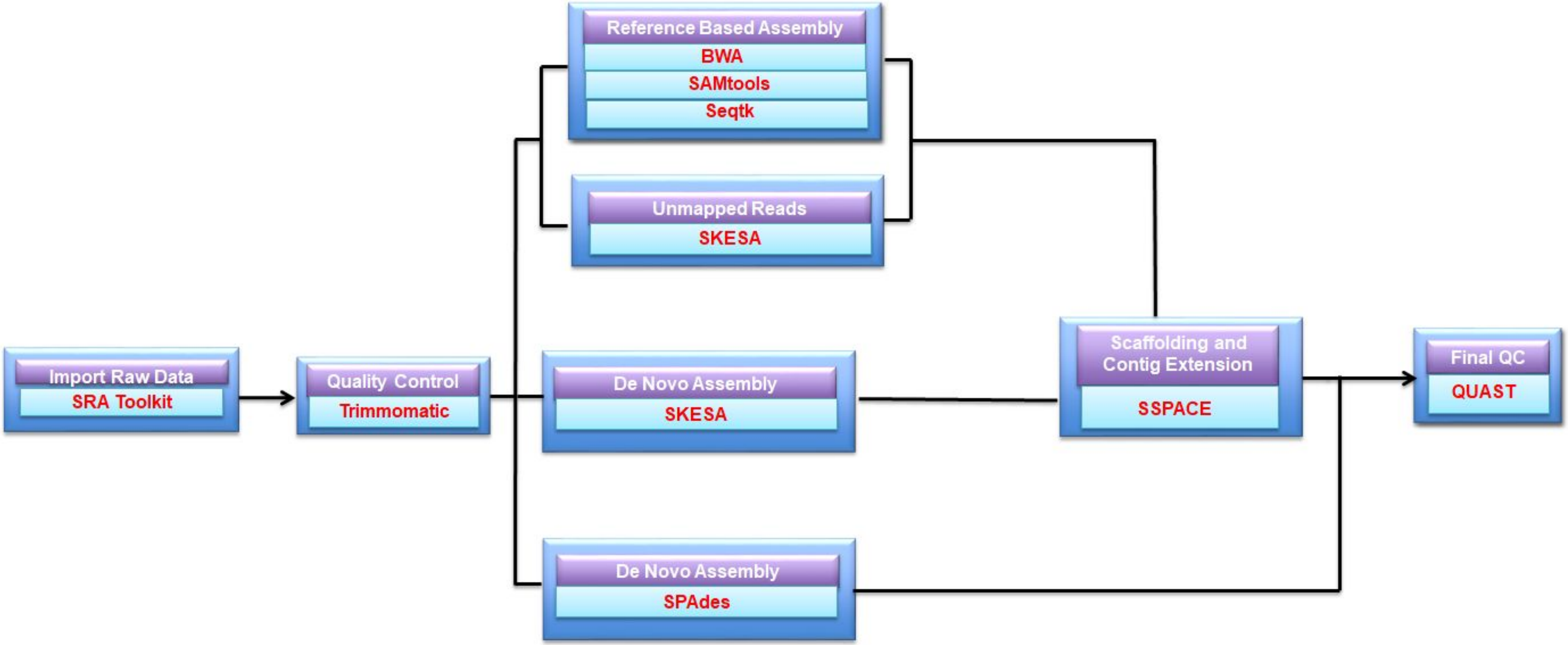
# Final Results, Genome Assembly

BIOL 7210: Computational Genomics - Spring 2018

Team-1 Members: Kunal Agarwal, Victoria Caban, Vasanta Chivukula, Seonggeon Cho, Siarhei Hladyshau, Hunter Seabolt, Nirav Shah, Tianze Song, Qinwei Zhuang

# Pipeline



Import Raw Data

SRA Toolkit

Quality Control

Trimmomatic

Reference Based Assembly

BWA

SAMtools

Seqtk

Unmapped Reads

SKESA

De Novo Assembly

SKESA

De Novo Assembly

SPAdes

Scaffolding and Contig Extension

SSPACE

Final QC

QUAST

# Trimming and Quality Control

Trimming raw data with Trimmomatic

- ILLUMINACLIP: trims adapter sequences in the reads
- SLIDINGWINDOW: trims the reads based on the threshold quality score set by a user

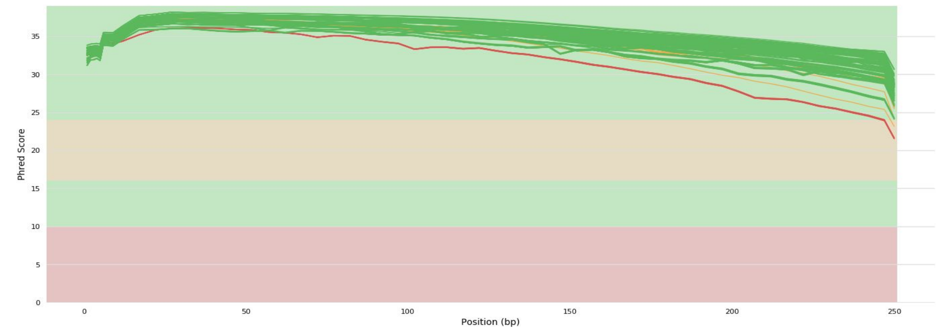
  *4:20 was used in our samples

- MINLEN: drops reads if they are below an assigned length

  *20 was set as the minimum length

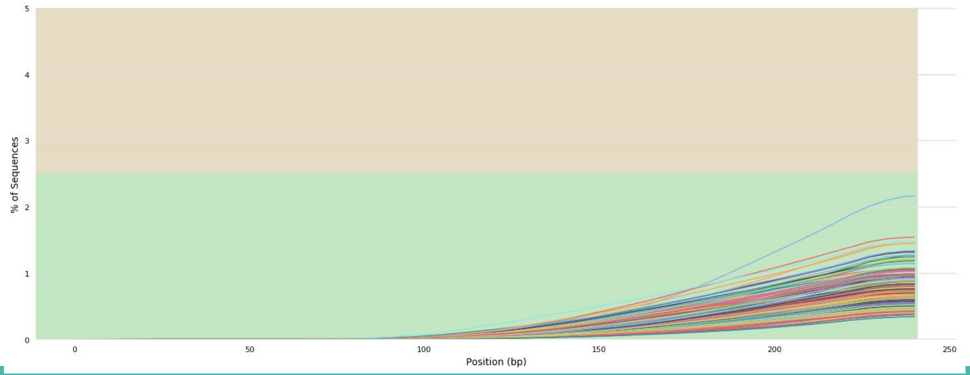
# Trimmomatic Successfully Removes Low Quality and Adapter Reads

FastQC: Mean Quality Scores

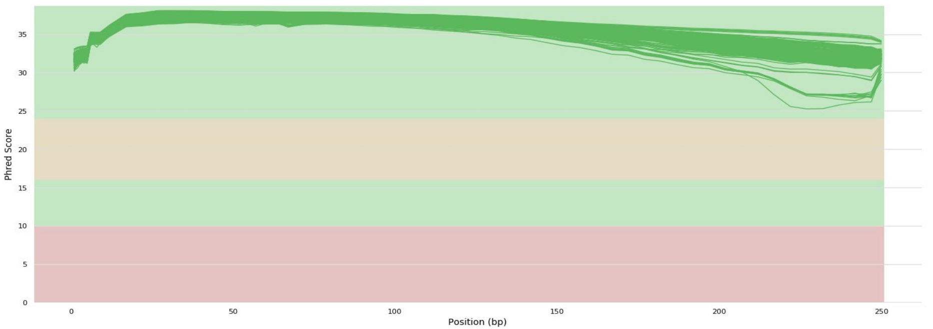


## Before Trimming

FastQC: Adapter Content

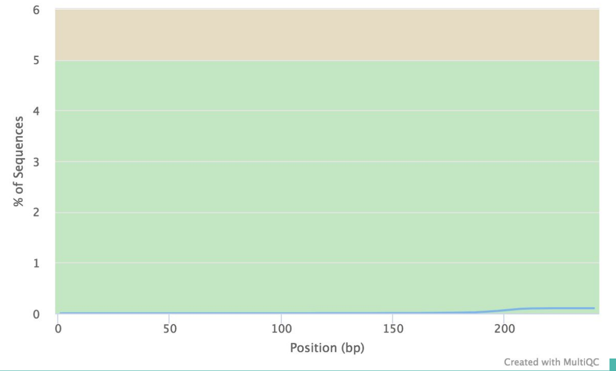


FastQC: Mean Quality Scores



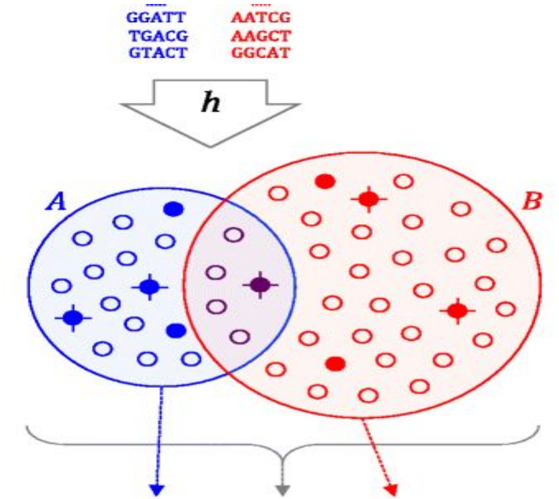
## After Trimming

FastQC: Adapter Content



# Reference Based Assembly

# MASH

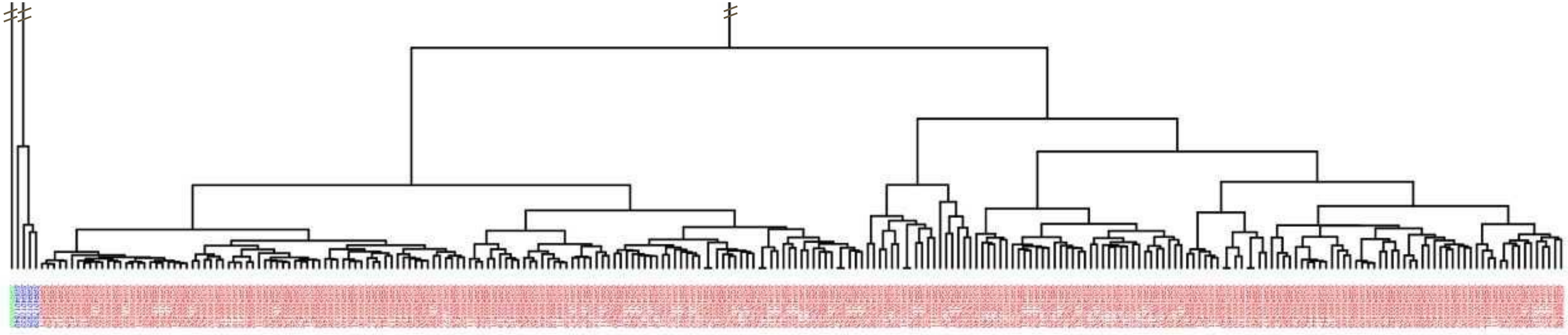


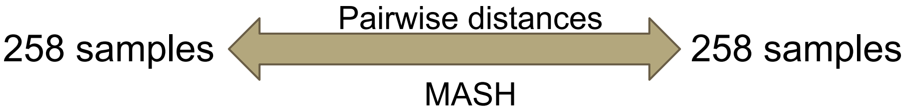
| $S(A)$ | $S(A \cup B)$ | $S(B)$ |
|---|---|---|
| 42 | 42 | 66 |
| 64 | 64 | 82 |
| 82 | 66 | 87 |
| 128 | 82 | 104 |
| 139 | 87 | 127 |

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

- MinHash Algorithm is used by MASH.
- MinHash algorithm provides an estimation of the Jaccard index.
- MASH evaluates mutation distance using Jaccard index between the genomes for similarity.

