Li and Xie *BMC Genomics* 2015, **16**(Suppl 2):S1
http://www.biomedcentral.com/qc/1471-2164/16/S2/S1

BMC Genomics

PROCEEDINGS Open Access

MixClone: a mixture model for inferring tumor subclonal populations

Yi Li[1], Xiaohui Xie[1,2,3*]

From The Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015)
HsinChu, Taiwan. 21-23 January 2015

Abstract

Background: Tumor genomes are often highly heterogeneous, consisting of genomes from multiple subclonal types. Complete characterization of all subclonal types is a fundamental need in tumor genome analysis. With the advancement of next-generation sequencing, computational methods have recently been developed to infer tumor subclonal populations directly from cancer genome sequencing data. Most of these methods are based on sequence information from somatic point mutations, However, the accuracy of these algorithms depends crucially on the quality of the somatic mutations returned by variant calling algorithms, and usually requires a deep coverage to achieve a reasonable level of accuracy.

Results: We describe a novel probabilistic mixture model, MixClone, for inferring the cellular prevalences of subclonal populations directly from whole genome sequencing of paired normal-tumor samples. MixClone integrates sequence information of somatic copy number alterations and allele frequencies within a unified probabilistic framework. We demonstrate the utility of the method using both simulated and real cancer sequencing datasets, and show that it significantly outperforms existing methods for inferring tumor subclonal populations. The MixClone package is written in Python and is publicly available at https://github.com/uci-cbcl/MixClone.

Conclusions: The probabilistic mixture model proposed here provides a new framework for subclonal analysis based on cancer genome sequencing data. By applying the method to both simulated and real cancer sequencing data, we show that integrating sequence information from both somatic copy number alterations and allele frequencies can significantly improve the accuracy of inferring tumor subclonal populations.

Background

Tumor genomes have been shown to present extensive cellular heterogeneity for decades since Nowell's original clonal theory for tumor progression [1]. Identifying tumor subclonal populations is important for both understanding the evolution of tumor cells, and for designing more effective treatments as pre-existing mutations occurring in some subclones could lead to drug resistance [2]. For example, a research in lymphocytic leukemia has shown links between the presences of driver mutations within subclones and adverse clinical outcomes [3].

With the advancement of next-generation sequencing (NGS) and launch of large-scale cancer genome sequencing projects [4], computational methods have recently been developed to infer tumor subclonal populations based on cancer genome sequencing data [5-9].

Most of these methods rely on sequence information from somatic point mutations, such as PyClone [5], EXPANDS [6], PhyloSub [7] and rec-BTP [8]. Methods in this category leverage the cluster pattern of allele frequencies at somatic point mutations to detect distinct subclonal populations. However, as the determination of somatic point mutations is imperfect and the inclusion of false-positives is unavoidable [10], deep sequencing with more than 100X coverage is often required for subclonal inferences with high sensitivity and specificity [5,7,8].

* Correspondence: xhx@ics.uci.edu
[1]Department of Computer Science, University of California, Irvine, CA 92697 US
Full list of author information is available at the end of the article

Other approaches utilizing the read depth information from genomic segments with somatic copy number alterations (SCNAs) to infer the cellular prevalences of subclonal populations have also been developed, such as THetA [9]. THetA explores all combinations of copy number changes across all segments to infer the most likely collection of subclonal populations [9]. However, with the copy number information alone, THetA suffers from the "identifiability problem", where distinct combinations of tumor purity and ploidy are able to explain the read depth information from SCNAs equally well [9]. Additionally, the running time of THetA scales exponentially with the number of genomic segments [9], and often takes a prohibitively long time to run under certain parameter settings.

In this article, we present a novel probabilistic mixture model, MixClone, to infer the cellular prevalences of subclonal populations. MixClone integrates both read depth information from genomic segments with SCNAs and allele frequency information from heterozygous single-nucleotide polymorphism (SNP) sites within a unified probabilistic framework. Such integrative framework has been shown to significantly improve the accuracy of tumor purity estimation in our previous work [11]. Here, we present that MixClone achieves two major advantages compared to the existing methods that (i) it does not require deep sequencing data, (ii) it resolves the identifiability problem. To demonstrate MixClone's utility, we conducted simulation studies and showed that it outperforms existing methods. We also applied MixClone on a breast cancer sequencing dataset [12], and showed that it was able to discover subclonal events not reported before.

Methods

In this section, we introduce the generative mixture model of MixClone, which is an extension of our previous work on tumor purity estimation[11]. First, we introduce the notations for input data. Then, we describe the probabilistic models for sequence information of both SCNAs and allele frequencies. Finally, we combine these two types of data into a single likelihood model, and describe an algorithm to solve the model.

Basic notations

The raw input data for MixClone are two aligned whole genome sequencing read sets of paired normal-tumor samples and a genome segmentation file based on the tumor sample. Following the notations from our previous work [11], we assume the tumor genome has been partitioned into J segments. We also assume there are I_j heterozygous SNP sites within segment j in the corresponding normal genome, and use (i, j) to index SNP site i within segment j. For each SNP site (i, j) we define the A allele to be the reference allele and the B to be the alternative allele, with respect to the reference genome. We also use a superscript N to denote data from normal samples and superscript T to denote data from tumor samples. Overall, the observed data are summarized in the following notations [13]:

b_{ij}^N = number of reads mapped to the B allele in the normal sample at site (i, j).
d_{ij}^N = reads depth of the normal sample at site (i, j).
D_j^N = total number of reads mapped to segment j of the normal sample.

