Heterogeneity Detector: finding heterogeneous positions in Phred/Phrap assemblies

W.H. Lee and V.B. Vega

Genome Institute of Singapore

ABSTRACT

Summary: A modification to phred and program to detect heterogeneous positions, which is particularly useful in the identification of mutations and other abnormalities in phred/phrap genome assemblies.

Availability: The package is made available at http://giscompute.gis.a-star.edu.sg/~charlie/DHetero.html

Contact: leewhc@gis.a-star.edu.sg

Phred is a base-calling software that reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files. The processed traces are displayed in the form of chromatograms consisting of four curves, each representing the signal for one of the four bases. The phred base-caller uses a four-phase procedure to determine a sequence of base-calls from the processed trace. This procedure involves the determination of predicted peaks, identification of observed peaks, comparison of the predicted and observed peaks and lastly, the insertion of unmatched peaks. It then assigns an error probability to each called base with an accuracy that is shown to be higher than that of the ABI base-caller (Ewing et al., 1998). The output files produced by Phred are then used by Phrap for assembly. Although the resultant consensus sequence is highly accurate, it is inadequate in reporting the existence of heterogeneous positions. We define heterogeneous positions to be positions which have strong signals of two or more bases that are not attributed to noise. Thus, a position with a primary signal representing the base ‘A’ is a heterogeneous position if it also has one or more relatively strong non-primary signals representing other bases, eg. base ‘T’. This observation can be particularly useful in identifying mutation sites or co-existence sites of different strains in a genome assembly, as demonstrated in our sequencing and assembly of the SARS virus recently (Ruan et al., 2003; Vega et al., in submission; Liu et al., in submission).

The Severe Acute Respiratory Syndrome (SARS) was first reported in November 2002. SARS is caused by a newly identified coronavirus named SARS-CoV (Kaizek et al., 2003; Marra et al., 2003; Rota et al., 2003). A RNA virus like SARS-CoV is known to have a high recombination and mutation rate. We had a unique opportunity to analyze a substantial number of SARS-CoV strains. Using capillary sequencing, traces for 17 SARS sequences were produced. The ABI trace files are then used as input for our modified version of Phred. The modified version of Phred gives the same output as the original Phred except that in addition, it extracts and stores the signal values of all positions for each trace file in a directory called “trace_dir”. The data stored in this directory will provide the basis for the analysis in our detector program.

The detection of heterogeneous positions by inspection of the Phrap consensus sequence displayed by Consed (Gordon et al., 1998) can be extremely tedious and impractical in genomes larger than a few thousand base pairs, eg SARS-CoV. This led to the development of our heterogeneity detector program, “DetectHetero”, to automatically generate a complete set of heterogeneous positions for an assembly.

“DetectHetero” has three selection phases. The first phase identifies all positions in all component traces in an assembly that exhibit the heterogeneity property. These chosen positions should be of high quality and have low noise. Unfortunately, imperfections of sequencing reactions, of gel electrophoresis and of trace processing will lead to problems in trace signals (Ewing et al., 1998a). In order to produce a sequence as accurate as possible despite of the data problems, Phred assigns a quality value, q, to each called base where p is the estimated error probability (Ewing and Green, 1998):

\[ q = -10 \times \log_{10}(p) \]

Using this quality score, we could eliminate positions which have a high error probability. In our program, users can specify the quality cut-off for the position selection. In our experiments using the SARS traces, we used a quality cut-off of 20. This would result in our selected positions to have at most a probability of 0.01 of being incorrect and effectively eliminate positions displaying the heterogeneity property due to sequencing errors from being selected.

The second phase involves a more stringent selection process from the list of all possible high quality heterogeneous positions selected through the first phase. It eliminates positions with noisy signals and selects the most likely heterogeneous positions from the initial list, thus providing a priority for scientists as to which positions would be most worth investigating first.
Noisy positions are determined by two factors. The first factor is the ratios of the non-primary peaks to the primary peak. The ratio of a non-primary peak to the primary peak should exceed a certain threshold, $t$, for it to be deemed as non-noise generated. The second factor is the localized average background noise percentage in the local region of the position. We define the localized average background noise percentage, $labnp$, to be:

$$labnp = \left( \frac{1}{w \cdot 2w} \sum_{i=x-w,i \times 1}^{x+w} \left( \frac{1}{w} \sum_{j=1}^{i} \left( \frac{h_{ij} \times 100}{p_{i}} \right) / 3 \right) \right) / 2w$$

where $x$ is the target position in the trace, $w$ is the number of adjacent positions we take into account on each side of the target position, $p_i$ is the primary peak value of position $i$ and $h_{ij}$ are the non-primary peak values of position $i$. Thus, a high $labnp$ value would indicate a high background noise level in the localized region surrounding the target position and vice versa. Therefore, a position is deemed as a non-noisy heterogeneous position iff one or more of its non-primary peak values exceed both $t$ and $labnp$. In our experiments, we set $t$ to be 0.5.

The third phase computes the number of traces in which a candidate position possesses heterogeneity properties. The intuition behind this is that if heterogeneity properties are present in many traces for the same position, then it would eliminate isolated abnormalities that may result from making a conclusion based on just a single trace. In this way, our program gives a higher priority to heterogeneous positions having multiple supportive traces over those which only have one supporting trace.

Experiments were done by applying the modified version of Phred and “DetectHetero” to all 17 sequences of SARS-CoV assembled in our institute. The heterogeneous positions generated by our program facilitated scientists to carry out wet lab experiments on the more interesting regions in the SARS-CoV virus. It takes hours or even days to look through all contigs and bases in a given 30kb SARS assembly using consed to identify all heterogeneous positions manually but our program took only seconds to do that. Moreover, it will not miss out any position. Ambiguities in our SARS-CoV assemblies were clarified by mass spectrometry validation (Liu et al., in submission). For example, we discovered that some strains of the SARS-CoV virus are actually the composition of two coexisting strains (Vega et al., in submission).

As shown in our experiments, heterogeneous positions are often the key regions to start with in an investigation of the characteristics and other interesting properties of a genome, in our case, the SARS-CoV virus. Thus, it is important that we find these positions accurately and in a scalable manner.

Our modification to Phred and program, DetectHetero, has shown to fulfill this criterion to find all high-quality and low noise heterogeneous positions in a genome assembly. This would greatly save time and cost in identifying the more interesting regions in a genome assembly to begin wet lab experiments with.

ACKNOWLEDGEMENTS

We would like to thank the sequencing department of the Genome Institute of Singapore for the trace data used in evaluation. Special thanks to Dr J.J. Liu for his efforts in validation of the heterogeneous positions.

REFERENCES


