Rhee and Kim *BMC Medical Genomics* 2018, **11**(Suppl 2):29
https://doi.org/10.1186/s12920-018-0352-z

BMC Medical Genomics

RESEARCH Open Access

Population-based statistical inference for temporal sequence of somatic mutations in cancer genomes

Je-Keun Rhee[1] and Tae-Min Kim[1,2*]

From The 28th International Conference on Genome Informatics
Seoul, Korea. 31 October - 3 November 2017

Abstract

Background: It is well recognized that accumulation of somatic mutations in cancer genomes plays a role in carcinogenesis; however, the temporal sequence and evolutionary relationship of somatic mutations remain largely unknown.

Methods: In this study, we built a population-based statistical framework to infer the temporal sequence of acquisition of somatic mutations. Using the model, we analyzed the mutation profiles of 1954 tumor specimens across eight tumor types.

Results: As a result, we identified tumor type-specific directed networks composed of 2-15 cancer-related genes (nodes) and their mutational orders (edges). The most common ancestors identified in pairwise comparison of somatic mutations were *TP53* mutations in breast, head/neck, and lung cancers. The known relationship of *KRAS* to *TP53* mutations in colorectal cancers was identified, as well as potential ancestors of *TP53* mutation such as *NOTCH1*, *EGFR*, and *PTEN* mutations in head/neck, lung and endometrial cancers, respectively. We also identified apoptosis-related genes enriched with ancestor mutations in lung cancers and a relationship between *APC* hotspot mutations and *TP53* mutations in colorectal cancers.

Conclusion: While evolutionary analysis of cancers has focused on clonal versus subclonal mutations identified in individual genomes, our analysis aims to further discriminate ancestor versus descendant mutations in population-scale mutation profiles that may help select cancer drivers with clinical relevance.

Keywords: Somatic mutation, Cancer genome, Mutation accumulation

Background

Various types of genomic aberrations accumulate in cancer genomes and play roles in the development and progression of the disease [1]. It has long been recognized that cancer genomes undergo a stepwise progression in which they acquire somatic mutations in a sequential order during their evolution. This model is relatively well established in colorectal cancer genomes [2], and may be true for other types of cancer. Recent advances in high-throughput sequencing technologies have enabled screening of cancer genomes for well-known cancer-related genomic aberrations such as somatic mutations, DNA copy number alterations, and chromosomal translocations [3]. Genomic snapshots of human solid tumors can only be obtained by surgical intervention and such procedures have limitations for a full understanding of the temporal or longitudinal evolution of individual cancer genomes. While current cancer genome studies are mainly focused on the identification of significant recurrent genomic aberrations as potential cancer drivers, the

*Correspondence: tmkim@catholic.ac.kr
[1]Cancer Research Institute, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, 06591 Seoul, Republic of Korea
[2]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, 06591 Seoul, Republic of Korea

inference of acquisition order of somatic mutations may provide mechanistic insights into the evolution of the cancer genome and have potential clinical relevance.

Several studies have been proposed to investigate the order of acquisition of genomic alterations. For example, Attolini et al. proposed a mathematical approach to determine the sequential order of *APC*, *KRAS*, and *TP53* mutations in 70 colorectal cancer samples [4]. They estimated the mutation rate per allele and predicted the temporal sequences for mutations acquired in these genes. Hu et al. tried to identify tumor driver genes using association rule mining [5]. In addition, some researchers tried to estimate the tumor initiation time. Tomasetti et al. modeled the process of mutation accumulation and verified that a substantial number of somatic mutations may have appeared before the onset of neoplasia [6]. Also, Foo et al. investigated driver mutations in the evolutionary processes of mutation accumulation using healthy and tumor tissues [7]. However, most of these previous reports used binary genomic data (e.g., calls for presence or absence of mutations or copy number alterations) and did not exploit information regarding the clonality of mutations (e.g., clonal vs. subclonal mutations).

Here, we inferred the acquisition order of somatic mutations (hereafter in this study, we define somatic mutations as non-silent single nucleotide variations [SNVs] including missense, nonsense, and splice site mutations) based on information of the cancer cell fraction (CCF) measured for each mutation, using 1954 tumor specimens from eight major tumor types of the Cancer Genome Atlas (TCGA) consortium: 90 samples from bladder urothelial carcinoma (BLCA); 733 from breast invasive carcinoma (BRCA); 246 from colorectal adenocarcinoma (COADREAD); 265 from head and neck squamous cell carcinoma (HNSC); 42 from kidney renal clear cell carcinoma (KIRC); 290 from lung adenocarcinoma (KIRC); 118 from lung squamous cell carcinoma (LUSC); and 170 from uterine corpus endometrial carcinoma (UCEC). The CCFs, as variant allele frequencies (VAFs) adjusted for the tumor purity and global/local ploidy, are a measure of the clonality of given somatic mutations. CCFs have been used to distinguish clonal or subclonal mutations in individual cancer genomes [8]. In theory, clonal mutations represent early genomic events that have occurred in a founder cell and are maintained during the clonal proliferation whereas subclonal mutations represent late genomic events that are not yet fixed by clonal amplification or clonal sweeps. Under the infinite sites model of genome evolution with no homoplasy, somatic mutations with lower CCFs cannot occur earlier than those with higher CCFs [9]. Although a recent study showed that the mutation acquisition order affects cancer and cancer therapy [10], it is still largely unclear how to aggregate the information on individual genomes such as the CCFs to facilitate population-scale inference of temporal ordering of somatic mutations. In this study, we established a statistical model to infer the temporal order of somatic mutations observed across multiple cancer genomes and applied the method to a pan-cancer landscape of somatic mutations of eight major tumor types.

