Decomposing the subclonal structure of tumors with two-way mixture models on copy number aberrations

An-Shun Tai[1], Chien-Hua Peng[1,*], Shih-Chi Peng[1], and Wen-Ping Hsieh[1,*]

1. Institute of Statistics, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.

* Corresponding author

Wen-Ping Hsieh[1] Institute of Statistics, National Tsing Hua University 101 sec. 2, Kwang-Fu Rd. , Hsin-Chu City, Taiwan, R.O.C. 30013; TEL: 886-35715131-33188; FAX:886-35728318;

E-mail:wphsieh@stat.nthu.edu.tw

Chien-Hua Peng [1] Institute of Statistics, National Tsing Hua University 101 sec. 2, Kwang-Fu Rd. , Hsin-Chu City, Taiwan, R.O.C. 30013; TEL: 886-35715131-33195; FAX:886-35728318;

E-mail: ariel0207@gmail.com

Abstract

Motivation: Multistage tumorigenesis is a dynamic process characterized by the accumulation of mutations. Thus, a tumor mass is composed of genetically divergent cell subclones. With the advancement of next-generation sequencing (NGS), mathematical models have been recently developed to de-mix tumor subclonal architecture among single-nucleotide variants (SNVs) from DNA sequencing data. However, somatic copy number aberrations (CNAs) also play critical roles in carcinogenesis. Therefore, further modeling subclonal CNAs composition would hold the promise to improve the analysis of tumor heterogeneity and cancer evolution.

Results: We developed a two-way mixture Poisson model, named CloneDeMix, for the deconvolution of read-depth information and inferred the subclonal copy number of each target region, mutational cellular prevalence (MCP), subclone composition, and the order in which mutations occurred in the evolutionary hierarchy. The performance of CloneDeMix was systematically assessed in simulations. Furthermore, we also demonstrated its applicability using head and neck cancer samples from TCGA. Our results inform about the extent of subclonal CNA diversity, and a group of candidate genes that probably initiate lymph node metastasis during tumor evolution was also discovered. Most importantly, these driver genes located at 11q13.3 which is highly susceptible to copy number change in head and neck cancer genomes. Briefly, this framework has implications for improved modeling of tumor evolution and the importance of inclusion of subclonal CNAs.

Availability and Implementation: The CloneDeMix R package is available at https://github.com/AshTai/CloneDeMix.

Introduction

Cancer, a dynamic disease, is characterized by unusual cells with somatic mutations. These mutations are caused by environmental factors accumulated during an individual's lifetime; this accumulation of mutational events results in a large degree of genetic heterogeneity among cancer cells. The intratumor heterogeneity causes difficulties in devising personalized treatment strategies.

To decipher intratumor heterogeneity, understanding how cancer evolves is a key step. The hypothesis for the somatic evolution of cancer was proposed in the 1970s (Nowell, 1976). It states that all tumor cells descend from a single founder cell, and cells with some advantageous mutations become more competitive than normal cells for growth and clonal expansion. This hypothesis could also be formed through random drift. Gradually, subsequent clonal expansion occurs, and the tumor evolves into an organization of multiple cell subpopulations. Understanding clonal evolution in cancer is one of the goals of cancer medicine (Kreso, et al., 2013). Presently, sequencing technology enables performing a large-scale molecular profiling of tumors to comprehend cancer development and determine disease progression. However, the process of evolution is not directly observed because tissues for measuring somatic mutations are typically obtained from patients at a single time point. Thus, the ancestral relationship among tumor subclones have to be inferred, and this is closely related to a well-studied problem, phylogenetic tree reconstruction. To construct a phylogenetic tree, the mutations in each cancer cell should be measured to infer evolutionary relationships among various cells. For addressing this concern, the current technology of single-cell sequencing seems appropriate (Hou, et al., 2012; Xu, et al., 2012).

However, this technology is not widely used because of some technical limitations and financial considerations (Wang, et al., 2014). Most studies on tumor evolution rely on DNA sequencing technology with a bulk tumor containing genetically different cells. Therefore, the cellular prevalence of each subclone have to be deconvolved through the relative read count information of the variants.

Single-nucleotide variants (SNVs) and copy number aberrations (CNAs) are widely used data types to study tumor evolution. Recently, studies inferring the population structure and clonal architecture have either focused on SNVs according to variant allele frequencies (VAFs) or on CNAs with read counts obtained through DNA sequencing (Ding, et al., 2013; Oesper, et al., 2013). Methods for either type of data can adopt the other type of data to improve their reconstruction, and most methods have developed corresponding computational tools.

The first category of method reconstructs models with only SNV data. AncesTree and clonality inference in tumors using phylogeny (CITUP) are the representatives of this category, and they build models based on heterozygous SNV to study cancer progression, assuming that the copy number is two (El-Kebir, et al., 2015; Malikic, et al., 2015). CITUP directly applies a somatic allele frequency to reconstruct the clonal evolutionary tree. It uses the quadratic integer programming and heuristic iterative method to address an optimization problem to select the tree structure. AncesTree defines a VAF factorization problem to determine the subclone numbers, the subclone proportions, and a tree describing the ancestral relationships among all subclones.

To relax the assumption of the normal copy number status, many studies have included CNAs to correct the baseline. Pyclone is one of the clonal inference

approaches, and it applies a hierarchical Bayes binomial distribution to model allelic counts (Roth, et al., 2014). This approach applies a Dirichlet process prior on group mutations and infers the posterior distribution to estimate the cellular prevalence, which is the fraction of cancer cells harboring a mutation. The concept of the Bayes clustering approach is also applied in both SciClone and Phylosub (Jiao, et al., 2014; Miller, et al., 2014). SciClone builds a Bayes beta mixture model on a VAF to track the spatial or temporal patterns of tumor evolution. Phylosub relies on the Dirichlet process mixture model with a tree-structured stick-breaking construction. It can provide clusters of mutations and simultaneously perform phylogenetic tree inference. The improved version of Phylosub is PhyloWGS, and the major improvement of PhyloWGS is to integrate VAFs with existing CNA-based subclonal reconstruction (Deshwar, et al., 2015).

