# Beehive Tutorial

Version: 0.1

## I. Introduction

Beehive is an algorithm to estimate the timing of chromosomal copy gain and clonal expansion during tumorigenesis. The estimation is based on somatic single nucleotide variants (SNVs) because they occur at each cell division. When the chromosomal gain occurs, the somatic SNVs are duplicated. The number of duplicated and unduplicated somatic SNVs can tell us the timing of chromosomal gain. Similarly, the number of clonal and subclonal somatic SNVs can tell us when the clonal expansion occurs. This tool works the best on whole-genome sequencing (WGS) data in which copy number, tumor purity and subclone structure can be accurately inferred. There are some caveats in this type of analysis and one shall interpret the results with caution. Please read our paper listed in Section V for more details. This document only covers technical aspects. A lot of examples will be used to demonstrate how to use beehive. We strongly recommend the users to read through all examples carefully, especially the examples in advanced mode, and make sure you understand the complexity of data in order to choose the proper modes and parameters.

## II. Prerequisites

### A. Software environment:

1. Unix/Linux system.
2. PERL 5.8.1 or above.
3. R 2.6.1 or above (you must be able to run Rscript from command line).

### B. Input file:

A text file of allele fractions is the required input file. The file shall have two columns: read count for alternative allele and total coverage. Each line contains information for one somatic SNV. See examples in the example folder. This file can easily be generated from a VCF file.

The evolutionary history for the region of interest has to be determined prior to running this tool. We will discuss how to infer and validate predicted evolutionary history from somatic SNVs. Sometimes if complex history or subclonal events are involved, it may not be possible to infer the history with high confidence.

## III. Standard mode

## A. Infer tumor purity and copy number

Before running Beehive, you need to determine tumor purity and integer copy number for each chromosomal region. There are plenty of tools available for this task. These tools typically use copy ratio between tumor tissue and match normal tissue, minor allele (B-allele) frequencies (BAFs) for germline heterozygous SNVs, and mutant allele fractions (MAFs) for somatic SNVs. Some of these tools work better than others. We highly recommend to manually review the results since the determination of copy number is critical for tumor evolutionary history reconstruction. The predicted tumor purity and copy numbers shall be compatible with copy ratio, germline SNV BAFs and somatic SNV MAFs. Beehive can also provide some hints on the tumor purity and copy number prediction. Figure 1 shows the genetic alteration profile for a colorectal tumor crc3913 described in our paper. The tumor purity is 0.74. The integer copy number can be inferred for most chromosomes.

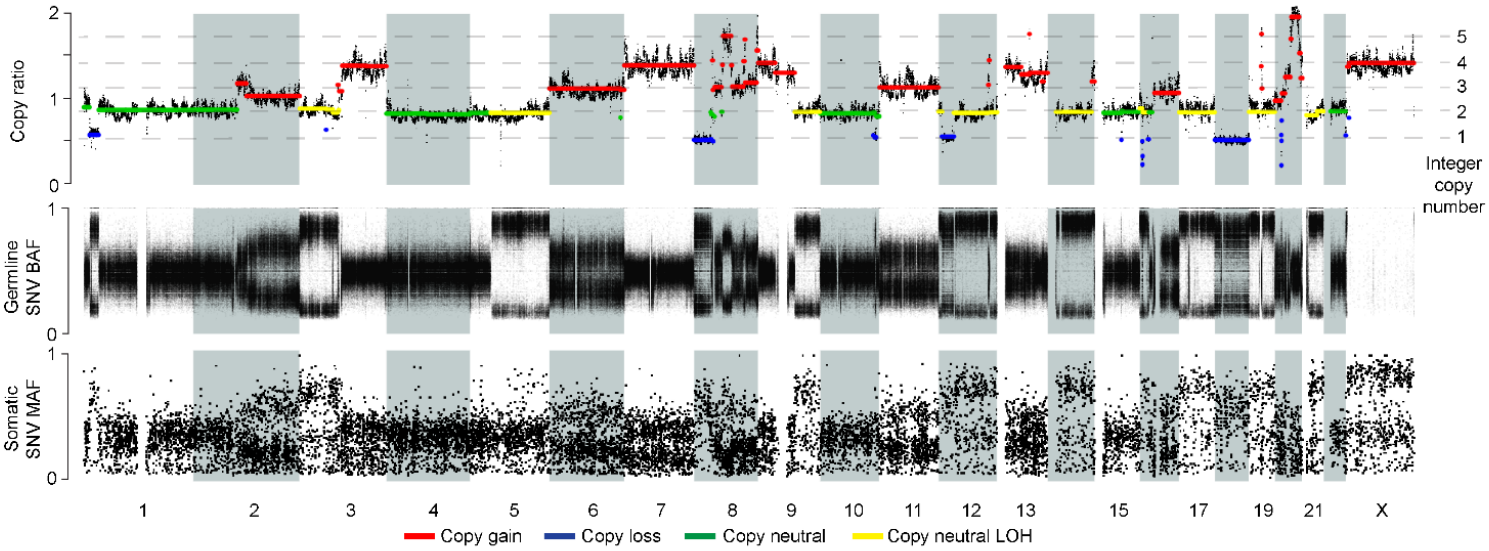

**Figure 1. Genetic alteration profile for tumor crc3913.**

## B. beehive.pl

Usage:

```
beehive.pl [options]
      -i FILE    input file, required
      -o FILE    output file prefix, default input file name
      -t STR     event type, format x+y, x: copy number of major
allele, y: copy number for minor allele
      -b INT     bootstrap, 0/1, default 0, set to 1 to enable
bootstrap
      -n INT     number of bootstrap, default 1000
      -s INT     spline smooth parameter, default 30
      -p STR     predefined peak locations, minimum 2 peaks, maximum 4
peaks, format: p1,p2,p3,p4
      -d STR     denominators for calculating timing based on
predefined peaks, minimum 2, maximum 4, format: d1,d2,d3,d4
```

```
      -m STR     multipliers for calculating timing based on predefined
peaks, minimum 1, maximum 6, format: m11,m21,m22,m31,m32,m33
      -h help
      version: 0.1
```

The input file contains a list of allele fractions of somatic SNVs in the given regions. If multiple chromosomes have the same copy number and evolutionary history, they can be merged together. A normal cell has two copies of each autosome, a paternal copy and a maternal copy. These two different copies can be differentiated based on germline heterozygous SNVs. The option t needs to be given as x+y format. “X” represents the copy number of major allele (the one with higher copy number) while “y” represents the copy number of minor allele (the one with lower copy number). In copy neutral wild-type region, it shall be given as “1+1”. Copy-neutral loss of heterozygosity (CN-LOH), one-copy loss, one-copy gain, mono-allelic two-copy gain and bi-allelic two-copy gain shall be given as “2+0”, “1+0”, “2+1”, “3+1” and “2+2”, respectively. Option s determines how smooth the spline curve is. Usually, you don’t need to change option s. In standard mode, we will use option t and assume copy gain occur at one time. Even with high copy number, such as “3+3”, “4+1”, etc., we will assume all copies are gained at the same time. If you believe there is complex history that multiple copies are gained at different time, you will need to use advanced mode which is discussed in Section IV. Now, we will use a few examples to demonstrate how the standard mode works.

