# Derivation of the distance between two parent/infant cases accounting for underreporting

Thibaut Jombart, Pierre Nouvellet, Anne Cori, ...

April 7, 2016

## Contents

# 1 Theoretical framework

## 1.1 Notations

In all the following, subscripts $t$, $s$ and $g$ are used to refer to time, geographical space, and genetic space respectively. Pdf stands for probability density function and pmf for probability mass function. For both, we use the same general notation $f$.

**Data**

We denote $i$ and $j$ two cases, so that $i$ is the closest ancestry of $j$ in the database. $i$ can be the infector of $j$, or the infector of its infector, etc... We call $\kappa_{i,j}$ the unoberved number of missing intermediates between $i$ and $j$, so that $\kappa_{i,j} = 0$ if $i$ is the infector of $j$, $\kappa_{i,j} = 1$ if $i$ is the infector of the infector of $j$, etc. We denote $x_i$ and $x_j$ information on the "location" of cases $i$ and $j$. These can be temporal information (times of swabbing, of times of symtpoms etc.), spatial information (geographical location of the cases) or genetic information (sequences of the cases). We denote $d(i,j)$ a measure of distance between these two locations.

**Parameters**

We denote $\pi$ the reporting probability, and $\phi$ the probability density function of the distance (in time, geographical space or genetic space) between an infector and an infected individual. This can be the serial interval, the spatial kernel, or the genetic signature.

## 1.2 Distribution of the distance between two cases accounting for underreporting

We are interested in computing the probability density function $f(d(i,j)|\pi,\phi)$, which can be decomposed according to the unobserved number of missing generations between $i$ and $j$ as follows:

$$f(d(i,j)|\pi,\phi) = \sum_{k=0}^{+\infty} f(d(i,j)|\kappa_{i,j} = k,\phi) f(\kappa_{i,j} = k|\pi)$$

The first factor in the sum is $f(d(i,j)|\kappa_{i,j} = k,\phi) = \phi^{(k+1)}(d(i,j))$, where $\phi^{(k+1)}$ denotes the convolution of $\phi$ with itself, $k+1$ times ($\phi^{(1)} = \phi$). Note this is assuming that the distance $d$ is additive, so that if $i$ infected $l$ who infected $j$, then $d(i,j) = d_{i,l} + d_{l,j}$. This is the case for time, and we assume it is the same for genetic distance. For spatial distance we perform a slightly more complicated reasoning (see below).

The second factor in the sum is the probability of $k$ intermediate cases having been unobserved, and the $k+1$ case (going back in time, that is $i$) having been observed. This is given by the geometric distribution $f(\kappa_{i,j} = k|\pi) = \pi(1-\pi)^k$.

Therefore:

$$f(d(i,j)|\pi,\phi) = \sum_{k=0}^{+\infty} \phi^{(k+1)}(d(i,j))\,\pi(1-\pi)^k$$

## 1.3 Special case: $\phi$ is the pdf of a Gamma distribution (typical for serial interval distribution)

In this section we consider the special case where $\phi$ is the pdf of a Gamma distribution with shape $\alpha$ and scale $\beta$: $\phi(x) = f_\gamma(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$. The sum of $k+1$ independent variables with same Gamma distribution with parameters $(\alpha,\beta)$ also has a Gamma distribution, with parameters $((k+1)\alpha,\beta)$. Therefore:

$$f(d(i,j)|\pi,\alpha,\beta) = \sum_{k=0}^{+\infty} \frac{1}{\Gamma((k+1)\alpha)\beta^{(k+1)\alpha}} d(i,j)^{(k+1)\alpha-1} e^{-d(i,j)/\beta}\pi(1-\pi)^k$$

### 1.4 Special case: $\phi$ is the pmf of a negative binomial distribution (typical for genetic signature distribution)

The genetic signature, the expected number of SNPs between two cases, will typically depend on the serial interval (amount of time available to evolve) and the mutation rate. Using a Gamma distribution for the serial interval (with shape $\alpha$ and scale $\beta$), and a Poisson distribution for the number of mutations in a given time interval $dt$ (with mean $\mu * dt$) leads to a negative binomial distribution with parameters $\left(\alpha, \frac{\beta\mu}{\beta\mu+1}\right)$ for the number of mutations between two cases. Indeed, if $i$ infected $j$ and $d_g\left(i,j\right)$ is the genetic ditstance between $i$ and $j$, $\phi_t$ the (Gamma) pdf of the serial interval, and $\mu$ the mutation rate, then the pmf of the genetic distance between $i$ and $j$ is:

$$
\begin{aligned}
\phi_g\left(d_g\left(i,j\right)\right) = f\left(d_g\left(i,j\right)|\kappa_{i,j}=0,\phi_t,\mu\right) &= \int_{t=0}^{+\infty} f\left(d_g\left(i,j\right)|\kappa_{i,j}=0,\mu,d_t\left(i,j\right)=t\right)\phi_t\left(t\right)dt \\
&= \int_{t=0}^{+\infty} \frac{\left(\mu t\right)^{d_g(i,j)}e^{-\mu t}}{d_g\left(i,j\right)!}\frac{1}{\Gamma(\alpha)\beta^\alpha}t^{\alpha-1}e^{-t/\beta}dt \\
&= \frac{\Gamma(\alpha+d_g\left(i,j\right))\mu^{d_g(i,j)}}{\left(\mu+1/\beta\right)^{\alpha+d_g(i,j)}d_g\left(i,j\right)!\Gamma(\alpha)\beta^\alpha}\int_{t=0}^{+\infty}\frac{t^{\alpha+d_g(i,j)-1}e^{-t(\mu+1/\beta)}}{\Gamma(\alpha+d_g\left(i,j\right))\left(\frac{1}{\mu+1/\beta}\right)^{\alpha+d_g(i,j)}}dt \\
&= \frac{\Gamma(\alpha+d_g\left(i,j\right))}{d_g\left(i,j\right)!\Gamma(\alpha)}\frac{\mu^{d_g(i,j)}}{\left(\mu+1/\beta\right)^{\alpha+d_g(i,j)}\beta^\alpha} \\
&= \binom{\alpha+d_g\left(i,j\right)-1}{d_g\left(i,j\right)}\left(\frac{\beta\mu}{\beta\mu+1}\right)^{d}_{g}(i,j)\left(1-\frac{\beta\mu}{\beta\mu+1}\right)^{\alpha}
\end{aligned}
$$