Unfortunately, the aforementioned algorithms only considered abnormal copy number states but do not infer the evolution of copy number changes. If we do not account for clonal evolutionary architecture, the estimation of CNAs would be inaccurate and just reported as an average of the CNAs of all tumor subclones. Hence, in contrast to the SNV-based models, some studies focus on subclonal CNA heterogeneity (Fischer, et al., 2014; Li and Xie, 2014; Li and Xie, 2015; Oesper, et al., 2013; Oesper, et al., 2014; Yu, et al., 2016). They recognize that subclonal CNAs could technically improve the analysis accuracy. THetA is one of the most popular tools for subclonal copy number decomposition (Oesper, et al., 2013); it searches all possible combinations of copy numbers across all segments and applies the maximum likelihood approach to infer the most likely subclonal structures. However, THetA has an

identifiability concern, such that several solutions of subclone structures and copy number status levels can explain the read-depth information equally well (Li and Xie, 2014; Li and Xie, 2015).

Integrating other data, such as single-nucleotide polymorphisms, to jointly analyze tumor progression is a solution to the identification problem. The methods developed on the basis of these integrated data types constitute another category of cancer subclone reconstruction approaches. In 2014, Oesper et al. modified THetA to THetA2, which designs a probabilistic model of B-allele frequencies (BAFs) to solve the identification problem and simultaneously improves the efficiency of the algorithm (Oesper, et al., 2014). Furthermore, PyLOH resolves the identifiability problem by integrating CNAs and loss of heterozygosity (LOH) within a unified probabilistic model (Li and Xie, 2014). PyLOH is aimed at determining the contamination from normal cells and evaluating tumor purity, which is the fraction of tumor cells within a tumor tissue. Instead of tumor purity, MixClone improves PyLOH with a more delicate measurement of tumor progression, the subclonal cellular prevalence (SCP) (Li and Xie, 2015). The major concept of PyLOH and MixClone is to use the Poisson and binomial models simultaneously to analyze the read depth and BAFs. CloneCNA is another tool for automatically detecting clonal and subclonal somatic CNAs through the whole-exome sequencing of paired tumor–normal samples and by building a hidden Markov model (Yu, et al., 2016) on BAFs. CloneHD uses a different approach by adopting hidden Markov models to reconstruct cancer progression and integrates both somatic SNVs and CNAs (Fischer, et al., 2014).

Most of the mentioned methods that reconstruct the process of copy number

evolution assume heterozygous SNV sites within chromosome segments. This assumption facilitates the decomposition of clonal CNAs, but it necessitates ignoring segments without any somatic SNVs. Therefore, to more effectively address this concern, we developed a new algorithm, called CloneDeMix, that considers subclonal copy number changes when inferring the clonal evolutionary structures. It requires only the read-depth information of loci of any sizes no matter SNVs are included or not. The input can be a predefined segment of the chromosome or simply a single nucleotide locus. CloneDeMix is a two-way clustering model that clusters each locus into an appropriate copy number state and a most likely clonal group. The procedure can simultaneously evaluate all loci and regions. The algorithm uses information from all samples and loci simultaneously to infer clone progression and can efficiently reduce the identification bias.

In this study, we demonstrated the performance of the developed algorithm with simulation data and applied it to a head and neck cancer dataset from The Cancer Genome Atlas (TCGA). The simulation demonstrated the accuracy of clone identification and subclonal copy number change detection, particularly in early mutational events, which could be the candidate of driver mutations. The specificity of the copy number detection exceeded 98%, and the sensitivity was nearly 93.5%. These simulations support that our approach can successfully identify the copy number mutation and deconvolute its amplification or deletion state from the clonal architecture.

Our results obtained for 75 paired normal–tumor samples recapitulated most of findings reported in head and neck cancer (Koboldt, et al., 2012; Krijgsman, et al., 2014; Network, 2015; Zack, et al., 2013). The novel subclonal CNAs have also been identified,

and their subclonal structure has been shown to facilitate the discovery of driver mutations for advanced tumor progression. Furthermore, we provide evidence for the association between tumor heterogeneity and metastasis. A large heterogeneity tends to promote tumor metastasis. To sum up, CloneDeMix demonstrated ability to accurately identify subclonal CNAs and clarify intratumor heterogeneity. This flexible model relaxes the restriction to sequence mutation types. It is useful complement to other methods for cancer evolution studies.

Materials and Methods

Two-way Poisson mixture model

We delineated the structure of cellular evolution based on two concepts: SCP and mutational cellular prevalence (MCP), as shown in Figure 1. The SCP is defined as the fraction of cells that are relatively homogeneous and carry the same set of mutations. The MCP is defined as the fraction of cells that carry a certain mutation. The SCPs can be added to match the MCPs according to the evolutionary structure of subclones (Figure 2a). The evolution matrix, an upper triangular matrix, in Figure 2a provides information on the ancestral relationship among the subclones. Consider five subclones in this dummy example; their relationship is shown in the evolution tree in Figure 2a. The percentages indicate the corresponding SCPs. In this evolutionary structure, six mutations create five subclones. For example, locus A exists in every tumor subclone because of its presence at the top of the tree. Hence, the MCP of this locus can be calculated as the sum of all SCPs. By contrast, locus G is a later mutation and only

exists in the leaf subclone C4. The corresponding MCP is equal to the SCP of C4.