Methods

Study dataset

All experiments were carried out using publicly available TCGA pan-cancer data for eight tumor types, BLCA, BRCA, COADREAD, HNSC, KIRC, LUAD, LUSC, and UCEC. All mutation data were obtained from mutation annotation format (MAF) files with available sequencing read abundance of mutant and wildtype alleles to calculate VAFs. Among these somatic mutations, only the non-silent mutations (nonsense, mutation, and splice site SNVs) were extracted. We further selected mutations with minimum number of variant alleles ≥ 5 and minimum number of total alleles ≥ 30. The integer-level copy number, tumor ploidy and purity values estimated by ABSOLUTE [11] were downloaded from the Synapse website for TCGA pancancer analysis (https://www.synapse.org/#!Synapse:syn1703335) and used for the estimation of CCFs.

Estimating cancer cell fraction

The CCF is defined as the proportion of cancer cells harboring the mutations for each variant, and can be estimated using a method outlined by Landau et al. [12]. Briefly, for a single point mutation m_i at a sample n, $P\left(C_i^n\right)$, the posterior distribution for the CCF C_i^n is obtained from binomial distribution of the observed VAF over the expected VAFs calculated using a uniform grid of 100 CCF values ($C_i^n \in [0.01, 1]$), and subsequently normalized. Then, the probability mass function of the $P\left(C_i^n\right)$ is represented as:

$$P\left(C_i^n\right) = \sum_{k=0.01}^{1} P\left(C_i^n = k\right) \quad (1)$$

Statistical inference of mutational temporal order of somatic mutations

The mutational order for a pair of somatic mutation, m_i and m_j, was determined using a generalized likelihood ratio test (GLRT). This examines whether the occurrence of somatic mutation m_i precedes that of another somatic mutation m_j. Then, a null hypothesis H_0 and an alternative hypothesis H_1 follow:

H_0 : m_i is an ancestor of $m_j (m_i \rightarrow m_j)$

H_1 : m_i is not an ancestor of m_j

Suppose that there are a total of N samples, which have somatic mutations both in m_i and m_j. The CCF for m_i and m_j is represented as a set of independent and identically distributed (i.i.d.) variables, as $C_i = \left(C_i^1, C_i^2, \ldots, C_i^N\right)$ and $C_j = \left(C_j^1, C_j^2, \ldots, C_j^N\right)$, respectively. In the n-th sample among the total N, the evolutionary precedence of the two mutations, m_i and m_j, was approximated from the comparison of their CCFs, C_i^n and C_j^n, respectively. That is, $C_i^n \geq C_j^n$ implies that mutation m_i is an ancestor of the mutation m_j in the sample n. Suppose that a random variable D_{ij}^n represents the difference of two variables, C_i^n and C_j^n, and $\hat{C}_i^n = P\left(C_i^n\right)$ and $\hat{C}_j^n = P\left(C_j^n\right)$ are the estimated distribution of the CCFs at a mutation m_i and m_j, respectively, in the sample n. Then the two hypotheses can be re-written as follows:

$$\mathrm{H}_0 : \hat{D}_{ij}^n \geq 0$$

$$\mathrm{H}_1 : \hat{D}_{ij}^n < 0,$$

where $\hat{D}_{ij}^n$ is the estimated distribution of D_{ij} at the sample n. By the definition of GLRT under i.i.d. condition, the statistics are represented as:

$$\begin{aligned}\Lambda &= \frac{max_{H_0} L\left(\hat{D}_{ij}^1, \hat{D}_{ij}^2, ..., \hat{D}_{ij}^N\right)}{max_{H_0 \cup H_1} L\left(\hat{D}_{ij}^1, \hat{D}_{ij}^2, ..., \hat{D}_{ij}^N\right)} \\ &= \prod_{n=1}^{N} \frac{max_{H_0} L\left(\hat{D}_{ij}^n | D_{ij}^n \in [0, +\infty]\right)}{max_{H_0 \cup H_1} L\left(\hat{D}_{ij}^n | D_{ij}^n \in [-\infty, +\infty]\right)}, \end{aligned} \quad (2)$$

where $L(\cdot)$ is a likelihood function. Using the characteristics of the convolution of two independent random variables, the probability mass function of the variable D_{ij}^n $\left(D_{ij}^n = C_i^n - C_j^n\right)$ is expressed as

$$P(D_{ij}^n = z) = \sum_{k=0}^{1} P\left(C_i^n = k\right) P\left(C_j^n = -z + k\right) \quad (3)$$

Then, Eq. 2 is rewritten with the property of Eq. 3 as follows:

$$\Lambda = \prod_{n=1}^{N} \frac{max_{H_0} \sum_{z=0}^{+\infty} P\left(\hat{D}_{ij}^n = z\right)}{max_{H_0 \cup H_1} \sum_{-\infty}^{+\infty} P\left(\hat{D}_{ij}^n = z\right)} \quad (4)$$

Equation 4 with k ∈ [0.01, 1] is rewritten as

$$\Lambda = \prod_{n=1}^{N} \frac{max_{H_0} \sum_{z=0}^{1} P\left(\hat{D}_{ij}^n = z\right)}{max_{H_0 \cup H_1} \sum_{-1}^{1} P\left(\hat{D}_{ij}^n = z\right)} \quad (5)$$

For the convenience of the calculation, the statistics of Eq. 5 are changed to a logarithmic value:

$$\begin{aligned} log(\Lambda)_{I_{D_{max}<0}} &= \sum_{n=1}^{N} log\left(\frac{max_{H_0} \sum_{z=0}^{1} P\left(\hat{D}_{ij}^n = 1\right)}{max_{H_0 \cup H_1} \sum_{z=-1}^{1} P\left(\hat{D}_{ij}^n = 1\right)}\right) I_{\hat{D_{ij}^n}_{max}<0} \\ &= \sum_{n=1}^{N} \left(log\left(max_{H_0} \sum_{z=0}^{1} P\left(\hat{D}_{ij}^n = 1\right)\right)\right. \\ &\quad \left. - log\left(max_{H_0 \cup H_1} \sum_{z=-1}^{1} P\left(\hat{D}_{ij}^n = 1\right)\right)\right) I_{\hat{D_{ij}^n}_{max}<0} \end{aligned} \quad (6)$$

$I_{\hat{D}_{ij_{max}}^n<0}$ is an indicator function for whether the maximum value of the likelihood function for $\hat{D}_{ij}^n$ is observed at $z < 0$. $I_{\hat{D}_{ij_{max}}^n<0} = 0$ means that the numerator and denominator values are identical. The null hypothesis (H_0) is not rejected if the log(Λ) is significantly large.

