorable performance when SIFT scores are used (and also when no FI is used).

3.2 Application to TCGA somatic mutation data

3.2.1 Ovarian cancer

MADGiC identified 19 genes with a posterior probability of being a driver greater than 0.95 (listed in Supplementary Section S5). Table 3 (top) displays the number of genes found by each method, along with the proportion of those found that were also on the list of putative drivers from Vogelstein *et al.* (2013) (the 'Putative Driver Rate'). YS identified 70 significant genes after adjustment for multiple comparisons, 14 of which were also found by our model. MutSigCV identified five significant genes after adjustment for multiple comparisons, two of which were identified by our model and YS. Six of the 21 genes identified by OncodriveFM and three of the 20 genes identified by OncodriveCLUST were also found by MADGiC. We note that 57.9% of the drivers identified by our model are contained in the list from Vogelstein *et al.* (2013), while

Table 2. Simulation results

					MADGiC		
			MutSigCV	YS	No FI	SIFT	Ideal FI
SIM I	Ovary	Power	0.05	0.30	0.42	0.51	0.86
		FDR	0.04	0.04	0.04	0.04	0.02
	Lung	Power	0.01	0.16	0.27	0.31	0.75
		FDR	0.07	0.08	0.07	0.06	0.03
SIM II	Ovary	Power	0.06	0.33	0.45	0.55	0.86
		FDR	0.02	0.32	0.08	0.09	0.04
	Lung	Power	0.02	0.36	0.30	0.34	0.77
		FDR	0.58	0.97	0.32	0.30	0.05

Note: Power and FDR averaged over 100 SIM I datasets, where dependence of mutation rate on replication timing and expression level is ignored and 100 SIM II datasets, where this dependence is preserved. The first set of simulations was designed to mimic TCGA ovarian data, which has a relatively large sample size, an average number of mutations and relatively little variability among sample-specific mutation rates; the second set is based on TCGA lung data, with smaller sample size, larger number of mutations and greater heterogeneity in sample-specific mutation rates.

the same figure is 12.9% for YS, 40.0% for MutSigCV, 38.1% for OncodriveFM and 25.0% for OncodriveCLUST. Of the five genes identified by MADGiC but not YS, four have five or fewer mutated samples and two of those are putative drivers.

Figure 3 displays the proportion of genes found by each method in each replication timing and expression level category. Here we see that MADGiC is not biased toward finding genes in the high background mutation categories (late replication timing or low expression) compared to the distribution of all genes. In contrast, YS finds the highest proportion of genes in the high mutation rate categories.

Table 3. Case study results

					Oncodrive	
		MADGiC	YS	MutSigCV	FM	CLUST
Ovary	Total found	19	70	5	21	20
	Put. driver fraction	0.579	0.129	0.400	0.381	0.250
Lung	Total found	47	585	7	85	55
	Put. driver fraction	0.213	0.019	0.571	0.153	0.145

Note: For each method applied to the two case studies (TCGA ovarian and lung), we report the total number of driver genes identified, along with the proportion of those found that are putative drivers [i.e. they are on the list identified by Vogelstein *et al.* (2013)].

3.2.2 Squamous cell lung carcinoma

Although the lung data set is structurally different than ovarian with a smaller sample size and much higher average mutation rate, the qualitative results from each method are similar. MADGiC identified 47 genes with a posterior probability of being a driver greater than 0.95 (listed in Supplementary Section S5). Table 3 (bottom) displays the number of genes found by each method, along with the proportion of those found that was also on the list of putative drivers from Vogelstein *et al.* (2013). YS identified 585 significant genes after adjustment for multiple comparisons, 45 of which were also found by our model. MutSigCV identified seven significant genes after adjustment for multiple comparisons, six of which were identified by MADGiC and YS. Eight of the 85 genes identified by OncodriveFM and seven of the 55 genes identified by OncodriveCLUST were also found by MADGiC. We note that 21.3% of the drivers identified by our model are contained in the list from Vogelstein *et al.* (2013), while the same figure is 1.9% for YS, 57.1% for MutSigCV, 15.3% for OncodriveFM and 14.5% for OncodriveCLUST. As in the ovarian case study, YS is biased toward identifying genes in the high background mutation rate categories (see Supplementary Section S5 for details). Specifically, of the 448 genes significant only by YS that also have complete replication timing and expression information, 400 (89%) are in either the late replicating region, the low expression category or both. In addition, only one of these additional genes was identified by Vogelstein *et al.* (2013). Of the two genes identified by MADGiC but not YS, one has five or fewer mutated samples and the other is a putative driver.