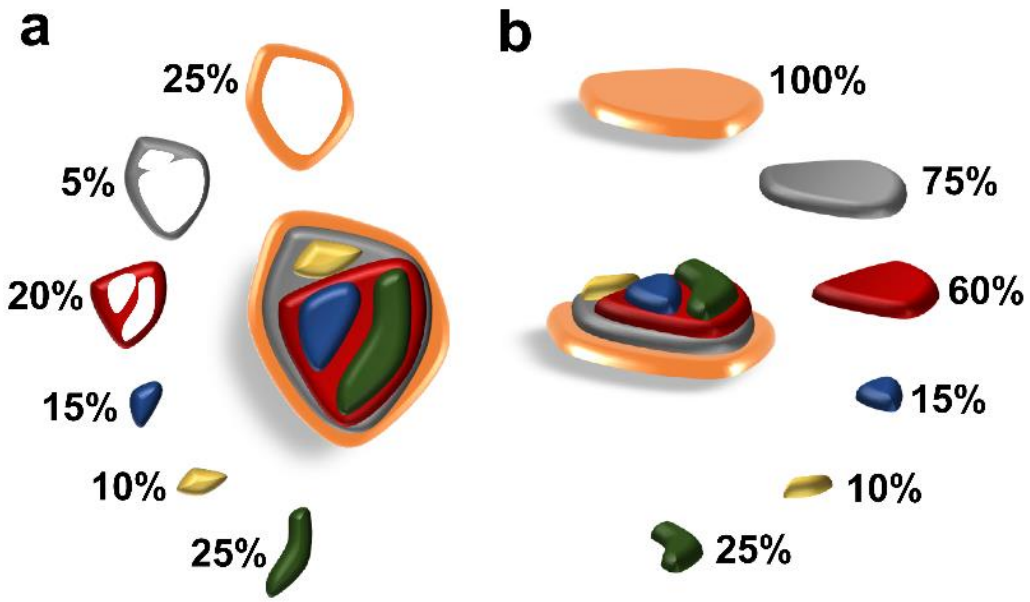
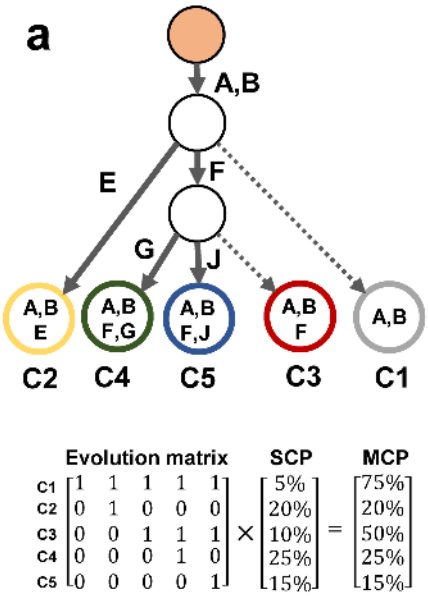


Figure 1 Illustration of SCP and MCP

A tissue has two decompositions. Panel (a) provides an overhead view that divides the cells into several groups according to their mutations. The cells in the same group are relatively homogeneous and carry the same set of mutations. The size of a group or the fraction of cells in it is called the SCP. In contrast to the SCP, panel (b) demonstrates the MCP, which is defined as the fraction of cells carrying a certain mutation.

The read depth of each locus is proportional to the copy number and MCP. To delineate the read depth of each somatic copy number variant into its copy number state and MCP, this study proposes a two-way mixture model (CloneDeMix). Any locus in a sample has only two cell types, namely normal and mutated cells; the proportion of both types differs across different loci. For example, locus E shows copy number changes in subclone C2 but not in the other subclones (Figure 2b). Hence, all other

subclones comprise the normal allele for locus E. Furthermore, locus F has copy number changes in C3, C4, and C5; hence, subclones C1 and C2 are classified as normal for this locus. CloneDeMix clusters all loci according to their copy number state and MCP. As shown in Figure 2c, all loci in this case are classified into five copy number states and simultaneously into five MCP levels. This thus results in 21 groups because we could not distinguish the MCP levels for the loci of two copies. The mean MCP for the five MCP groups is unknown and is estimated using CloneDeMix. Thus, CloneDeMix provides the copy number and MCP for each locus.



b

Clone (SCP)	Locus									
	A	B	C	D	E	F	G	H	I	J
C1 (5%)	3	1	2	2	2	2	2	2	2	2
C2 (20%)	3	1	2	2	0	2	2	2	2	2
C3 (10%)	3	1	2	2	2	4	2	2	2	2
C4 (25%)	3	1	2	2	2	4	0	2	2	2
C5 (15%)	3	1	2	2	2	4	2	2	2	3

MCP	75%	75%	0%	0%	20%	50%	25%	0%	0%	15%

c

Copy number state	MCP				
	r1 (75%)	r2 (50%)	r3 (25%)	r4 (20%)	r5 (15%)
0-copy				E	J
1-copy	B				
2-copy (Normal State)	C, D, H, I				
3-copy	A		G		
4-copy		F			

Figure 2 Two-way mixture model for inferring tumor progression by using copy numbers

(a) A dummy example for tumor progression of five distinct subclones. Six of the ten loci (A, B, E, F, G, and J) have gained or lost copy numbers, and the remaining loci (C, D, H, and I) show no copy number change. The mutation in each locus forms a new subclone. MCPs can be determined by multiplying the SCP and evolution matrix. (b) The copy number status of each locus is listed in the table, and the MCP of each locus is listed under the table. (c) Each locus belongs to one of the 21 clusters in CloneDeMix. The columns represent five MCP levels, and the rows represent five copy number states considered in the example.

The input in CloneDeMix is the read depth of each analyzed locus. When the locus represents a segment, such as an exon or a predefined amplicon, the average read depth is adopted. Let X_i be the read depth of locus i or the integer by truncating average read depth in region i, and assume that it follows a two-way mixture Poisson distribution.

$$P\left(X_i \middle| \{r_h\}, \{\pi_{kh}\}, a_{base,i}\right) = \sum_{k=1}^{m_1} \sum_{h=1}^{m_2} \pi_{kh} f_{kh}(X_i | r_h, a_{base,i}) \ \forall\, i$$

Each component $f_{kh}(X_i)$ in the model represents the distribution of read depths sampled from the k-th and h-th groups of the copy number state and MCPs, respectively. The read count for each combined group is specified as a Poisson distribution; the mean of this distribution is proportional to a function of the mutated copy number and the MCP. It is specified as

$$\mu_{hk} = a_{base,i} \times \left(2(1 - r_h) + c_k r_h \right),$$

where r_h is the MCP for the h-th group, c_k is the copy number of the k-th copy number group, and $a_{base,i}$ is a normalization number for locus i. The corresponding mixture weight is denoted as π_{kh}. Without further evidence, the copy number of the normal cells can be considered to be two in CloneDeMix. The number of groups for copy numbers and cellular proportions are prespecified as m_1 and m_2, respectively; we select m_1 and m_2 according to the Akaike information criterion (AIC).