# Evaluation of distance between samples



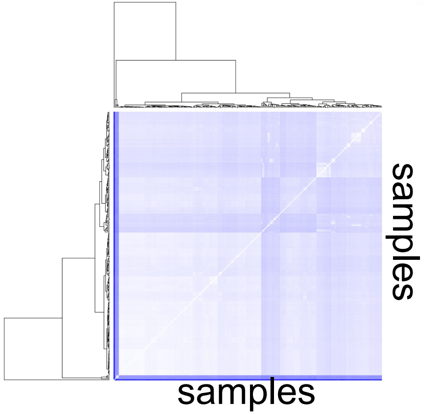


258 samples

Pairwise distances

MASH

258 samples



# Choosing Reference Genomes

NCBI

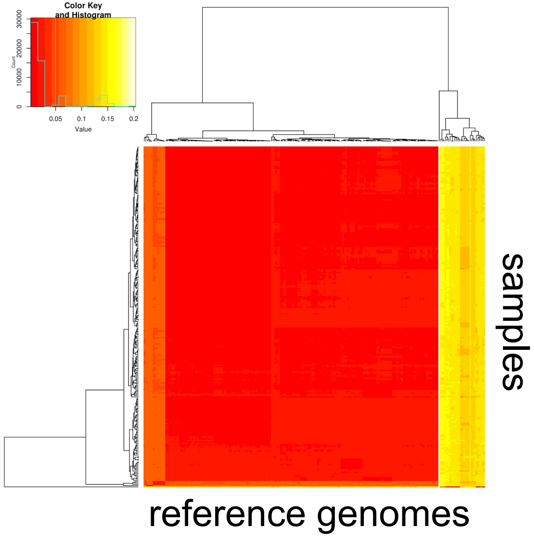
220 complete, reference genomes of *Klebsiella* spp

Pairwise distances

MASH

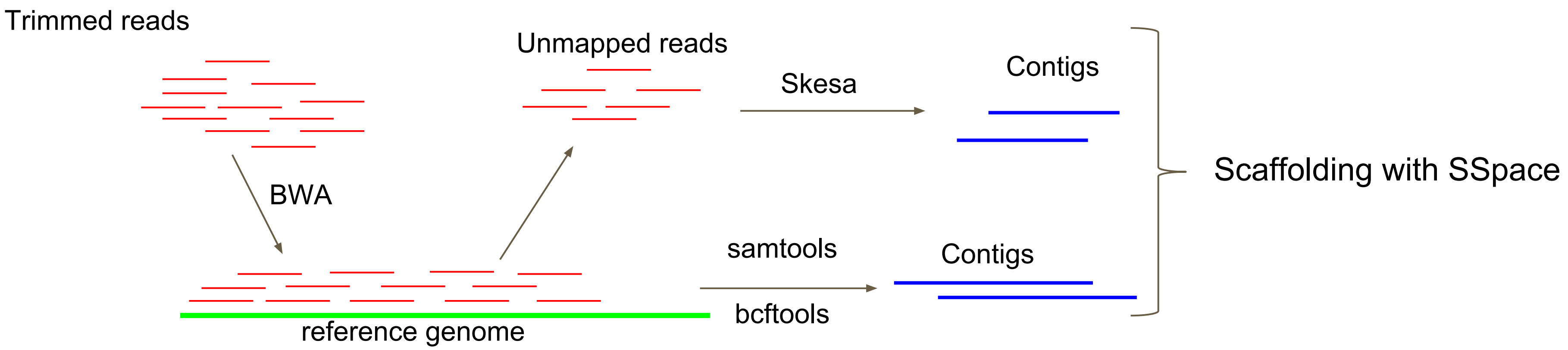
258 samples

Choice of best genome reference genome for every sample



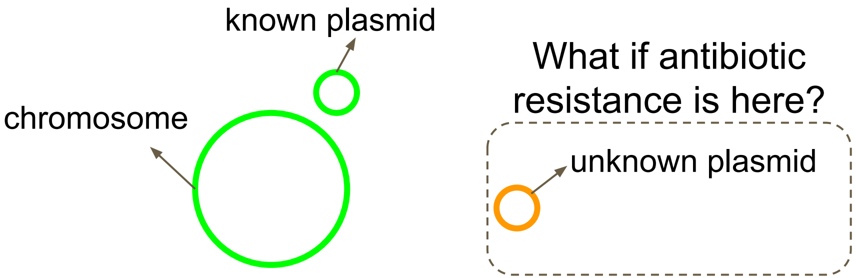
| | |
|---|---|
| 252 samples: | *Klebsiella pneumoniae* |
| 4 samples: | *Klebsiella variicola* |
| 1 sample: | *Klebsiella oxytoca* |
| 1 sample: | *Klebsiella* sp. 2N3 |

# Reference Based Assembly



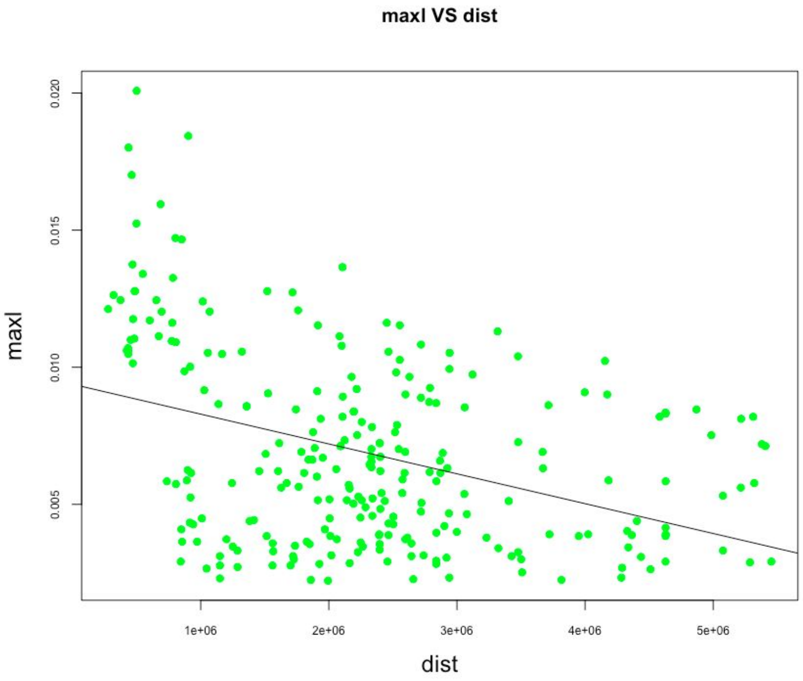
?

Why to assemble unmapped reads separately?



What if antibiotic resistance is here?

# Importance of reference genome



### Linear Regression

```
Residuals:
     Min       1Q   Median       3Q      Max
-2002720  -836508  -166123   576530  3190016

Coefficients:
                    Estimate          Std. Error          t value   Pr(>|t|)
(Intercept)          3246617              157595          20.601   < 2e-16 ***
all_data$dist -137253228            20458750          -6.709   1.25e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’

Residual standard error: 1146000 on 256 degrees of freedom
Multiple R-squared:  0.1495,       Adjusted R-squared:  0.1462
F-statistic: 45.01 on 1 and 256 DF,  p-value: 1.252e-10
```

### Pearson's product-moment correlation

```
t = -6.7088, df = 256, p-value = 1.252e-10
alternative hypothesis: true, correlation is not equal to 0
95 percent confidence interval: -0.4858635, -0.2776702
sample estimates: cor, -0.3866826
```

# de Novo Assembly

# de Novo Assembly Using SPAdes

It is an assembler that works based on DeBruijn graphs

Designed to assemble small genome

Do scaffolding by itself

Supports paired-ends and unpaired reads

Give flexibility in Kmer selection

```
Spades.py --careful -k kmer size --pe1-1 forward_paired.fq --pe1-2 reverse_paired.fq --pe1-s forward_unpaired.fq --pe1-s reverse_unpaired.fq -o output_directory
```

SPAdes 3.11.1 User Mannul
http://cab.spbu.ru/files/release3.11.1/manual.html#sec3.4

# SPAdes Pipeline

Read Error Correction--BayersHammer

Assemble--Spades

Mismatch Correction--improves mismatch and short indel rates in resulting contigs and scaffolds; this module uses the BWA tool, activated by --careful

Sergey I. Nikolenko, Anton I. Korobeynikov, Max A. Alekseyev
*BMC Genomics* 2013, **14**(Suppl 1):S7

SPAdes 3.11.1 User Mannul
http://cab.spbu.ru/files/release3.11.1/manual.html#sec3.4

# SPAdes Kmer

If we give many kmers in one command line like this:

```
spades.py -k 41,77,99,127 --careful <your reads> -o spades_output
```

Output is the assembly with best N50.

```
spades.py -k 41 --careful <your reads> -o spades_output
```