The sum of $k+1$ independent variables with same Gamma distribution with parameters $(r,p)$ also has a Gamma distribution, with parameters $\left(\left(k+1\right)r,p\right)$. Therefore:

$$
\begin{aligned}
f\left(d_g\left(i,j\right)|\pi,\alpha,\beta,\mu\right) &= \sum_{k=0}^{+\infty}\phi_g^{(k+1)}\left(d_g\left(i,j\right)\right)\pi\left(1-\pi\right)^k \\
&= \sum_{k=0}^{+\infty}\binom{\left(k+1\right)\alpha+d_g\left(i,j\right)-1}{d_g\left(i,j\right)}\left(\frac{\beta\mu}{\beta\mu+1}\right)^{d_g(i,j)}\left(\frac{1}{\beta\mu+1}\right)^{(k+1)\alpha}\pi\left(1-\pi\right)^k
\end{aligned}
$$

### 1.5 Special case: $\phi$ is the pdf of a Rayleigh distribution (typical for spatial kernel distribution)

We assume that the geographical location of an individual $i$ is given by coordinates $(x_i, y_i)$ in an orthonormed system. We assume that the coordinates of an inividual $j$ infected by $i$ are so that $x_j - x_i$ and $y_j - y_i$ are independent and identically distributed according to a centered normal distribution $\mathcal{N}\left(0,\sigma^2\right)$:

$$
f\left(x_j - x_i|\kappa_{i,j}=0,\sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_j-x_i)^2}{2\sigma^2}}
$$

Now if $i$ infected an unobserved case $l$ who infected $j$, then $x_j - x_i = x_j - x_l + x_l - x_i$ so the distribution of $x_j - x_i$ is that of the sum of two independent identical normal variables. This is true as well for $y_j - y_i$, and easily extends to more than 1 unobserved intermediate case. Now, the sum of $k+1$ independent univariate Normally distributed variables with same mean $\nu$ and same variance $\sigma^2$ is a univariate Normally distributed variable with mean $\left(k+1\right)\nu$ and variance $\left(k+1\right)\sigma^2$ (see here http://www.tina-vision.net/docs/memos/2003-003.pdf for a proof for $k+1=2$). Therefore,

$$
f\left(x_j - x_i|\kappa_{i,j}=k,\sigma\right) = \frac{1}{\sqrt{2\pi\left(k+1\right)\sigma^2}}e^{-\frac{(x_j-x_i)^2}{2(k+1)\sigma^2}}
$$

Using the change of variable technique, one can show that the pdf of $(x_j - x_i)^2$, conditionnal on $\kappa_{i,j} = k$ and $\sigma$ is $\frac{1}{\sqrt{2\pi(k+1)\sigma^2}} e^{-\frac{s}{2(k+1)\sigma^2}} \frac{1}{2\sqrt{s}}$. Therefore, conditionnal on $\kappa_{i,j} = k$ and $\sigma$, $(x_j - x_i)^2$ follows a Gamma distribution with shape $\frac{1}{2}$ and scale $2(k+1)\sigma^2$.

The same reasoning holds for $(y_j - y_i)^2$.

The square of the Euclidian distance between individuals $i$ and $j$ can be computed as $d_s(i,j)^2 = (x_j - x_i)^2 + (y_j - y_i)^2$. Conditionnal on $\kappa_{i,j} = k$ and $\sigma$, this is the sum of the squares of 2 independent Gamma distributed variables with same shape $\frac{1}{2}$ and scale $2(k+1)\sigma^2$. Therefore, conditionnal on $\kappa_{i,j} = k$ and $\sigma$, $d_s(i,j)^2$ is also Gamma distributed with shape $2 \times \frac{1}{2} = 1$ and scale $2(k+1)\sigma^2$, that is it is Exponentially distributed with rate $\frac{1}{2(k+1)\sigma^2}$.

This means the Euclidian distance between individuals $i$ and $j$, $d_s(i,j)$ follows a Rayleigh distribution with scale $\sigma\sqrt{k+1}$ (see general proof here http://www.math.wm.edu/ leemis/chart/UDR/PDFs/ExponentialRayleigh.pdf).

