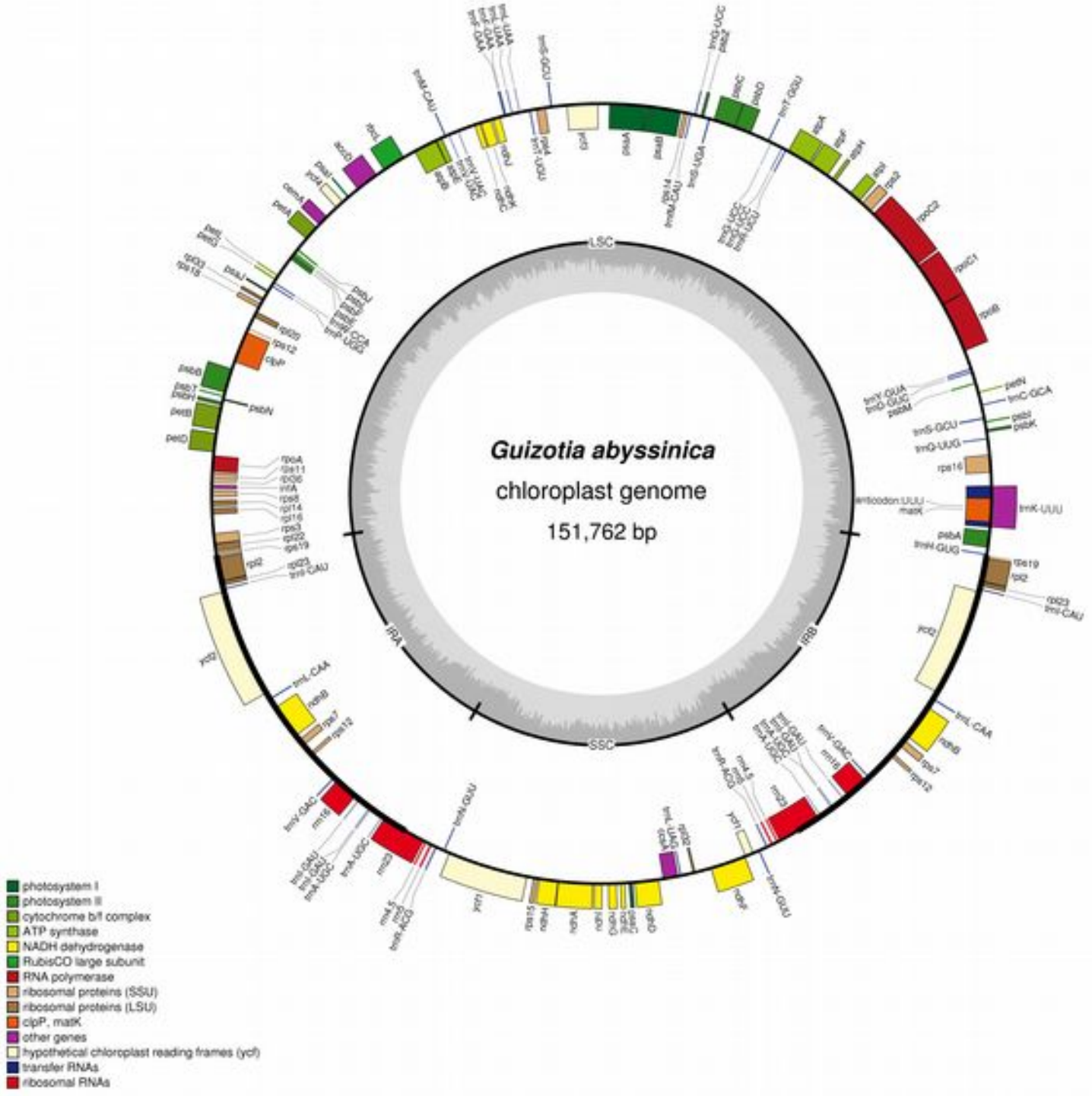
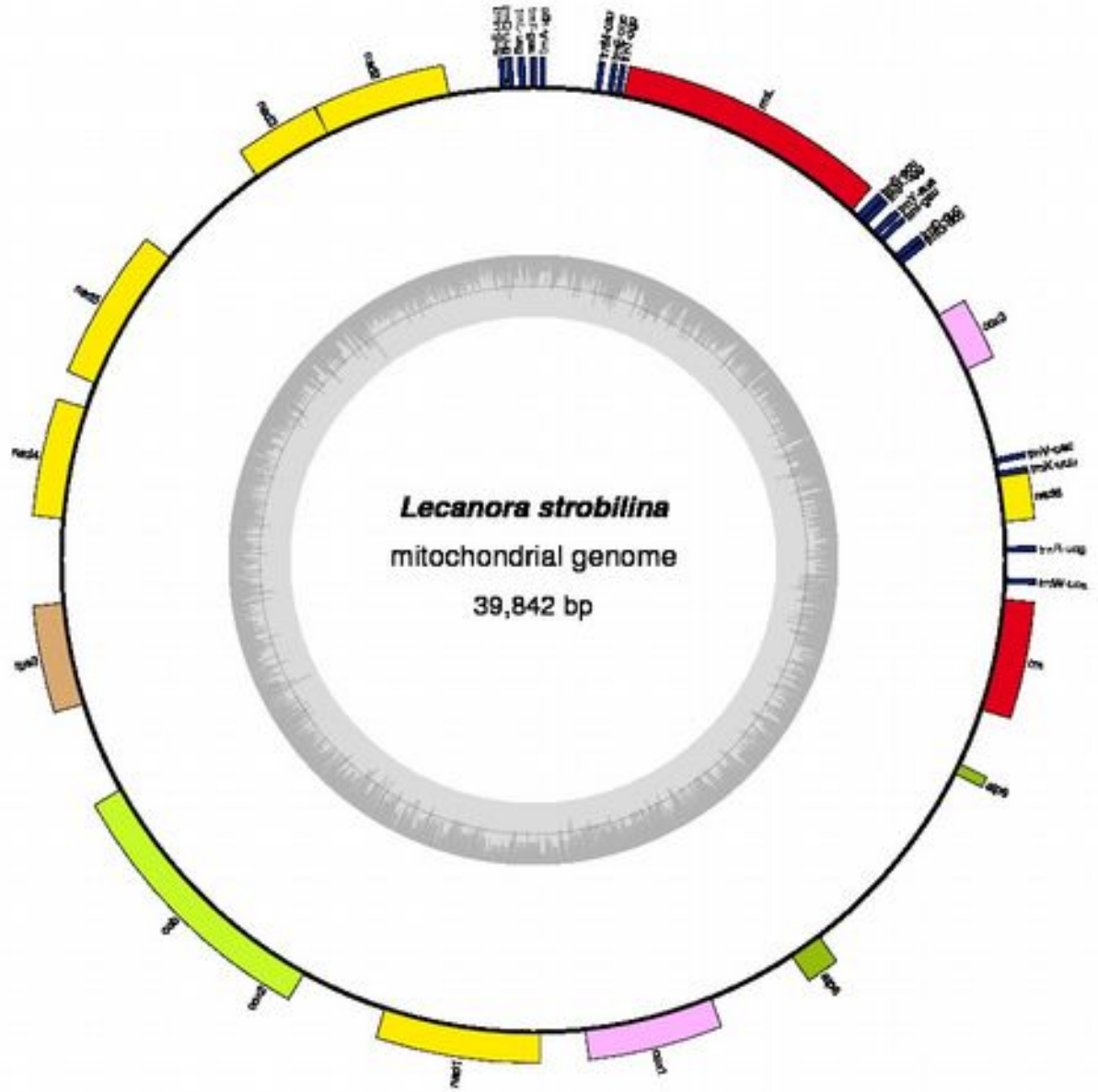
De novo genome assembly

```
@HWI-ST765:7:1101:1318:2091#0/1
GGCCACCTATGACCGGCTCGCGCCGCTCGTCGGGGAGCGGCTGCTCGTCGTCACCGGGGGCGCGCCCGCGGACGCCGTCCGCGGCCCGCTCCGCGCGCCCC
+
___ccccgggggghhhhhh^b^^c__UZFLZWacdBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST765:7:1101:1628:2156#0/1
TCTTCGCGAGTATGTCTGTTGATGGCGCTGTGTCCTATCTGCTCAAGGAAAGCAGCCCAACTCAATGTGTTACGCATTAGCGGCATTTGCTACATAATCCG
+
___eeeeefggggf_bddgeafgihdgehhgfeghhhifbgfhhhhihhhihhdhigggfeede`d`]bbdbccccccccccccccccbcb`b`bdbcbcc
@HWI-ST765:7:1101:2627:2192#0/1
ATTATGAAGACTGGAGAAAGCCCTATATTTATTGTATTTCTTTTCTGGATCACAAAATCCTCCCCTCTGAAACAAAAGATGTAGTTGGAATAAATAAAAGG
+
bbbeeeeegfgeggfhfffefghiiiihiihhhhfghhicegihiihhiihfhiiiiihfihiifhihhihihfdggeceeecee_bdddbccbcddbccb_
@HWI-ST765:7:1101:3236:2246#0/1
GCGGAAAGAGGGCTTGAGGATGACTTCCCTCATAGACTGGGACCCCCACTTTGAGGTGGCTGACGTAGCCTTTAAACGGAGTCCCCGCATTCCCGGTATCT
+
bbbeeeeegfgggiiihiihhiiiiiiiihiiihiiiiiihiiiiiiiiiiighhgggfeeeec`cdccccc`bccccccˆbcccac]aacdccc[_ccd
@HWI-ST765:7:1101:3400:2241#0/1
GCGGACAGCTAATGCGTTCCACTTATTGAACAGGGTTCTATGGTCGGTCCGTGACCCCCGGATGCCGAAGGCGTCCTTGGGGTAATCTCGTAGTTCCTACG
+
___cacc_eeaegfffZa`e]]de`egdfg[cgffcgZf]eˆaXˆG[Ze_agfffddgc`bXZˆ[]_aaa_GTTTTW_SX`aTX]`_bbaa_aacY`bbRO
@HWI-ST765:7:1101:4139:2060#0/1
NCTTCTCTCTTCATCAGAGAGTAGAGGTTGGGGCAATTGTGGGATCACGACGGGGACAGGGGCAGGTGCGGGCGGCGTCTCCGGTTGAGGAAGAGGCTGCC
+
BS\cceeeggggiiiiiihifgiiiiffhiiiiiiiiighiihiiiiiiiiggeccccccccccT___acX_c]][]acc_cT[_`bcbaa``caaaˆˆ
@HWI-ST765:7:1101:4188:2089#0/1
ACAAGATATATTTGATATACTAAGATGATAGCTAGAGACTAGAGATGAGAGTGCAGGATCTAGATTTGTAACAAATATTCGACTTTGCTTATGCAAACTGT
+
bbbeeeeeggggiiiiiiiiiiiiiiiiiiiihifghiiiihiiiiiifghiiiiiiiiihiiiiiihiiihhiihiihhggggeeeeeeddddddcc
```