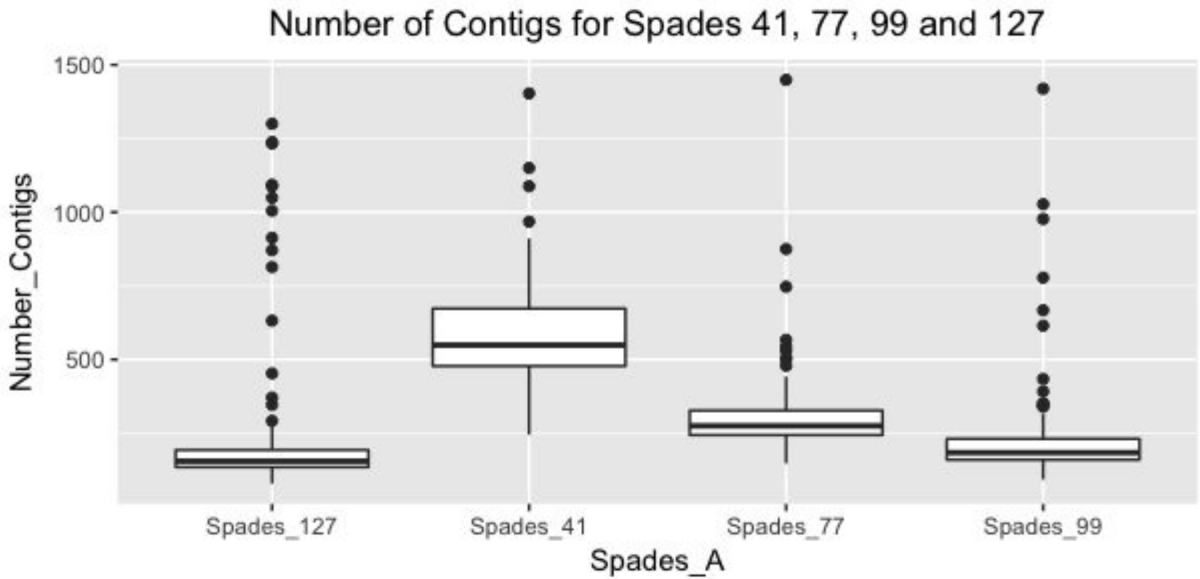
```
spades.py -k 77 --careful <your reads> -o spades_output
```

```
spades.py -k 99 --careful <your reads> -o spades_output
```

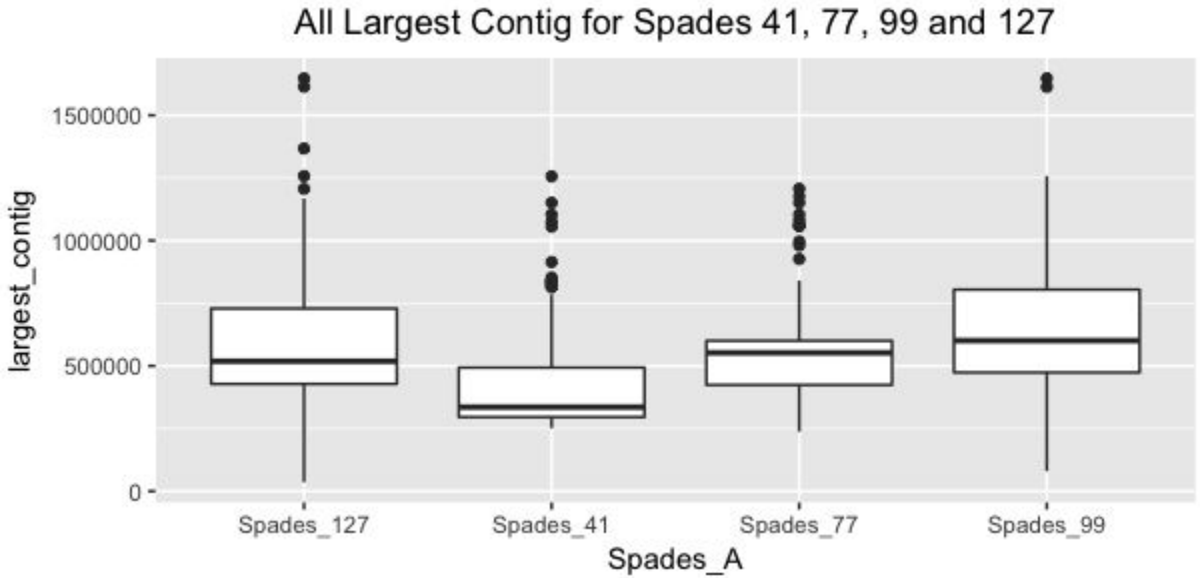
```
spades.py -k 127 --careful <your reads> -o spades_output
```

Select the best assembly by multi-parameters

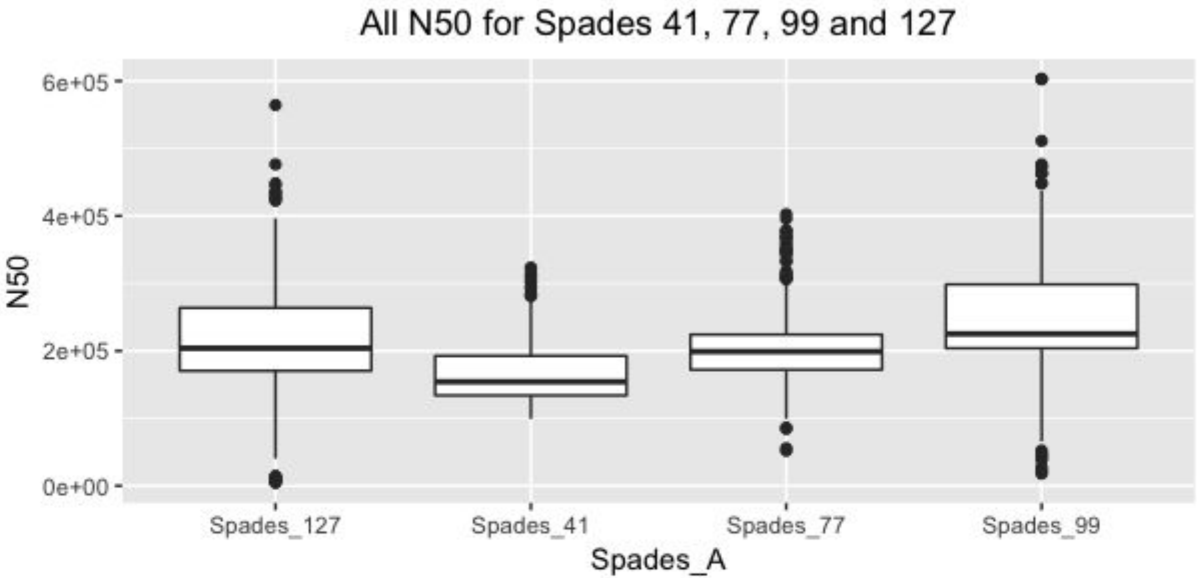
# SPAdes: number of contigs



# SPAdes: largest contig

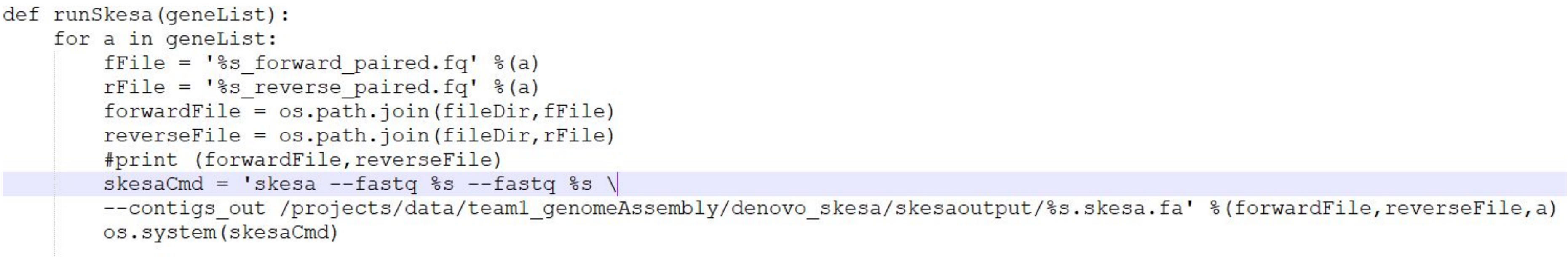


# SPAdes: N50


All N50 for Spades 41, 77, 99 and 127
N50
6e+05
4e+05
2e+05
0e+00
Spades_127
Spades_41
Spades_77
Spades_99
Spades_A

# de Novo Assembly Using Skesa

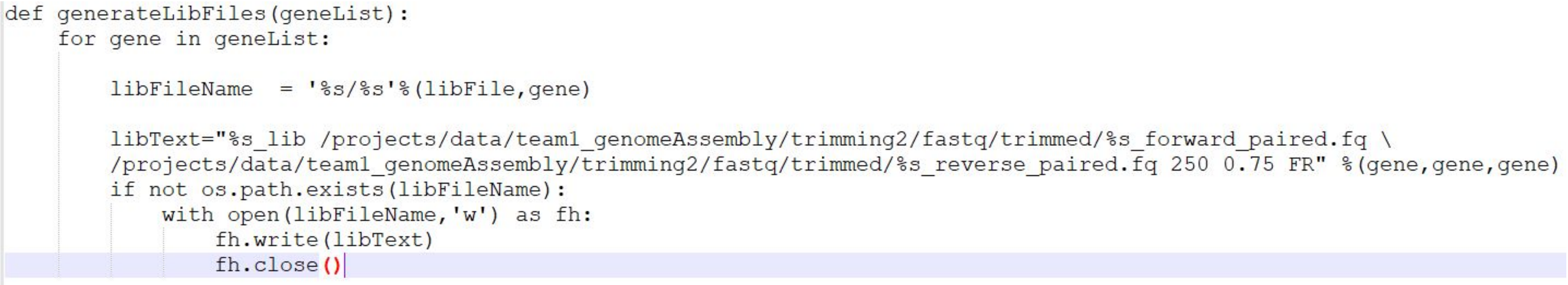
- The binary for Skesa was provided by CDC
- It is an assembler that works based on DeBruijn graphs
- It is designed for haploid genomes sequenced using Illumina
- Creates breaks at repeat regions in genomes
- Multi-threaded application - so good for scaling



```
def runSkesa(geneList):
    for a in geneList:
        fFile = '%s_forward_paired.fq' %(a)
        rFile = '%s_reverse_paired.fq' %(a)
        forwardFile = os.path.join(fileDir,fFile)
        reverseFile = os.path.join(fileDir,rFile)
        #print (forwardFile,reverseFile)
        skesaCmd = 'skesa --fastq %s --fastq %s \
        --contigs_out /projects/data/team1_genomeAssembly/denovo_skesa/skesaoutput/%s.skesa.fa' %(forwardFile,reverseFile,a)
        os.system(skesaCmd)
```

# Scaffolding Using SSPACE

- Scaffolding Pre-Assemblies After Contig Extension (SSPACE)
- Extends and scaffolds pre-assembled contigs
- Uses Bowtie to map all reads to the pre-assembled contigs
- A library file containing library name, read 1, read 2, insert size (500), error (0.75), FR



