

RESEARCH ARTICLE

Open Access



Somatic evolutionary timings of driver mutations

Karen Gomez^{1,2†}, Sayaka Miura^{1,2†}, Louise A. Huuki¹, Brianna S. Spell^{1,2}, Jeffrey P. Townsend^{3,4,5} and Sudhir Kumar^{1,2,6*}

Abstract

Background: A unified analysis of DNA sequences from hundreds of tumors concluded that the driver mutations primarily occur in the earliest stages of cancer formation, with relatively few driver mutation events detected in the late-arising subclones. However, emerging evidence from the sequencing of multiple tumors and tumor regions per individual suggests that late-arising subclones with additional driver mutations are underestimated in single-sample analyses.

Methods: To test whether driver mutations generally map to early tumor development, we examined multi-regional tumor sequencing data from 101 individuals reported in 11 published studies. Following previous studies, we annotated mutations as early-arising when all tumors/regions had those mutations (ubiquitous). We then inferred the fraction of mutations occurring early and compared it with late-arising mutations that were found in only single tumors/regions.

Results: While a large fraction of driver mutations in tumors occurred relatively early in cancers, later driver mutations occurred at least as frequently as the early drivers in a substantial number of patients. This result was robust to many different approaches to annotate driver mutations. The relative frequency of early and late driver mutations varied among patients of the same cancer type and in different cancer types. We found that previous reports of the preponderance of early driver mutations were primarily informed by analysis of single tumor variant allele profiles, with which it is challenging to clearly distinguish between early and late drivers.

Conclusions: The origin and preponderance of new driver mutations are not limited to early stages of tumor evolution, with different tumors and regions showing distinct driver mutations and, consequently, distinct characteristics. Therefore, tumors with extensive intratumor heterogeneity appear to have many newly acquired drivers.

Keywords: Driver mutation, Ubiquitous mutation, Private mutation, Somatic mutation

Background

Tumor cells accumulate numerous somatic mutations. Some of these mutations directly contribute to tumor growth and progression and are commonly referred to as driver mutations. Knowledge of the relative timing of driver mutations is essential for understanding cancer progression as a whole and for optimizing treatment for individual patients [1–4]. For this reason, much attention has been paid to identifying the distribution and

frequency of driver mutations among tumor cell populations [5–7].

Mutations found in most cells of a tumor can be detected by estimating the fraction of cancer cells harboring a particular variant, or cancer cell fraction (CCF) [8]. This approach has been utilized to analyze the extensive data available through the Cancer Genome Atlas (TCGA), a comprehensive database that contains genomic changes in hundreds of thousands of tumors (single tumor genome sequencing results) from 33 types of cancers [9]. When CCFs were estimated for driver mutations found in tumors from TCGA, most driver mutations were present at high CCF, meaning most driver mutations were found in the majority of cells in a

* Correspondence: s.kumar@temple.edu

†Equal contributors

¹Institute for Genomics and Evolutionary Medicine, Sudhir Kumar, SERC 602A, 1925 N. 12th Street, Philadelphia, PA 19122, USA

²Department of Biology, Temple University, Philadelphia, PA 19122, USA

Full list of author information is available at the end of the article

tumor [10]. Mutations found in a large proportion of tumor cells are likely to have occurred at the earlier stages of tumor growth, because such early-arising mutations are inherited by all cells in a tumor following clonal evolution [2, 11]. Therefore, it was inferred that the majority (>70%) of driver mutations occur early in tumor growth for all types of cancer surveyed [12].

However, tumors consist of heterogeneous cell lineages, each of which may be driven by different driver mutations [1–3]. For example, a study of renal cell carcinoma reported 73–75% of driver mutations to have likely occurred at a later time in tumor evolution [13]. This study employed a different methodology to detect early mutations, where variants from multiple regions of a given tumor (M-seq) were analyzed. Using an M-seq approach, variants found in a majority of sampled regions are classified as early (ubiquitous) mutations [6]. Additional studies have employed M-seq methodology to study different types of cancer, and these studies have identified many mutations that are private to one or only a few regions of a tumor [14–21], or mutations that are present at different points in time [22, 23]. This identification of many private mutations indicates that single-tumor profiles, such as those in the TCGA database, do not completely capture the spectrum of late-arising driver mutations. Therefore, conclusions based on analyses of TCGA data [12] may not apply to all the cancers and/or patients.

Now that M-seq data from various cancers are available, it is possible to comprehensively explore the relative preponderance of driver mutations arising in early and late stages of tumor growth. Here we perform a meta-analysis using samples from 101 individuals representing various cancer types [13, 14, 24–32], which revealed that the fraction of driver mutations occurring early in tumor growth varies extensively among cancers as well as among individuals. We evaluate the frequency of late-arising driver mutations in primary and metastatic tumors.

Methods

We obtained sequencing read counts of mutant and wild type alleles, and their chromosomal positions, for 101 tumor data sets from 11 published studies that contained at least three tumor samples per patient (Table 1) [13, 14, 24–32]. Our analysis focused on single nucleotide variants (SNVs) and insertions and deletions (indels) that arose somatically in the tumors of individual patients. To identify mutations that likely affected cancer progression or development (driver mutations), we first extracted mutations in coding regions of the genome by mapping the chromosomal position of each variant onto the reference human genome (hg19) from the Ensembl database [33]. We excluded mutations in intergenic regions, because their functional effects on cancer development are rarely known. We also excluded synonymous mutations, because

Table 1 Summary of data sources analyzed in the present study

Cancer type	Study (Reference)	Number of Patients in Study
Mixed: Lung, pancreatic, and 11 other types	[32]	37
Brain (glioblastoma)	[25]	10
Colorectal	[27]	9
Prostate	[31]	9
Kidney (clear cell renal cell carcinoma)	[13]	8
Lung	[14]	7
Endometrial	[28]	6
Esophageal (with Barrett's esophagus)	[29]	5
Brain (glioma)	[24]	5
Breast	[26]	4
Ovarian	[30]	1

these mutations are not expected to significantly affect protein function.