Guizotia abyssinica
chloroplast genome
151,762 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs


Lecanora strobilina
mitochondrial genome
39,842 bp
complex I (NADH dehydrogenase)
complex III (ubichinol cytochrome c reductase)
complex IV (cytochrome c oxidase)
ATP synthase
ribosomal proteins (SSU)
transfer RNAs
ribosomal RNAs

De novo genome assembly

```
@HWI-ST765:7:1101:1318:2091#0/1
GGCCACCTATGACCGGCTCGCGCCGCTCGTCGGGGAGCGGCTGCTCGTCGTCACCGGGGGCGCGCCCGCGGACGCCGTCCGCGGCCCGCTCCGCGCGCCCC
+
___ccccgggggghhhhhh^b^^c__UZFLZWacdBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST765:7:1101:1628:2156#0/1
TCTTCGCGAGTATGTCTGTTGATGGCGCTGTGTCCTATCTGCTCAAGGAAAGCAGCCCAACTCAATGTGTTACGCATTAGCGGCATTTGCTACATAATCCG
+
___eeeeefggggf_bddgeafgihdgehhgfeghhhifbgfhhhhihhhihhdhigggfeede`d`]bbdbccccccccccccccccbcb`b`bdbcbcc
@HWI-ST765:7:1101:2627:2192#0/1
ATTATGAAGACTGGAGAAAGCCCTATATTTATTGTATTTCTTTTCTGGATCACAAAATCCTCCCCTCTGAAACAAAAGATGTAGTTGGAATAAATAAAAGG
+
bbbeeeeegfgeggfhfffefghiiiihiihhhhfghhicegihiihhiihfhiiiiihfihiifhihhihihfdggeceeecee_bdddbccbcddbccb_
@HWI-ST765:7:1101:3236:2246#0/1
GCGGAAAGAGGGCTTGAGGATGACTTCCCTCATAGACTGGGACCCCCACTTTGAGGTGGCTGACGTAGCCTTTAAACGGAGTCCCCGCATTCCCGGTATCT
+
bbbeeeeegfgggiiihiihhiiiiiiiihiiihiiiiiihiiiiiiiiiiighhgggfeeeec`cdccccc`bccccccˆbcccac]aacdccc[_ccd
@HWI-ST765:7:1101:3400:2241#0/1
GCGGACAGCTAATGCGTTCCACTTATTGAACAGGGTTCTATGGTCGGTCCGTGACCCCCGGATGCCGAAGGCGTCCTTGGGGTAATCTCGTAGTTCCTACG
+
___cacc_eeaegfffZa`e]]de`egdfg[cgffcgZf]eˆaXˆG[Ze_agfffddgc`bXZˆ[]_aaa_GTTTTW_SX`aTX]`_bbaa_aacY`bbRO
@HWI-ST765:7:1101:4139:2060#0/1
NCTTCTCTCTTCATCAGAGAGTAGAGGTTGGGGCAATTGTGGGATCACGACGGGGACAGGGGCAGGTGCGGGCGGCGTCTCCGGTTGAGGAAGAGGCTGCC
+
BS\cceeeggggiiiiiihifgiiiiffhiiiiiiiiighiihiiiiiiiiggeccccccccccT___acX_c]][]acc_cT[_`bcbaa``caaaˆˆ
@HWI-ST765:7:1101:4188:2089#0/1
ACAAGATATATTTGATATACTAAGATGATAGCTAGAGACTAGAGATGAGAGTGCAGGATCTAGATTTGTAACAAATATTCGACTTTGCTTATGCAAACTGT
+
bbbeeeeeggggiiiiiiiiiiiiiiiiiiiihifghiiiihiiiiiifghiiiiiiiiihiiiiiihiiihhiihiihhggggeeeeeeddddddcc
```

Look at your dataset

- Look at the amount and quality of your reads
  - How much data do you have?
  - How good is it?

Look at your dataset

- Look at the amount and quality of your reads
  - How much data do you have?
  - How good is it?
    - fastqc

Trimming and cleaning Illumina

http://www.usadellab.org/cms/index.php?page=trimmomatic

java -jar /home/nkane/Trimmomatic-0.32/trimmomatic-0.32.jar SE -threads 4 -phred33 sra_data.fastq trimmed.fq LEADING:30 TRAILING:30 MINLEN:35

Trimming and cleaning Illumina

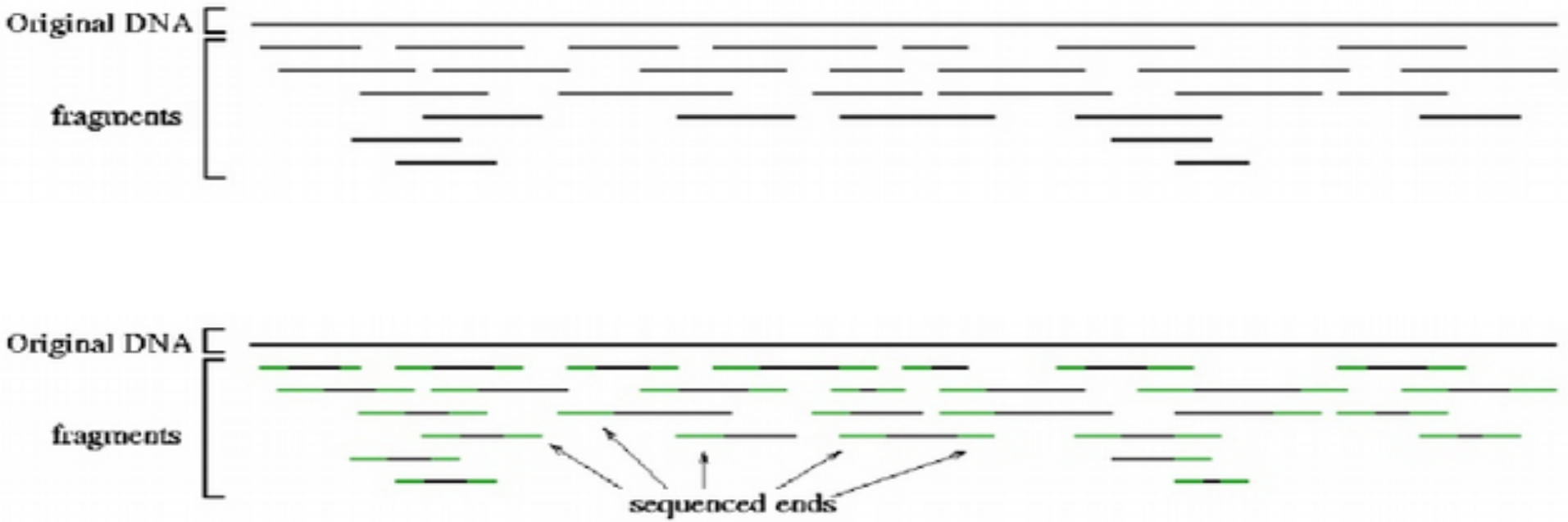
http://www.usadellab.org/cms/index.php?page=trimmomatic

