

Case Study #4: Gene-Trap Sequence Tags Processing

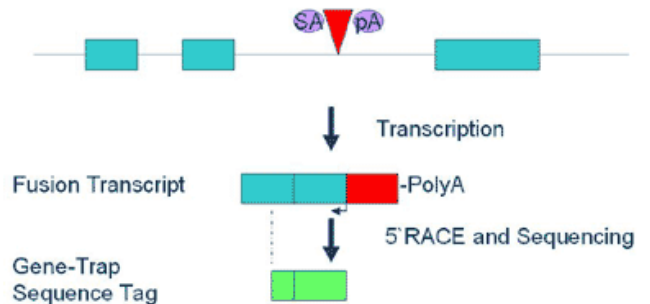
In this study short tags are mapped against the human genome. Using a tool like BLAST invariably generates huge numbers of random hits that take up valuable disk space and subsequently need to be parsed-out using specialized scripts. Vmatch solves the problem by not generating the random hits in the first place!

This example is representative of gene-trap sequence tag processing, as depicted in the scheme shown in the figure below (or in similar approaches). Genome mapping of the sequenced tags will either identify annotated exons or may indicate bona fide new genes or novel transcript isoforms. This process is a simple problem if tag sequences are long (greater than 40 bp), because most long tags can typically be uniquely matched to the genome even with polymorphisms taken into account. However, long tags are prohibitively expensive, and thus it is expedient to use shorter tags in combination with computationally advanced tools to solve the matching task.

The mapping of short tags can be problematic, particularly if the tags are broken up by long introns. Vmatch solves this problem by easily linking-in user defined match-constraints which get executed while the search is actually executing (versus having to run post-process scripts to parse-out the unwanted random hits that get generated in the search). In our particular case we want to impose restraints such that: 1) genomic matches end exactly at the 3'-terminus of the tag; 2) the genomic sequence downstream of the match starts with the consensus intron donor dinucleotide GT (98% of introns); and 3) we are selecting only the "best" match. All these criteria may be relaxed, of course, if higher sensitivity is more important than high specificity.

We take advantage of one other Vmatch feature to improve overall performance. It is likely that within our set of tags there will be many duplicates resulting from transcripts of highly expressed genes. Duplicate sequences will only slow down the overall sequence matching process. So we use the **-nonredundant** parameter in Vmatch to extract all of the **unique** tags within our original collection, discarding the duplicates. We then match these unique tags against the entire human genome.

A. Gene-Trap Sequence Tag



B. Gene-Trap Sequence Tag Mapping

