Designing Meaningful Measures of Read Length for Data Produced by DNA Sequencers

David O. Nelson and Jane Fridlyand

Abstract

Nearly everyone uses "the number of Q20 bases" as a rough measure of the effective length of a given DNA sequence produced by the base-caller PHRED. This metric simply counts the number of bases in a read in which the PHRED quality score is at least 20. While the number of Q20 bases is a simple, easy to implement rule-of-thumb, it does not have much else going for it: it consistently underestimates the number of usable bases in the read. In this short paper, we develop and evaluate an alternative metric that uses more of the PHRED quality data in a read to predict how many bases from that read would make it into the eventual consensus sequence of an assembly. The metric was developed by evaluating a set of pre-existing, high-quality assembled contigs. The resulting predictor is a simple function of the histogram of PHRED quality values already produced by sequencing software and performs nearly as well as a more complex additive model that uses regression splines.

Keywords: DNA read length; genomics; predicting progress; PHRED; PHRAP

1 Introduction

Large-scale genome sequencing projects have become increasingly common over the last fifteen years. Many recent papers, starting with Lander and Waterman in 1988 [6], have described mathematical models for predicting the progress of such sequencing projects. These different "Lander-Waterman" analyses arise in response to different approaches to sequencing large genomes. They model the sequencing process as a coverage process like those described by Hall [5] and derive predictions of mean coverage, depth, expected number of gaps, and the like, as a function of the number of clones sequenced N, the genome size G, and the length of sequence L obtained from an individual clone chosen for sequencing. These predictions are then used to estimate the number of clones required to obtain an assembled genome to a given depth or coverage. Conversely, statistics on coverage and read length gathered during the sequencing effort are used with these models to track progress, detect problems, and refine estimates of the remaining work required.

The approximate genome size G can be determined in advance, and number of clones sequenced N is easy to obtain from daily production statistics. However, what