Inferring Clonal Structure From SNP Copy Number Data

Kevin R. Coombes and Mark Zucker

July 6, 2017

Contents

We start by loading the package into the current R session.

```
> library(CloneFinder)
```

We introduce the term *compartment* to describe a pure (undiluted, homogeneous) copy number state. For modeling purposes, we assume that there is a fixed number of pure compartments. In particular, we do not model high amplifications in copy number, mainly because they are simply indistinguishable in SNP copy number data.

We consider the following comparments:

- Most segments of the genome appear in two different (i.e., heterozygous) copies. In this state, we expect the true log R ratio (LRR) to equal zero and the true B allele frequency (BAF) to equal one-half.
- Some segments contain two identical (homozygous) copies of the genomic material, either through inheritance (identical by descent) or because of a somatic loss of heterozygosity (LOH). In this case, the true $LRR = 0$ and the true $BAF = 0$.
- Another compartment arises when all cells in the sample being measured have lost one copy of a genomic segment. In this case, the true $LRR = \log(1/2)$ and the true $BAF = 0$.

- Similarly, it is possible for all cells in the sample to acquire a gain of the same genomic segment. In this case, the true $LRR = \log(3/2)$ and the true $BAF = 1/3$.
- A gain of two copies of the same piece of a chromosome has true $LRR = \log(2)$ and true $BAF = 1/4$. (For modeling purpoises, we ignore the tetraploidy case when both parental chromosomes are duplicated, leading to $LRR = \log(2)$ and $BAF = 1/2$.) We also ignore higher copy number gains.
- The case when both copies of a genomic segment is problematic, since the true $LRR = \log(0) = -\infty$ and the true $BAF = 0/0$ is undefined.

Code Example A compartment is modeled in the `CloneFinder` package by the `CompartmentModel` class, which is implemented as:

```
> showClass("CompartmentModel")

Class "CompartmentModel" [package "CloneFinder"]

Slots:

Name:      markers pureCenters      sigma0
Class:     numeric  data.frame     numeric

Known Subclasses: "Tumor"
```

In this preliminary implementation, instead of using actual (LRR, BAF) pairs, we instead simulate and model the data as though it comes from a pair of independent normal distributions. For example,

```
> set.seed(2726642) # for reproducible examples
> nSeg <-  1000     # number of segments supposedly found by CBS
> markers <- round(runif(nSeg, 25, 1000))  # numbers of markers
> # set 'known' centers for the pure compartments
> xy <- data.frame(x = c(0.2, 0.7, 0.8, 0.1, 0.4),
+                  y = c(0.2, 0.3, 0.5, 0.9, 0.7))
> # build the model. sigma0 = std dev at one marker
> baseModel <- CompartmentModel(markers, xy, sigma0=0.25)
> rm(nSeg, xy)
```

Before we can show you how the modeling works, we have to simulate data from a tumor.

```
> wts <- rev(5^(1:5))
> wts <- wts/sum(wts)  # prevalence of differnt compartments
> fracs <- c(5, 3, 1)  # relative frequency of subclones
> # length of 'fracs' is the number of clones
```

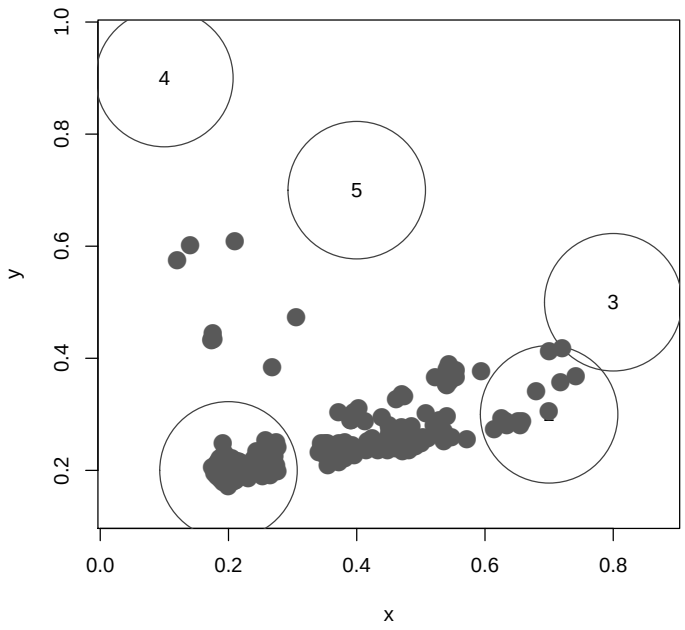

Figure 1: Sized scatter plot of simulated data.

```
> # now simulate a tumor;
> tumor <- Tumor(baseModel, fracs, wts)
> rm(wts, fracs, markers, baseModel)
> class(tumor)

[1] "Tumor"
attr(,"package")
[1] "CloneFinder"
```