Finally,

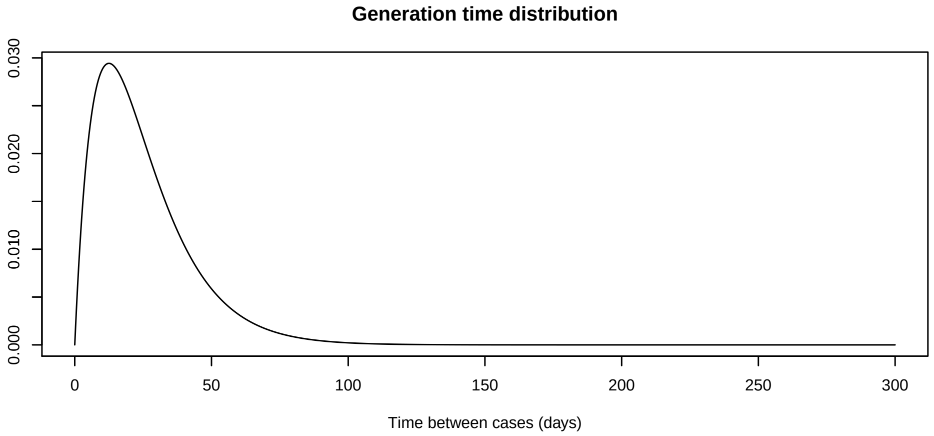
$$\begin{aligned} f\left(d_s(i,j) \,|\, \pi, \sigma\right) &= \sum_{k=0}^{+\infty} \phi_s^{(k+1)}\left(d_s(i,j)\right) \pi \left(1-\pi\right)^k \\ &= \sum_{k=0}^{+\infty} \frac{d_s(i,j)}{(k+1)\sigma^2} e^{-\frac{d_s(i,j)^2}{2(k+1)\sigma^2}} \pi \left(1-\pi\right)^k \end{aligned}$$

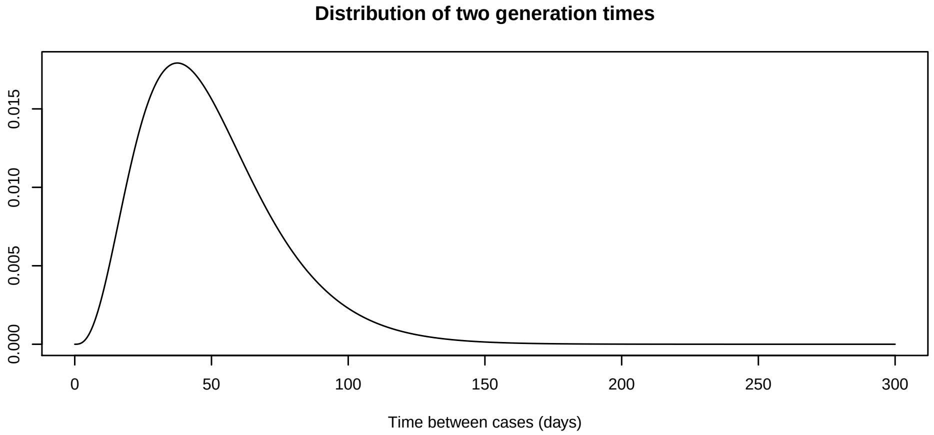
# 2 Implementation of the distributions

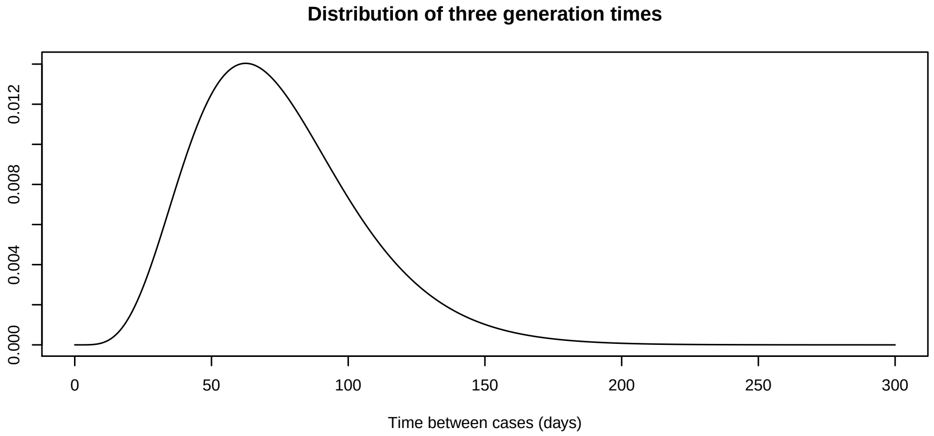
## 2.1 Serial interval gamma distributed

We assume a Gamma distributed generation time with shape 2 and scale 25/2 (measured in days), and we can compute the distribution of the time between a case and its infector, a case and its infector's infector, etc...

```
# assuming a given Gamma distribution for the generation time #
shape_t <- 2
scale_t <- 25/2 # parameters taken from main vignette
f.gentime <- function(x) dgamma(x, shape=shape_t, scale=scale_t)
f.gentime.convol <- function(x,k) dgamma(x, shape=shape_t*(k+1), scale=scale_t)
dt <- 0.1
times <- seq(0,300,dt)
par(mfrow=c(3,1))
plot(times,f.gentime(times), main="Generation time distribution",
     xlab="Time between cases (days)", ylab="",type="l")
# checking this is the same as the time convolved with no intermediate cases missed:
# lines(times,f.gentime.convol(times,k=0),col="red")
plot(times,f.gentime.convol(times,k=1), main="Distribution of two generation times",
     xlab="Time between cases (days)", ylab="",type="l")
plot(times,f.gentime.convol(times,k=2), main="Distribution of three generation times",
     xlab="Time between cases (days)", ylab="",type="l")
```



