BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Using passenger mutations to estimate the timing of driver mutations and identify mutator alterations

Ahrim Youn and Richard Simon*

Abstract

Background: Recent developments in high-throughput genomic technologies make it possible to have a comprehensive view of genomic alterations in tumors on a whole genome scale. Only a small number of somatic alterations detected in tumor genomes are driver alterations which drive tumorigenesis. Most of the somatic alterations are passengers that are neutral to tumor cell selection. Although most research efforts are focused on analyzing driver alterations, the passenger alterations also provide valuable information about the history of tumor development.

Results: In this paper, we develop a method for estimating the age of the tumor lineage and the timing of the driver alterations based on the number of passenger alterations. This method also identifies mutator genes which increase genomic instability when they are altered and provides estimates of the increased rate of alterations caused by each mutator gene. We applied this method to copy number data and DNA sequencing data for ovarian and lung tumors. We identified well known mutators such as TP53, PRKDC, BRCA1/2 as well as new mutator candidates PPP2R2A and the chromosomal region 22q13.33. We found that most mutator genes alter early during tumorigenesis and were able to estimate the age of individual tumor lineage in cell generations.

Conclusions: This is the first computational method to identify mutator genes and to take into account the increase of the alteration rate by mutator genes, providing more accurate estimates of the tumor age and the timing of driver alterations.

Keywords: Probabilistic modeling of tumor development, Estimating the order of mutations during tumorigenesis, Identifying mutator genes

Background

Recent developments in high-throughput genomic technologies are providing a comprehensive view of genomic alterations in tumors, including DNA copy number changes and nucleotide mutations on a whole genome scale. Although a large number of somatic alterations are detected in tumor genomes, only a small number of those are considered driver alterations which drive clonal expansion and invasion. Most of the somatic alterations appear to be passengers that are neutral for tumor cell selection [1]. Currently, most research efforts are put into distinguishing and analyzing driver alterations although an in-depth understanding of the driver alterations in the early stages of tumorigenesis has not emerged for most cancer types.

The passenger alterations can provide valuable information about the tumor. The number of passenger somatic alterations accumulated in the tumor can provide information about the approximate age of the tumor lineage, which is the number of cell divisions in the dominant clone's lineage from the birth of the patient until the biopsy. Somatic alterations are acquired at each cell division with small probability, therefore, tumor samples which have undergone many cell divisions tend to accumulate many passenger alterations [2].

In addition to providing information on the age of the tumor lineage, passenger alterations can also give information about the approximate timing of the driver alterations occuring during tumorigenesis. Although the

*Correspondence: rsimon@mail.nih.gov
Biometric Research Branch, National Cancer Institute, Bethesda, Maryland, USA

driver alterations and their order of occurrence differ among tumors, elucidating this information can be important for understanding tumorigenesis. Tumors evolve through a sequence of somatic driver alterations [3]. Mutations occur randomly and are selected for in cellular clonal evolution. For example, during early tumorigenesis, mutations which confer growth advantage may be selected for, however, as the tumor expands, mutations which give advantage in the condition of cellular crowding and substrate limitations due to reduced blood flow will be selected [4]. Early mutations may represent important therapeutic targets because they occur in all tumor clones and late mutations may play important roles in metastasis. Due to the importance of understanding the temporal order of driver alterations during tumorigenesis, several computational methods have been developed to estimate this order [5-9], but no previous methods have used passenger alterations for that purpose.

If a driver alteration occurs late during tumorigenesis, it will be found mainly in tumors with a large number of passenger somatic alterations. If it occurs early, the number of passenger alterations should be smaller. One important caveat, however, is that the rate of formation and accumulation of new passenger alterations may increase during tumorigenesis.

The most frequently observed genomic instability is chromosome instability (CIN), which refers to a high rate of chromosome structure alteration in cancer cells. Another form of genomic instability is characterized by increased frequencies of nucleotide mutations. Microsatellite instability (MIN), which is a special case of this genomic instability is characterized by the expansion or contraction of the number of oligonucleotide repeats present in microsatellite sequences [10].

A higher rate of nucleotide mutations or chromosome alterations during tumorigenesis is caused by alterations in genes that maintain genomic stability. These so called mutator genes which increase genomic instability when altered, are involved in the processes of DNA sysnthesis and repair, chromosome segregation, damage surveillance, cell cycle checkpoints, and apoptosis [11-13].

Since alterations of mutator genes increase the rate of alterations, the samples in which mutator genes are altered tend to accumulate many passenger alterations. Therefore, if one does not take into account the increase of the rate of alterations due to mutator genes, the timing of the mutator gene alterations as well as the tumor age will be overestimated.

Here, we propose a method which estimates the age of the tumor lineage and the timing of the driver alterations from the number of passenger alterations. The alterations include point mutations, short insertions and deletions detected in sequencing data and copy number alterations detected in copy number data. This method also identifies mutator genes that induce increase of chromosome alteration or point mutations and estimates the increased rate of chromosome alterations or point mutations caused by the mutator gene during tumorigenesis.

In the Methods section, we introduce the data types to which this method can be applied and then describe the probability model used. We then present the results obtained from applying the method to ovarian cancer data and lung cancer data in the Results section.

Methods

Data types

Our method can be applied to sequencing data as well as copy number data. Sequencing data provide point mutations and short insertions or deletions (INDEL). Copy number data provide copy number alterations (CNA), either deletions or amplifications spanning a range of chromosomal regions. To apply our method, we first need to distinguish driver from passenger alterations.

For sequencing data, we first define driver genes as those which are more frequently mutated than expected by the background mutation rate. There are various methods to find driver genes and we use the method of Youn *et al.* [14] in this paper. We define the mutations detected in driver genes as driver alterations and those detected in non-driver genes as passenger alterations.

For copy number data, we use the segmented copy number obtained from the circular binary segmentation algorithm [15]. The circular binary segmentation algorithm splits the chromosomes of each sample into contiguous regions of constant copy number taking into account the noise in the data. We only consider the segmented regions whose value of the log_2 copy number change are larger than 1 or less than -1 as CNAs. Of these CNAs, we define driver and passenger CNAs as follows.

We define the CNAs which occur 10^6 base pairs away from each end of any GISTIC region (the chromosomal regions that are focally amplified or deleted recurrently, found by the algorithm GISTIC [16]) as passenger CNAs. If there are multiple passenger CNAs of the same type (amplification or deletion) that are close to each other (less than 10^5 base pairs), we merge them.

