Finding Structural Variations with Pair-End Sequences and a Sliding Window [1]

October 18, 2009

Structural variations are sequence differences between a sample genome and a reference genome. Variations include insertions, deletions, and translocations. To find these variations we use pair end tags and a sliding window. Pair end tags compare the local progression of the two sequences, and the sliding window aggregates local progressions to detect variations.

1. Sequence sample paired tags

   In this step we sequence a sample for structural analysis. The result is a set of sequence pairs, where the distance between elements of each pair is similar.

   $$\begin{gathered} S = \{(s_1, s_1'), (s_2, s_2'), (s_3, s_3'), \dots\} \\ \text{where } (\&s_i - \&s_i') \sim (\&s_j - \&s_j') \\ \text{for every } (s_i, s_i'), (s_j, s_j') \in S \end{gathered}$$

   Note: The method used to sequence the pairs guarantees the distances are similar even known the address of each end ($\&s$ and $\&s'$) is not known.

2. Map pair ends to the reference genome

   We map each element of each pair to a reference genome. The result of each mapping is a set of addresses in the reference genome.

Sample	Reference
$s_1 \rightarrow$	$\{\&r_a, \&r_b, \dots\}$
$s_1' \rightarrow$	$\{\&r_i, \&r_j, \dots\}$
$\vdots$	$\vdots$

   This mapping is used to compare small sequences in the sample to the reference. However, when either element of a sample pair maps to multiple addresses in the reference, we have no way to distinguish between the different possible pairings. For this reason, we only keep pairs when both elements map to only one address.

   Considering this filter, we have a mapping from each sample pair to a pair of addresses in the reference.

   $$\begin{gathered} (s_1, s_1') \rightarrow (\&r_i, \&r_j) \\ (s_2, s_2') \rightarrow (\&r_k, \&r_l) \\ \vdots \end{gathered}$$

3. Quantify *normal* pair distances

   Real that all sample pairs $(s, s') \in S$ have similar distances $(\&s_i - \&s'_i)$. Given this, insertions and deletions can be characterized by sample pairs that map to reference pairs whose distance $(\&r - \&r')$ is larger than normal and smaller normal distance, respectively.

   Determining normal requires some subjective judgments. The distribution of the reference pair distances is found, and a cutoff is made to the right and left of the mean. Any pair with a distance inside this cutoff is normal, otherwise it is *aberrant.*

4. Classify *aberrant* tags by distance

   Short reference pairs are classified as deletions, and *long* reference pairs are classified as insertions.

5. Identify portions of the genome with many aberrant tags

   The address of each end of an *aberrant* pair is annotated along the reference sequence. A *sliding window* is then used to scan the sequence for clusters of aberrant pair ends. A windows with are significant number of aberrant pair ends is called a *high density window.*

6. Identify aberrant linkages

   To complete the analysis we would like to know what *type* of insertion or deletion exists in that sample. That is, which sequence was inserted or deleted. To do this we must *follow* each high density window.

   Within a high density window there are a number of pair ends whose matching pair is in another window. It is likely that some pairs in a window disagree as to which window they should map to. Again, we cannot determine which tag is *right*, so the majority rules.

   The window with the highest incidence is determined to be the matching window. The sequence information of these *aberrant linkages* is used to analyze the impact of the given variation. Ideally we can find causes of cancers and other diseases from such an analysis.

References

[1] Y. Shibata, A. Malhotra, S. Bekiranov, A. Dutta, Yeast genome analysis identifies chromosomal translocation, gene conversion events, and several sites of Ty element insertion, *Nucleic Acids Resarch*, Aug 2009.