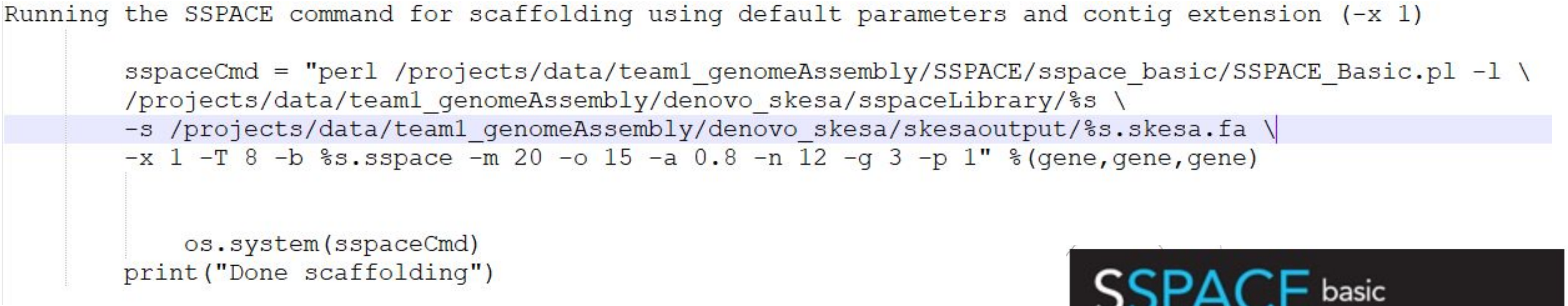
```
def generateLibFiles(geneList):
    for gene in geneList:

        libFileName  = '%s/%s'%(libFile,gene)

        libText="%s_lib /projects/data/team1_genomeAssembly/trimming2/fastq/trimmed/%s_forward_paired.fq \
        /projects/data/team1_genomeAssembly/trimming2/fastq/trimmed/%s_reverse_paired.fq 250 0.75 FR" %(gene,gene,gene)
        if not os.path.exists(libFileName):
            with open(libFileName,'w') as fh:
                fh.write(libText)
                fh.close()
```

SSPACE basic

FOR 100% BASECLEAR

# Scaffolding continued

- Contig extension was performed using SSAKE method by changing the standard -x 0 to 1
- This is followed by building scaffolds and merging contigs
- The output contains final scaffolds in fasta format, scaffolds with initial numbered contigs, a log file and a summary file



```
Running the SSPACE command for scaffolding using default parameters and contig extension (-x 1)

    sspaceCmd = "perl /projects/data/team1_genomeAssembly/SSPACE/sspace_basic/SSPACE_Basic.pl -l \
    /projects/data/team1_genomeAssembly/denovo_skesa/sspaceLibrary/%s \
    -s /projects/data/team1_genomeAssembly/denovo_skesa/skesaoutput/%s.skesa.fa \
    -x 1 -T 8 -b %s.sspace -m 20 -o 15 -a 0.8 -n 12 -g 3 -p 1" %(gene,gene,gene)


        os.system(sspaceCmd)
    print("Done scaffolding")
```

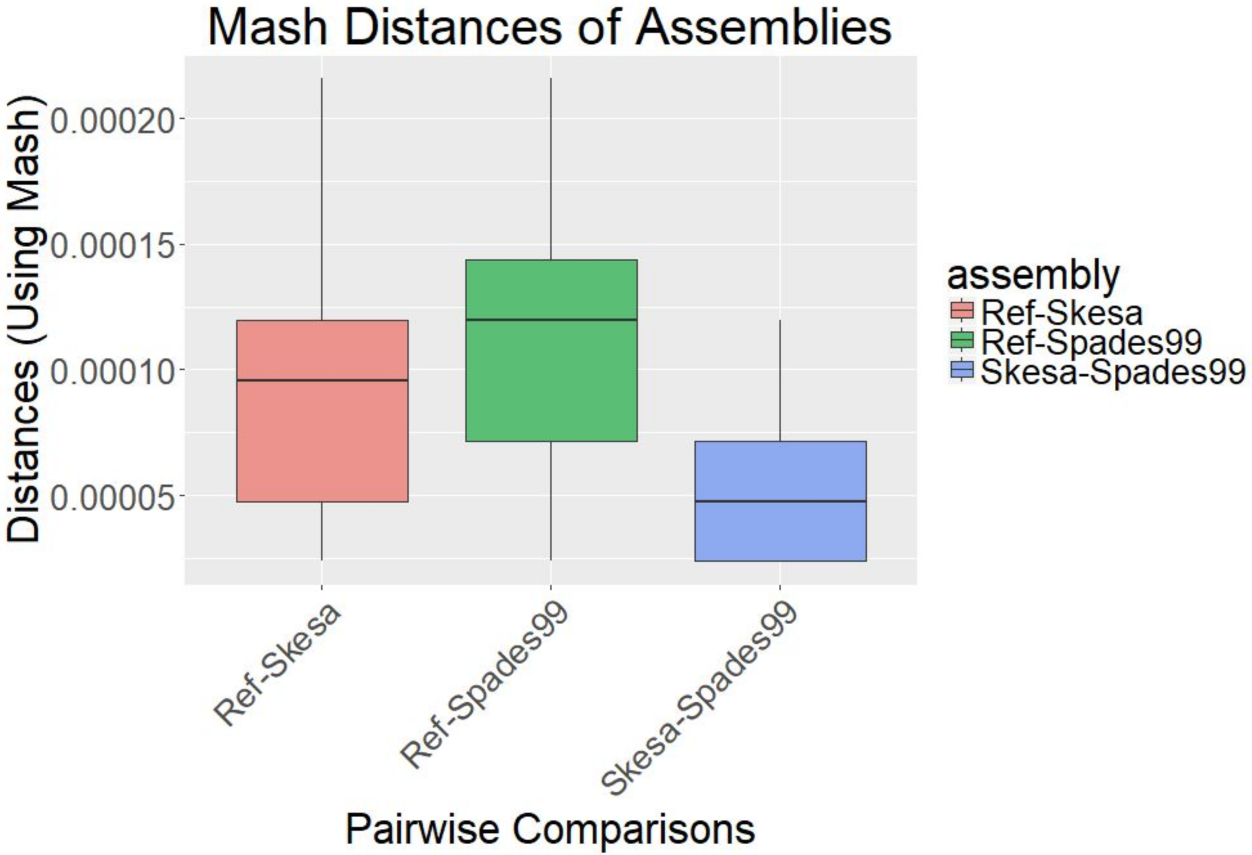
SSPACE basic

FOR 100% BASECLEAR

# Comparison between Spades and Skesa

| Parameters | Average SPAdes | Average Skesa | P value |
|---|---|---|---|
| N50 | 250137 | 229259 | 0.19592 |
| # Contigs | 212 | 123 | 1.55E-10*** |
| Largest Contigs | 645324 | 609123 | 0.063028 |
| Total Length | 5588948 | 5601627 | 0.44905 |
| N’s per 100kbp | 2.781 | 11.456 | 0.000104*** |

# Merging assemblies

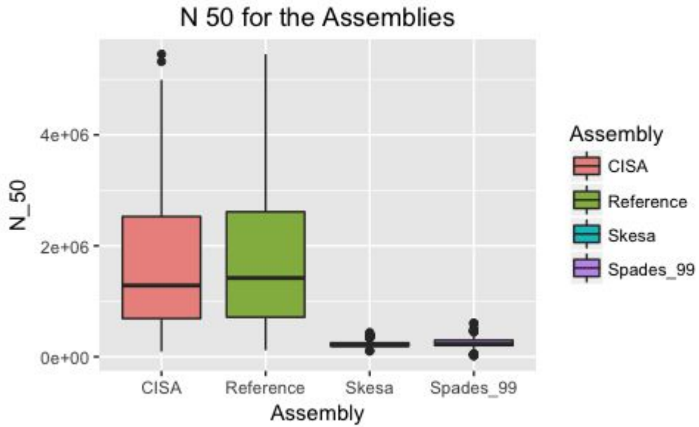


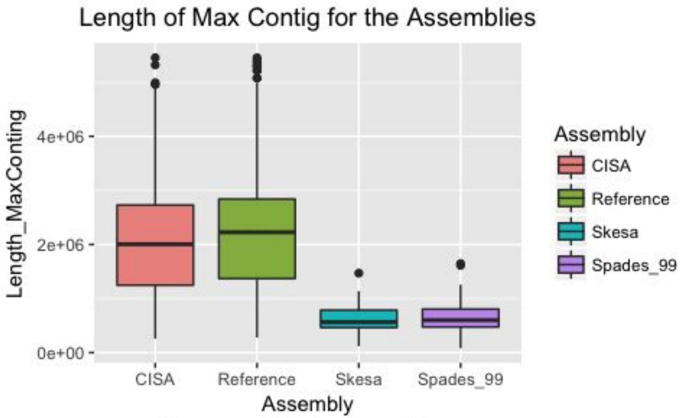
Reference based

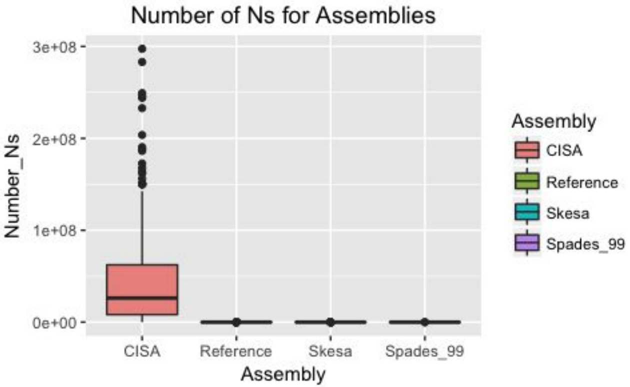
Skesa

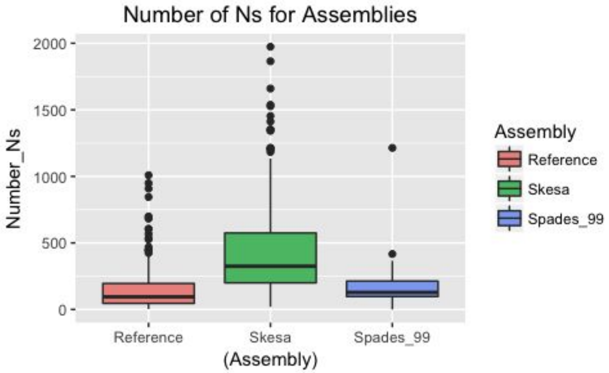
} CISA

# Quality of assemblies


N 50 for the Assemblies


Length of Max Contig for the Assemblies


Number of Ns for Assemblies


Number of Ns for Assemblies

# References

Bankevich, Anton et al. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology* 19.5 (2012): 455–477. *PMC*. Web. 6 Mar. 2018.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30.15 (2014): 2114–2120. *PMC*. Web. 6 Mar. 2018.

