new_report

Siwei Zhang

Our data include 7003 observations with each individual correponds an intensity value, of which is the proportion that belongs to the interval of (0,1). For example, if A and B represent the alleles of a specific variant. SNP genotypes are determined by comparing A and B intensities. AA means A fluorescence intensity is stronger than B, AB means intensities are similar, BB means B fluorescence intensity is stronger than A. A scanner is used to measure the fluorescence intensity of hybridized A and B probes for each SNP on the array, these data are referred as the raw intensities of the A and B allels, SNP genotypes are determined by comparing A and B intensities.

From the intensity-age plots of three mutations, we estimate our data belong to a mixture of beta regression with two components, of which the first component's intensities is constant over age across individuals, and the second component's intensities varies over age across individuals.

Algorithm

Estimating the finite mixture models using the EM algorithm, the missing data for mixture models is the information to which component an observation belongs to.

The latent variable of the information to which component an observation belongs to is

$$z_{ik} = \begin{cases} 1, & \text{if y comes from component k} \\ 0, & \text{otherwise} \end{cases}$$

The vector of latent variable is $z_i = (z_{i1}...z_{ik})$.

The prior distribution for z_i is $\pi = (\pi_1...\pi_k)$.

Since $y \in (0, 1)$, we estimate it is beta-distributed, of which the density function of the beta distribution is

$$f_k(y|\alpha_k, \beta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} y^{\alpha_k - 1}(1-y)^{\beta_k - 1}$$

Under a mixture model of beta distributions, the unknown parameters is $\theta = (\alpha_1, \beta_1, ..., \alpha_k, \beta_k)$.

The finite mixture models with K components

$$h(y_i|x_i,) = \sum_{k=1}^{K} \pi_k f(y_i|x_i, \alpha_k, \beta_k)$$

The likelihood function is

$$L(\theta, \pi, z) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f_k(y_i|\alpha_k, \beta_k)]$$

The loglikelihood function is

$$l(\theta, \pi, z) = \sum_{i=1}^{n} \sum_{k=1}^{K} [log \pi_k f_k(y_i|\alpha_k, \beta_k)]$$

EM algorithm

The M-step consists of maximizing the "complete-data"" log-likelihood where the missing component memberships are replaced by the current posterior probabilities(given the group assignment, maximize the loglikelihood with respect to the parameters and obtaine the maximizer of alpha and beta); The posterior probabilities of an observation to be from each component given the current parameter estimates are determined in the E-step(given the parameters estimates from the M step, compute the latent group variable)

The number of K is 2. The EM algorithm can be started by either initializing the algorithm with a set of initial parameters and then E step, or by starting with a set of initial weights and then M step. The initial parameters or weights can be chosen randomly. (betamix function in betareg package: default is to begin with an integer vector with the random assignment into k clusters) Given the random assignments into k clusters, we know the $z_i = (z_{i1}...z_{ik})$, so the estimated prior probability $\hat{\pi_k} = \frac{\sum_{i=1}^{n} z_{ik}}{n}$, start with this set of initial weights

M step:

Maximize the log-likelihood for each component separately by using the posterior probabilities as weights and obtain the estimates of α_k and β_k

$$max \sum_{i=1}^{n} \sum_{k=1}^{K} P(k|\hat{x}, y, \theta) log f(y_i|x_i, \alpha_k, \beta_k)$$

and maximize the log-likelihood for the prior probability by using the posterior probabilities as weights and obtain the estimates of prior probability π_k

$$max \sum_{i=1}^{n} \sum_{k=1}^{K} P(k|\hat{x}, y, \theta) log(\hat{\pi_k})$$

then can derive the estimated prior probability used in the next iteration is $\hat{\pi_k} = 1/n \sum P(k|\hat{x}, y, \theta)$ of which the posterior class probabilities for each individual is

$$P(k|\hat{x}, y, \theta) = \frac{\hat{\pi_k} f(y|\hat{\alpha_k}, \hat{\beta_k})}{\sum_k \hat{\pi_k} f(y|\hat{\alpha_k}, \hat{\beta_k})}$$

At first iteration, use the estimated prior probabilities as weights, then as for the later iterations, can use the posterior probability as weights

E step:

Given the current parameters in the (i-1)th iteration: $\theta = (\alpha_1, \beta_1, ..., \alpha_k, \beta_k)$ and the prior distribution $\hat{\pi_k}$. Compute the estimated posterior probabilities

$$P(k|\hat{x}, y, \theta) = \frac{\hat{\pi_k} f(y|\hat{\alpha_k}, \hat{\beta_k})}{\sum_k \hat{\pi_k} f(y|\hat{\alpha_k}, \hat{\beta_k})}$$

Repeat M and E step until the change in the value of log-likelihood is negligible.

specific to beta regression>>>betareg package

Let $y_1, \ldots, y_n$ be a random sample, when $\mu = \alpha/(\alpha+\beta), \phi = \alpha + \beta\$$, then the density function of beta distribution is

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

The mixture model with K components is

$$h(y|x,z,c,\theta) = \sum_{k=1}^{K} \pi_k f(y|g_1^{-1}(x^T\beta_k), g_2^{-1}(z^T\gamma_k))$$

The beta regression model is defined as the mean submodel and the precision submodel. The latent groups can be assumed to differ in their mean and in their precision.

The mean submodel is

$$g_1(\mu_i) = \eta_i = x_i^T\beta$$

and the link function $g_1()$ can be logit, probit, complementary log-log, log-log, and cauchy;

The precision submodel is

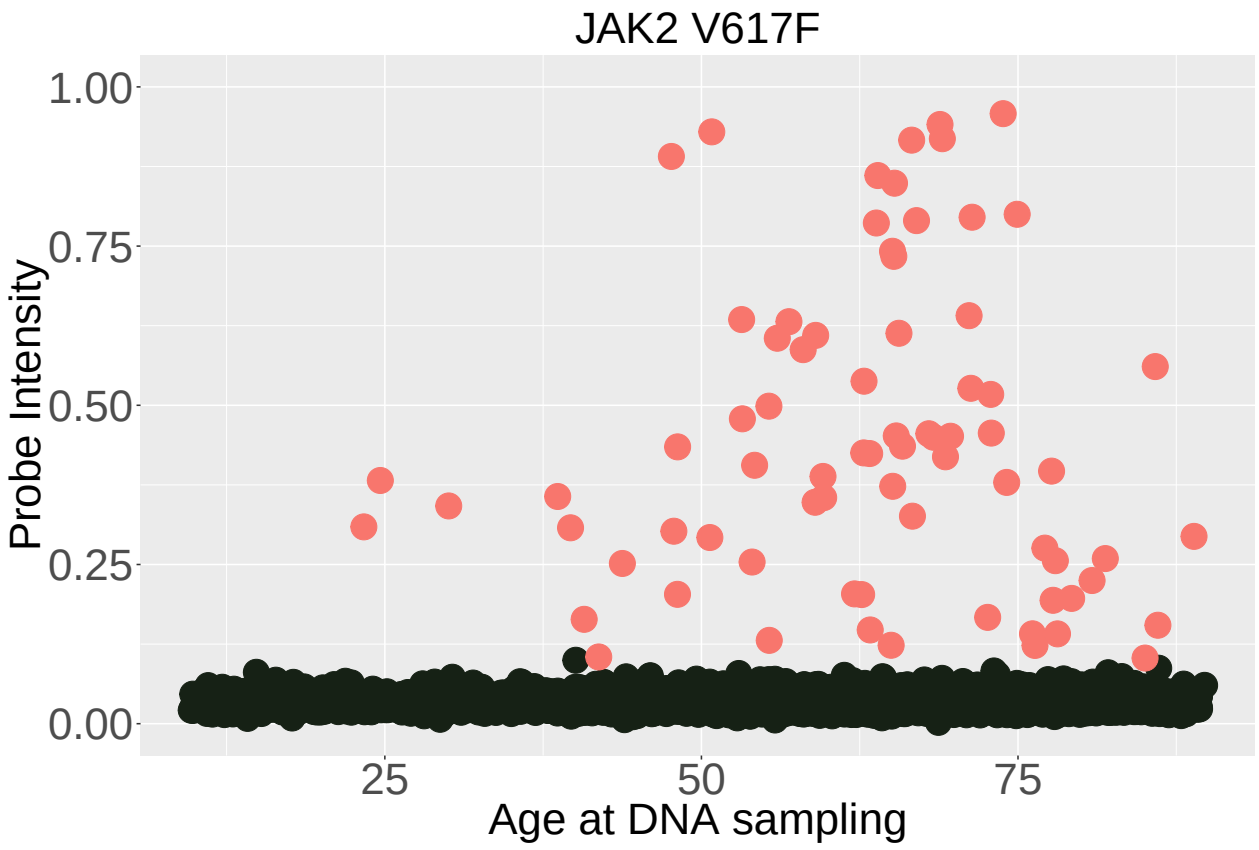
$$g_2(\phi_i) = \zeta_i = z_i^T\gamma$$

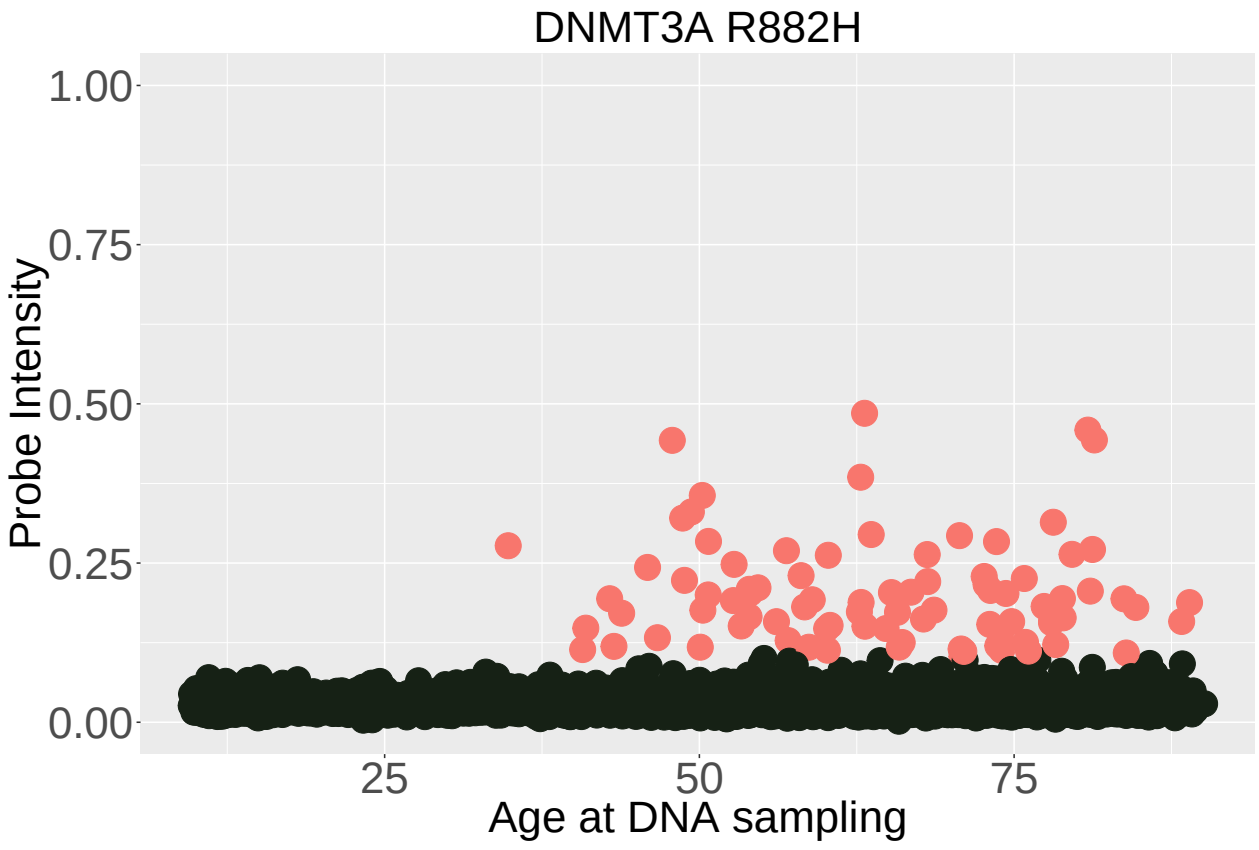
where x_{i} and z_{i} are p and q dimensional vectors of covariates, the coefficient vectors of β and γ are estimated by ML, the inference is based on the CLT with its associated asymptotic tests(likelihood ratio, Wald, score)

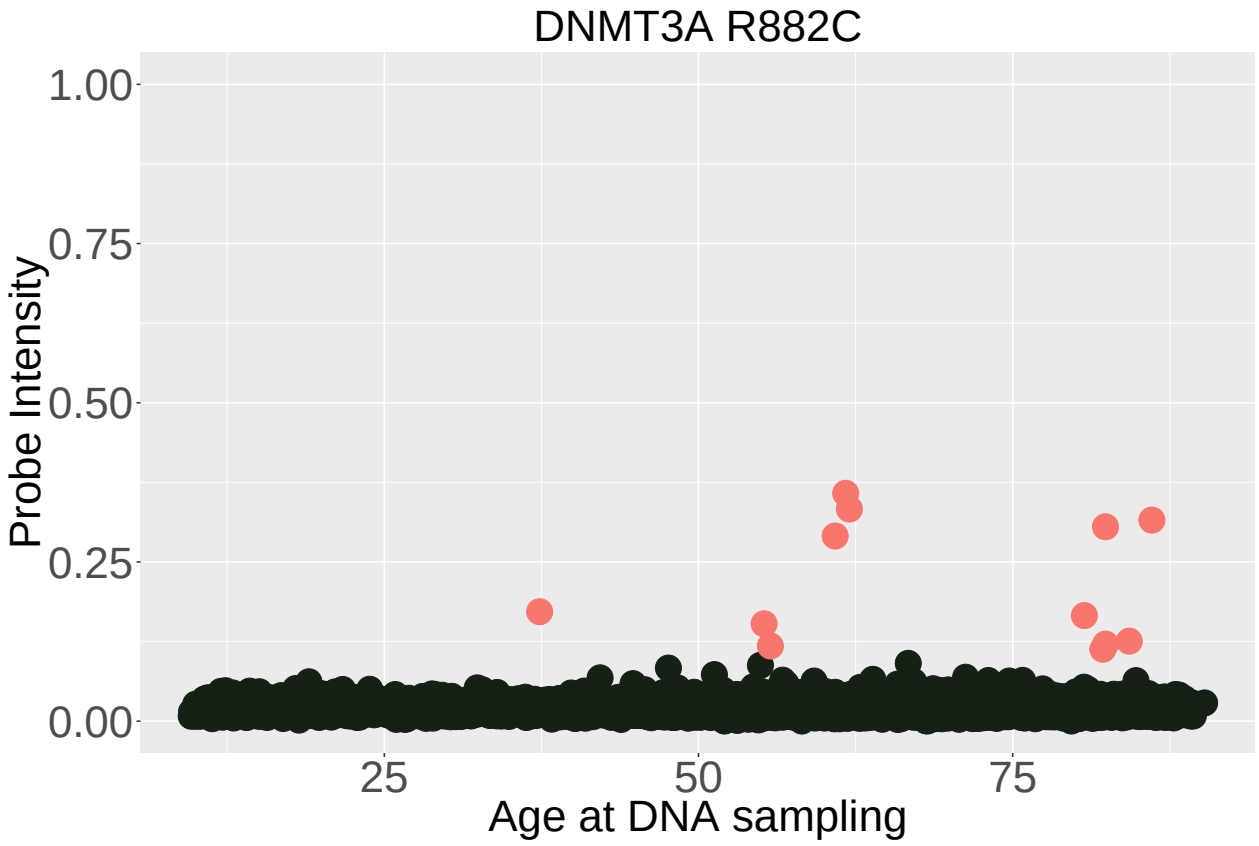
Concomitant variable: The component weights(prior probability or posterior probability mentioned in EM) are assumed to be determined from a vector of covariates c by the concomitant variable model. which is a multinomial logit model.

$$\pi(k;c,\delta) = \frac{exp(c^T\delta_k)}{\sum_{u=1}^{K} exp(c^T\delta_u)}$$

Below are three figures reflect the intensity levels over age across individuals.


