Wang *et al. BMC Medical Genomics* 2019, **12**(Suppl 1):27
https://doi.org/10.1186/s12920-018-0457-4

BMC Medical Genomics

RESEARCH

Open Access

A graph-based algorithm for estimating clonal haplotypes of tumor sample from sequencing data

Yixuan Wang[1,2], Xuanping Zhang[1,2*], Shuai Ding[3], Yu Geng[1,2], Jianye Liu[1,2], Zhongmeng Zhao[1,2], Rong Zhang[1,2], Xiao Xiao[2,4] and Jiayin Wang[1,2*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2018
Los Angeles, CA, USA. 10-12 June 2018

Abstract

Background: Haplotype phasing is an important step in many bioinformatics workflows. In cancer genomics, it is suggested that reconstructing the clonal haplotypes of a tumor sample could facilitate a comprehensive understanding of its clonal architecture and further provide valuable reference in clinical diagnosis and treatment. However, the sequencing data is an admixture of reads sampled from different clonal haplotypes, which complicates the computational problem by exponentially increasing the solution-space and leads the existing algorithms to an unacceptable time-/space- complexity. In addition, the evolutionary process among clonal haplotypes further weakens those algorithms by bringing indistinguishable candidate solutions.

Results: To improve the algorithmic performance of phasing clonal haplotypes, in this article, we propose *MixSubHap*, which is a graph-based computational pipeline working on cancer sequencing data. To reduce the computation complexity, *MixSubHap* adopts three bounding strategies to limit the solution space and filter out false positive candidates. It first estimates the global clonal structure by clustering the variant allelic frequencies on sampled point mutations. This offers a priori on the number of clonal haplotypes when copy-number variations are not considered. Then, it utilizes a greedy extension algorithm to approximately find the longest linkage of the locally assembled contigs. Finally, it incorporates a read-depth stripping algorithm to filter out false linkages according to the posterior estimation of tumor purity and the estimated percentage of each sub-clone in the sample. A series of experiments are conducted to verify the performance of the proposed pipeline.

Conclusions: The results demonstrate that *MixSubHap* is able to identify about 90% on average of the preset clonal haplotypes under different simulation configurations. Especially, *MixSubHap* is robust when decreasing the mutation rates, in which cases the longest assembled contig could reach to 10kbps, while the accuracy of assigning a mutation to its haplotype still keeps more than 60% on average. *MixSubHap* is considered as a practical algorithm to reconstruct clonal haplotypes from cancer sequencing data. The source codes have been uploaded and maintained at https://github.com/YixuanWang1120/MixSubHap for academic use only.

Keywords: Cancer genomics, Haplotype phasing, Clonal haplotype, Computational pipeline, Sequencing data analysis

*Correspondence: zxp@mail.xjtu.edu.cn; wangjiayin@mail.xjtu.edu.cn
[1] Department of Computer Science and Technology, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710048, China
[2] Shaanxi Engineering Research Center of Medical and Health Big Data, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710048, China
Full list of author information is available at the end of the article

Background

Modern canceration theory summarizes that tumor heterogeneity is one of the key results of tumor proliferation and evolution [1]. Any macroscopic tumor tissue is considered as an admixture of cancerous and non-cancerous cells, where the cancerous cells, in many cases, could be further clustered into multiple sub-clones, according to their somatic mutational events [2, 3]. These somatic mutations, interacting with germline variations, often underlie different deleterious selective advantages, which may further contribute to drug resistance, tumor recurrence and metastasis, and many other phenotypes [4–6]. For example, it is observed that the presence of multiple sub-clones could be associated with poor clinical outcomes in a group of chronic lymphocytic leukemia cases [7]. It is also reported that the clonal competition for predominance occurs spontaneously in multiple myeloma cases and the heterogeneous clonal mixtures may shift predominant clones with therapeutic selection [8]. Not only in blood cancer cases, similar conclusions could be drawn in many other cancer types, such as esophageal adenocarcinoma [9], lung adenocarcinoma [10] and renal clear cell carcinoma [11]. It is now a popular opinion that a comprehensive understanding on tumor heterogeneity benefits clinical diagnosis and potential precision treatments.

Genomic analysis on tumor heterogeneity has two levels: genotype level and haplotype level. The genotype-level bioinformatics pipelines differentiate homozygous mutational events, including loss of heterozygosity (LOH), from heterozygous ones [12–14], and then cluster them into sub-clones [3, 15]. On this basis, the haplotype level analysis requires locating each heterozygous mutation on the corresponding chromosomal sequence of alleles, named haplotype, whose computational problem is often called *haplotype phasing*. Actually, haplotype phasing has already been an important step in many bioinformatics workflows besides cancer research [16], but its importance in cancer genomics is recently emphasized [17]. Understanding haplotype heterogeneity is suggested not only to elucidate a series of critical genome-to-transcriptome events, e.g. gene fusion transcripts and their driver partners [18], but to facilitate the studies on the interactions among different germline and somatic variations, e.g. two-hit events and allelic amplifications [4, 5, 17]. Such results could significantly benefit downstream analyses and studies in many fields, including disease association studies [6, 19], clinical decision-support with electronic medical record data [20–22], drug and treatment designs and improvements [23, 24], etc.

Benefiting from the second generation sequencing technology, tens of thousands of cancer patients have been sequenced, and the cancer sequencing data have been accumulating rapidly as well, which greatly promotes the studies on clonal heterogeneity and expansion [25] and the developments of related computational approaches [26]. Nowadays, tumor heterogeneity analyses are almost built up on cancer sequencing data. Although the existing approaches differ in models, algorithms, and the evaluation standards for "well suited", based on our best knowledge, the core algorithms may be roughly divided into two categories: the phylogenetic model-based methods and model-free ones. The phylogenetic model-based methods, just as the name implies, usually focus on the computational problem of inferring phylogenetic trees that describe clonal expansion and evolution [27–32]: *TrAp* proposed an expanded algorithm from a brute-force algorithm for sub-clonal deconvolution, which generated the evolutionary tree(s) by comprising the maximum number of first-generation trees [27]. *PhyloWGS* established a probabilistic model for phylogeny inference, which incorporated the information of variant allele frequencies and the estimations of allelic amplifications and LOHs [28]. *BitPhylogeny* designed a graphical model, and then it adopted two strategies, which were a Markov chain Monte Carlo (MCMC) sampling and a maximum posterior method on expected adjusted rand, to solve the possible phylogenies [29]. *SPRUCE* utilized a bounded enumeration strategy to search the solution space of candidate perfect phylogenies which were consistent with the given data set [30]. *Canopy* improved the statistical framework and was capable of handling the data sequenced from temporally and/or spatially separated samples from the same patient to reconstruct tumor phylogeny [31]. A recent published method further addressed the lack of methods for tumor deconvolution and phylogenetics of diverse classes of structural variations at base-pair resolution [32].

