Statistical model for k-mer counts.

1 Overview

The distance of k-th k-mer count n_k from haplogroup h is measured by standardized square distance:

$$d_{h,k} = \frac{(n_k - \mu_{h,k} \cdot cov)^2}{\hat{\sigma}_k^2},$$

where $\mu_{k,h}$ is the average number of times k-th k-mer is present in haplogroup h genomes, cov is the sequencing coverage of the sample (estimated coverage for as CG-rich region as the k-th k-mer) and σ_k^2 is the estimated variability of the k-mer count for particular sample given the haplogroup, sample/sequencing quality and coverage.

The total raw distance of the investigated sample from the haplogroup h is measured by the average distance over all k-mers included in the haplogroup model:

$$d_h = \frac{1}{n_{model\ k-mers}} \sum_{i=1}^{n_{model\ k-mers}} d_{h,k}. \tag{1}$$

The haplogroup with smallest distance from the sample is proposed as the most likely haplogroup.

2 Variance σ_k^2 estimation

In estimating the variance for k-mer counts we relay on the following model for k-mer counts. The counts for k-mer k for an individual i from haplogroup h in a sample s are assumed to follow the Poisson distribution for given k-mer and sample:

$$N_{k,s}|k, s \sim Poi(\lambda_{h,i,s,k}),$$

where $\lambda_{h,i,s,k} = (\mu_{h,k} + \gamma_{i,k} + \varepsilon_{s,k}) \cdot cov_s$. The haplogroup mean $\mu_{h,k}$ indicates how many times k-th k-mer appears, on average, in haplogroup h genomes.

The i-th individual has k-th k-mer in his/her genome $\mu_{h,k} + \gamma_{i,k}$ times (repeats). The person-specific random effect $\gamma_{i,k}, \mathrm{E}(\gamma_{i,k}) = 0$ indicates how many more or less copies of this k-mer the given individual has compared to haplogroup average. The random effect $\varepsilon_{s,k}, \mathrm{E}(\varepsilon_{s,k}) = 0$ allows to account for uneven sequencing coverage of different regions in some samples.

If one limits its attention to unique NIPT k-mers (k-mers that are present in genomes of all individuals, but are present in only one location — their copy number is always one), then $\mu_{h,k} + \gamma_{i,k} = 1$ and the NIPT k-mer counts can be modelled as

$$N_{s,k}|s, k \sim Poi((1 + \varepsilon_{s,k}) \cdot cov_s).$$

Hence the variance of k-mer counts for all NIPT k-mers in a sample can be calculated as:

$$\begin{aligned}
\mathrm{Var}(N_{s,k}|s) &= \mathrm{E}\left(\mathrm{Var}(N_{s,k}|s,k)\,|s\right) + \mathrm{Var}\left(\mathrm{E}(N_{s,k}|s,k)\,|s\right) \\
&= \mathrm{E}\left((1+\varepsilon_{s,k}) \cdot cov_s\,|s\right) + \mathrm{Var}\left((1+\varepsilon_{s,k}) \cdot cov_s\,|s\right) \\
&= cov_s + cov_s^2 \cdot \mathrm{Var}\left(\varepsilon_{s,k}|s\right) \\
&= cov_s + cov_s^2 \cdot \sigma_s^2.
\end{aligned}$$

The variance of $\varepsilon_{s,k}$, the $\sigma_s^2 := \mathrm{Var}\left(\varepsilon_{s,k}|s\right)$, measures the overdispersion of k-mer counts compared to the Poisson distribution (the variance of all NIPT k-mer counts from particular sample may have higher variance than cov_s, σ_s^2 can be viewed as a measure of the overdispersion parameter for negative binomial distribution).

If one estimates the variance of NIPT k-mer counts, $\hat{\mathrm{Var}}(N_{s,k}|s)$, and the average sample coverage ($\hat{coverage}_s$, mean NIPT k-mer count), one can calculate an estimate for σ_s^2 for a given sample:

$$\hat{\sigma}_s^2 = (\hat{\mathrm{Var}}(N_{s,k}|s) - \hat{cov}_s)/\hat{cov}_s^2.$$

If one predicts the coverage for a given k-mer based on its CG-content (taking into account the possible CG-related bias in sequencing or higher substitution rate in ancient DNA), the formula for estimating the $\hat{\sigma}_s^2$ changes to

$$\hat{\sigma}_s^2 = (\hat{\mathrm{Var}}\left(N_{s,k}|s\right) - \mathrm{E}\left(\hat{cov}_{s,k}|s\right) - \mathrm{Var}\left(\hat{cov}_{s,k}|s\right))/\mathrm{E}\left(\hat{cov}_{s,k}^2|s\right). \qquad (2)$$

If the estimate from Equation 2 is negative it is replaced with zero.