The notations for the observed data from tumor samples are similarly defined, e.g. D_j^T denotes total number of reads mapped to segment j of the tumor sample.

Modeling SCNAs

Next, we describe the probabilistic model for SCNAs data. For each segment j, we define an allelic configuration H_j to represent its underlying allele-specific copy number status. For example, if the absolute copy number of segment j is 2, then the compatible allelic configurations are PP, MM and PM, where P and M denotes the paternal and maternal allele of the tumor genome, respectively. Since PP and MM are not distinguishable based on sequence information alone as the reference human genome is not phased, we define the set of all possible allelic configuration as

$$H_j \in \mathcal{H} = \{\emptyset, \text{P/M}, \text{PP/MM}, \text{PM}, \text{PPP/MMM}, \text{PPM/PMM}\} \quad (1)$$

assuming the maximum copy number for each segment is 3. The corresponding copy number associated with each allelic configuration in $\mathcal{H}$ is then

$$n_h = \{0, 1, 2, 2, 3, 3\} \quad (2)$$

MixClone allows the user to specify the maximum copy number and the default value is 6 in the released package [11]. We further assume there are K subclonal populations within the tumor sample, each of which has an associated cellular prevalence $\varphi_k \in 0[1]$. The subclonal type of each segment j is denoted as

$$Z_j \in \mathcal{Z} = \{1, 2, \cdots, K\} \quad (3)$$

representing one of the K possible subclonal populations. Given the allelic configuration $H_j = h$ and the subclonal type $Z_j = k$, the average copy number of segment j within the tumor sample, taking into account the subclonal cellular prevalence φ_k, is

$$\bar{C}_j = \phi_k n_h + (1 - \phi_k)2 \quad (4)$$

Based on the Lander-Waterman model [14], the probability of sampling a read from a given segment j

depends on three main factors: 1) its copy number, 2) its total genomic length, and 3) its mappability, which depends on factors such as repetitive sequence and GC content [9]. For each segment j, we associate a coefficient θ_j to account for the effect of its mappability and genomic length. Thus the expected read counts mapped to segment j, which is denoted as λ_j, is proportional to $\bar{C}_j\theta_j$. For example, for segment x and segment y, we have

$$\frac{\lambda_x}{\lambda_y} = \frac{\bar{C}_x\theta_x}{\bar{C}_y\theta_y} \tag{5}$$

Because the mappability coefficients (θ_j's) matter only in a relative sense, we take $\theta_x/\theta_y = D_x^N/D_y^N$, as these segments should have the same sequence properties between the normal and tumor samples.

Additionally, to determine the absolute value of λ_j, we curate a list of segments which contain no loss of heterozygosity according to their allele frequencies information. Based on the observed number of reads mapped to each segment, we further remove "outlier" segments from the list if their copy numbers are different from the bulk of the segments' copy numbers in the list. Finally, we call the remaining segments in the list as "baseline segments" and denote the set of these segments as S. We assume the allelic configurations of all the baseline segments are PM with copy number $n_s = 2$. Other possible allelic configurations for baseline segments, which have equal copy numbers for each allele (e.g. ϕ, PPMM), are likely to be rare, and currently we do not model them. Then based on n_s, we specify λ_j as follows

$$\lambda_j = \frac{1}{|S|}\sum_{s\in S}\frac{\bar{C}_j\theta_j}{n_s\theta_s}D_s^T \tag{6}$$

where D_s^T denotes the number of reads mapped to segment s of the tumor sample.

Finally, we model the number of reads mapped to segment j in the tumor sample as a Poisson distribution, given H_j and Z_j

$$D_j^T|H_j,\ Z_j \sim \text{Poisson}\ (\lambda_j) \tag{7}$$

Details on curating the baseline segments are given in Supplementary, Additional file 1.

Modeling allele frequencies

Next, we describe the probabilistic model used for allele frequencies of heterozygous SNP data. For each SNP site i within segment j, we denote its tumor genotype as G_{ij}, which is selected from the set of all possible tumor genotypes up to a maximum copy number alteration, e.g.

$$\mathcal{G} = \{\phi, \text{A}, \text{B}, \text{AA}, \text{AB}, \text{BB}, \text{AAA}, \text{AAB}, \text{ABB}, \text{BBB}\} \tag{8}$$

assuming the maximum copy number is 3. The corresponding B allele frequencies (BAF) for all the genotypes in$\mathcal{G}$ are

$$\mu_g = \{\frac{1}{2}, \epsilon, 1-\epsilon, \epsilon, \frac{1}{2}, 1-\epsilon, \epsilon, \frac{1}{3}, \frac{2}{3}, 1-\epsilon\} \tag{9}$$

in which, $\varepsilon \ll 1$ is a small random deviation accounting for general sequencing errors. We choose $E = 0.01$, which is equivalent to a Phred quality of 20 [15].