On the other hand, it is argued that the specific features of tumor evolution may challenge the direct applications of classical phylogenetic models [33]. One of the key issues occurs when classical phylogenetic approaches require a priori on the number of sub-clones, which is an unknown parameter for cancer sequencing data. To overcome such issues, the model-free methods often focus on the clonal structures with the maximum likelihood on global variant allelic frequencies [3, 34–38]: *THetA* designed a convex optimization algorithm to solve the maximum likelihood mixture decomposition, which optimized the multinomial probability [34]. *PhyloSub* proposed a series of topological constraint rules to limit the possible phylogenies that were able to explain the frequency changes [35]. *PyClone* introduced a Bayesian clustering method, which integrated the estimations on cellular prevalences, normal-cell contamination and segmental copy-number changes [36]. *SciClone* adopted a variational Bayesian mixture model to provide a global estimation of clonal architecture across all of the given copy-number neutral regions [3]. *TITAN* established a graphical model to estimate sub-populations based on copy number alterations and loss of heterozygosity events [37]. Automate learning was also incorporated for deconvolution of genomic mixtures, where the RNA expression data was involved in addition to improve the performance [38]. In general, there is no clear boundary between the two categories, and several comprehensive reviews compared the advantages among the existing approaches [26, 33].

However, most of the existing approaches are not able to deepen the analyses to haplotype level efficiently. When multiple haplotypes are considered, the evolutionary process should be represented by a set of parallel phylogenetic trees rather than possible single phylogenies, which is different from the hypothesis on which most of the existing methods, the phylogenetic model-based methods or model-free ones, rely [3, 28–30, 32, 34, 36–38]. For those methods considering concurrent evolutionary processes, haplotype phasing algorithms are needed to locate heterozygous mutations prior to inferring clonal structure [27, 31, 35]. Moreover, phasing multiple haplotypes is also a quite challenging computational problem because the solution space of possible haplotypes is exponentially increased along with the increasing of sub-clones. For example, for k sub-clones each with n heterozygous variation sites, the solution space of $2k$ clonal haplotypes (allelic imbalance events are not considered) reaches $O\left(2^{(2k-1)n}\right)$ [39]. The polyploid phasing problem has already been suggested as an NP-hard problem [39–44], hence probabilistic algorithms and heuristic strategies are commonly used to approximate optimization solutions, such as Gibbs sampling [39], greedy binning algorithm [41], branch-and-bound scheme by maximum likelihoods [42], semi-definite programming [43], sparse tensor decomposition [44], etc.

Different from polyploid haplotypes, the haplotypes of multiple sub-clones from the same sample always imply the information of its clonal structure. Thus, to enhance the efficiency of the existing computational pipeline, e.g. [17], we consider to incorporate the priori of clonal structure to bound the solution space, and then polish the clonal structure with estimated clonal haplotypes. To achieve this, in this article, we propose *MixSubHap*, a computational pipeline for phasing clonal haplotypes as well as estimating clonal structure. To reduce the computation complexity, *MixSubHap* adopts three bounding strategies to limit the solution space and filter out false positive candidates. It first estimates the global clonal structure by clustering the variant allelic frequencies on sampled point mutations. This offers a priori on the number of clonal haplotypes when copy-number variations are not considered. Then, it utilizes a greedy extension algorithm to approximately find the longest linkage of the locally assembled contigs. Finally, it incorporates a read-depth stripping algorithm to filter out false linkages according to the posterior estimation of tumor purity and the estimated percentage of each sub-clone in the sample.

Methods

Suppose that we are given a set of paired-end sequencing data with mapping information, and the outputs of the proposed pipeline include both the number of sub-clones and the haplotypes of each sub-clone. The given data is first pre-processed: a read is retained if it brings at least one point mutations, while a read-pair is retained if it brings at least two point mutations. Each read-pair that passed the filter is then collapsed to a much shorter sequence, named VPE as in [45], by extracting the sites with point mutations from this read-pair, as shown in Fig. 1. A VPE consists of

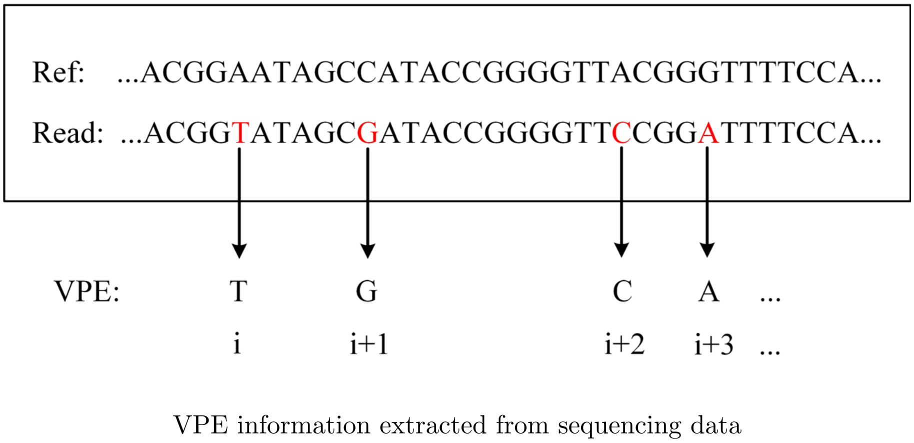

Ref: ...ACGGAATAGCCATACCGGGGTTACGGGTTTTCCA...
Read: ...ACGGTATAGCGATACCGGGGTTCCGGATTTTCCA...
VPE: T G C A ...
i i+1 i+2 i+3 ...
VPE information extracted from sequencing data

Fig. 1 Extracting a VPE from a pair of mapped reads. When a read-pair is mapped to the reference genome, the corresponding VPE consists of the sites with variants

only the sites with variants from the corresponding read-pair. In the current stage of this research, the structural variations, including the allelic imbalance events on point mutations, are temporarily not considered. According to these reads (VPEs), the variant allelic frequency (VAF) of each variant is calculated. We adopt a model-free method to provide a priori of the clonal structure according to global variant allelic frequencies. A series of model-free method could achieve this. Here, we incorporate *SciClone* [3], a popular method, into the proposed pipeline.

