

White Paper: Calculating Genotype Likelihoods

Introduction

It is often useful to determine the most likely genotype by examining all of the reads that have a base aligned at a base position. Most often, the observed allele frequency at that base position will very closely match the reference that the reads were aligned to; however, you may be interested in identifying SNPs located in genes of interest or differently expressed alleles between samples.

There are at least a few reasons you might observe a base call that does not match the reference at a given location — a bad base call when reading the sequence, a bad alignment resolution of multiple possible alignments, or a genuine SNP. Because phred quality scores only help resolve the first cause of mismatches, Partek allows you to specify an error probability that is assumed to be a constant probability of identifying a SNP based on read or alignment error.

When a homozygous genotype would be observed, you would expect to observe nearly 100% of the homozygous allele. When the base frequencies of a heterozygous base are examined, you expect to observe nearly 50% for each allele. The observed base frequencies may deviate slightly from these numbers because call and alignment errors.

An example of expected allele probabilities using for an error probability of .01 is given below. If an error occurred (caused by base calling or mapping), assume each of the 4 alleles are equally likely to be observed with probability P_{error} (including alleles compatible with the genotype). P_{hom} is the expected probability of observing the allele matching a homozygous genotype. P_{het} is the probability of observing each of the two alleles of a heterozygous genotype.

$$\begin{aligned}P_{\text{error}} &= .01 / 4 \\P_{\text{hom}} &= 1.0 - 3 * P_{\text{error}} \\P_{\text{het}} &= .5 - P_{\text{error}}\end{aligned}$$

The likelihood of a homozygous genotype AA given an observed base frequency $F = \{F_A, F_C, F_G, F_T\}$ can be expressed as:

$$L(\text{AA} | F, P_{\text{error}}) = P_{\text{hom}}^{F_A} * P_{\text{error}}^{(F_C + F_G + F_T)}$$

The likelihood of a heterozygous genotype CT given an observed base frequency F can be expressed as:

$$L(\text{CT} | F, P_{\text{error}}) = P_{\text{het}}^{(F_C + F_T)} * P_{\text{error}}^{(F_A + F_G)}$$

The genotype, G, is assigned using maximum likelihood, and a log (base 10) odds ratio is calculated to aid in sorting.

$$\begin{aligned}G_{\text{max}} &= \text{argmax} \{L(G | F, P_{\text{error}})\} \\ \text{Log Odds} &= \log (L(G_{\text{max}} | F, P_{\text{error}}) / (1.0 - L(G_{\text{max}} | F, P_{\text{error}})))\end{aligned}$$

If the Log Odds are undefined because of machine numeric representation limitations, then the log odds are capped at 10^6 .