We define the CNAs which overlap the GISTIC region of the same type (amplification or deletion) for longer than two thirds of the region as driver CNAs. In other words, we say that a sample contains an amplified driver CNA associated with a given focally amplified GISTIC region if the amplified segment (segment whose log_2 copy number change is larger than 1) in the sample overlaps more than two thirds of the amplified GISTIC region. Similarly, a sample contains a deleted driver CNA associated with a given focally deleted GISTIC region if the deleted segment

(segment whose log_2 copy number change is less than -1) in the sample overlaps more than two thirds of the deleted GISTIC region.

Probability model

For each tumor sample i, we know the number of passenger somatic alterations N_i, the age of the patient S_i, whether an alteration occurred in driver gene/region j or not (denoted by $A_{i,j} = 1 \text{or } 0$) and whether it is germline or somatic (denoted by $G_{i,j} = 1 \text{or } 0$).

From these data, we want to infer when the driver gene/region j alters in sample i and if altered, how much it increases the alteration rate of other genes or regions. We also want to estimate the age of tumor lineage T_i. We define it as the number of cell divisions between the birth of the patient and the biopsy of the tumor in the lineage containing the founder cell of the dominant clone for sample i. We will use the Bayesian probabilitstic model defined below.

We model the accumulation of passenger somatic alterations in the lineage of tumor founder cell by a Poisson process. In the tumor cell lineage, we assume that new passenger alterations are acquired with rate λ at each cell division. Therefore, for the cell which has gone through T_i cell divisions, N_i follows Poisson distribution with rate λT_i if the alteration rate stays constant. (Figure 1(a)) In order to permit the increase of alterations by unknown factors such as exposure to mutagens by smoking or UV radiation, we add E_i and therefore, the number of passenger somatic alterations N_i follows a Poisson distribution with rate $\lambda T_i + E_i$

When a driver gene or region j is altered in sample i ($A_{ij} = 1$), we assume that it increases the alteration rate by Δ_j. It is positive if the gene/region is a mutator and 0 otherwise (Figure 1(b)). Suppose the alteration of the driver gene or region j occurred in sample i at time $X_{i,j}$. We assume the increase Δ_j is independent of when the driver j is altered. Then until the time $X_{i,j}$, the alteration rate per cell division is λ and after that, it becomes $\lambda + \Delta_j$. Therefore the number of passenger alterations follows a Poisson distribution with rate $\lambda X_{i,j} + (\lambda + \Delta_j)(T_i - X_{i,j}) + E_i = \lambda T_i + \Delta_j(T_i - X_{i,j}) + E_i$.

In general, when there are multiple driver genes/regions $j \in J$, each of which increases the alteration rate by Δ_j, the number of passenger somatic alterations follows a Poisson distribution with rate $\lambda T_i + \sum_{j \in J} \Delta_j(T_i - X_{i,j}) + E_i$ (The derivation of this is provided in the Additional file 1). This means that the alteration of each driver gene/region j increases the average number of passenger alterations accumulated in the sample by $\Delta_j(T_i - X_{i,j})$ additively. The value of $X_{i,j}$ is unknown, but the values of $A_{i,j}$ and $G_{i,j}$ give some information about $X_{i,j}$ since $A_{i,j} = 1$ implies $X_{i,j} \leqq T_i$ and $A_{i,j} = 0$ implies $X_{i,j} > T_i$. Also, $G_{i,j} = 1$ implies $X_{i,j} = 0$ since the alteration existed from the birth of the patient. $G_{i,j} = 0$ implies $X_{i,j} > 0$.

Since we cannot estimate $X_{i,j}$ and T_i for each sample separately, we use a Bayesian approach and assume a prior distribution for $X_{i,j}$ and T_i. We assume T_i follows a Gamma distribution with an unknown shape and rate parameter α, β. We restrict the range of values it can assume for each sample to be between 50 and the age of the patient S_i divided by the tumor cell division time r for

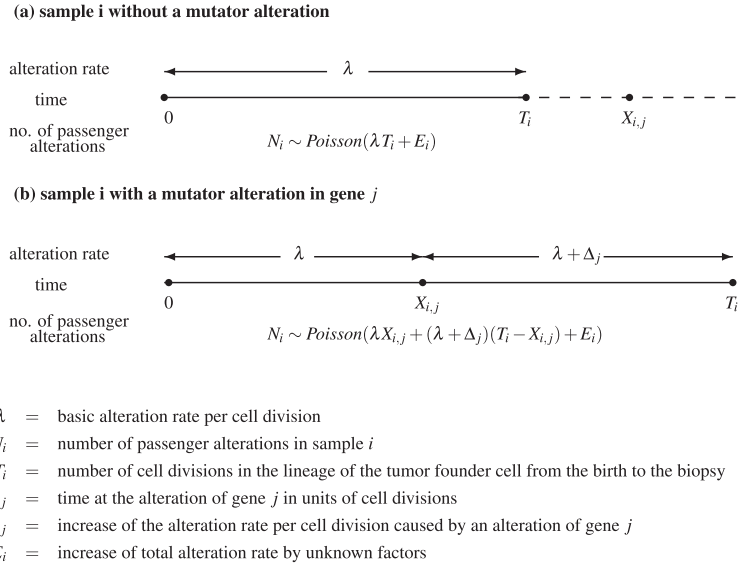

Figure 1 Probability model for generation of passenger somatic alterations. (a) for samples without mutator alterations, **(b)** for samples with a mutator alteration.

the specific tissue. This is because the number of cell divisions in the tumor lineage is unlikely to be less than 50 or larger than S_i/r since cell divides most frequently after the onset of neoplasia in the lineage of the founder cell.

We assume $X_{i,j}$ follows a Gamma distribution, however since it is possible that the alteration may never occur, we assume $\Pr(X_{i,j} = \infty) = p_j > 0$. When $0 < X_{i,j} < \infty$, we assume $X_{i,j}$ follows a Gamma distribution with shape and rate parameter α_j, β_j. We assume E_i follows an exponential distribution with a parameter ρ.