## C. Examples

### 1. CN-LOH

In CN-LOH regions, one parental copy is lost and the other copy is duplicated. The loss and duplication may occur at different times. We are only able to measure the duplication here. Figure 2 shows CN-LOH of chromosomes 3p,9q,12,14,17,19,21 in tumor crc3913. All somatic SNVs in these chromosomes are merged. The green chromosome is lost and the orange chromosome is duplicated (left panel of Figure 2). The red SNVs have occurred before the CN-

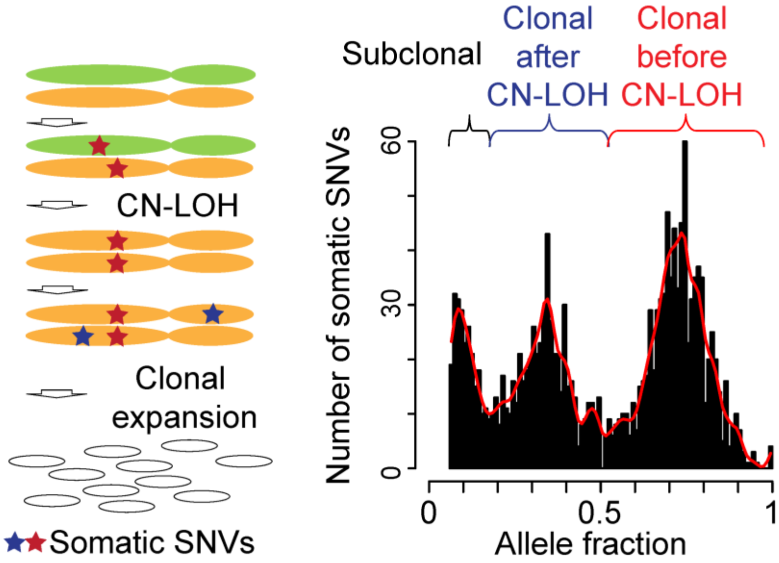

**Figure 2. CN-LOH regions in crc3913.**

LOH are duplicated. The distribution of MAFs on the right indicates there are three peaks: red SNVs occur before CN-LOH, blue SNVs occur after CN-LOH and black SNVs (subclonal). The red SNVs are present on both copies of the chromosomes while the blue SNVs are present on only one of the two copies of the chromosomes. So the MAFs of blue SNVs should be half of red SNVs. The "2+0" copy number prediction is compatible with the MAF profile. The expected MAFs for red and blue SNVs depend on tumor purity. If the tumor is 100% pure, the red and blue SNVs shall have MAFs of 1 and 0.5. We then can use the deviation of MAFs to predict tumor purity.

When we run the following command:

```
perl beehive.pl -i ./examples/crc3913.2+0 -t 2+0
```

We will get a pdf plot shown in Figure 3 and the following output:

```
Peaks called: 0.09 0.35 0.74
Valleys called:  0.18 0.52
Number of SNVs in bin1: 844
Number of SNVs in bin2: 544
Total number of SNVs: 1632
Peak position for amplified alleles: 0.74
Peak position for unamplified alleles: 0.35
Predicted clonality based on event type: 0.74
Timing of copy change: 0.68 (, )
Timing of clonal expansion: 0.90 (, )
```

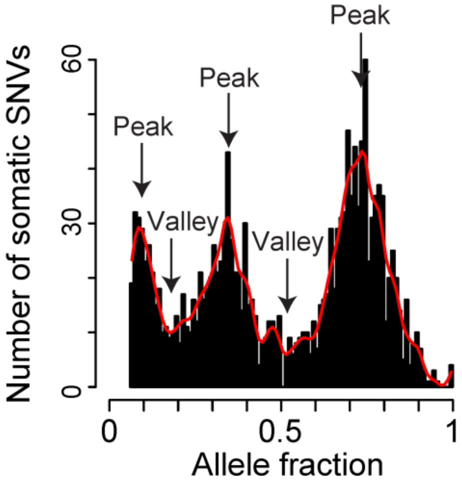

**Figure 3. Modeling MAFs.**

Beehive fits a spline smoothing curve, identifies the peaks and the valleys in the MAF profile (Figure 3) and assign somatic SNVs into three categories: clonal amplified, clonal unamplified, and subclonal (Figure 2 right panel). As shown in the output, 3 peaks are called at 0.74, 0.35 and 0.09, 2 valleys are called at 0.52 and 0.18. As expected, the peak location of clonal unamplified alleles (0.35) is roughly half of the peak of clonal amplified alleles (0.74). Based on the peak location, clonality is predicted as 0.74 and is exactly the same as tumor purity 0.74. These results indicate the prediction of tumor purity and copy number are compatible with somatic SNVs. The timing of copy change is measured to be 0.68 as the proportion of somatic SNVs occurred before CN-LOH. The timing of clonal expansion is measured to be 0.90 as the proportion of clonal SNVs. If we turn on the bootstrap function as follow:

```
perl beehive.pl -i ./examples/crc3913.2+0 -t 2+0 -b 1
```

We will get the 95% confidence intervals of timing estimations:

```
Peaks called: 0.09 0.35 0.74
Valleys called:  0.18 0.52
Number of SNVs in bin1: 844
Number of SNVs in bin2: 544
Total number of SNVs: 1632
Peak position for amplified alleles: 0.74
Peak position for unamplified alleles: 0.35
Predicted clonality based on event type: 0.74
Timing of copy change: 0.68 (0.66, 0.91)
Timing of clonal expansion: 0.90 (0.89, 0.91)
```

Confidence intervals will be given for both timing of copy change and clonal expansion in the parentheses.

## 2. Copy neutral

For copy neutral regions (chromosomes 1,2p,4,10,15,22) in crc3913 (Figure 4), we run:

```
perl beehive.pl -i ./examples/crc3913.1+1 -t 1+1
```

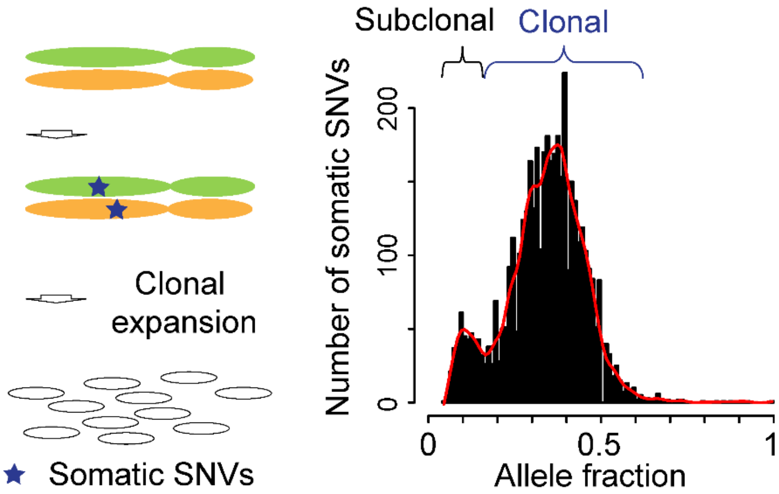

**Figure 4. Copy neutral regions in crc3913.**

And the output will be:

```
Peaks called: 0.11 0.38
Valleys called:   0.17
Number of SNVs in bin1: 3911
Total number of SNVs: 4357
Peak position for amplified alleles: 0.38
```

```
Predicted clonality based on event type: 0.76
Timing of clonal expansion: 0.90 (, )
```

Two peaks are identifiable. The clonality is predicted to be 0.76 based on peak location which is very close to the tumor purity 0.74. The timing of clonal expansion is estimated to be 0.90, same as CN-LOH regions.

### 3. One-copy gain

For one copy gain regions (chromosomes 1,2p,4,10,15,22) in crc3913 (Figure 5), we run:

```
perl beehive.pl -i ./examples/crc3913.2+1 -t 2+1
```

And the output will be:

```
Peaks called: 0.09 0.25 0.54
Valleys called:  0.13 0.42
Number of SNVs in bin1: 555
Number of SNVs in bin2: 1244
Total number of SNVs: 1956
Peak position for amplified alleles: 0.54
Peak position for unamplified alleles: 0.25
Predicted clonality based on event type: 0.74
Timing of copy change: 0.66 (, )
Timing of clonal expansion: 0.94 (, )
```

Three peaks are identified (Figure 5). The peak position of unamplified alleles (0.25) is roughly half of the peak of amplified alleles (0.54). Note that the middle peak around 0.25 is a mixture of red and blue SNVs. The red SNVs on the green chromosome are unamplified although they occurred before copy gain. Clonality is predicted to be 0.74 which is identical to the tumor purity 0.74. All these results suggest the predicted one-copy gain and tumor purity are compatible with the somatic SNVs. The timing of copy gain is 0.66 which is very close to the

