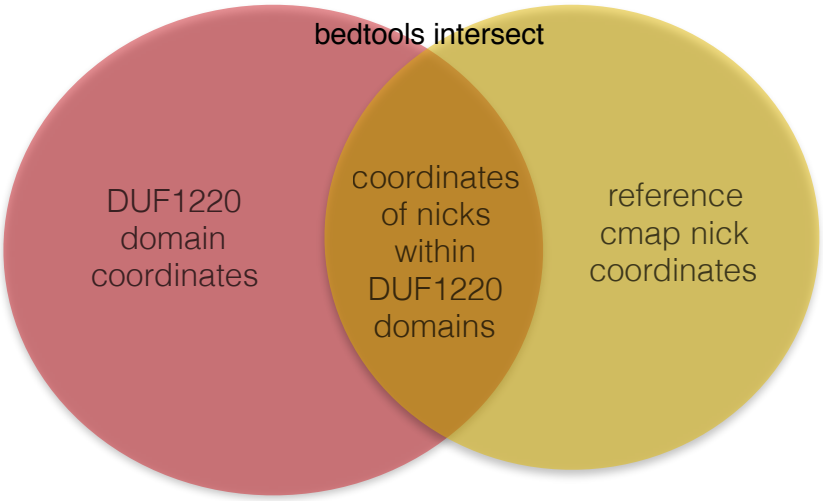
Irys DUF1220 SV Caller
version 1

Documentation - 12/23/2016

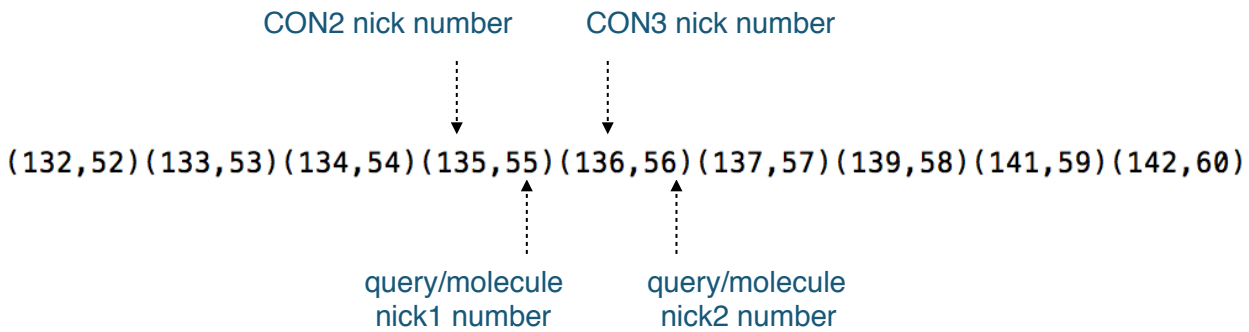
1. Identifies base pair positions of reference nicks of interest



2. Identifies the reference nick *number* of the reference nick of interest (based on bp position) (Uses reference cmap file).

```
#h CMapId       ContigLength    NumSites        SiteID  LabelChannel    Position        StdDev  Coverage        Occurrence
#f int  float   int     int     int     float   float   float   float   float   float
   1    248956422.0     28429   1       1       17623.0 0.0     6       6
   1    248956422.0     28429   2       1       19742.0 0.0     7       3
```

3. Identifies molecules aligned to the two reference nick numbers of interest for each gene *and* identifies the query/molecule nick numbers aligned to those reference nicks (Uses the alignment string in the xmap)

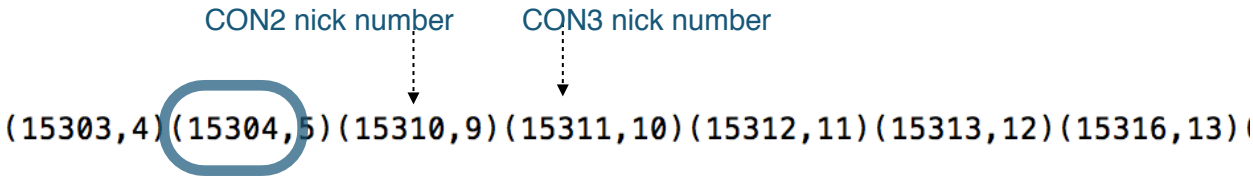


4. For each molecule with an alignment to both reference nicks of interest, calculates the distance between the two relevant query nicks *on the molecule*.

Filters out: Molecules with a confidence score less than 8, and molecules where alignments adjacent to the nicks of interest are not consecutive (checks to the left and right of the molecules of interest in both the reference and the query).

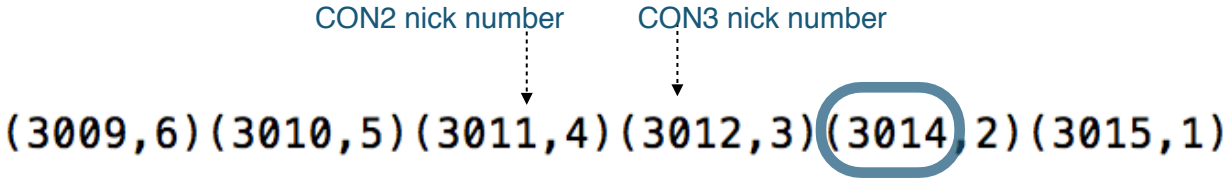
- Since false negatives or false positives/novel nick sites in the query could produce this condition, molecules are retained (with a flag) if they meet certain conditions

Filtered out



The circled reference nick is not adjacent to the CON2 nick — 6 reference nicks are skipped

Retained with "S" flag (to indicate a nick was skipped)



Although the circled nick is not adjacent to the CON3 nick, only one reference nick is skipped, and the next nick, "3015" is adjacent to to 3014. This condition could be created by a false negative in the molecule. This molecule is retained with a note.

6. Calls “peaks” by stringing together/bins molecules where the distance between the nicks of interest is less than 1kb apart.

Molecule ID	Distance b/w nicks on molecule	
1001	5000	Each of these molecules has a distance b/w nicks that is less than 1kb different from another molecule. These molecules merged into one “bin” that ranges from 5000-6800.
2232	5200	
246	5900	
5003	6800	
210	8700	These molecules are more than 1kb different from the other “bin” and less than 1kb different from one another, so they form their own bin.
472	9500	

7. Calls structural variations. Compares the distance between nicks in the reference and the distance between nicks in the molecules (the median distance is used for binned molecules). The molecules are called as “reference” if the differ from the reference distance by less than 2kb, and are called as “insertion” or “deletion” if they are more than 2kb larger or smaller than the reference distance.

Criteria for molecules to be considered: (1) a bin with at least two molecules in it, (2) a single molecule that has good alignments both 1 nick to the left and 1 nick to the right of the nicks of interest.