Gurevich, Alexey et al. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29.8 (2013): 1072–1075. *PMC*. Web. 6 Mar. 2018

Heng Li, Richard Durbin; Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, Volume 25, Issue 14, 15 July 2009, Pages 1754–1760, https://doi.org/10.1093/bioinformatics/btp324

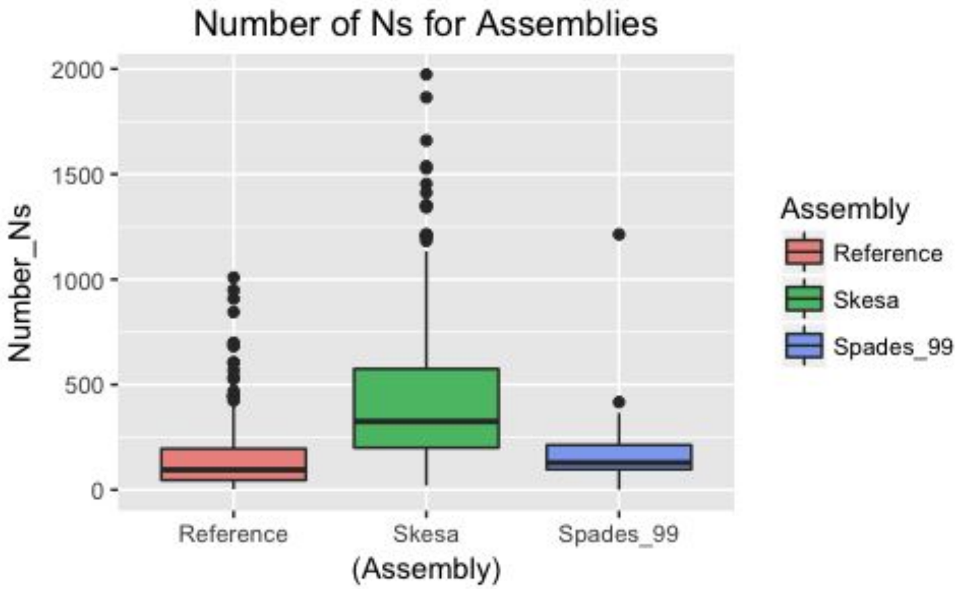
Boetzer M, Henkel CV, Jansen HJ, Butler D and Pirovano W. 2010. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 27(4):578-579

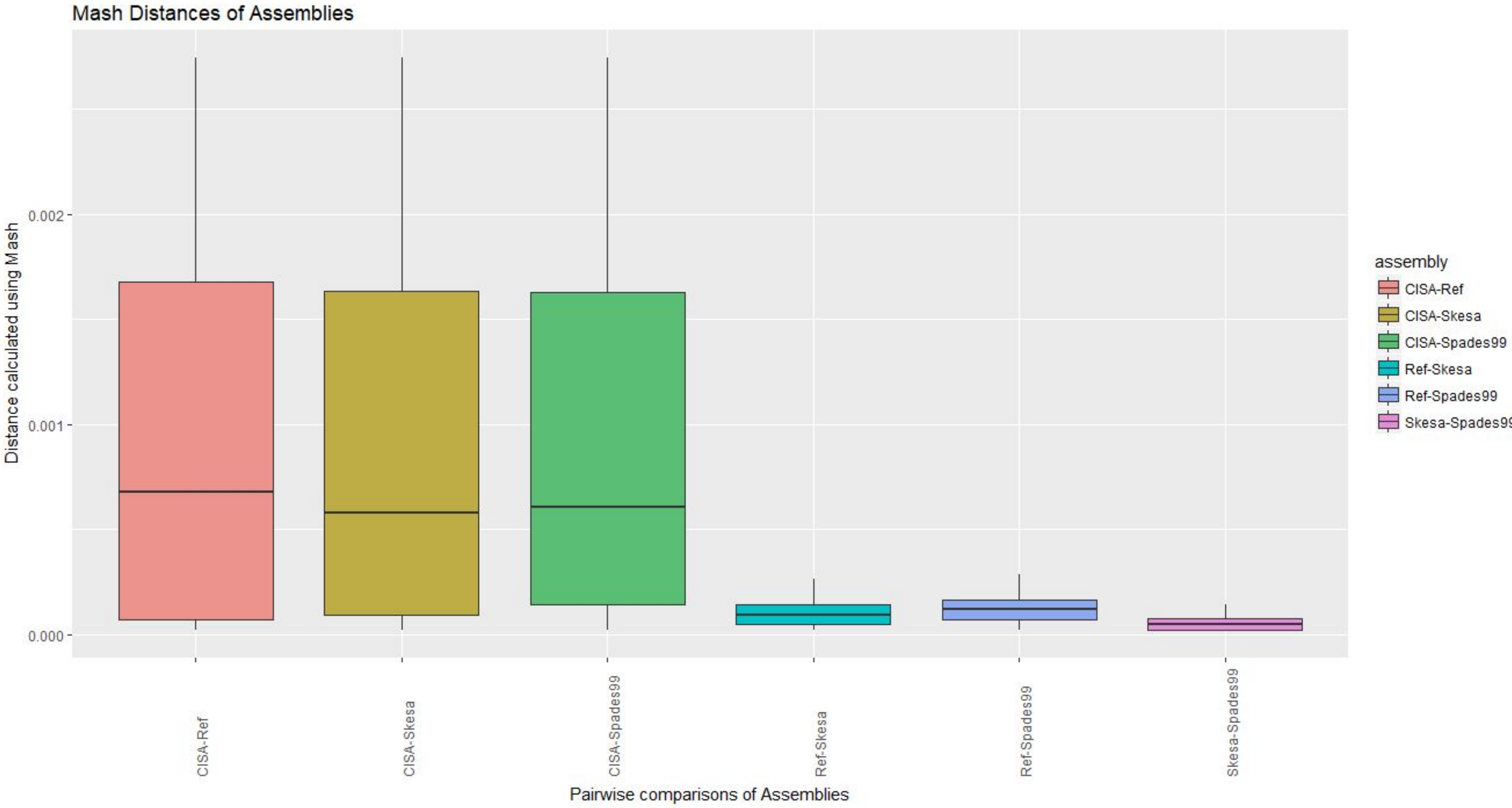
http://bioinf.spbau.ru/spades

http://sb.nhri.org.tw/CISA/en/Instruction

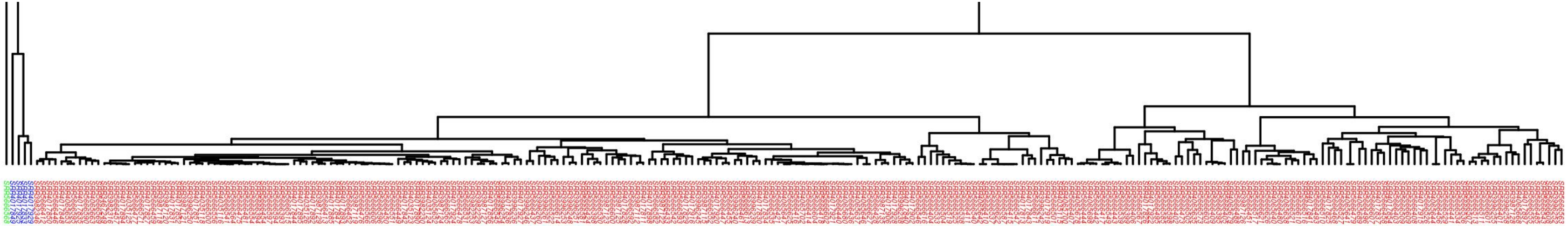
https://github.com/enormandeau

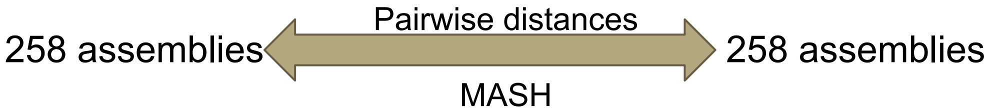
Thank you for your attention!


Number of Ns for Assemblies
Number_Ns
2000
1500
1000
500
0
Reference
Skesa
Spades_99
(Assembly)
Assembly
Reference
Skesa
Spades_99


Mash Distances of Assemblies
Distance calculated using Mash
0.002
0.001
0.000
CISA-Ref
CISA-Skesa
CISA-Spades99
Ref-Skesa
Ref-Spades99
Skesa-Spades99
Pairwise comparisons of Assemblies
assembly
CISA-Ref
CISA-Skesa
CISA-Spades99
Ref-Skesa
Ref-Spades99
Skesa-Spades9

# Evaluation of assemblies



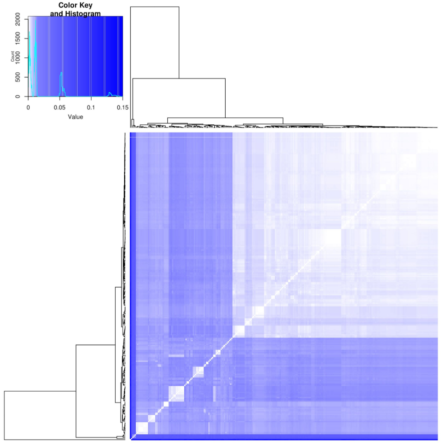


258 assemblies

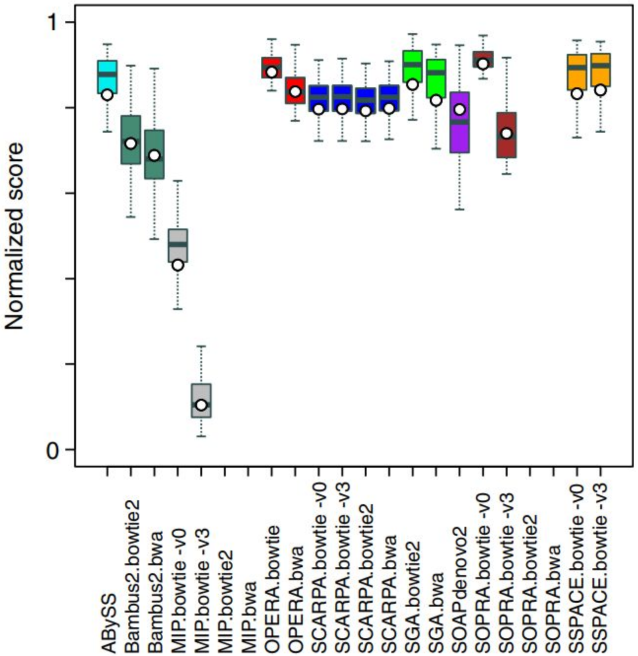
Pairwise distances

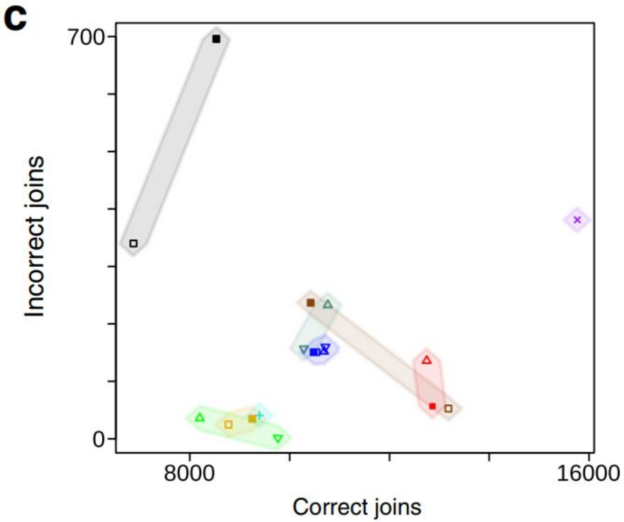
MASH

258 assemblies



# Choice of scaffolding tool





Genome Biol. 2014; 15(3),Martin Hunt, el.al.