To ensure that our findings regarding the distribution of drivers are robust to methods of driver determination, we used five schemes to determine mutations that are cancer drivers. We first determined driver mutations based on whether the affected gene has been previously implicated in cancer. We annotated every mutation occurring in a cancer-associated gene listed in the COSMIC cancer gene census [34, 35] as a driver (Driver annotation type I). Among these genes, those without functional annotation in cancer (oncogene or tumor suppressor gene) could be false-positives. Therefore, we applied a second, more stringent approach, in which we used only known oncogenes and tumor suppressor genes listed in COSMIC (Driver annotation type II).

Some sites in the genome are more frequently mutated in cancer than others, and these somatic variant hot spots are believed to play a role in cancer [36]. In our third approach, we annotated driver mutations when mutations were located within 15 nucleotides of somatic variant hot spot (Driver annotation type III), because mutations at hot spots and neighboring regions may be cancer drivers. Variant hot spots were those identified as individual substitution hot spots as presented in Chang et al. [36], those mutated >10 individuals in the COSMIC database, or those mutated in at least two individuals in our 101 datasets.

Furthermore, many computational methods have been created to determine variants with functional roles in cancer [37]. In the fourth approach, we used one such driver mutation prediction tool, IntOGen [38] (<http://www.intogen.org/analysis>). The IntOGen pipeline examines genes that frequently have mutations with high functional impact and regions of the protein sequences where mutations frequently occur (Driver annotation type

IV). To predict driver mutations with IntOGen, we input all mutations (point mutations and indels) with chromosomal positions, wild type and mutant nucleotides, and strand information obtained from Ensembl database.

Lastly, it has been observed that genes that are causal in one cancer type many not be causal in every cancer type. Therefore, in annotation type V, we obtained a list of cancer associated genes for each cancer type from IntOGen database [38] in order to designate driver mutations following the approach outlined in annotation type I above. For analysis of data from Zhao et al. [32], we used cancer-associated genes identified for each patient by Zhao et al. to maintain consistency with their analysis.

Before distinguishing between early vs. late-occurring mutations, we made use of the extra power conferred by the M-seq approach to detect potential sequencing errors. Using Treeomics [39], we estimated the posterior probability that a putative variant read is actually present in a tumor given the reference and mutant allele read counts for multiple tumor samples. For each variant within a tumor sector, we annotated a mutation as present if the Treeomics-inferred posterior probability was greater than 0.95. We removed variants with a posterior probability less than 0.95 in all tumor samples from a patient. Similarly, a mutation was annotated to be absent if the Treeomics-inferred posterior probability was less than 0.05.

To distinguish early from late-arising mutations, we defined early-arising mutations as those that are found in all samples (ubiquitous mutations), and late-arising mutations as those that are private to only one sample (region-specific or private mutations). This definition is more stringent than the definition used in the previous TCGA analysis [12], because it excludes mutations found in some, but not all tumors. Consequently, mutations annotated to be early-arising are expected to be found in the progenitor of all tumor cells in a patient, and those designated late-arising status are expected to have arisen in only one

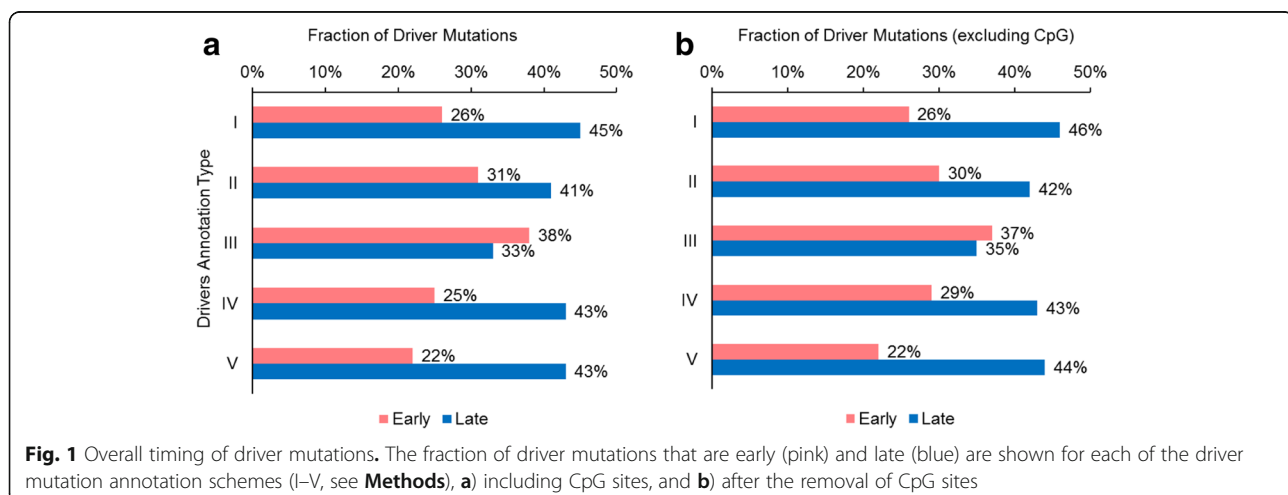
tumor in a patient. We refer to all other drivers to be of intermediate origin. Note that early and late designations refer to relative timing of occurrence of mutations, they are not meant to convey absolute times. Supplementary Additional file 1 shows designations of drivers using five annotation schemes, chromosomal positions, count of samples with mutant allele after Treeomics treatment, patient ID, and study information.