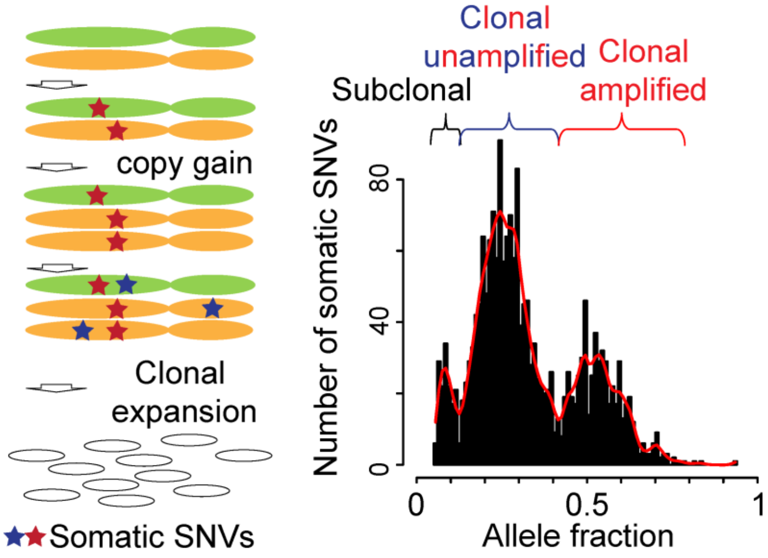

**Figure 5. One-copy gain regions in crc3913.**

timing of CN-LOH 0.68. The one-copy gain most likely have occurred at the same time as CN-LOH. The timing of clonal expansion is less reliable when there are more than two copies of chromosomes because the MAFs are pushed towards the left and few subclonal SNVs are detectable. Therefore, we suggest to only use two-copy or one-copy region for clonal expansion time estimation.

### 4. Bi-allelic two-copy gain

For bi-allelic two-copy gain regions (chromosomes 3q,7,13) in crc3913 (Figure 6), we run:

```
perl beehive.pl -i ./examples/crc3913.2+2 -t 2+2
```

And the output will be:

```
Peaks called: 0.08 0.2 0.44
Valleys called:  0.09 0.29
Number of SNVs in bin1: 1052
Number of SNVs in bin2: 759
Total number of SNVs: 1914
Peak position for amplified alleles: 0.44
Peak position for unamplified alleles: 0.2
Predicted clonality based on event type: 0.79
Timing of copy change: 0.71 (, )
Timing of clonal expansion: 0.97 (, )
```

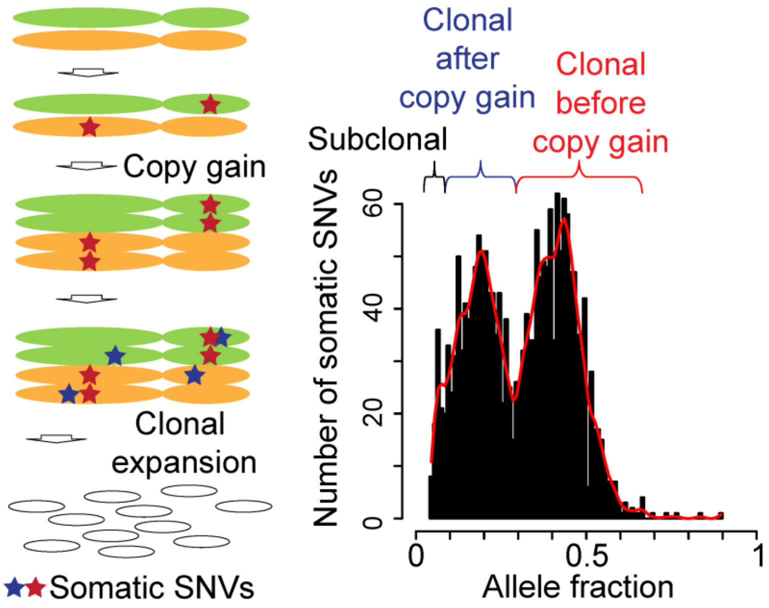

**Figure 6. Bi-allelic two-copy gain regions in crc3913.**

Although three peaks are called (Figure 6), the subclonal SNVs (peak of 0.08) are barely distinguishable from the clonal SNVs occurred after copy gain (blue SNVs). So we shall not use the timing estimation of clonal expansion. The peak location of unamplified alleles (0.2) is roughly half of the peak of amplified alleles (0.44). Clonality is predicted to be 0.79 which is quite close to the tumor purity 0.74. These results suggest the copy number and tumor purity is compatible with the somatic SNVs. The predicted timing of copy gain is 0.71 which is also very close to the timing of CN-LOH 0.68.

To demonstrate the utility of confirming copy number and tumor purity prediction using somatic SNVs. Let's run this bi-allelic two-copy gain regions as if they are one-copy gain:

```
    perl beehive.pl -i ./examples/crc3913.2+2 -t 2+1
```

And the output will be:

```
Peaks called: 0.08 0.2 0.44
Valleys called:  0.09 0.29
Number of SNVs in bin1: 1052
Number of SNVs in bin2: 759
Total number of SNVs: 1914
Peak position for amplified alleles: 0.44
Peak position for unamplified alleles: 0.2
Predicted clonality based on event type: 0.56
Timing of copy change: 1.06 (, )
Timing of clonal expansion: 0.97 (, )
```

Again, three peaks are identified, but the peak of 0.08 is unreliable. The peak location of unamplified allele (0.2) is roughly half of amplified allele (0.44), which is compatible with one-copy gain prediction. However, the predicted clonality 0.56 is quite different from tumor purity 0.74. In addition, the timing of copy change is 1.06. Timing of copy change and clonal expansion shall always be between 0 and 1. The incompatible clonality and timing suggest one-copy gain is unlikely to be the case for these chromosomes.

### 5. Mono-allelic two-copy gain

For mono-allelic two-copy gain regions (chromosomes 3q,6,7,10,12) in crc2683 (Figure 7, note this is a different tumor from the previous examples), we run:

```
perl beehive.pl -i ./examples/crc2683.3+1 -t 3+1
```

And the output will be:

```
Peaks called: 0.23 0.23 0.71
Valleys called:  0.05 0.47
Number of SNVs in bin1: 1479
Number of SNVs in bin2: 5871
Total number of SNVs: 7351
Peak position for amplified alleles: 0.71
Peak position for unamplified alleles: 0.23
Predicted clonality based on event type: 0.90
Timing of copy change: 0.57 (, )
Timing of clonal expansion: 1.00 (, )
```

The tumor purity of this sample crc2683 is 0.98. Two peaks are called at 0.71 and 0.23 (Figure 7). The subclonal peak is not distinguishable. The peak location of unamplified alleles (0.23) is roughly one third of the peak of amplified alleles (0.71). Again, the unamplified SNVs include both red and blue SNVs. Clonality is predicted to be 0.90 which is close to the tumor purity 0.98. These results suggest the copy number and tumor purity is compatible with the somatic SNVs.

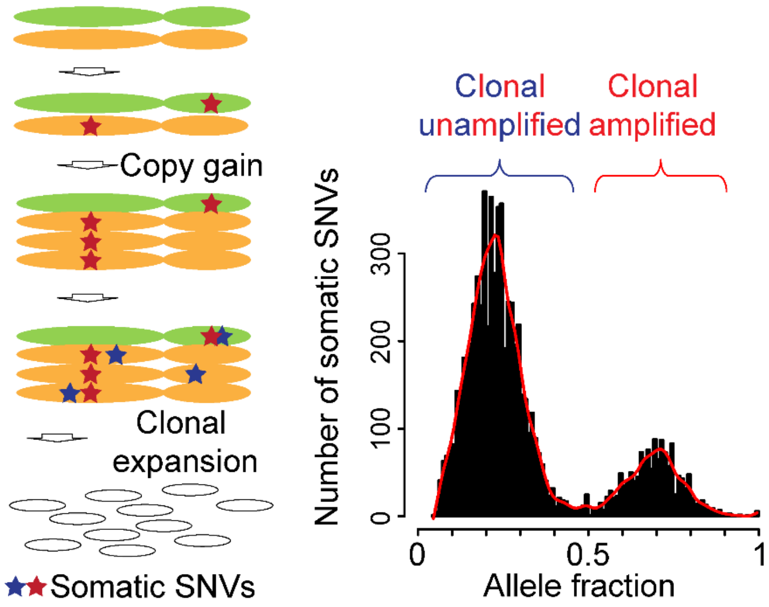

**Figure 7. Mono-allelic two-copy gain regions in crc2683.**

# IV. Advanced mode

In Section III, if there are more than one duplications, we assume all duplications occur at the same time. If the duplications occur at different time, you will need to use advanced mode described in this section. If the peaks are called incorrectly, you shall use advanced mode as well. Note that every region runs in standard mode can be run in advanced mode. Advanced mode offers more flexibility to handle situations that standard mode cannot deal with. Now we will use a few examples to demonstrate how to use advanced mode. In advanced mode, you will not use option t, but instead, use options p, d and m. Option p takes pre-defined peak locations. A minimum of two and maximum of four peaks are allowed here. The locations given do not need to be precise, Beehive will search for the peak in the neighboring region you specified. Options d and m are denominators and multipliers you need to provide to describe a specific evolutionary history. When two peaks are defined, two denominators and one multiplier are required. When three peaks are defined, three denominators and three multipliers are required. When four peaks are defined, four denominators and six multipliers are required. Once again, the evolutionary history has to be determined prior to applying either standard mode or advanced mode to measure timing.