The rate of alteration λ differs for nucleotide mutations (point mutations and short INDELs) detected in sequencing data and CNAs detected in copy number data. The rate of nucleotide mutations per cell division λ^{MUT} is calculated using the experimentally obtained mutation rate per cell division and per base pair, 10^{-9} [17,18].

$$\lambda^{MUT} = 10^{-9} \times \text{number of base pairs sequenced for non-driver genes}$$

The rate of CNAs per cell division, λ^{CNA} is unknown. The ratio of λ^{MUT} vs. λ^{CNA} is,

$$R = \frac{\lambda^{CNA}}{\lambda^{MUT}} = \frac{\text{rate of CNAs per cell division}}{\text{rate of mutations per cell division}}$$
$$= \frac{\text{no. of CNAs/no. of cell divisions}}{\text{no. of mutations/no. of cell divisions}} = \frac{\text{no. of CNAs}}{\text{no. of mutations}}$$
$$\cong \frac{\text{average no. of passenger CNAs per sample}}{\text{average no. of passenger mutations per sample}}$$

Therefore, we estimate λ^{CNA} as $R \cdot \lambda^{MUT}$.

The unknown values of the parameters $\alpha, \beta, \alpha_j, \beta_j, \Delta_j, p_j, \rho$ are estimated by maximizing the likelihood of the observed data: the number of passenger somatic alterations N_i and occurrences of driver alterations j in sample i ($A_{i,j}$) given their germline status ($G_{i,j}$) and the age of the patient S_i.

For given values of the times $x_{i,j}$ of alterations of gene/region $j \in J$ and the age of the tumor lineage t_i, the number of passenger somatic alterations N_i in sample i would have a Poisson distribution with mean

$$\mu(x_{i,k}, t_i, e_i) = \sum_{k, A_{i,k}=1} \Delta_k(t_i - x_{i,k}) + \lambda t_i + e_i$$

Then, the likelihood of observing N_i and $A_{i,j}$ given their germline status $G_{i,j}$ and age of the patient S_i is obtained by integrating Poisson$(n_i; \mu(x_{i,k}, t_i, e_i))$ times probability density functions of $X_{i,j}$, T_i, E_i over the ranges of $X_{i,j}$, T_i and E_i corresponding to $A_{i,j} = a_{i,j}, G_{i,j} = g_{i,j}, \forall j$. When $G_{i,j} = 1, X_{i,j}$ is zero. When $G_{i,j} = 0, X_{i,j}$ is between 0 and T_i if $A_{i,j} = 1$; otherwise $X_{i,j}$ is larger than T_i. T_i takes values from 50 to the age of the patient i divided by the tumor cell division time r and E_i takes values from 0 to infinity. For the derivation of the likelihood function and the details of parameter estimation, see the Additional file 1.

This model can be applied to both sequencing and copy number data. For sequencing data, the number of passenger somatic mutations for sample i, N_i^{MUT} is assumed to follow a Poisson distribution with rate $\lambda^{MUT} T_i + \sum_{j \in J} \Delta_j^{MUT}(T_i - X_{i,j}) + E_i^{MUT}$ where λ^{MUT} is the basic nucleotide mutation rate and Δ_j^{MUT} is the increase of nucleotide mutation rate by the alteration of driver j. For copy number data, the number of passenger somatic CNAs, N_i^{CNA} is assumed to follow a Poisson distribution with rate $\lambda^{CNA} T_i + \sum_{j \in J} \Delta_j^{CNA}(T_i - X_{i,j}) + E_i^{CNA}$ where λ^{CNA} is the basic CNA rate and Δ_j^{CNA} is the increase of CNA rate by the alteration of driver j.

With these parameters, we can obtain the posterior mean of T_i and $X_{i,j}$ for each sample i given the data $N_i, A_{i,j}$ and $G_{i,j}$. Also, using the posterior mean of $X_{i,j}$, we can order the sequence of driver alterations which occurred for each sample i. In the Results section, we present the result obtained for ovarian and lung cancer data.

Results and discussion

Ovarian cancer data

We applied our method to the ovarian cancer data from The Cancer Genome Atlas (TCGA) [19], which analysed DNA copy number and whole exome sequences in 316 high-grade serous ovarian adenocarcinomas.

We first identified driver genes by applying the method of Youn *et al.* [14] to the whole exome sequencing data. We further select genes mutated in more than ten samples and obtained CSMD3, FAT3, NF1, TP53, USH2A, BRCA1 and BRCA2. The genes BRCA1 and BRCA2 have somatic mutations in 11 and 10 samples, but they have germline mutations in 27 and 20 samples, respectively.

Second, we used GISTIC to identify 63 regions of focal amplification and 50 regions of focal deletion from the copy number data. Although GISTIC identified 113 driver regions, we found that many of the regions show correlated pattern of alterations as shown in Figure 2. It is a heatmap of amplification patterns of focal amplification regions amplified in more than ten samples. Columns represent amplification regions and rows represent tumor samples. The yellow color indicates that the region is amplified in the corresponding tumor samples. Columns are sorted by their chromosome locations. Figure 2 shows that the amplification patterns of columns 9, 10, 12 are clustered around that of column 11. Although GISTIC found four separate regions, it seems that the amplifications in columns 9, 10, 12 are not separate events from the amplification in the column 11. The fact that column 11 contains a well known driver gene MYC while the other