Results

Meta-analysis of driver mutation timing

We first pooled driver mutations from 101 patient data sets from 11 studies [13, 14, 24–32] and identified early drivers. Analyzing type I drivers (i.e., any mutation in a cancer-associated gene [34, 35]), we found that only 26% of all driver gene mutations have arisen early (Fig. 1a), which indicates that most driver mutations did not occur at the earliest stages of tumor growth. In fact, we found that 74% of driver mutations were not early, which is in contrast to results from TCGA analysis [12], but consistent with the findings of Gerlinger et al. [13] in clear cell renal carcinoma samples. However, our data sample is 15 times larger than Gerlinger et al. [13] and includes tumors from 21 types of cancers. These findings were robust to the use of more stringent approaches to driver gene determination (annotation type II to V): the number of non-early driver mutations was always greater than the number of early driver mutations. Across annotation types, only about one-third of all driver mutations occurred early.

Interestingly, we found relatively large numbers of driver mutations to be late-occurring (33–45%), with the number of late-driver mutations similar to or even greater than the number of early-occurring driver mutations (22–38%) (Fig. 1a). The remaining driver mutations had intermediate origins. In these analyses, only annotation type III predicted early driver mutations (38%) to be



larger than late driver mutations (33%), but still they are very similar in numbers. That is, across M-seq cases, late-occurring drivers are generally more frequently observed than, or are equal in frequency to, early-arising drivers. Analysis using the less stringent definition of driver mutation (annotation type I) produced results similar to the more stringent definitions (annotation types II–V). Overall, we found the numbers of late-arising driver mutations to be substantial.

False-positive detections of driver mutations are expected to be high when mutations are located on genomic positions with higher mutation rate, e.g., CpG sites [40, 41]. Therefore, we obtained a list of CpG sites in the human genome from UCSC sequence hg19 using the Bioconductor R package, and removed driver mutations that were located at these sites. We still observed fewer driver mutations to have occurred early (22–37%) than late (35–46%; Fig. 1b). So, we expect that our inferences have not been affected by false-positive detection of driver mutations at mutational hot spots.

Driver mutation timing by cancer type and individual differences

Although the total number of late driver mutations in our data was similar to or greater than that of early driver mutations, we found that the fraction of driver mutations occurring at early stages varied extensively among patients, studies, and cancer types. Very few early drivers were detected in esophageal adenocarcinoma data sets (average 6%, range 0–18%), but a large fraction of drivers were early in breast cancer data sets (average 69%, range 50–100%; Fig. 2a). Similarly, the average fraction of late mutations had a wide range, from 13% in ovarian cancer to 81% (range 0–91%) in recurrent glioblastoma.

Although patients showed similar fractions of early and late driver mutations within specific cancer types, some cancer types did not. For example, there is extensive variation in the fraction of early and late mutations identified in both glioblastoma data sets sequenced at primary tumor and recurrence [24, 25]. Similarly,

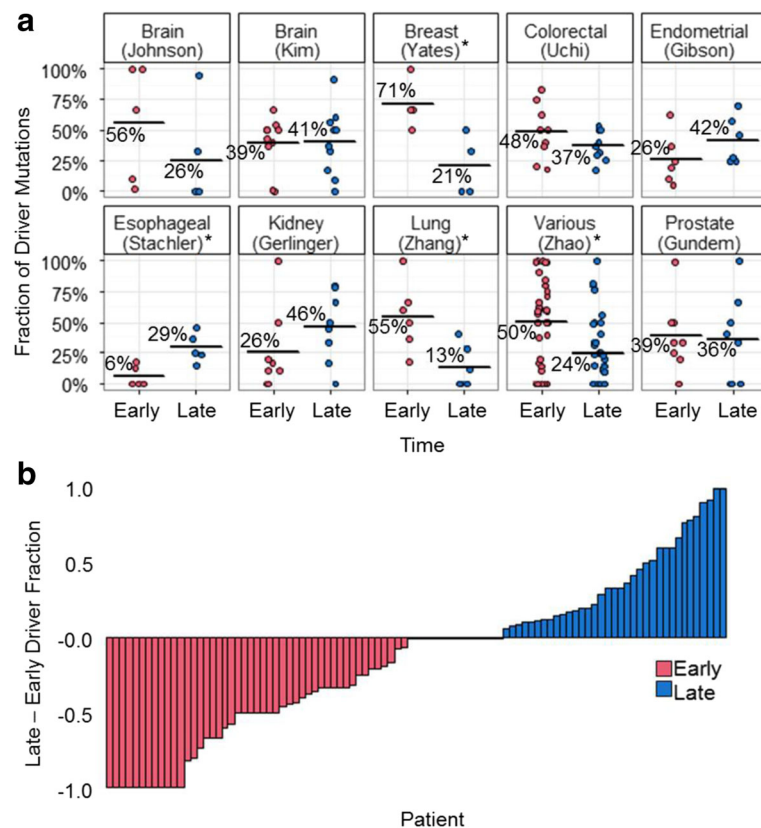


Fig. 2 Fraction of driver mutations occurring at early and late time. Driver mutations were annotated as those found in cancer-associated genes. **a** Fraction of all driver mutations that occurred early and late as inferred from multi-sample profiles. Each dot refers to data from one patient from a study, and a bar shows the average. Statistical tests (paired *t*-test) were performed to test if the fraction of early-driver mutations is significantly different from late-driver mutations for a cancer type. Cancer types that have significantly different fractions ($P \leq 0.05$) are shown with asterisks. **b** The difference in early and late driver mutation fractions for individual patients. Zero difference was found for 15 patients. Three data sets were removed because there were zero driver mutations after removing variants absent from all tumors after the application of Treeomics software (see **Methods**)

analyses of primary tumor and multiple metastatic tumors for each patient (Zhao's [32] data) revealed extensive variation among patients in the fraction of early and late mutations. About half of the patients (48 patients) exhibited a larger fraction of early driver mutations than late driver mutations (Fig. 2b). Furthermore, we found that a large number of patients (35 patients) exhibited a greater fraction of late driver mutations than early driver mutations. Therefore, the relative counts of early to late driver mutations in a tumor varied both by tumor type as well as on an individual basis.