Given the tumor genotype $G_{ij} = g$, the allelic configuration $H_j = h$, and the subclonal type $Z_j = k$, the average BAF of site (i, j) within the tumor sample, taking into account the subclonal cellular prevalence ϕ_k, is

$$\bar{\mu}_{ij} = \frac{\phi_k n_h \mu_g + (1-\phi_k)2\mu_0}{\phi_k n_h + (1-\phi_k)2} \tag{10}$$

in which $\mu_0 = 0.5$ is the BAF of heterozygous SNP sites in the normal sample. Finally, we model the distribution of the B allele count b_{ij}^T at site (i, j) as a binomial distribution, given G_{ij} , H_j and Z_j

$$b_{ij}^T|d_{ij}^T, G_{ij}, H_j, Z_j \sim \text{Binomial}(d_{ij}^T, \bar{\mu}_{ij}) \tag{11}$$

Combining SCNAs and allele frequencies

Now, we combine sequence information from both SCNAs and heterozygous SNP sites. For all the heterozygous SNP sites within the same segment, their genotypes should be consistent with the underlying allelic configuration of the segment. We model this consistency through a predefined conditional probability $Q_{gh} = \mathbb{P}(G_{ij} = g|H_j = h)$. If the genotype g is inconsistent with the allelic configuration h, e.g. `AA` is inconsistent with `PM`, we assign a small probability σ as Q_{gh}, otherwise we assign equal probabilities to genotypes that are consistent with the allelic configuration.

Conditional on the underlying allelic configuration H_j and subclonal type Z_j, the probability of observing B allele read count b_{ij}^T at site (i, j) is given as

$$\mathbb{P}(b_{ij}^T|H_j = h, Z_j = k) = \sum_{g\in\mathcal{G}} Q_{gh}\mathbb{P}(b_{ij}^T|G_{ij} = g, H_j = h, Z_j = k) \tag{12}$$

We assume that conditional on the allelic configuration H_j , the B allele read counts $\{b_{ij}^T\}_{i=1}^{I_j}$ at different sites within the same segment j are independent of each other, and are also independent of the total read count D_j^T of the segment. Then, the joint probability of observing the two types of read counts information of segment j is

$$\begin{aligned}&\mathbb{P}(D_j^T, \{b_{ij}^T\}_{i=1}^{I_j}|H_j = h, Z_j = k)\\ =&\mathbb{P}(D_j^T|H_j = h, Z_j = k)\times\prod_{i=1}^{I_j}\sum_{g\in\mathcal{G}} Qgh\mathbb{P}(b_{ij}^T|G_{ij} = g, H_j = h, Z_j = k)\end{aligned} \tag{13}$$

Likelihood model

We have specified the joint distribution of the two types of read counts information of segment j. We then further model the allelic configuration H_j and the subclonal type Z_j of segment j as random variables that follow categorical distributions

$$H_j|\rho_j \sim \text{Categorical}(\rho_j) \tag{14}$$

$$Z_j|\pi \sim \text{Categorical}(\pi) \tag{15}$$

$\rho_j = (\rho_{j\varnothing}, ..., \rho_{j\text{PPM/PMM}})$, where $\rho_{jh} = \mathbb{P}(H_j = h)$ is the probability of observing h as the allelic configuration of segment j. $\pi = (\pi_1, ..., \pi_K)$, where $\pi_k = \mathbb{P}(Z_j = k)$ is the probability of observing subclonal type k for all the segments. The model parameters Θ is defined as

$$\Theta = (\{\rho_j\}_{j=1}^{J}, \{\pi_k\}_{k=1}^{K}, \{\phi_k\}_{k=1}^{K}) \tag{16}$$

And the model likelihood of observing all the data is then

$$\begin{aligned}
&\mathbb{P}(\{D_j^T\}_{j=1}^{J}, \{b_{ij}^T\}_{i=1,j=1}^{I_j,J}|\Theta) \\
&= \prod_{j=1}^{J}\sum_{k=1}^{K}\sum_{h\in\mathcal{H}} \mathbb{P}(Z_j = k)\mathbb{P}(H_j = h)\mathbb{P}(D_j^T|H_j = h, Z_j = k) \\
&\times \prod_{i=1}^{Ij}\sum_{g\in\mathcal{G}} Q_{gh}\mathbb{P}(b_{ij}^T|G_{ij} = g, H_j = h, Z_j = k) \\
&= \prod_{j=1}^{J}\sum_{k=1}^{K}\sum_{h\in\mathcal{H}} \pi_k\rho_{jh}\frac{\lambda_j^{D_j^T}e^{-\lambda_j}}{D_j^T!} \\
&\times \prod_{i=1}^{I_j}\sum_{g\in\mathcal{G}} Q_{gh}\binom{d_{ij}^T}{b_{ij}^T}\bar{\mu}_{ij}^{b_{ij}^T}(1-\bar{\mu}_{ij})d_{ij}^T - b_{ij}^T
\end{aligned} \tag{17}$$

We use Expectation-Maximization (EM) algorithm [16] to find the maximum likelihood estimation of Θ. The complete details of the EM updates are given in Supplementary, Additional file 1.

