SolSNP

August 15, 2010

1 Some preliminaries on Kolmogorov-Smirnov statistic

Kolmogorov-Smirnov statistic is a distance measure that determines whether two empirical distributions are same.

For a given set X and a subset $A \subset X$, the characteristic function

$$\chi_A : X \to \mathbb{R}$$

is defined by

$$\begin{aligned} \chi_A(x) &= 1 \text{ if } x \in A \\ &= 0 \text{ otherwise} \end{aligned}$$

We denote the characteristic function of the interval $(-\infty, \alpha), \alpha \in \mathbb{R}$ by χ_α for brevity. For n observations of a random variable X the cumulative distribution function F^X is given by

$$F^X(\alpha) = \frac{1}{n} \sum_{x \in X} \chi_\alpha(x) \text{ where } x \in X$$

Given two random variables X and X' with n observations the Kolmogorov statistic is given by

$$D(X, X') = sup_\alpha |F^X(\alpha) - F^{X'}(\alpha)|$$

2 Adapting Kolmogorov-Smirnov statistic for SolSNP

Given a sampling distribution X_S with a pileup length of n then we consider the expected distributions X_{ref} and X_{nonref} in case of a haploid genome and

an additional X_{het} in case of diploid. These expected distributions have same length as X_S.

To adapt the above statistic to determine a variant we define the function f_α to be

$$f_\alpha(x) = (1 - p_\epsilon(x))\chi_\alpha(x)$$

where $x \in X$ and $p_\epsilon(x)$ is the probability of error in the base call as indicated by the quality score.

We replace χ_α by f_α in the definition of F^X. We use the same pileup of quality scores to determine F^{X_i} for $i \in I = \{ref, nonref, het\}$. Let D_i denote $D(X_S, X_i)$ for $i \in I$ then we define

$$D = min\{D_i, i \in I\}$$

If D corresponds to X_{nonref} or X_{het} on both the strands then a variant is determined. The associated confidence score is then

$$-10 \log_{10}(\frac{D_F + D_R}{2})$$

where D_F and D_R are the D values from forward and reverse strand respectively.