The statistical tests were carried out using mutant gene pairs with number of cases > 10. The statistical cutoff for the log(Λ) was obtained by random re-arrangement of the original data to generate a background distribution of the GLRT statistics. The cutoff was determined as a 5th percentile value from the background distribution of the 100,000 randomized experiments.

Results

Co-occurring pairs of somatic mutations

We first examined genes with frequent mutations and their co-occurrence patterns using mutation profiles of 1954 patients across eight TCGA tumor types (BLCA, BRCA, COADREAD, HNSC, KIRC, LUAD, LUSC, and UCEC). A mean number of 120 somatic mutations (non-silent SNVs; missense, nonsense and splice site mutations) were observed (1 to 1597 mutations per case; median of 59 mutations; Table 1). Figure 1(a) shows the distribution of somatic mutations for 10 genes with the most frequent somatic mutations across all cases examined. To investigate gene pairs, we employed a scoring system of mutation co-occurrence. The score of the co-occurrence, SCORE-CO is calculated by summing the outputs of the

Table 1 Number of non-silent mutations in each tumor type

	Nonsense mutation	Missense mutation	Splice site mutation
BLCA	1430	14,685	560
BRCA	2512	30,499	815
COADREAD	2348	37,702	765
HNSC	2110	26,040	1179
KIRC	185	2592	513
LUAD	4310	53,898	3090
LUSC	2004	23,485	640
UCEC	1790	21,327	538

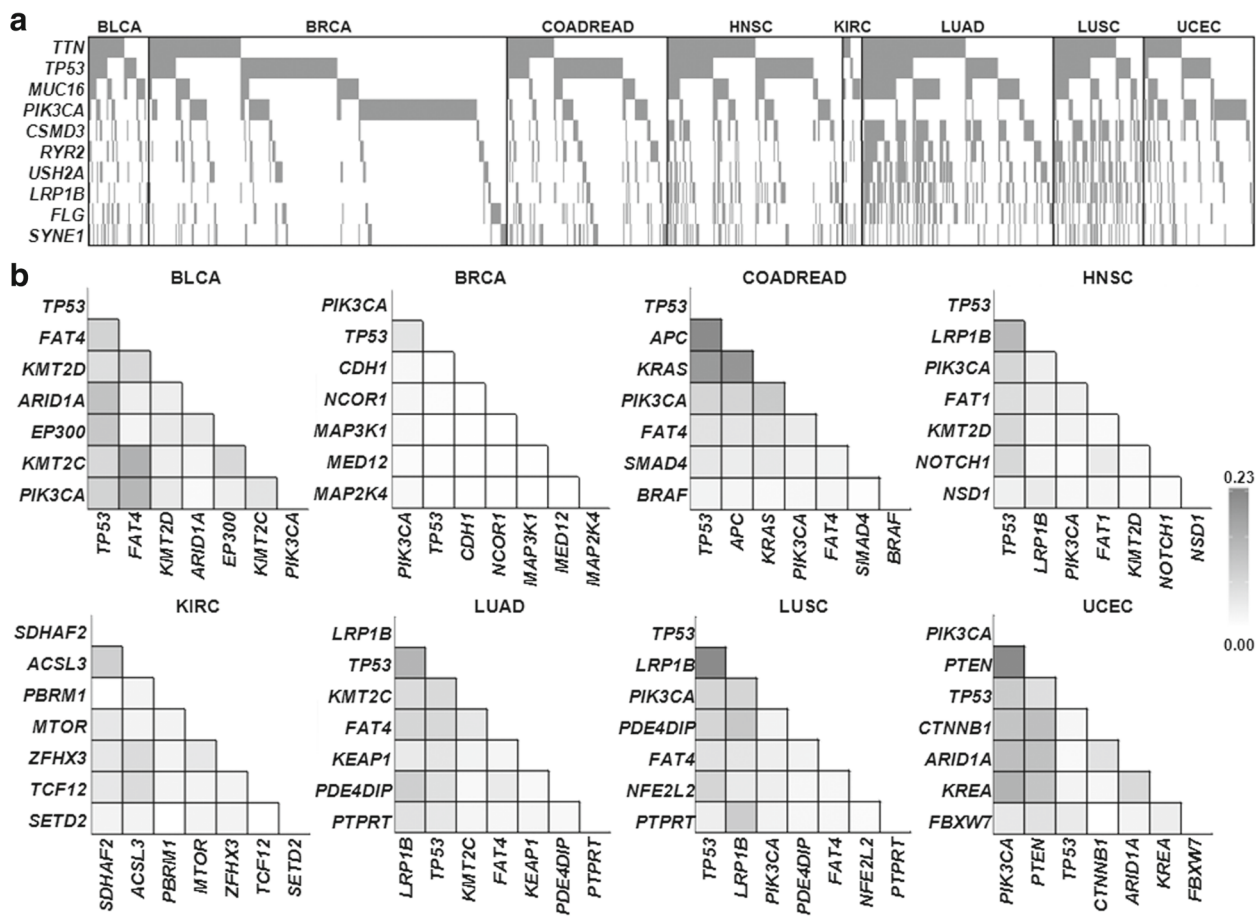

Fig. 1 Frequent somatic mutations and co-occurring mutation pairs. **a** Distribution of somatic mutations observed in all cases examined. The plot is shown for the top 10 frequently mutated genes across all cases. A tumor patient with a somatic mutation is marked as gray in each column. **b** Frequently observed gene pairs of somatic mutations are shown for each tumor type. The heatmaps of SCORE-CO are shown for the top 7 genes among CGC in a given tumor type. The color indicator of SCORE-CO is shown separately (right)