JAK2 V617F
1.00
0.75
0.50
0.25
0.00
Probe Intensity
25
50
75
Age at DNA sampling


DNMT3A R882H
1.00
0.75
0.50
0.25
0.00
Probe Intensity
25
50
75
Age at DNA sampling



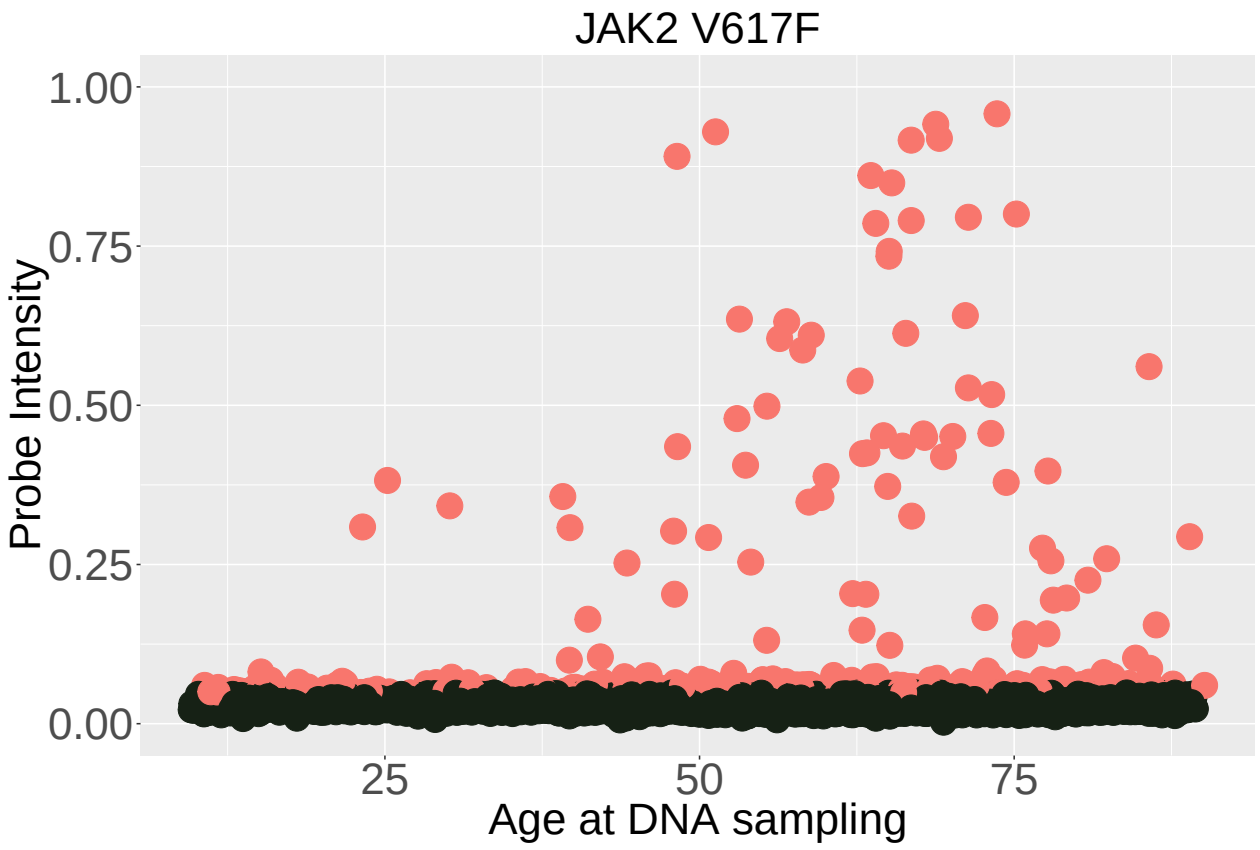
Firstly, check the age distribution of this population, we can see that most people are over 50. It seems bimodal distribution for age. Then check the distribution of the intensity. We can see that nearly all intensities are in the interval of (0,0.1] no matter which age group.

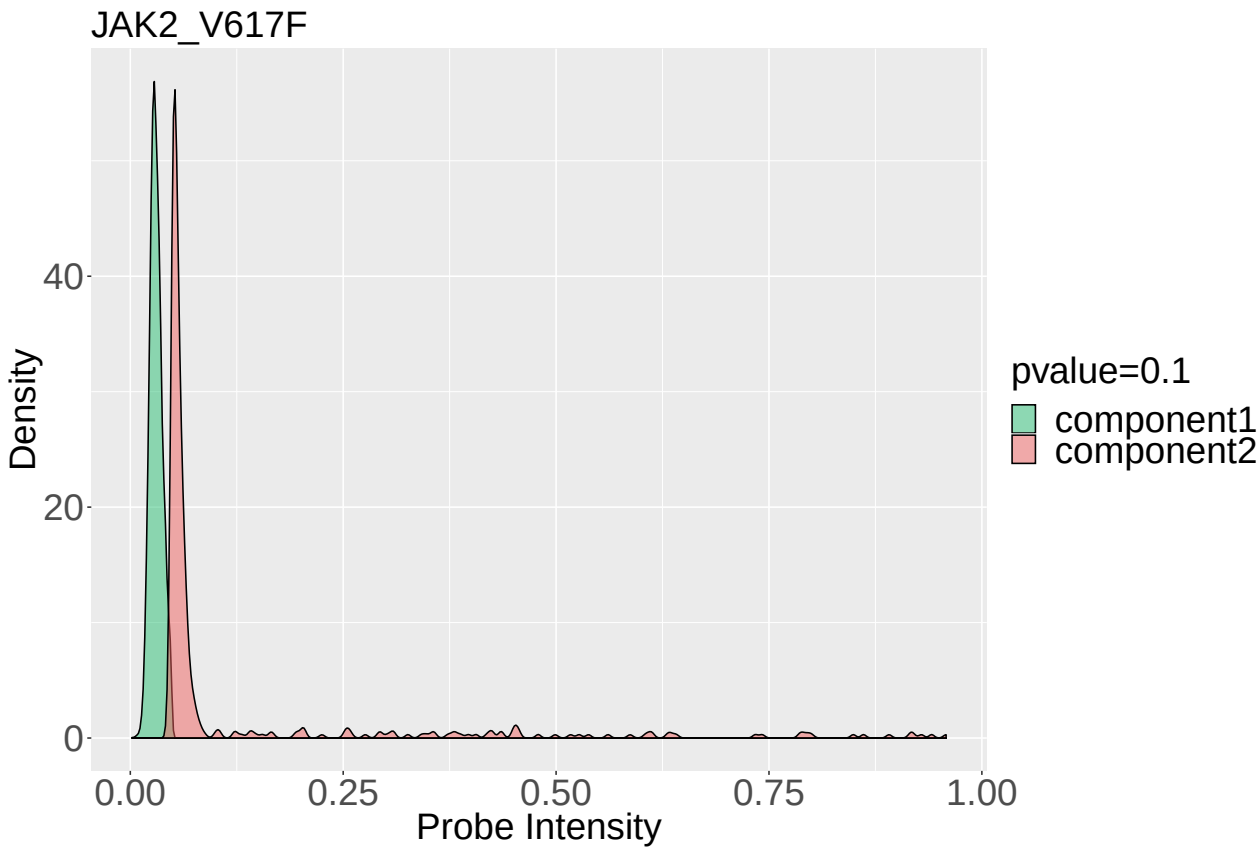
Given a pre-defined cutoff of age, separate individuals into two groups firstly based on the cut-off. Find the parameters of beta distribution for the people aged<=40. Apply this beta distribution on the whole population. If we set the pvalue as 0.1,0.05,0.01,0.005,0.001, the thresholds of the intensity is 0.04895787, 0.05486497, 0.0670443, 0.07186142 and 0.08239699. Different visualizations are below. More orange means more likely this people belong to the second component. We can see an apparent change compared to the Option1.

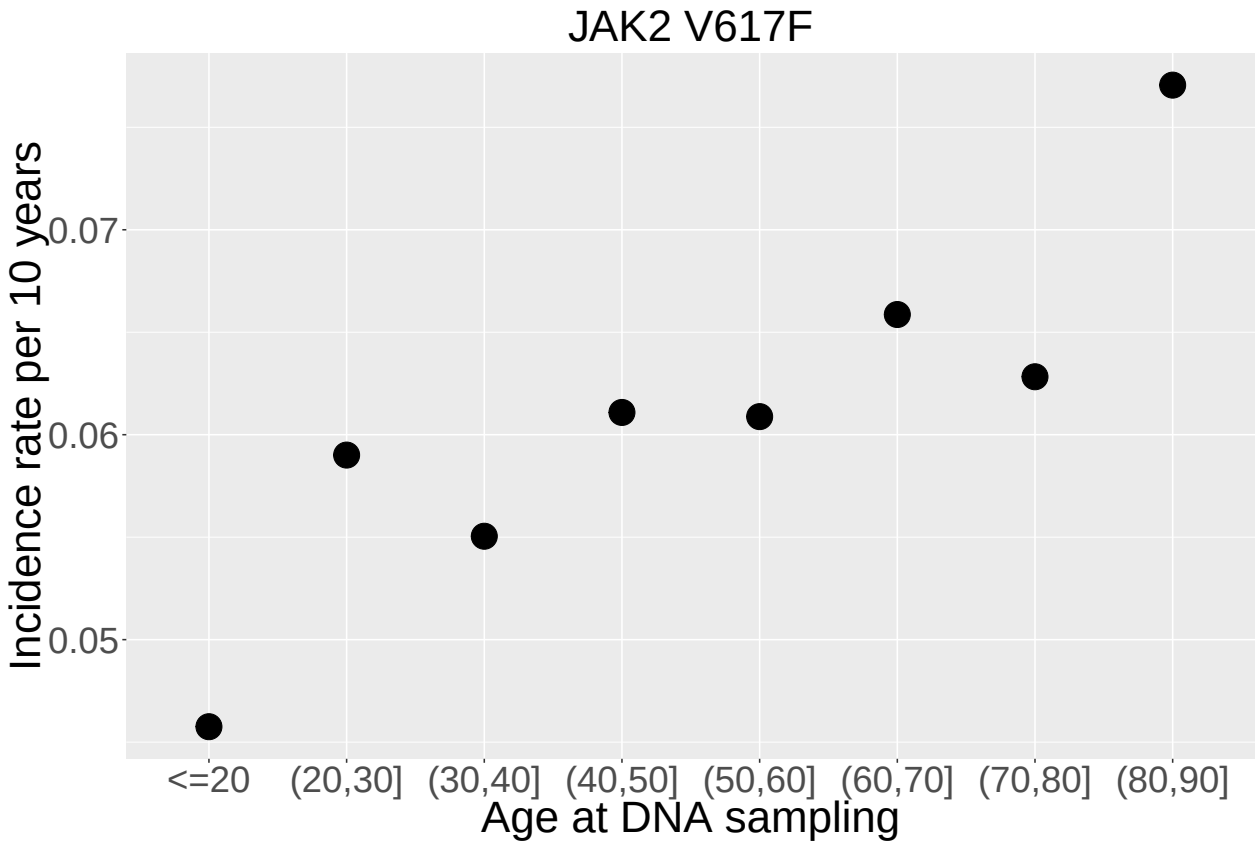
As for the group <=40, y(instensity), examine the distribution of intensity