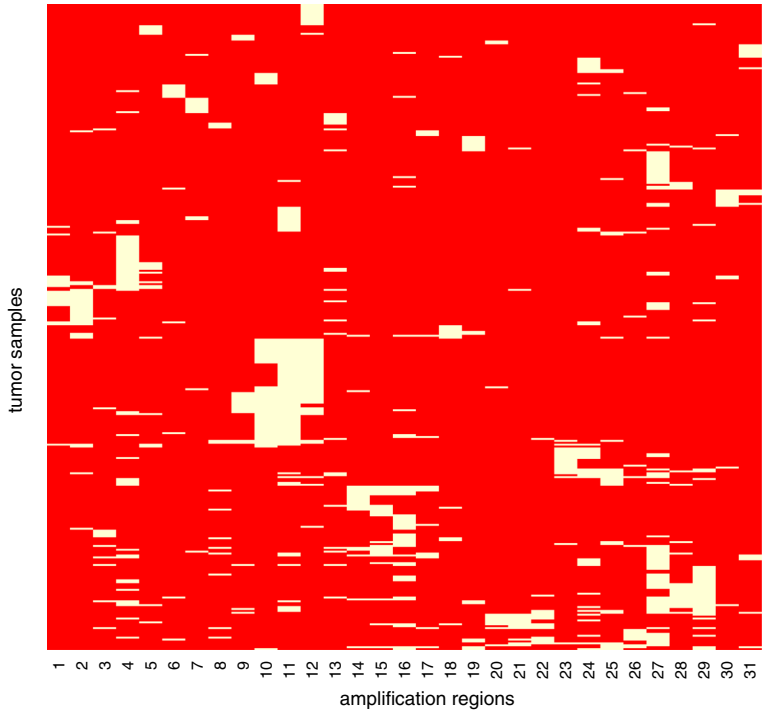

Figure 2 Heatmap of amplification patterns of the significantly amplified regions in ovarian cancer found by GISTIC. Columns represent amplified regions and rows represent tumor samples. The yellow color indicates that the region is amplified in the corresponding tumor samples. Columns are sorted by their chromosome locations.

columns 9, 10, 12 do not contain such genes also support this claim. Therefore, we removed such satellite chromosomal regions and also removed the regions altered in less than or equal to ten samples which leaves 14 driver regions.

Then, we applied our model to the selected driver genes/regions. For each sample i, we obtain the posterior mean of the age of tumor lineage T_i and the time of driver gene/region alteration $X_{i,j}$. In Additional file 2: Table S1, we present these values with their 90% confidence interval (CI) obtained by performing 400 bootstraps.

The average value of the posterior mean of tumor ages is 1113 cell divisions. The 10th percentile is 563 and the 90th percentile is 1839 cell divisions. We removed the gene TP53 from this analysis since it is mutated in almost all samples (95%) and with very few samples in which TP53 is not mutated, it is difficult to estimate the parameters for TP53 correctly. This may have caused the overestimation of the age of tumor lineage since we ignored the possible increase of the alteration rate by the mutation of TP53. Note that the estimated age of tumor lineage is inversely proportional to the alteration rate.

Identified mutators

We estimated the increase of mutation rate Δ_j^{MUT} and CNA rate Δ_j^{CNA} by the alteration of the gene/region j and also obtained their 90% CI from 400 bootstraps. The genes BRCA1, BRCA2 and the chromosomal region 16q23.1 are estimated to increase the mutation rate by 30%, 50% and 120%, respectively. However, only BRCA1 and BRCA2 have 90% CIs which do not include zero. Therefore, we can say only reliably that BRCA1 and BRCA2 genes increase mutation rate. They are well known mutator genes that play key roles in repairing double-strand breaks in DNA [20].

The chromosomal regions 8p21.2, 8q24.21, 16q23.1, 19q12, 22q13.33 are estimated to increase the CNA rate by 70%, 30%, 40%, 30% and 50%, respectively. Only the region 8p21.2 and 22q13.33 have 90% CI that do not include zero, implying only they increase CNA rate.

The region 8p21.2 (chromosome 8 between 26165916 bp and 26284094 bp) includes 12 genes, one of which is a tumor suppressor gene PPP2R2A. PPP2R2A is frequently deleted or downregulated in prostate, breast, lung and thyroid cancer [21]. Kalev *et al.* [22] recently revealed that PPP2R2A plays a critical role in double strand break repair through dephosphorylation of ATM. Moreover, they idenfied PPP2R2A as a novel predictive marker for the efficiency of treatment with PARP inhibitors.

The region 22q13.33 (chromosome 22 between 49481137 bp and 49498777 bp) is the most significantly deleted regions of all regions found by GISTIC and all

alterations involving this region were telomere loss. The loss of 22q13.33 is the cause of Phelan-McDermid Syndrome characterized by global developmental delay, absent or severely delayed speech, and normal to accelerated growth [23]. Although the role of the deletion of this region in tumorigenesis is not known, telomere loss in general is observed frequently in cancer cells and it is suggested to play an important role in driving the chromosome instability associated with cancer. The telomere loss on the chromosome leads to chromosome fusions between two sister chromatids during mitosis facilitating the accumulation of genetic changes [24,25]. The list of genes included in these regions is provided in the Additional file 1.

Of the four mutator gene/regions selected by our method, three are associated with double strand break repair pathways. The other one is a telomere loss which is known to lead to chromosome instability by chromosome fusion. This provides a degree of validation of our method.

Timing of driver alterations

We calculate the posterior mean of the alteration time of each gene/region for each sample *i*. The posterior mean alteration time of the gene/region *j* for the given sample depends on the estimated parameters of the prior distribution for the gene/region, other alterations which occurred in the same sample and the number of passenger alterations in the sample. Table 1 gives the posterior mean alteration time of the gene/region *j* averaged among samples in which *j* is altered and their 90% CIs. Each region is represented by its chromosome location, the candidate target genes included in the region and the type of alteration (amplification or deletion).

Based on the posterior mean of the alteration time of each gene/region for each sample *i*, we have inferred the order of driver alterations. We estimated the confidence of the sequence by the proportion the same sequence occurred out of 400 sequences obtained from 400 bootstraps. We present the order and its confidence for each sample in Additional file 2: Table S1. Figure 3 shows a summary of the inferred order of alterations occurring in tumor samples represented as a tree structure. The number in parentheses beside each alteration represents the number of samples which have the same inferred order up to that alteration. Figure 3 shows only cases in which the inferred order of the first two driver alterations occurs more than once.

Figure 3 shows that all four mutator gene/regions have the smallest posterior mean time of alterations for most of the samples in which they were altered. One of the main questions in the area of genetic instability in cancer is whether it arises early or late during tumorigenesis. It was suggested that a mutator phenotype would need to be expressed early to generate the causally associated mutations driving tumorigenesis [13], however there has been little evidence supporting this hypothesis. Our result supports the claim that alterations resulting in a mutator phenotype occur early during tumorigenesis. It also shows that in the samples in which the mutator regions 8p21.2 and 22q13.33 are altered, many driver alterations occur afterwards, confirming their roles as mutators.