logical conjunction ('AND' gate) for the binary input data as the presence or absence of the somatic mutations in the given pair of two genes and by dividing the value by the number of total cases in the dataset. Table S1 shows the co-occurring pairs of somatic mutations observed in no fewer than 10 cases per tumor type (Additional file 1: Table S1). The gene pairs with high SCORE-CO include *TTN* and *MUC16* whose frequent mutations are largely due to their large gene size (36,800 and 14,500 amino acids, respectively) rather than their functional significance. Thus, we focused on mutations in known cancer-related genes or the Cancer Gene Census (CGC) [13] (Fig. 1(b)). The co-occurring mutation gene pairs with high SCORE-CO were tumor type-specific, e.g., gene pairs of *TP53* and *PIK3CA* were highly ranked in BLCA, BRCA, COADREAD, HNSC, LUSC, UCEC (SCORE-CO = 0.089 for 8 cases with the co-occurrence / total 90 patients, 0.055 for 40 cases, 0.085 for 21 cases, 0.083 for 22 cases, 0.085 for 10 cases, 0.106 for 18 cases, respectively) and to a lesser extent in LUAD (SCORE-CO = 0.024 for 7 cases). However, the pair of *TP53* and *PIK3CA* was not identified in KIRC. *LRP1B* mutants frequently co-occurred with *TP53* mutants in HNSC, LUAD and LUSC. *APC* mutants were frequently observed with *TP53* or *KRAS* mutations in COADREAD. In addition, some of the mutation occurrences were tumor type-specific, e.g., *PIK3CA* mutations showed co-occurrence with *FAT4* mutations with a high frequency in BLCA, but mainly co-occurred with *PTEN* mutations in UCEC.

Temporal sequence of mutations in cancer-related genes

To infer the temporal sequence of the somatic mutations in cancer-related genes, we first distinguished two types of mutations, clonal and subclonal mutations, based on CCF. The distinction was made by a criterion proposed by Landau et al. [12], and the measure of CCF was clearly higher for the clonal mutations than for subclonal mutations (P value $< 1.0 * 10^{-20}$).

Based on the CCF, we established a statistical framework for the population-based inference of temporal order between somatic mutations in a gene pair and applied the method for the mutation profiles from individual tumor types. Using the results from permutation tests to determine the minimum case number to identify the statistically significant mutation pairs with sequential orders (i.e., $log(\Lambda) \approx 0$ at the 5th percentile of the 100,000 re-sampling experiments at the small number of cases), we performed the test for all pairs of somatic mutations observed in no fewer than 10 cases (Additional file 2: Figure S1 (a)). Additional file 2: Figure S1 (b) shows the distribution of the number of the mutation pairs in each tumor type. The ancestor-descendant relationship in a mutation pair was then inferred by GLRT statistics and the significance was estimated for each direction. Figure 2 shows the mutation orders of cancer-related genes for six tumor types except for BLCA and KIRC, which did not have any significant mutant pairs within cancer-related genes (*P* value cutoff was 0.05).

The mutation order of gene pairs was largely distinct across tumor types, suggesting that the accumulation patterns or hierarchy of somatic mutations are lineage-dependent. However, some of the mutation pairs and their orders were consistently observed across tumor types. For example, a frequently co-occurred mutation pair of *TP53* and *LRP1B* (Fig. 1(b)) was observed as ancestor (*TP53*) - descendant (*LRP1B*) pairs (*TP53* → *LRP1B*) in COADREAD, HNSC, LUAD, and LUSC, with statistical significance (*P* value $= 3.0 * 10^{-5}$, 0.00044, 0.00029 and 0.00058, respectively). In addition, *KMT2C* mutation was consistently observed as a descendant of *TP53* mutations (*TP53* → *KMT2C*) in HNSC and LUAD (*P* value $= 3.0 * 10^{-5}$ and $5.0 * 10^{-5}$, respectively).

When we investigated mutation pairs in each tumor type, three ordered pairs were identified with statistical significance in BRCA (*P* value < 0.05) (Fig. 2). The hierarchy of the three genes (*TP53* → *PIK3CA* → *CDH1*) in BRCA suggests that *TP53* mutations represent early events that are followed by subsequent *PIK3CA* mutations (*P* value $= 0.034$), then *CDH1* mutations (*P* value $= 3.0 * 10^{-5}$). This mutation sequence can be functionally interpreted as follows: genomic integrity is disrupted with *TP53* mutations followed by cancer cell proliferation stimulated by *PIK3CA* mutations and the acquisition of later invasive/metastatic potential with *CDH1* mutations. This mutation order between *TP53* and *PIK3CA* was also found in HNSC (*P* value $= 3.0 * 10^{-5}$), suggesting that the *TP53* → *PIK3CA* axis may play important roles in the development of epithelial tumors. *NCOR1* mutation was also observed as an ancestor for *PIK3CA* mutation (*P* value $= 0.00036$) and it has been reported that functional inactivation of *NCOR1* as a *HDAC3* cofactor may produce genomic instability, which is functionally equivalent to the loss of *TP53* [14].