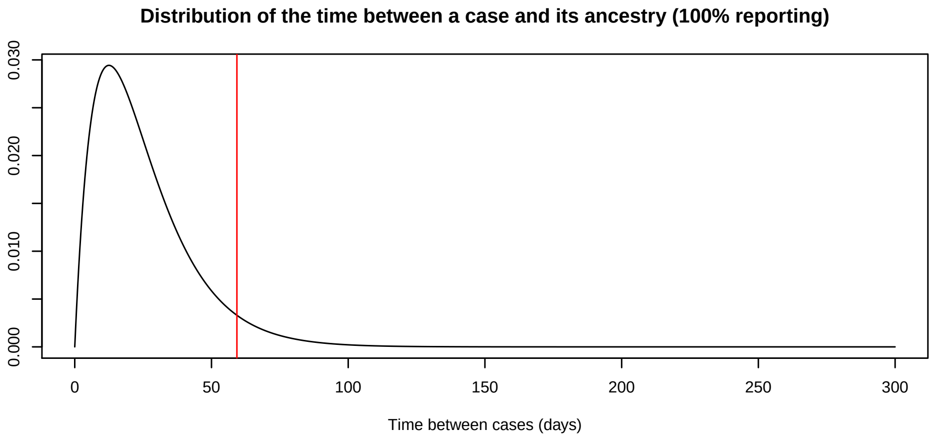




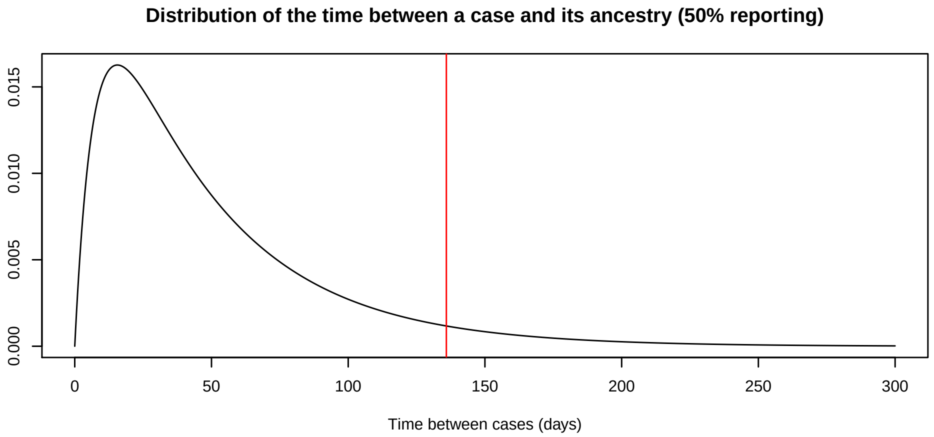
Now assume the reporting probability is not 100%. We can readily compute the distribution of time between a case and its closest ancestry in the database, accounting for underreporting. The red bars in the figures below show the 95% quantiles of the distributions.

```
f.gentime.with.under.reporting <- function(x,pi) sapply(x, function(e)
  dconvol(e,type="t",reporting_prob=pi, shape_t=shape_t, scale_t=scale_t) )
### TO DO: change code so dconvol deals with vectors as well as unique numeric objects?

par(mfrow=c(3,1))
```

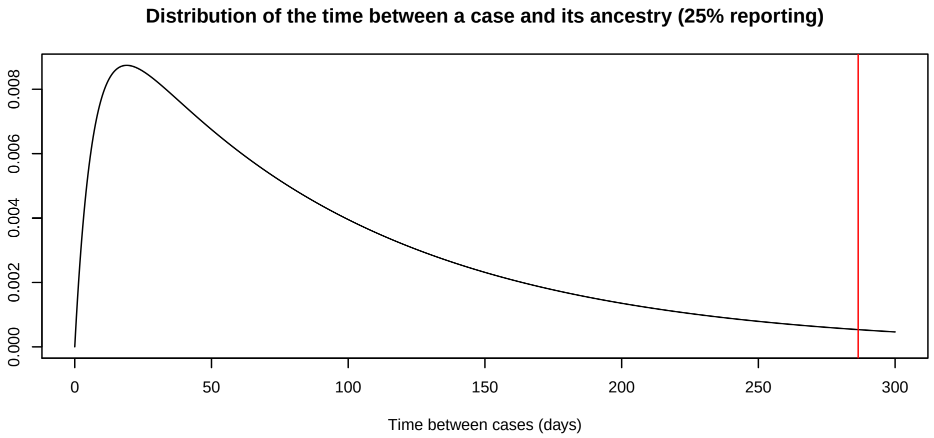
```r
# 100% reporting:
y <- f.gentime(times)
plot(times,y,
     main="Distribution of the time between a case and its ancestry (100% reporting)",
     xlab="Time between cases (days)", ylab="",type="l")
Q95 <- times[min(which(cumsum(y)*dt>0.95))]
abline(v=Q95,col="red")
# checking the simple serial interval is the same as that with 100% reporting:
# lines(times,f.gentime.with.under.reporting(times,pi=1),col="red")

# 50% reporting:
y <- f.gentime.with.under.reporting(times,pi=0.50)
plot(times,f.gentime.with.under.reporting(times,pi=0.5),
     main="Distribution of the time between a case and its ancestry (50% reporting)",
     xlab="Time between cases (days)", ylab="",type="l")
Q95 <- times[min(which(cumsum(y)*dt>0.95))]
abline(v=Q95,col="red")

# 25% reporting:
y <- f.gentime.with.under.reporting(times,pi=0.25)
plot(times,y,
     main="Distribution of the time between a case and its ancestry (25% reporting)",
     xlab="Time between cases (days)", ylab="",type="l")
Q95 <- times[min(which(cumsum(y)*dt>0.95))]
abline(v=Q95,col="red")
```