Model selection

One of the key issues in subclonal analysis is to determine the number of subclonal populations K. PyClone and PhyloSub use posterior sampling methods to estimate K [5,7], while THetA requires users to specify K as an input [9]. Since the probabilistic model of MixClone is a generative mixture model, the model complexity and the corresponding log-likelihood increases as K increases. Therefore, we use a criterion based on the increase of the log-likelihood to select K. Practically, Mix-Clone allows the user to specify K. If K is not specified, MixClone runs the mixture model five times with different K in range of 1 to 5. We denote the log-likelihoods under the five different settings as $\{L_K\}_{K=1}^{5}$, and the total log-likelihood increase as

$$\Delta = L_5 - L_1 \tag{18}$$

If $|\Delta/L_1| < 0.01$, which means the ratio of total log-likelihood increase is less than 0.01, MixClone predicts there is no subclonal event in the tumor sample and selects $K = 1$ as the number of subclonal populations. If $|\Delta\ /L_1| \geq 0.01$, MixClone further calculates another quantity

$$\delta_i = |L_i - L_1|\Delta, i \in [2, 5] \tag{19}$$

which is the cumulative log-likelihood increase from $K = 1$ to $K = i$ as a percentage regarding to the total increase Δ. If $\delta_i \geq 0.9$ and $\delta_{i-1} < 0.9$, MixClone selects $K = i$ as the number of subclonal populations.

In practice, we suggest users use this criterion as a heuristic guide when analyzing real data, and determine the number of subclonal populations in conjunction with regard to other external information.

MixClone software package

Figure 1 is the general workflow of MixClone. MixClone is a comprehensive software package, including subclonal cellular prevalences estimation, allelic configuration estimation, absolute copy number estimation and a few visualization tools. This package is implemented in Python and is built on top of the PyLOH package, previously released by us [11]. It also utilizes some features from the software package JointSNVMix [13], which have been explicitly indicated in the source code.

Results

In this section, we evaluate the performance of MixClone on both simulated and real datasets and compare its performance with two published algorithms: (i) PyClone, a method based on somatic point mutations, and (ii) THetA, a method based on somatic copy number alterations.

Results from simulated data

To generate simulation data, we simulated ten sets of NGS reads from chromosome 1 of artificial paired normal-tumor samples, each with 60X coverage. Heterozygous SNP sites from dbSNP [17] were inserted to the reference human genome to create the artificial normal genome. Both heterozygous SNP sites and somatic point mutations from [18] were inserted to the reference human genome to create artificial tumor genomes. Five of the artificial tumor genomes contain two subclonal populations and the other five contain three subclonal populations. Each artificial tumor genome was randomly assigned with segmentations, allelic configurations and subclonal cellular prevalences. We used segmentations based on both ground truth and BIC-seq [19] as the input for MixClone. We used ground truth somatic point