The core of timing estimation is to count the number of somatic SNVs during each time period. In the example shown in Figure 8, four peaks are present. There are four time periods we can model: t1, t2, t3 and t4. The timing is estimated by the proportion of somatic SNVs, but we need to adjust the number of SNVs based on how many chromosomes can independently acquire SNVs during each time period. Let’s use n1, n2, n3 and n4 to denote number of SNVs observed in each peak; adj_n1, adj_n2, adj_n3 and adj_n4 to denote the adjusted number of SNVs during time period t1 to t4; d1, d2, d3 and d4 to denote number of chromosomes independently acquiring SNVs and contributing to each peak. We will use the following formulae to calculate timing:

$$\mathrm{adj\_n1} = \mathrm{n1} / \mathrm{d1} \tag{1}$$
$$\mathrm{adj\_n2} = (\mathrm{n2} - \mathrm{m11} * \mathrm{adj\_n1}) / \mathrm{d2} \tag{2}$$
$$\mathrm{adj\_n3} = (\mathrm{n3} - \mathrm{m22} * \mathrm{adj\_n2} - \mathrm{m21} * \mathrm{adj\_n1}) / \mathrm{d3} \tag{3}$$
$$\mathrm{adj\_n4} = (\mathrm{n4} - \mathrm{m33} * \mathrm{adj\_n3} - \mathrm{m32} * \mathrm{adj\_n2} - \mathrm{m31} * \mathrm{adj\_n1}) / \mathrm{d4} \tag{4}$$
$$\mathrm{adj\_n} = \mathrm{adj\_n1} + \mathrm{adj\_n2} + \mathrm{adj\_n3} + \mathrm{adj\_n4}$$
$$\mathrm{t1} = \mathrm{adj\_n1} / \mathrm{adj\_n}$$
$$\mathrm{t1} + \mathrm{t2} = (\mathrm{adj\_n1} + \mathrm{adj\_n2}) / \mathrm{adj\_n}$$
$$\mathrm{t1} + \mathrm{t2} + \mathrm{t3} = (\mathrm{adj\_n1} + \mathrm{adj\_n2} + \mathrm{adj\_n3}) / \mathrm{adj\_n}$$
$$\mathrm{t1} + \mathrm{t2} + \mathrm{t3} + \mathrm{t4} = 1$$

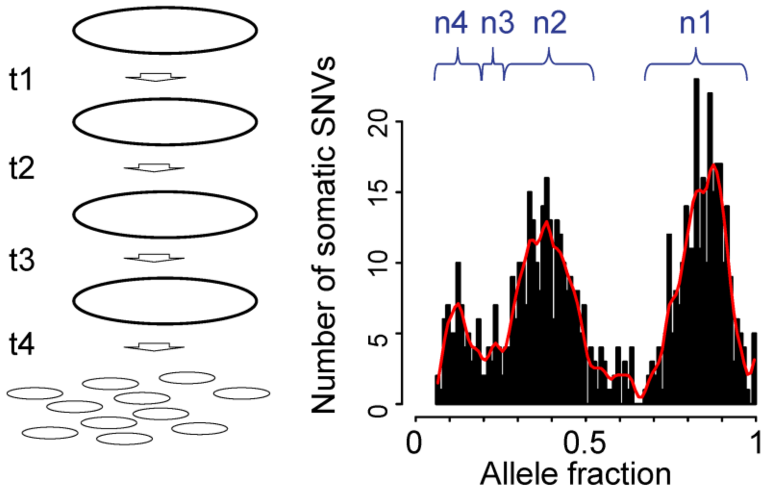

**Figure 8. Four peaks of somatic SNVs occurred during four time periods.**

m11, m21, m22, m31, m32 and m33 are multipliers. We will show how to use them in examples. For each evolutionary history, you will need to determine d and m parameters in formulae (1) to (4) provide them to beehive.pl. Option d needs to be given in the format of “d1,d2,d3,d4”. Option m needs to be given in the format of “m11,m21,m22,m31,m32,m33”. If there are only two or three peaks, we will only use the first two or three formulae rather than all four.

## 1. One-copy loss

Let’s start with a simplest case. For one-copy loss regions (chromosomes 8p,18) in crc3913 (Figure 9), we can run the standard mode:

```
perl beehive.pl -i ./examples/ crc3913.1+0 -t 1+0
```

And the output will be:

```
Peaks called: 0.38 0.59
Valleys called:   0.41
Number of SNVs in bin1: 303
Total number of SNVs: 394
Peak position for amplified alleles: 0.59
Predicted clonality based on event type: 0.74
Timing of clonal expansion: 0.77 (, )
```

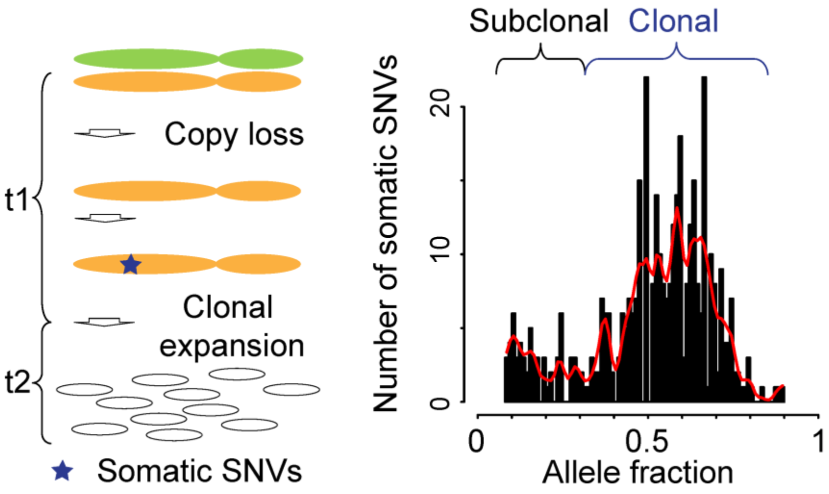

**Figure 9. One-copy loss regions in crc23913.**

Two peaks are called at 0.59 and 0.38. The subclonal peak is called at a wrong location. It should be at around 0.1 rather than at 0.38. The clonal expansion time estimation will be wrong when peaks are called wrong. To manually specify the correct peak location, we will need to use the advanced mode. There is only one chromosome accumulating somatic SNVs, so d1 and d2 are both 1. We will set m11 as 0 for now, and will explain the meaning of multipliers later. We run:

```
perl beehive.pl -i ./examples/crc3913.1+0 -p 0.59,0.1 -d 1,1 -m 0
```

And the output will be:

```
Peaks called: 0.11 0.59
Valleys called:   0.32
Number of SNVs in bin1: 332
Total number of SNVs: 394
Timing of 1st event: 0.84 (, )
```

Note that the actual subclonal peak location is 0.11. Although the peak we gave is 0.1, Beehive is able to find the true peak in the surrounding region. The timing of 1st event which is clonal expansion is 0.84 which is close to the estimation based on CN-LOH regions (0.90) and copy neutral regions (0.90). The clonal expansion timing estimation from one-copy region is almost always smaller than two-copy region because more subclonal SNVs can be detected in one-copy region. In most tumors, the number of somatic SNVs in one-copy regions is small, so we recommend to use two-copy regions to estimate the timing of clonal expansion.