Distribution of the time between a case and its ancestry (100% reporting)



Distribution of the time between a case and its ancestry (50% reporting)



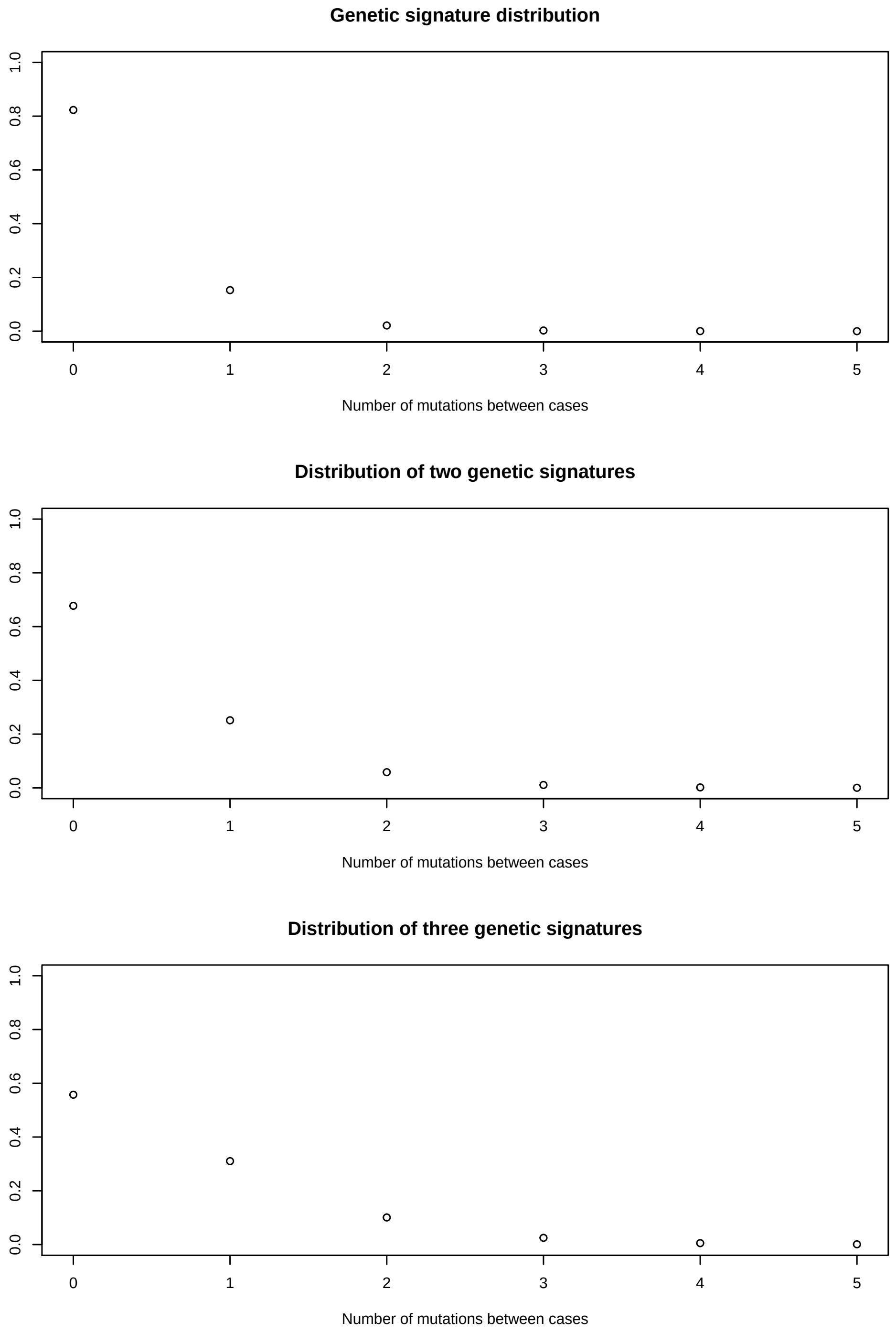
Distribution of the time between a case and its ancestry (25% reporting)

## 2.2 Genetic signature NegBin distributed

We assume the same generation time as above, and a mutation rate of 5.9x10$^{-4}$ substitutions/site/year (see main vignette), so that:

```
## rate per day and sequence
mu.day.whole <- (5.9e-4 * 5063 / 365) # 5063 is the sequence length
                                      # (ncol(dna) in main vignette)
```

```r
f.genetic <- function(x) dnbinom(x, size=shape_t,
                        prob=1-scale_t*mu.day.whole/(scale_t*mu.day.whole+1))
f.genetic.convol <- function(x,k) dnbinom(x, size=shape_t*(k+1),
                        prob=1-scale_t*mu.day.whole/(scale_t*mu.day.whole+1))
mutations <- 0:5
par(mfrow=c(3,1))
plot(mutations,f.genetic(mutations), main="Genetic signature distribution",
     xlab="Number of mutations between cases", ylab="",ylim=c(0,1))
# checking this is the same as the time convolved with no intermediate cases missed:
# points(mutations,f.genetic.convol(mutations,k=0),col="red")
plot(mutations,f.genetic.convol(mutations,k=1), main="Distribution of two genetic signatures",
     xlab="Number of mutations between cases", ylab="",ylim=c(0,1))
plot(mutations,f.genetic.convol(mutations,k=2), main="Distribution of three genetic signatures",
     xlab="Number of mutations between cases", ylab="",ylim=c(0,1))
```