Estimating MCPs and copy number by using expectation–maximization algorithm

The parameters include the normalization constants $a_{base,i}$, MCPs $\mathbf{r} = \{r_h\}$, and weights $\Pi = \{\pi_{kh}\}$ of CloneDeMix. The plug-in estimator of $a_{base,i}$ is estimated from the paired normal sample of each tumor sample. Because all samples are assumed to be globally normalized, and the sample-specific variation is removed before the analysis, the read depth of locus i in the normal sample represents an unbiased estimator of the mean read depth in the tumor sample when the copy number is two. Hence, we use half of the read depth of locus i in the paired normal sample as the estimator of $a_{base,i}$. In case of no paired normal sample, we suggest taking half of the sample mean across all existing normal samples to estimate $a_{base,i}$. All other parameters are estimated using the

expectation–maximization (EM) algorithm to approximate the maximum likelihood estimation (MLE). The full derivations used in the EM algorithm are presented in the supplement.

Determining the order of copy number variants

Based on the subclone size inferred using two-way cluster modeling, we can determine the order of any pairs of recurrent mutations existing across multiple samples. Herein, we use the notation MCP $\hat{r}_{ij}$ to indicate the estimated MCP of mutation *i* from the model of sample *j*. If this pair of mutations is recurrent in tumors with a fixed order, the relative size of their estimated MCPs should be consistent. For any two loci a and b with somatic mutations, the MCP profiles across p samples are $(\hat{r}_{a1}, \dots \hat{r}_{ap})$ and $(\hat{r}_{b1}, \dots \hat{r}_{bp})$. To determine whether the two mutations are highly related, the Wilcoxon signed-rank test can be applied to the profiles of the two mutations. In the event of significant inequality, when one mutation is more common in cells than the other mutation, it indicates a recurrent evolutionary order between the two mutations.

Results

We evaluated CloneDeMix by applying it to both simulated and real datasets. In the simulation study, we generated several scattered regions and assigned their copy number states and MCP groups. The performance was evaluated for the prediction accuracy. We also applied CloneDeMix on the whole-exon sequencing data of 75 pairs of tumor–control samples from TCGA to identify recurrent copy number mutation.

Simulation

The simulation considered four variant states of copy numbers, namely 0, 1, 3, and 4 copies. Four MCPs were included: 0.1, 0.3, 0.5, 0.7, and 0.9. Each combination was repeated three times, thus resulting in 60 regions with copy number changes. Furthermore, each region was assigned 20 bases generated with a Poisson distribution whose mean value was determined by its assigned copy number state and MCP. In addition to the mutated regions, 100 normal regions were scattered among the mutated regions; their copy number state was two. The simulation generated depths for a long region with 2200 bases for each of the 60 samples. CloneDeMix was subsequently applied to the simulation data to reconstruct respective copy number states and MCP groups. The entire simulation process was repeated 10,000 times to obtain a conclusion.

The simulation was performed to evaluate the model estimation accuracy. Table 1 shows the mean and standard deviation (SD) of simulation results for MCP estimation, and Figure 3 demonstrates the accuracy of assignments for copy number states. As presented in Table 1, the MCP estimates were very close to the underlying truth, indicating high performance for MCP estimation. Notably, the bias decreased as the ground value of the MCP increased. Detecting mutations of low cellular prevalence was relatively difficult because the signal was not adequately strong.

As illustrated in Figure 3, the accuracy of the copy number assignment under each condition was calculated from 10,000 simulations. The specificity of CloneDeMix was found to exceed 98%, and the sensitivity levels for amplification and deletion were 0.9365 and 0.9389, respectively. Thus, CloneDeMix represents a high specificity and efficiently controls false positive results. As mentioned in the discussion on MCP estimation, estimating mutations of low cellular prevalence was biased. These biased

MCPs directly caused the misclassification of the copy number state and reduced the model sensitivity. In conclusion, these simulations support that CloneDeMix can successfully identify the potential copy number mutations and deconvolute its amplification or deletion state from the clonal architecture.

Table 1. Mean and SD of MCP estimation

	Summary statistics	
True value	**Mean**	**SD**
0.1	0.100212	0.006723
0.3	0.300050	0.003620
0.5	0.499991	0.002218
0.7	0.699984	0.001451
0.9	0.900004	0.000881

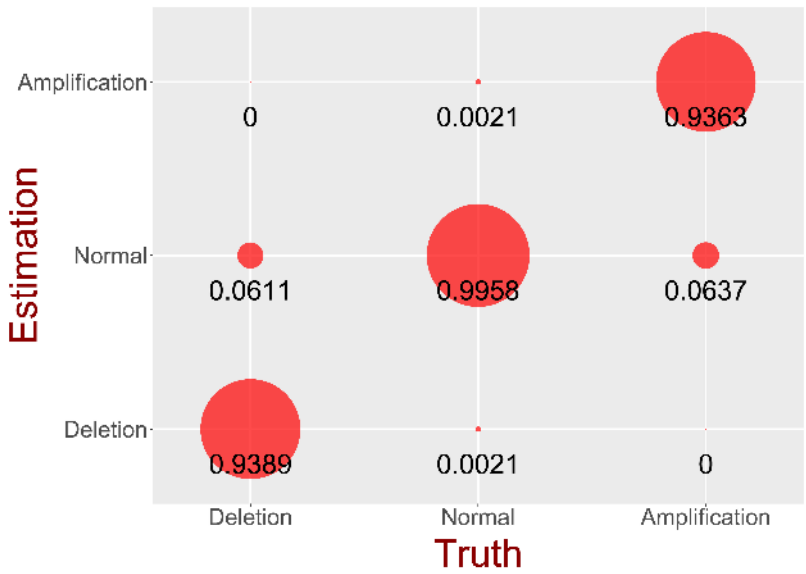


Figure 3 Result of copy number estimation

The size of the circle is proportional to the number of loci assigned to each estimated status from 10,000 simulations. The CNA status is divided into three conditions: deletion, amplification, and normal conditions.

Preprocessing of TCGA data

We analyzed the whole-exon sequencing data of 75 pairs of normal and head and neck tumor samples from TCGA (http://cancergenome.nih.gov/). This dataset includes a total of 20,846 genes with 180,243 exons. We assumed the copy number state of a single exon to be homogeneous. Each exon was represented by the mean read depth. The read-depth profile of a tumor sample was normalized with loess transformation against its paired normal sample. The baseline parameter $a_{base,i}$ for exon i was estimated from the paired normal sample by using half of the read depth of the normal sample at the same locus. Because the normal sample could also have an abnormal copy number status, we checked it against all other normal samples. The target normal sample was first normalized against all other normal samples by using the cyclic loess method and was subsequently processed through CloneDeMix to identify the copy number status at each locus. In this step, the average profile of all other normal samples could be taken as the baseline. If, for example, an abnormal copy number is found to be k, the raw read depth of this locus would be divided by k to provide the estimate of $a_{base,i}$ for tumor modeling.