## 2. CN-LOH

We've shown how to use standard mode on CN-LOH regions. In advanced mode, we can run:

```
perl beehive.pl -i ./examples/crc3913.2+0 -p 0.75,0.4,0.1 -d 1,2,2 -m
0,0,0
```

And the output will be:

```
Peaks called: 0.09 0.35 0.74
Valleys called:  0.18 0.52
Number of SNVs in bin1: 844
Number of SNVs in bin2: 544
Total number of SNVs: 1632
Timing of 1st event: 0.68 (, )
Timing of 2nd event: 0.90 (, )
```

In this case, three peaks are called and represent somatic SNVs accumulated during t1, t2 and t3 (Figure 10). Although there are two chromosomes independently accumulating SNVs during t1, the SNVs contribute to the right-most peak only from one chromosome. In contrast, SNVs are accumulating on two chromosomes independently after CN-LOH and in subclone. All these SNVs contribute to middle and left-most peaks. So the denominators are 1, 2 and 2. The multipliers are still set as 0. The timing of 1st and 2nd events are CN-LOH and clonal expansion, respectively, and are 0.68 and 0.90, same as in standard mode.

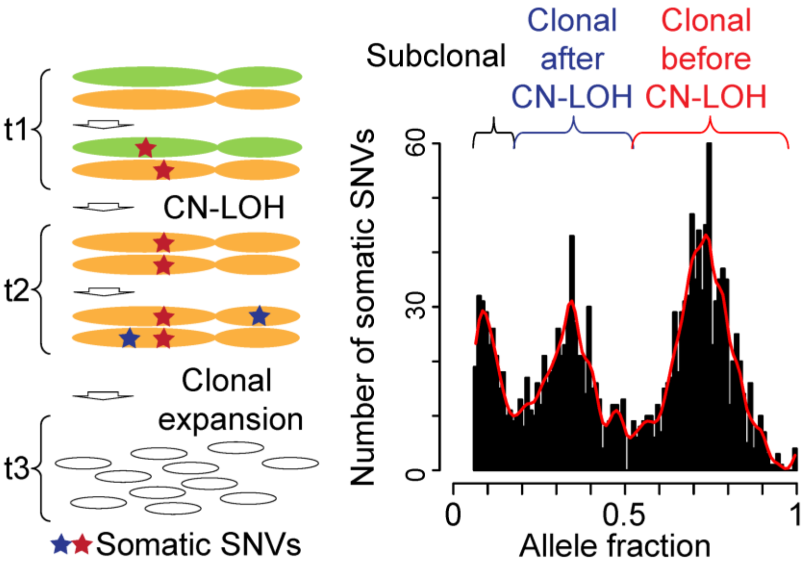

**Figure 10. CN-LOH regions in crc3913.**

Next, let's look at another CN-LOH region (chromosome 5q) in crc3913 (Figure 11). There are very few red SNVs in this region suggesting the CN-LOH is a very early event. When we run standard mode, the peak at around 0.8 is not called because it is too small to be recognized. Usually, it is a sign of miscalling copy number or tumor purity or both. However, after reviewing the copy ratio, germline SNV BAFs (Figure 1), we are quite sure it is a CN-LOH region in major clone. To estimate the timing properly, we will need to use advanced mode to define peak locations. We run:

```
perl beehive.pl -i ./examples/crc3913.2+0.s -p 0.8,0.4,0.1 -d 1,2,2 -m
0,0,0
```

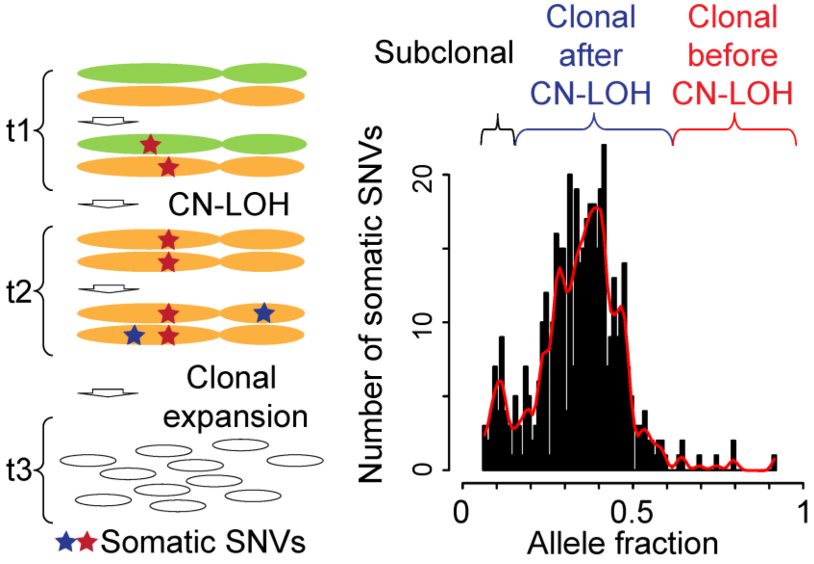

**Figure 11. Early CN-LOH in crc3913.**

And the output will be:

```
Peaks called: 0.11 0.4 0.8
Valleys called:  0.15 0.68
Number of SNVs in bin1: 5
Number of SNVs in bin2: 397
Total number of SNVs: 439
Timing of 1st event: 0.02 (, )
Timing of 2nd event: 0.92 (, )
```

Now the peaks are identified properly. The timing of CN-LOH is 0.02 which is much earlier than other CN-LOH chromosomes (3p,9q,12,14,17,19,21). The timing of clonal expansion is 0.92 which is very similar to the estimation (0.90) from chromosomes 3p,9q,12,14,17,19,21. This suggests the CN-LOH call is compatible with somatic SNVs. This is a case that cannot be properly handled by standard mode, but ideal for advanced mode.

## 3. Copy neutral

We've shown how to use standard mode on copy neutral regions. In advanced mode, we can run:

```
perl beehive.pl -i ./examples/crc3913.1+1 -p 0.4,0.1 -d 2,2 -m 0
```

And the output will be:

```
Peaks called: 0.11 0.38
Valleys called:   0.17
Number of SNVs in bin1: 3911
```

```
Total number of SNVs: 4357
Timing of 1st event: 0.90 (, )
```

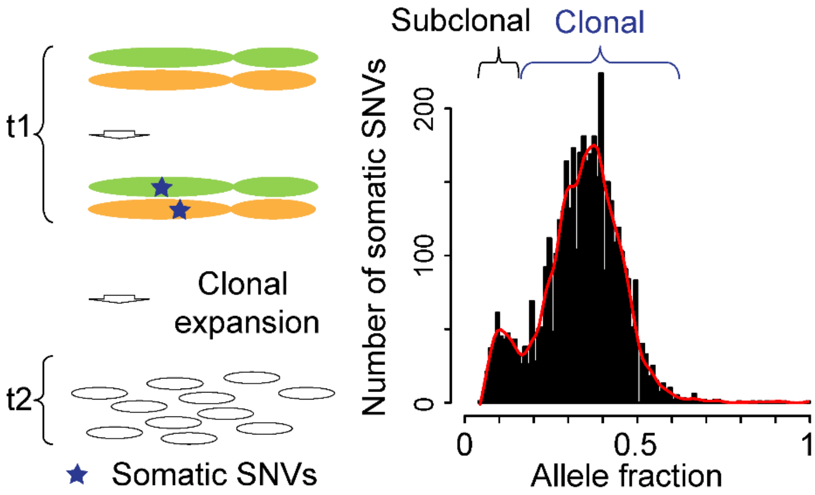

**Figure 12. Copy neutral regions in crc3913.**

Two peaks are called as clonal and subclonal SNVs (Figure 12). Two chromosomes are accumulating SNVs in both major clone and subclone. So the denominators are 2 and 2. The multipliers are still set as 0. The timing of 1st evens is clonal expansion, and is 0.90, same as in standard mode.