Fit the beta distribution for the group of people aged <=40

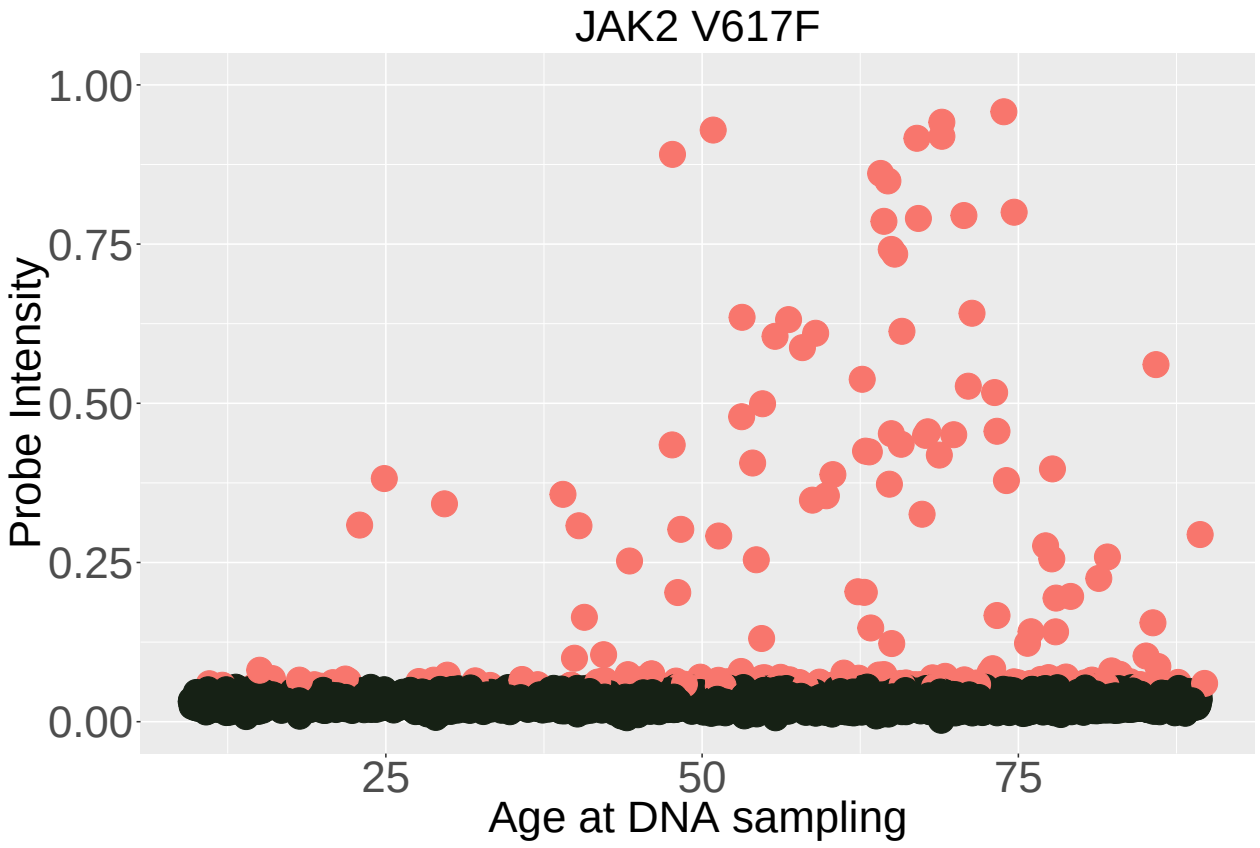
When p-value is 0.1

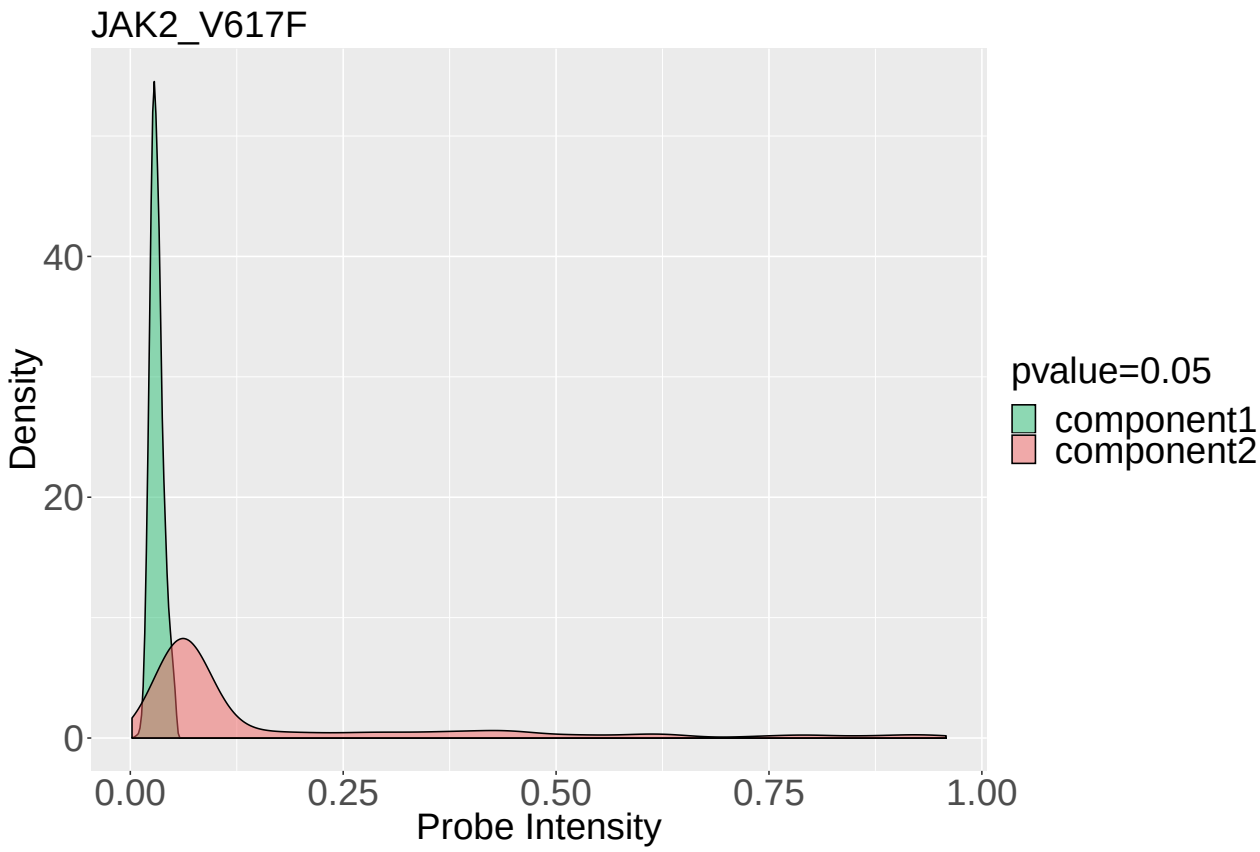

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

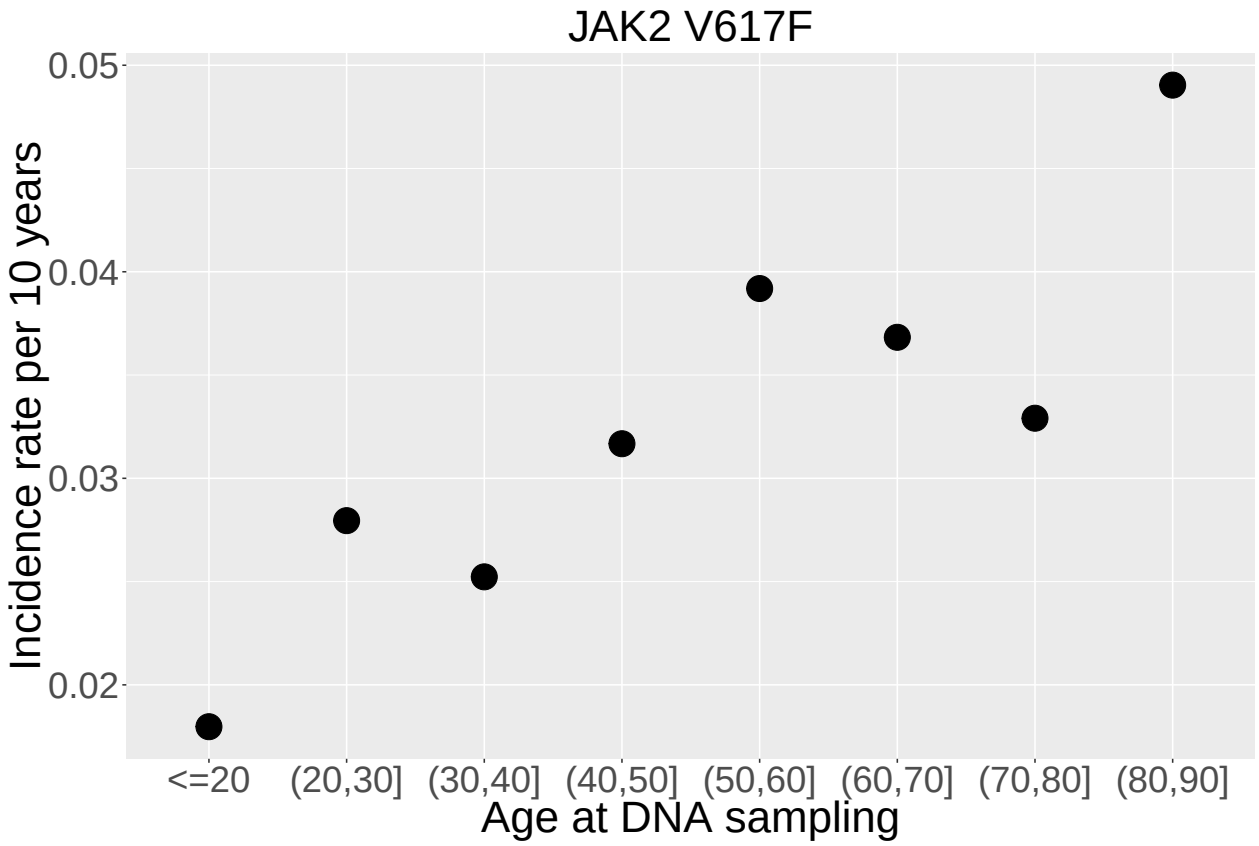

JAK2_V617F
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.1
component1
component2



