1. **Create bam files**

* **Align fastq files**
  + bwa mem [reference genome file] [fastq files] > [sam file]
* **Convert sam file to bam file**
  + samtools view -S -h -b [sam file] > [bam file]
* **Sort bam file**
  + samtools sort [bam file] > [sorted bam file]

1. **Find insert size**

* **Script and instructions for finding insert size for bam/sam files can be found at**
  + <http://allaboutbioinfo.blogspot.com/2012/04/estimating-paired-end-read-insert.html>
  + Script can also be downloaded from https://github.com/mam288/JonesLab.git
* **Run the script on the bam file (per instructions)**
  + samtools view [sorted bam file] | python3 getinsertsize.py –

1. **Find variants using Pindel**

* **Download and install Pindel**
  + <http://gmt.genome.wustl.edu/packages/pindel/user-manual.html>
* **Create a config file (per Pindel instructions)**
  + File should contain [bam file name] [insert size] [line name]
  + File contents should look like:

../Talia\_sequences/net-cn\_bwa\_dm6.sort.bam 381 net-cn

../Talia\_sequences/wt\_bwa\_dm6.sort.bam 382 wt

CL100070073\_L1\_B5RDDROawkRAAAAA-501\_bwa\_dm6.sort.bam 236 501

CL100070073\_L1\_B5RDDROawkRAAAAA-502\_bwa\_dm6.sort.bam 237 502

CL100070073\_L1\_B5RDDROawkRAAAAA-503\_bwa\_dm6.sort.bam 236 503

* **Run Pindel on bam files**
  + ~/pindel/pindel -f [reference genome file] -i [config file] -c ALL -o [output file name]

1. **Filter Pindel Results**

* **Use awk to filter Pindel\_D output file**
  + awk ' ((int($3)>=5) && ($2=="D") && (int($30)==1) && (int($16)>=5) && ((int($49)>=1) || (int($51)>=1) || (int($42)>=1) || (int($44)>=1) || (int($35)>=1) || (int($37)>=1))) { for (i=1;i<=66;i++) {printf ("%s ",$i)} }' [pindel\_D file] > test.txt
  + (int($3) >= 5)
    1. Column 3 of the Pindel\_D file is the variant length
    2. Filters for all variants with a length >= 5
  + ($2=="D")
    1. Column 2 of the Pindel\_D file is the type of variant
    2. Filters for all deletions
  + (int($30)==1)
    1. Column 30 of the Pindel\_D file is the number of samples supporting the variant
    2. Filters for all variants supported by one sample
  + (int($16)>=5)
    1. Column 16 of the Pindel\_D file is the number of readssupporting the variant
    2. Filters for all variants supported by at least 5 reads
  + ((int($49)>=1) || (int($51)>=1) || (int($42)>=1) || (int($44)>=1) || (int($35)>=1) || (int($37)>=1)))
    1. Column numbers correspond to number of supporting reads in each offspring line.
    2. Filters for variants supported by one of the offspring lines (as opposed to parent lines)