## 4. One-copy gain

We've shown how to use standard mode on one-copy gain regions. In advanced mode, we can run:

```
perl beehive.pl -i ./examples/crc3913.2+1 -p 0.5,0.25,0.1 -d 1,3,3 -m
1,0,0
```

And the output will be:

```
Peaks called: 0.09 0.25 0.5
Valleys called:  0.13 0.42
Number of SNVs in bin1: 555
Number of SNVs in bin2: 1244
Total number of SNVs: 1956
Timing of 1st event: 0.66 (, )
Timing of 2nd event: 0.94 (, )
```

In advanced mode, multipliers will need to be used in this case. In the MAF profile, there are three peaks (Figure 13). The red SNVs on green chromosome and all blue SNVs are present on one of the three chromosomes, while the red SNVs on orange chromosome are present on two of the three chromosomes with higher MAF. The right-most peak contains SNVs that are amplified. The middle peak is a mixture of SNVs occurred both before (red) and after (blue) copy gain. When we adjust the number of SNVs in the middle peak to calculate adj_n2 during t2, we need to subtract the number of SNVs occurred before copy gain in formula (2). For the subclonal peak, we don't need to adjust the number of SNVs. So the m option will be "1,0,0".

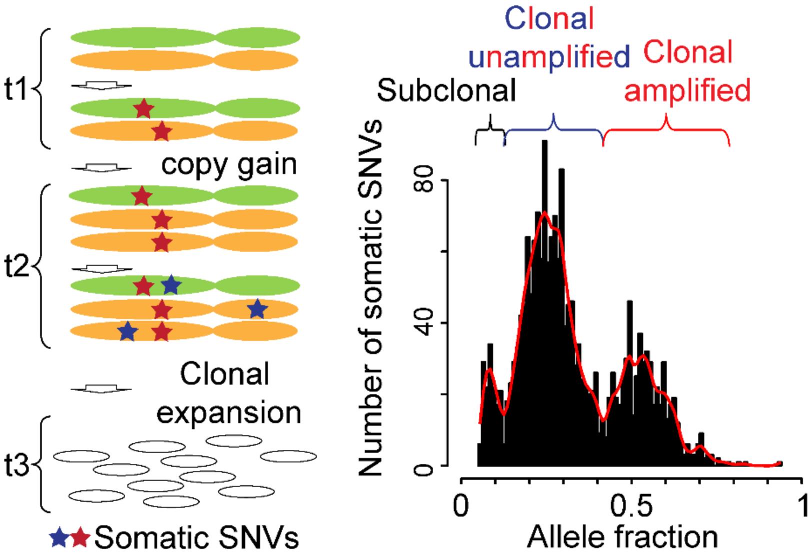

**Figure 13. One-copy gain regions in crc3913.**

The amplified red SNVs are accumulating on one chromosome, the clonal SNVs after copy gain (blue SNVs) are accumulating independently on three chromosomes, and similarly the subclonal SNVs are accumulating on three chromosomes. So the d option will be given as “1,3,3”.

## 5. Bi-allelic two-copy gain

We’ve shown how to use standard mode on bi-allelic two-copy gain regions. Previously, a subclonal peak is called at 0.08, but the peak is really small. So here, we will not call that peak (Figure 14). In advanced mode, we can run:

```
perl beehive.pl -i ./examples/crc3913.2+2 -p 0.44,0.2 -d 2,4 -m 0
```

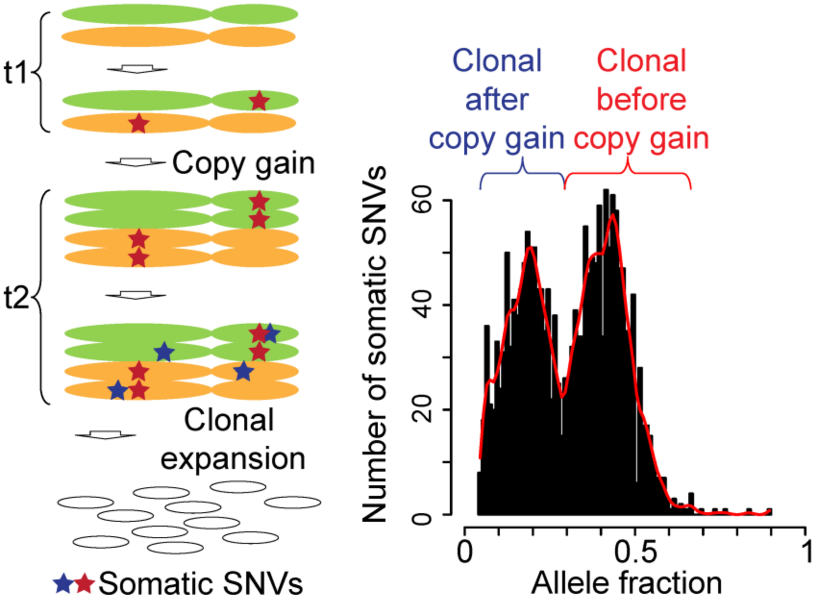

**Figure 14. Bi-allelic two-copy gain regions in crc3913.**

And the output will be:

```
Peaks called: 0.2 0.44
Valleys called:   0.29
Number of SNVs in bin1: 1052
Total number of SNVs: 1914
Timing of 1st event: 0.71 (, )
```

There are two chromosomes independently accumulating SNVs and contributing to the amplified peak while there are four chromosomes accumulating SNVs after the copy gain. So we set d option as “2,4”. No SNV needs to be adjusted, so we set m option as “0”.

## 6. Mono-allelic two-copy gain

We have used standard mode on mono-allelic two-copy gain regions before on chromosomes 3q,6,7,10,12 in tumor crc2683. In advanced mode, we run:

```
perl beehive.pl -i ./examples/crc2683.3+1 -p 0.71,0.23 -d 1,4 -m 1
```

And the output will be:

```
Peaks called: 0.23 0.71
Valleys called:   0.47
Number of SNVs in bin1: 1479
Total number of SNVs: 7351
Timing of 1st event: 0.57 (, )
```

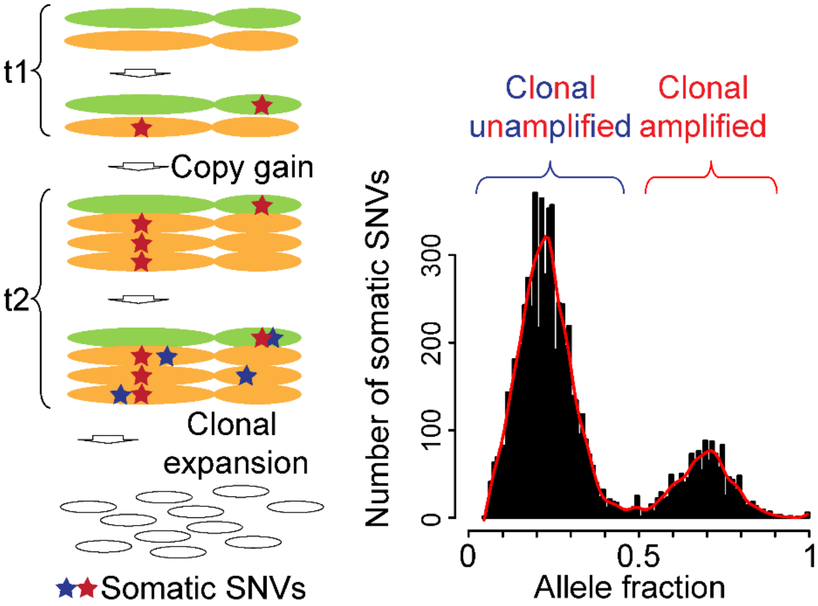

**Figure 15. Mono-allelic two-copy gain regions in crc2683.**

We call two peaks at 0.71 and 0.23 (Figure 15). The peak of 0.71 contains amplified SNVs. One chromosome is accumulating SNVs and contributing to this peak. The peak of 0.23 is a mixture of red and blue SNVs. So we need to subtract the red SNVs to adjust the number. During time period t2, four chromosomes are accumulating SNVs. Therefore, we set d option as “1,4” and m option as “1”.