One can derive the Equation 2 using similar logic as before, just by assuming that after using information about CG% content in estimating coverage the additional unexplained variability in k-mer counts do not depend on k-mer CG%, namely $\mathrm{Var}(\varepsilon_{s,k}|CG) = \sigma_s^2$ for all possible CG% values. Also one has to notice, that the estimated coverage of a k-mer, $\mathrm{cov}_{s,k}$, is the same for all k-mers with the same CG%, $\mathrm{Var}(\mathrm{cov}_{s,k}|s, CG) = 0$:

$$\begin{aligned}
\mathrm{Var}(N_{s,k}|s) =&\mathrm{Var}(cov_{s,k}(1+\varepsilon_{s,k})|s)\\
=&\mathrm{Var}(\mathrm{E}(cov_{s,k}(1+\varepsilon_{k,s})|s,k)|s)+\\
&\mathrm{E}(\mathrm{Var}(cov_{s,k}(1+\varepsilon_{k,s})|s,k)|s)\\
=&\mathrm{Var}(cov_{s,k}(1+\varepsilon_{k,s})|s)+\mathrm{E}(cov_{s,k}(1+\varepsilon_{s,k})|s)\\
=&\mathrm{Var}(cov_{s,k}(1+\varepsilon_{s,k})|s)+\mathrm{E}(cov_{s,k}|s)\\
=&\mathrm{Var}(\mathrm{E}(cov_{s,k}(1+\varepsilon_{s,k})|CG,s)|s)+\\
&\mathrm{E}(\mathrm{Var}(cov_{s,k}(1+\varepsilon_{s,k})|CG,k)|k)+\mathrm{E}(cov_{s,k}|s)\\
=&\mathrm{Var}(cov_{s,k}|s)+\mathrm{E}(cov_{s,k}^2\cdot\sigma_s^2|s)+\mathrm{E}(cov_{s,k}|s)\\
=&\mathrm{Var}(cov_{s,k}|s)+\sigma_s^2\cdot\mathrm{E}(cov_{s,k}^2)+\mathrm{E}(cov_{s,k}|s).
\end{aligned}$$

One can solve this equation for σ_s^2 and replace remaining terms with estimates calculated using sample NIPT k-mers do get the Equation 2.

After estimating the sample-specific σ_s^2 one can proceed with the modelling of the variances for model k-mers (k-mers for which the copy number can vary between individuals and between haplogroups). As stated previously, we assume

$$N_{h,i,s,k}|s,k \sim Poi\left((\mu_{h,k}+\gamma_{i,k}+\varepsilon_{s,k})\cdot cov_{s,k}\right),$$

and we are interested in the variance of the k-mer counts given the haplogroup, estimated sample quality (σ_s^2) and estimated k-mer sequencing coverage $cov_{s,k}$ (estimated from the k-mer CG%):

$$\mathrm{Var}(\,N_{h,i,s,k}\,|\,h,\sigma_s^2,cov_{s,k},k\,).$$

Conditioning on the sample quality here is interpreted as a condition on the $\varepsilon_{s,k}$ values, eg $\mathrm{Var}(\varepsilon_{s,k}|\sigma_s^2) = \sigma_s^2$.

The k-mer count variance for k-th model k-mer, σ_k^2, can be expressed as

$$\begin{aligned}
\sigma_k^2 :=& \mathrm{Var}\left(N_{h,i,s,k} | h, \sigma_s^2, cov_{s,k}, k\right) \\
=& \mathrm{Var}\left(\mathrm{E}(N_{h,i,s,k} | h, i, s, k) | h, \sigma_s^2, cov_{s,k}, k\right) + \\
& + \mathrm{E}\left(\mathrm{Var}(N_{h,i,s,k} | h, i, s, k) | h, \sigma_s^2, cov_{s,k}, k\right) \\
=& \mathrm{Var}\left((\mu_{h,k} + \gamma_{i,k} + \varepsilon_{s,k}) \cdot cov_{s,k} | h, \sigma_s^2, cov_{s,k}, k\right) + \\
& + \mathrm{E}\left((\mu_{h,k} + \gamma_{i,k} + \varepsilon_{s,k}) \cdot cov_{s,k} | h, \sigma_s^2, cov_{s,k}, k\right) \\
=& \mathrm{cov}_{s,k}^2 \cdot \mathrm{Var}\left(\gamma_{i,k} + \varepsilon | h, \sigma_s, cov_{s,k}, k\right) + \\
& + \ cov_{s,k} \cdot \mathrm{E}\left((\mu_{h,k} + \gamma_{h,i} + \varepsilon_{s,k}) | h, \sigma_s, cov_{s,k}, k\right) \\
=& \mathrm{cov}_{s,k}^2 \cdot (\sigma_h^2 + \sigma_s^2) + cov_{s,k} \cdot \mu_{h,k}.
\end{aligned} \tag{3}$$

The k-mer copy number variability within the haplogroup h, σ_h^2, is estimated during the model building as the variability in k-mer copy numbers in the reference sample. The value of $\mu_{h,k}$ is estimated as the mean copy number for the given haplogroup.