# Example of pipeline for reference guided assembly

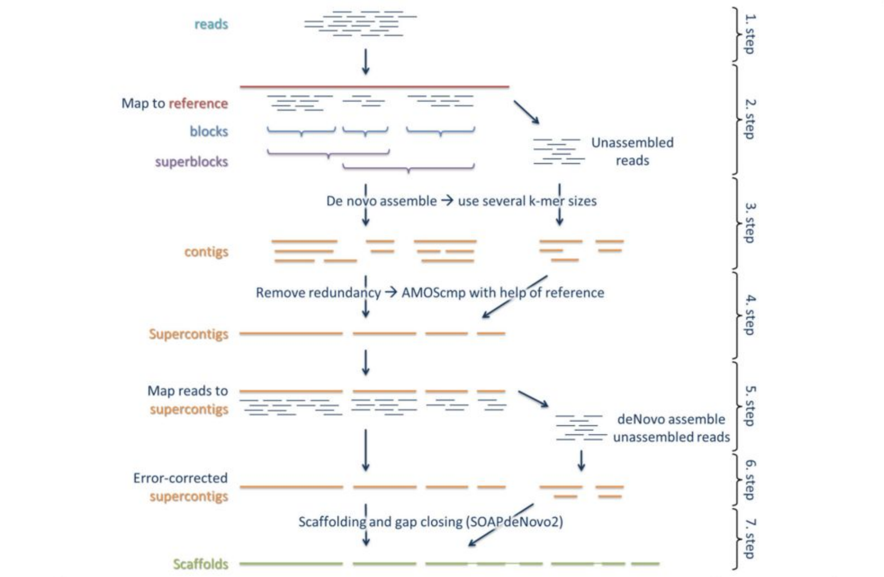

reads
Map to reference
blocks
superblocks
Unassembled reads
De novo assemble → use several k-mer sizes
contigs
Remove redundancy → AMOScmp with help of reference
Supercontigs
Map reads to supercontigs
deNovo assemble unassembled reads
Error-corrected supercontigs
Scaffolding and gap closing (SOAPdeNovo2)
Scaffolds
1. step
2. step
3. step
4. step
5. step
6. step
7. step

**Fig. 1** Reference-guided de novo assembly pipeline. Raw reads get quality trimmed (1. step) and mapped against a reference (2. step). Reference mapped reads are grouped into blocks with continuous read coverage. These blocks are then combined into superblocks until a total length of at least 12 kb is reached. Superblocks are overlapping by at least one block. Each superblock and all unmapped reads are separately de novo assembled (3. step). Resulting contigs are merged into non-redundant supercontigs (4. step). In the fifth step, reads are mapped back to the supercontigs and unmapped reads are de novo assembled to get additional supercontigs. All supercontigs are error corrected with back mapped reads (6. step) and afterwards used for scaffolding and gap closing (7. step)

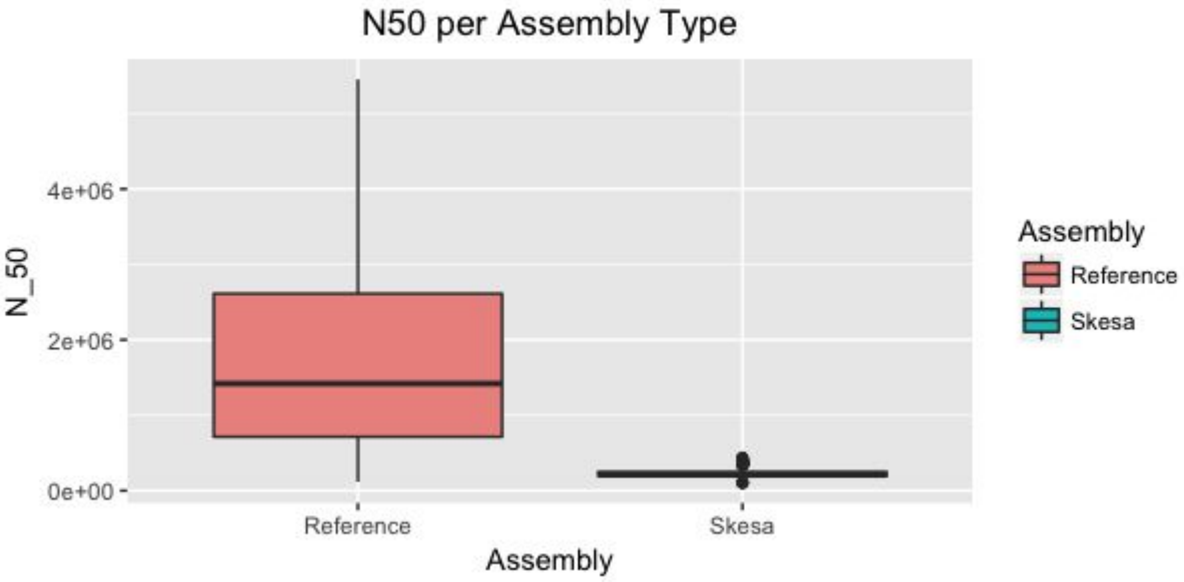
BMC Bioinformatics. 2017 Nov 10;18(1):474.

Lischer HEL, Shimizu KK.

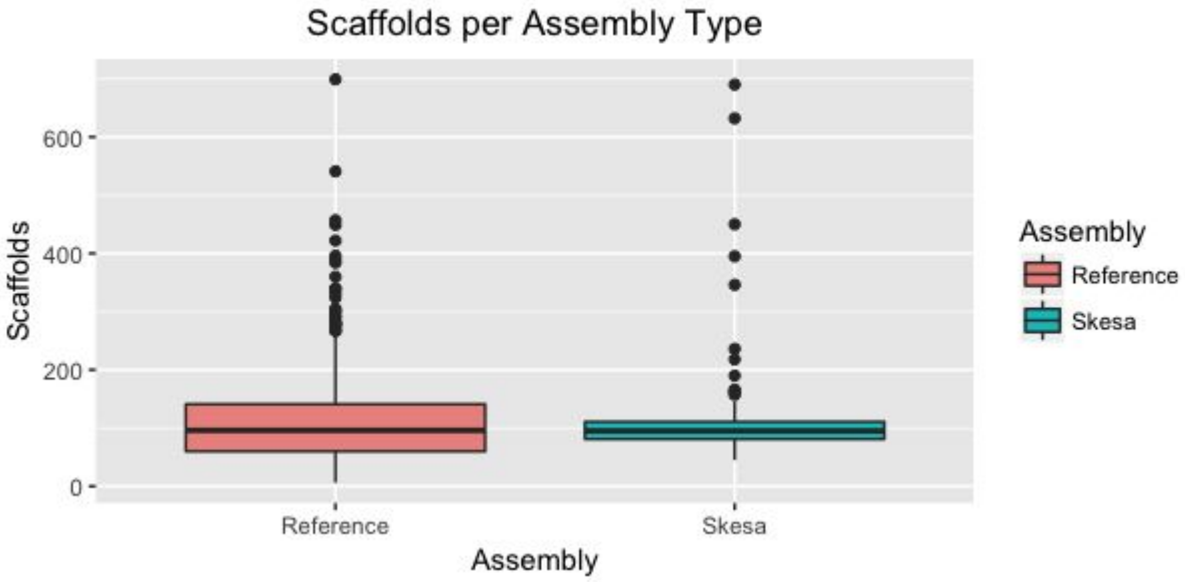
# Pipeline for reference based assembly

```
bwa index -a is [reference genome]
bwa mem [reference genome] [forward and reverse reads] > [output.sam]
samtools sort [output.sam] > [output_sorted.bam]
samtools index [output_sorted.bam]
samtools view -b -f 4 [output_sorted.bam] > [unmapped.bam]
samtools bam2fq [unmapped.bam]> [unmapped.fastq]
samtools mpileup -v --no-BAQ -f [reference genome] [output_sorted.bam] |
        bcftools call -c | vcfutils.pl vcf2fq | seqtk seq -A > [assembly.fasta]
```

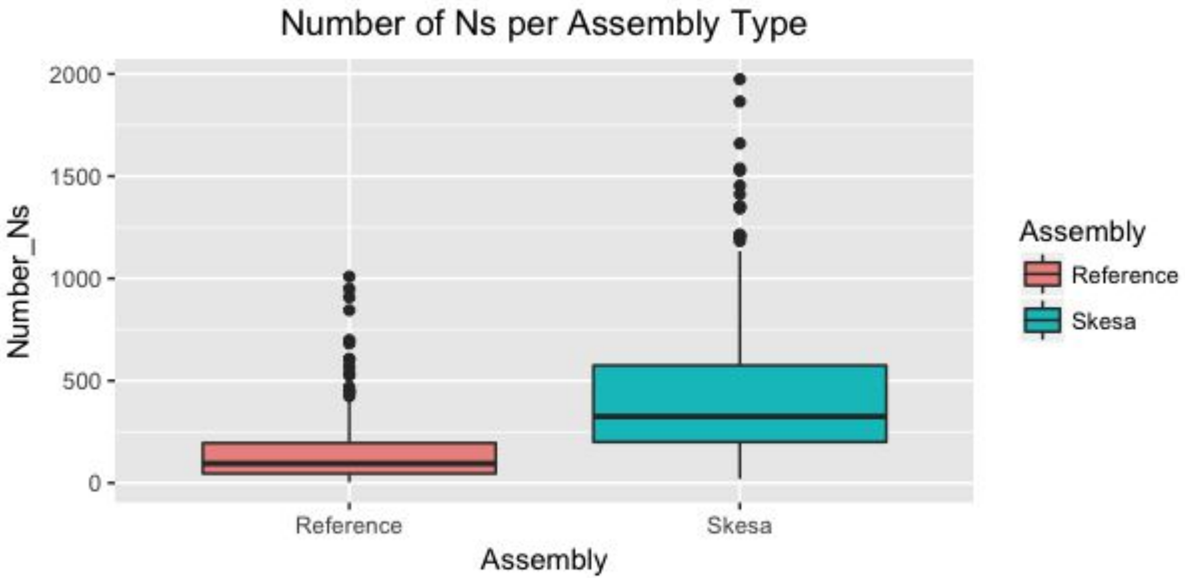
# N50 for Referenced Based and Skesa Assembly