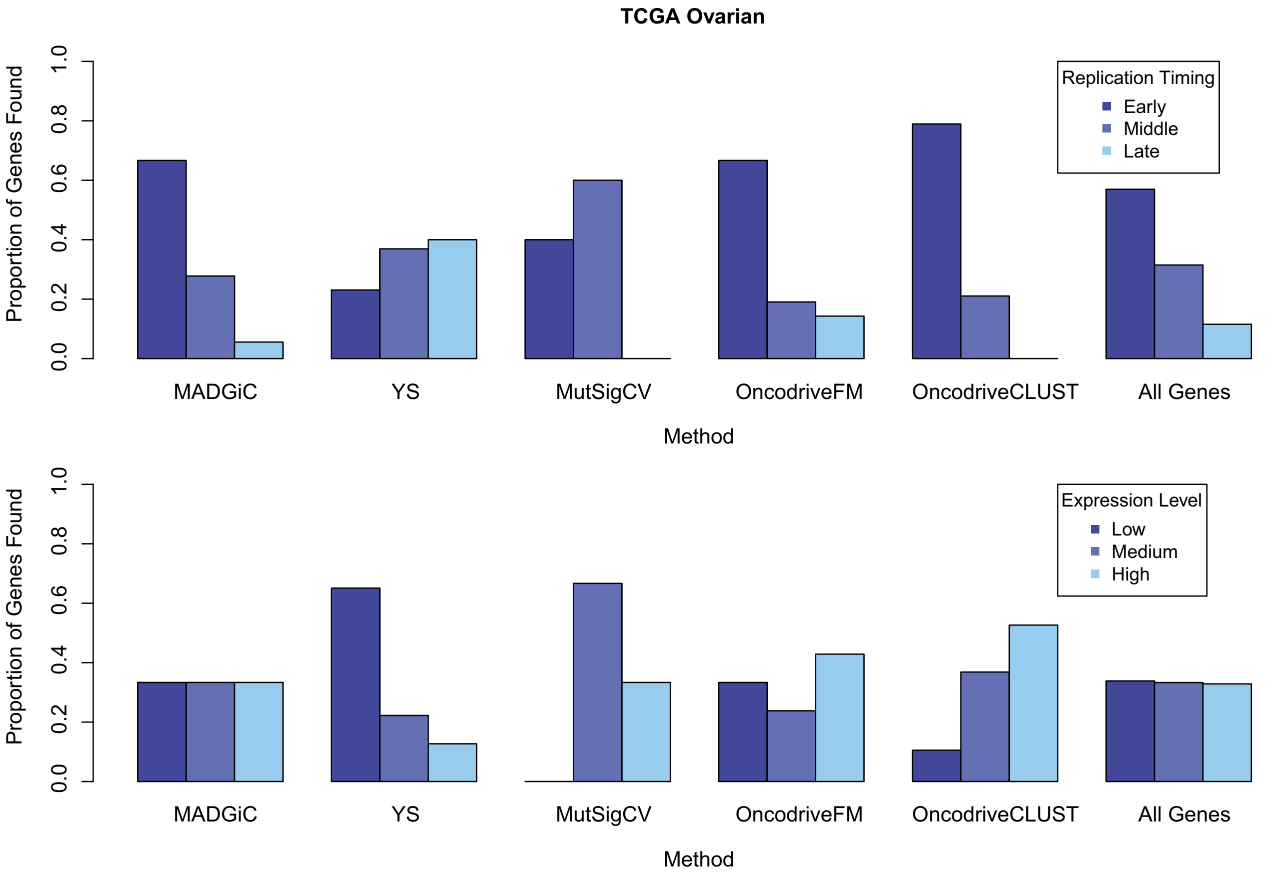

Fig. 3. The proportion of driver genes identified by each method in each replication timing (top) and expression level (bottom) category for the TCGA ovarian case study. Refer to Supplementary Section S5 for similar results from the lung case study

Note that the results presented here for MutSigCV are slightly different than those observed in Lawrence *et al.* (2013) since we have removed 10 samples, used a q-value threshold of 0.05 instead of 0.10 and used the most updated version of MutSigCV (see Sections 2.1.2 and 2.4 for details).

4 Discussion

MADGiC is an integrative model that provides posterior probabilities for improved inference for driver gene identification. The empirical Bayesian framework provides a natural way to incorporate several critical features together that were previously only considered in isolation. In addition to modeling key features of the observed mutation data, MADGiC also leverages the non-random mutational patterns observed across many cancer types in the COSMIC database to inform the prior probability of driver activity. Until recently, these spatial patterns were only evident in well-studied cancer genes that were the focus of targeted sequencing studies. Over the past few years, however, the COSMIC database has accumulated data from thousands of whole genomes and whole exomes, enabling a systematic search over all genes. The use of a database that collects mutation position data from multiple studies for each cancer type is vital, as the characteristic spatial patterns observed across thousands of cancers are not discernible when analyzing data from a single cancer in isolation.

The performance of MADGiC shows promise both in simulations and case studies. The simulation studies suggest that MADGiC has favorable operating characteristics relative to existing methods, and further highlights specifically the amount of advantage gained by incorporating functional impact scores. As the quality of these scores improves, so too should the power of MADGiC. In addition, the simulation study demonstrates that the operating characteristics of all approaches can vary widely with sample size, mutation frequency and heterogeneity in sample-specific mutation rates. It also demonstrates that MADGiC's integration of data across multiple sources facilitates the identification of putative driver genes showing relatively few mutations, a result also observed in the case studies. Specifically, as seen in Supplementary Tables S7 and S8, there are several genes with only three to five samples mutated that are identified as drivers by MADGiC but not other approaches. The fact that many of these are also on the putative driver list of Vogelstein *et al.* (2013) suggests that they are not false positives.