Additional problems might arise if in a small reference set a given k-mer is never observed. If one uses the estimated values for $\mu_{h,k}$, $\hat{\mu}_{h,k} = 0$ and for $\hat{\sigma}_h^2 = 0$ in the Equation 3 to calculate the expected variability of the k-mer count one could get the estimated variance for k-mer counts to be exactly equal to zero (especially for high quality samples for which $\sigma_s^2 = 0$). However, if due to a sequencing error one does observe this particular k-mer at least once, then the distance from the given haplogroup as calculated by Equation 1 would be infinitely large. To account for additional variability of the k-mer counts due to sequencing errors/imprecise estimation of the haplogroup means, additional correction factors are added when estimating the variance:

$$\begin{aligned}
\hat{\sigma}_k^2 =& cov_{s,k}^2 \cdot (\hat{\sigma}_h^2 + \hat{\sigma}_s^2) + cov_{s,k} \cdot \hat{\mu}_{h,k} + \\
& + (\hat{\mu}_{h,k}^2 / \hat{\mathrm{E}}(cov|h) + \hat{\sigma}_h^2)/n_h \cdot cov_{s,k}^2 + \lambda_{err} \cdot cov_{s,k},
\end{aligned} \tag{4}$$

where λ_{err} is the average number of times we encounter a k-mer due to sequencing errors if sequencing with coverage 1. The default value used in this implementation is $\lambda_{err} = 0.003$. The correction term $\lambda_{err} \cdot cov_{s,k}$ intends to account for the possible effect of sequencing errors.

The correction term $(\hat{\mu}_{h,k}^2 / \mathrm{E}(cov_{h,i,s,k}|h) + \hat{\sigma}_h^2)/n_h \cdot cov_{s,k}^2$ tries to account for the additional possible variability in k-mer counts due to possible misestimation of $\mu_{h,k}$. If the number of reference samples from the haplogroup

h, n_h, is large, then this correction term vanishes to zero. But if there are just a few reference samples from this haplogroup, then this correction term tries to account for additional uncertainty due to possible misestimation of the haplogroup mean.

3 Statistical haplogroup testing

Sometimes, especially for low-coverage and low-quality samples, the selected haplogroup h might not be the correct haplogroup. To investigate the reliability of the call one might want to test the hypothesis:

$$\begin{aligned} &H_0 : \text{some other haplogroup is the correct haplogroup} \\ &H_1 : \text{the called haplogroup is the correct haplogroup.} \end{aligned}$$

Notice, that the null hypothesis is a composite hypothesis and by rejecting the null hypothesis one has proven that the proposed haplogroup is really the correct haplogroup. However, the reference samples used for building the haplogroup model might not be really representative for the samples one might want to use the model for. For example, the high quality reference samples might come from currently living representatives of a haplogroup, but the ancient DNA sample one might want to test might come from a period when that haplogroup reprentatives had somewhat different Y-chromosomes. One might accidentally prove the incorrect call just because the ancient DNA sample from haplogroup h' is considerably different from contemporary reference samples from haplogroup h' (there could exist other possible problems of the same kind, for example if one uses one sequencing method for reference samples and another for the sample tested one potentially might obtain significantly different k-mer distributions just because the difference in sequencing methods). To allow the reference samples to be from different population than the tested sample one might want to slightly rephrase the hypothesis. The modified alternative hypothesis is H_1: the tested individual Y-chromosome is more similar (closer) to the called haplogroup (given unlimitied coverage and no sequencing errors) than to any other tested haplogroup vs the null hypothesis stating that some other haplogroup might actually be more similar to the tested individual Y-chromosome.

3.1 Testing in case of two alternatives

If there exists just two possibilities, the investigated individual might belong either into haplogroup h or haplogroup h'. Lets assume, that according to the sample haplogroup h was the closest haplogroup, $d_h < d_{h'}$, and we want to prove that the haplogroup h is actually the closest haplogroup to the investigated individual Y-chromosome, $\mathrm{E}\left(d_h - d_{h'}\right) < 0$, eg one can test the pair of hypothesis

$$H_0 : \mathrm{E}\left(d_h - d_{h'}\right) \geq 0$$
$$H_1 : \mathrm{E}\left(d_h - d_{h'}\right) < 0.$$

If one can reject the null hypothesis then one has proven the haplogroup h to be the closer haplogroup. To test these hypothesis one can look at the pairwise differences for each k-mer:

$$z_k = d_{h,k} - d_{h',k}.$$

Then, by using the one-sided z-test, we can test the hypothesis $H_0 : \mathrm{E}z_k \geq 0$. Small p-value would indicate, that the called haplogroup is really the correct haplogroup (tested individual Y-chromosome is really more similar, closer, to the haplogroup h than to the alternative).

3.2 Testing in case of more than two alternatives

When testing the composite null hypothesis (no other haplotype can be closer to the sample than the haplotype called), first for all the alternatives pairwise tests are performed. Then the largest p-value obtained is reported as the p-value for the composite null hypothesis. This type of testing composite hypothesis is a conservative approach from statistical perspective — selecting the largest p-value as the p-value for the composite test guarantees that the type I-error probability will not exceed α for a given significance level α, but can sometimes actually be considerably smaller than α.