## 7. Two different copy gain with LOH events

In our paper, we have described two different histories of copy gain with LOH in tumor crc2683. The one on chromosome 2 is a mono-allelic two-copy gain right CN-LOH or three-copy gain after losing one parental copy (Figure 16). If we believe all copy gains occur at the same time, we can use the standard mode:

```
perl beehive.pl -i ./examples/crc2683.4+0 -t 4+0
```

And the output will be:

```
Peaks called: 0.24 0.24 0.94
Valleys called:  0.04 0.66
Number of SNVs in bin1: 462
Number of SNVs in bin2: 1377
Total number of SNVs: 1841
Peak position for amplified alleles: 0.94
Peak position for unamplified alleles: 0.24
Predicted clonality based on event type: 0.89
Timing of copy change: 0.57 (, )
Timing of clonal expansion: 1.00 (, )
```

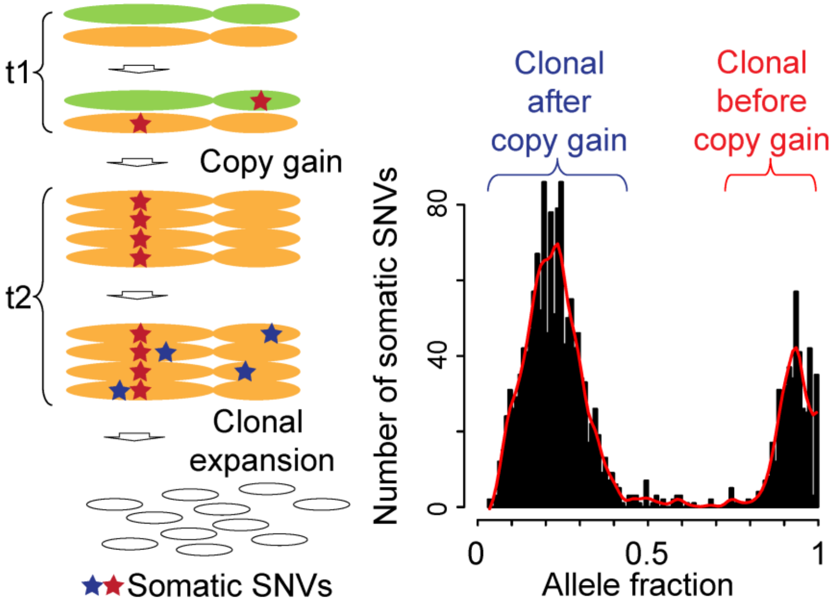

**Figure 16. Copy gain with LOH.**

Two peaks are called as before and after copy gain. The unamplified peak is about a quarter of the amplified peak. Subclonal peak is not distinguishable. The clonality estimate is 0.89 which is a bit far from the tumor purity 0.98, but not too far. All these suggest "4+0" prediction is reasonable. The fact that we do not see many somatic SNVs between these two major peaks suggests that the copy gains occur almost at the same time. Otherwise we shall see a substantial number of SNVs on two or three out of four copies of the chromosome.

On the other hand, the MAF profile of chromosome 9 of crc2683 is completely different

(Figure 17). Based on the profile, three peaks are present at 0.96, 0.74 and 0.24. The 0.96 peak is composed of SNVs present on all four copies of the chromosome. The 0.24 peak contains unamplified SNVs. The 0.74 peak is composed of SNVs present on three out of four copies of the chromosome. Since the chromosome is also characterized as LOH based on germline SNVs, we can conclude the most likely history is a mono-allelic two copy gain after CN-LOH (Figure 17). In this case, the CN-LOH and mono-allelic two-copy gain do not occur at the same time, so we need to use advanced mode to estimate timing. We run:

```
perl beehive.pl -i ./examples/crc2683.4+0.s -p 0.97,0.74,0.24 -d 1,1,4
-m 0,0,1
```

And the output will be:

```
Peaks called: 0.24 0.74 0.96
Valleys called:  0.52 0.86
Number of SNVs in bin1: 17
Number of SNVs in bin2: 209
Total number of SNVs: 1044
Timing of 1st event: 0.04 (, )
Timing of 2nd event: 0.60 (, )
```

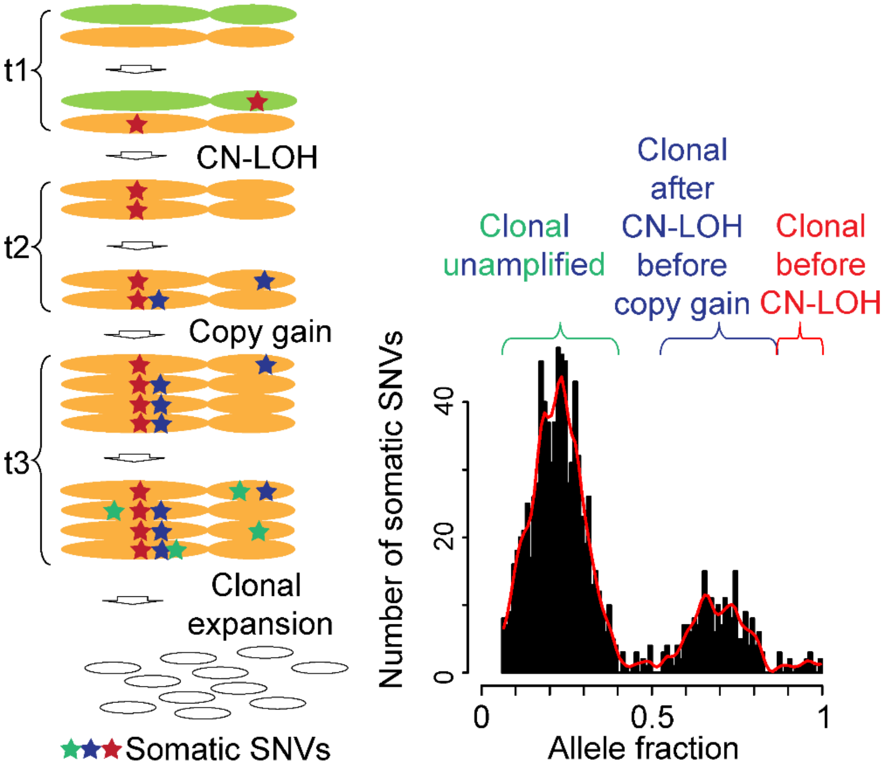

**Figure 17. Mono-allelic two-copy gain after CN-LOH.**

In this case, somatic SNVs are accumulating on one chromosome before CN-LOH and contributing to the right-most peak, on one chromosome after CN-LOH before mono-allelic two-copy gain and contributing to the middle peak, and on four chromosomes after mono-allelic two-copy gain. So the d option is given as “1,1,4”. The left-most peak is composed of both blue and green SNVs. There is one chromosome with blue SNVs not amplified. So those blue SNVs will be present on one out of four copies of the chromosome. We need to subtract the number of blue SNVs when adjusting number of green SNVs occurring during t3. So the m option is given as “0,0,1”. Based on above results, the copy gain of chromosome 2 and the mono-allelic two-copy gain of chromosome 9 occur roughly at the same time (0.57 and 0.60 respectively), while the CN-LOH of chromosome 9 occur quite early (0.04).