Table 1 Estimates of the mean time of alteration in cell divisions with its 90% CI from ovarian data

Gene or region	Mean time of alteration in cell divisions	90% CI
1p34.2(MYCL1), Amp	307	(67,731)
3q26.2(MECOM), Amp	473	(413,688)
8p21.2(PPP2R2A), Del	6	(0,326)
8q24.21(MYC), Amp	10	(0,383)
10q23.31(PTEN), Del	545	(215,922)
11q14.1(ALG8), Amp	382	(167,830)
12p12.1(KRAS), Amp	62	(47,252)
13q14.2(RB1), Del	256	(196,602)
16q23.1(WWOX), Del	790	(101,851)
17q11.2(NF1), Del	375	(282,637)
19p13.13, Amp	445	(5,729)
19q12(CCNE1), Amp	280	(5,453)
20q13.12(ZMYND8), Amp	111	(81,359)
22q13.33, Del	0	(0,0)
BRCA1	113	(2,132)
BRCA2	2	(0,2)
CSMD3	426	(350,548)
FAT3	338	(288,745)
NF1	177	(73,684)
USH2A	521	(45,690)

The mean time of alteration for each gene/region is calculated by averaging the posterior mean of the alteration time of the gene/region among samples in which it is altered.

The non-mutator genes/regions containing MYC, KRAS, CCNE1 and RB1 have the smallest posterior mean time of alterations in most samples while the driver gene CSMD3, USH2A and the region containing MECOM, WWOX have large posterior means.

Lung cancer data

We applied our method to lung tumor sequencing data from Ding et al. (2008) [26] who sequenced the coding exons and splice sites of 623 candidate cancer genes in 188 samples from patients with lung adenocarcinomas. We applied our method to the driver genes mutated in more than ten samples : APC, ATM, EGFR, KRAS, LRP1B, NF1,PTPRD, STK11, TP53. We also included the gene

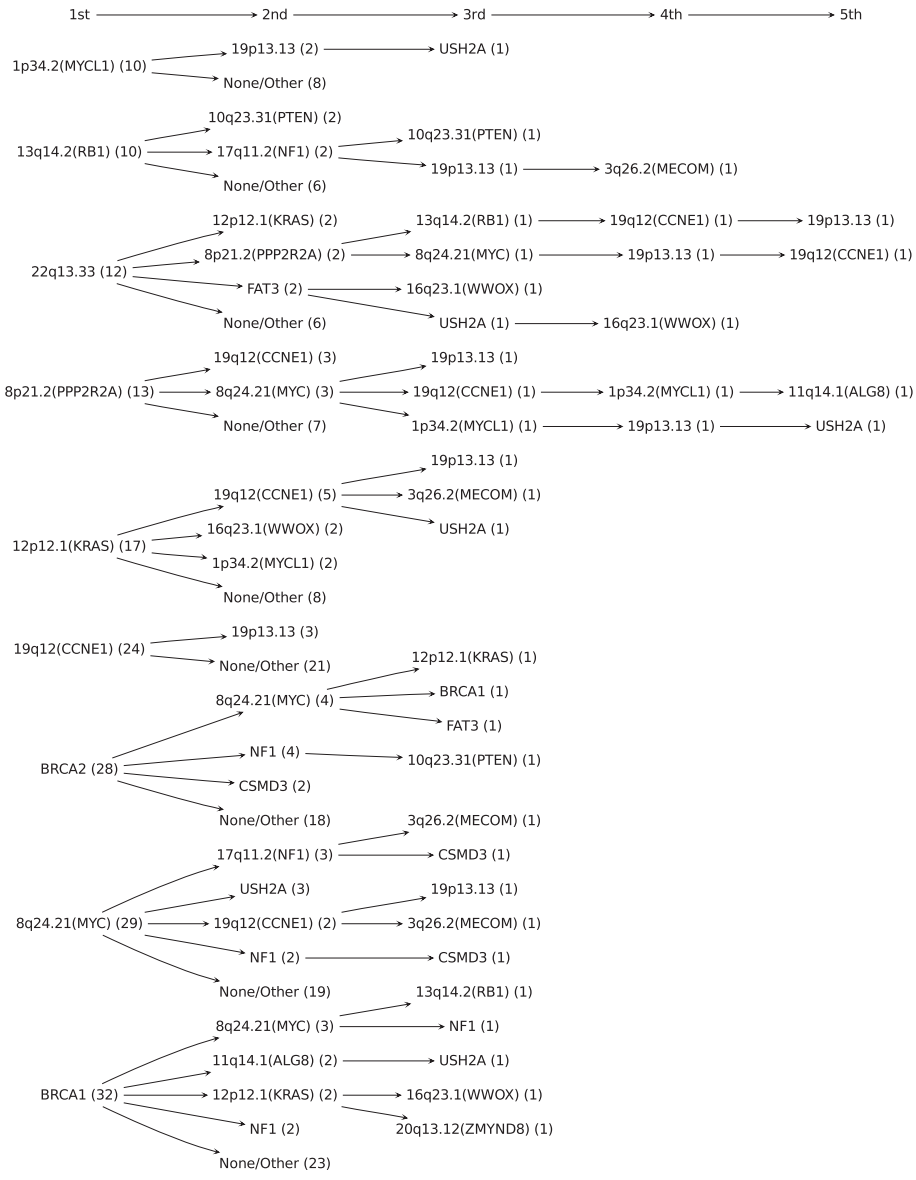

Figure 3 Order of alterations occurring in ovarian tumor samples represented as a tree structure. The number in parentheses beside each alteration represents the number of samples which have the same order up to that alteration.

PRKDC which is mutated in eight samples since it is a well known mutator gene.

In Additional file 3: Table S2, we present the posterior mean of the age of tumor lineage T_i and the alteration time of gene j, $X_{i,j}$ with their 90% CI for each sample i. The average value of the posterior mean of tumor ages is 749 cell generations. The $10th$ percentile is 236 and the $90th$ percentile is 1617 cell divisions.

Identified mutators

We estimated the increase of mutation rate Δ_j^{MUT} by the alteration of the gene j and also obtained their 90% CI from 400 bootstraps. Only two genes, TP53 and PRKDC, were found to increase mutation rates. TP53 increases mutation rate by 170% while PRKDC increases mutation rate by 670%. The 90% CI for Δ_j^{MUT} of both genes do not include zero.

Both of the genes TP53 and PRKDC are well known mutator genes. A new finding from our method is that PRKDC increases mutation rate much greater than TP53. TP53 activates DNA repair proteins when DNA has sustained damage or it initiates apoptosis if DNA damage is irreparable. PRKDC encodes a protein involved in the repair of double-stranded DNA breaks.