Objects of the `Tumor` class are basically multivariate distributions; you have to make another function call to sample/simulate data from them.

```
> simdata <- generateData(tumor)
```

3.2 Segment-Level Modeling

After the SNP copy number data has been segmented (typically by applying something like the circular binary segmentaion algorithm implemented in the `DNAcopy` R package), we model the data as arising from a mixture (in terms of

cells in the biological sample) of the pure compartments. We use the following notation:

- Let K denote the number of pure compartments.
- For $k \in 1, \ldots, K$, let C_k be the statistical distribution modeling the data observed from a single SNP marker in a region consisting of cells from the k^{th} compartment.
- Let S be the number of segments.
- For $s \in 1, \ldots, S$, let M_s be the number of SNP markers contained in the s^{th} segment.
- Becuase the data in the s^{th} segment is obtained by averaging over M_s markers, the observed data even for a pure compartment should arise from a modified distribution $C_{s,k} = C_k\{M_s\}$, which typically involves dividing the standard deviation by $\sqrt{M_s}$.

Now the observed pair of measurements $X = (LRR, BAF)$ on each segment s is modeled by an equation of the form

$$X_s \sim \sum_{k=1}^{K} \varphi_{s,k} C_k\{M_s\},$$

with the obvious constraints that every parameter satisfies $0 \leq \varphi_{s,k} \leq 1$ and

$$\forall s, \sum_{k=1}^{K} \varphi_{s,k} = 1.$$

3.2.1 Preliminary Estimate of Compartment Frequencies (φ)

In spite of having only one observed data point per segment, we can still get an estimate of the vector $\bar{\varphi}_s = (\varphi_{s,1}, \ldots, \varphi_{s,k})$ by using the following Bayesian procedure. First, we use a prior distribution that says that every vector in the simplex defined by the constraints above is equally likely. We then sample potential vectors $\bar{\varphi}$ uniformly from the simplex. (The sampling step is implemented in the function `sampleSimplex`, which implements the method described in Wolfgang Huber's answer to a question on the Cross Validated part of the web site Stack Exchange: `http://stats.stackexchange.com/questions/14059/generate-uniformly-distributed-weights-that-sum-to-unity`.

Next, we compute the likelihood ($Prob(X_s|\bar{\varphi}_s, C_k, M_s)$) of the observed data at each of the sampled vectors $\bar{\varphi}$, and record the vector with the maximum likelihood. Since the prior and the sampling scheme are uniform, this is the same as the vector with the maximum posterior probability.

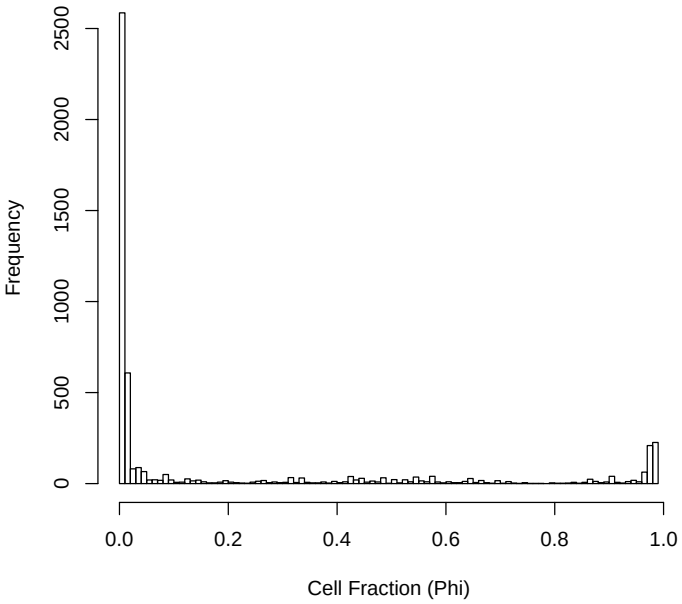

Figure 2: Histogram og components of phi-vectors from first pass at modeling the simulated data.