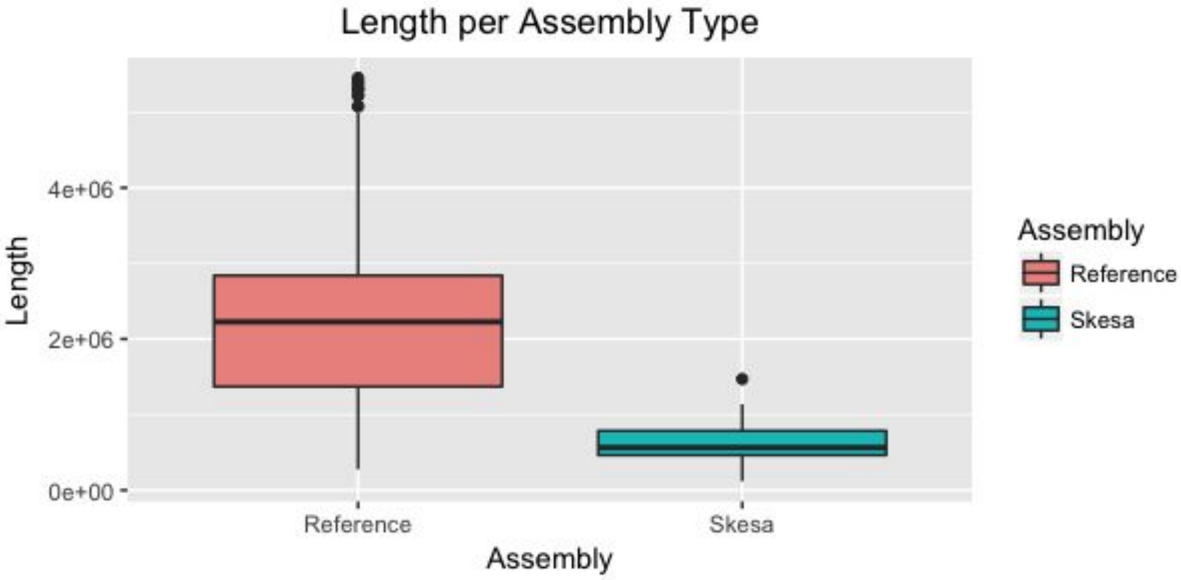
# Scaffolds for Referenced Based and Skesa Assembly



# Number of Ns for Referenced Based and Skesa Assembly



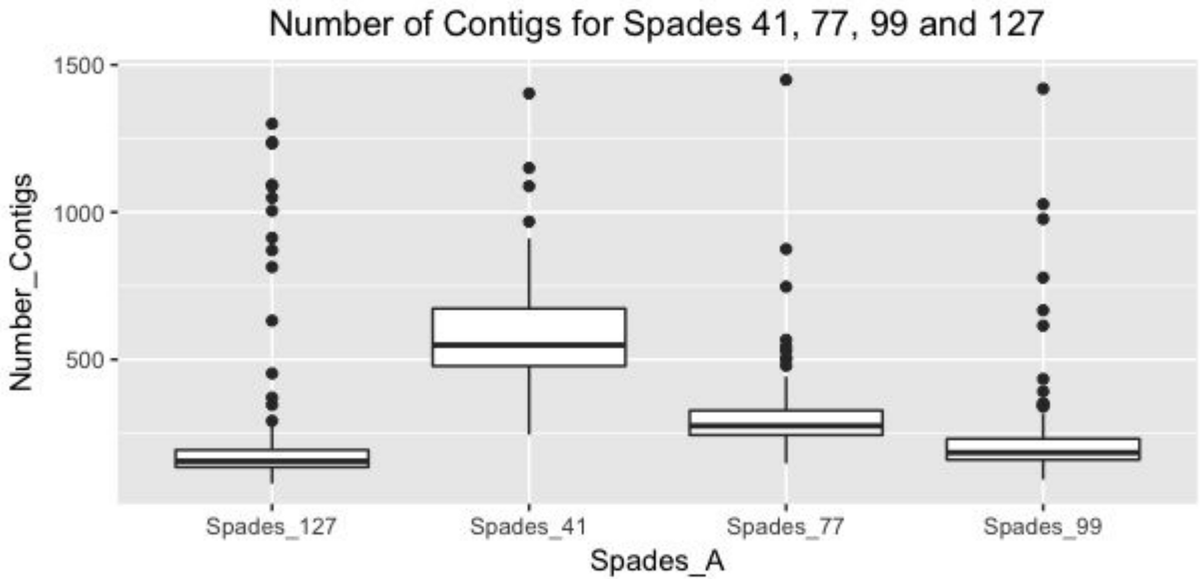
# Length for Referenced Based and Skesa Assembly



# References

https://doi.org/10.1186/s13059-016-0997-x

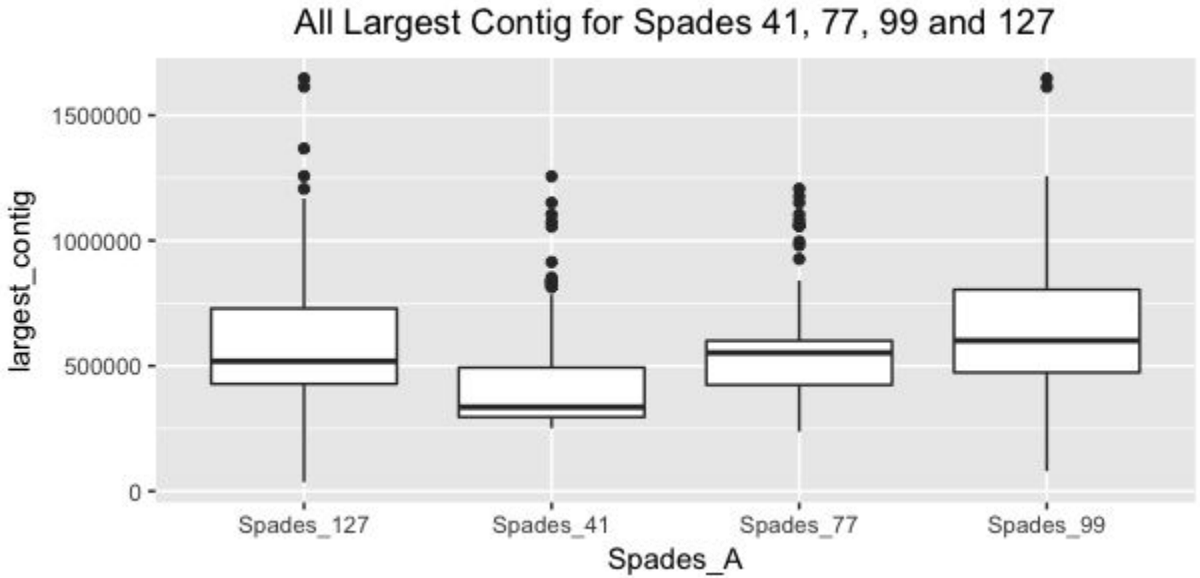
# Number of Contigs for Spades Assembly



```
Parameter
Table Analyzed spades_number_of_contigs

One-way analysis of variance
  P value            < 0.0001
  P value summary        ***
  Are means signif. different? (P < 0.05) Yes
  Number of groups       4
  F          146.7
  R square           0.3006

Bartlett's test for equal variances
  Bartlett's statistic (corrected)  28.16
  P value            < 0.0001
  P value summary        ***
  Do the variances differ signif. (P < 0.05)          Yes

ANOVA Table                             SS           df        MS
  Treatment (between columns)           2.524e+007   3         8.414e+006
  Residual (within columns)      5.872e+007   1024      57343
  Total                          8.396e+007   1027

Tukey's Multiple Comparison Test        Mean Diff.   q         Significant? P < 0.05?         95% CI of diff
  SPAdes 41 vs SPAdes 77                292.1        19.56            Yes      ***            237.3 to 346.9
  SPAdes 41 vs SPAdes 99                379.2        25.38            Yes      ***            324.4 to 434.0
  SPAdes 41 vs SPAdes 127               384.3        25.72            Yes      ***            329.5 to 439.1
  SPAdes 77 vs SPAdes 99                87.03        5.826            Yes      ***            32.23 to 141.8
  SPAdes 77 vs SPAdes 127               92.12        6.167            Yes      ***            37.32 to 146.9
  SPAdes 99 vs SPAdes 127               5.089        0.3407           No       ns             -49.71 to 59.89
```

Summary: SPAdes 99 and SPAdes 127 have the significantly lower contig number compared to other kmer size

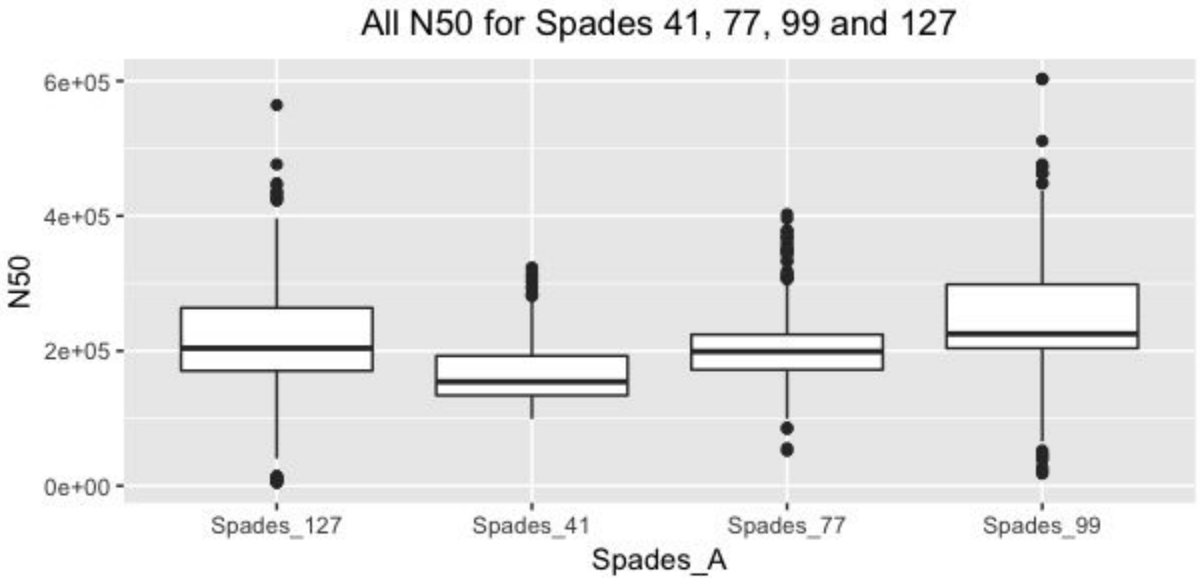
# Largest Contig for Spades Assembly



Summary: SPAdes 99 has the significantly longer contig length compared to other kmer size

