

# Jackson Labs Long Read Sequencing Workshop

Informatics Demonstration

Preparation Work (already done for today’s demonstration, for your reference)

1. Install SMRT Tools

**mkdir smrtlink\_tools\_only**

**cd smrtlink\_tools\_only**

**wget** [**https://downloads.pacbcloud.com/public/software/installers/smrtlink\_5.1.0.26412.zip**](https://downloads.pacbcloud.com/public/software/installers/smrtlink_5.1.0.26412.zip)

**unzip smrtlink\_5.1.0.26412.zip**

**./smrtlink\_5.1.0.26412.run --rootdir smrtlink --smrttools-only**

1. Source for publicly available datasets (**DO NOT DOWNLOAD DURING EXERCISE**)
   1. NA12878 PacBio data (for SV analysis):

<https://www.pacb.com/blog/identifying-structural-variants-na12878-low-fold-coverage-sequencing-pacbio-sequel-system/>

* + 1. (Follow the download link, and retrieve from DNA Nexus)
  1. Human UHRR + Lexogen SIRV synthetic spike in (for IsoSeq2 walkthrough)

<https://github.com/PacificBiosciences/IsoSeq_SA3nUP/wiki/Iso-Seq-in-house-datasets#human-uhrr--lexogen-sirv-synthetic-spike-in-released-june-2017>

Setup: Set up command line environment

**PATH=$PATH:~/smrtlink\_tools\_only/smrtlink/smrtcmds/bin**

Part I: SV analysis using pbsv

1. Prepare reference

**gunzip hs37d5.fa.gz**

1. Generate config file and edit

**pbsv generate\_config > my\_config.cfg**

**tmp\_dir=/home/ubuntu/NA12878**

1. Run SV alignment using NGM-LR (pbsv align)

**pbsv align --cfg\_fn my\_config.cfg hs37d5.fa NA12878.reads.ngm.bam aligned\_NA12878.bam**

1. Run SV calling using pbsv call

**pbsv call –reference\_regions chr21: 100-200000 hs37d5.fa aligned\_NA12878.bam chr21\_100\_200000\_NA12878\_svcalls.bed**

Part II: Iso-Seq2 analysis

1. Get tiny dataset (demo of bamsieve - https://github.com/PacificBiosciences/pbcoretools)

**curl -L https://www.dropbox.com/sh/m64cvsrqqd7acj7/AAAjl0x-RRTssEzW8cu7oZNqa?dl=1 > download.zip**

**unzip download.zip**

1. Prepare dataset for analysis

**dataset create --type SubreadSet --generateIndices --name human\_SIRV\_control\_RNA\_tiny sirv\_rna\_tiny.subreadset.xml iso\_seq\_tiny.bam**

**dataset summarize sirv\_rna\_tiny.subreadset.xml**

1. Generate CCS

**ccs --minPasses 0 --minPredictedAccuracy .8 --minIdentity 0 --minZScore NaN --noPolish --numThreads 4 sirv\_rna\_tiny.subreadset.xml sirv\_rna\_tiny\_ccs.xml**

1. Iso-Seq2 classify command (generate list of full length iso-forms)-

**pbtranscript classify --flnc isoseq\_flnc.fasta --nfl isoseq\_nfl.fasta -d classifyOut --cpus 4 --min\_seq\_len 100 isoseq\_draft.fasta**