MixSubHap consists of three major components: assembling local VPEs, expanding local contigs and iteratively stripping clonal read-depth. The flowchart is shown in Fig. 2.

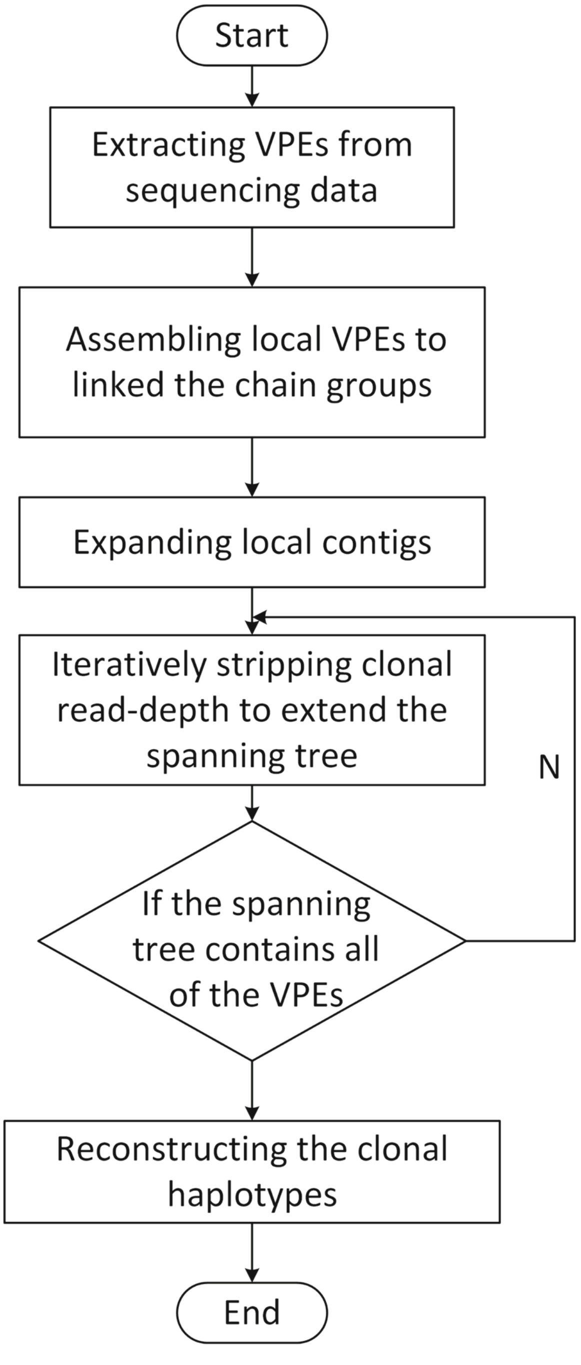
- Assembling local VPEs: As the first component, a divide-and-conquer strategy is adopted to assemble the VPEs to multiple groups of contigs. Different from assembling reads, VPEs are first clustered by starting sites, and then for each cluster, the VPEs with the same starting site are processed together to form a small set of contigs. Note that, multiple reads whose mapping positions are slightly different may collapse to the same VPE sequence if they bring the same variants. In addition, several constraints which imply the inheritance principle across sub-clones are applied. The details of these constraints are further discussed in Discussion section.
- Expanding local contigs: To link the contigs across different clusters, an efficient greedy algorithm is conducted. An undirected weighted graph is established. Each vertex represents a variant site. An edge exists between two vertexes if at least one VPE supports the linkage of them. The edges are weighted according to the likelihoods of possible linkage modes. The greedy algorithm first generates a maximum spanning tree on the sub-graph that consists of the vertexes from the founding clone. Here, the estimation of founding clone is provided by *SciClone*. Moreover, the computation is also simplified by the inheritance constraint, which limits the linkage mode in all of the descendant sub-clones.
- Stripping each clonal read-depth: For the variants do not occur in the founding clone, different genotypes are brought by different sub-clones. In these cases, the likelihoods are interfered by allelic clonal haplotypes. To filter out the bias on the likelihoods, a thickness stripping algorithm is designed. According to the VAFs and the estimations on clonal structure, the VPEs that have higher probabilities of sampling from other sub-clones are removed, and then the spanning tree is extended based on the corrected edge weights. The parental haplotypes guide the clonal haplotypes of its descendants. When the tree is traversed, the clonal haplotypes are then reconstructed.

Assembling local VPEs

Let **V** denote the set of given variant sites. VPEs are defined as the information which contain the base states of all the mutated sites along with their actual positions in reads and relative positions in set **V**. When extracting VPE information, it is necessary to ensure that the number of mutations brought by the read or read-pair is not less than two. Sequencing reads that contain only one mutation do not reflect the interrelationship between mutations and do not provide useful information for haplotype reconstruction, thus we do not take such information into consideration. It is assumed that each sequence library is well prepared and the insert-size obeys a normal distribution with a small variance. Therefore, libraries with different lengths are introduced, more variant sites can be extracted correctly.

In general, the rare somatic mutations on cancer susceptibility genes may reach 10,000, which is a huge challenge for the computational capability of ordinary computer. At this time, the computational complexity is too high due to the large amount of variants. Therefore, the *MixSubHap* uses the Dividing and Assembling strategy to process the VPE information. The strategy appropriately clusters the length of the division and designs the connection ways, so that the resulting short chains can accurately exhibit fragments of the sub-clonal haplotypes. To achieve this, the VPE is first mapped to the reference by mapping the base state of corresponding variant against the reference according to its actual position. After that, the aligned VPE are divided into M groups according to its starting position, then the algorithm processes group by group. Assume that the number of sub-clones I is known, then the number of haplotypes is determined, which is $2I$, and the total number of variants is N. The process consists of the following steps:

- Sorting the VPEs via the starting positions.
- Constructing initial short chain groups. VPE alignments at the same starting position are processed together to form a group of short chains. Specifically, this is an integer programming problem, where the goal of the programming is that the cardinality of a set of short chains (the number of short chains in the group) is minimal, and the constraint of the programming is that the short-chain group must be able to support all the corresponding VPEs, carrying the maximum number of variant sites. During solving this, the greedy strategy is used to minimize the number of short chains, and at the same time to support all the VPE information of this group. In order to ensure that the ambiguity chain


Start
Extracting VPEs from sequencing data
Assembling local VPEs to linked the chain groups
Expanding local contigs
Iteratively stripping clonal read-depth to extend the spanning tree
If the spanning tree contains all of the VPEs
N
Reconstructing the clonal haplotypes
End

The flowchart of *MixSubHap*

Fig. 2 The flowchart of the proposed pipeline

does not produce redundant, we add each VPE into a short chain group only when the VPE not contained in the existing short chains, otherwise, the VPE remains in the candidate set and the short chain group retains the same.