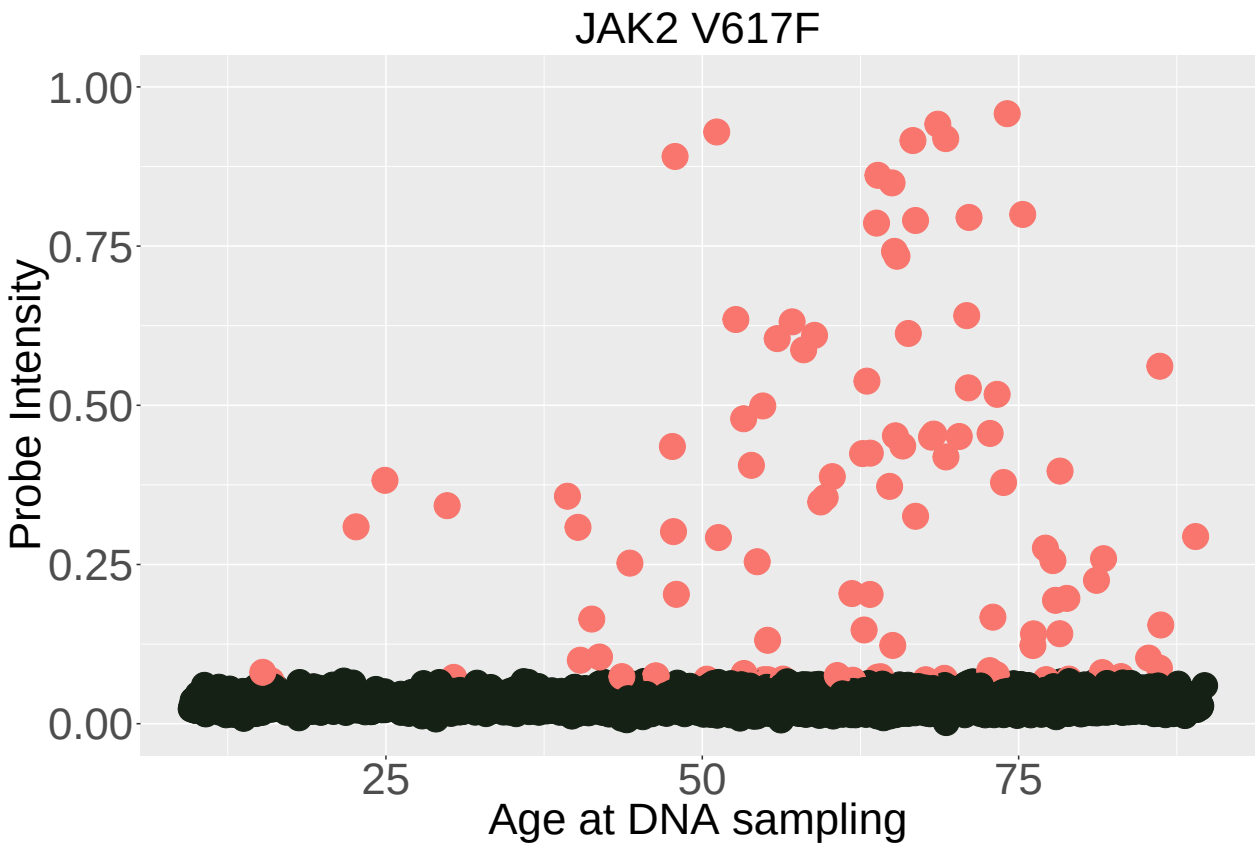
When p-value is 0.05

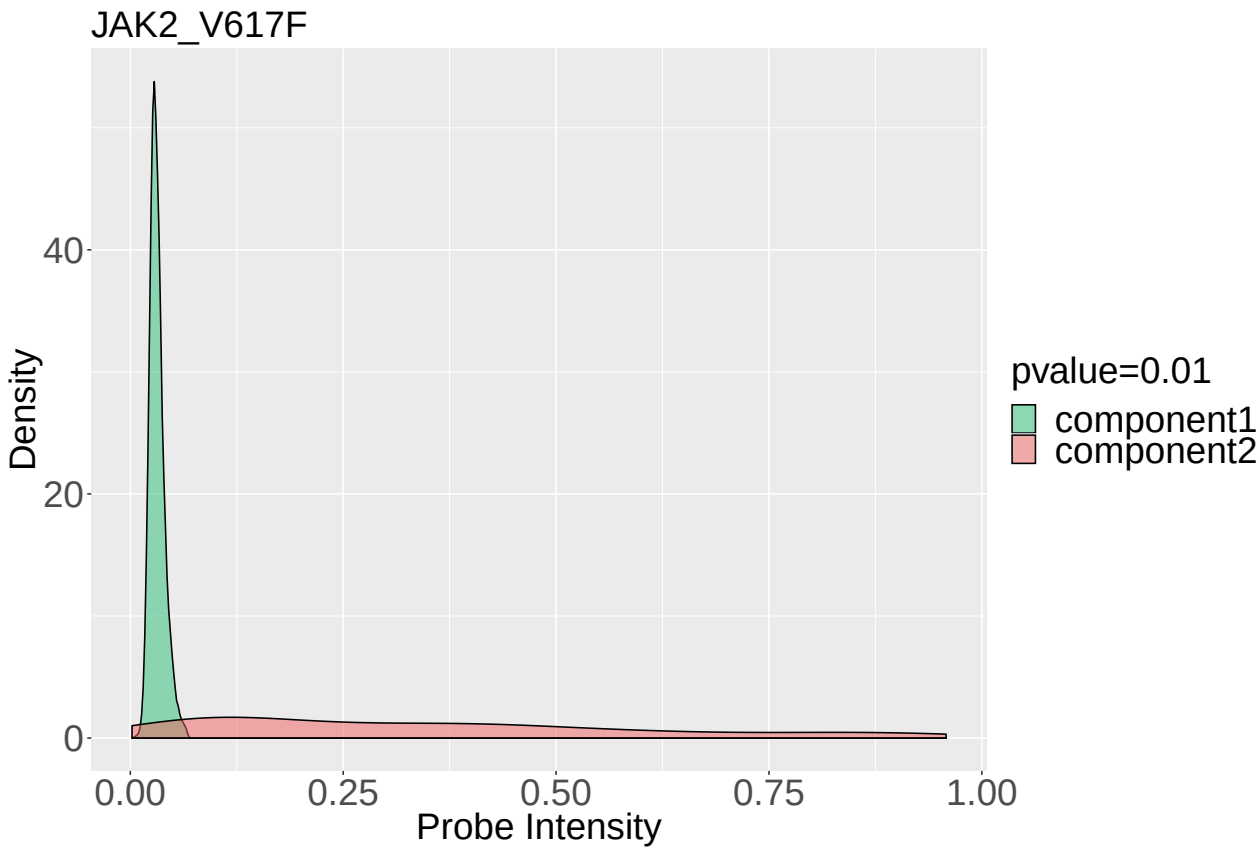

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

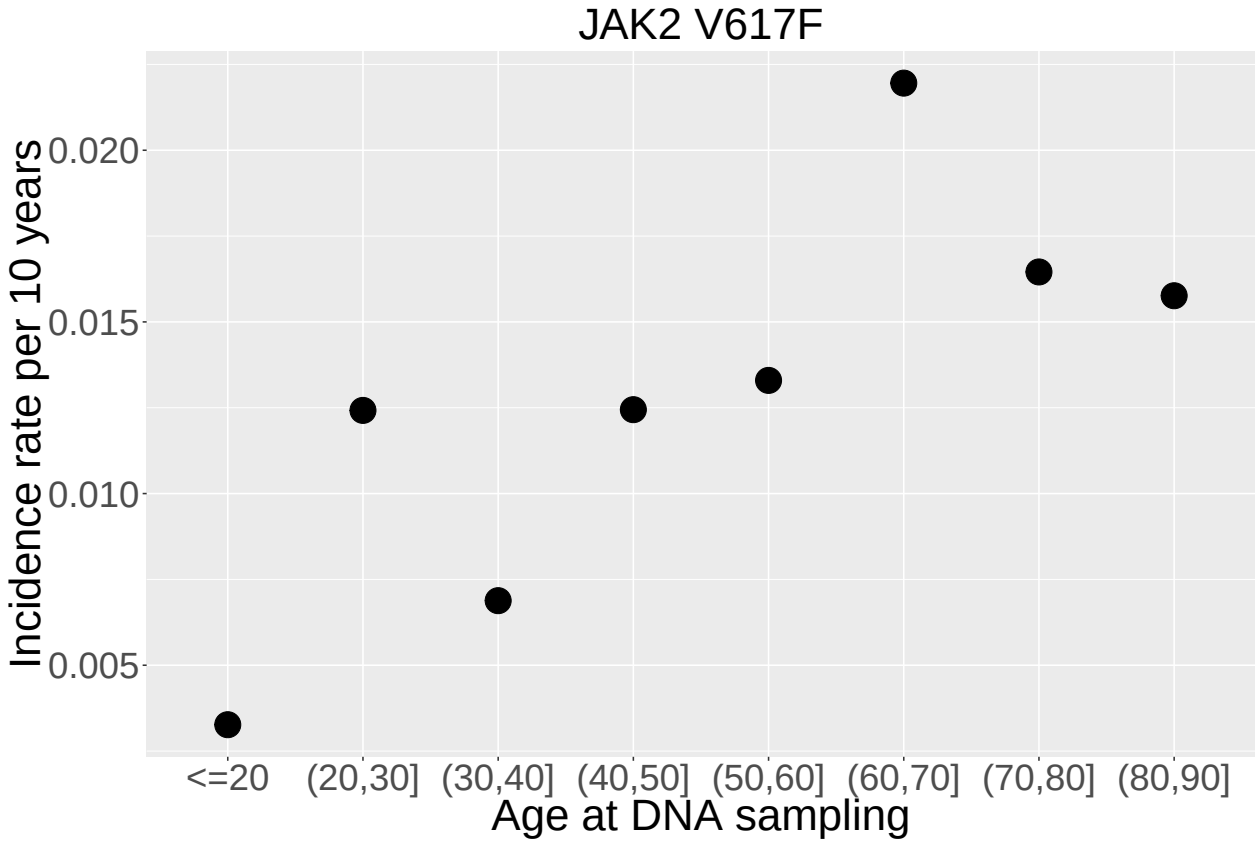

JAK2_V617F
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.05
component1
component2



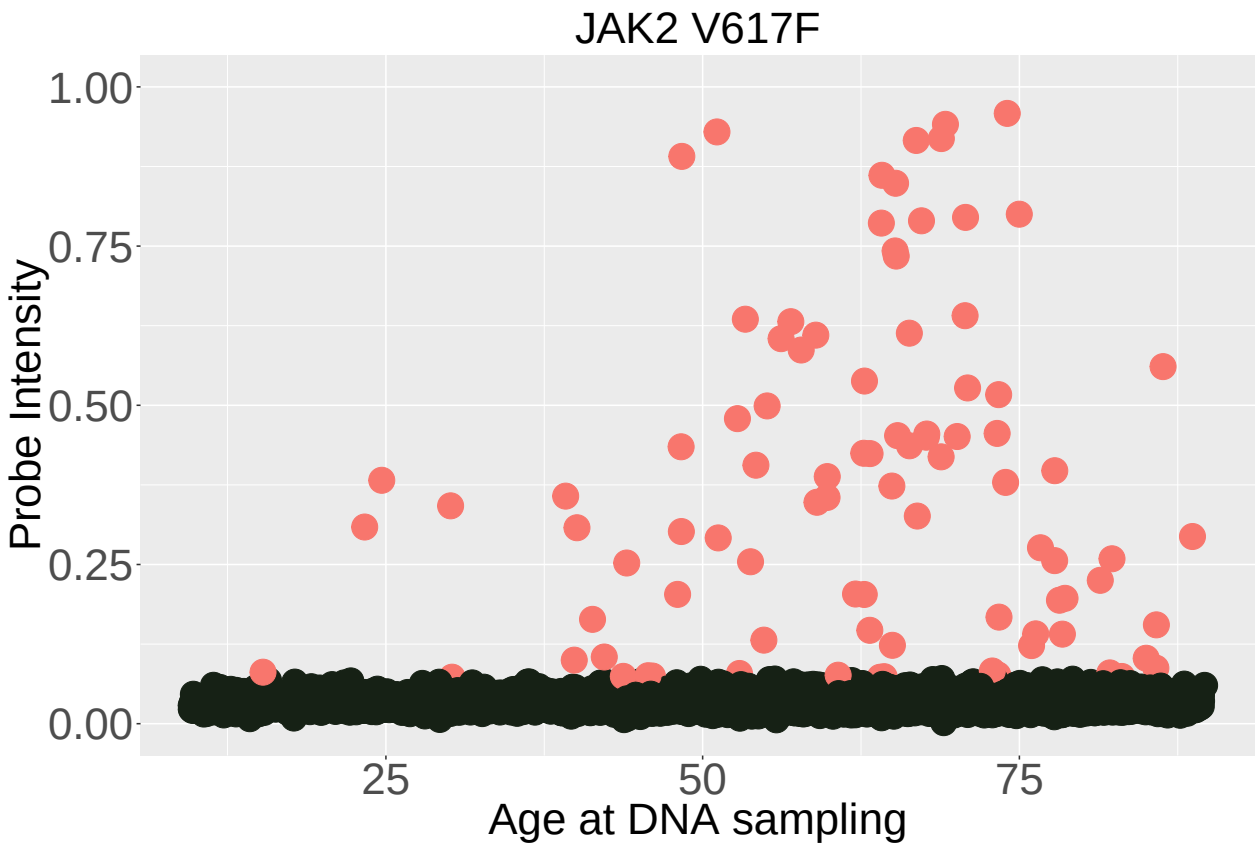
When p-value is 0.01

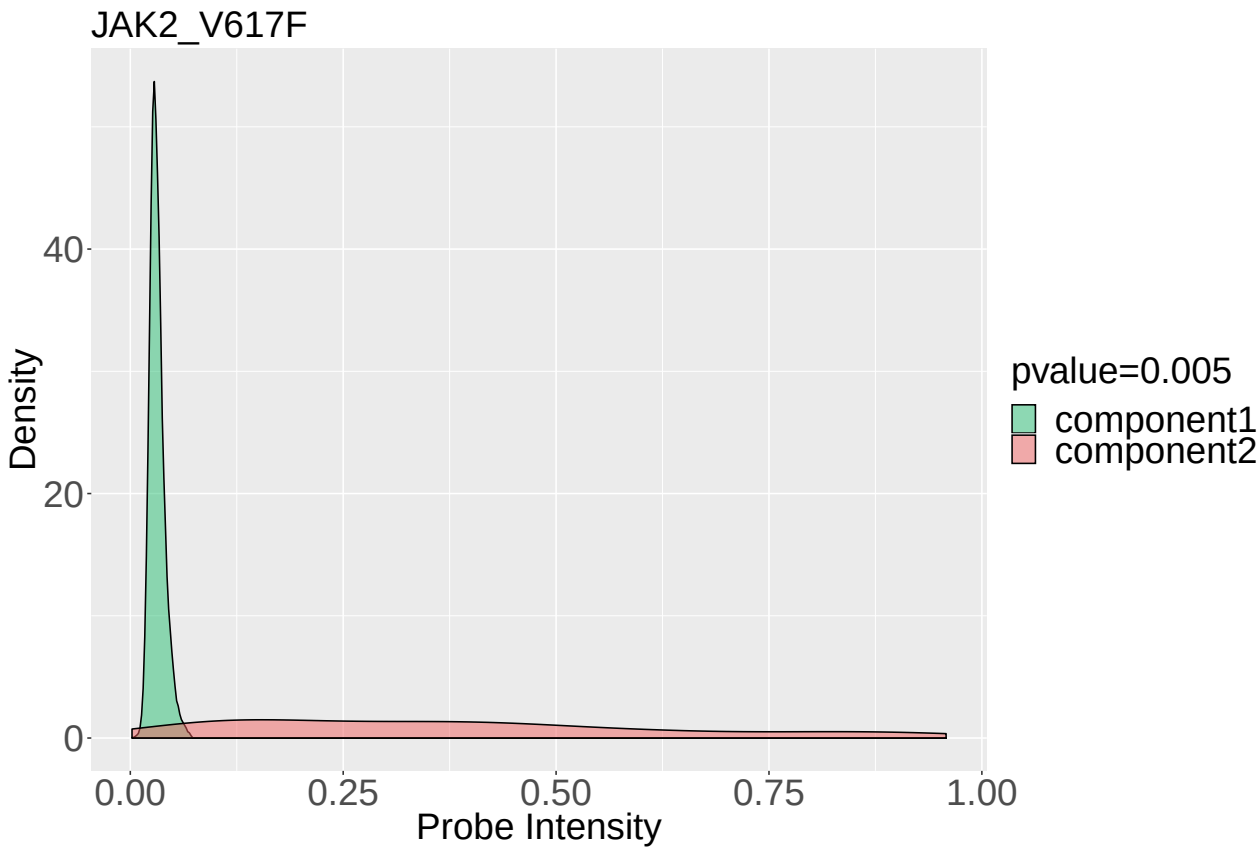

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

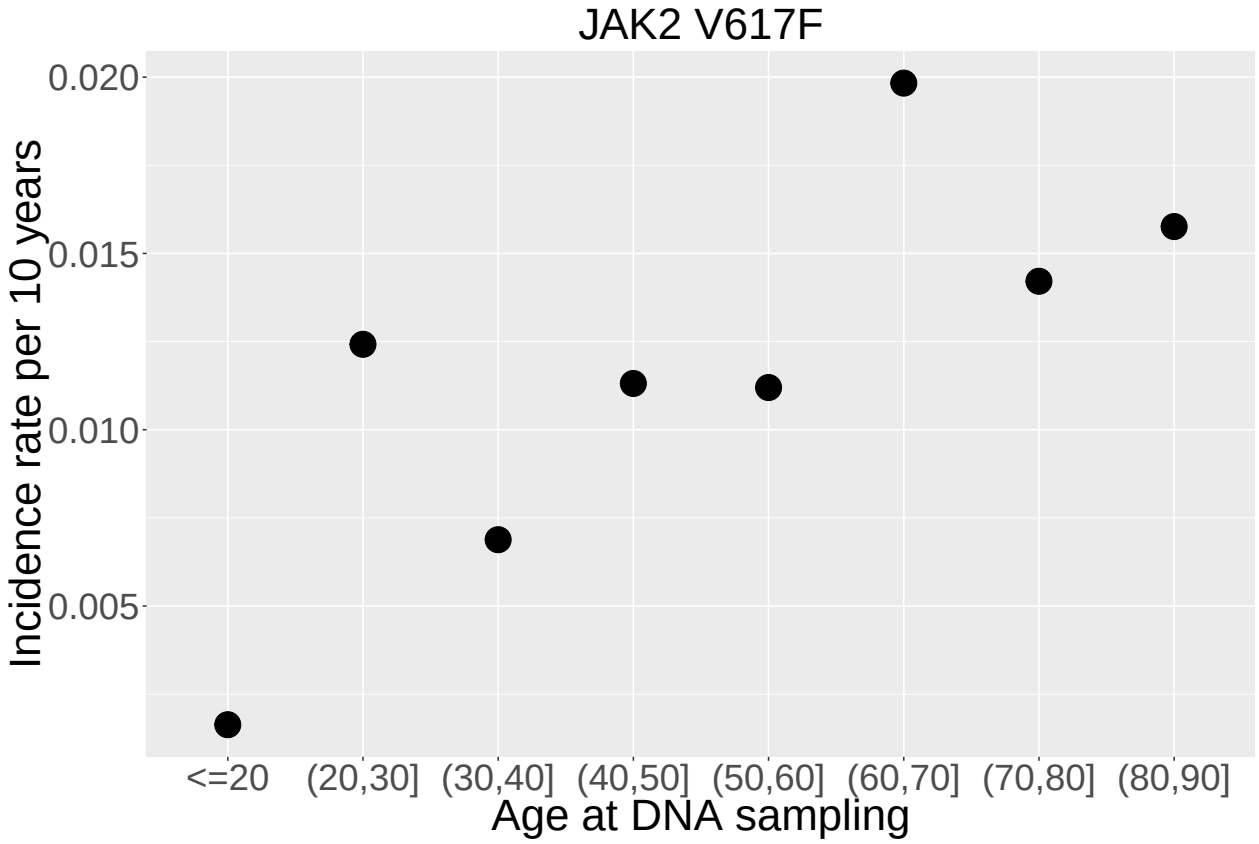

JAK2_V617F
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.01
component1
component2