Code Example This step of the algorithm is implemented in the function `PrefitCloneModel`.

```
> pcm <- PrefitCloneModel(simdata, tumor)
> summary(pcm)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       1       1       1       1       1
```

3.2.2 Refining the Vectors $\bar{\varphi}$

The precision of the estimates of $\bar{\varphi}_s$ arising from the previous step depend on how many vectors we initially sampled from the simplex. In principle, one could improve the precision either by sampling much more deeply or by implementing a full-blown Markov Chain Monte Carlo (MCMC) algorithm. In the present circumstance, however, we can be much more efficient by borrowing strength across segments.

The idea is that the biological sample as a whole is composed of a fairly small number of clones. (Specifically, we do not believe that SNP copy number

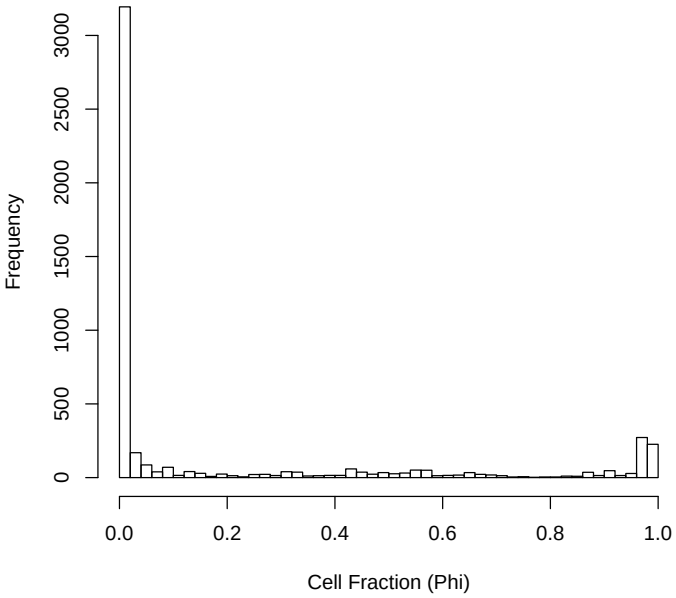

Figure 3: Histogram of components of phi-vectors from first pass at modeling the simulated data.

data is capable of distinguishing more than about five subclones.) We assume that there are N subclones with frequencies ψ_i for $i \in 1, \ldots, N$, subject to the constraint that $\sum_{i=1}^{N} \psi = 1$. Then, for each segment, the true fraction of cells in each pure compartment must be given by a sum of a subset of the ψ_i values, and so there is a finite (and reasonably small) number of true vectors $\bar{\varphi}_s$.

OBSOLETE! So, we use the posterior distribution of the "maximum likelihood" $\bar{\varphi}$ vactors as a new prior distribution. We sample vectors from this distribution (therefore ensuring that we sample more vectors in a neighborhood of the most common vectors from the first pass), and repeat the computation of likelihoods and the idenitification for each segment of the maximum likelihood vector. In principle, this step could be repeated more than once; in practice, we have not yet seen any reason to do so. This step is implemented in the function `updatePhiVectors`.

Code Example

3.3 Clone-Level Modeling, When The Number of Clones is Known

Let $\Phi = [\varphi_{s,k}]$ denote the matrix of all segment-wise compartment frequencies. In modeling terms, we can express the idea that the data all arise from a mixture of N subclones by writing

$$\Phi = \sum_{i=1}^{N} \psi_i Z_i,$$

where the ψ_i are as above and each Z_i is an indicator matrix. That is, each Z_i is a matrix with S rows (one per segment) and K columns (one per pure compartment). Moreover, every entry in Z_i is equal either to 0 or to 1, and every row must contain exactly one entry equal to 1. In other words, each $Z - i$ specifies which pure compartment is contained in the i^{th} subclone for each segment.

3.3.1 Initial Estimate of ψ

As noted above, the true value of each $\phi_{s,k}$ should be a linear combination of the ψ_i. So, the distribution of the components of $\bar{\varphi}$ shown in hist.second should contain at least some information about the values of ψ_i. So, we shoudl start by trying to either cluster those data or to identify the peaks in the histogram.

3.3.2 Refined Estimate of ψ

3.4 Inferring the Number of Clones

It would be really nice to know how to do this....

4 Conclusions