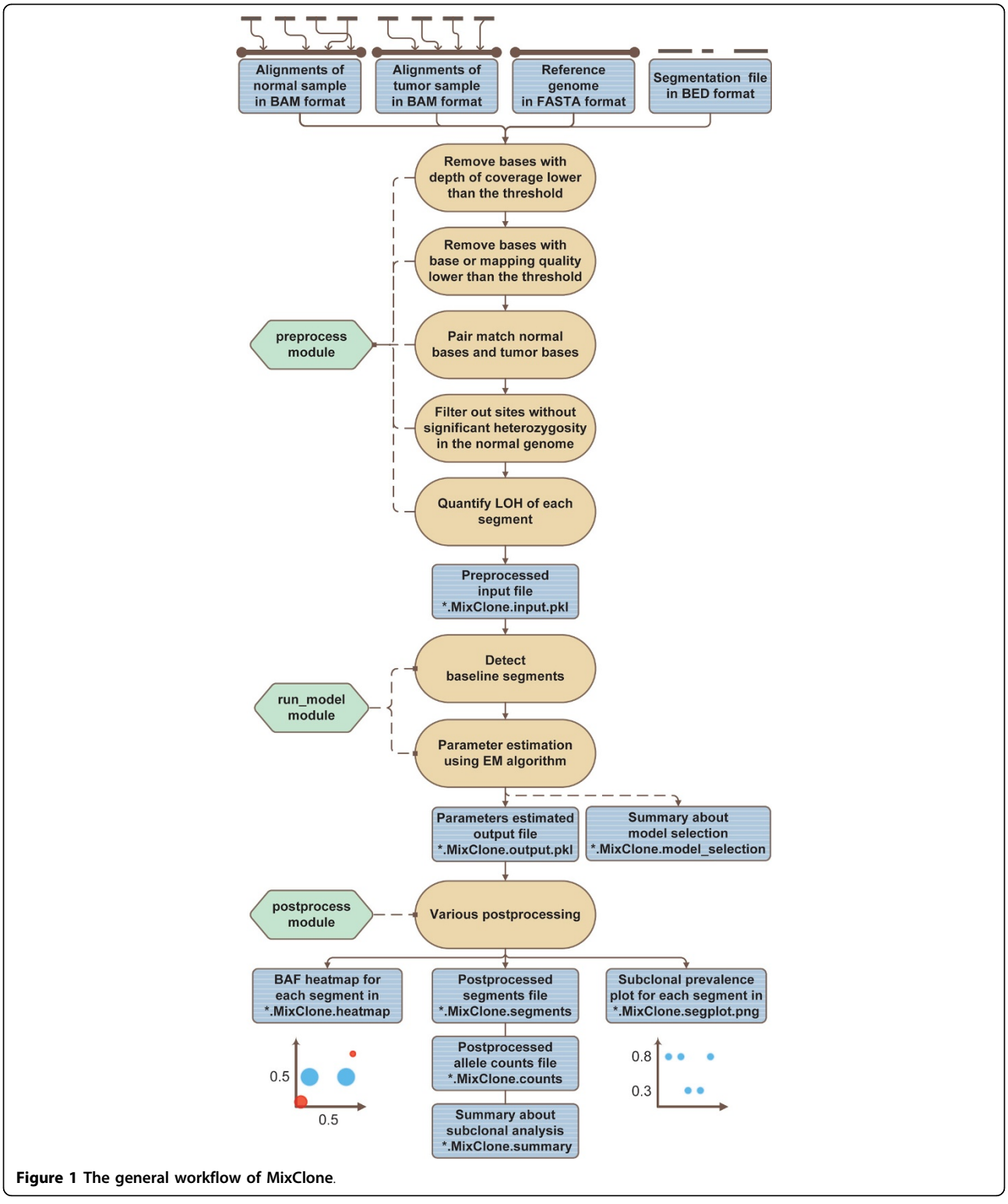

Figure 1 The general workflow of MixClone.

mutation sites and copy numbers as the input for PyClone and THetA. Details on how reads were simulated and preprocessed are given in Supplementary, Additional file 1.

MixClone is able to identify the correct subclonal populations for all the simulated datasets based on ground truth segmentations. Figure 2 shows the result of simulated dataset with two subclonal populations.

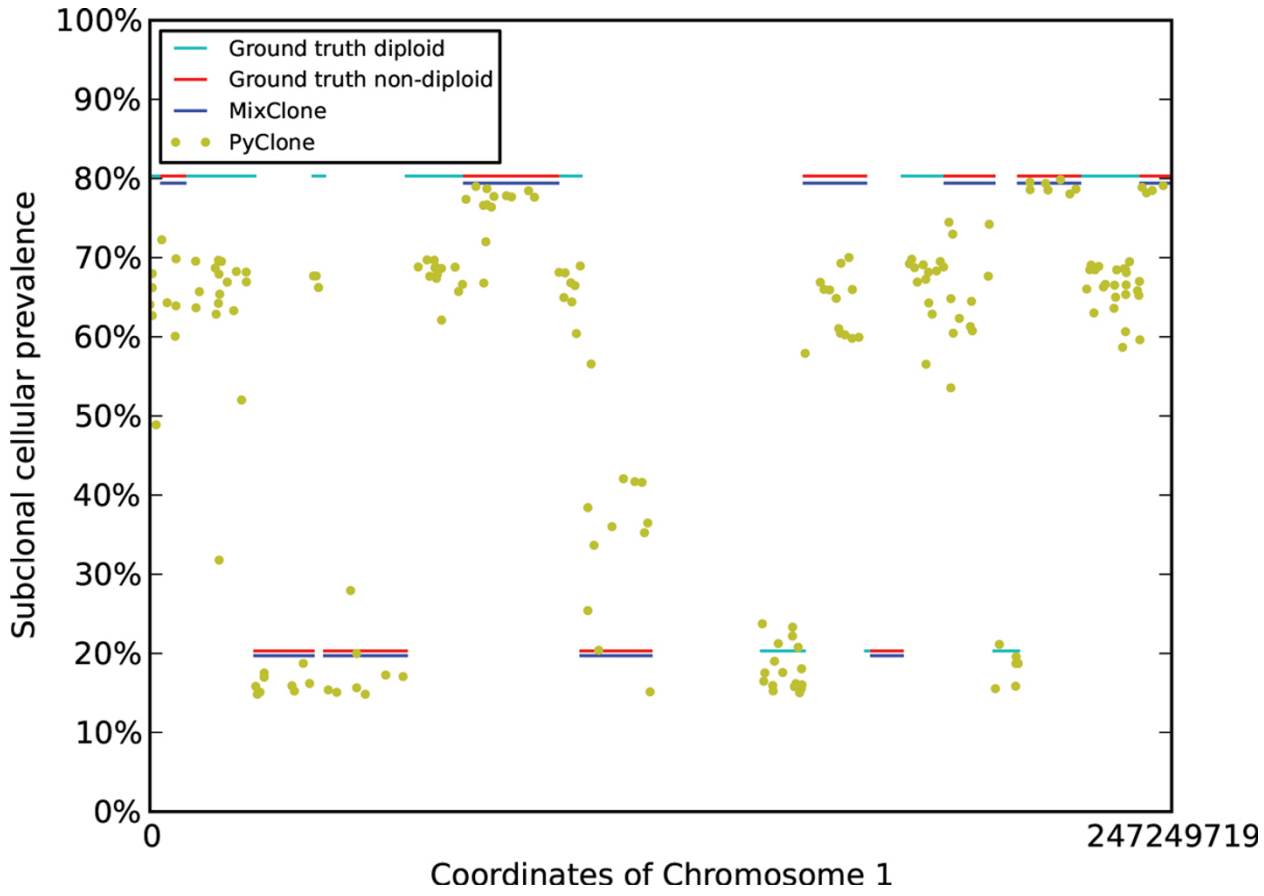

Figure 2 Subclonal inference results by MixClone and PyClone on a simulated dataset with two subclonal populations. The x-axis are the coordinates of Chromosome 1, and the y-axis are subclonal cellular prevalences. The blue horizontal bars represent the subclonal cellular prevalences estimated by MixClone based on non-diploid segments. Cyan and red horizontal bars represent the ground truth subclonal cellular prevalences of diploid and non-diploid segments. Yellow dots represent the subclonal cellular prevalences estimated by PyClone based on somatic point mutations.

