FILE RENAMING

1. Download data to MacPro into /Volumes/NG_Sequence_data/original_data_files/
    a. BaseSpace (used by MOGene) has a Download application (I understand they may be trying to change the way they store data, but until then some extra steps are needed here):
        i. Files should be downloaded to the ./basespace subdirectory of the above path
        ii. Files should then be moved into their own directory as:
        cmd: cd /Volumes/NG_Sequence_data/original_data_files/basespace
        cmd: mkdir <descriptive_directory_name_here>
        cmd: find ./ -name "*fastq.gz" -exec mv {}
        <descriptive_directory_name_here> \;
        for example:

```
bartlabmacpro1:basespace/ $ cd
bartlabmacpro1:~/ $ cd /Volumes/NG_Sequence_data/original_data_files/basespace
bartlabmacpro1:basespace/ $ ls ANU_NEB\ AMPLICON-16655640/1-26212291/Data/Intensities/BaseCalls/
1_S1_L001_R1_001.fastq.gz 1_S1_L001_R2_001.fastq.gz
bartlabmacpro1:basespace/ $ mkdir ./mogene_blibpool10_20150626
bartlabmacpro1:basespace/ $ find ./ANU_NEB\ AMPLICON-16655640 -name "*fastq.gz" -exec mv {} ./mogene_blibpool10_20150626 \;
```

NOTE: MOGene files need to be differentiated from each other for the pattern matching to work as shown in the screenshot here:

```
bartlabmacpro1:basespace/ $ cd mogene_blibpool10_20150626
bartlabmacpro1:mogene_blibpool10_20150626/ $ for i in *; do mv ${i} brun002_${i}; done
bartlabmacpro1:mogene_blibpool10_20150626/ $ ll
total 14684568
drwxr-xr-x  2 mawilson  staff   884B Jun 26 20:05 .
drwxr-xr-x  5 mawilson  staff   204B Jun 26 19:34 ..
-rw-r--r--  1 mawilson  staff   201M Jun 26 15:32 brun002_10_S10_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   215M Jun 26 15:33 brun002_10_S10_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   273M Jun 26 15:33 brun002_11_S11_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   295M Jun 26 15:33 brun002_11_S11_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   285M Jun 26 15:33 brun002_12_S12_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   305M Jun 26 15:33 brun002_12_S12_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   283M Jun 26 15:29 brun002_1_S1_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   302M Jun 26 15:29 brun002_1_S1_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   332M Jun 26 15:29 brun002_2_S2_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   357M Jun 26 15:29 brun002_2_S2_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   356M Jun 26 15:30 brun002_3_S3_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   379M Jun 26 15:30 brun002_3_S3_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   287M Jun 26 15:30 brun002_4_S4_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   306M Jun 26 15:30 brun002_4_S4_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   327M Jun 26 15:31 brun002_5_S5_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   350M Jun 26 15:31 brun002_5_S5_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   205M Jun 26 15:30 brun002_6_S6_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   220M Jun 26 15:31 brun002_6_S6_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   295M Jun 26 15:31 brun002_7_S7_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   315M Jun 26 15:31 brun002_7_S7_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   362M Jun 26 15:32 brun002_8_S8_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   391M Jun 26 15:32 brun002_8_S8_L001_R2_001.fastq.gz
-rw-r--r--  1 mawilson  staff   255M Jun 26 15:32 brun002_9_S9_L001_R1_001.fastq.gz
-rw-r--r--  1 mawilson  staff   275M Jun 26 15:32 brun002_9_S9_L001_R2_001.fastq.gz
```