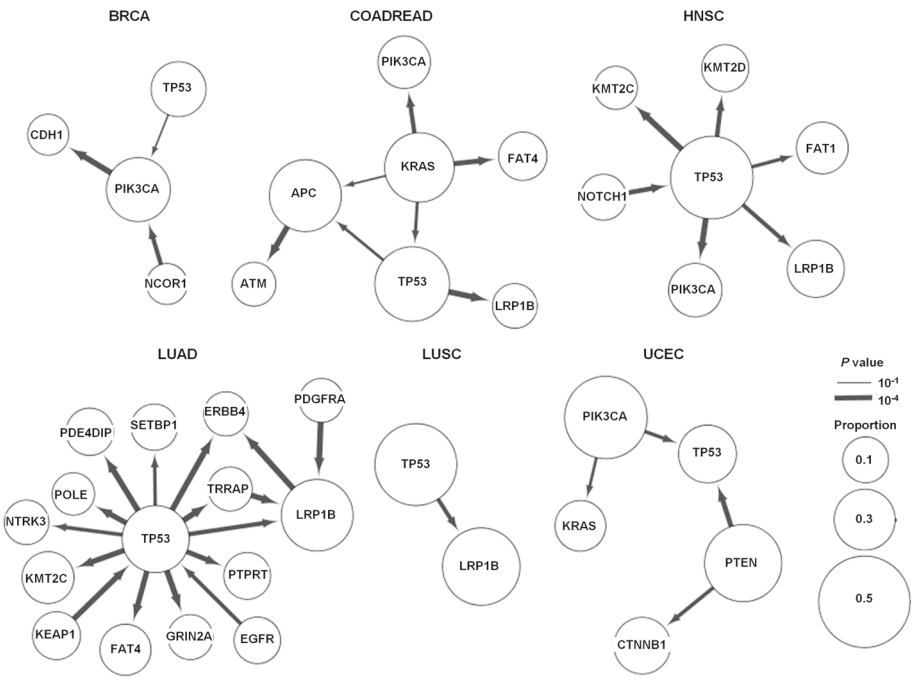

Fig. 2 Tumor type-specific networks representing the mutation order among cancer-related genes. The thickness of the edges is proportional to the $-logP$ value and the size of the nodes corresponds to the proportion of the genes with somatic mutations

For LUAD, 16 ordered pairs were identified with statistical significance. The elevated mutation abundance (mean of 211 mutations per LUAD case vs. mean of 120 mutations for total cases) and the relatively large size of the cohort (290 cases) may explain this number, but only one mutation pair was observed in LUSC with similar mutation abundance (average 221 mutations per cases) and a smaller number of cohorts (118 cases). *TP53* mutation appeared as a hub in the 16 edge-based network of LUAD and was identified as ancestor in most mutation pairs. *TP53* mutations have been implicated in tumor development and progression across many tumor types [15–17]. Our analysis also suggests that *EGFR* mutations may be earlier genomic events among the mutations in the LUAD pathogenesis [18]. A substantial fraction of *EGFR* mutations in LUAD are considered to be early addicted targets of targeted therapy [19, 20], suggesting that they represent early genomic aberrations together with *TP53* mutations. In the case of LUSC, the *TP53* $\rightarrow$ *LRP1B* ordered mutation pair was solely observed.

In HNSC, *NOTCH1* mutations may be earlier events than *TP53* mutations. Although it is mutated at a lower frequency than *TP53* in HNSC, *NOTCH1* has been highlighted as a potential cancer driver and tumor suppressor in HNSC [21].

For COADREAD, ordered pairs of somatic mutations involving *APC*, *KRAS* and *TP53* are observed and have been recognized to have pivotal roles associated with colorectal carcinogenesis [2]. Colorectal carcinogenesis is one of the well-established stepwise cancer progression models and involves sequential acquisition of *APC*, *KRAS*, and *TP53* mutations at colorectal dysplasia, adenoma, and carcinoma stages, respectively [22, 23]. Our inferred hierarchy from somatic mutations suggested that *KRAS* mutations were the earliest events in colorectal carcinogenesis. Given that our statistical model only considered the SNV, tumor suppressors that can be inactivated by chromosomal deletions, such as *APC*, may not be adequately assessed for order of mutation.

Accumulation of somatic mutations including non-CGC gene

We next performed analysis beyond the known cancer-related genes (Additional file 3: Table S2). For individual genes, we calculated SCORE-AN by subtracting the number of genes marked as a descendant from the number of genes marked as an ancestor in a given tumor type. SCORE-AN for the genes is provided in Additional file 4: Table S3 and Fig. 3. A large positive value of the SCORE-AN means that the corresponding gene is more likely to be an ancestor or early clonal event and can be regarded as a potential driver of the corresponding tumor type. In contrast, the genes with a large negative value would be passengers or late mutation events.

Consistent with our assumption, many of the genes with high SCORE-AN values or genes with more calls for ancestors were cancer-related genes listed in CGC (Fig. 3). For example, the genes with a positive SCORE-AN were *TP53* and *KMT2D* in BLCA and *TP53*, *PIK3CA* and *NCOR1* in BRCA. However, the two longest genes in the human genome, *TTN* and *MUC16*, which are likely to be passengers without putative roles generally showed negative values of SCORE-AN even though these two genes showed frequent mutations as shown in Fig. 1(a). In the case of UCEC, *TTN* and *TP53* showed a positive and negative value, respectively. For this tumor type, *TP53* mutation was marked as a descendant for *PIK3CA* and *PTEN* mutations. It was also noted that *TP53* mutations were also observed as descendants of *KRAS* mutations in COADREAD. For theses tumor types (UCEC and COADREAD), an unusual tendency for elevated mutation rates was observed and this might be responsible for the unique evolutionary position of *TP53* [24].

To further evaluate the potential functional significance of the SCORE-AN, we carried out Gene Set Enrichment Analysis (GSEA) [25] to identify the functional gene sets significantly enriched for genes with high or low SCORE-AN. Additional file 5: Table S4, Additional file 6: Table S5, Additional file 7: Table S6, and Additional file 8: Table S7 show the GSEA results only for LUAD, in which there was a high enough number of the ranked genes (427 genes) for analysis. The results for positively ranked genes on C5bp predefined gene sets (gene ontology, biological process in MSigDB), showed that most of the enriched gene sets were related to cell cycle and cellular transport (Additional file 5: Table S4). For gene sets enriched with low SCORE-AN, the significantly enriched genes were related to epidermal or epithelial development (Additional file 6: Table S5). On C2cp gene sets (canonical pathway in MSigDB), the results for the positively ranked genes or functions enriched with high SCORE-AN were also obviously enriched for cancer-related functionalities (Additional file 7: Table S6) but no gene sets with statistical significance were observed for genes with low SCORE-AN (Additional file 8: Table S7).