Timing of driver alterations

We have inferred the order of driver alterations by the posterior mean of the alteration time of each gene for each sample *i*. We present the inferred order of driver mutations and its confidence for each sample in Additional file 3: Table S2. The posterior mean alteration time of gene *j* averaged among samples in which *j* is altered and their 90% CIs are given in Table 2. Figure 4 shows a summary of the inferred order of mutations occurring in tumor samples represented as a tree structure.

It shows that EGFR, TP53, KRAS and STK11 have the smallest posterior mean time of alterations for most of the samples in which they were altered. In our analysis with ovarian cancer data, KRAS amplification was also identified as an early event.

There is much evidence supporting the finding that alterations of KRAS, EGFR, STK11 and TP53 are early events in many cancer types [9]. Figure 4 also shows that LRP1B and PTPRD tend to have the largest posterior mean time of alterations for most of the samples in which they were altered. This suggests that these genes may play important roles in invasion or metastasis. This is supported by the study suggesting LRP1B may be involved in cellular invasion/metastasis [27] and the study showing the association between deletion of PTPRD and cutaneous squamous cell carcinoma metastasis [28].

Conclusions

We have developed a method which estimates the age of the tumor lineage and the timing of the driver alterations. This method also identifies mutator genes and estimates the increase in rate of alterations caused by the mutator gene during tumorigenesis. We applied this method to TCGA ovarian cancer and lung cancer data. For ovarian cancer data, we used both sequencing and copy number data and found that BRCA1 and BRCA2 increase the rate of point mutations and the chromosomal regions 8p21.2 and 22q13.33 increase the rate of copy number alterations. We found that alterations in genes/regions resulting in a mutator phenotype tend to occur early. For the non-mutator genes/regions, the regions containing MYC, KRAS, CCNE1 and RB1 tend to alter early while the gene CSMD3, USH2A and the region containing MECOM, WWOX tend to alter late.

For lung data, we applied this method to only sequencing data and found that TP53 and PRKDC increase mutation rate. We found that EGFR, KRAS, STK11 and TP53 tend to mutate early while LRP1B and PTPRD tend to mutate late.

This is the first attempt to identify genes that increase the mutation rate or CNA rate using computational methods. Finding mutator genes simply based on the correlation between the number of passenger alterations and the alteration status of a driver gene generates many false positives since it cannot distinguish a mutator gene and a gene that alters late. For both genes, there are high correlations between the number of passenger alterations and their alteration status. For example, if we test for each driver *j* whether there is a difference in the mean between the number of passenger alterations in samples in which driver *j* is altered and those in other samples, we find that LRP1B, NF1, PRKDC, PTPRD, TP53 genes have p values less than 0.01 for lung sequencing data. For ovarian sequencing data, we found that 16q23.1, 19q12, BRCA2, FAT3, USH2A have p values less than 0.01. For ovarian copy number data, we found that 8p21.2, 8q24.21, 16q23.1, 19p13.13, 19q12, 22q13.33 have p values less than 0.01. Note that this method finds many more mutator candidates compared to our method while missing an important mutator BRCA1. The mutator candidates found by the correlation, such as PTPRD or LRP1B whose p-values are $2 \cdot 10^{-6}$ and $3 \cdot 10^{-5}$ are estimated to be simply altered late by our method. There is no evidence supporting their role in increasing genomic instability, implying they could be false positives.

It is well known that genomic instability can be caused by dysfunction of DNA repair genes and cell cycle checkpoint control genes. The DNA repair genes which have been found to be altered in cancers include BRCA1/2, MSH2/6, MLH1/2, BLM, RAD50, MRE11, NBS1, PRKDC, NBS1, BLM, RECQL4, BAP1, WRN, RAD51L3, RAD52, FANCA, and PALB2 [10-12]. Of the mutator genes identified in our analysis of lung and ovarian cancer, BRCA1/2, PRKDC, and PPP2R2A gene in the region 8p21.2 belong to this category although the role of PPP2R2A in inducing chromosomal instability in ovarian cancer was previously unknown. Other DNA repair genes are rarely altered in our dataset. The genes in the cell cycle

Table 2 Estimates of the mean time of alteration in cell divisions with its 90% CI for the driver genes from lung data

Gene	Mean time of alteration in cell divisions	90% CI
APC	379	(44,801)
ATM	594	(93,805)
EGFR	23	(18,93)
KRAS	280	(158,392)
LRP1B	549	(443,917)
NF1	505	(60,744)
PRKDC	466	(324,1637)
PTPRD	801	(394,1228)
STK11	259	(108,455)
TP53	323	(208,456)

The mean time of alteration for each gene is calculated by averaging the posterior mean of the alteration time of the gene among samples in which it is altered.

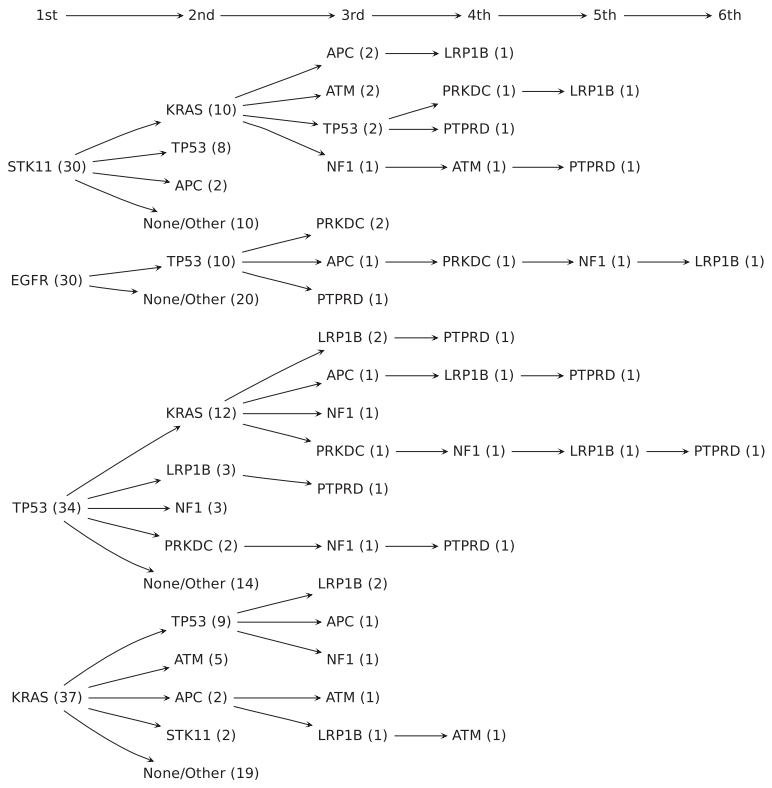

Figure 4 Order of alterations occurring in lung tumor samples represented as a tree structure. The number in parentheses beside each alteration represents the number of samples which have the same order up to that alteration.

checkpoint control pathway which have been found to be altered in cancers include TP53, ATM, MDM2/4, BUB1, and STK12. Of the mutator genes we identified, TP53 belongs to this category.