MixClone also correctly estimates the subclonal cellular prevalences of all the segments with SCNAs except for one small segment in tumor genome case 4 with three subclonal populations. For results based on BIC-seq segmentations, MixClone still correctly estimates the subclonal cellular prevalences of the majority of the segments with SCNAs, except for those with copy-neutral loss of heterozygosity. This is likely due to the incorrect segmentations of BIC-seq, as BIC-seq relies on copy number changes and is unable to detect segments with copy-neutral loss of heterozygosity when they are adjacent to diploid segments. The complete results of all the simulated datasets based on both ground truth and BIC-seq segmentations are shown online through the github website associated with MixClone. As a comparison, we also run PyClone and THetA on the same datasets. We were unable to obtain THetA results after running it for more than 72 hours, likely due to its exponential scalability with the number of segments. In Figure 2, PyClone detects one of the two subclonal populations, whose ground truth cellular prevalence is 20%, but misestimates the other subclonal population, whose ground truth cellular prevalence is 80%, except for a few segments. The performance of MixClone on the other simulated datasets also significantly outperforms PyClone. One possible reason might be that the reads coverage of simulated datasets is not deep enough to support PyClone's non-parametric method [5], thus PyClone tends to report more subclonal populations due to the statistical variance.

Results from breast cancer sequencing data

We also applied MixClone on a whole-genome breast cancer sequencing dataset [12]. The details on data preprocessing are described in Supplementary, Additional file 1.

Figure 3a shows the subclonal inference results of sample MB-116. One estimated subclonal cellular prevalence

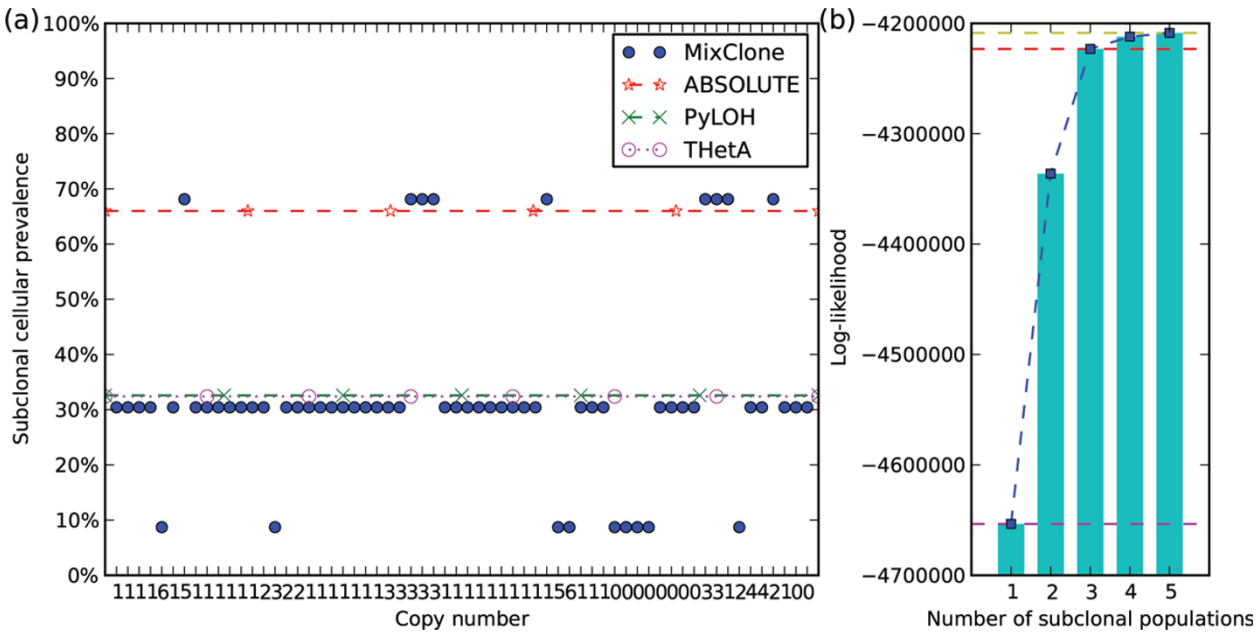

Figure 3 Subclonal inference results of sample MB-116. (a) The subclonal cellular prevalences estimated by MixClone, the tumor purities estimated by PyLOH, THetA [11], and the tumor purities estimated by ABSOLUTE [20] reported in [12] of sample MB-116. Each blue dot represents a segment. The x-axis is the estimated absolute copy number of the segment, and the y-axis is the estimated subclonal cellular prevalence of the segment. (b) The five log-likelihoods of MB-116 under different number of subclonal populations.