Genetic signature distribution

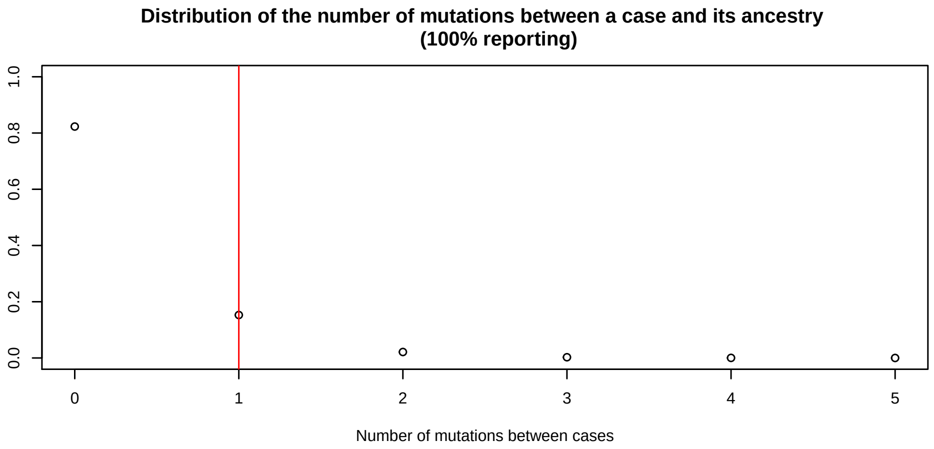
Distribution of two genetic signatures

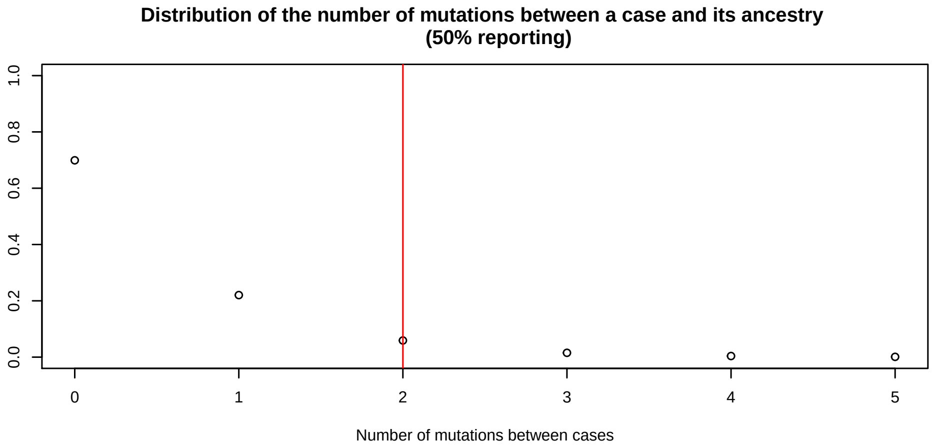
Distribution of three genetic signatures

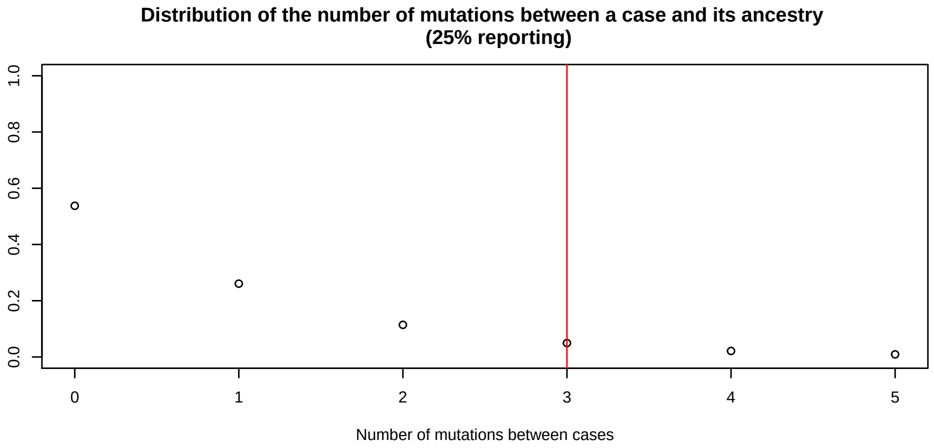
Now assume the reporting probability is not 100%. We can readily compute the distribution of the number of mutations between a case and its closest ancestry in the database, accounting for underreporting. The red bars in the figures below show the 95% quantiles of the distributions.

```r
f.genetic.with.under.reporting <- function(x,pi) sapply(x, function(e)
  dconvol(e,type="g",reporting_prob=pi, shape_t=shape_t, scale_t=scale_t, mu=mu.day.whole) )
### TO DO: change code so dconvol deals with vectors as well as unique numeric objects?

par(mfrow=c(3,1))
```

```
# 100% reporting:
y <- f.genetic(mutations)
plot(mutations,y,
     main="Distribution of the number of mutations between a case and its ancestry
     (100% reporting)",
     xlab="Number of mutations between cases", ylab="",ylim=c(0,1))
Q95 <- mutations[min(which(cumsum(y)>0.95))]
abline(v=Q95,col="red")
# checking the simple serial interval is the same as that with 100% reporting:
# points(mutations,f.genetic.with.under.reporting(mutations,pi=1),col="red")

# 50% reporting:
y <- f.genetic.with.under.reporting(mutations,pi=0.50)
plot(mutations,y,
     main="Distribution of the number of mutations between a case and its ancestry
     (50% reporting)",
     xlab="Number of mutations between cases", ylab="",ylim=c(0,1))
Q95 <- mutations[min(which(cumsum(y)>0.95))]
abline(v=Q95,col="red")

# 25% reporting:
y <- f.genetic.with.under.reporting(mutations,pi=0.25)
plot(mutations,y,
     main="Distribution of the number of mutations between a case and its ancestry
     (25% reporting)",
     xlab="Number of mutations between cases", ylab="",ylim=c(0,1))
Q95 <- mutations[min(which(cumsum(y)>0.95))]
abline(v=Q95,col="red")
```