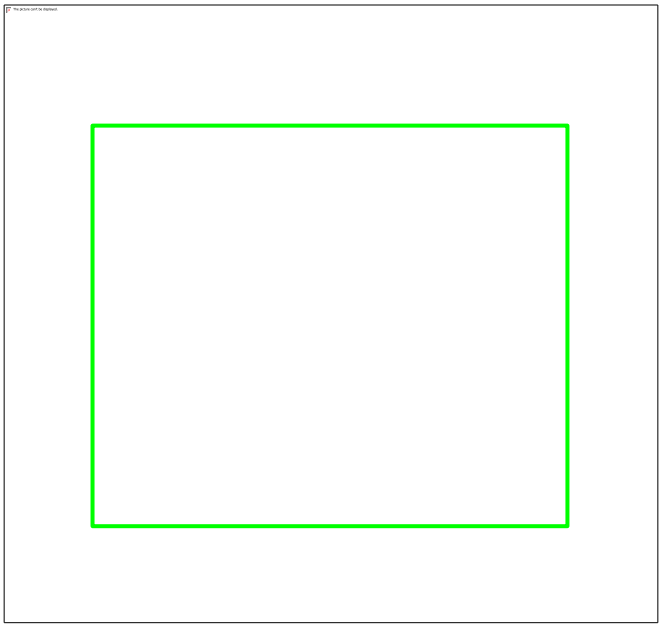
    b. GTAC is easily downloaded using wget to the above path

2. Copy files to a temporary directory on apollo such as /home/bbart/data/blib/tmp_raw_data
    for example, from apollo:
        scp blab:/Volumes/NG_Sequence_data/original_data_files/basespace/mogene_blibpool10_20150626/* /home/bbart/data/blib/tmp_raw_data

3. Create regular expression patterns for the file renaming (this part is a little weird):
    a. The patterns are stored in JSON format in /home/bbart/data/blib/match.patterns which is similar to a Python dictionary. (JSON files don't allow comments, otherwise there would be some more information in that file) The JSON is broken into two main parts:
        i. "run" - this contains the pattern to match the raw data filenames
            1. the key field of each entry is the pattern and contains two parenthesized parts that are used as keys to indicate:
                a. library
                b. paired read file indicator
            2. the data field of each entry is a list of 3 things:
                a. key indicating the data pool in the "pool" section described below
                b. string containing sequencing type and date to be inserted in the filenames
                c. list indicating order the parenthesized sections of the pattern come in since sequencing facilities do this differently
        ii. "pool" - this contains the data pools to associate blib numbers with the files. Each entry in the pools contain:
            1. key field that comes from the input filenames
            2. data field is a list of a blib number and a dictionary that associates read numbers to 1 or 2 since GTAC assigns 1 and 3 to their paired end read files
    b. New entries should be created for new datasets. Care should be taken that each pattern is different in the "run" section and that each pool indicator is different in the "pool" section:
        i. "run" section for the example used in this guide:

```
{
  "run":{
    "this_file_group_([ATCG]{6})\\.([0-9])\\.fq": ["poolnum", "hiseq012345"],
    "run_1497_s_2_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool2", "hiseq040115", [2,1]],
    "run_1499_s_2_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool2", "hiseq040215", [2,1]],
    "run_1500_s_1_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool3", "hiseq040115", [2,1]],
    "run_1500_s_2_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool3", "hiseq040215", [2,1]],
    "run_1521_s_1_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool6", "rnaseq042115", [2,1]],
    "run_1521_s_2_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool7", "rnaseq042115", [2,1]],
    "run_1535_s_1_([1,3])_withindex_sequence\\.txt_([ATCG]{6})\\.fq\\.gz": ["pool8", "rnaseq042115", [2,1]],
    "rnd1_amplicon([0-9])_S[0-9]_L001_R([1,2])_001\\.fastq\\.gz": ["amplicon_pool1", "miseq101214", [1,2]],
    "rnd2_amplicon([0-9])_S[0-9]_L001_R([1,2])_001\\.fastq\\.gz": ["amplicon_pool1", "miseq101714", [1,2]],
    "brun001_([0-9]{1,2})_S[0-9]{1,2}_L001_R([1,2])_001\\.fastq\\.gz": ["brun001", "miseq050415", [1,2]],
    "brun002_([0-9]{1,2})_S[0-9]{1,2}_L001_R([1,2])_001\\.fastq\\.gz": ["brun002", "miseq062610", [1,2]],
    "(RBS[0-9]{1,2})_[ATCG]{6}_L001_R([1,2])_001\\.fastq\\.gz": ["becky_xam_data", "oldseq000000", [1,2]]
  },
  "pool":{
    "poolnum":{
      "AATCAG":["blib000",{"1":1,"3":2}]
    },
```

ii. “pool” section for the example used in this guide:



4. Create blib directories if necessary. Example:

```
cylon0:blib/ $ for i in {131..142}; do mkdir blib${i}.xam; done
```

