**Steps to generate long read files suitable for isoform prediction or fusion detection**

1. Generate a FASTA file of all sub reads with > .75 accuracy

**SMRT analysis package**

2. Generate a FASTA file of all ccs reads with > .95 accuracy

**SMRT analysis package**

3. Generate a FASTA file of all ccs reads with > .9 accuracy

**SMRT analysis package**

4. Get the longest sub read from each molecule. Use the outputs from **step #1** as an input.

**Au-public/iron/utilities/  
 pacbio\_get\_longest\_fasta\_per\_molecule.pl <sub75fasta>**

5. Construct a set of reads that excludes any ccs reads with > 0.95 accuracy but that includes ccs reads with greater than 0.9 accuracy and less than 0.95 accuracy. And any remaining > 0.75 longest sub reads not yet considered. Use the outputs from **step #2**, **step #3**, and **step #4** as inputs.

**Au-public/iron/utilities/  
 pacbio\_make\_ccs90-95\_sub75\_set.py <cc95fasta> <ccs90fasta> <sub75fasta>**

6. Perform LSC on the FASTA output from **step #5**. Subsequent steps will use the following outputs:

**corrected\_LR.fa**

**full\_LR.fa**

7. Replace **corrected\_LR.fa** entries with **full\_LR.fa** entries when the length of the corrected is 90% or greater the length of the **full\_LR.fa**. The purpose is to maintain the adaptor sequences when possible.

**IDP/utilities/**

**replace\_LSC\_corrected\_with\_full\_when\_similar\_length.py <full\_LR.fa> <corrected\_LR.fa> <threshold (i.e. 0.9)> <output fasta> <output list>**

8. Assemble a non-redundant set of reads for fusion detection that includes the high quality ccs reads from **step #2**, the reads that were used as an input to LSC (output of **step #5**), this allows us to recover any reads that were not operated on by LSC, and finally, the swapped **FASTA output from step #7**

**Au-public/iron/utilities/**

**assemble\_IDP-fusion\_read\_set.pl <ccs95 fasta> <pre-LSC fasta> <LSC swapped fasta>**

9. Assemble a set of reads for isoform prediction that includes both the corrected\_LR.fa and full\_LR.fa entries in cases when the length of the corrected is within 90% the length of the full\_LR.fa. This introduces some redundancies that will be removed during the execution of IDP isoform prediction since actual quantifications will be based on the short read counts. Similar to **step #8**, this step requires the high quality ccs reads from **step #2**, the reads that were used as an input to LSC (output of **step #5**), the full\_LR.fa output from LSC in **step #6**, the swapped **FASTA output from** **step #7**, and the **list output from** **step #7**.

**Au-public/iron/utilities/**

**assemble\_IDP-isoform\_read\_set.pl <ccs95 fasta> pre-LSC fasta> <full\_LR.fa> <LSC swapped fasta> <LSC swapped list>**