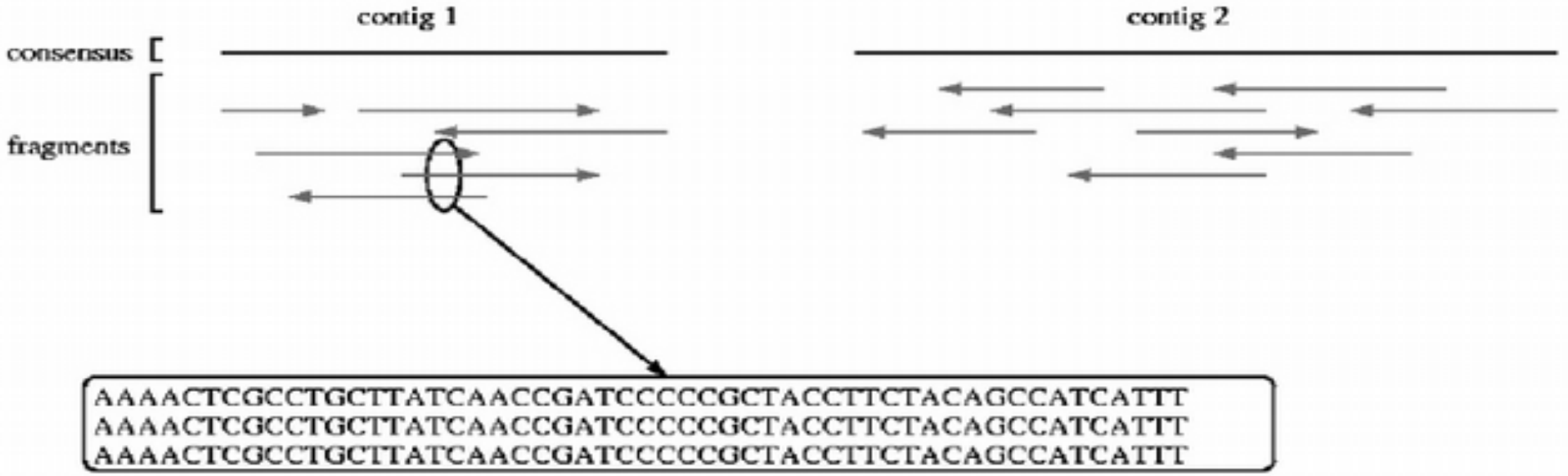
java -jar /home/nkane/Trimmomatic-0.32/trimmomatic-0.32.jar PE -threads 4 -phred33 Species_name_1.fq Species_name_2.fq Species_trim_1_paired.fq.gz Species_trim_1_unpaired.fq.gz Species_trim_2_paired.fq.gz Species_trim_2_unpaired.fq.gz LEADING:30 TRAILING:30 MINLEN:35

fastqc

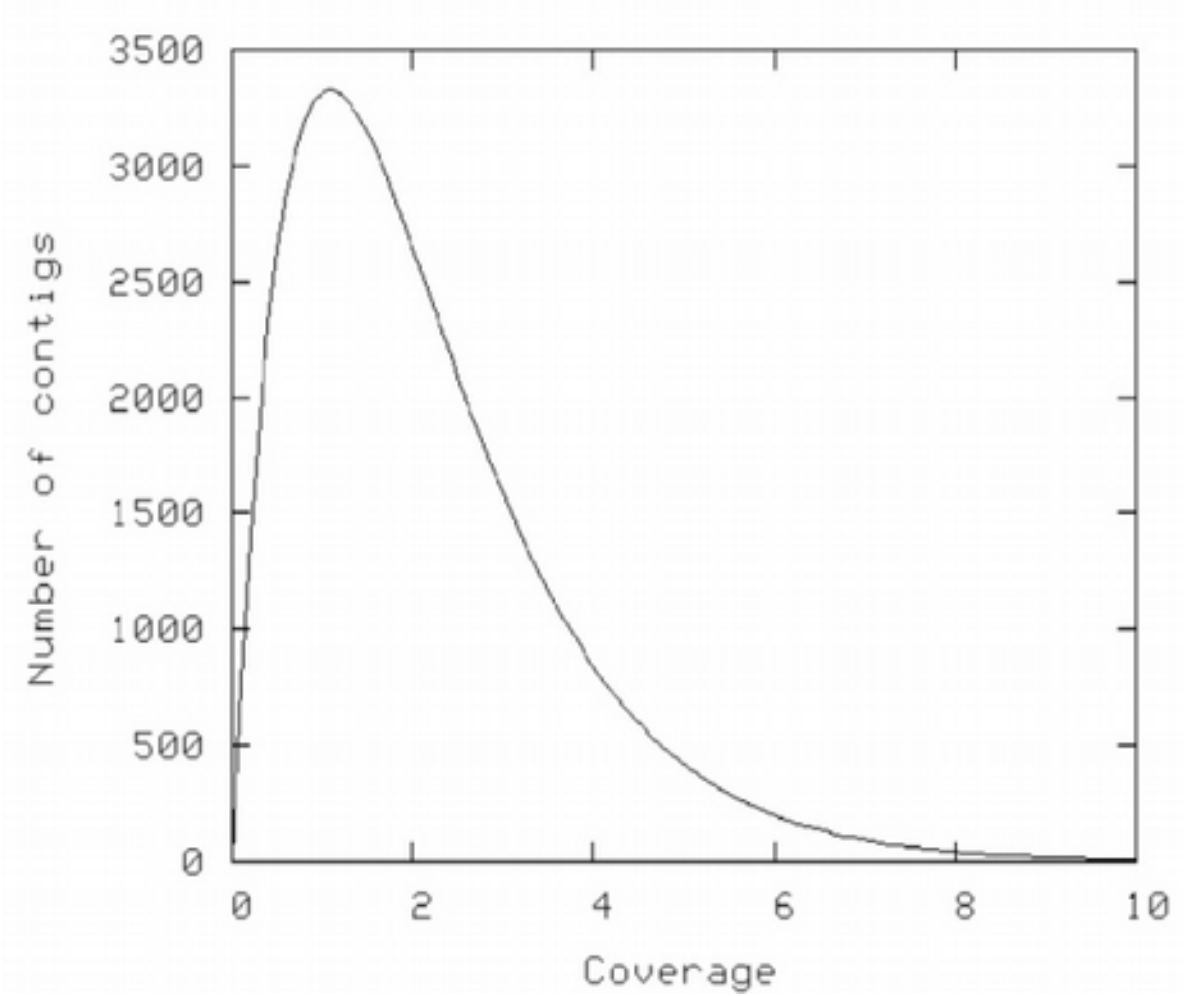
- Look at the quality of your reads again after trimming!

DNA extraction, sequencing, assembly

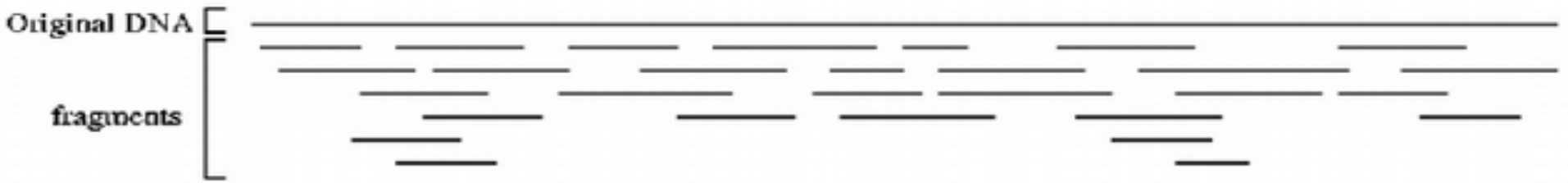




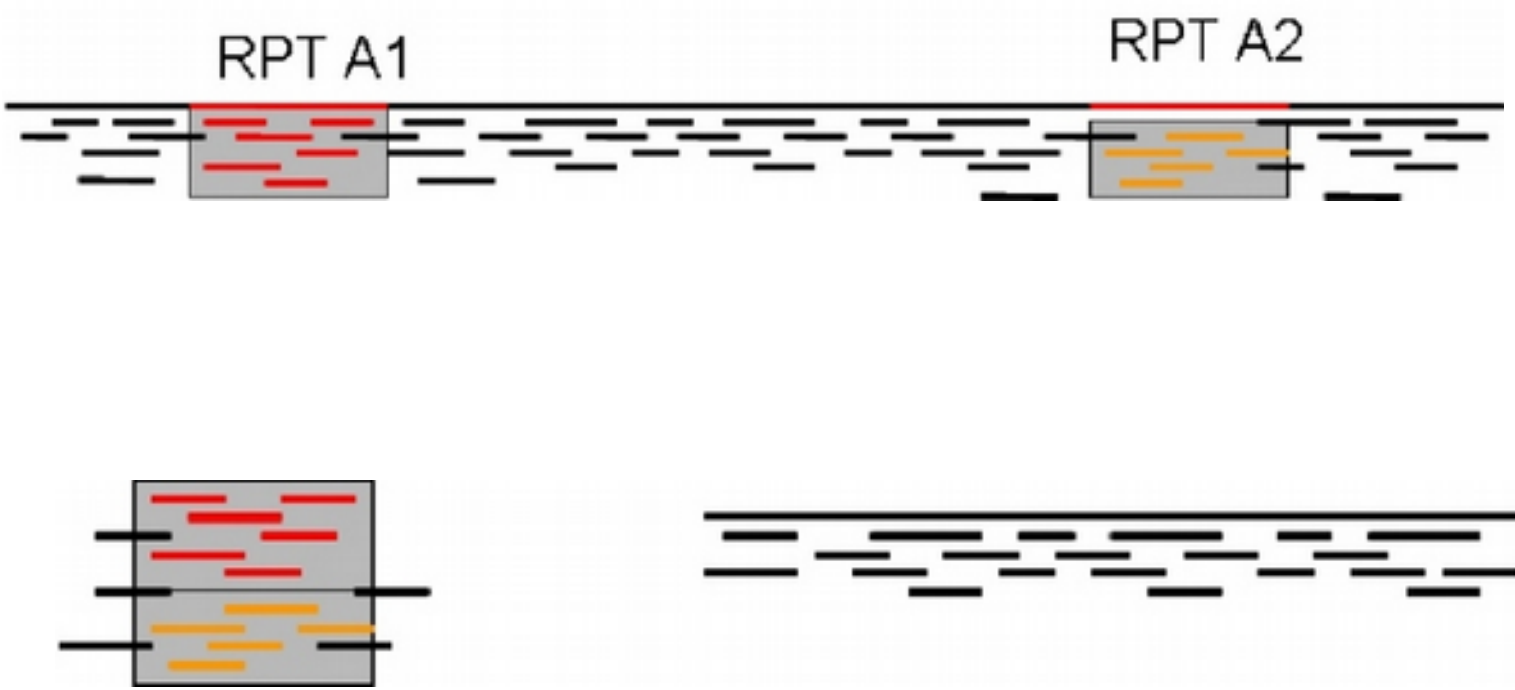
Number of contigs vs. genome coverage



DNA extraction, sequencing, assembly



Repeats can cause challenges



Assembly algorithms

- Overlap-layout-consensus

Assembly algorithms

- De Bruin graph

What is a k-mer?