In addition to the DNA repair and cell cycle chekpoint processes, there are many other processes involved in genomic stability. These include DNA replication, deoxynucleotide metabolism, chromosome condensation, sister chromatid cohesion, kinetochore structure and function and centrosome/microtubule formation. Therefore, in principle, there are many genes that could induce genomic instability. Other than these processes, telomere erosion is known to be able to lead to chromosome instability. In our analysis of ovarian data, we found that the deletion of 22q13.33 is telemere loss which leads to chromosome instability. This is a new finding that supports the role of telomere erosion in CIN of ovarian cancer.

Our method also provides an estimate of tumor age and timing of driver alterations which can be obtained only through computational methods. The age of the tumor lineage is the number of cell divisions in the dominant clone's lineage from the birth of the patient until the biopsy. Some tissues such as pancreatic epithelia do not self-renew, therefore, most of the cell divisions in the lineage of the pancreatic tumor occur after the onset of neoplasia [29]. Therefore the age of the tumor lineage corresponds approximately to the tumor age, the time interval from the onset of neoplasis to the tumor detection in units of cell generations. Some tissues such as skin or gastrointestinal epithelia regulary self-renew. In these cases, the number of cell divisions in the lineage is the sum of the number of cell divisions before the onset of neoplasia and that after the onset of neoplasia. If the cell division rate has been constant throughout a life, the age of the tumor lineage corresponds to the age of the patient. In this paper, we estimated the average age of the tumor lineage for the ovarian tumor is 1113 cell divisions and that for the lung tumor is 749 cell generations. Ovarian elithelia regularly self renew [29], while lung epithelia renew slowly and are stimulated to self-renew upon injury [30], therefore, the age of the tumor lineage for lung tumor is close to the tumor age. The cell division time for a lung tumor cell is known to be approximately 8 days [31]. Therefore it takes $749 \cdot 8$ days $= 16.4$ years on average from the beginning of the tumor to the detection of lung tumor.

Estimates of tumor age, together with clinical data such as tumor stage can provide information for how long it takes for a benign tumor to develop into invasive and metastatic tumors. Estimating when metastasis occurs during tumorigenesis is particularly important

since metastasis is responsible for most cancer related deaths although it is the least understood process. Understanding this can help planning early detection programs for cancer since it is critical to know how early you have to detect the tumor in order to have an effect. If the tumor metastasizes before detection, then early detection of the primary tumor may not help the patient. For example, cancer screening has been successful for both colon and cervical cancers in reducing death rate but results for breast cancer are less successful, indicating that screening breast mammography fails to detect cancer until after they have spread [32]. Although this problem of estimating when metastasis occurs has not been dealt with in this paper, it is an important future work that our method can be used to answer.

Estimates of tumor age also provide insight into the biology of tumor cell populations, may help to understand intra-tumor heterogeneity and differences in prognosis and responsiveness to therapy. A previous attempt to estimate the tumor age using the number of passenger mutations [2] did not take into account the increase of the mutation rate by the alteration of mutator genes, and hence their estimate of tumor age may be somewhat overestimated. We believe our method will be a useful contribution for better understanding the process of tumorigenesis.

Availability of supporting data

The software and data are available at: https://sites.google.com/site/ahrimy2013/home/software.zip.

Additional files

Additional file 1: Supplementary Materials.

Additional file 2: Table S1.

Additional file 3: Table S2.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AY and RS conceived the study. AY implemented the algorithm. AY and RS wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
We thank Dr. Kazimierz O. Wrzeszczynski for his helpful feedback.