Mutational hotspots and accumulation order

We next investigated somatic mutations occurring on known mutation hotspots and their temporal mutation order. Chang et al. previously defined 470 mutational hotspots in 275 genes and we investigated all of the pairs of hotspot mutations to detect the temporal sequence of mutations on known hotspots [26]. However, the number of mutation pairs harboring the hotspot mutations was too small for our GLRT-based statistical models (number of cases ≤ 5).

It has been previously reported that *APC* mutations may initiate the process of colon cancer development

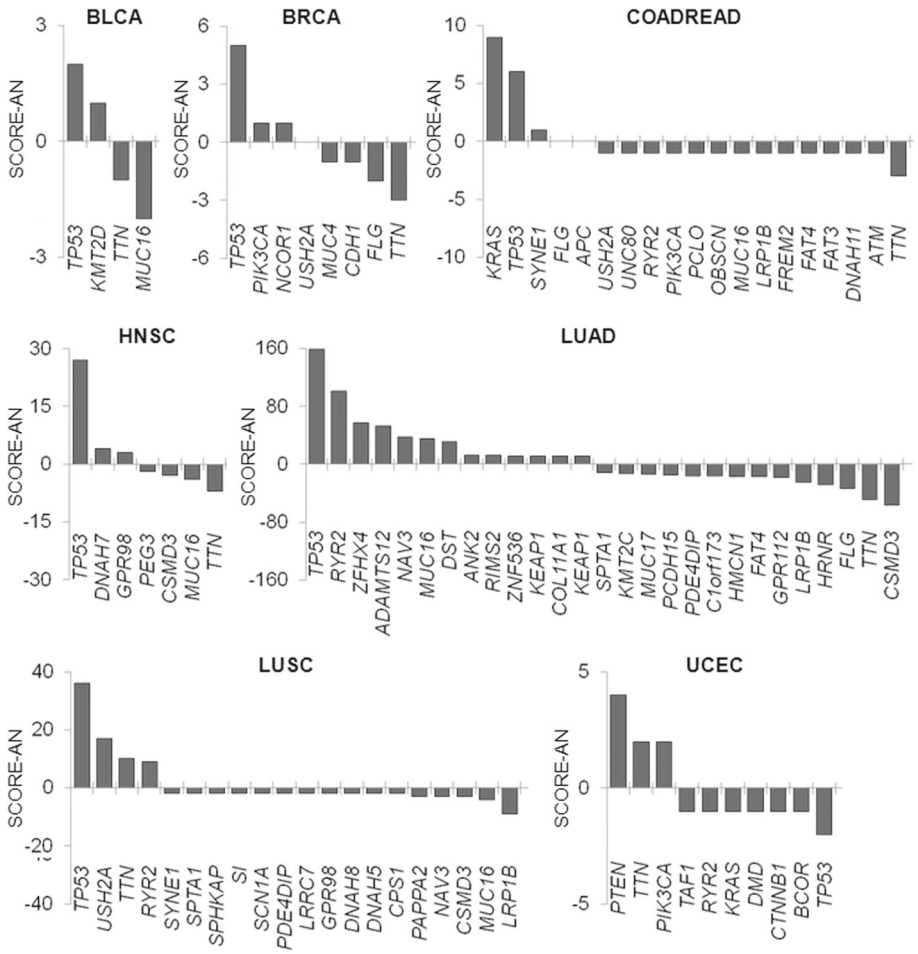

Fig. 3 Ancestral scores of genes in each tumor type. The figure shows SCORE-AN (y-axis), calculated as the number of the ancestors minus the number of descendants for each gene. Genes are sorted in descending order of the SCORE-AN for each tumor type (x-axis). For LUAD, genes with absolute number of SCORE-AN > 10 are plotted, and for LUSC genes with absolute number of SCORE-AN > 2 are shown

as one of the earliest genomic aberrations [27], but we did not obtain clear results for the early occurrence of *APC* mutations in the gene-level experiments shown in the previous section. We divided the *APC* mutations into hotspot mutations and non-hotspot mutations and then investigated the CCF distribution. The *APC* non-hotspot mutations showed relatively low CCF values compared with the *APC* hotspot mutations (Additional file 9: Figure S2(a)). When we further examined four cases harboring both *APC:Q1387* hotspot mutations and *TP53* mutations, the *APC:Q1387* hotspot mutations had higher CCF values compared with *TP53* mutation, and it is reasonable to assume that the *APC:Q1387* mutations would be an ancestor of the *TP53* mutations in these cases (Additional file 9: Figure S2(b)).

Discussion

The identification of known and novel cancer drivers with moderate-to-high population-level frequency of somatic mutations has been one of the major goals in cancer genome analyses [28, 29]. However, the temporal sequence of somatic mutations, i.e., which somatic mutations occurred earlier in the evolution of cancer genomes than others, is still largely unknown. Early- and late-occurring somatic mutations have different biological and clinical implications-the early addicted somatic mutations may serve as appropriate targets for therapeutic intervention while late-occurring cancer drivers have been associated with therapeutic resistance or disease progression. The distinction of such early and late genomic events, especially for somatic mutations, has been previously investigated using VAF or CCF. However, VAF- or CCF-based discrimination of early/clonal and late/subclonal mutations is limited to an individual genome and may miss information inferring the temporal relationship between mutations. To solve this problem, we built a GLRT-based statistical framework to determine the temporal sequence of somatic mutations from