- A k-mer is a string (sequence of letters) of length k
- ATGTAATAATG
- ATGT

  TGTA

  GTAA

  TAAT

  AATA

  ATAA

  TAAT

  AATG

Assembly algorithms

- it was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

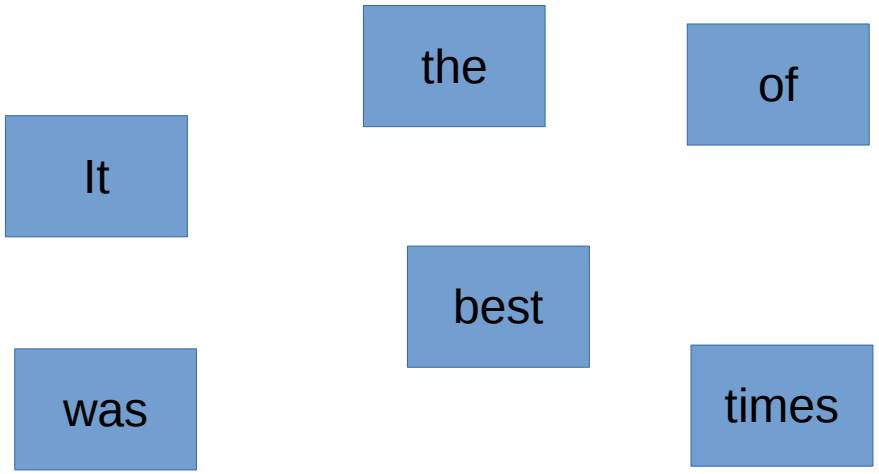
Assembly algorithms

- it was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

it was the best	was the age of
it was the age	it was the
age of foolishness	wisdom, it was
it was the worst	was the best of
times, it was	the best of times

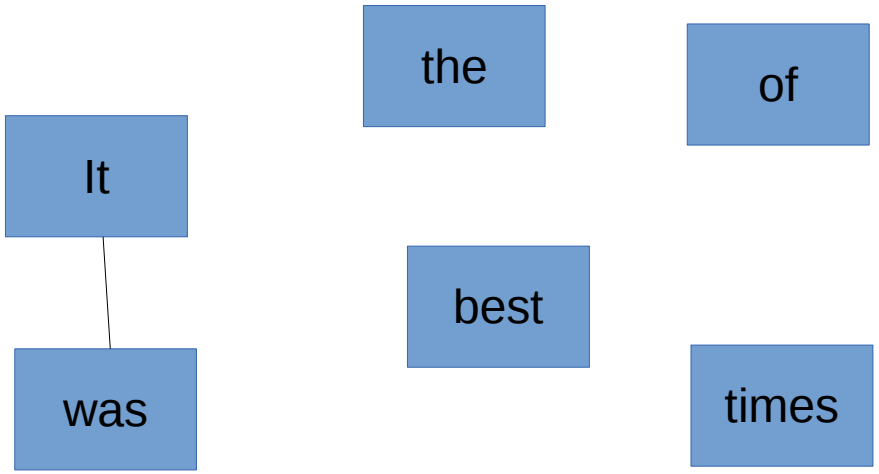
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



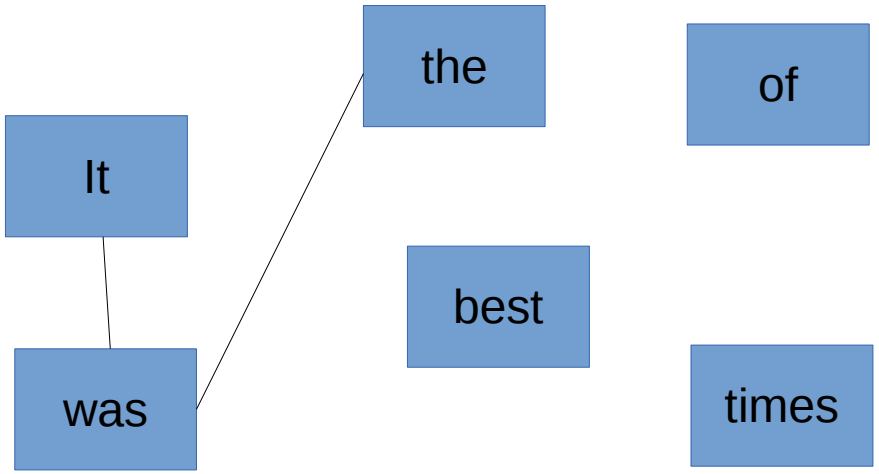
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



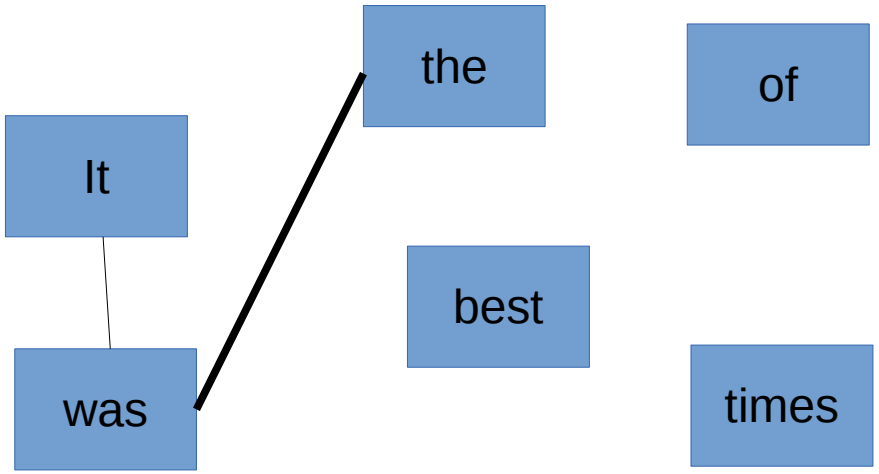
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



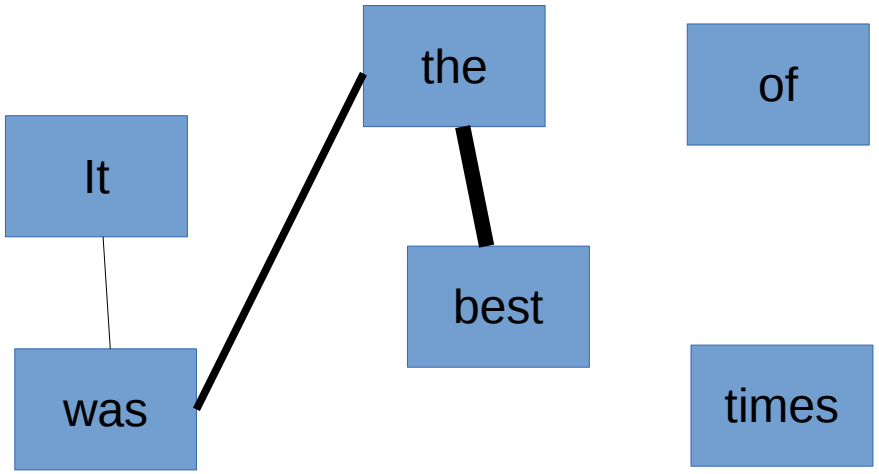
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



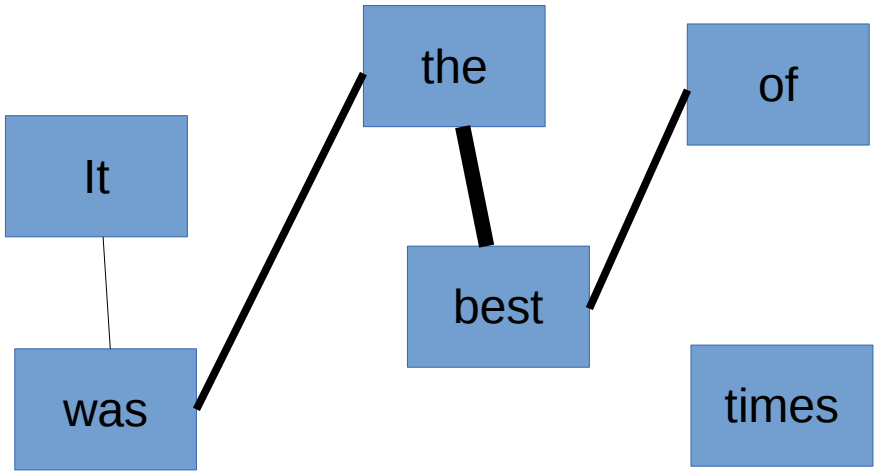
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