Received: 19 August 2013 Accepted: 10 December 2013
Published: 13 December 2013

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW: **Cancer Genome landscapes.** *Science* 2013, **339**(6127):1546–1558.
2. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, Kamiyama M, Hruban RH, Eshleman JR, Nowak MA, Velculescu VE, Kinzler KW, Vogelstein B, Iacobuzio-Donahue CA: **Distant metastasis occurs late during the genetic evolution of pancreatic cancer.** *Nature* 2010, **467**(7319):1114–1117.
3. Weinberg RA: *The Biology of Cancer, 1st edn.* Garland Science; 2006.
4. Gatenby RA, Maini PK: **Mathematical oncology: cancer summed up.** *Nature* 2003, **421**(6921):321. http://dx.doi.org/10.1038/421321a.
5. Desper R, Jiang F, Kallioniemi OP, Branch CG, Moch H, Papadimitriou CH, Schaffer AA: **Inferring tree models for oncogenesis from comparative genome hybridization data.** *J Comput Biol* 1999, **6:**37–51.
6. Attolini CSO, Cheng YK, Beroukhim R, Getz G, Abdel-Wahab O, Levine RL, Mellinghoff IK, Michor F: **A mathematical framework to determine the temporal sequence of somatic genetic events in cancer.** *Proc Natl Acad Sci* 2010. http://www.pnas.org/content/early/2010/09/17/1009117107.abstract.
7. Durinck S, Ho C, Wang NJ, Liao W, Jakkula LR, Collisson EA, Pons J, Chan SW, Lam ET, Chu C, Park K, Hong Sw, Hur JS, Huh N, Neuhaus IM, Yu SS, Grekin RC, Mauro TM, Cleaver JE, Kwok PY, LeBoit PE, Getz G, Cibulskis K, Aster JC, Huang H, Purdom E, Li J, Bolund L, Arron ST, Gray JW, Spellman PT, Cho RJ: **Temporal dissection of tumorigenesis in primary cancers.** *Cancer Discov* 2011, **1**(2):137–143. http://dx.doi.org/10.1158/2159-8290.CD-11-0028.
8. Gerstung M, Eriksson N, Lin J, Vogelstein B, Beerenwinkel N: **The temporal order of genetic and pathway alterations in tumorigenesis.** *PLoS ONE* 2011, **6**(11):e27136. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0027136.
9. Youn A, Simon R: **Estimating the order of mutations during tumorigenesis from tumor genome sequencing data.** *Bioinformatics* 2012, **28**(12):1555–1561. http://bioinformatics.oxfordjournals.org/content/28/12/1555.abstract.
10. Negrini S, Gorgoulis VG, Halazonetis TD: **Genomic instability — an evolving hallmark of cancer.** *Nature Rev Mol Cell Biol* 2010, **11**(3):220–228.
11. Lengauer C, Kinzler KW, Vogelstein B: **Genetic instabilities in human cancers.** *Nature* 1998, **396**(6712):643–649. http://dx.doi.org/10.1038/25292.
12. Loeb LA: **A mutator phenotype in cancer.** *Cancer Res* 2001, **61**(8):3230–3239. http://cancerres.aacrjournals.org/content/61/8/3230.abstract.
13. Loeb LA, Loeb KR, Anderson JP: **Multiple mutations and cancer.** *Proc Natl Acad Sci USA* 2003, **100**(3):776–781. http://dx.doi.org/10.1073/pnas.0334858100.
14. Youn A, Simon R: **Identifying cancer driver genes in tumor genome sequencing studies.** *Bioinformatics* 2011, **27**(2):175–181. http://dx.doi.org/10.1093/bioinformatics/btq630.
15. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data..** *Biostatistics* 2004, **5**(4):557–572. http://view.ncbi.nlm.nih.gov/pubmed/15475419.
16. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome Biol* 2011, **12**(4):R41. http://www.biomedsearch.com/nih/GISTIC20-facilitates-sensitive-confident-localization/21527027.html.
17. Lichtenauer-Kaligis EG, Thijssen J, den Dulk H, van de Putte P, de Jong JGT, Giphart-Gassler M: **Comparison of spontaneous hprt mutation spectra at the nucleotide sequence level in the endogenous hprt gene and five other genomic positions.** *Mut Res/Fundam Mol Mech Mutagen* 1996, **351**(2):147–155. http://www.sciencedirect.com/science/article/pii/0027510795002197.
18. Araten DJ, Golde DW, Zhang RH, Thaler HT, Gargiulo L, Notaro R, Luzzatto L: **A quantitative measurement of the human somatic mutation rate.** *Cancer Res* 2005, **65**(18):8111–8117. http://www.biomedsearch.com/nih/quantitative-measurement-human-somatic-mutation/16166284.html.
19. Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474**(7353):609–615. http://dx.doi.org/10.1038/nature10166.
20. Yoshida K, Miki Y: **Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage.** *Cancer Sci* 2004, **95**(11):866–871.
21. Cheng Y, Liu W, Kim ST, Sun J, Lu L, Sun J, Zheng SL, Isaacs WB, Xu J: **Evaluation of {PPP2R2A} as a prostate cancer susceptibility gene: a comprehensive germline and somatic study.** *Cancer Genet* 2011, **204**(7):375–381. http://www.sciencedirect.com/science/article/pii/S2210776211001190.

22. Kalev P, Simicek M, Vazquez I, Munck S, Chen L, Soin T, Danda N, Chen W, Sablina A: **Loss of PPP2R2A Inhibits Homologous Recombination DNA Repair and Predicts Tumor Sensitivity to PARP Inhibition.** *Cancer Res* 2012, **72**(24):6414–6424. http://cancerres.aacrjournals.org/content/72/24/6414.abstract.
23. Phelan K, McDermid H: **The 22q13.3 deletion syndrome (Phelan-McDermid Syndrome).** *Mol Syndromol* 2012, **2**(3–5):186–201.
24. Murnane JP: **Telomere loss as a mechanism for chromosome instability in human cancer.** *Cancer Res* 2010, 0008–5472. CAN–09–4357+. http://dx.doi.org/10.1158/0008-5472.can-09-4357.
25. Maser RS, Depinho RA: **Connecting chromosomes, crisis, and cancer.** *Science* 2002, **297**(5581):565–569.
26. Ding L, Getz G, Wheeler DA, Mardis ER, Mclellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, et al.: **Somatic mutations affect key pathways in lung adenocarcinoma.** *Nature* 2008, **455**(7216):1069–1075. http://dx.doi.org/10.1038/nature07423.
27. Liu CX, Li Y, Obermoeller-McCormick LM, Schwartz AL, Bu G: **The putative tumor suppressor LRP1B, a novel member of the Low Density Lipoprotein (LDL) receptor family, exhibits both overlapping and distinct properties with the LDL receptor-related protein.** *J Biol Chem* 2001, **276**(31):28889–28896. http://www.jbc.org/content/276/31/28889.abstract.
28. Lambert S, Harwood C, Purdie K, Gulati A, Matin R, Romanowska M, Cerio R, Kelsell D, Leigh I, Proby C: **Metastatic cutaneous squamous cell carcinoma shows frequent deletion in the protein tyrosine phosphatase receptor Type D gene.** *Int J Cancer* 2012, **131**(3):E216–E226.
29. Tomasetti C, Vogelstein B, Parmigiani G: **Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation.** *Proc Natl Acad Sci* 2013, **110**(6):1999–2004. http://www.pnas.org/content/110/6/1999.abstract.
30. Stripp BR, Reynolds SD: **Maintenance and repair of the bronchiolar epithelium.** *Proc Am Thoracic Soc* 2008, **5**(3):328–333.
31. Tinnemans M, Schutte B, Lenders M, Ten Velde G, Ramaekers F, Blijham G: **Cytokinetic analysis of lung cancer by in vivo bromodeoxyuridine labelling.** *Br J Cancer* 1993, **67**(6):1217–1222.
32. Griffith O, Gray J: **Omic approaches to preventing or managing metastatic breast cancer.** *Breast Cancer Res* 2011, **13**(6):230. http://breast-cancer-research.com/content/13/6/230.

doi:10.1186/1471-2105-14-363
Cite this article as: Youn and Simon: Using passenger mutations to estimate the timing of driver mutations and identify mutator alterations. *BMC Bioinformatics* 2013 **14**:363.