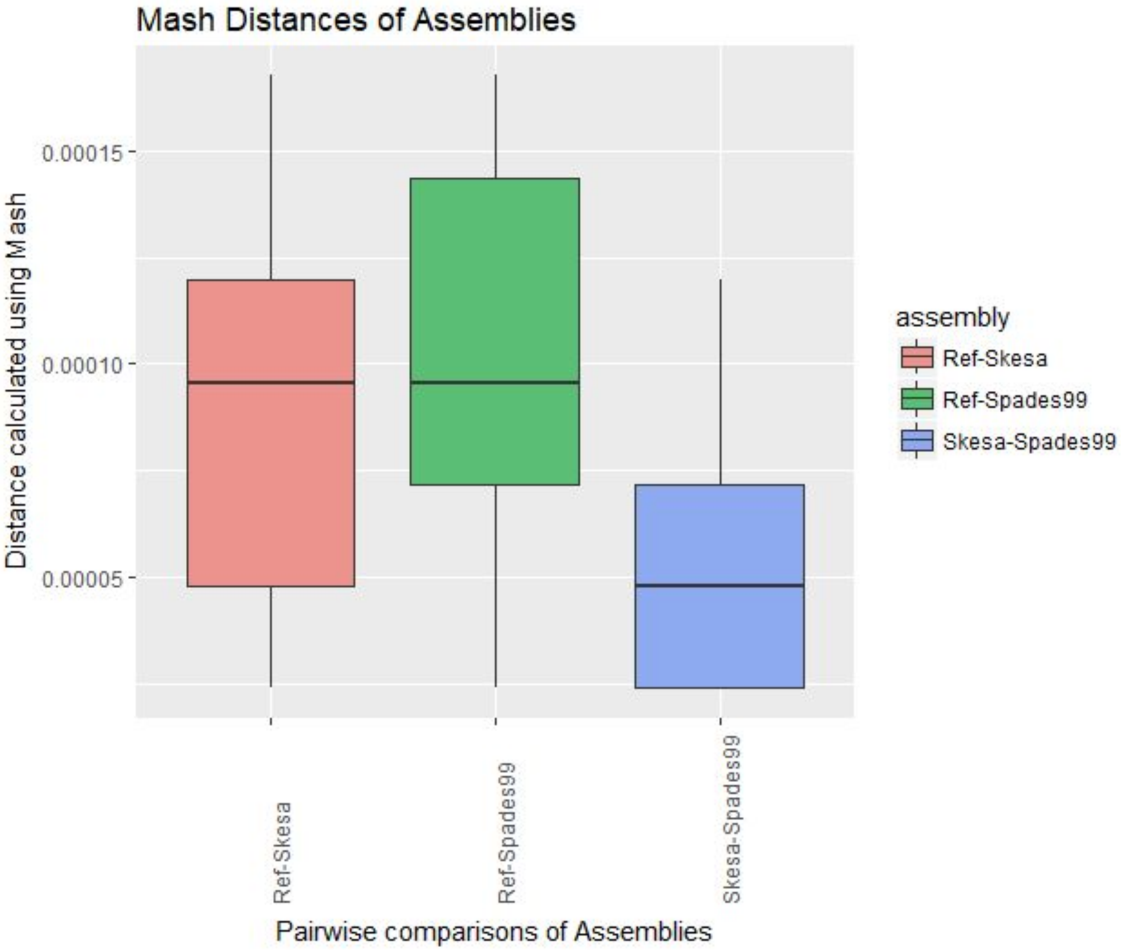
```
Parameter
Table Analyzed spades_large_contigs

One-way analysis of variance
  P value            < 0.0001
  P value summary         ***
  Are means signif. different? (P < 0.05) Yes
  Number of groups        4
  F          47.50
  R square           0.1216

Bartlett's test for equal variances
  Bartlett's statistic (corrected)  52.39
  P value            < 0.0001
  P value summary         ***
  Do the variances differ signif. (P < 0.05)          Yes

ANOVA Table                          SS          df          MS
  Treatment (between columns)        6.592e+012  3           2.197e+012
  Residual (within columns)     4.760e+013  1029   4.626e+010
  Total                         5.419e+013  1032

Tukey's Multiple Comparison Test     Mean Diff.    q        Significant? P < 0.05?   95% CI of diff
  SPAdes 41 vs SPAdes 77             -124376       9.280             Yes     ***      -173547 to -75205
  SPAdes 41 vs SPAdes 99             -220155       16.47             Yes     ***      -269185 to -171126
  SPAdes 41 vs SPAdes 127            -153780       11.46             Yes     ***      -202999 to -104562
  SPAdes 77 vs SPAdes 99             -95779        7.174             Yes     ***      -144761 to -46797
  SPAdes 77 vs SPAdes 127            -29404        2.194             No      ns       -78575 to 19767
  SPAdes 99 vs SPAdes 127            66375         4.966             Yes     **       17345 to 115405
```

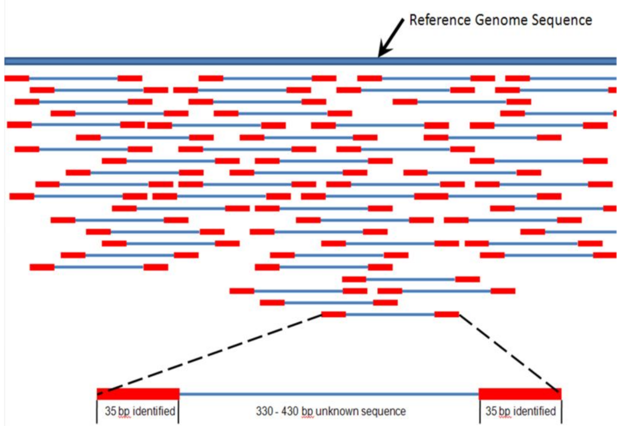
# N50 for Spades Assembly



Summary: SPAdes 99 has the significantly higher N50 compared to other kmer size

```
Parameter
Table Analyzed N50

One-way analysis of variance
  P value           < 0.0001
  P value summary        ***
  Are means signif. different? (P < 0.05) Yes
  Number of groups       4
  F        44.23
  R square          0.1147

Bartlett's test for equal variances
  Bartlett's statistic (corrected)  143.6
  P value           < 0.0001
  P value summary        ***
  Do the variances differ signif. (P < 0.05)           Yes

ANOVA Table    SS         df          MS
  Treatment (between columns)          7.819e+011       3          2.606e+011
  Residual (within columns)       6.034e+012       1024      5.893e+009
  Total   6.816e+012       1027

Tukey's Multiple Comparison Test      Mean Diff.       q          Significant? P < 0.05?            95% CI of diff
  SPAdes 41 vs SPAdes 77              -36846           7.695             Yes       ***              -54414 to -19279
  SPAdes 41 vs SPAdes 99              -77679           16.22             Yes       ***              -95247 to -60111
  SPAdes 41 vs SPAdes 127             -43688           9.124             Yes       ***              -61256 to -26121
  SPAdes 77 vs SPAdes 99              -40833           8.527             Yes       ***              -58401 to -23265
  SPAdes 77 vs SPAdes 127             -6842            1.429             No        ns               -24410 to 10725
  SPAdes 99 vs SPAdes 127             33991            7.098             Yes       ***              16423 to 51558
```


Mash Distances of Assemblies
Distance calculated using Mash
0.00015
0.00010
0.00005
Ref-Skesa
Ref-Spades99
Skesa-Spades99
Pairwise comparisons of Assemblies
assembly
Ref-Skesa
Ref-Spades99
Skesa-Spades99

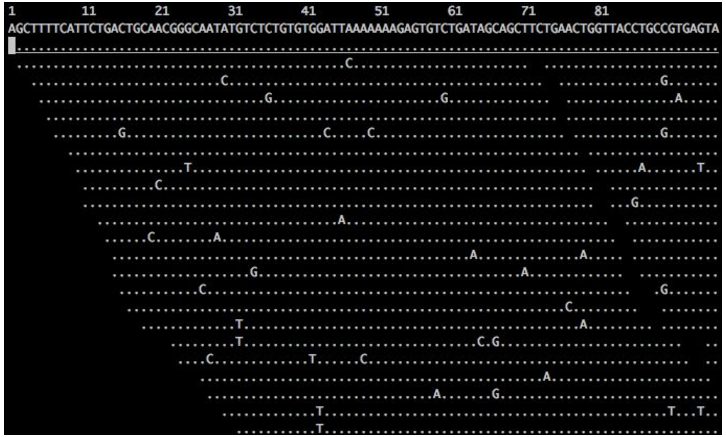


Reference Genome Sequence

35 bp identified

330 - 430 bp unknown sequence

35 bp identified



```
1         11        21        31        41        51        61        71        81
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTA
█................................................................................................
 .............................................C........................  ........................
   ..........................C...........................................  ..............G.......
    ...............................G.......................G..............  ...............A.....
     ...................................................................... .....................
      .........G...........................C.....C.......................... ............G.......
        ...................................................................... ..................
         ...............T...................................................... ......A.......T..
          ..........C...........................................................  ...............
          ......................................................................  ...G...........
            .................................A....................................  .............
             ......C........A......................................................  ............
              .................................................A..............A..... ............
              ...................G....................................A.............  ...........
               ...........C..........................................................   .G.......
                ............................................................C.........   ........
                  .............T..............................................A......... ........
                      .........T................................C.G.........................   ..
                       ....C.............T......C............................................  ..
                          ...............................................A......................
                           ...............................A.......G..............................
                             .............T...............................................T...T..
                               ...........T......................................................
```

# Supplementary: SPAdes Kmer Selection

For multicell paired end 250bp data:

It suggests:

```
spades.py -k 21,33,55,77,99,127 --careful <your reads> -o spades_output
```

Kmer selection can be tricky.