Copy number distribution and clone structure

We applied CloneDeMix to each normalized sample and estimated the copy number state of each locus as well as the corresponding MCPs. Figure 4 shows the chromosomes that were mutated most frequently, and the results of all other chromosomes are shown in S1 Figure. S1 Figure presents the copy number events across 180,243 exons for each of the 75 tumor–control sample pair. The proportion of exons with a normal copy number was high in all samples, and it was close to 100% in the control samples. The proportion was significantly decreased in the tumor samples,

indicating considerable structural variations during cancer development.

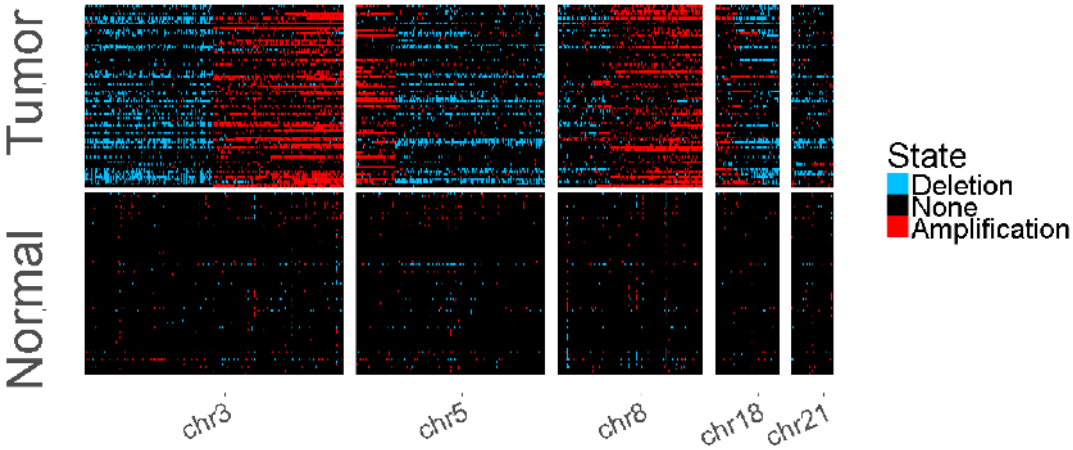


Figure 4 Copy number estimation of chromosomes with high mutation rates

The estimated copy number states for exons across the genome are presented by different colors. Light blue and red represent the deletion and amplification events, respectively. Black indicates no copy number changes.

On average, 4.7% and 8.7% of exons were estimated to have deletion and amplification, respectively. We also found that the exons located at 3p, 21p, and 18q were deleted most frequently, and the average proportions of deletion within these chromosomal arms were 19%, 17%, and 13%, respectively. Conversely, the estimated amplification frequently occurred at 3q, 8q, and 5p, with average frequency levels of 29%, 24%, and 23%, respectively. Previous studies have reported a loss of 3p and 8p as well as gains of 3q, 5p, and 8q not only in head and neck cancer but also in most tumors (Koboldt, et al., 2012; Krijgsman, et al., 2014; Network, 2015; Zack, et al., 2013); these results are concordant with our findings. Other novel subclonal CNA regions that were not reported in pan-cancer data analysis (Koboldt, et al., 2012;

Krijgsman, et al., 2014; Network, 2015; Zack, et al., 2013) were identified as multiple tumor subpopulations were considered (e.g. Deletion in 21p, S1 Fig). These subclonal CNA signals may be diluted in the previous studies that assumed only one homogeneous tumor clone and inferred CNAs from the average of whole tumor information. Our results confirm the identification strength of large-scale structural variations based on clonal evolution.

Figure 5 presents a summary of MCP estimation. The number of MCPs was determined using the model selection criterion AIC. We associated the number of subclones in the tumors with clinical outcomes because this number is closely related to tumor heterogeneity. The target phenotype included tumor invasion and metastasis, which are particularly ominous signs of poor prognosis in head and neck cancer. The association analysis was applied to only 68 samples because the clinical records of the other samples were incomplete in TCGA. Figure 6 illustrates the box plot of the number of MCP groups under each clinical group. In this figure, a sample is denoted as “NO” if no record of either invasion or metastasis exists; otherwise, it is denoted as “YES.” There appeared to be a tendency of increased tumor heterogeneity for tumors with invasion or metastasis. The variation of numbers of MCPs was larger for this group. To more comprehensively clarify this factor, we dichotomized the number of MCPs into two groups. The number of MCPs exceeding 4 indicated strong tumor heterogeneity, whereas a lower number indicated less heterogeneity. The contingency table (Table 2) shows the dichotomization of tumor heterogeneity associated with the clinical outcomes. The corresponding odds ratio was 3.64, and the p value evaluated with logistic regression was 0.029. For the samples with higher tumor heterogeneity, the

odds of invasion and metastasis were 3.64 times higher than those for the samples with lower tumor heterogeneity.

Table 2. Contingency table for dichotomization of tumor heterogeneity and clinical outcomes

	Clinical trait (Invasion/Metastasis)	
	NO	YES
Number of MCPs $\leq$ 4	16	22
Number of MCPs $>$ 5	5	25
Odds Ratio (p-value)	3.64 (p=0.029*)	