- According to the principle of inheritance, arrange all short chain groups and keep qualified arrangement. Process VPEs whose starting sites are p ($p \in N$), then process group by group, until all VPEs are processed.

Until now, all the VPEs are rationally and effectively connected, and a large number of short chain groups carrying haplotype information are formed. Next, we will present a reasonable and efficient strategy to further construct the haplotypes of each sub-clone.

Expanding local contigs

According to the hypothesis of linear evolution mode between tumor sub-clones (see in Discussion), we know that, once the connection mode of a pair of variants from the founding clone S_0 is fixed in the tumor evolutionary process, the differentiation of subsequent sub-clone S_i, $i \in \{1, 2, \ldots, I-1\}$ will inherit the same connection without any change. In the same way, once the connection between a pair of variants from sub-clone S_i at any level is confirmed, it will not change in its descendant clone $S_j\,(i < j)$. According to the linear evolution mode, the mutation sites in the clones are separated layer by layer. Thus, the parental clonal haplotype structure can be used as a known condition to guide the construction of the descendant clonal haplotypes.

In order to recognize the clonal haplotypes efficiently and accurately, *MixSubHap* algorithm first clusters all the variant sites according to the VAF of each site. The variants in the same cluster are from the same sub-clone [3], and then the clustering results are considered as the basis to initialize the clonal haplotypes. The clustering method used in this paper is *SciClone* version 1.1.0 [3], which is reported to be relatively accurate in clustering the somatic mutations by clonal structure.

MixSubHap algorithm mainly generates a maximum spanning tree based on the short chain groups we obtained. Let M be the set of all variants, and $\langle p_i, p_j \rangle$ represent two adjacent allelic sites, $p_i, p_j \in M$. Let $H_{S_k}^{p_i,p_j}$ represent the connection mode between the p_ith variant site and the p_jth variant site from the sub-clone S_k. A stands for the same base as the reference genome, while B stands for a base different from the reference genome, namely B stands for a mutation. Therefore, for any two sites from the same sub-clone, possible values for $H_{S_k}^{p_i,p_j}$ are $\{(A,A),\ (B,B)\}$ *or* $\{(A,B),\ (B,A)\}$. For the founding clone S_0, according to the short chain groups, we can get many different values of $H_{S_0}^{p_i,p_j}$ and add these different values respectively to the corresponding coverage of variant pair $\langle p_i, p_j \rangle$. In order to separate the sub-clones layer by layer, variant pair $\langle p_i, p_j \rangle$ from every sub-clone and the corresponding coverage level $c_{\langle i,j \rangle}$ are used to estimate the clonal haplotypes. For variant pairs, the coverage of each pair and the probability of various possible connection modes between the two variants are calculated together. According to the probabilities of various connection patterns, the corresponding undirected weighted graphs are established. When calculating the coverage level $c_{\langle i,j \rangle}$ of allelic sites $\langle p_i, p_j \rangle$, only short chain groups are considered. Let $F(i,j)$ represent the set of short chains containing both allelic sites $\langle p_i, p_j \rangle$. $|F(i,j)|$ represents the number of chains in this set. We set the connection probability of $\{(A,A),\ (B,B)\}$ to be positive and set the connection probability of $\{(A,B),\ (B,A)\}$ to be negative. Let G represent a weighted undirected graph where the variant sites are the vertexes of the graph. Here, we set the coverage threshold Cov, with a default value of 2. If $|F(i,j)| \geq Cov$, there will be an edge between variants $\langle p_i, p_j \rangle$ and the weight of edge will be the probability of the connection. The formula of weight is:

$$W_{\langle i,j \rangle} = f\left(b_{p_i}, b_{p_j}, A\right) - f\left(b_{p_i}, b_{p_j}, B\right)$$

$$N^{b_{p_i},b_{p_j}}(p_i, p_j) = \sum_{r \in F(i,j)} I\left(r(p_i, p_j) = \left(b_{p_i}, b_{p_j}\right)\right)$$

Where, $N^{b_{p_i},b_{p_j}}(p_i, p_j)$ represents the number of chains across allelic sites $\langle p_i, p_j \rangle$ corresponding to the connection mode (b_{p_i}, b_{p_j}). b_{p_i} and b_{p_j} respectively indicate the base states of the variant site p_i and p_j. $r(p_i, p_j)$ represents alleles for variant site p_i and p_j. There are four kinds of joint states for allele p_i and p_j, where $(b_{p_i}, b_{p_j}) \in \{(A,B), (B,A), (A,A), (B,B)\}$. $I(.)$ is the indicator function. When $r(p_i, p_j) = (b_{p_i}, b_{p_j})$ is true, the function value is 1, 0 otherwise. For variant site p_i, consider the sequencing error and alignment error of each site ε.

Let

$$N^A(p_i, p_j) = N^{A,A}(p_i, p_j) + N^{B,B}(p_i, p_j)$$

$$N^B(p_i, p_j) = N^{A,B}(p_i, p_j) + N^{B,A}(p_i, p_j)$$

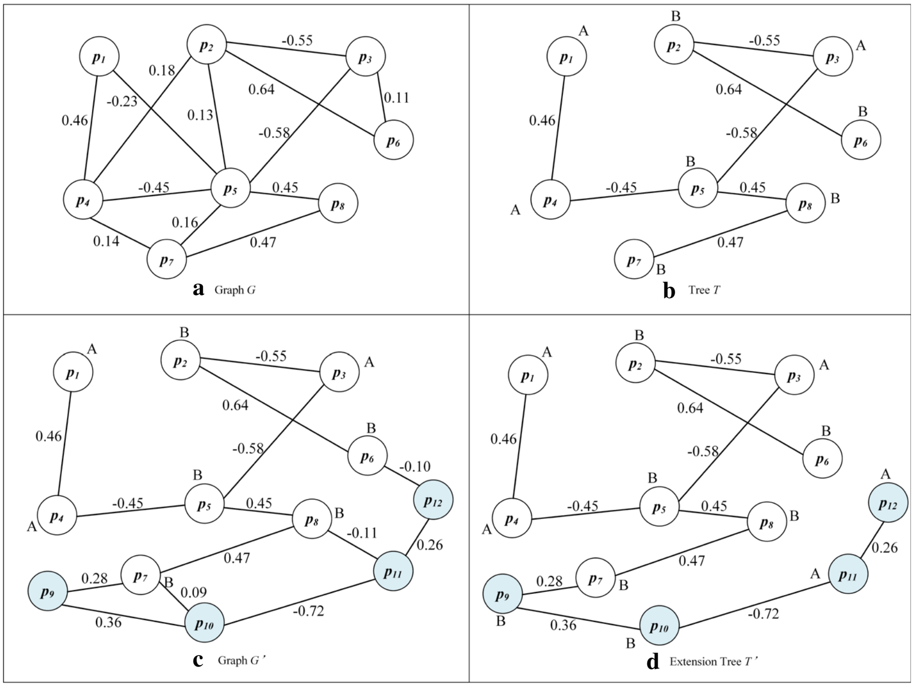
Two connection probabilities of the paired variant sites $\langle p_i, p_j \rangle$ are,

