Suppose we are interested in identifying the genotype

$$G \in \{AA, AC, AG, AT, CC, CG, CT, GG, GT, TT\}$$

of an individual at a given position p of the reference genome. Let $R \in \{A, C, G, T\}$ be the nucleotide in the reference genome at that position. Suppose that we have n reads overlapping this position and let $D = D_1, D_2, ...D_n$ be the n nucleotides observed aligned to position p. The question we are interested in is: which of the 10 genotypes is most likely at position p, given the knowledge we have about R and D? In reasoning about this, we will need to keep in mind that sequencers can make mistakes in calling bases. Let us assume that all base calls have the same probability ϵ of error (this is not very accurate, but if we wanted we could instead make use of the q-values of the reads to modify this).

To answer this question, we need to calculate, for each possible choice of genotype G, the posterior probability of G, given the reference base R and the observed data D. Remembering Bayes rule, we get

$$\begin{aligned}
\Pr[G|D,R] &= \Pr[G,D,R]/\Pr[D,R] \\
&= \Pr[D|G,R] \cdot \Pr[G,R]/\Pr[D,R] \\
&= \Pr[D|G,R] \cdot \Pr[G|R] \cdot \Pr[R]/(\Pr[D|R \cdot \Pr[R]) \\
&= \Pr[D|G,R] \cdot \Pr[G|R]/\Pr[D|R] \\
&= \Pr[D|G] \cdot \Pr[G|R]/\Pr[D|R]
\end{aligned}$$

where the last step (dropping R from the condition) is possible because, given G, D is conditionally independent of R. More intuitively, if you tell me what G is, the knowledge of R does not affect the probability of D.

How do we calculate each of the three terms in this equation? Let's look at them one by one. First, since we assume that reads are generated independently from one another, we get:

$$\Pr[D|G] = \Pr[D_1, D_2, ..., D_n|G] = \prod_{i=1}^{n} \Pr[D_i|G]$$

So what is $\Pr[D_i|G]$? This is the probability that read i has base D_i, given that the true genotype is G. Let's first look at the case where G is homozygous xx (where x is any of the four nucleotides). Then either $D_i = x$ (there was no sequencing error) or $D_i \neq x$ (there was a sequencing error). Thus, $\Pr[D_i = x|G = xx] = (1 - \epsilon)$, and $\Pr[D_i = y|G = xx] = (1 - \epsilon)$ for any $y \neq x$. If G is heterozygous xy, then $\Pr[D_i = x|G = xy] = (1 - \epsilon)/2 + (\epsilon/3)/2$. Similarly, $\Pr[D_i = y|G = xy] = (\epsilon/3)/2 + (1 - \epsilon)/2$. Finally, $\Pr[D_i = z|G = xy] = (\epsilon/3)/2 + (\epsilon/3)/2 = \epsilon/3$.

Now, let's look at how to calculate $\Pr[G|R]$, the prior probability of the genotype, given that the base in the reference genome is R. First, $\Pr[G = xx|R = x]$ is the probability of a homozygous-reference base. This will depend on the species and the population the sample comes from, but let's set this probability to 0.999. The probability of a heterozygous-reference site, $\Pr[G =$

$xy|R = x]$, also depends on the same factors; let's set it to 0.0008. Finally, let's set the probability of homozygous non-reference genotypes to $\Pr[G = xy|R = x] = 0.0002$. We will assume that the probability of tri-allelic sites is zero: $\Pr[G = xy|R = z] = 0$.

The last term to calculate is $\Pr[D|R]$. Because this term is independent of the genotype G, it just acts as a scaling constant for the numerator $\Pr[D|G] * \Pr[G|R]$. This means that we don't really need calculate it, and instead we can just obtain the desired probability $\Pr[G|D, R]$ by setting it to $\Pr[D|G] * Pr[G|R]$ and then normalizing the numbers so that $\sum_G \Pr[G|D, R] = 1$. In other words, define $p[G|D, R] = \Pr[D|G] * Pr[G|R]$, then $\Pr[G|D, R] = p(G|D, R)/\sum'_G p(G'|D, R)$.

In conclusion, we can use the math above to calculate, for each of the 10 possible values of the genotype, its posterior probability, given the observed read data and the reference base. We would then report the most likely genotype.