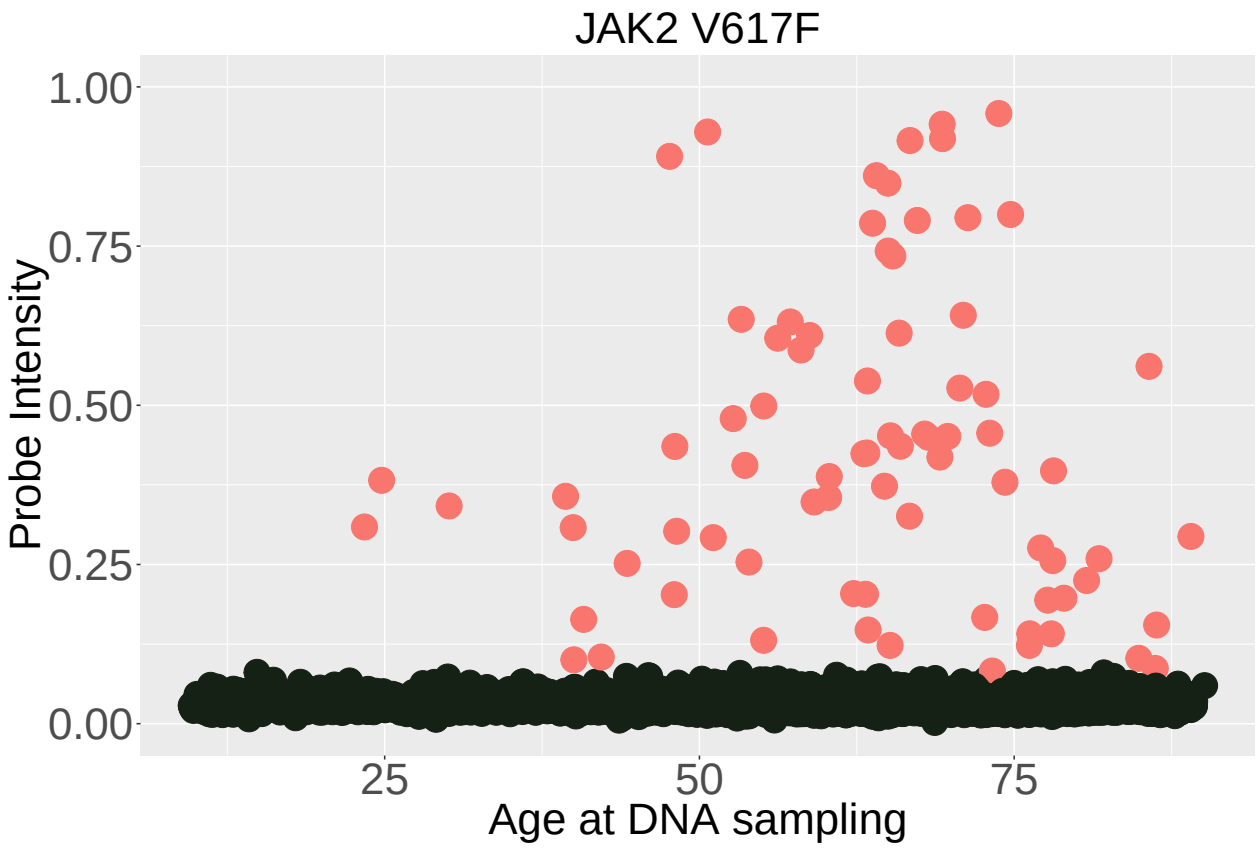
When p-value is 0.005,

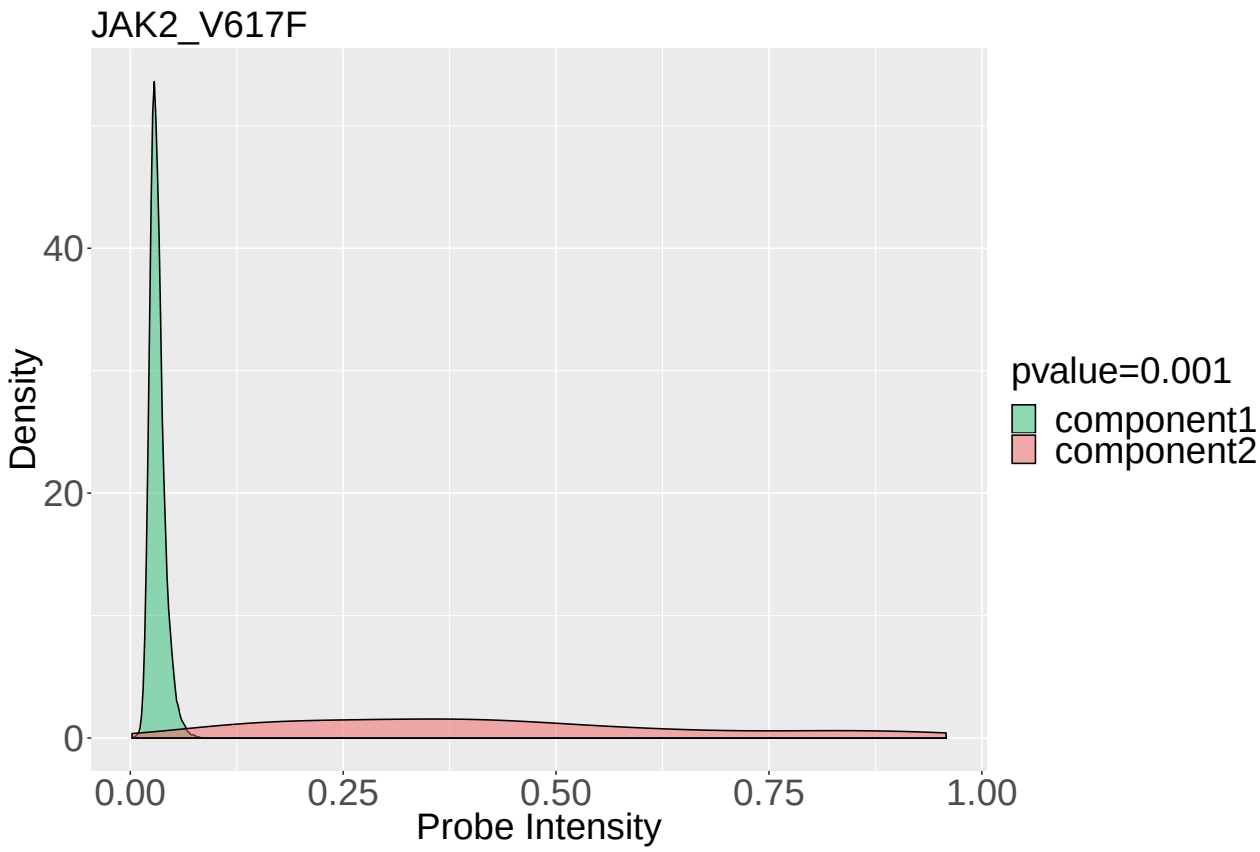

JAK2 V617F
1.00
0.75
0.50
0.25
0.00
Probe Intensity
25
50
75
Age at DNA sampling

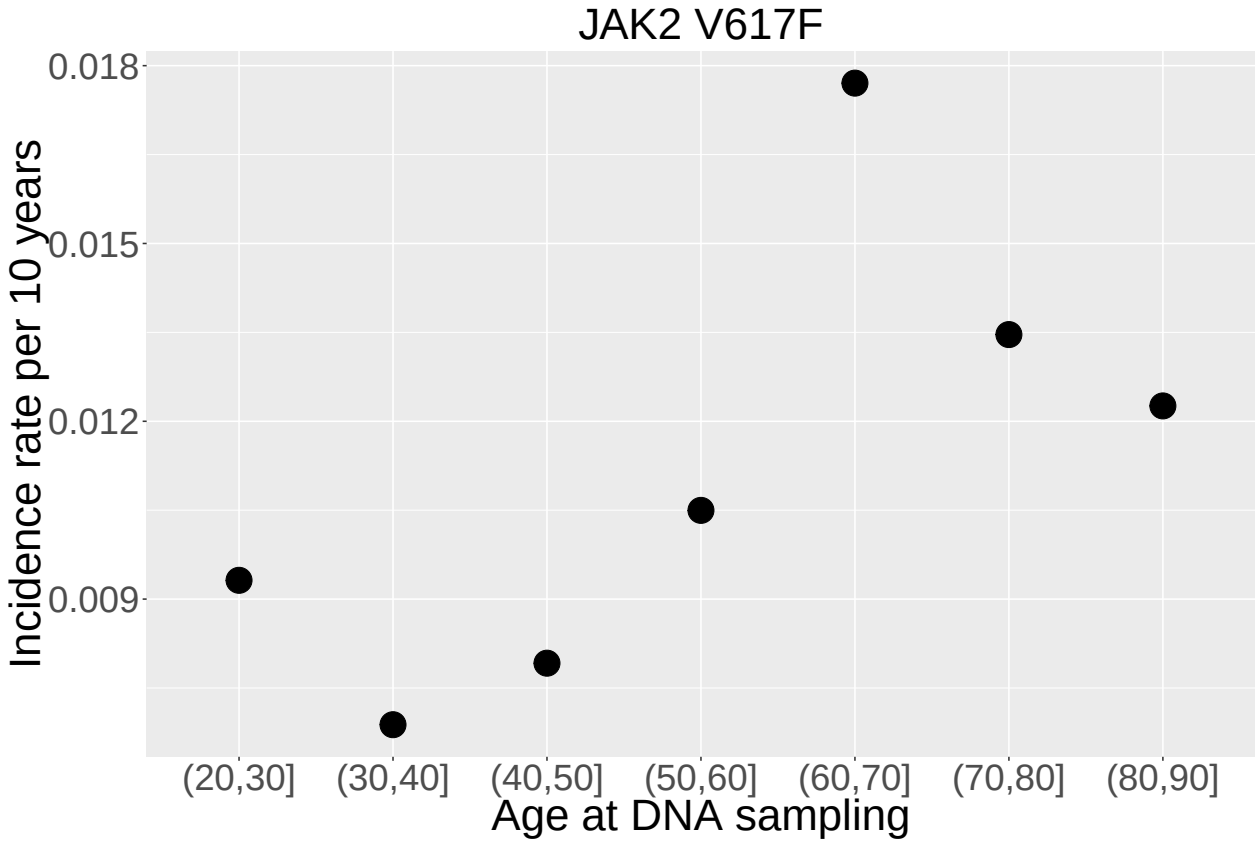

JAK2_V617F
40
20
0
Density
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.005
component1
component2



When p-value is 0.001,

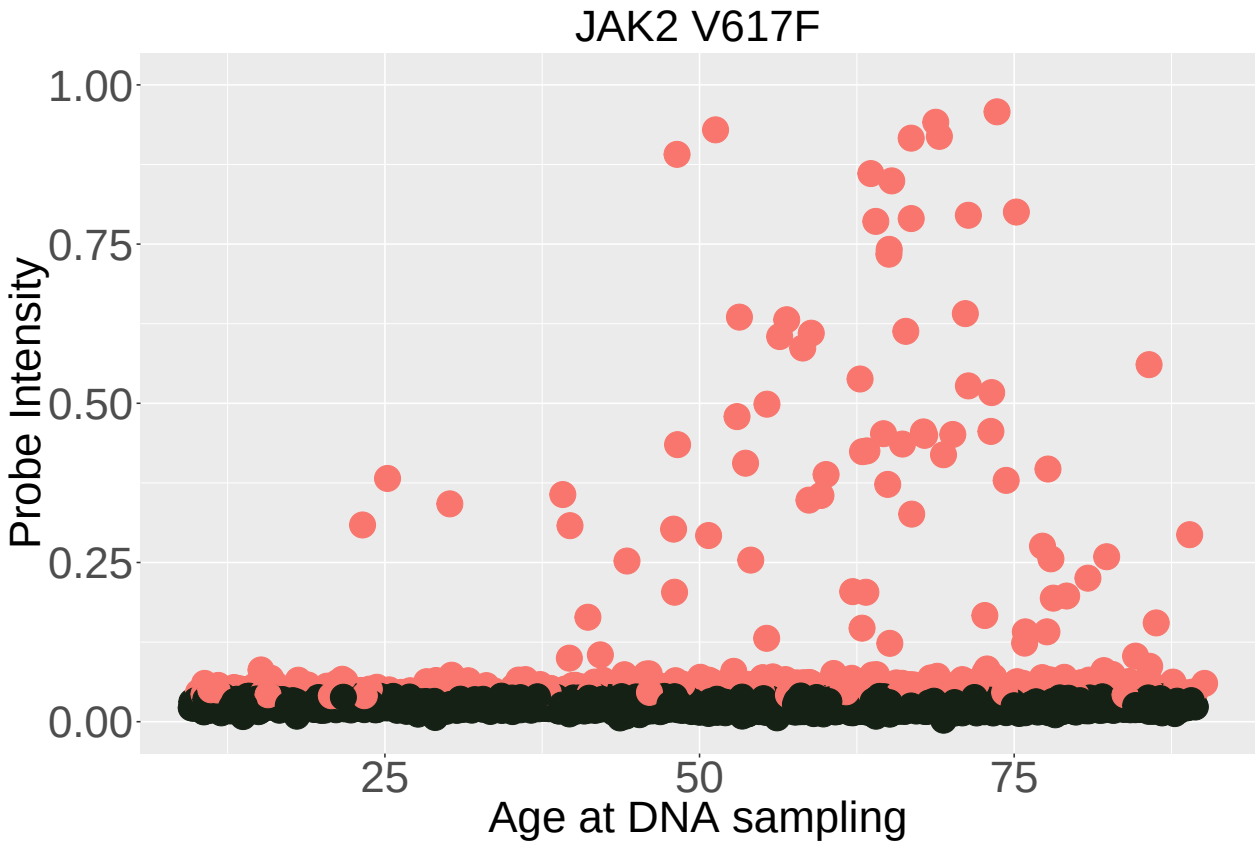

JAK2 V617F
1.00
0.75
0.50
0.25
0.00
Probe Intensity
25
50
75
Age at DNA sampling

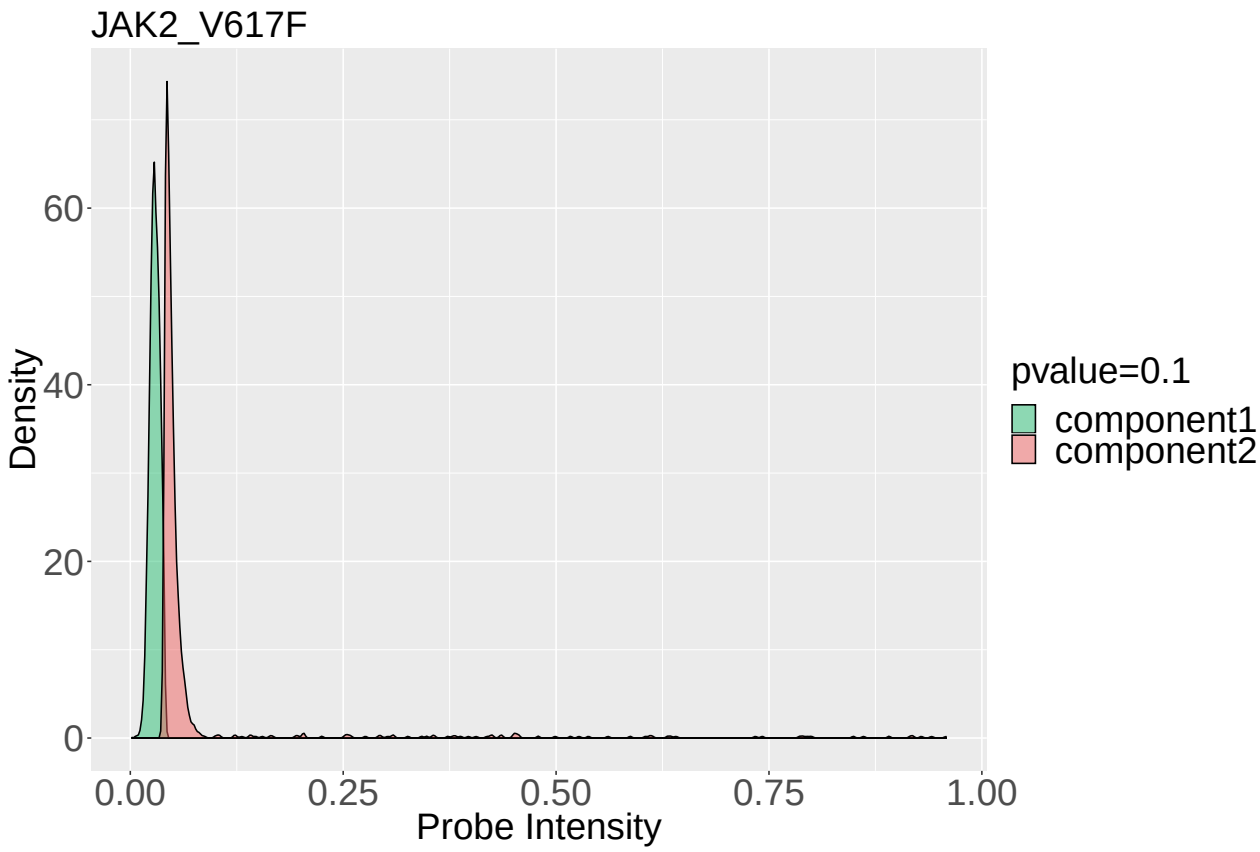

JAK2_V617F
40
20
0
Density
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.001
component1
component2