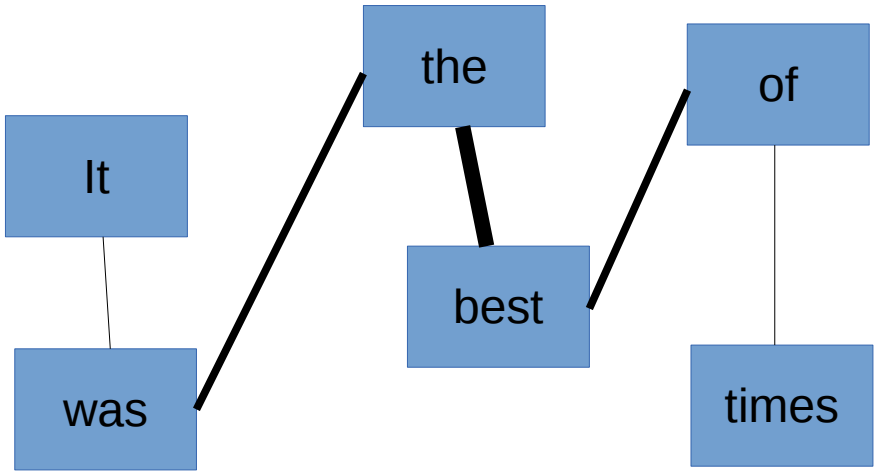
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



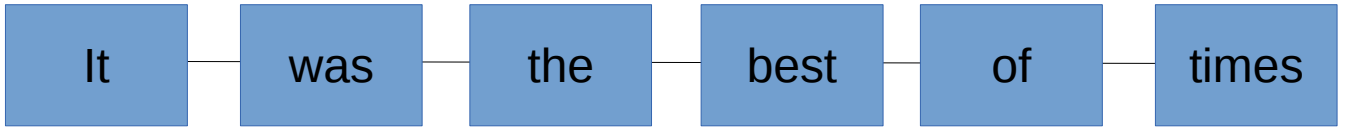
Joining reads with k-mers

- It was the best
- was the best of
- the best of times



Joining reads with k-mers

- It was the best
- was the best of
- the best of times



Assembly algorithms

- it was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

it was the best	was the age of
it was the age	it was the
age of foolishness	wisdom, it was
it was the worst	was the best of
times, it was	the best of times

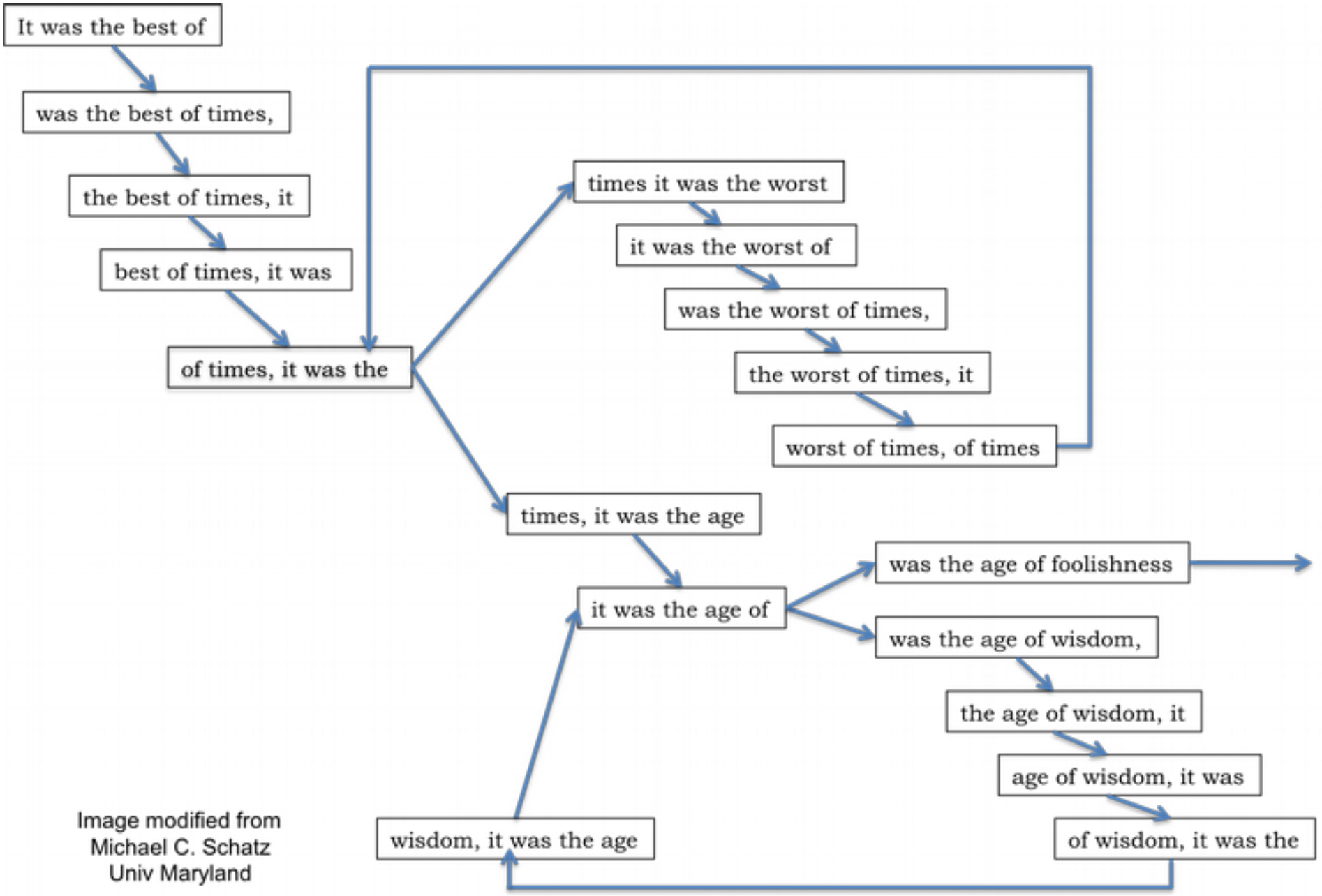
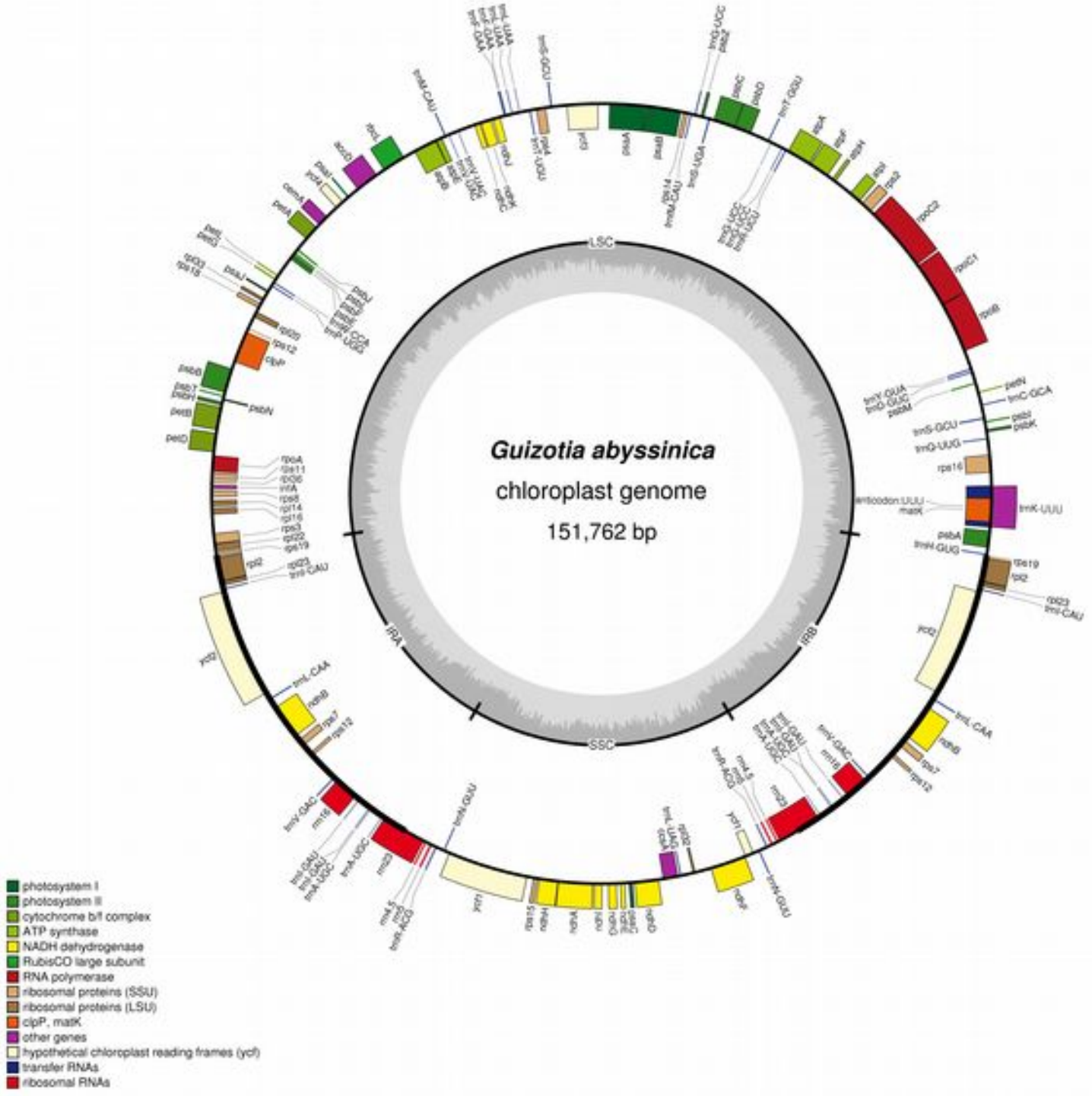
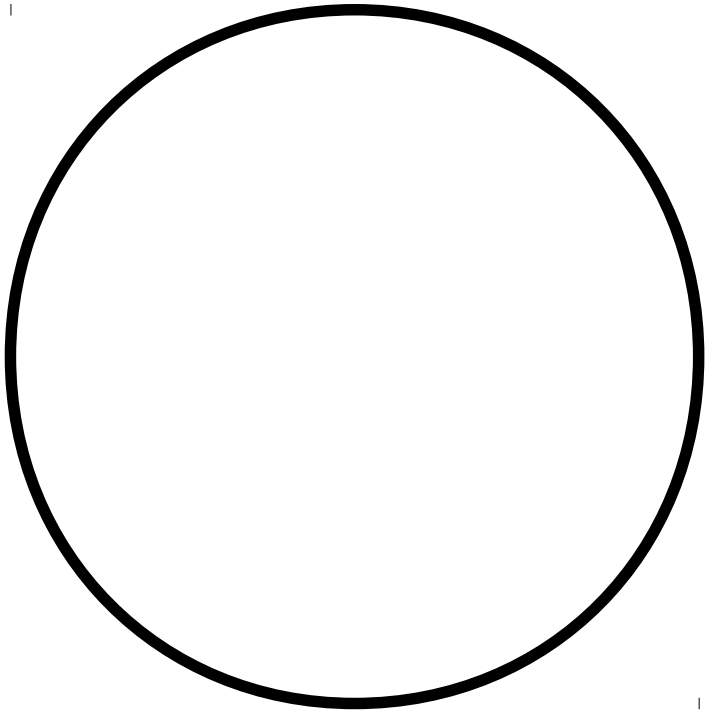