We also analyzed the preponderance of late drivers found in metastatic tumors, because mutations found in metastatic tumors can be classified as occurring late with greater certainty than those mutations found in the primary tumors as well. We found that the fraction of early driver mutations detected in this way remained similar to those reported above (32%), and late drivers still occurred at a high frequency (27%; Fig. 3a). That is, the fraction of late driver mutations was only slightly smaller than early driver mutations (27 and 32%), with some patients showing larger numbers of late mutations than early mutations (Fig. 3b).

Single versus multiple tumor profiles

We tested the hypothesis that the use of only a single tumor per patient in previous analyses is the primary

reason for the difference between our results and those reported earlier (e.g., [12]). Using just one tumor sample for each patient in our datasets and applying the driver annotation scheme as in McGranahan et al. [12], we found that 66% of the drivers were inferred to be early, which is consistent with McGranahan et al. [12]'s finding of 70% or more of drivers originating early-on (Fig. 4a). This fraction decreased dramatically (to 45%) when multiple samples are used for each patient from the same data set. Therefore, the power to detect late driver mutations is strongly dependent on the use of multiple samples per patient.

In addition, very few patients had a greater fraction of late-drivers compared to early-drivers when single samples were used (Fig. 4b). This comparison reveals that the use of a single sample per patient leads to a different result from that obtained using multiple samples, and the power to detect late-occurring drivers increases with additional sampling. This result is consistent with those reported previously [6, 32, 42]: multiple sequenced regions are necessary to determine the numbers of early and late driver mutations [43]. Overall, the use of single tumor samples provides poor scope to differentiate driver mutation events that happen early in tumor growth from late-arising driver mutations.

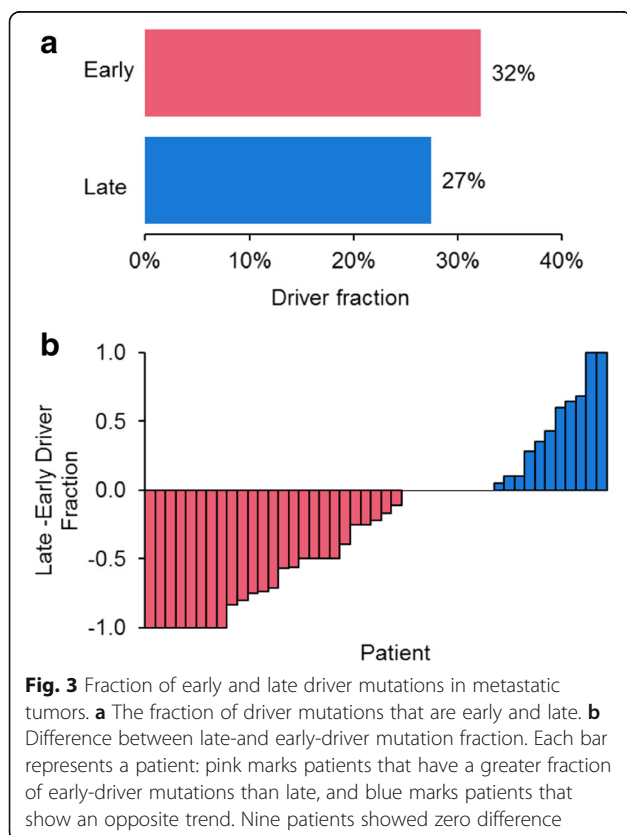
Robustness of early vs. late driver occurrence patterns

While the above patterns consistently showed that the numbers of driver mutations occurring late are comparable to those that occurred early-on, it is important to assess their robustness to a number of factors that complicate analysis and interpretation of tumor genome variation.

First, the observed variability in the timing of driver mutation occurrence among patients may be caused by technical issues, such as mutation calling methods, tumor purity, and sequencing depth. This was the reason for our use of Treeomics to exclude low quality SNVs due to low sequencing depth.

Second, it is possible that the differences observed between studies (cancer types) were caused by the differences in mutation calling methods among the studies, as some studies may be able to detect mutations with lower SNV frequencies than others. However, we often observed that the fraction of early driver mutations as well as late driver mutations varied among patients from the same study analyzed with the same methodologies. Therefore, any systematic error based on methodology would appear to be minor, and such technical issues should not strongly affect our conclusion.