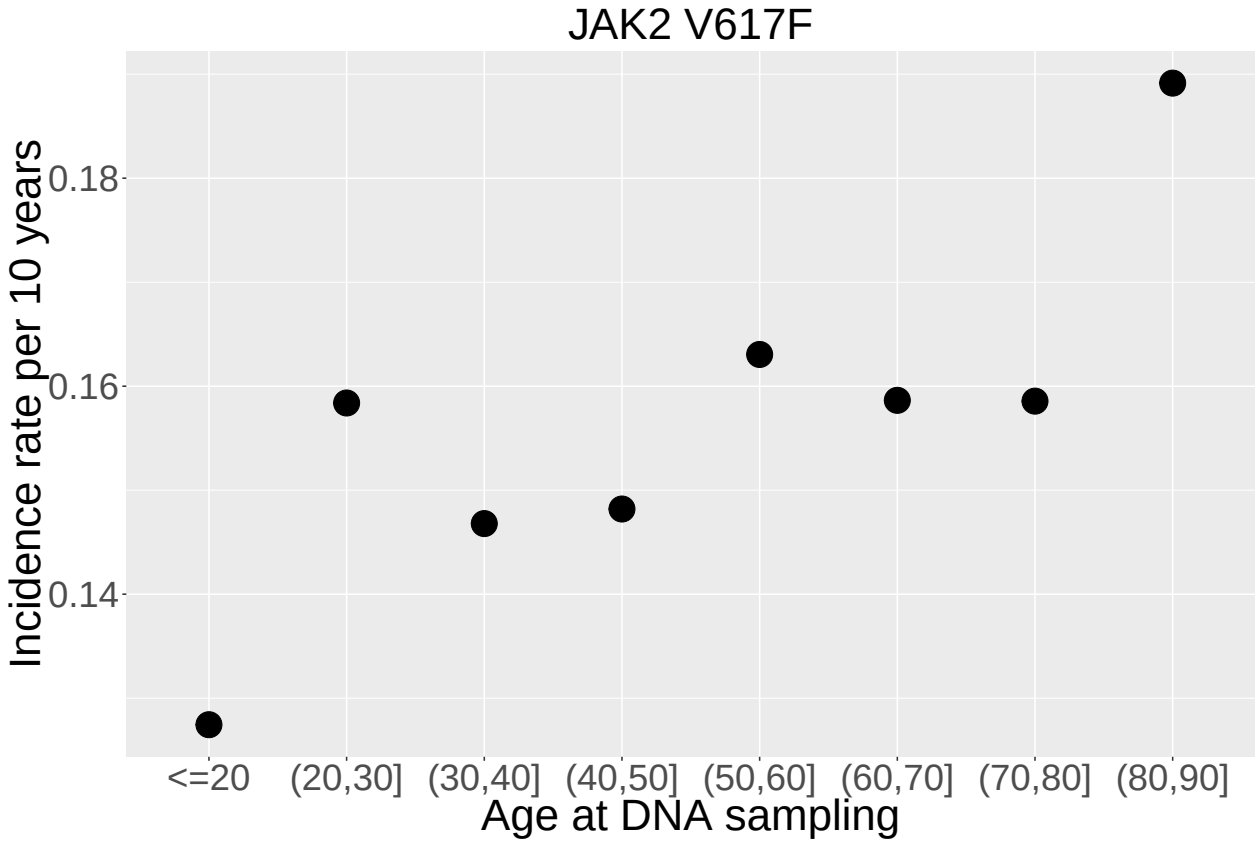


fit normal distribution for the people aged<=40

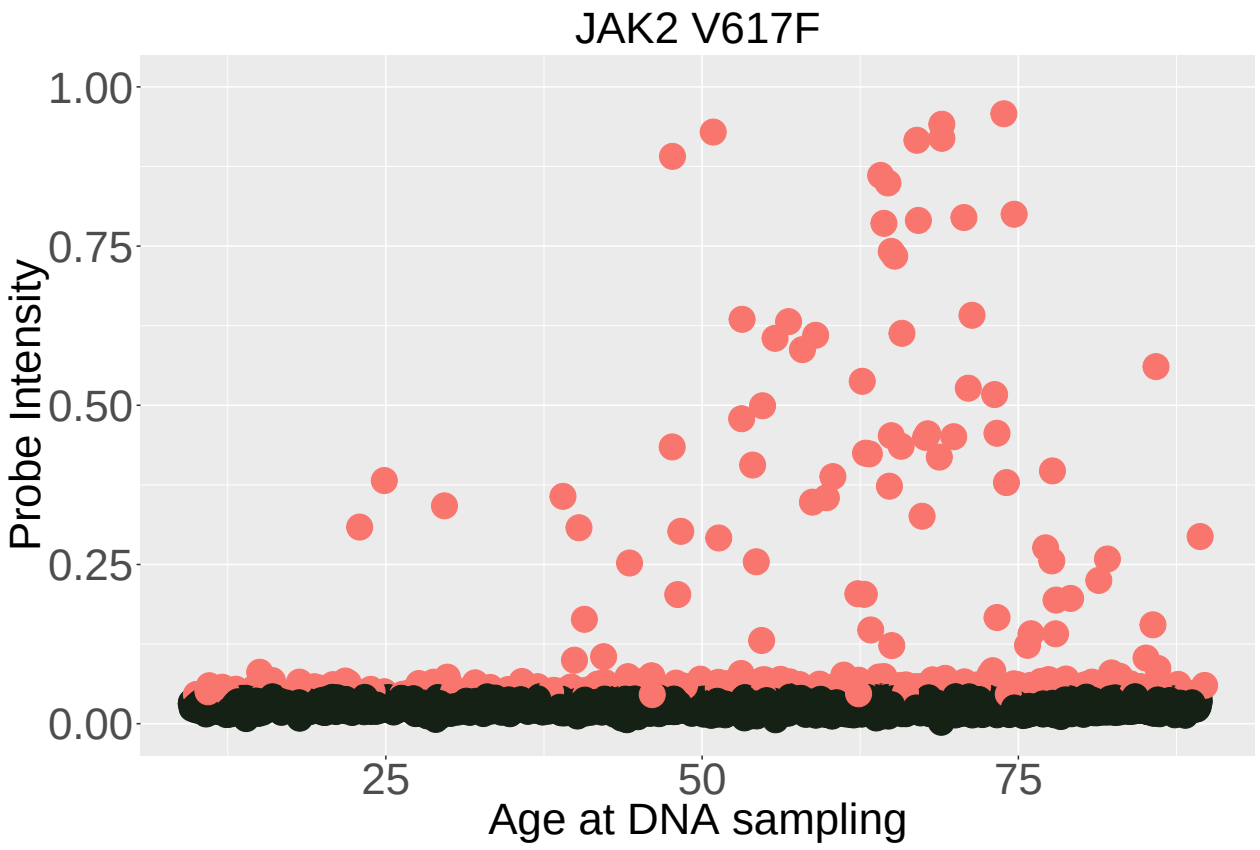
when p-value is 0.1

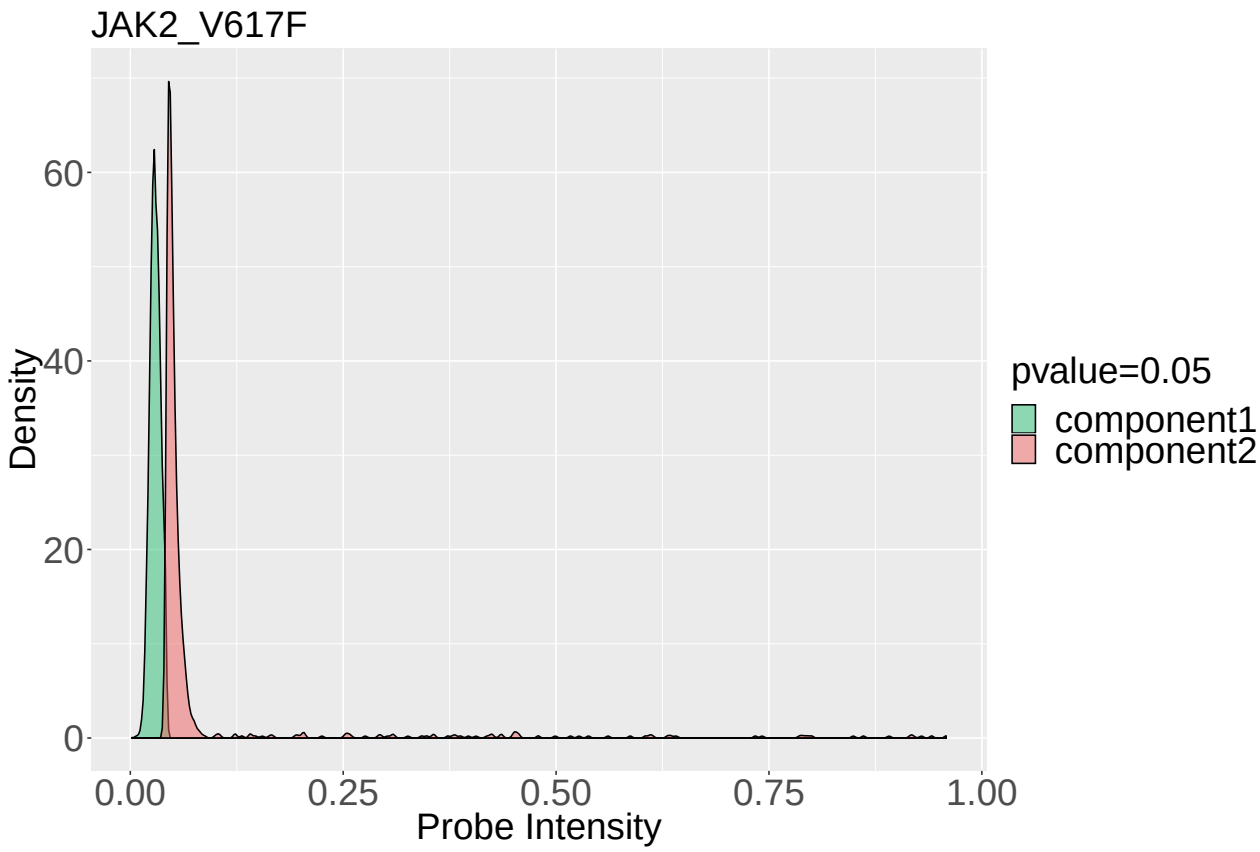

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

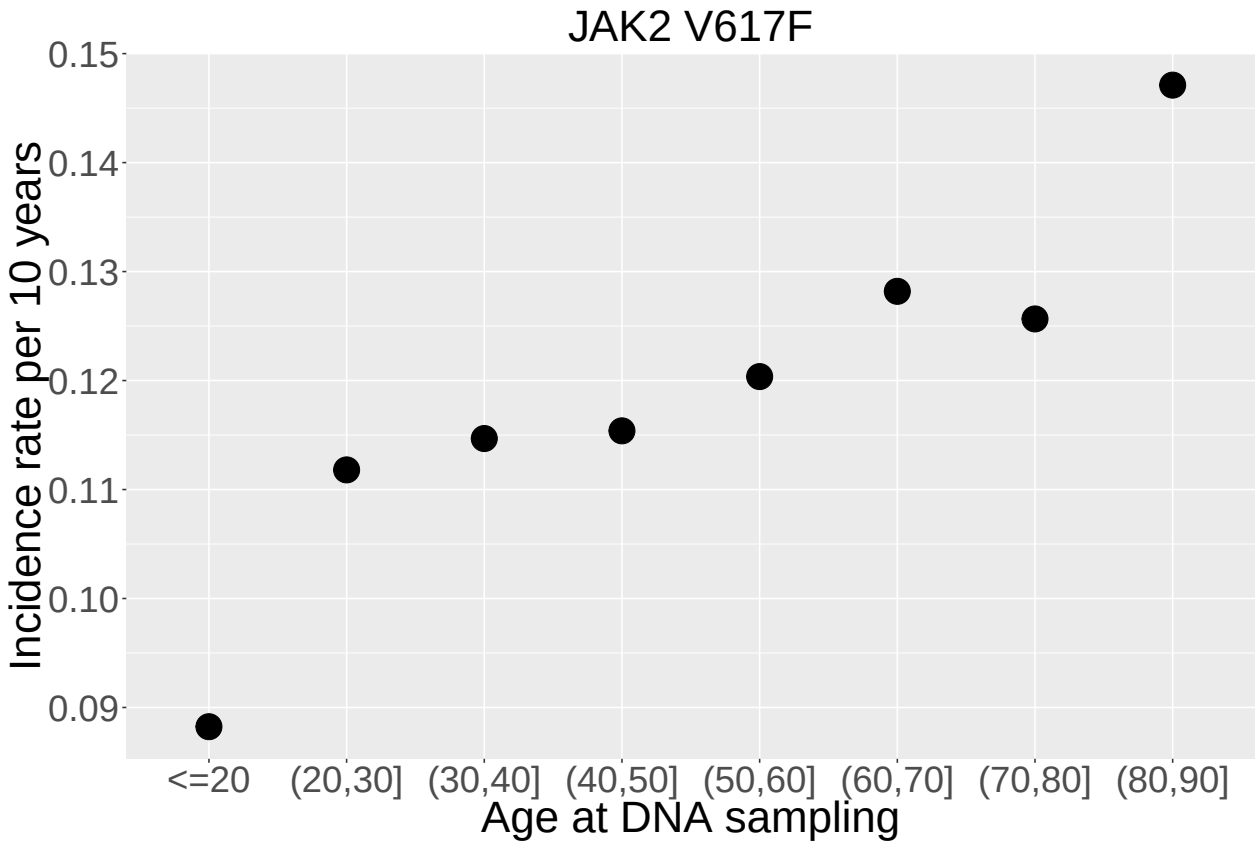

JAK2_V617F
Density
60
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.1
component1
component2