mutation profiles of multiple individuals (population-level genomics data). This population-scale analysis may capture the temporal sequence of somatic mutations and identify the temporal sequence or hierarchy of somatic mutations of a given tumor type. Similar approaches have been previously proposed in which genomic data of multiple tumors (i.e., binary calls of chromosomal amplifications or deletions) at their fully transformed stages may be used to deduce the temporal sequence of genomic events (RESIC [4]); however, we extended this idea by exploiting the distinction of clonal versus subclonal mutations based on CCF estimates of individual mutations. We applied our method to publicly available mutation profiles of eight major human tumor types. For this, we carried out pairwise comparisons between somatic mutations based on a statistical test to infer the temporal order among them. Thus, it would be impossible to detect the effects of multiple factors on mutation acquisition. Although technical innovations have been proposed to solve this issue, e.g. single cell sequencing from a bulk tumor genome or longitudinal biopsies, these methods are largely limited in terms of cost or patient safety issues. Given that sequencing-based large-scale mutation profiles are currently available to the research community, such as those from the TCGA consortium used in our study, our method can be further applied to other datasets or tumor types. In addition, we assumed that C_i and C_j were under i.i.d. condition, and applied GLRT. Biologically, this assumption of independency between two mutations may not be valid given the crosstalk or interplay between genomic alterations in cancer cells. Other statistics for inference of mutational orders, such as an order statistic, can be considered as an alternative to GLRT.

Among the tumor types examined, we identified *TP53* mutations as a recurrently observed hub connected with other cancer-related genes, consistent with its prevalent and known roles in tumorigenesis across multiple cancer types [30]. Among the mutation pairs involving *TP53*, we observed the mutation pair of *KRAS* → *TP53* as a well-recognized mutation sequence in the stepwise colorectal carcinogenesis [2]. In addition, we recurrently observed the mutation pairs of *TP53* → *LRP1B* and *TP53* → *KMT2C* across multiple tumor types. Inactivation of *LRP1B* increased the invasive potential in an in vitro setting, implicating a role of *LRP1B* mutations in the later stages of carcinogenesis [31]. Whether the mutations in epigenetic modifiers are early or late events drivers is a subject of debate, with lines of evidence supporting early events for *TET2* mutations [32] or late events for *SETD2* mutations [33]. Our results suggest that *KMT2C* mutations are descendant genomic events relative to *TP53* mutations in LUAD and HNSC, but further experimental validation in terms of multiregion sequencing or other method is required. Moreover, as we expected, non-CGC genes were commonly observed as descendants of a CGC gene in multiple tumor types (Fig. 3 and Additional File 10: Table S8). For example, *USH2A* mutation was frequently observed in several tumor types as shown in Fig. 1a, but it was a late event occurring after *TP53* or *KRAS* mutation.

In the case of COADREAD, the mutation pair of *KRAS* → *APC* was observed even though it is generally recognized that *APC* mutations occur early, before *KRAS* and *TP53* mutations. One limitation of our methodology is that only SNVs available for CCF can be used as input of the algorithm. In the case of *APC*, chromosomal deletions or frameshifting indels may be also responsible for *APC* inactivation and our methods may not adequately evaluate the genetic hierarchy of tumor suppressors such as *APC*. When we limit the *APC* mutations to those on a known mutation hotspot (*APC:Q1387*) accompanying *TP53* mutations (four COADREAD cases), the CCF values of *APC* mutations were higher than those of *TP53* mutations suggesting that *APC* mutation may have occurred earlier than *TP53* mutation in those cases.

The population scale inference of mutational orders assumes that the mutation processes are uniform across the cases, or at least for the majority of cases. This assumption, and the related results, should be interpreted with caution since the sequence of mutation accumulation can be specific in individual cancer genomes and distinctive to patient subgroups, according to their tumor subtype or other clinical features. By collecting many more samples with information, the specific accumulation patterns (e.g., candidate sets of mutation orders) may be further investigated and might help elucidate individual features such as the treatment response.

Conclusions

In spite of several limitations, our results inferred a genetic hierarchy between somatic mutations as part of the cancer genome evolution. We found some ordered pairs of genes within cancer-related genes in each tumor and this information will provide mechanistic insights into the tumor initiation process. We also demonstrated that the scores of mutation co-occurrence (SCORE-CO) or ancestor/descendant ratio (SCORE-AN) may help identify or prioritize new candidates of driver mutations in each tumor. Furthermore, the study on mutation hotspot information may be more robust in that the hotspot mutations represent functionally relevant cancer drivers as shown in the example of *APC* mutations in COADREAD. In summary, our proposed statistical framework can be used to infer the temporal sequence of somatic mutations in population-scale cancer genomics data, providing information regarding the timing of mutation occurrence in given tumor types.

Additional files

Additional file 1: Table S1. SCORE-CO for the pair of genes in each tumor type. (XLSX 696 kb)

Additional file 2: Figure S1. Determination of the minimum sample size for the experiments (a) 5 percentile of the 100,000 random re-sampling experiments with the number of cases (b) Distribution of the frequency for the mutant gene pairs. (PDF 23 kb)

Additional file 3: Table S2. Ordered pairs with statistical significance in each tumor type. (XLSX 61 kb)

Additional file 4: Table S3. SCORE-AN for genes observed in each tumor type. (XLSX 22 kb)

Additional file 5: Table S4. Enrichment genesets on C5bp for genes with positive SCORE-AN. The results were obtained by GSEApreranked on C5bp (Gene Ontology, biological process). BLCA, BRCA and KIRC were no results with nominal P value < 0.05. (XLSX 16 kb)

Additional file 6: Table S5. Enrichment genesets on C5bp for genes with negative SCORE-AN. The results were obtained by GSEApreranked on C5bp (Gene Ontology, biological process). BLCA, BRCA, COADREAD, HNSC, KIRC and LUSC were no results with nominal P value < 0.05. (XLSX 10 kb)

Additional file 7: Table S6. Enrichment genesets on C2cp for genes with positive SCORE-AN. The results were obtained by GSEApreranked on C2cp (canonical pathway). BLCA, BRCA, COADREAD, HNSC and KIRC did not did not detect any genesets. (XLSX 12 kb)