32% is consistent with the tumor purities estimated by PyLOH and THetA [11], and another estimated cellular prevalence 66% is consistent with the tumor purity estimated by ABSOLUTE [20] reported in [12].

Figure 3b shows the five log-likelihoods of MB-116 under different numbers of sub-clonal populations. The magenta, red and yellow curves represent the log-likelihoods corresponding to number 1, 3, and 5, respectively. Because the distance between the magenta and red curves (the cumulative log-likelihood increase from 1 to 3) is greater than 0.9 of the distance between the magenta and yellow curves (the total log-likelihood increase from 1 to 5), MixClone selected $K = 3$ as the number of subclonal populations for MB-116.

For samples without significant subclonal events, MixClone selected one as the number of subclonal populations, e.g. MB-106 (Figure 4). In Figure 4b, the ratio of total log-likelihood increase from 1 to 5 is 1.4×10^{-4}, which is less than the threshold of 0.01. Therefore, MixClone selected $K = 1$ as the number of subclonal populations for MB-106. The estimated cellular prevalence of this single population is 83%, which is also consistent with the tumor purities estimated by PyLOH, ABSOLUTE and one result of THetA [11] (Figure 4a).

Besides MB-116, MixClone also detected significant subclonal events in MB-45 and MB-123. Results of MB-45 and MB-123 are given in Supplementary, Additional file 1.

Discussion

In this article, we demonstrated MixClone's utility using whole genome sequencing data. However, most of the existing cancer genome sequencing data are from exome sequencing. An important future direction is to extend the current methodology to handle the exome sequencing data. Yet, extending MixClone to whole exome sequencing data is not trivial, as reads coverage on targeted exonic regions are no longer randomly distributed due to probe's variable efficiency [21]. Instead of Poisson distribution, using Gaussian distribution to model reads depth ratios between tumor and normal samples might be more appropriate to account for such additional variances, which has been demonstrated in whole exome sequencing based copy number analysis [21].

Another important future direction to extend MixClone is to implement joint analysis based on multiple samples, which is supported by PyClone and PhyloSub [5,7]. Multiple samples have been obtained for a single heterogeneous tumor tissue both temporally and spatially, and joint analysis based on these samples may reveal additional patterns of the history of tumor progression [5].

Currently, MixClone runs the subclonal analysis five times with different number of subclonal populations in range of 1 to 5 by default. In reality, larger numbers of subclonal populations may coexist within one tumor sample, but in this case some of the populations are

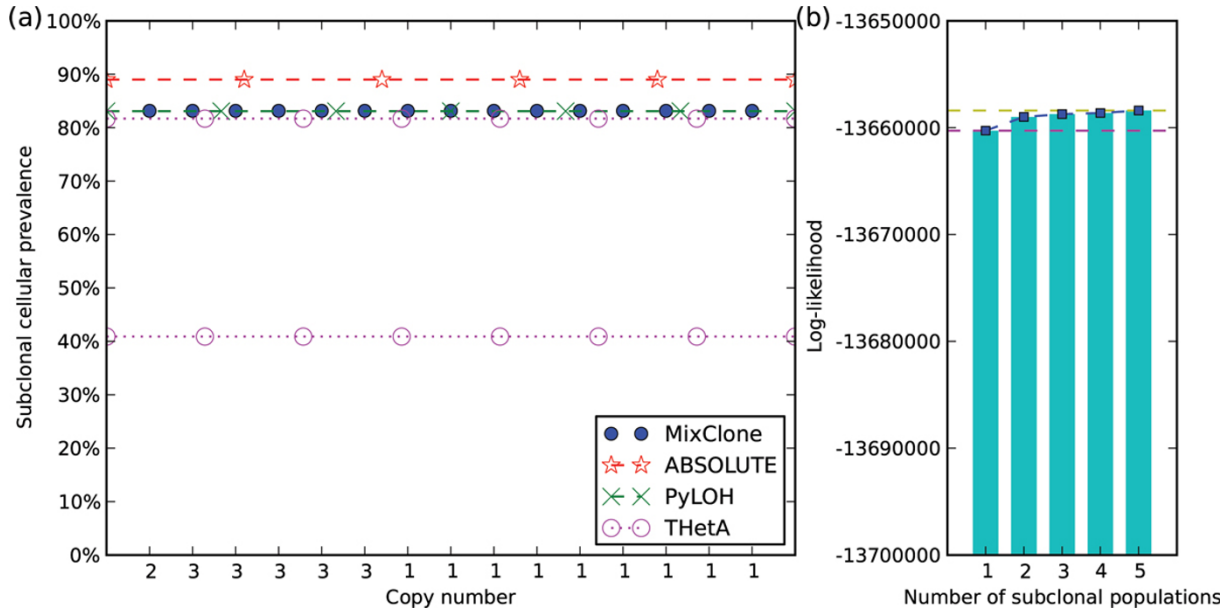

Figure 4 Subclonal inference results of sample MB-106. (a) The subclonal cellular prevalences estimated by MixClone, the tumor purities estimated by PyLOH, THetA [11], and the tumor purities estimated by ABSOLUTE [20] reported in [12] of sample MB-106. Each blue dot represents a segment. The x-axis is the estimated absolute copy number of the segment, and the y-axis is the estimated subclonal cellular prevalence of the segment. (b) The five log-likelihoods of MB-106 under different number of subclonal populations.

very likely to share similar cellular prevalences. Since Mix-Clone defines different subclonal populations based on distinct cellular prevalences, those populations with similar cellular prevalences may not be differentiated by MixClone. To achieve finer resolution of subclonal populations, subclonal lineages information would be necessary to further differentiate each population in addition to cellular prevalences. And phylogenetic methods may be possible solutions to explicitly incorporate subclonal lineages information [7].