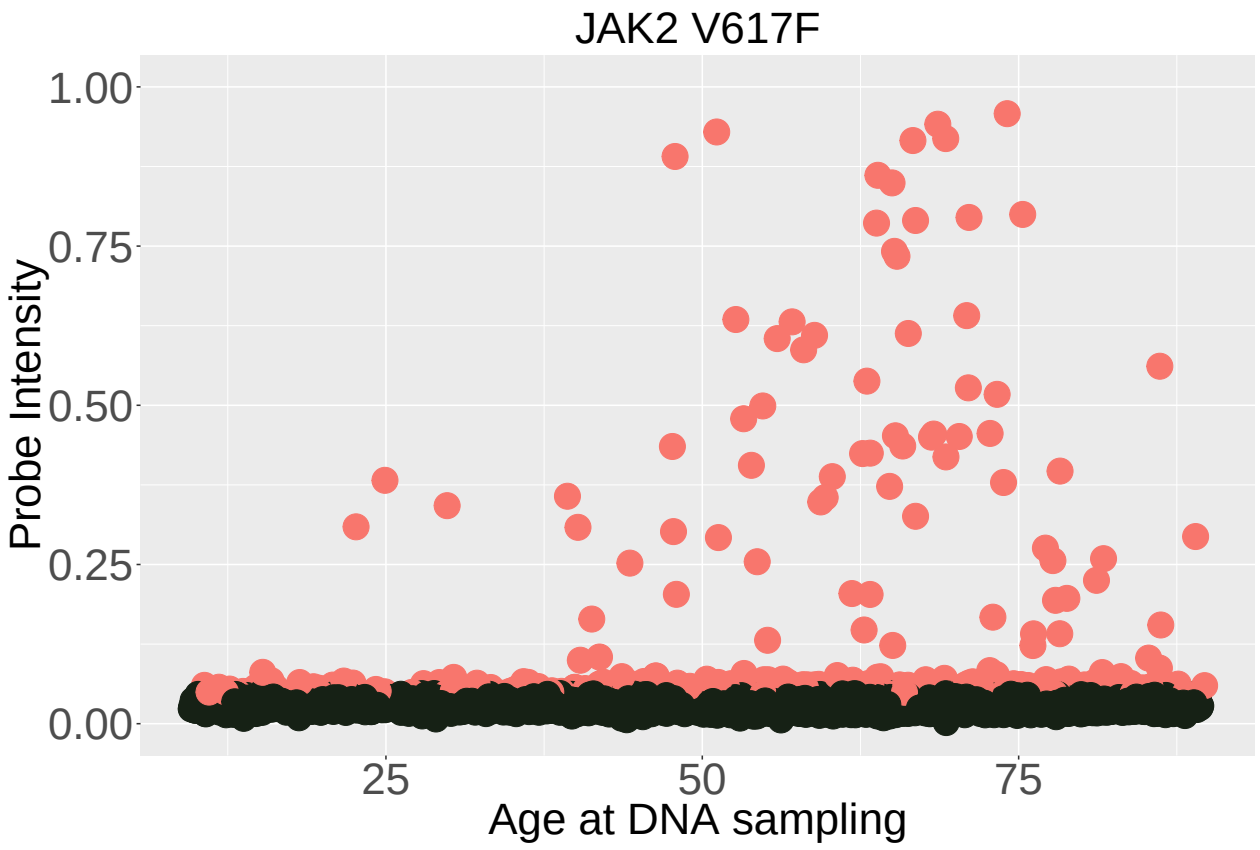
when p-value is 0.05

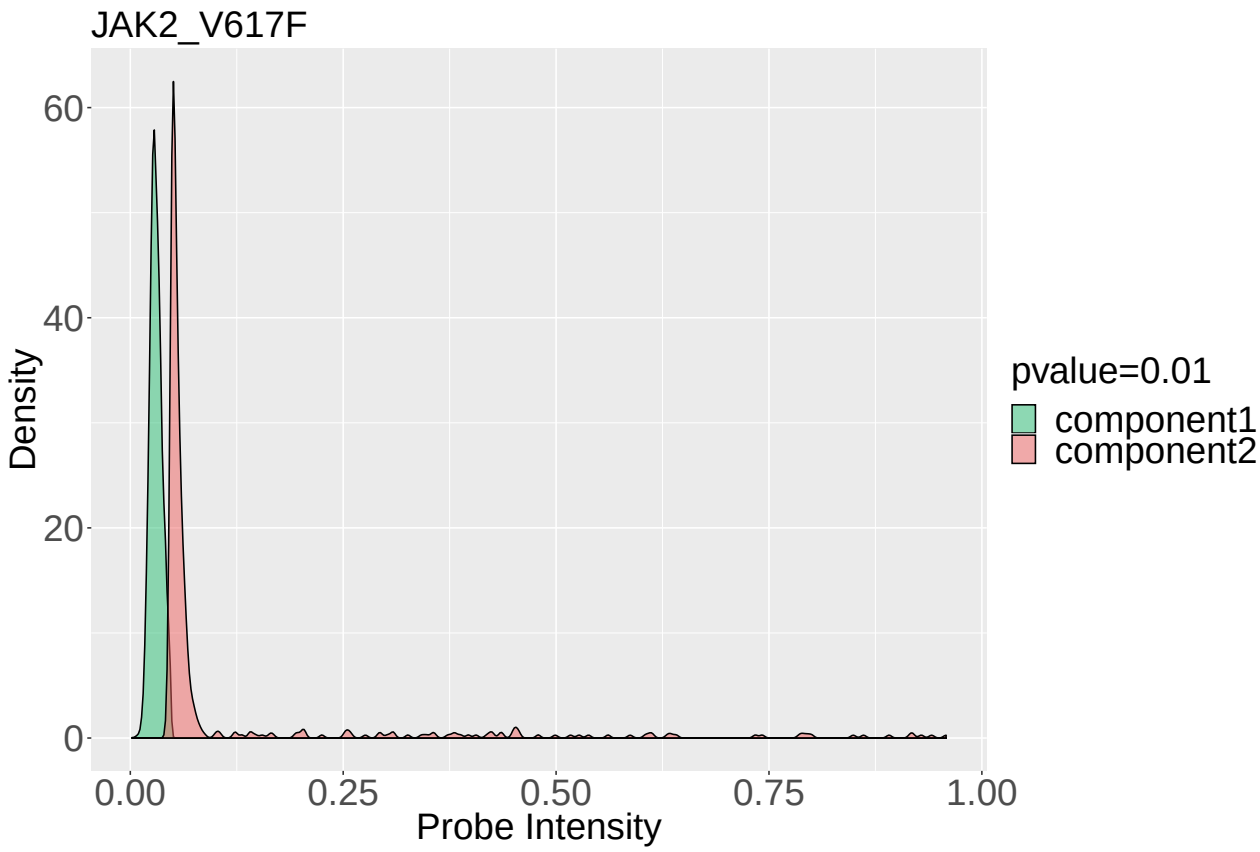

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

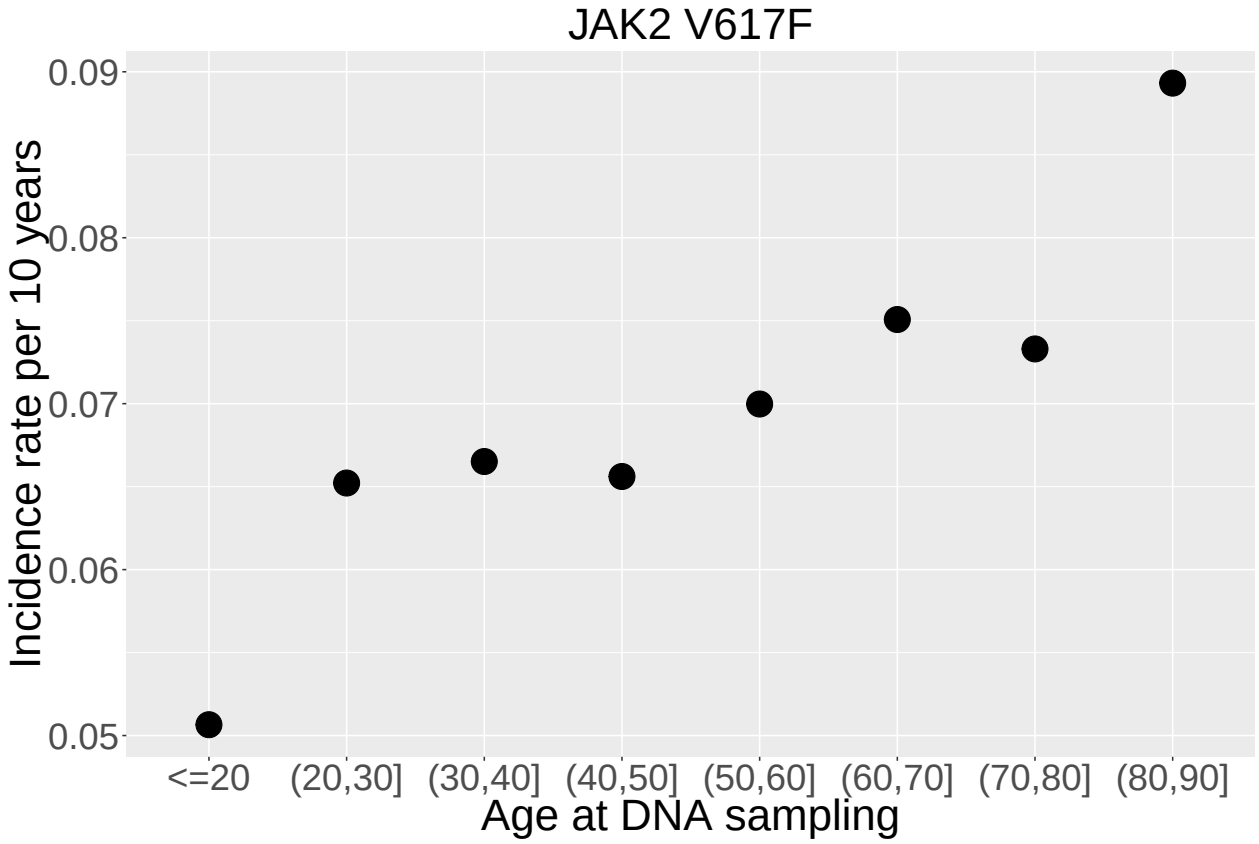

JAK2_V617F
Density
60
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.05
component1
component2