A limitation of all methods investigated stems from our assumption that the somatic mutation calls are complete and accurate. While it has been observed that properties of tumor samples (e.g. low allelic fraction) are responsible for introducing systematic sequencing bias, methods for improving the sensitivity of mutation callers have been developed (Yost *et al.*, 2013). As these methods continue to improve, so too will results from MADGiC. A further limitation of frequency-based methods that was noted in Lawrence *et al.* (2013) is the bias toward longer genes. Although MADGiC, YS and MutSigCV each account for gene length, the driver genes are still enriched for longer genes in all three methods in both case studies except for MutSigCV in the lung cancer study (see Supplementary Section S5 for details). However, MutSigCV is more conservative than the other two methods and the bias reappears as the gene list size increases. The bias is likely a result of additional, perhaps unknown factors that affect the rate of mutation of these longer genes. The fact that none of the methods are able to completely overcome this bias demonstrates that this is an ongoing challenge for frequency-based methods.

So far, we have only considered modeling one gene at a time. Thus, when computing the posterior probability that a given gene is a driver, no information pertaining to any other genes is considered, beyond that used to estimate the parameters in the background mutation model. However, a non-silent mutation in any one of a group of coordinately regulated genes (e.g. A, B and C) could cause the same selective advantage to a cancer cell. In this situation, evidence of driver activity of gene A would increase given non-silent mutations in genes B and C in other samples. A number of methods are available for identifying pathways containing driver genes (Ciriello *et al.*, 2012; Vaske *et al.*, 2010; Vandin *et al.*, 2012). Extensions of MADGiC to accommodate pathway structure should further improve our ability to identify drivers of cancer.

Acknowledgements

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/. We thank the anonymous reviewers for their comments and suggestions, which have served to improve the manuscript.

Funding

This work was supported by NIH [GM102756].

Conflict of Interest: none declared.

References

Adzhubei,I.A. *et al.* (2010). A method and server for predicting damaging missense mutations. *Nat. Methods,* **7**, 248–249.

Barretina,J. *et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature,* **483**, 603607.

Birkbak,N.J. *et al.* (2013). Tumor mutation burden forecasts outcome in ovarian cancer with brca1 or brca2 mutations. *PLoS One,* **8**, e80023.

Bozic,I. *et al.* (2010). Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA*, **107**, 18545–18550.

Chapman,M.A. *et al.* (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature,* **471**, 476–472.

Chen,C.-L. *et al.* (2010). Impact of replication timing on non-cpg and cpg substitution rates in mammalian genomes. *Genome Res.*, **20**, 447–457.

Ciriello,G. *et al.* (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.

Dees,N.D. *et al.* (2012). MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.

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

Efron,B. *et al.* (2001). Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Forbes,S.A. *et al.* (2011). COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**(Suppl. 1), D945–D950.

Futreal,P.A. *et al.* (2004). A census of human cancer genes. *Nat. Rev. Cancer,* **4**, 177–183.

Gonzalez-Perez,A. and Lopez-Bigas,N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.

Hasin,Y. *et al.* (2008). High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet.*, **4**, e1000249.

Henikoff,S. and Henikoff,J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Kenfield,S.A. *et al.* (2008). Comparison of aspects of smoking among the four histological types of lung cancer. *Tob. Control,* **17**, 198–204.

Kinzler,K.W. and Vogelstein,B. (1997). Gatekeepers and caretakers. *Nature,* **386**, 761–763.

Koren,A. *et al.* (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.*, **91**, 1033–1040.

Kumar,P. *et al.* (2009). Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nat. Protoc.*, **4**, 1073–1081.

Lawrence,M.S. *et al.* (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* **499**, 214–218.

Liu,X. *et al.* (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.*, **32**, 894–899.

Ng,P.C. and Henikoff,S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

Pleasance,E.D. *et al.* (2009). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature,* **463**, 184–190.

Reva,B. *et al.* (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.*, **39**, e18.

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

Tamborero,D. *et al.* (2013). Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics,* **29**, 2238–2244.

Vandin,F. *et al.* (2012). De novo discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375–385.

Vaske,C.J. *et al.* (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics,* **26**, i237–i245.

Vogelstein,B. *et al.* (2013). Cancer genome landscapes. *Science,* **339**, 1546–1558.

Vogelstein,B. and Kinzler,K.W. (2004). Cancer genes and the pathways they control. *Nat. Med.*, **10**, 789–799.

Woo,Y.H. and Li,W.-H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.*, **3**, 1004.

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

Yost,S.E. *et al.* (2013). Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing. *Bioinformatics,* **29**, 1908–1909.

Youn,A. and Simon,R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics,* **27**, 175–181.