Conclusions

In summary, we have developed a new method for inferring tumor subclonal populations by integrating sequence information gathered from SCNAs and heterozygous SNP sites. We showed that our method outperforms existing ones on simulation data, and applying it to a real breast cancer dataset is able to reveal new subclonal events not discovered before. Compared with existing methods, our method requires no additional deep sequencing of somatic point mutation sites.

Additional material

Additional file 1: Complete details of (1) detecting heterozygous SNP sites, (2) curating the baseline segments, (3) the EM updates of MixClone, (4) reads simulation for simulated data and (5) reads preprocessing for both simulated data and breast cancer sequencing data.

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

Authors' contributions
Designed the experiments: YL and XX; Performed the experiments: YL; Wrote the paper: YL and XX; All authors contributed to the analysis, and approved the paper.

Acknowledgements
The work was partly supported by National Institute of Health grant R01HG006870. The authors would also like to acknowledge dbGaP repository for providing the cancer sequencing datasets. The accession numbers for the breast cancer and prostate cancer datasets are phs000369.v1.p1 and phs000447.v1.p1, respectively.
This article has been published as part of *BMC Genomics* Volume 16 Supplement 2, 2015: Selected articles from the Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S2

Authors' details
[1]Department of Computer Science, University of California, Irvine, CA 92697 US. [2]Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697 US. [3]Center for Machine Learning and Intelligent Systems, University of California, Irvine, CA 92697 US.

Published: 21 January 2015

References

1. Nowell PC: **The clonal evolution of tumor cell populations.** *Science* 1976, **194(4260)**:23-28.
2. Garraway LA, Lander ES: **Lessons from the cancer genome.** *Cell* 2013, **153(1)**:17-37.
3. 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)**:714-726.

4. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, Bhan M, Calvo F, Eerola I, Gerhard DS, *et al*: **International network of cancer genome projects.** *Nature* 2010, **464(7291)**:993-998.
5. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, Shah SP: **Pyclone: statistical inference of clonal population structure in cancer.** *Nature methods* 2014, **11(4)**:396-398.
6. Andor N, Harness JV, Müller S, Mewes HW, Petritsch C: **Expands: expanding ploidy and allele frequency on nested subpopulations.** *Bioinformatics* 2014, **30(1)**:50-60.
7. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q: **Inferring clonal evolution of tumors from single nucleotide somatic mutations.** *BMC Bioinformatics* 2014, **15(1)**:35.
8. Hajirasouliha I, Mahmoody A, Raphael BJ: **A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data.** *Bioinformatics* 2014, **30(12)**:78-86.
9. Oesper L, Mahmoody A, Raphael BJ: **Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data.** *Genome biology* 2013, **14(7)**:80-80.
10. Roberts ND, Kortschak RD, Parker WT, Schreiber AW, Branford S, Scott HS, Glonek G, Adelson DL: **A comparative analysis of algorithms for somatic snv detection in cancer.** *Bioinformatics* 2013, **29(18)**:2223-2230.
11. Li Y, Xie X: **Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity.** *Bioinformatics* 2014, **174.**
12. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, *et al*: **Sequence analysis of mutations and translocations across breast cancer subtypes.** *Nature* 2012, **486(7403)**:405-409.
13. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, *et al*: **Jointsnvmix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data.** *Bioinformatics* 2012, **28(7)**:907-913.
14. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2(3)**:231-239.
15. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. ii. error probabilities.** *Genome research* 1998, **8(3)**:186-194.
16. Dempster AP, Laird NM, Rubin DB, *et al*: **Maximum likelihood from incomplete data via the em algorithm.** *Journal of the Royal statistical Society* 1977, **39(1)**:1-38.
17. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbsnp: the ncbi database of genetic variation.** *Nucleic acids research* 2001, **29(1)**:308-311.
18. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, *et al*: **The genomic complexity of primary human prostate cancer.** *Nature* 2011, **470(7333)**:214-220.
19. Xi R, Hadjipanayis AG, Luquette LJ, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, *et al*: **Copy number variation detection in whole-genome sequencing data using the bayesian information criterion.** *Proceedings of the National Academy of Sciences* 2011, **108(46)**:1128-1136.
20. 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.** *Nature biotechnology* 2012, **30(5)**:413-421.
21. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF: **Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv.** *Bioinformatics* 2011, **27(19)**:2648-2654.

doi:10.1186/1471-2164-16-S2-S1
Cite this article as: Li and Xie: MixClone: a mixture model for inferring tumor subclonal populations. *BMC Genomics* 2015 **16**(Suppl 2):S1.