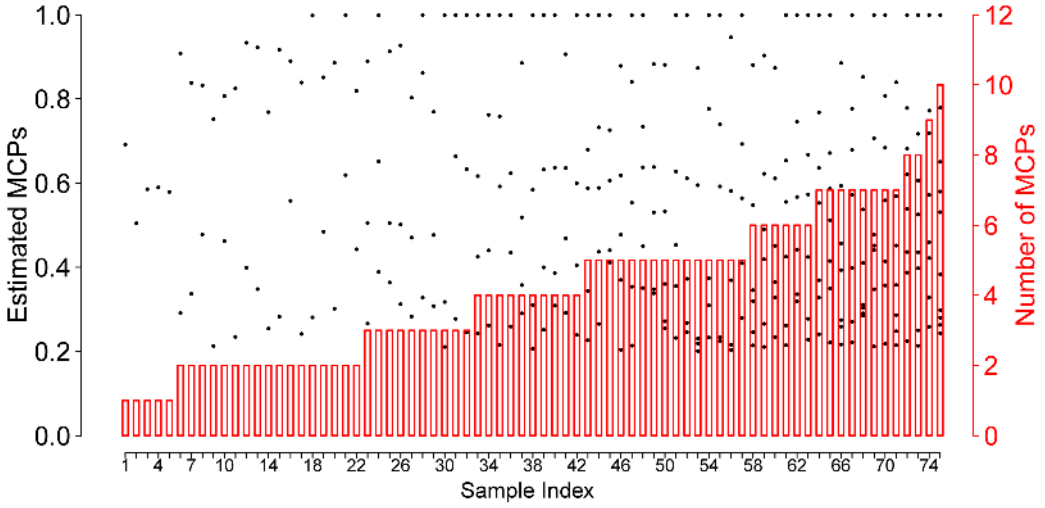


Figure 5 Summary of MCP estimation

The black dots indicate the estimated MCPs with respect to the left axis. The red bars represent the number of MCPs and range from 1 to 10 with respect to the right axis.

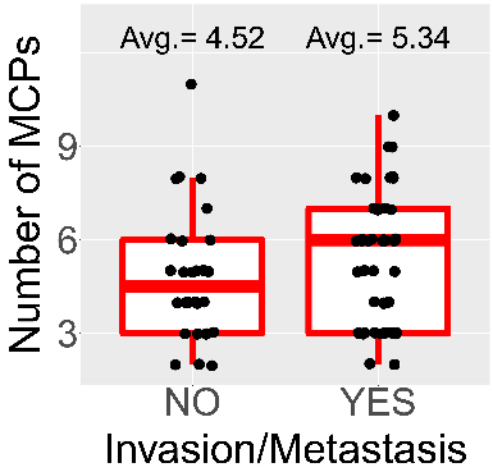


Figure 6 Comparison for the number of MCPs in different clinical groups
The box plot for the number of MCPs with and without invasion or metastasis. The number of MCPs in each sample is represented by a black point jittered around the box.

Inference of evolutionary order of mutations

As stated in the Methods section, we inferred the evolutionary order of recurrent variants with multiple samples. For easy comprehension, we demonstrated the result at the gene level through a series of summary steps. We first selected the genes with consistent amplification or deletion states in more than 25% of the exons within at least one sample. A total of 3,244 genes were included in this demonstration, and this set is called the background gene set. For each sample, the MCP of a gene was represented by the mean MCP of its exons. We then performed the Wilcoxon signed-rank test using the gene-level MCP of any two genes across the samples to derive all pairwise evolutionary relationships. For example, if the MCP of gene *i* was larger than that of gene *j* ($p = 0.05$), the mutation on gene *i* was considered to be more likely to be an earlier event than that on gene *j*. This relationship was marked as 1; otherwise, it was marked as 0. The 0–1 matrices of pairwise evolutionary relationships were separately

calculated for samples with and without nodal metastasis, and they could be denoted as a matrix M_{neg} and M_{pos}. The element of the matrix could be denoted as $M_{E,ij}$, representing the evolutionary order of mutations on gene *i* against mutations on gene *j* inferred with samples under the E condition, which could be neg or pos.

The evolutionary order matrix can be used to construct an evolutionary tree of all mutations. However, a tree of 3,244 genes is highly complicated, rendering the comparison of different clinical traits difficult. Therefore, for simplification, we proposed a progression score to summarize the relative position of a mutation on the evolutionary tree of tumor formation. The scores of a gene in advanced tumors can be compared with those of genes in newly developed tumors to select the ones that recurrently occur in the early stage of tumor development. The P score of gene *i* under condition E is thus defined as a summary statistic from the evolution matrix and is formulated as follows:

$$\text{P score (gene } i \text{ |E)} = \sum_{j \neq i} M_{E,ij} / \left(\sum_{j \neq i} M_{E,ij} + \sum_{k \neq i} M_{E,ki}\right). \quad (13)$$

Among all relations of gene *i* with other genes, the P score provides the number of times the mutation in gene *i* is more likely to occur before that in other genes. If a gene is close to the root of an evolutionary tree, its corresponding P score must be higher than that of its descending gene.

We compared the P score between the samples with and without nodal metastasis by plotting a scatter plot (Figure 7). Most background genes are more likely to be mutated in a random order that is not related to tumor progression. According to our P score definition, we postulated that the driving genes of lymph node metastasis would be scattered above the diagonal line, particularly in the upper left corner. The genes

above the diagonal line of the plot are more likely to acquire mutations at an earlier stage of tumor formation and occupy a significant proportion of the tumor at its advanced stage. This would yield higher P scores when only the samples with lymph node metastasis are considered. By contrast, the prevalence of mutations in those genes might be low in the samples without lymph node metastasis and hence yield lower P scores.

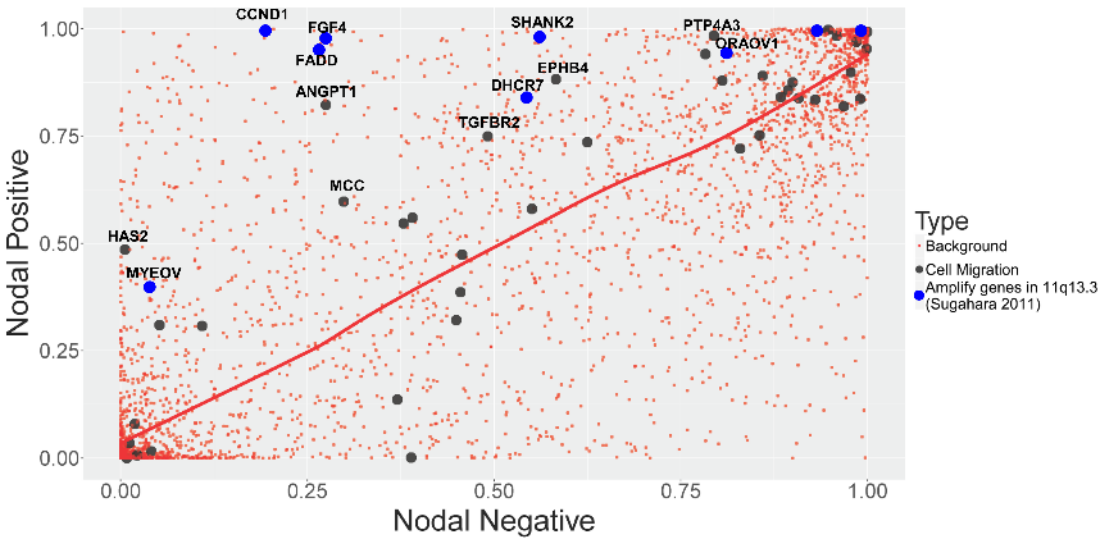


Figure 7 Scatter plot of P scores between nodal positive and negative samples
The red points indicate the background gene set. The red curve indicates the loess smoothing curve constructed using all points in the figure. Genes related to cell migration are marked in black. The genes from 11q13.3 are marked in blue. The literature supporting genes are labeled.