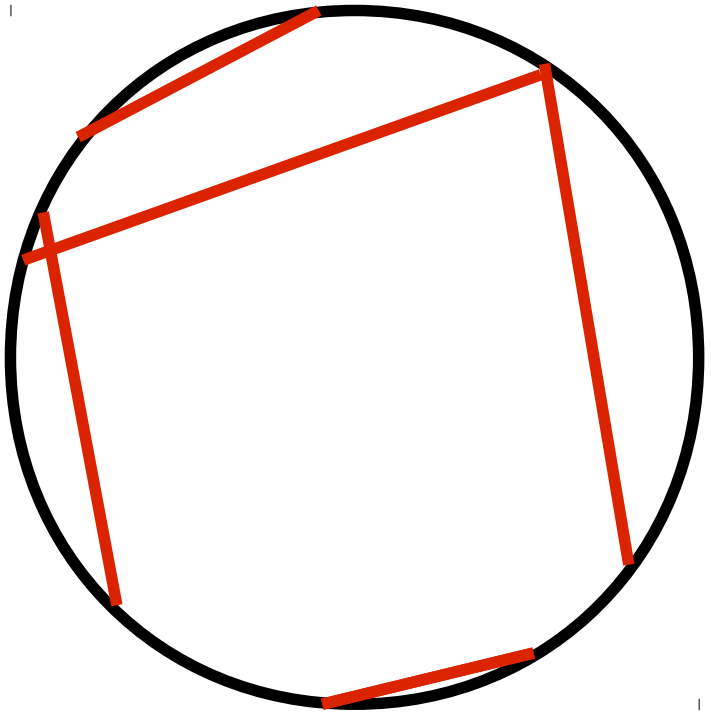
Image modified from
Michael C. Schatz
Univ Maryland

Assembly algorithms

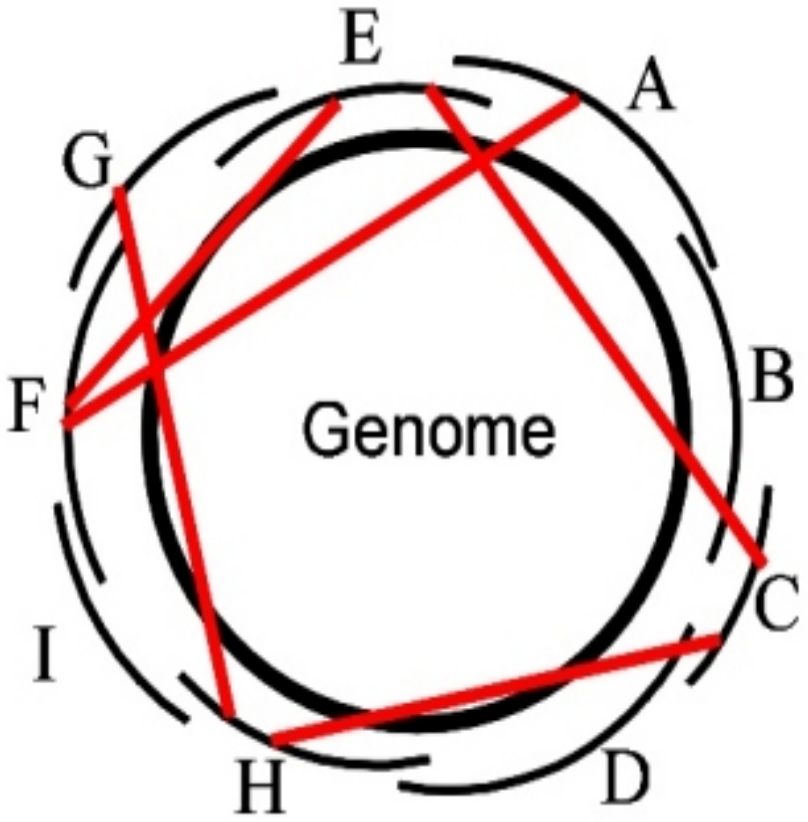
- It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way


Guizotia abyssinica
chloroplast genome
151,762 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs