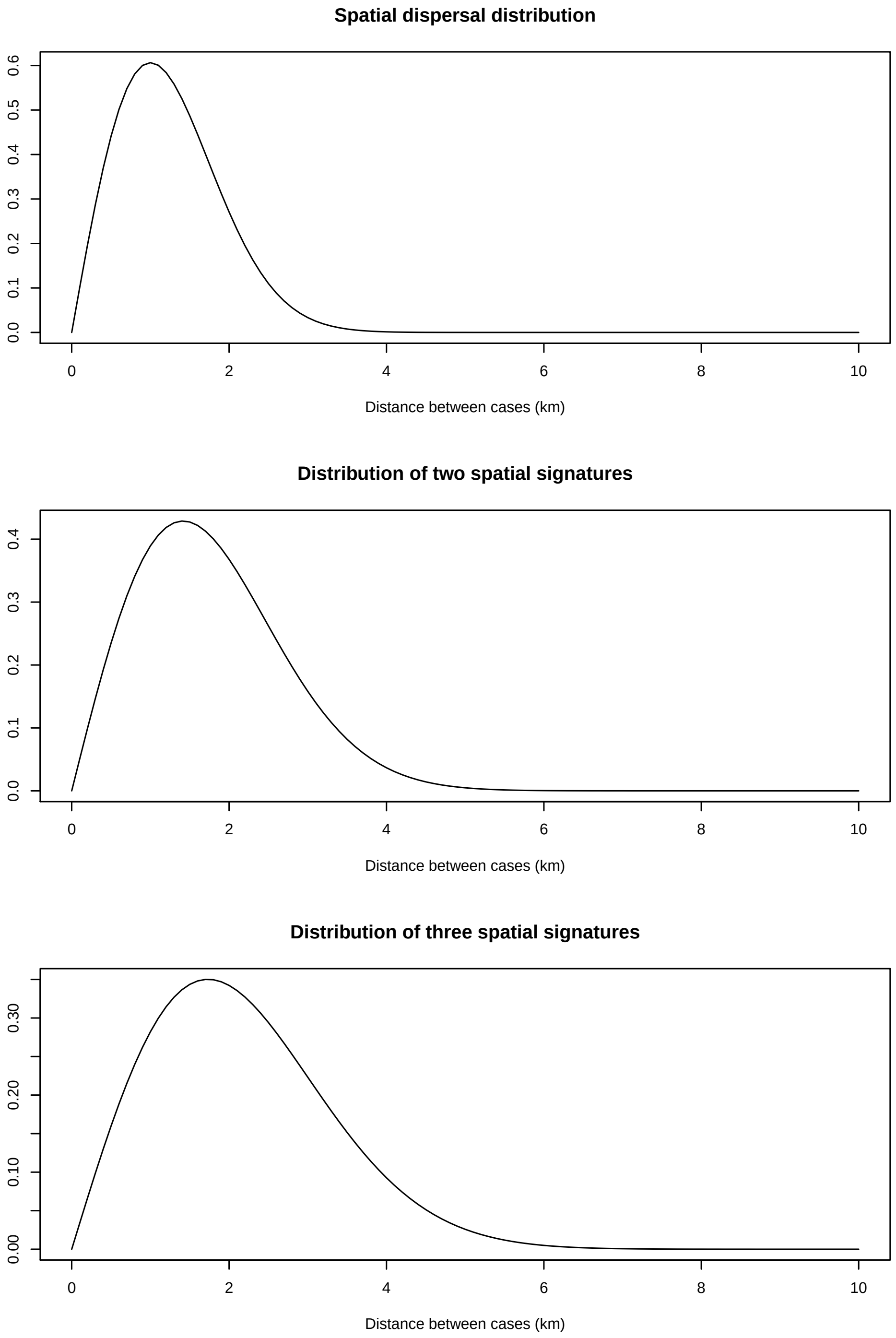


## 2.3 Spatial kernel Rayleigh distributed

We assume a standard deviation for the spatial diffusion in both directions of 1km.

```
## standard deviation for the spatial diffusion in both directions
sigma_s <- 1

f.spatial <- function(x) drayleigh(x,scale=sigma_s)
```