Third, tumor purity could impact the annotation of early and late drivers. Generally, though not necessarily, the late-arising subclones will be in lower frequency when the tumor purity is low. Therefore, if purity were



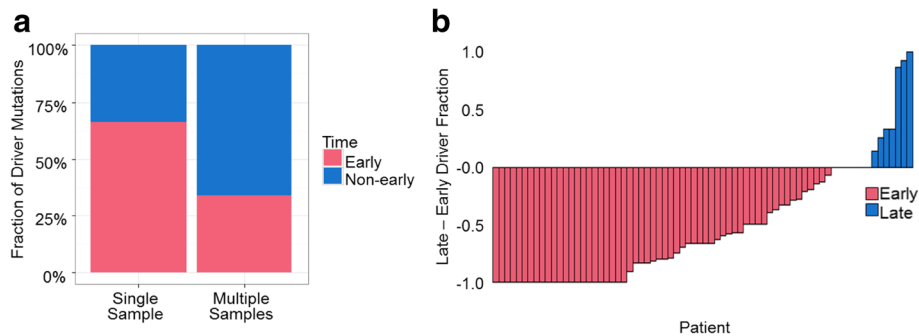


Fig. 4 Timing of driver mutations using single and multiple tumor samples. Driver mutations were annotated as those found in driver genes identified in the previous report [12]. **a** Fraction of driver mutations occurring at early time. For the single sample data set (left), we generated 100 replicates, where we randomly selected a single sector per patient. For each replicate, we pooled driver mutations and computed the fraction of early driver mutations (mean: 66%). For multiple samples (right), all samples available for each data set were used to compute the fraction of early driver mutations (45%). The fraction of early drivers found in 100 replicates of single-tumor sampling was statistically greater than the early driver fraction found using multiple samples by single single-sample t test ($P < 10^{-15}$). **b** Difference between late and early-driver mutation fraction calculated using single-tumor samples (one replicate is shown). Each bar represents a patient: pink marks patients that have a greater fraction of early-driver mutations than late, and blue marks patients that show an opposite trend. Eleven patients contained equal proportions of early and late drivers, and 7 patients were removed as no driver mutations were identified

an issue, we would expect to experience a lesser power to detect late drivers as compared to early drivers, as early drivers are expected to manifest at higher frequencies. Thus, our estimates of the relative excess of late drivers are likely to be conservative.

Fourth, the number of early driver mutations may be underestimated, because sequencing reads indicating true early driver mutations may not be observed in one of many samples by chance, despite high overall coverages. While this dropout can occur, we expect it to be far less common for early drivers than it would be for private mutations, because they will generally occur with lower frequency and in fewer multiregion samples than the early drivers. Once again, our observation of the relative excess of late drivers is conservative.

Fifth, copy number alternations (CNAs) will likely cause difficulty in designating some early-arising drivers, because the drivers can be lost by the loss of genomic segments in some tumor samples. Ideally, a reanalysis of all the primary data will be desired to identify this effect fully. However, currently available methods are only modestly accurate [44, 45]. Furthermore, CNAs can occur multiple times during the clonal evolution, which will result in complex evolutionary trajectories for SNVs involved in CNAs. In general, we expect our results to be not severely impacted by CNAs, because the number of SNVs affected by loss of mutant alleles due to CNAs is expected to be small due to the fact that most of CNAs will not affect the presence of mutant alleles, i.e., mutant alleles will be lost only when segmental losses or losses of heterozygosity (LOHs) lead to the loss of mutant alleles. To examine the potential effect of CNAs on the counts of early driver mutations, we annotated

mutations as ‘early,’ when >80, >70, and >60% of samples had mutant alleles. Although the number of early driver mutations was increased as we used a less stringent criterion (i.e., allowing some samples without mutant alleles), the number did not exceed the number of late driver mutations (Fig. 5). Therefore, our conclusion should be robust to CNAs.

Discussion

Our results establish that the fraction of driver mutations occurring in the earliest stages of cancer varies among patients as well as cancer types. We have shown that, overall, the number of late driver mutations are equal to or greater than early drivers in 44% of the patients with metastatic tumors. This conclusion differs from some previous reports arguing that the majority of

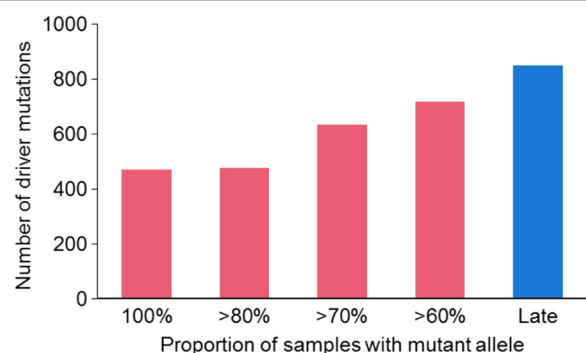


Fig. 5 The number of early driver mutations when some samples may have wild-type alleles. We annotated mutations as early mutations, when 100% (all), >80, >70, and >60% of samples had mutant alleles. The number of late driver mutations are shown with the blue bar

driver mutations happen early in cancer progression [12] or that tumors follow a neutral pattern of evolution after initial growth propelled by the effects of early driver mutations, i.e., intratumor heterogeneity is caused by passenger mutations [46].

Our observation that the number of late driver mutations is similar to early driver mutations does not inform us about the rate of driver mutation occurrence per cell or about the relative degrees of selective advantage conferred by early and late drivers. The number of cells that arise late in cancer progression (subclonal cells that have subclonal variants) is expected to exceed the number of clonal cells, so the number of subclonal variants is expected to exceed the number of clonal variants, which would result in the increased preponderance of mutations. In fact, the number of driver mutations was linearly correlated with the number of passenger mutations for both early (Fig. 6a) and for late (Fig. 6b) mutations (see also [32]). Actually, the number of early driver mutations per all early mutations (fraction of driver mutations) was similar to the fraction of late driver mutations (Fig. 6c). This pattern was different from the fractions of early and late driver mutations (Fig. 1). Thus, even when the numbers of early and late driver mutations are similar, it will not mean that the rate of driver mutation occurrence or accumulation per cell is the same.