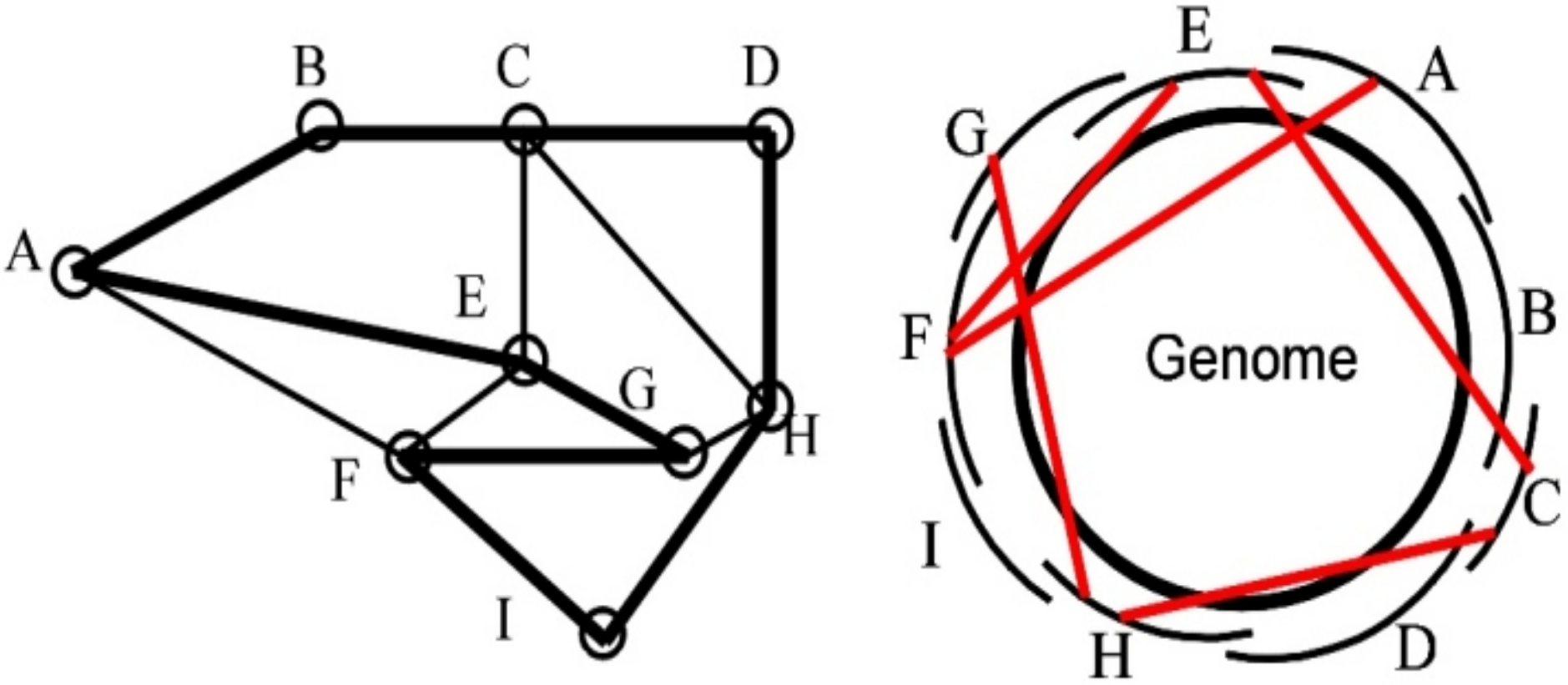




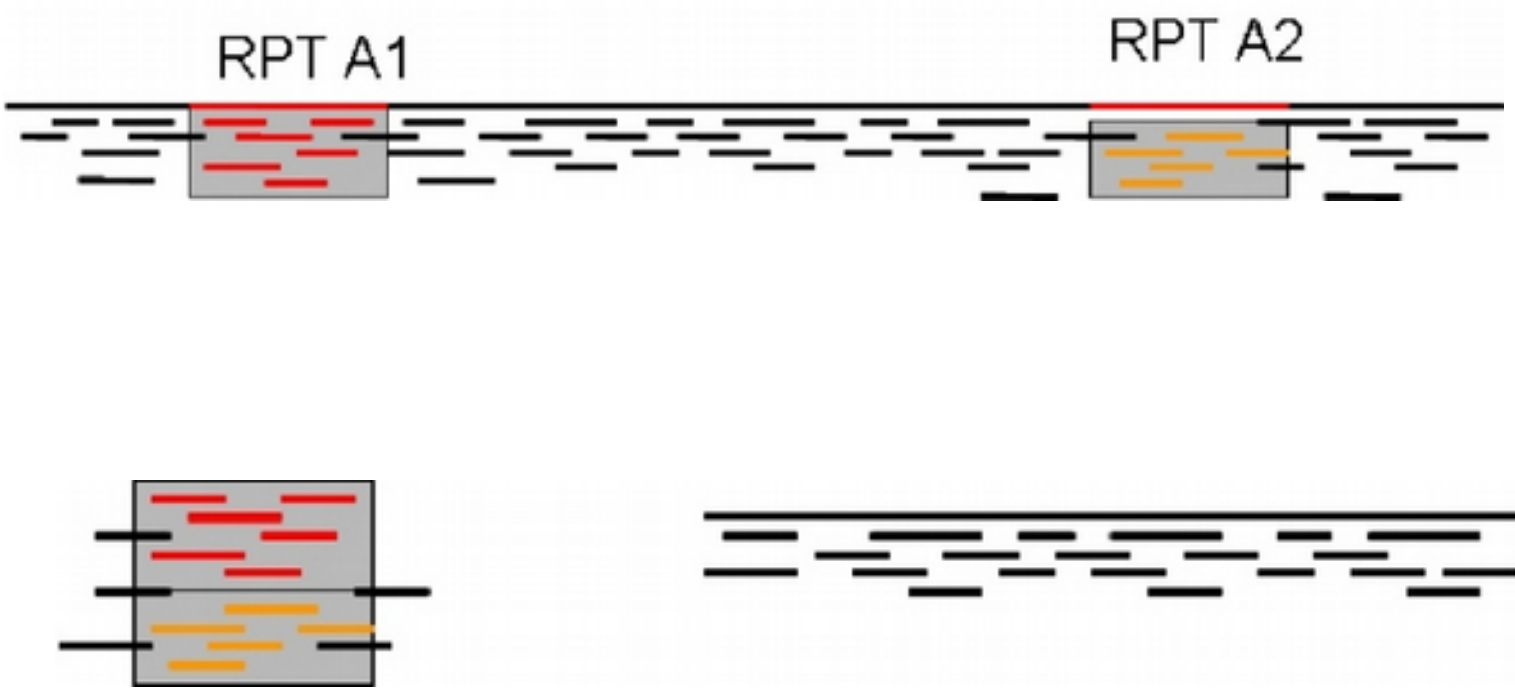
K-mers connect reads -> assembly



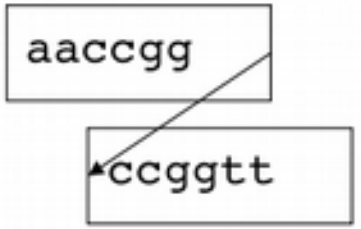
K-mers connect reads -> assembly



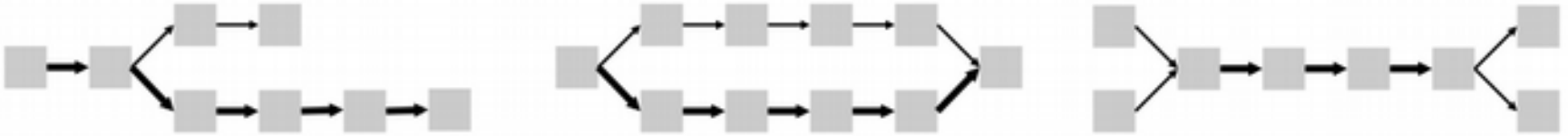
Repeats can cause errors



Graph Algorithms



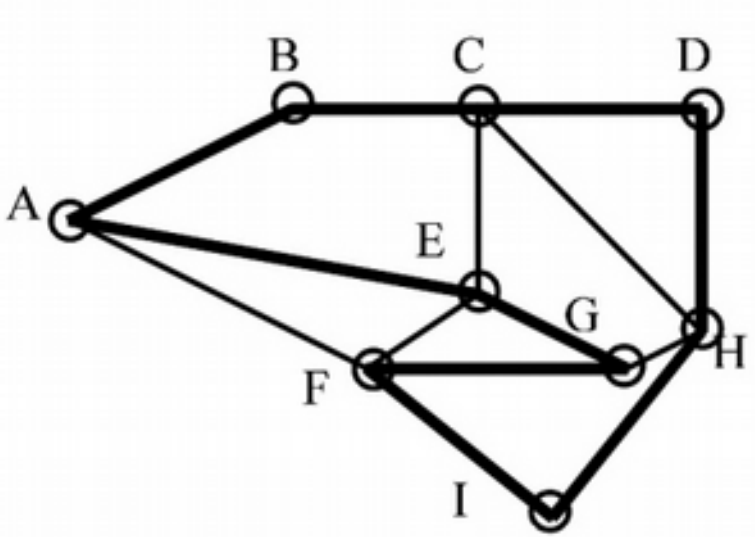
Assembly algorithms construct graphs, which are nodes connected by edges. In traditional assemblers (left) nodes are reads, edges are overlaps. In de Bruijn assemblers (right) nodes are K-mers, edges are exact matches of K-1.

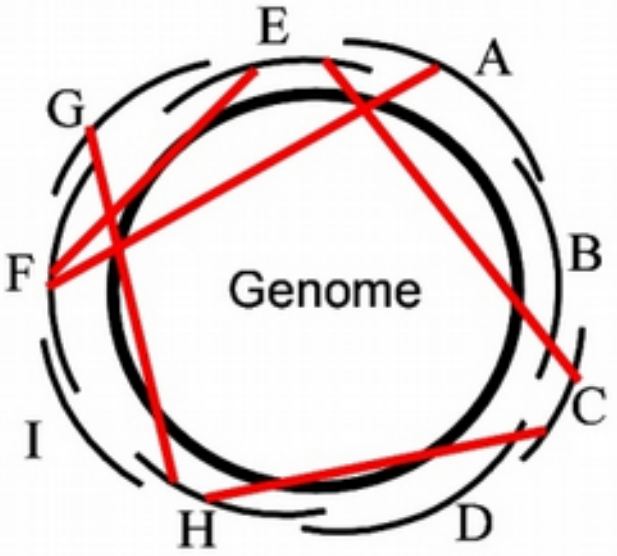


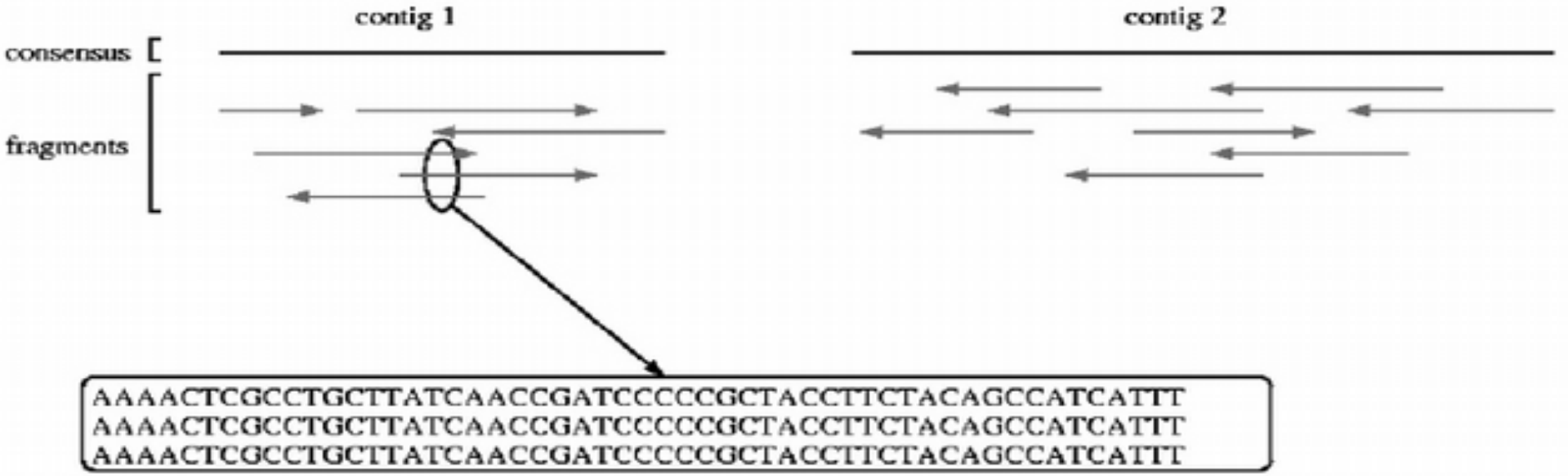
Assembly algorithms reduce graph complexity. Boxes are reads or K-mers. Edges are overlaps or matches. Edge thickness can indicate amount of support in reads.