$$f(b_{p_i}, b_{p_j}, A) = \frac{\left((1-\epsilon)^2 + \epsilon^2\right) \times N^A(p_i, p_j) + 2\epsilon \times (1-\epsilon) \times N^B(p_i, p_j)}{|F(i,j)|}$$

$$f(b_{p_i}, b_{p_j}, B) = \frac{\left((1-\epsilon)^2 + \epsilon^2\right) \times N^B(p_i, p_j) + 2\epsilon \times (1-\epsilon) \times N^A(p_i, p_j)}{|F(i,j)|}$$

Where $W_{\langle i,j \rangle} > 0$ indicates that the connection mode of allelic sites $\langle p_i, p_j \rangle$ is $H_{S_0}^{A,A}/H_{S_0}^{B,B}$; otherwise, $W_{\langle i,j \rangle} < 0$ indicates that the connection mode of allelic sites $\langle p_i, p_j \rangle$ is $H_{S_0}^{A,B}/H_{S_0}^{B,A}$. The greater the absolute value of $W_{\langle i,j \rangle}$, the higher the reliability of the corresponding connection pattern.

After constructing the undirected weighted graph G of the founding clone is constructed, the number of vertexes in the graph is N_v. All the vertexes in the graph G are all derived from variant sites in the founding clone S_0. The algorithm selects a vertex, whose base state is known, as the starting point *sp* for constructing the sub-clonal haplotypes and generating the initial maximum spanning tree T corresponding to graph G. The processing steps are as follows, as shown in Fig. 3:



Extension of Tree T' Construction Process

Fig. 3 Extension of Tree T' Construction Process

- Find all variant sites connected to *sp* from graph G;
- Select the corresponding edge of $max\left(\left|W_{\langle i,j\rangle}\right|\right)$ as one edge of the maximum spanning tree T;
- According to the positive or negative of the edge weight, the state of another variant connected to *sp* can be determined and these variants will be considered as new known vertexes;
- Conduct the second step successively for other unkonwn variants and repeat adding border process to maximum spanning tree T till all the variants have appeared in T or cannot be added any more.

In actual, due to the sparseness of the founding clone variants and the limitation of the read length of the second generation sequencing data, the generated undirected weighted graph G is often not a connected graph, but consists of several mutually disconnected subgraphs. *MixSubHap* algorithm generates an equal number of subtrees on these subgraphs and identifies the base states, thus guides the later extension of the spanning tree.

Stripping each clonal read-depth

After the initial maximum spanning tree T is built, the connection patterns of partial variant sites on the haplotypes have been determined. However, there are a large quantity of variant sites not included in the initial maximum spanning tree T. The connection modes of these variant sites are relatively complex, including three types of linkages $(B,A),(A,A),(A,B)$ with varying proportion of each according to the sub-clone proportion in the tumor sample. So we adopt the thickness stripping strategy to strip the read depth level by level from the founding clone to the uppermost layer descendant sub-clone. The remaining sub-clones can be processed in accordance with the method of constructing tree in the founding clone.

Thickness stripping strategy refers to finding the separation point that identifies the current sub-clone and divides all the clones into two parts in the direction of alleles. If sub-clone S_i is the first sub-clone which two variants have been mutated, this sub-clone should be considered as the demarcation line and the upper parental clone of this sub-clone should be separated at a mixed ratio. With the process of evolution, the variant allelic frequency is decreasing, so the initial value of sub-clone proportion should be estimated according to the mean of allelic frequency of each sub-clone. After separation, there are only two types of connection ways between allelic sites for remaining sub-clones and the weights of the two ways are the same in positive and negative. So the connection between the allelic sites can be clearly judged. For the portion of sub-clones to be stripped, they can be separated from the mixed sequencing data according to the mixed ratio of the sub-clones. Set the number of clones to be N_s, and then the thickness stripping formula is:

$$P^A(p_i,p_j) = \left|F(p_i,p_j)\right| \times \left(\sum_{0\leq i\leq S_k} r_i + \sum_{S_{k+1}\leq i\leq S_{N_s}} \frac{1}{2} r_i \right)$$

$$P^B(p_i,p_j) = \left|F(p_i,p_j)\right| \times \left(\sum_{S_{k+1}\leq i\leq S_{N_s}} \frac{1}{2} r_i \right)$$

Adjust the coverage of the allelic site $\langle p_i,p_j\rangle$, we have

$$\hat{N}^A(p_i,p_j) = N^A(p_i,p_j) - P^A(p_i,p_j)$$

$$\hat{N}^B(p_i,p_j) = N^B(p_i,p_j) - P^B(p_i,p_j)$$

After separating the data, the new undirected weighted graph G' is established by using the same method for the undirected weighted graph G and the weight calculation formula becomes

$$W'\langle p_i,p_j\rangle = f'\left(b_{p_i},b_{p_j},A\right) - f'\left(b_{p_i},b_{p_j},B\right)$$

$$f'(b_{p_i},b_{p_j},A) = \frac{\left((1-\epsilon)^2+\epsilon^2\right)\times \widehat{N^A}(p_i,p_j) + 2\epsilon\times(1-\epsilon)\times \widehat{N^B}(p_i,p_j)}{\widehat{N^A}(p_i,p_j)+\widehat{N^B}(p_i,p_j)}$$

$$f'(b_{p_i},b_{p_j},B) = \frac{\left((1-\epsilon)^2+\epsilon^2\right)\times \widehat{N^B}(p_i,p_j) + 2\epsilon\times(1-\epsilon)\times \widehat{N^A}(p_i,p_j)}{\widehat{N^A}(p_i,p_j)+\widehat{N^B}(p_i,p_j)}$$

Where $\left|W'\langle p_i,p_j\rangle\right| > \delta$ represents an effective edge between $\langle p_i,p_j\rangle$ allelic sites. When adding it to the graph G', the corresponding weight is $W'\langle p_i,p_j\rangle$. The default value of δ is 0.1. When the graph G' is constructed, it can basically contain all the variation sites from the reference sequence. Then the maximum spanning tree T is extended to T' according to the graph G' following the steps in section Expanding local contigs.