Additional file 8: Table S7. Enrichment genesets on C2cp for genes with negative SCORE-AN. The results were obtained by GSEApreranked on C2cp (canonical pathway). BLCA, BRCA, COADREAD, KIRC and UCEC did not detect any genesets. (XLSX 11 kb)

Additional file 9: Figure S2. Hotspot mutation and CCF. (a) Maximum value of CCF in COADREAD. Each dot means a tumor patient with a somatic mutation in APC hotspots. (b) Distribution of CCFs for APC and TP53 mutations in COADREAD. The figure is plotted for the patients with APC:Q1387 hotspot mutation. Red is CCF distribution for APC and blue is for TP53. (PDF 35 kb)

Additional file 10: Table S8. Common temporal orders in multiple tumor types. (XLSX 11 kb)

Acknowledgements
The authors wish to acknowledge the financial support of the Catholic Medical Center Research Foundation made in the program year of 2015.

Funding
The publication cost of this article was funded by the Korea Health Technology R&D Project via the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant no. HI15C3224) and by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT, and Future Planning, Republic of Korea (grant no. NRF-2015R1C1A1A01053824) in part.

Availability of data and materials
The datasets analysed during the current study are available in the TCGA repository. Also, the source code used in the experiments is available at https://github.com/jkrhee/MutationTemporalSequence.

About this supplement
This article has been published as part of *BMC Medical Genomics* Volume 11 Supplement 2, 2018: Proceedings of the 28th International Conference on Genome Informatics: medical genomics. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-11-supplement-2.

Authors' contributions
JKR and TMK conceived the study, analyzed the experimental results, and wrote the manuscript. JKR collected the data, implemented the tool and performed the computational experiments. Both 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

Published: 20 April 2018

References

1. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153(1):17–37.
2. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5):759–67.
3. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010;11(10):685–96.
4. Attolini CS-O, 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;107(41):17604–9.
5. Hu W, Li X, Wang T, Zheng S. Association mining of mutated cancer genes in different clinical stages across 11 cancer types. Oncotarget. 2016;7(42):68270.
6. 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.
7. Foo J, Liu LL, Leder K, Riester M, Iwasa Y, Lengauer C, Michor F. An evolutionary approach for identifying driver mutations in colorectal cancer. PLoS Comput Biol. 2015;11(9):1004350.
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(283):283–5428354.
9. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci. 2016;113(37):5528–37.
10. Ortmann CA, Kent DG, Nangalia J, Silber Y, Wedge DC, Grinfeld J, Baxter EJ, Massie CE, Papaemmanuil E, Menon S, et al. Effect of mutation order on myeloproliferative neoplasms. N Engl J Med. 2015;372(7):601–12.
11. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. Absolute quantification of somatic dna alterations in human cancer. Nat Biotechnol. 2012;30(5):413–21.
12. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152(4):714–26.
13. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177–83.
14. Bhaskara S, Knutson SK, Jiang G, Chandrasekharan MB, Wilson AJ, Zheng S, Yenamandra A, Locke K, Yuan J-l, Bonine-Summers AR, et al. Hdac3 is essential for the maintenance of chromatin structure and genome stability. Cancer Cell. 2010;18(5):436–47.
15. Rivlin N, Brosh R, Oren M, Rotter V. Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis. Genes Cancer. 2011;2(4):466–74.
16. Petitjean A, Achatz M, Borresen-Dale A, Hainaut P, Olivier M. Tp53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene. 2007;26(15):2157–65.
17. Wang Y, Zhang Z, Lubet R, You M. A mouse model for tumor progression of lung cancer in ras and p53 transgenic mice. Oncogene. 2006;25(8):1277–80.
18. Anoosha P, Huang LT, Sakthivel R, Karunagaran D, Gromiha MM. Discrimination of driver and passenger mutations in epidermal growth

factor receptor in cancer. Mutat Res Fundam Mol Mech Mutagen. 2015;780:24–34.
19. Chan BA, Hughes BG. Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. Transl Lung Cancer Res. 2015;4(1):36.
20. Zhang Z, Stiegler AL, Boggon TJ, Kobayashi S, Halmos B. Egfr-mutated lung cancer: a paradigm of molecular oncology. Oncotarget. 2010; 1(7):497.
21. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in notch1. Science. 2011;333(6046):1154–7.
22. Conlin A, Smith G, Carey FA, Wolf CR, Steele RJ. The prognostic significance of k-ras, p53, and apc mutations in colorectal carcinoma. Gut. 2005;54(9):1283–6.
23. Smith G, Carey FA, Beattie J, Wilkie MJ, Lightfoot TJ, Coxhead J, Garner RC, Steele RJ, Wolf CR. Mutations in apc, kirsten-ras, and p53-alternative genetic pathways to colorectal cancer. Proc Natl Acad Sci. 2002;99(14):9433–8.
24. Supek F, Lehner B. Differential dna mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521(7550):81–4.
25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50.
26. 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(2):155–63.
27. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. Cell. 1996;87(2):159–70.
28. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
29. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, et al. Music: identifying mutational significance in cancer genomes. Genome Res. 2012;22(8): 1589–98.
30. Olivier M, Hollstein M, Hainaut P. Tp53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol. 2010;2(1):001008.
31. Prazeres H, Torres J, Rodrigues F, Pinto M, Pastoriza M, Gomes D, Cameselle-Teijeiro J, Vidal A, Martins T, Sobrinho-Simoes M, et al. Chromosomal, epigenetic and microrna-mediated inactivation of lrp1b, a modulator of the extracellular environment of thyroid cancer cells. Oncogene. 2011;30(11):1302–17.
32. Itzykson R, Kosmider O, Renneville A, Morabito M, Preudhomme C, Berthon C, Adès L, Fenaux P, Platzbecker U, Gagey O, et al. Clonal architecture of chronic myelomonocytic leukemias. Blood. 2013;121(12): 2186–98.
33. 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(10):883–92.