5. Move files using ~/code/scriptbox/bartlab_file_mgmt.py This command is meant to be run in two steps:
    a. DRY RUN: This mode is available to check that the patterns were entered correctly and files aren’t overwritten or moved improperly:
    cmd: cd /home/bbart/data/blib/
    cmd: ~/code/scriptbox/bartlab_file_mgmt.py -d ./ -r ./tmp_raw_data -p ~bbart/data/blib/match.patterns

        i. If you have forgotten to add the blib directories, this is what happens:

```
cylon0:blib/ $ ~/code/scriptbox/bartlab_file_mgmt.py -d ./ -r ./tmp_raw_data -p ~bbart/data/blib/match.patterns
DRY RUN
Error: ./blib140.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib140.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib141.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib141.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib142.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib142.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib131.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib131.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib132.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib132.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib133.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib133.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib134.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib134.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib135.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib135.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib136.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib136.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib137.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib137.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib138.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib138.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib139.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
Error: ./blib139.xam does not exist. Please check the directory you specified or create the blib directories needed for this file.
```

ii. If the Dry Run is successful:

```
cylon0:blib/ $ ~/code/scriptbox/bartlab_file_mgmt.py -d ./ -r ./tmp_raw_data -p ~bbart/data/blib/match.patterns
DRY RUN
Directory created at ./blib140.xam/raw_reads
Moving brun002_10_S10_L001_R1_001.fastq.gz to blib140.xam.miseq062610.r1.fq.gz
Moving brun002_10_S10_L001_R2_001.fastq.gz to blib140.xam.miseq062610.r2.fq.gz
Directory created at ./blib141.xam/raw_reads
Moving brun002_11_S11_L001_R1_001.fastq.gz to blib141.xam.miseq062610.r1.fq.gz
Moving brun002_11_S11_L001_R2_001.fastq.gz to blib141.xam.miseq062610.r2.fq.gz
Directory created at ./blib142.xam/raw_reads
Moving brun002_12_S12_L001_R1_001.fastq.gz to blib142.xam.miseq062610.r1.fq.gz
Moving brun002_12_S12_L001_R2_001.fastq.gz to blib142.xam.miseq062610.r2.fq.gz
Directory created at ./blib131.xam/raw_reads
Moving brun002_1_S1_L001_R1_001.fastq.gz to blib131.xam.miseq062610.r1.fq.gz
Moving brun002_1_S1_L001_R2_001.fastq.gz to blib131.xam.miseq062610.r2.fq.gz
Directory created at ./blib132.xam/raw_reads
Moving brun002_2_S2_L001_R1_001.fastq.gz to blib132.xam.miseq062610.r1.fq.gz
Moving brun002_2_S2_L001_R2_001.fastq.gz to blib132.xam.miseq062610.r2.fq.gz
Directory created at ./blib133.xam/raw_reads
Moving brun002_3_S3_L001_R1_001.fastq.gz to blib133.xam.miseq062610.r1.fq.gz
Moving brun002_3_S3_L001_R2_001.fastq.gz to blib133.xam.miseq062610.r2.fq.gz
Directory created at ./blib134.xam/raw_reads
Moving brun002_4_S4_L001_R1_001.fastq.gz to blib134.xam.miseq062610.r1.fq.gz
Moving brun002_4_S4_L001_R2_001.fastq.gz to blib134.xam.miseq062610.r2.fq.gz
Directory created at ./blib135.xam/raw_reads
Moving brun002_5_S5_L001_R1_001.fastq.gz to blib135.xam.miseq062610.r1.fq.gz
Moving brun002_5_S5_L001_R2_001.fastq.gz to blib135.xam.miseq062610.r2.fq.gz
Directory created at ./blib136.xam/raw_reads
Moving brun002_6_S6_L001_R1_001.fastq.gz to blib136.xam.miseq062610.r1.fq.gz
Moving brun002_6_S6_L001_R2_001.fastq.gz to blib136.xam.miseq062610.r2.fq.gz
Directory created at ./blib137.xam/raw_reads
Moving brun002_7_S7_L001_R1_001.fastq.gz to blib137.xam.miseq062610.r1.fq.gz
Moving brun002_7_S7_L001_R2_001.fastq.gz to blib137.xam.miseq062610.r2.fq.gz
Directory created at ./blib138.xam/raw_reads
Moving brun002_8_S8_L001_R1_001.fastq.gz to blib138.xam.miseq062610.r1.fq.gz
Moving brun002_8_S8_L001_R2_001.fastq.gz to blib138.xam.miseq062610.r2.fq.gz
Directory created at ./blib139.xam/raw_reads
Moving brun002_9_S9_L001_R1_001.fastq.gz to blib139.xam.miseq062610.r1.fq.gz
Moving brun002_9_S9_L001_R2_001.fastq.gz to blib139.xam.miseq062610.r2.fq.gz
```