We found that subclones with late drivers occur with significant frequencies; the average observed mutant frequency of late-arising mutations was 19% (with a standard deviation of 14%). In fact, 34% of late mutations were present at frequencies greater than 20% (Fig. 7a). However, the relative degrees of selective advantage conferred by early and late drivers is complex to assess from such frequency data, as for example, subclonal expansions may be caused by spatial constrain without positive selection [46]. Furthermore, a comparison of the frequency of late driver and passenger mutations is not able to inform about positive selection, because passenger mutations hitchhike with driver mutations—which would result in similar observed mutant frequencies for both [47]. As expected, the distribution of the observed mutant frequencies of late drivers was similar to that of late passengers. This pattern was also observed when all the data from all late mutations was pooled together (Fig. 7a) and when the comparison was restricted to individual regions that contained at least 10 late driver mutations (>10 mutations; Fig. 7b, c). However, it does appear that higher intratumor heterogeneity in the late stages is a result of the continued occurrence of genuine driver mutations with functional effects on tumor growth, because recent studies have found subclone-specific driver mutations in tumors using single-cell sequencing techniques. For

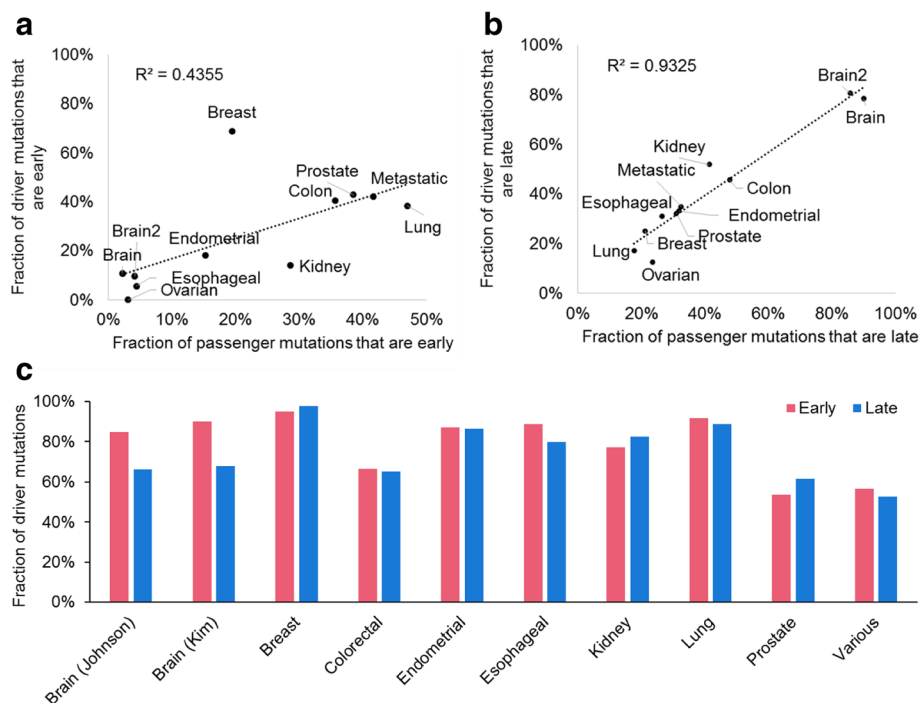
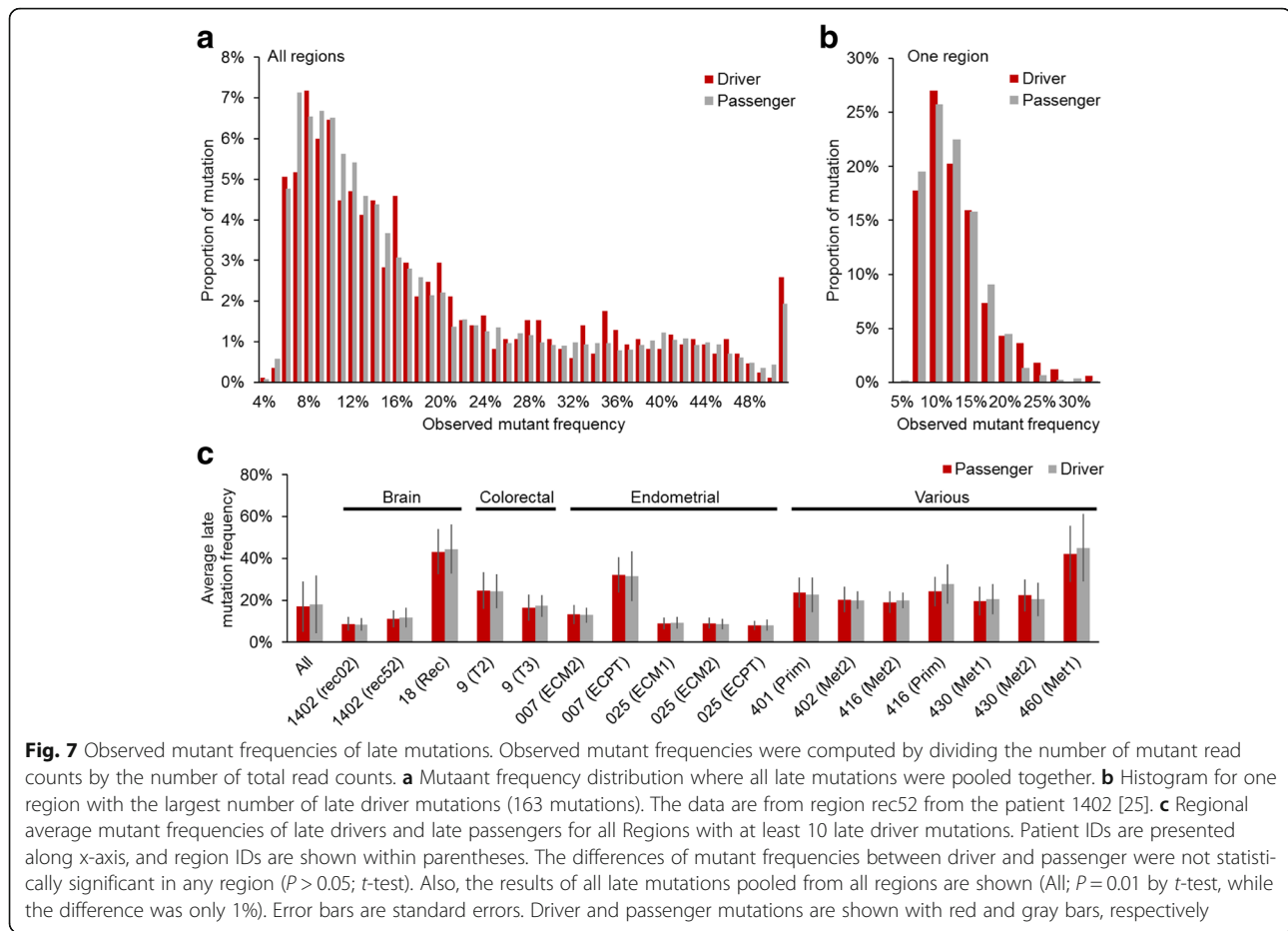