To confirm our conjecture, we selected the genes by their biological functions using ConsensusPathDB web (http://cpdb.molgen.mpg.de/) and investigated whether genes related to metastasis in the literature are more likely to be distributed above the diagonal line. Because cell migration is a crucial step in the metastatic cascade, we selected cell-

migration-related genes, which are marked as black in Figure 7. Consequently, we found that 43 genes had the function of cell migration. Most of these genes were distributed above the diagonal line of the P score scatter plot, whereas some were distributed below the diagonal line. Recurrent mutations in these cell migration genes are expected to be the driving forces for the initiation of lymph node metastasis, consistent with our observations. For example, HAS2 is a member of the gene family encoding putative hyaluronan synthases, which control the biosynthesis of hyaluronan and critically modulate the tumor microenvironment. Several studies have shown that the inhibition of HAS2 reduced the invasion of oral squamous cell carcinoma (Bernert, et al., 2011; Li, et al., 2015; Zhang, et al., 2016). Similar to HAS2, ANGPT1 is located in the upper left corner and has been recently investigated for the mechanism of lymph node metastasis (Fagiani, et al., 2011; Holopainen, et al., 2009; Li, et al., 2012; Mangiola, et al., 2016). ANGPT1 plays an important role in the regulation of vascular development and maintenance of vessel integrity. A study showed that the activity of ANGPT1 induced the enlargement of tumor blood vessels to facilitate tumor cell dissemination and increased the ability of metastasis in tumors (Li, et al., 2012). Fibroblast growth factor (FGF)-4 is another notable example. The P score of FGF4 significantly differs in nodal positive and negative patients. FGF4 is a member of the FGF family and possesses broad mitogenic and cell survival activities. It has been proposed to be involved in tumor growth, cell proliferation, and lymph node metastasis (Muller, et al., 1994; Qi, et al., 2016; Sugahara, et al., 2011). In contrast to the black genes located in the upper left corner of the plot in Figure 7, few studies have reported any relationship between the black genes located in the lower right corner and lymph node metastasis, although they have the same biological function. A complete literature

review of the genes associated with cell migration and tumor metastasis is presented in S1 Table. These observations suggest that our inference of the clonal evolutionary order is relevant and can be applied for identifying causal drivers.

Another notable observation is about the neighboring genes of FGF4. As mentioned, FGF4 is an important gene for driving lymph node metastasis. It is located in 11q13.3, which is frequently amplified in head and neck squamous cell carcinoma (Sugahara, et al., 2011). Sugahara also listed several other genes in 11q13.3 that are related to cancer development, namely *TPCN2*, *MYEOV*, *CCND1*, *ORAOV1*, *TMEM16A*, *FADD*, *PPFIA1*, *CTTN*, *SHANK2*, and *DHCR7*. We also assessed their status by using the P score analysis; the genes are indicated in blue in Figure 7. All these genes were above the diagonal line. Their corresponding P scores showed considerably significant differences between patients with and without nodal metastasis. Hence, we postulated that those genes in 11q13.3 are possibly related to lymph node metastasis in head and neck cancer. Several previous studies have confirmed this observation, as reported in S2 Table.

Discussion

In this study, we developed CloneDeMix for the deconvolution of tumor progression through high-throughput DNA sequencing data. The features of CloneDeMix are as follows. First, it reconstructs an evolutionary structure of copy number changes during tumorigenesis. Most existing methods for cancer evolution discuss the history of single-nucleotide changes and derive the potential driver genes.

However, the importance of CNAs is growing and its influence on disease and cancer development is clearly established (Valsesia, et al., 2013). Therefore, the reconstruction of copy number evolution in tumor progression is in demand. Second, CloneDeMix provides the MCP as a measure of the evolutionary structure. This measurement is used to estimate the fraction of cells containing a specific set of mutational events. According to the definition of the MCP, it provides a more direct evolutionary reconstruction than does the SCP, which is defined as the size of a subclone in a tumor. For instance, the MCPs of early mutations in cancer must exceed those of other mutations, but no such structural relationship exists for SCPs. Third, our model exhibits high flexibility. CloneDeMix can identify the copy number state of any type of variant, from a single nucleotide to a moderate size of regions. Furthermore, the model facilitates the simultaneous analysis of multiple types of targets because it depends on only the summary information of each locus.

The simulation study revealed that CloneDeMix can identify the current clonal structures of a tumor. The accuracy of copy number states was nearly 93%, and the MCP was also accurately restored (Table 1). Furthermore, the application of CloneDeMix to head and neck cancer data from TCGA yielded promising putative CNAs. The deletions observed on chromosomes 3p, 18q, and 21p and the amplifications on chromosomes 3q, 5p, and 8q are consistent with most cancer studies on copy number identification (Koboldt, et al., 2012; Krijgsman, et al., 2014; Network, 2015; Zack, et al., 2013). This observation strongly supports our CNA inference procedure.

When the estimation accuracy reaches a certain level, the most important concern

is to understand the relationship between tumor heterogeneity and disease progression. Tumor clone dynamics have been associated with clinical outcomes for different types of cancer (Maley, et al., 2006; McGranahan, et al., 2016; Park, et al., 2010). Our method provides a quantitative measurement of clonality, and it is associated with tumor invasion and metastasis development in TCGA database. Tumors with more subclones are a result of complex branched evolution, implying a series of adaptations to a new environment. These newly emerged subclones may contribute to metastatic initiation or acquire a new ability to invade the lymphatic or vascular system. Thus, the strong prognostic association of the number of MCPs with invasion or metastasis reinforces its clinical relevance; this index appears to be a novel feature for further exploration.