In order to finally reconstruct the haplotype that contains the variants as many as possible, it is necessary to ensure that the extended tree T' contains more variants. Lower coverage and higher threshold of coverage will cause some variation sites be left out. Thus, we automatically adjust to lower coverage threshold and edge weight threshold to ensure lower false-negative as much as possible. We adopt depth-first traversal of all vertexes in extended tree T' and sort according to the relative positions in the order of reference sequence. After sorting, the state set of vertexes is a haplotype of the last sub-clone. Assuming that all variation sites are heterozygous, another haplotype of the last sub-clone is easily obtained. According to the linear evolution relationship among the sub-clones, the haplotypes of the remaining sub-clones are obtained by the following formula.

$$h_{2i,j} = \begin{cases} A, h_{2(i+1),j} = A \\ B, h_{2(i+1),j} = B \; and \; sub(j) \leq i \\ A, h_{2(i+1),j} = B \; and \; sub(j) > i \end{cases}$$

Where, *i* represents the label of sub-clone and *j* represents variation site number. *sub*(*j*) represents the sub-clone label of the *j*th variation site, $h_{2i,j}$ represents the base state of the *j*th site from father chain on the *i*th sub-clone. By the above formula, we can find base state of the site from one haplotype corresponding to the paired haplotype.

$$h_{2i,j} = \begin{cases} h_{2i,j} \oplus 1, sub(j) \leq i \\ h_{2i,j}, i < sub(j) \leq I-1 \end{cases}$$

The construction process is shown in Fig. 4.

Results

To generate simulation datasets, a chromosome is randomly selected from the human reference genome as a reference sequence. Simulating single point variation, germline mutation rate is set to 0.1%, and the somatic mutation rate is 1%. Consider the purity of the tumor sample: the founding clone S_0 and two descendant sub-clone S_1 and S_2, the ratio of them is set to $3:5:2$. Several parameters, such as coverage, the number of libraries, the length of read, have impact on the performance of *MixSubHap* algorithm, only one parameter value was changed for each experiment.

Varying read length

Three libraries were set with different insert-sizes of 1000 bp, 1500 bp, and 2000 bp, respectively. Three subclonal mixing ratio is $3:5:2$. A priori of sub-clone haplotype depends on the result of *SciClone*, whereas the accuracy of *SciClone* mainly depends on the library coverage and the sequence deviation of VAF. The library used for computing VAF is called base library. The coverage of the base library is set to be 100×, the coverage of other libraries is 50×, and the lengths of the paired-end reads are 100 bp , 150 bp , 200 bp , and 250 bp .

We have caculated the propotion of the clonal haplotypes we can recognize among all the variants, the accuracy rate of reconstruction, the longest length of fragment we can assemble, and the number of fragments assembled. The results are shown in Table 1. *MixSubHap* shows strong robustness, which can recognize over 90% clonal haplotypes, with the longest assembled fragments longer than 10 kbp under different read lengths. With the increase of read length, the recognition rate can be improved steadily.

Analysis on the influence of new library coverage

When selecting libraries and coverage, the base library and its coverage are generally determined firstly, from which the pair-end reads information covering at least two variants are extracted. Then we introduce new libraries until most of the variant sites are included. When the library is replaced, the library length is incremented by 500 bp by default. Influence of new library coverage on the accuracy of haplotype reconstruction and recognition

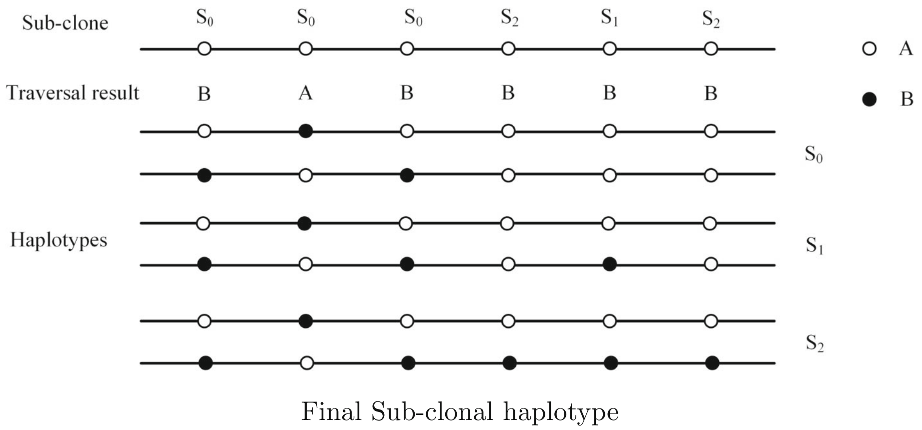


Fig. 4 Final Sub-clonal haplotype

rate in sub-clones are shown in Table 2. All the recognition rate is over 91%, and the longest assembled fragment is longer than 10077 bp. When the coverage of new library changes from 20× to 100×, the change of recognition rate in sub-clone recognition is inconspicuous (less than 1%), indicating that a minor effect is caused by new libraries.

Analysis on the influence of the number of the libraries

We found that the coverage of the new library almost had no effect on the accuracy of the clonal haplotype reconstruction, so the next experiment was performed on the number of introduced libraries. The coverage of base library is set to 100×, and the other libraries are all set to 50×, which can reduce the cost of sequencing. The effect of the libraries number on the recognition accuracy of sub-clonal haplotypes is shown in the following Table 3.

As can be seen, the base library cannot recogize the clonal haplotypes well. The recognition rate is less than 90%, the accuracy is less than 60%, and the longest assembled fragment is shorter than 10kbp. Meanwhile, we can see that the more libraries introduced to *MixSubHap*, the better the algorithm performs. Since the cost of sequencing sharply increases when the new library is added, we recommend two libraries with different insert-sizes to reconstruct the clonal haplotypes.