Fig. 6 Numbers of driver mutations and passenger mutations. The number of mutations were pooled for each study. **a** and **b** The fractions of driver and passenger mutations that are **(a)** early and **(b)** late. **c** The fractions of driver mutations over total mutations (driver and passenger mutations) for early (pink) and late (blue)



example, putative driver mutations were identified that are unique to a subset of the clones of an individual bladder tumor detected through single cell sequencing [48]. We expect more detailed studies in the future to test the patterns that we have observed in the meta-analysis presented here.

Conclusions

In a meta-analysis of genome variation data from multiple tumor in each patient, we find that the numbers of late driver mutations are substantial: they often exceed the number of early drivers. No previous study has conclusively demonstrated this pattern, even though they have indicated presence of driver mutations in tumors. These results implicate driver mutations in the continued development of aggressive tumor growth and in progression during later events such as recurrence, metastasis well beyond the initial founding of the tumor. Finally, these results highlight the importance of accounting for intratumor heterogeneity when evaluating the mutational histories of tumor cell populations.

Additional file

Additional file 1: Our designations of drivers using five annotation schemes, chromosomal positions, count of samples with mutant allele after Treeomics treatment, patient ID, and study information are shown. (XLSX 2387 kb)

Abbreviations

CCF: Cancer cell fraction; Indel: Insertions and deletions; M-seq: Multi-region sequencing; SNV: Single nucleotide variants; TCGA: The cancer genome atlas

Acknowledgements

We thank Dr. Heather Rowe for many scientific and editorial comments on this manuscript.

Funding

This research was supported in part by a grant from Temple University and from the National Institutes of Health to S.K. (LM012487-02) and S.M. (LM012758-01). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All the data sets analyzed were obtained from the supplementary materials of the published research articles [13, 14, 24–32]. Our designations of drivers using five annotation schemes, chromosomal positions, count of samples with mutant allele after Treeomics treatment, patient ID, and study information are available in the supplementary information.

Authors' contributions

SK conceived the study. SK, KG and SM designed the study. KG, SM, LAH, and BSS analyzed the data. SK, KG SM, and JPT interpreted results and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Institute for Genomics and Evolutionary Medicine, Sudhir Kumar, SERC 602A, 1925 N. 12th Street, Philadelphia, PA 19122, USA. ²Department of Biology, Temple University, Philadelphia, PA 19122, USA. ³Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06510, USA. ⁴Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06511, USA. ⁵Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06511, USA. ⁶Center for Genomic Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia.