## 8. Three-copy gain of X

In tumor crc3913, chromosome X has 4 copies according to copy ratio. This is a male patient with only one X chromosome as wild-type. If we want to use the standard mode, we run:

```
perl beehive.pl -i ./examples/crc3913.4+0 -t 4+0
```

And the output will be:

```
Peaks called: 0.24 0.39 0.88
Valleys called:  0.26 0.67
Number of SNVs in bin1: 284
Number of SNVs in bin2: 256
Total number of SNVs: 629
Peak position for amplified alleles: 0.88
Peak position for unamplified alleles: 0.39
Predicted clonality based on event type: 0.79
Timing of copy change: 0.77 (, )
```

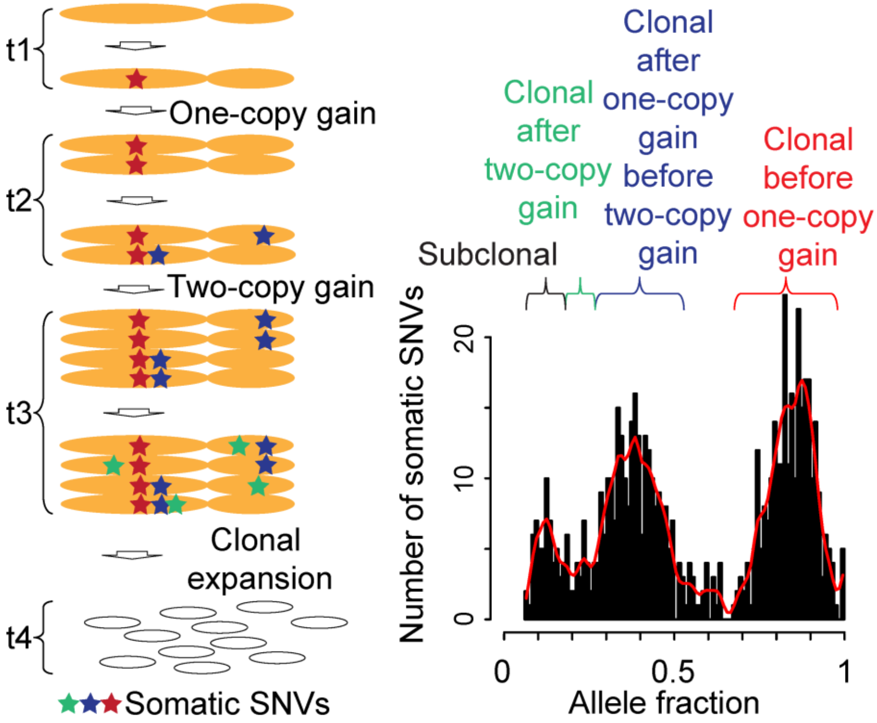

**Figure 18. Bi-allelic two-copy gain after one-copy gain of chromosome X in male.**

```
Timing of clonal expansion: 0.94 (, )
```

Three peaks are called (Figure 18). The middle peak of 0.39 is roughly half of the right-most peak of 0.88. The peak of 0.24 is very small and roughly a quarter of the right-most peak. These suggest the three identified peaks represent SNVs present on four, two and one copy out of four copies of the chromosome. We can infer the history to be bi-allelic two-copy gain after one-copy gain. These copy gains do not occur at the same time, so we should not use the standard mode. In addition, there is a very big peak around 0.12 but not called which is the subclonal peak. In this case, the peak at 0.24 is small, so the subclonal peak is identifiable. In the examples shown in Figure 14,15,16 and 17, the unamplified peak is quite large which makes the subclonal peak indistinguishable. The small peak of 0.24 suggest that time period t3 is short. In advanced mode, we run:

```
perl beehive.pl -i ./examples/crc3913.4+0 -p 0.88,0.44,0.22,0.1 -d
1,2,4,4 -m 0,0,0,0,0,0
```

And the output will be:

```
Peaks called: 0.13 0.24 0.39 0.88
Valleys called: 0.21 0.26 0.67
284, 128, 5, 17.25, 434.25
Number of SNVs in bin1: 284
Number of SNVs in bin2: 256
Number of SNVs in bin3: 20
Total number of SNVs: 629
Timing of 1st event: 0.65 (, )
Timing of 2nd event: 0.95 (, )
Timing of 3rd event: 0.96 (, )
```

In the peak of 0.39, two chromosomes are independently accumulating SNVs and both contributing to the peak. For peak of 0.24 and 0.13, there are four chromosomes accumulating SNVs. No SNVs need to be adjusted. So we set d as “1,2,4,4” and m as “0,0,0,0,0,0”. The 2nd event (bi-allelic two-copy gain) is very close to the 3rd event (clonal expansion).

### 9. Complex copy gains (4+1)

Let’s look at another more complex history. Figure 19 shows chromosome 1q of a third tumor crc3593. It is predicted to have 5 copies. The ratio between two parental alleles is 4:1 based on germline SNVs (not shown). There are three main peaks in MAF profile: 0.68, 0.36 and 0.18. In fact, at around 0.36, there are two peaks, one at 0.36 and the other at 0.3. Usually we do not call them as separate peaks since they are too close to each other. For this particular region, we choose 0.36 as the peak location because it is roughly half of the right-most peak of 0.68 and roughly twice of the left-most peak of 0.18. Given the relationship of these peak locations, we can infer the history to be bi-allelic two-copy gain after one-copy gain. Since the copy gains do not occur at the same time, we will use the advanced mode:

```
perl beehive.pl -i ./examples/crc3593.4+1 -p 0.67,0.36,0.18 -d 1,2,5 -
m 0,1,1
```

And the output will be:

```
Peaks called: 0.18 0.36 0.68
Valleys called:  0.33 0.45
Number of SNVs in bin1: 234
Number of SNVs in bin2: 78
Total number of SNVs: 846
Timing of 1st event: 0.72 (, )
Timing of 2nd event: 0.84 (, )
```

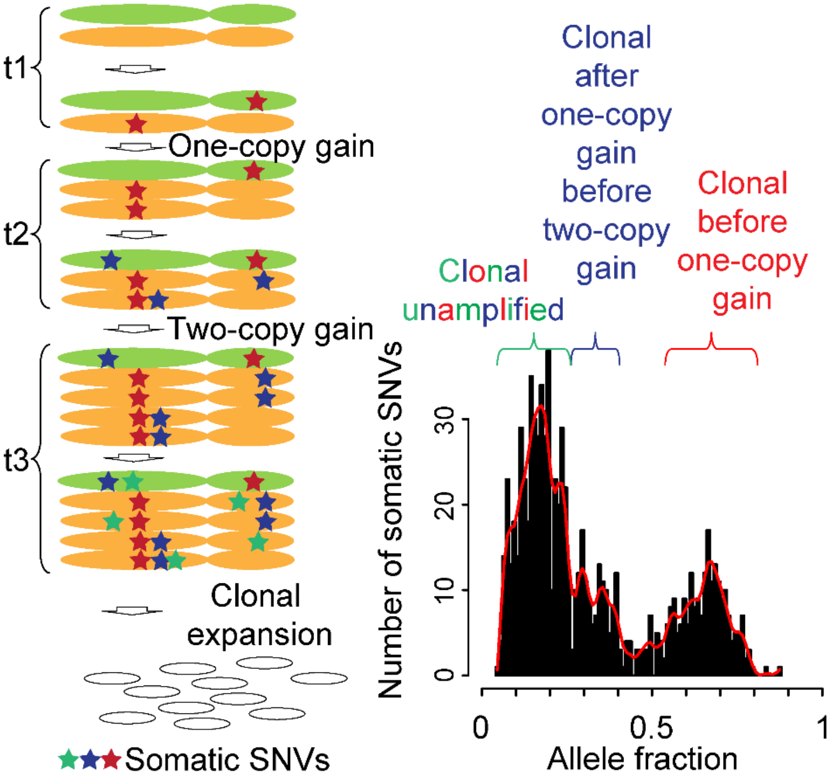

**Figure 19. Bi-allelic two-copy gain after one-copy gain of an autosome.**

The red SNVs are amplified on only one copy of the chromosome; the blue SNVs are amplified on two copies of the chromosome; and the green SNVs are accumulating on five copies of the chromosome. So we set option d as “1,2,5”. The red SNVs and blue SNVs on the green chromosome are not amplified and will be present in the left-most peak. So we need to subtract both of them when adjusting the number of SNVs during time period t3. So we set option m as “0,1,1”.

# V. Reference

Yang L., S. Wang, J. Lee, S. Lee, E. Lee, E. Shinbrot, D.A. Wheeler, R. Kucherlapati, and P.J. Park. An enhanced genetic model of colorectal cancer progression history. In review.

# VI. Contact

Questions and comments are welcome. Before contacting the author for questions, please read the tutorial carefully. There are a lot of contents in this document. Your question may already be answered.

Author:

Lixing Yang, PhD
Ben May Department for Cancer Research, The University of Chicago
Chicago, IL 60637, USA
Email: lixingyang@uchicago.edu or ylixing@gmail.com