Supplementary Notes for

VirusSeq: Software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue

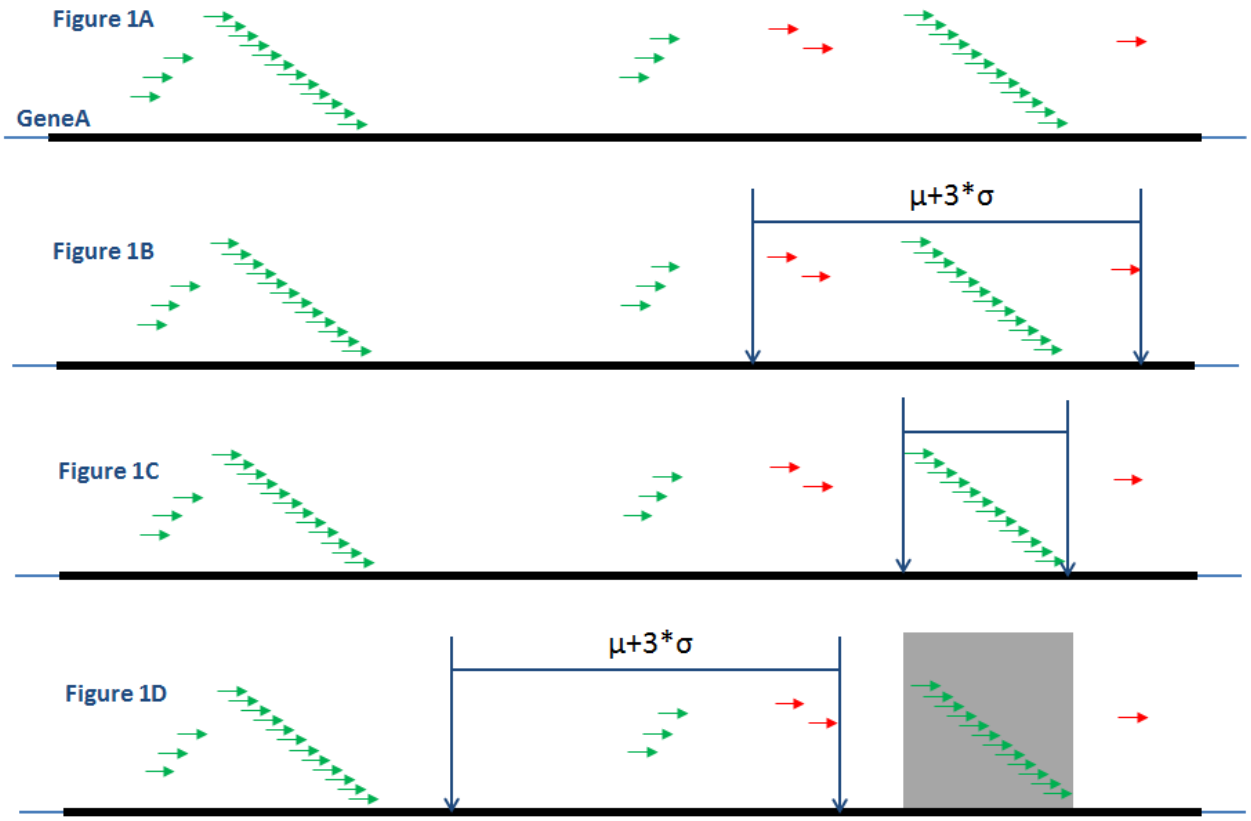
1 Dynamic Clustering Procedure

VirusSeq clusters the discordant read pairs that support the same integration (fusion) sites (e.g., HBV-MLL4) by implementing a dynamic clustering procedure.

We illustrate our dynamic clustering procedure in a schematic (Figure 1A-D). In this procedure, each gene is a unit, and the boundary of each gene is also extended in order to cover the entire region of human genome (lots of virus integration sites are located at intergenic regions). Specifically, all intergenic regions are annotated by the adjacent genes either as a gene's 5-prime region or a gene's 3-prime region based upon the distance to gene's start and end location in refSeq annotation.

In Figure 1A, each small arrow represents a discordant read, which was forwardly mapped to GeneA. For this set of forwardly-aligned discordant reads on GeneA, our clustering procedure assumes that the genomic position of the most-right discordant read is the start boundary of the first cluster on GeneA, and the size of this cluster is initially defined as the library insert size (fragment length) plus three standard deviations (Figure 1B). Within this cluster, the outliers of discordant reads, illustrated by the red arrows, exist. In order to remove them, we implemented the robust "extreme studentized deviate (ESD)" multiple-outlier detection procedure. For instance, in Figure 1B, the ESD procedure detects one discordant read (shown in red) from the most right end of the cluster as an outlier, which is removed from the further consideration. Meanwhile, the ESD procedure also detects two discordant reads (shown in red) from the left end of the cluster as outliers, which are excluded from this cluster. Once no more outliers can be detected in this cluster, the dynamic clustering procedure re-sets the boundary for this cluster (Figure1C). Next, a new cluster (Figure 1D) starts at the read adjacent to the left end of the previous cluster. The same procedure as above is applied until no more clusters can be detected on GeneA.

For the reversely-aligned discordant reads, the clustering process starts with the most-left mapped discordant read, and the genomic coordinate for the most-left read is used to define the start boundary of the discordant read cluster with the same outlier detection/removal procedure. For either side of the candidate fusion partner (gene or virus), the clustering process is performed independently.



Schematic of dynamic clustering