Received: 26 July 2017 Accepted: 21 December 2017

Published online: 18 January 2018

References

- Ryu D, Joung JG, Kim NK, Kim KT, Park WY. Deciphering intratumor heterogeneity using cancer genome analysis. *Hum Genet.* 2016;135:635–42.
- Mroz EA, Rocco JW. The challenges of tumor genetic diversity. *Cancer.* 2017; 123:917–27.
- Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc Natl Acad Sci U S A.* 2008;105:13081–6.
- Hiley C, de Bruin EC, McGranahan N, Swanton C. Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. *Genome Biol.* 2014;15:453.
- Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature.* 2009;458: 719–24.
- Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* 2012;366:883–92.
- Youn A, Simon R. Estimating the order of mutations during tumorigenesis from tumor genome sequencing data. *Bioinformatics.* 2012;28:1555–61.
- Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, et al. The life history of 21 breast cancers. *Cell.* 2012;149:994–1007.
- Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013;45:1113–20.
- McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell.* 2015;27:15–26.
- Nowell PC. The clonal evolution of tumor cell populations. *Science.* 1976; 194:23–8.
- McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med.* 2015;7:283ra254.
- Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet.* 2014;46:225–33.
- Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, Seth S, Chow CW, Cao Y, Gumbs C, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science.* 2014;346:256–9.
- Newburger DE, Kashef-Haghighi D, Weng Z, Salari R, Sweeney RT, Brunner AL, Zhu SX, Guo X, Varma S, Troxell ML, et al. Genome evolution during progression to breast cancer. *Genome Res.* 2013;23:1097–108.
- Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, Rosner J, Shumansky K, Kalloger S, Senz J, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol.* 2013;231:21–34.
- Zhang LL, Kan M, Zhang MM, Yu SS, Xie HJ, Gu ZH, Wang HN, Zhao SX, Zhou GB, Song HD, Zheng CX. Multiregion sequencing reveals the intratumor heterogeneity of driver mutations in TP53-driven non-small cell lung cancer. *Int J Cancer.* 2017;140:103–8.
- Cao W, Wu W, Yan M, Tian F, Ma C, Zhang Q, Li X, Han P, Liu Z, Gu J, Biddle FG. Multiple region whole-exome sequencing reveals dramatically evolving intratumor genomic heterogeneity in esophageal squamous cell carcinoma. *Oncogene.* 2015;4:e175.
- Mehine M, Heinonen HR, Sarvilinna N, Pitkanen E, Makinen N, Katainen R, Tuupainen S, Butzow R, Sjoberg J, Aaltonen LA. Clonally related uterine leiomyomas are common and display branched tumor evolution. *Hum Mol Genet.* 2015;24:4407–16.
- Hardiman KM, Ulintz PJ, Kuick RD, Hovelson DH, Gates CM, Bhasi A, Rodrigues Grant A, Liu J, Cani AK, Greenson JK, et al. Intra-tumor genetic heterogeneity in rectal cancer. *Lab Invest.* 2016;96:4–15.
- Harbst K, Lauss M, Cirenajwis H, Isaksson K, Rosengren F, Torngren T, Kvist A, Johansson MC, Vallon-Christersson J, Baldetorp B, et al. Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma. *Cancer Res.* 2016;76:4765–74.
- Green MR, Gentles AJ, Nair RV, Irish JM, Kihira S, Liu CL, Kela I, Hopmans ES, Myklebust JH, Ji H, et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood.* 2013;121:1604–11.
- Kroigard AB, Larsen MJ, Laenkholm AV, Knoop AS, Jensen JD, Bak M, Mollenhauer J, Kruse TA, Thomassen M. Clonal expansion and linear genome evolution through breast cancer progression from pre-invasive stages to asynchronous metastasis. *Oncotarget.* 2015;6:5634–49.
- Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science.* 2014;343:189–93.
- Kim H, Zheng S, Amini SS, Virk SM, Mikkelsen T, Brat DJ, Grimsby J, Sougnez C, Muller F, Hu J, et al. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* 2015;25:316–327.
- Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med.* 2015; 21:751–9.
- Uchi R, Takahashi Y, Niida A, Shimamura T, Hirata H, Sugimachi K, Sawada G, Iwaya T, Kurashige J, Shinden Y, et al. Integrated multiregional analysis proposing a new model of colorectal cancer evolution. *PLoS Genet.* 2016; 12:e1005778.
- Gibson WJ, Hoivik EA, Halle MK, Taylor-Weiner A, Cherniack AD, Berg A, Holst F, Zack TI, Werner HM, Staby KM, et al. The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat Genet.* 2016;48:848–55.
- Stachler MD, Taylor-Weiner A, Peng S, McKenna A, Agoston AT, Odze RD, Davison JM, Nason KS, Loda M, Leshchiner I, et al. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat Genet.* 2015;47:1047–55.
- Lee JY, Yoon JK, Kim B, Kim S, Kim MA, Lim H, Bang D, Song YS. Tumor evolution and intratumor heterogeneity of an epithelial ovarian cancer investigated using next-generation sequencing. *BMC Cancer.* 2015;15:85.
- Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JM, Papaemmanuil E, Brewer DS, Kallio HM, Hognas G, Annala M, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature.* 2015;520:353–7.
- Zhao ZM, Zhao B, Bai Y, Iamarino A, Gaffney SG, Schlessinger J, Lifton RP, Rimm DL, Townsend JP. Early and multiple origins of metastatic lineages within primary tumors. *Proc Natl Acad Sci U S A.* 2016;113:2140–5.
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. Ensembl 2016. *Nucleic Acids Res.* 2016;44:D710–6.
- COSMIC: Catalogue of somatic mutations in cancer. <http://cancers.sanger.ac.uk/cosmic>.

35. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 2015; 43:D805–11.
36. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, Gao J, Socci ND, Solit DB, Olshen AB, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol.* 2016;34:155–63.
37. Yi S, Lin S, Li Y, Zhao W, Mills GB, Sahni N. Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat Rev Genet.* 2017;18:395–410.
38. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods.* 2013;10:1081–2.
39. Reiter JG, Makohon-Moore AP, Gerold JM, Bozic I, Chatterjee K, Iacobuzio-Donahue CA, Vogelstein B, Nowak MA. Reconstructing metastatic seeding patterns of human cancers. *Nat Commun.* 2017;8:14114.
40. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. Patterns of somatic mutation in human cancer genomes. *Nature.* 2007;446:153–8.
41. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, Aoki M, Hosono N, Kubo M, Miya F, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet.* 2012;44:760–4.
42. Chen L, Shern JF, Wei JS, Yohe ME, Song YK, Hurd L, Liao H, Catchpole D, Skapek SX, Barr FG, et al. Clonality and evolutionary history of rhabdomyosarcoma. *PLoS Genet.* 2015;11:e1005075.
43. Hong WS, Shpak M, Townsend JP. Inferring the origin of metastases from cancer phylogenies. *Cancer Res.* 2015;75:4021–5.
44. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, Liu S. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget.* 2013;4(11):1868–81.
45. Yao R, Zhang C, Yu T, Li N, Hu X, Wang X, Wang J, Shen Y. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Mol Cytogenet.* 2017;10:30.
46. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat Genet.* 2016;48:238–44.
47. Illingworth CJ, Mustonen V. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics.* 2011;189(3): 989–1000.
48. Li Y, Xu X, Song L, Hou Y, Li Z, Tsang S, Li F, Im KM, Wu K, Wu H, et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience.* 2012;1:12.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