Discussion

The proposed pipeline follows several constraints which imply the inheritance principle: Suppose that the micro evolution process of tumor tissue which satisfies the phylogenetic tree model [35, 45]. Assume that sub-clones in the tumor samples are in a linear evolution mode, the somatic mutations in the evolutionary process satisfying the two hit hypothesis, with the selective advantage, and the sites having the repair mechanism are not considered. In another word, one locus varies at most once in the process of evolution and the mutated site cannot be recovered. Thus, VAF is an important index to distinguish the various sub-clones, following certain inheritance principles in the process of sub-clone differentiation. For any variant site of p, VAF V_p is the number of reads supporting the mutation accounted for the proportion of the site's sequenced depth, which can be statistically calculated from the sequencing reads data. We set the collection of sub-clone S_i 's somatic mutation sites to be M_i, among which $i \in \{0, 1, \ldots, I-1\}$ and I is the total number of sub-clones. Clone S_0 represents sub-clone with the largest common ancestor characteristics estimated from sequencing data, called the founding clone. The set of all the somatic mutations in sequencing samples is $M = \bigcup_i M_i$. Let G_i^p represents the genotype of site p on subclone S_i, S_j represents the descendant clone of S_i, where

Table 1 The accuracy of haplotype reconstruction under different read length

Read length	Recognition rate	Accuracy rate	Longest length	Fragments num
100bp	91.58%	60.50%	10068bp	4
150bp	91.84%	60.41%	10092bp	5
200bp	91.85%	60.50%	10091bp	5
250bp	92.15%	60.39%	10087bp	20

Table 2 The accuracy of haplotype reconstruction under different library coverage

Library num	Cov.(Base)	Cov.	Recognition	Accuracy	Longest length	Fragment num
3	100	20	91.61%	60.46%	10077bp	10
3	100	50	92.15%	60.39%	10087bp	20
3	100	80	92.00%	60.63%	10087bp	13
3	100	100	92.16%	60.35%	10091bp	19

Table 3 The accuracy of haplotype reconstruction under different number of libraries

Library num	Cov.(Base)	Cov.	Recognition	Accuracy	Longest length	Fragment num
1	100	50	89.00%	59.80%	9768bp	8
2	100	50	91.93%	60.10%	10097bp	12
3	100	50	92.15%	60.39%	10087bp	20
4	100	50	92.48%	60.28%	10091bp	32

$i < j$. If $V_p = 1$, then $p \in M_0$ and p is a homozygous mutation, while if $V_p = 0.5$, then $p \in M_0$ and p is a heterozygous mutation. V_p and V_q are the VAF of variant p and q, if $V_p \leq V_q$, then variant q belongs to the subclone which must be the descendant of the clone variant p belongs to. In addition, haplotype heterogeneity follows the inheritance principle that homozygous variation sites which are different from the reference sequence will not appear in clone. So $\sum_k V_{i,p,k} \leq 1$ and inheritance relationship in evolution process: If $V_{i,p,k} = 1$, then for all $i' \geq i$ we have $V_{i',p,k} = 1$, while if $V_{i,p,k} = 0$, then for all $i' \leq i$ we have $V_{i',p,k} = 0$. Among them, $V_{i,p,k}$ indicates the VAF of variation site p is on the kth haplotype of subclone S_i. The reconstruction of clonal haplotypes in tumor sample must satisfy the inheritance principle among sub-clones.

In addition, most of algorithms are based on the second generation sequencing data using linkage disequilibrium of haplotype. Different data input should also be considered [46]. the proposed pipeline consists of three components, which are assembling local VPEs, expanding local contigs and iteratively stripping clonal read-depth. These components are also prolongable to other types of sequencing data once the VPEs can be generated. Based on the second generation sequencing data, the VPEs are short relative to the sub-clonal haplotype, and the uncertainty is quite large: On the aspect of time complexity, if the tumor tissue contains n sites, there are 2^n haplotypes. *MixSubHap* processes VPEs that defines a proper partition length and connection, the short chains after partition can be accurately show sub-clonal haplotypes, and the time complexity is $O((2I)! \times M)$. Among them, M is the number of groups divided by the start position of the variation sites in VPE and I is the number of sub-clones. If the VPEs are extended, M may decrease significantly.

Conclusions

The heterogeneity patterns on haplotypes are suggested to provide not only comprehensive information on tumor evolution and micro-environment, but valuable clinical implications as well. Most of the existing methods investigated the heterogeneity on genotype level, while the computational methods that facilitate the analyses on clonal haplotypes are in urgent needs. In this article, we presented *MixSubHap*, which is a computational pipeline for reconstructing clonal haplotypes. *MixSubHap* is able to identify about 90% on average of the preset clonal haplotypes under different simulation configurations. Especially, *MixSubHap* is robust when decreasing the mutation rates, in which cases the longest assembled contig could reach to 10kbps, while the accuracy of assigning a mutation to its haplotype still keeps more than 60% on average. According to the experimental results on the simulation datasets, we may conclude that the proposed pipeline is a practical tool working on cancer sequencing data. On the other hand, we also notice that two components of *MixSubHap* use the estimation of clonal structure to reduce the solution space. However, such estimation provided by the model-free methods yields more or less errors, especially when the VAFs are close among different sub-clones. This transferred errors could hurt the accuracy on reconstructing clonal haplotypes. Current version only considers point mutations, which will be further extended to structural variations and more complicated cases.

Abbreviations

VAF: Variant allelic frequency

Acknowledgements

The authors would like to thank the conference organizers of the International Conference on Intelligent Biology and Medicine (ICIBM 2018). We also would like to thank the reviewers for their valuable comments and suggestions, which guide us to improve the work and manuscript.

Funding

Publication of this article is funded by the National Science Foundation of China (Grant Nos: 31701150, 91846107, 71571058) and the Fundamental Research Funds for the Central Universities (CXTD2017003), China Postdoctoral Science Foundation funded project 2018M643684.

Availability of data and materials

The source codes has been uploaded and maintained at https://github.com/YixuanWang1120/MixSubHap.git for academic use only.

About this supplement

This article has been published as part of *BMC Medical Genomics Volume 12 Supplement 1, 2019: Selected articles from the International Conference on Intelligent Biology and Medicine (ICIBM) 2018: medical genomics*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-1.

Authors' contributions

JYW and ZXP conducted this research. YXW, SD, YG, JYL and ZMZ designed the algorithms and the pipeline. YXW applied the simulation experiments. YXW, RZ, XX and JYW wrote this manuscript. All authors read and approved the final version of this manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

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

Author details

[1]Department of Computer Science and Technology, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710048, China. [2]Shaanxi Engineering Research Center of Medical and Health Big Data, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710048, China. [3]School of Management, Ministry of Education Key Laboratory of Process Optimization and Intelligent Decision-Making, Hefei University of Technology, Hefei 23009, China. [4]Institute of Health Administration and Policy, School of Public Policy and Administration, Xi'an Jiaotong University, Xi'an 710048, China.