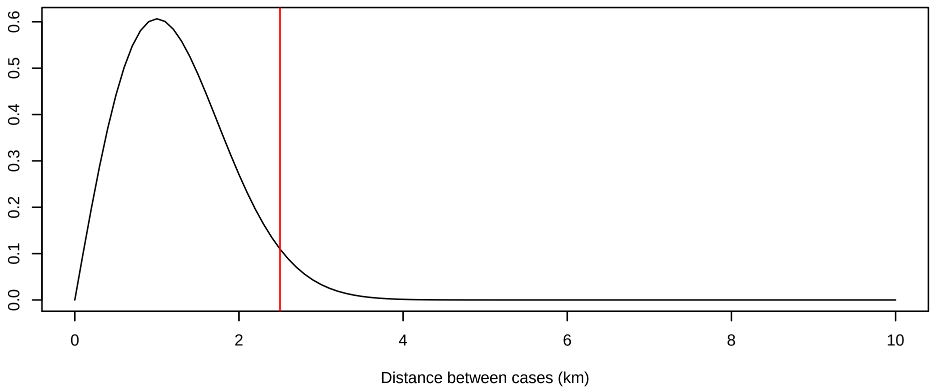
```
f.spatial.convol <- function(x,k) drayleigh(x,sigma_s*sqrt(k+1))

ds <- 0.1
distances <- seq(0,10,ds)
par(mfrow=c(3,1))
plot(distances,f.spatial(distances), main="Spatial dispersal distribution",
     xlab="Distance between cases (km)", ylab="",type="l")
# checking this is the same as the time convolved with no intermediate cases missed:
# lines(distances,f.spatial.convol(distances,k=0),col="red")
plot(distances,f.spatial.convol(distances,k=1), main="Distribution of two spatial signatures",
     xlab="Distance between cases (km)", ylab="",type="l")
plot(distances,f.spatial.convol(distances,k=2), main="Distribution of three spatial signatures",
     xlab="Distance between cases (km)", ylab="",type="l")
```



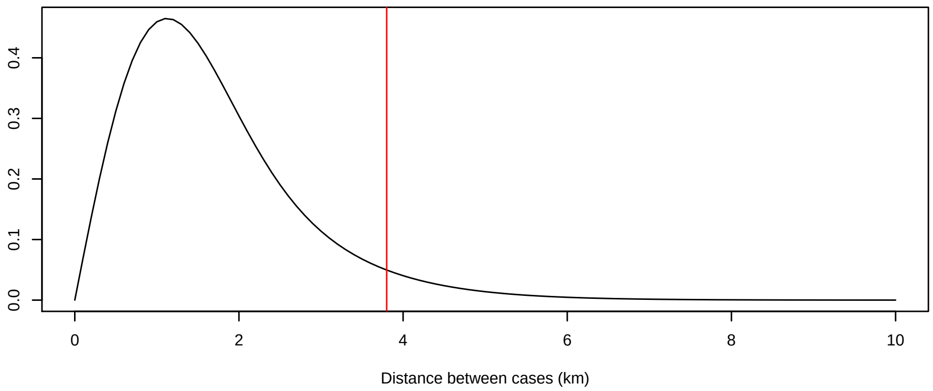
Now assume the reporting probability is not 100%. We can readily compute the distribution of the distance between a case and its closest ancestry in the database, accounting for underreporting. The red bars in the figures below show the 95% quantiles of the distributions.

```
f.spatial.with.under.reporting <- function(x,pi) sapply(x, function(e)
  dconvol(e,type="s",reporting_prob=pi, sigma_s=sigma_s) )
### TO DO: change code so dconvol deals with vectors as well as unique numeric objects?

par(mfrow=c(3,1))
```

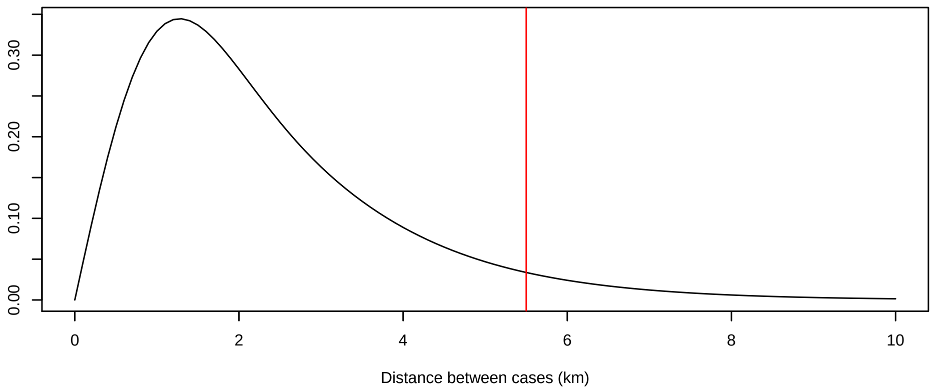
```
# 100% reporting:
y <- f.spatial(distances)
plot(distances,y,
     main="Distribution of the distance between a case and its ancestry (100% reporting)",
     xlab="Distance between cases (km)", ylab="",type="l")
Q95 <- distances[min(which(cumsum(y)*dt>0.95))]
abline(v=Q95,col="red")
# checking the simple serial interval is the same as that with 100% reporting:
# lines(distances,f.gentime.with.under.reporting(times,pi=1),col="red")

# 50% reporting:
y <- f.spatial.with.under.reporting(distances,pi=0.50)
plot(distances,y,
     main="Distribution of the distance between a case and its ancestry (50% reporting)",
     xlab="Distance between cases (km)", ylab="",type="l")
Q95 <- distances[min(which(cumsum(y)*dt>0.95))]
abline(v=Q95,col="red")

# 25% reporting:
y <- f.spatial.with.under.reporting(distances,pi=0.25)
plot(distances,y,
     main="Distribution of the distance between a case and its ancestry (25% reporting)",
     xlab="Distance between cases (km)", ylab="",type="l")
Q95 <- distances[min(which(cumsum(y)*dt>0.95))]
abline(v=Q95,col="red")
```



Distribution of the distance between a case and its ancestry (100% reporting)



Distribution of the distance between a case and its ancestry (50% reporting)



Distribution of the distance between a case and its ancestry (25% reporting)