b. WET RUN: Once the Dry Run is successful and you've checked that the filenames are changing as expected, you can do a real run by adding the --run flag to the command:
cmd: ~/code/scriptbox/bartlab_file_mgmt.py -d ./ -r ./tmp_raw_data -p ~bbart/data/blib/match.patterns --run

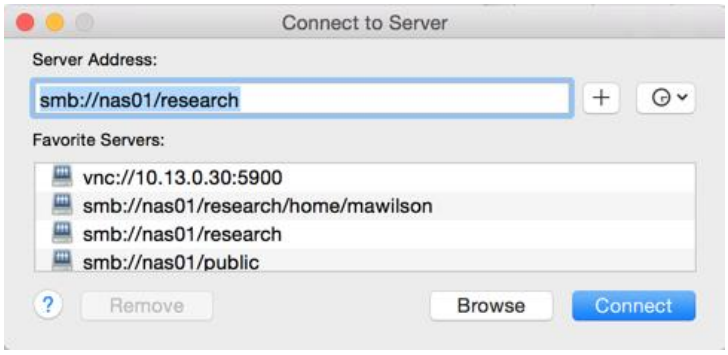
   i. Successful:

```
cylon0:blib/ $ ~/code/scriptbox/bartlab_file_mgmt.py -d ./ -r ./tmp_raw_data -p ~bbart/data/blib/match.patterns --run
Moving brun002_10_S10_L001_R1_001.fastq.gz to blib140.xam.miseq062610.r1.fq.gz
Moving brun002_10_S10_L001_R2_001.fastq.gz to blib140.xam.miseq062610.r2.fq.gz
Moving brun002_11_S11_L001_R1_001.fastq.gz to blib141.xam.miseq062610.r1.fq.gz
Moving brun002_11_S11_L001_R2_001.fastq.gz to blib141.xam.miseq062610.r2.fq.gz
Moving brun002_12_S12_L001_R1_001.fastq.gz to blib142.xam.miseq062610.r1.fq.gz
Moving brun002_12_S12_L001_R2_001.fastq.gz to blib142.xam.miseq062610.r2.fq.gz
Moving brun002_1_S1_L001_R1_001.fastq.gz to blib131.xam.miseq062610.r1.fq.gz
Moving brun002_1_S1_L001_R2_001.fastq.gz to blib131.xam.miseq062610.r2.fq.gz
Moving brun002_2_S2_L001_R1_001.fastq.gz to blib132.xam.miseq062610.r1.fq.gz
Moving brun002_2_S2_L001_R2_001.fastq.gz to blib132.xam.miseq062610.r2.fq.gz
Moving brun002_3_S3_L001_R1_001.fastq.gz to blib133.xam.miseq062610.r1.fq.gz
Moving brun002_3_S3_L001_R2_001.fastq.gz to blib133.xam.miseq062610.r2.fq.gz
Moving brun002_4_S4_L001_R1_001.fastq.gz to blib134.xam.miseq062610.r1.fq.gz
Moving brun002_4_S4_L001_R2_001.fastq.gz to blib134.xam.miseq062610.r2.fq.gz
Moving brun002_5_S5_L001_R1_001.fastq.gz to blib135.xam.miseq062610.r1.fq.gz
Moving brun002_5_S5_L001_R2_001.fastq.gz to blib135.xam.miseq062610.r2.fq.gz
Moving brun002_6_S6_L001_R1_001.fastq.gz to blib136.xam.miseq062610.r1.fq.gz
Moving brun002_6_S6_L001_R2_001.fastq.gz to blib136.xam.miseq062610.r2.fq.gz
Moving brun002_7_S7_L001_R1_001.fastq.gz to blib137.xam.miseq062610.r1.fq.gz
Moving brun002_7_S7_L001_R2_001.fastq.gz to blib137.xam.miseq062610.r2.fq.gz
Moving brun002_8_S8_L001_R1_001.fastq.gz to blib138.xam.miseq062610.r1.fq.gz
Moving brun002_8_S8_L001_R2_001.fastq.gz to blib138.xam.miseq062610.r2.fq.gz
Moving brun002_9_S9_L001_R1_001.fastq.gz to blib139.xam.miseq062610.r1.fq.gz
Moving brun002_9_S9_L001_R2_001.fastq.gz to blib139.xam.miseq062610.r2.fq.gz
```