Left: A spur is induced by bad sequence at a read end or low coverage and polymorphism.
Middle: A bubble is induced by polymorphism or sequencing error.
Right: A collapsed repeat might get teased apart or isolated or multiply placed.

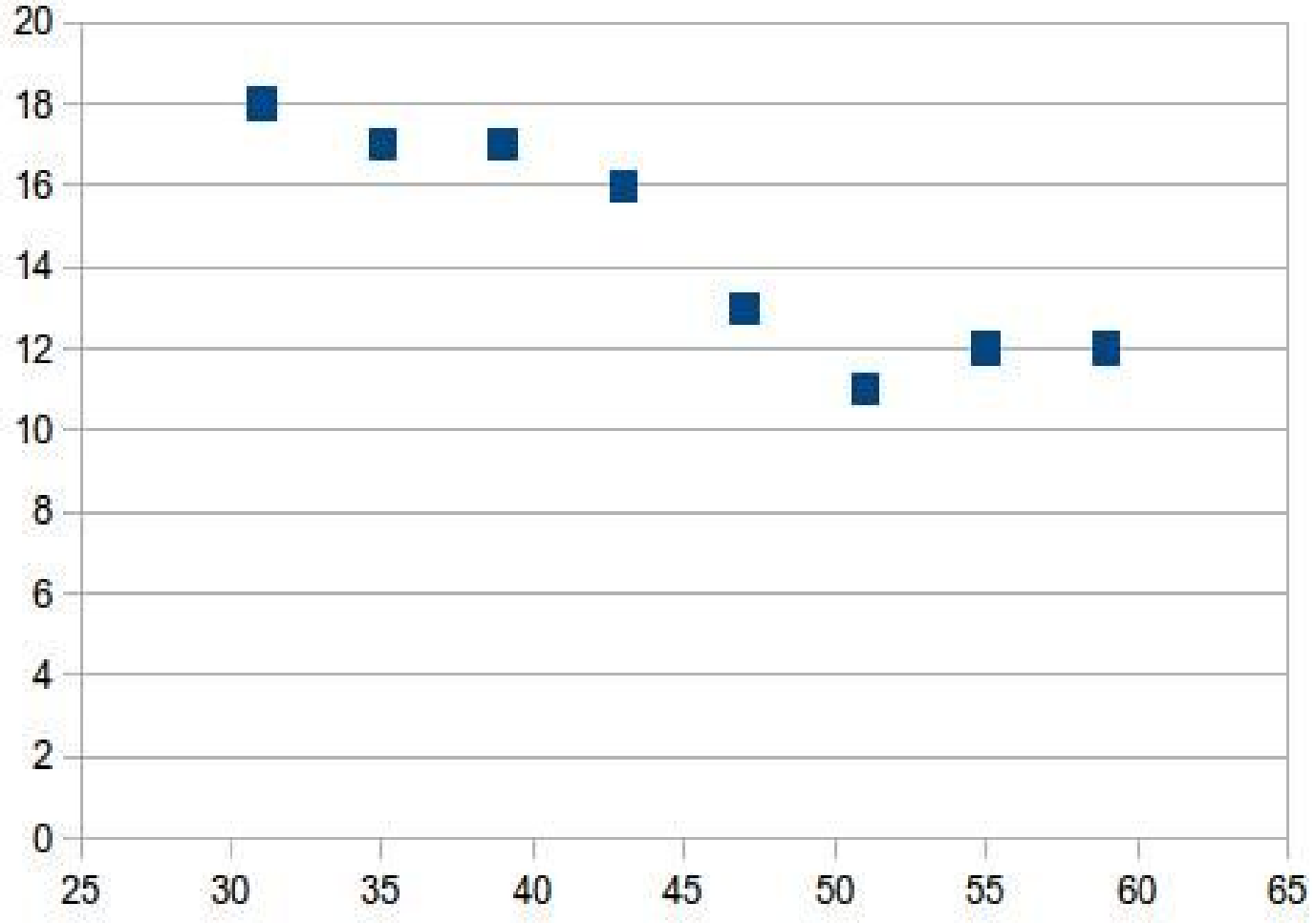
Evaluating the assembly – is it right? Which assembly is better?

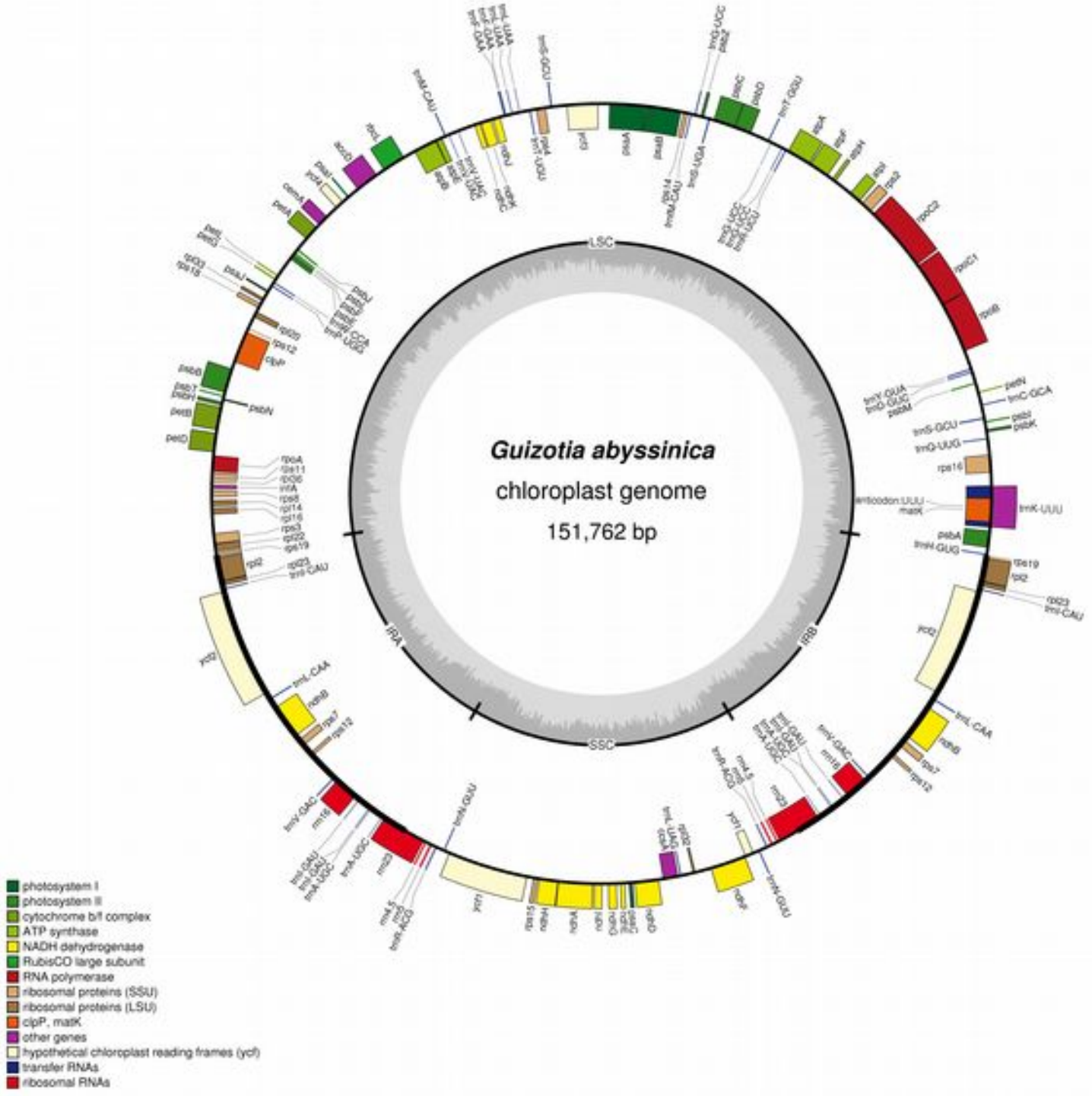






Assembly: varying kmer size




Guizotia abyssinica
chloroplast genome
151,762 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)
transfer RNAs
ribosomal RNAs