Published: 31 January 2019

References

1. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci. 1971;68(4):820–3.
2. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333–9.
3. Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10(8):e1003665.1–15.
4. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MD, et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun. 2015;10086(6):1–13.
5. Huang KL, Mashl RJ, Wu Y, et al. Pathogenic germline variants in 10,389 adult cancers. Cell. 2018;173(2):355–70.
6. Geng Y, Zhao Z, Zhang X, et al. An improved burden-test pipeline for identifying associations from rare germline and somatic variants. BMC Genomics. 2017;18(7:753):55–62.
7. Landau DA, Carter SL, Stojanov P, et al. Evolution and impact of subclonal mutations in Chronic Lymphocytic Leukemia. Cell. 2013;152(4):714–26.
8. Keats JJ, Chesi M, Egan JB, et al. Clonal competition with alternating dominance in multiple myeloma. Blood. 2012;120(5):1067–76.
9. Merlo LM, Shah NA, Li X, et al. A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. Cancer Prev Res. 2010;3(11):1388–97.
10. Turke AB, Zejnullahu K, Wu YL, et al. Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. Cancer Cell. 2010;17(1):77–88.
11. Huang Y, Wang J, Jia P, et al. Clonal architectures predict clinical outcome in clear cell renal cell carcinoma. Nat Commun. 2018. accepted and in publication.
12. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
13. Ye K, Wang J, Jayasinghe R, et al. Systematic discovery of complex insertions and deletions in human cancers. Nat Med. 2016;22(1):97–104.
14. Zheng T, Li Y, Geng Y, et al. CIGenotyper: A machine learning approach for genotyping complex indel calls. In: Rojas I, Ortuño F, editors. Bioinformatics and Biomedical Engineering. IWBBIO 2018: Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science, Vol 10813. Cham: Springer; 2018. p. 473–85.
15. Geng Y, Zhao Z, Xu J, et al. Identifying heterogeneity patterns of allelic imbalance on germline variants to infer clonal architecture. In: Huang DS, Jo KH, Figueroa-García J, editors. Intelligent Computing Theories and Application. ICIC 2017: Intelligent Computing Theories and Application. Lecture Notes in Computer Science, Vol 10362. Cham: Springer; 2017. p. 286–297.
16. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12(10):703–14.
17. Aguiar D, Wong W, Istrail S. Tumor haplotype assembly algorithms for cancer genomics. In: Proceedings of the 2014 Pacific Symposium on Biocomputing (PSB 2014). Big Island. Biocomputing 2014, pp. 3-14 (2013) https://doi.org/10.1142/9789814583220_0002.
18. Xu M, Zhao Z, Zhang X, et al. Synstable Fusion: a network-based algorithm for estimating driver genes in fusion structures. Molecules. 2018;2055(8):1–20.
19. Xu H, Guan Y. Detecting local haplotype sharing and haplotype association. Genetics. 2014;197(3):823–38.
20. Samwald M, Miñarro Giménez JA, Boyce RD, et al. Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. BMC Med Inform Decis Making. 2015;15(12):1–10.
21. Wang H, Ding S, Wu D, et al. Smart connected electronic gastroscope system for gastric cancer screening using multi-column convolutional neural networks. Int J Prod Res. online publication. https://doi.org/0.1080/00207543.2018.1464232.
22. Ding S, Li Y, Wu D, et al. Time-aware cloud service recommendation using similarity-enhanced collaborative filtering and ARIMA. Decis Support Syst. 2018;107:103–15.
23. Hoehe MR, Timmermann B, Lehrach H. Human inter-individual DNA sequence variation in candidate genes, drug targets, the importance of haplotypes and pharmacogenomics. Curr Pharm Biotechnol. 2003;4(6): 351–78.
24. McLeod HL. Cancer pharmacogenomics: early promise, but concerted effort needed. Science. 2013;339(6127):1563–6.
25. Xie M, Lu C, Wang J, et al. Age-related cancer mutations associated with clonal hematopoietic expansion and malignancies. Nat Med. 2014;20(12): 1472–8.
26. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 2017;18(4):213–29.
27. Strino F, Parisi F, Micsinai M, et al. TrAp: a tree approach for fingerprinting subclonal tumor composition. Nucleic Acids Res. 2013;e165(17):1–15.
28. Deshwar AG, Vembu S, Yung CK, et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16(35):1–20.
29. Yuan K, Sakoparnig T, Markowetz F, et al. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biol. 2015;16(36):1–16.
30. El-Kebir M, Satas G, Oesper L, et al. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Syst. 2016;3(1): 43–53.
31. Jiang Y, Qiu Y, Minn A, et al. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016;113(37): E5528–37.
32. Eaton J, Wang J, Schwartz R. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. Bioinformatics. 2018;34(13):i357—65.
33. Beerenwinkel N, Schwarz RF, Gerstung M, et al. Cancer evolution: mathematical models and computational inference. Syst Biol. 2015;64(1): e1–25.
34. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013;R80(7):1–21.
35. Jiao W, Vembu S, Deshwar AG, et al. Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics. 2014;15(35):1–16.
36. Roth A, Khattra J, Yap D. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11(4):396–8.
37. Ha G, Roth A, Khattra J, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014;24(11):1881–93.
38. Roman T, Xie L, Schwartz R. Automated deconvolution of structured mixtures from heterogeneous tumor genomic data. PLoS Comput Biol. 2017;13(10):e1005815.1-23.
39. He D, Saha S, Finkers R, et al. Efficient algorithms for polyploid haplotype phasing. BMC Genomics. 2018;19(Suppl 2):110.1–10.
40. Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics. 2008;24(16):i153–9.
41. Aguiar D, Istrail S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. Bioinformatics. 2013;29(13):i352–60.
42. Berger E, Yorukoglu D, Peng J, et al. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. PLoS Comput Biol. 2014;10(3):e1003502.1-10.
43. Das S, Vikalo H. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. BMC Genomics. 2015;16(260):1–16.

44. Hashemi A, Zhu B, Vikalo H. Sparse tensor decomposition for haplotype assembly of diploids and polyploids. BMC Genomics. 2018;19(Suppl 4): 191.1–15.
45. Geng Y, Zhao Z, Liu J, et al. An algorithm with base-pair resolution for identifying cancer heterogeneity by estimating multiple clonal haplotypes. J Xi'an Jiaotong Univ. 2017;51(6):92–6.
46. Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. 2017;27(5): 801–12.