6. Now the files have been moved to apollo and renamed according to the Bart Lab way

BACKUP ORIGINAL FILES

1. Connect the MacPro to the research drive:
   a. Open Finder
   b. Press Command-K
      i. This window should open:



   c. Under Server Address enter smb://nas01/research and hit connect

    d. You will be prompted for credentials, enter your Danforth username and password as you would to access the W3 or your email
2. Sync the original_data_files with the research drive:
    a. Open a terminal
    b. Use rsync to sync the data:
    cmd: rsync -arvh /Volumes/NG_Sequence_data/original_data_files/ /Volumes/research/bart_lab/original_data_files_bkup

TRIM READS

1. ssh to apollo
2. ssh to one of the cylons
3. cd /home/bbart/data/blib
4. For new blibs make sure the blib directories have a trimmed_reads directory, for example:
   cmd: for i in {131..142}; do mkdir blib${i}/trimmed_reads; done
5. This example of the command is built to loop through multiple blib datasets. In order to use this with just one blib, instead of a range entered with {092..101}, you can replace that with just the number. Notes about the command:
    a. On apollo a directory different from /tmp is often needed for JAVA commands, I use ~/tmp. This isn't accessible to anyone else so make sure you change this to your own home directory once you create your own
    b. Make sure the number of threads you've selected is appropriate for the machine you're using, however this command never seems to use more than 3 CPUs
    c. This command is set up to use the example parameters provided on the Trimmomatic home page. This may not be appropriate for your dataset, but it seems to be fine most of the time
    d. This command is set up as 3 nested for loops to make it easier to change things like blib number, species suffix, and run info. This allows you to change this information in one place instead of several

   cmd: for i in {131..142}; do for k in xam; do echo blib${i}.${k}; for j in miseq062615; do echo ${j}; java -Djava.io.tmpdir=/home/~/tmp -Xmx4g -jar /nfs4shares/bioinfosw/installs_current/Trimmomatic-0.32/trimmomatic-0.32.jar PE -threads 13 -phred33 blib${i}.${k}/raw_reads/blib${i}.${k}.${j}.r1.fq.gz blib${i}.${k}/raw_reads/blib${i}.${k}.${j}.r2.fq.gz blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.p1.fq.gz blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.u1.fq.gz blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.p2.fq.gz blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.u2.fq.gz ILLUMINACLIP:/nfs4shares/bioinfosw/installs_current/Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:36 &> blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.trim; done; done; done

ASSEMBLE READS

1. ssh to apollo
2. ssh to one of the cylons
3. cd /home/bbart/data/blib
4. For new blibs make sure the blib directories have an assembly directory, for example:
   cmd: for i in {131..142}; do mkdir blib${i}/assembly; done
5. Notes on loops and threads in the TRIM READS section apply here as well. Assemble reads:
   cmd: for i in {131..142}; do for j in miseq062615; do for k in xam; do echo ${i} ${k} ${j} spades run; /nfs4shares/bioinfosw/installs_current/SPAdes-3.0.0-Linux/bin/spades.py -t 10 -o blib${i}.${k}/assembly/spades_v1 --pe1-1 blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.p1.fq.gz --pe1-2 blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.p2.fq.gz --pe1-s blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.u1.fq.gz --pe1-s blib${i}.${k}/trimmed_reads/blib${i}.${k}.${j}.u2.fq.gz -k 33,53,73,93 &> blib${i}.${k}/assembly/blib${i}.${k}.${j}_spades.out; done; done; done