We established a novel score, the P score, for evaluating the order of a recurrent mutation in the evolutionary hierarchy by analyzing multiple samples. By comparing the P scores of a somatic variant between different clinical groups, we could identify the copy number mutations that occur early in the tumor stage and expand the accompanied subclones with time. The utility of P scores was also demonstrated in the head and neck cancer data according to the sample status of metastasis. Furthermore, we identified a group of genes that matched this condition. Specifically, the genes located at 11q13.3 are well known to be frequently amplified in head and neck squamous cell carcinomas. Their P scores in our analysis were particularly high for the samples with lymph node metastasis and relatively low for those without metastasis. Accordingly, those genes amplifications are potential causal mutations to drive metastatic cascade in head and neck cancer. Hence, screening for genes that differ considerably in their P scores is meaningful for driver gene detection.

The success of our approach highly depends on the coverage of DNA sequencing. A higher read depth can more efficiently reflect the clonal structure and copy number changes of different loci. Currently, CloneDeMix makes an independent assumption without considering the dependency among closely located loci. Hence, the neighboring segments are not grouped into the same copy number events. This can be an advantage as well as a disadvantage because there is no clear understanding about the range covered by a copy number event. Technically, we can still integrate the correlation structure into CloneDeMix to improve the flexibility; this is an ongoing project for our next version of the R package.

CloneDeMix can easily integrate different types of somatic mutations detected in sequencing data. For example, the well-studied SNVs carry extensive information about the clonal expansion in tumors. CloneDeMix can consider the copy number status of two alleles individually if the detection of each allele is optimized. Therefore, we expect CloneDeMix to be useful in understanding tumor heterogeneity and how it evolves to the current status. Moreover, CloneDeMix has high specificity for detecting early mutations in tumor progression; these early mutations would be good candidates for disease driver genes and targeted therapies.

Acknowledgments

This work was supported by the Ministry of Science and Technology [MOST 105-2118-M-007-001].

Reference

Bernert, B., Porsch, H. and Heldin, P. Hyaluronan synthase 2 (HAS2) promotes breast cancer cell invasion by suppression of tissue metalloproteinase inhibitor 1 (TIMP-1). *Journal of Biological Chemistry* 2011;286(49):42349-42359.

Deshwar, A.G., *et al.* PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology* 2015;16(1):35.

Ding, L., *et al.* Advances for studying clonal evolution in cancer. *Cancer letters* 2013;340(2):212-219.

El-Kebir, M., *et al.* Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 2015;31(12):i62-i70.

Fagiani, E., *et al.* Angiopoietin-1 and-2 exert antagonistic functions in tumor angiogenesis, yet both induce lymphangiogenesis. *Cancer research* 2011;71(17):5717-5727.

Fischer, A., *et al.* High-definition reconstruction of clonal composition in cancer. *Cell reports* 2014;7(5):1740-1752.

Holopainen, T., *et al.* Angiopoietin-1 overexpression modulates vascular endothelium to facilitate tumor cell dissemination and metastasis establishment. *Cancer research* 2009;69(11):4656-4664.

Hou, Y., *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* 2012;148(5):873-885.

Jiao, W., *et al.* Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics* 2014;15(1):35.

Koboldt, D.C., *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 2012;22(3):568-576.

Kreso, A., *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* 2013;339(6119):543-548.

Krijgsman, O., *et al.* Focal chromosomal copy number aberrations in cancer—Needles in a genome haystack. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 2014;1843(11):2698-2704.

Li, P., *et al.* Hyaluronan synthase 2 overexpression is correlated with the tumorigenesis and metastasis of human breast cancer. *International journal of clinical and experimental pathology* 2015;8(10):12101.

Li, T., *et al.* Molecular regulation of lymphangiogenesis in development and tumor microenvironment. *Cancer Microenvironment* 2012;5(3):249-260.

Li, Y. and Xie, X. Deconvolving tumor purity and ploidy by integrating copy number

alterations and loss of heterozygosity. *Bioinformatics* 2014:btu174.

Li, Y. and Xie, X. MixClone: a mixture model for inferring tumor subclonal populations. *BMC genomics* 2015;16(Suppl 2):S1.

Maley, C.C., *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics* 2006;38(4):468-473.

Malikic, S., *et al.* Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* 2015;31(9):1349-1356.

Mangiola, S., *et al.* Comparing nodal versus bony metastatic spread using tumour phylogenies. *Scientific Reports* 2016;6.

McGranahan, N., *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 2016;351(6280):1463-1469.

Miller, C.A., *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. 2014.

Muller, D., *et al.* Frequent amplification of 11q13 DNA markers is associated with lymph node involvement in human head and neck squamous cell carcinomas. *European Journal of Cancer Part B: Oral Oncology* 1994;30(2):113-120.

Network, C.G.A. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;517(7536):576-582.

Nowell, P.C. The clonal evolution of tumor cell populations. *Science* 1976;194(4260):23-28.

Oesper, L., Mahmoody, A. and Raphael, B.J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* 2013;14(7):R80.

Oesper, L., Satas, G. and Raphael, B.J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* 2014;30(24):3532-3540.

Park, S.Y., *et al.* Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *The Journal of clinical investigation* 2010;120(2):636-644.

Qi, L., *et al.* FGF4 induces epithelial-mesenchymal transition by inducing store-operated calcium entry in lung adenocarcinoma. *Oncotarget* 2016;7(45).

Roth, A., *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature methods* 2014;11(4):396-398.

Sugahara, K., *et al.* Combination effects of distinct cores in 11q13 amplification region on cervical lymph node metastasis of oral squamous cell carcinoma. *International journal of oncology* 2011;39(4):761.

Valsesia, A., *et al.* The growing importance of CNVs: new insights for detection and

clinical interpretation. *Frontiers in genetics* 2013;4:92.

Wang, Y., *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 2014;512(7513):155-160.

Xu, X., *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 2012;148(5):886-895.

Yu, Z., Li, A. and Wang, M. CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. *BMC bioinformatics* 2016;17(1):310.

Zack, T.I., *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* 2013;45(10):1134-1140.

Zhang, Z., *et al.* Hyaluronan synthase 2 expressed by cancer-associated fibroblasts promotes oral cancer invasion. *Journal of Experimental & Clinical Cancer Research* 2016;35(1):181.