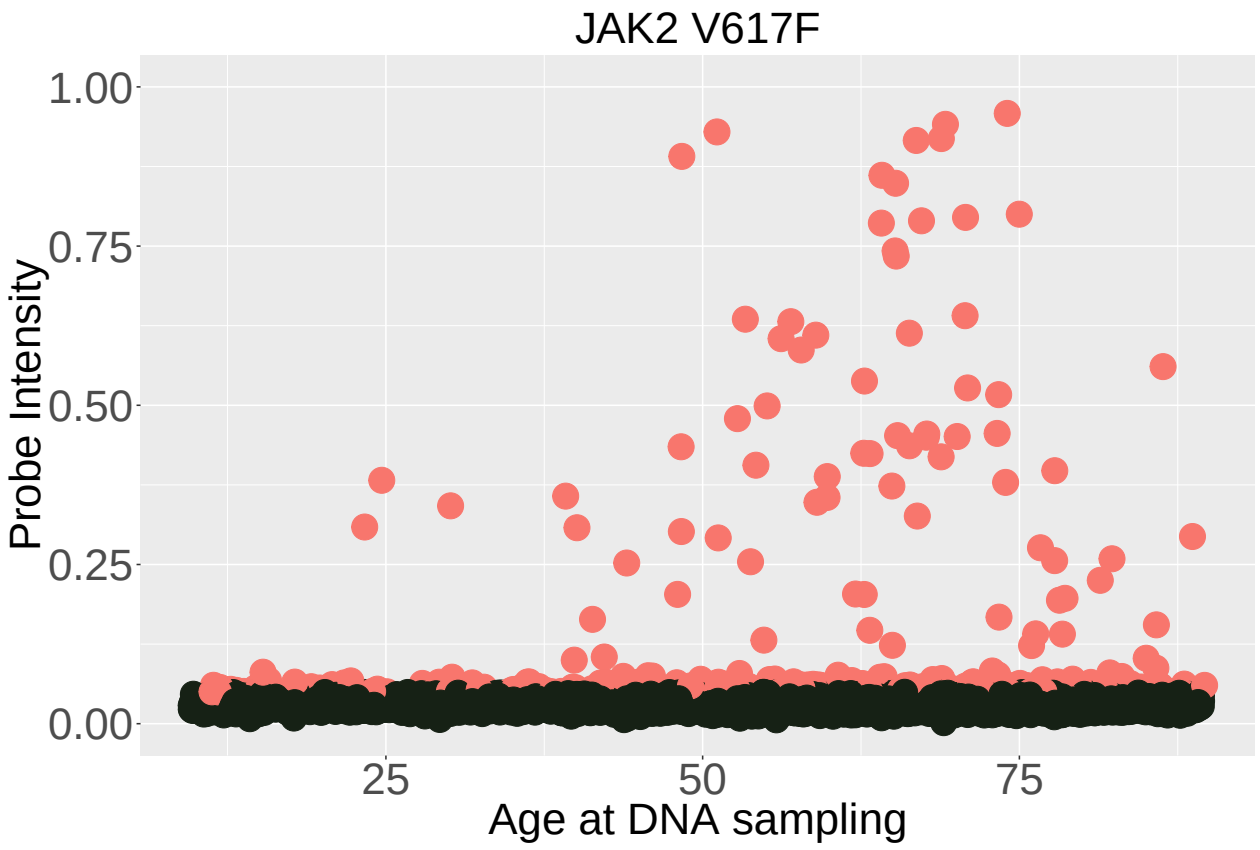
when p-value is 0.01

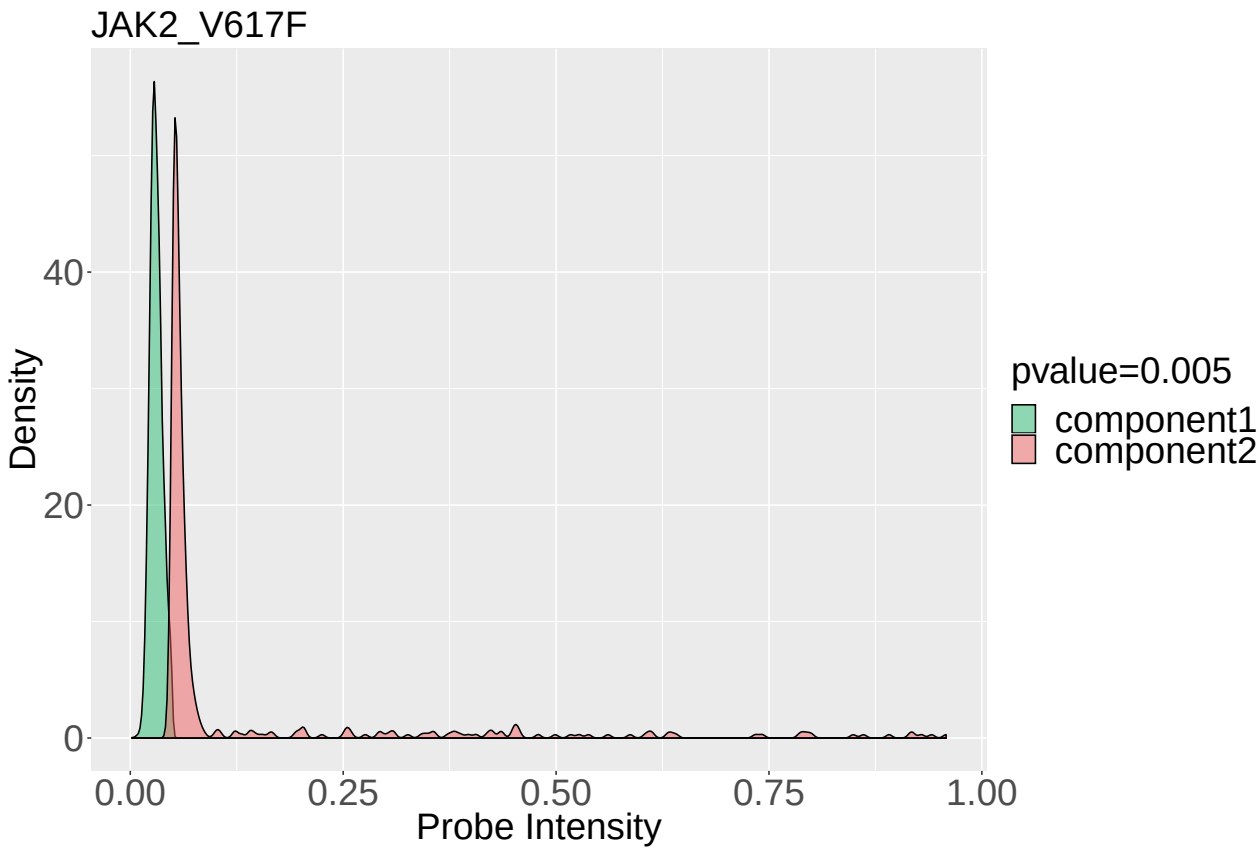

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

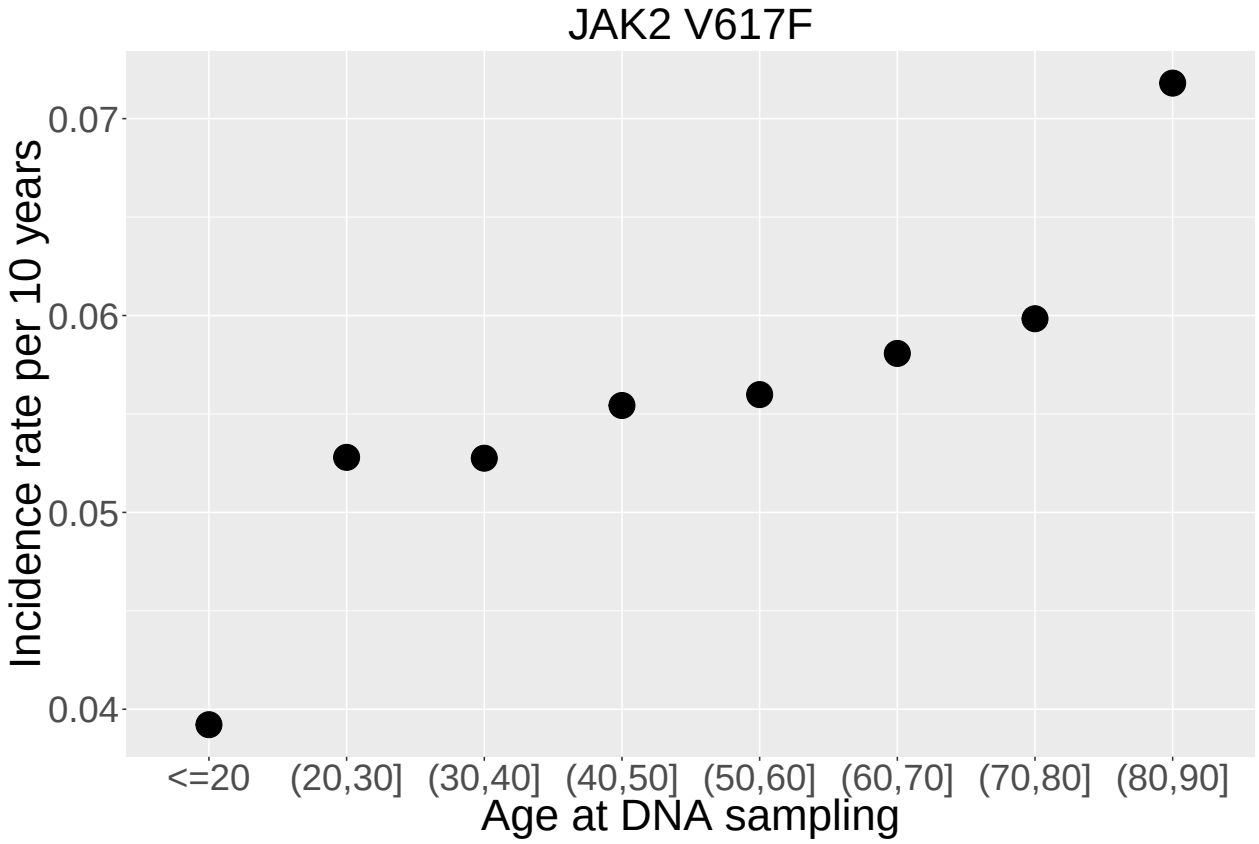

JAK2_V617F
Density
60
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.01
component1
component2



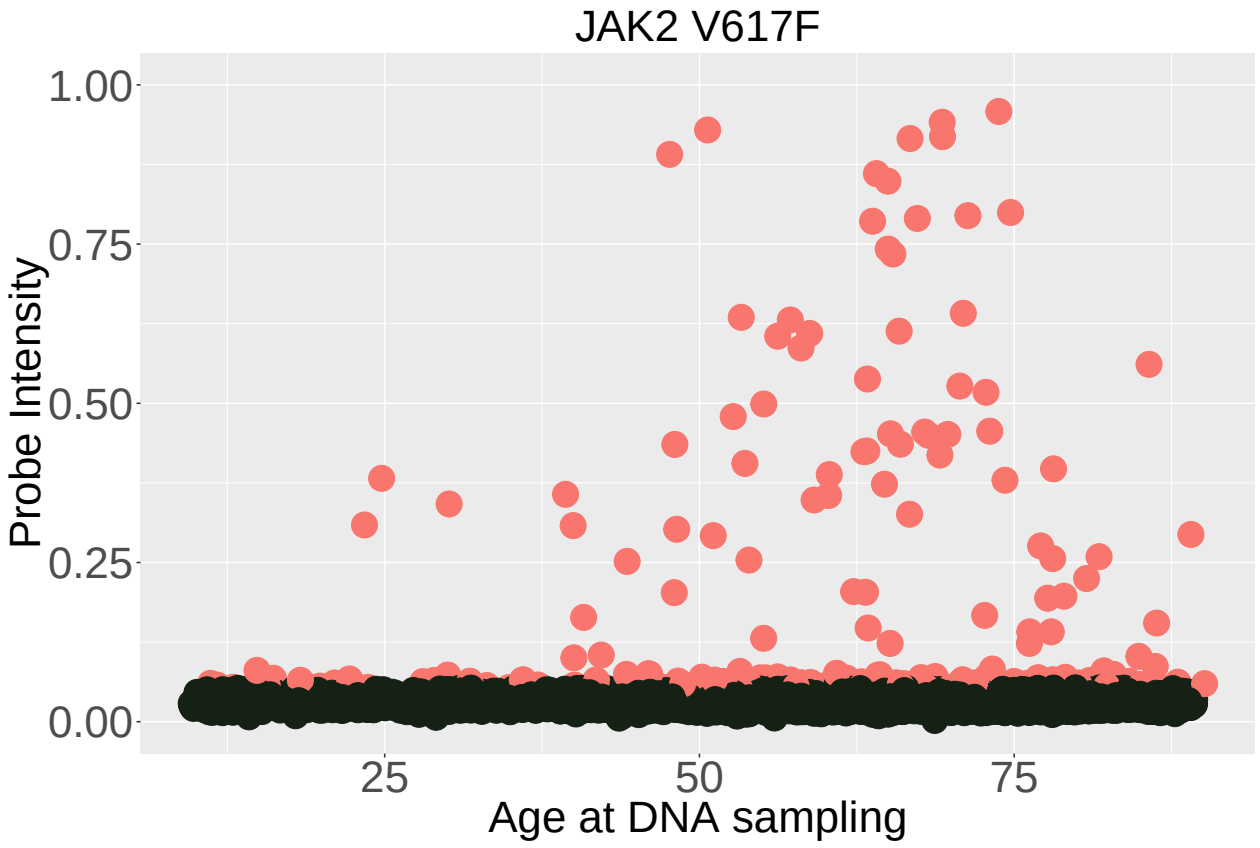
When p-value is 0.005,

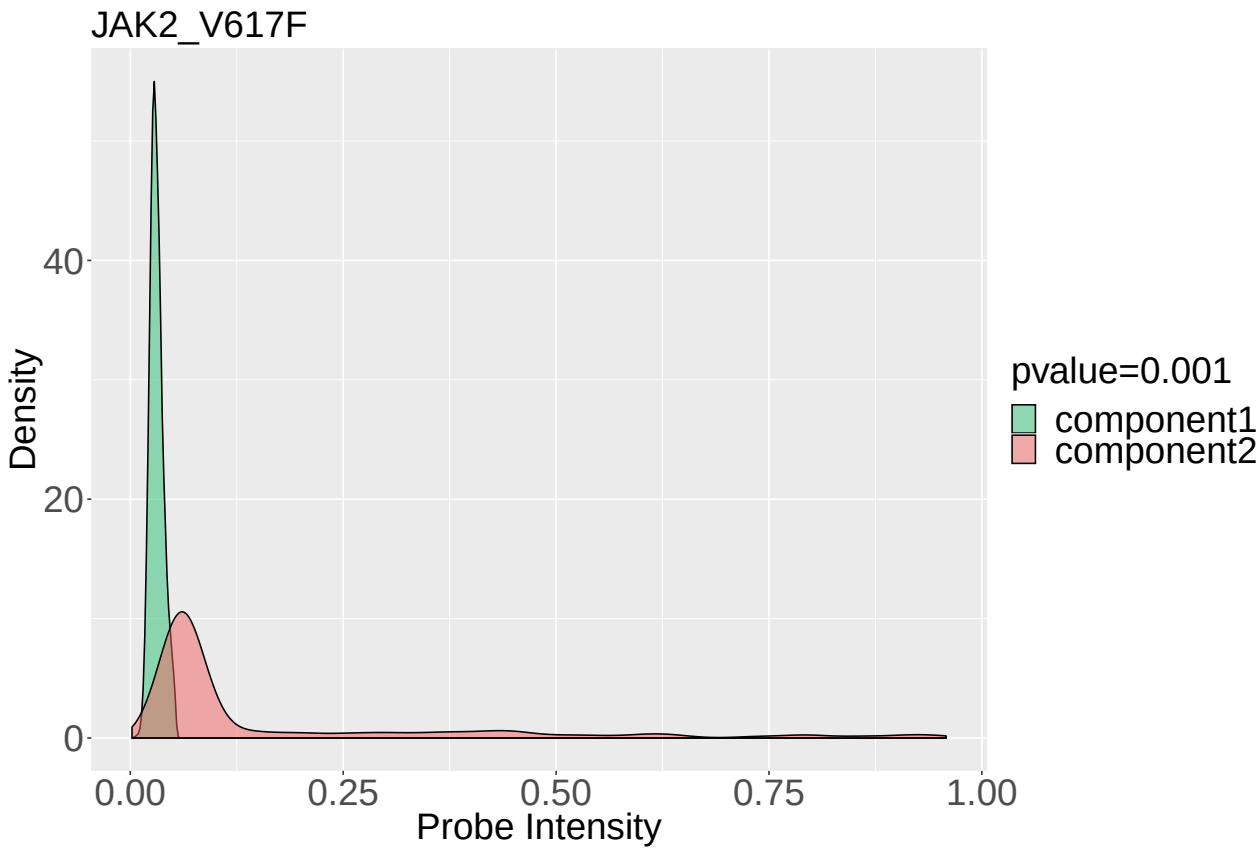

JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling

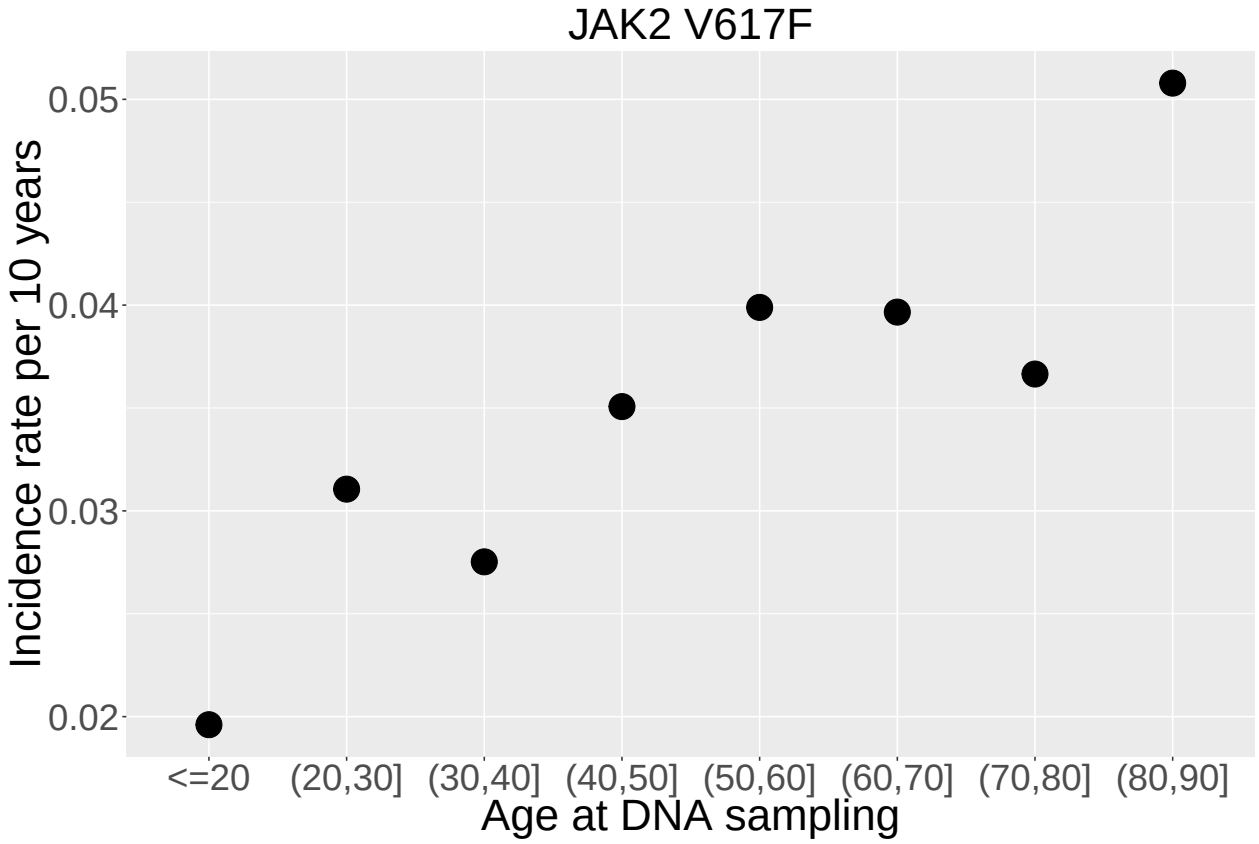

JAK2_V617F
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.005
component1
component2



When p-value is 0.001


JAK2 V617F
Probe Intensity
1.00
0.75
0.50
0.25
0.00
25
50
75
Age at DNA sampling


JAK2_V617F
Density
40
20
0
0.00
0.25
0.50
0.75
1.00
Probe Intensity
pvalue=0.001
component1
component2


JAK2 V617F
Incidence rate per 10 years
0.05
0.04
0.03
0.02
<=20
(20,30]
(30,40]
(40,50]
(50,60]
(60,70]
(70,80]
(80,90]
Age at DNA sampling