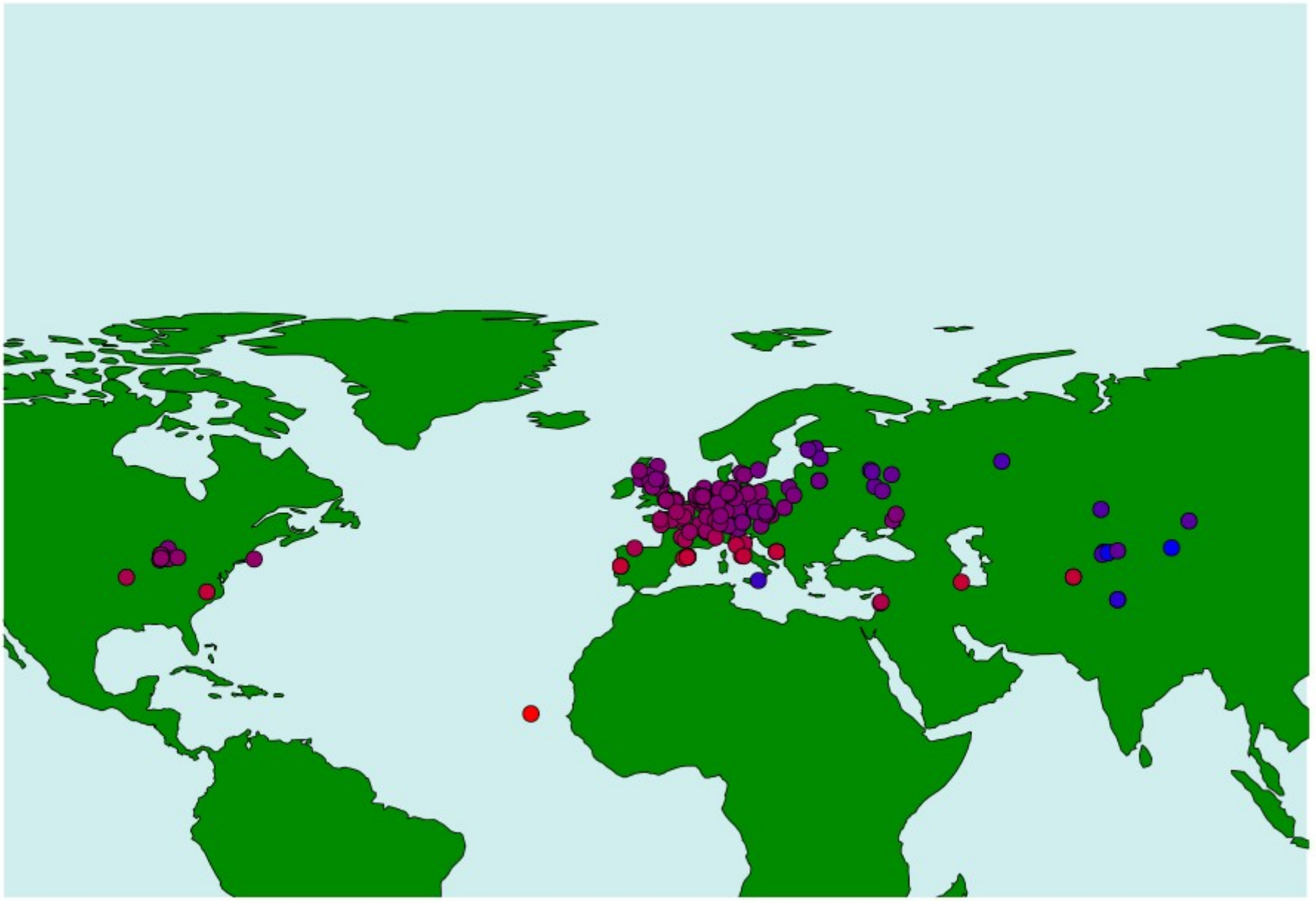
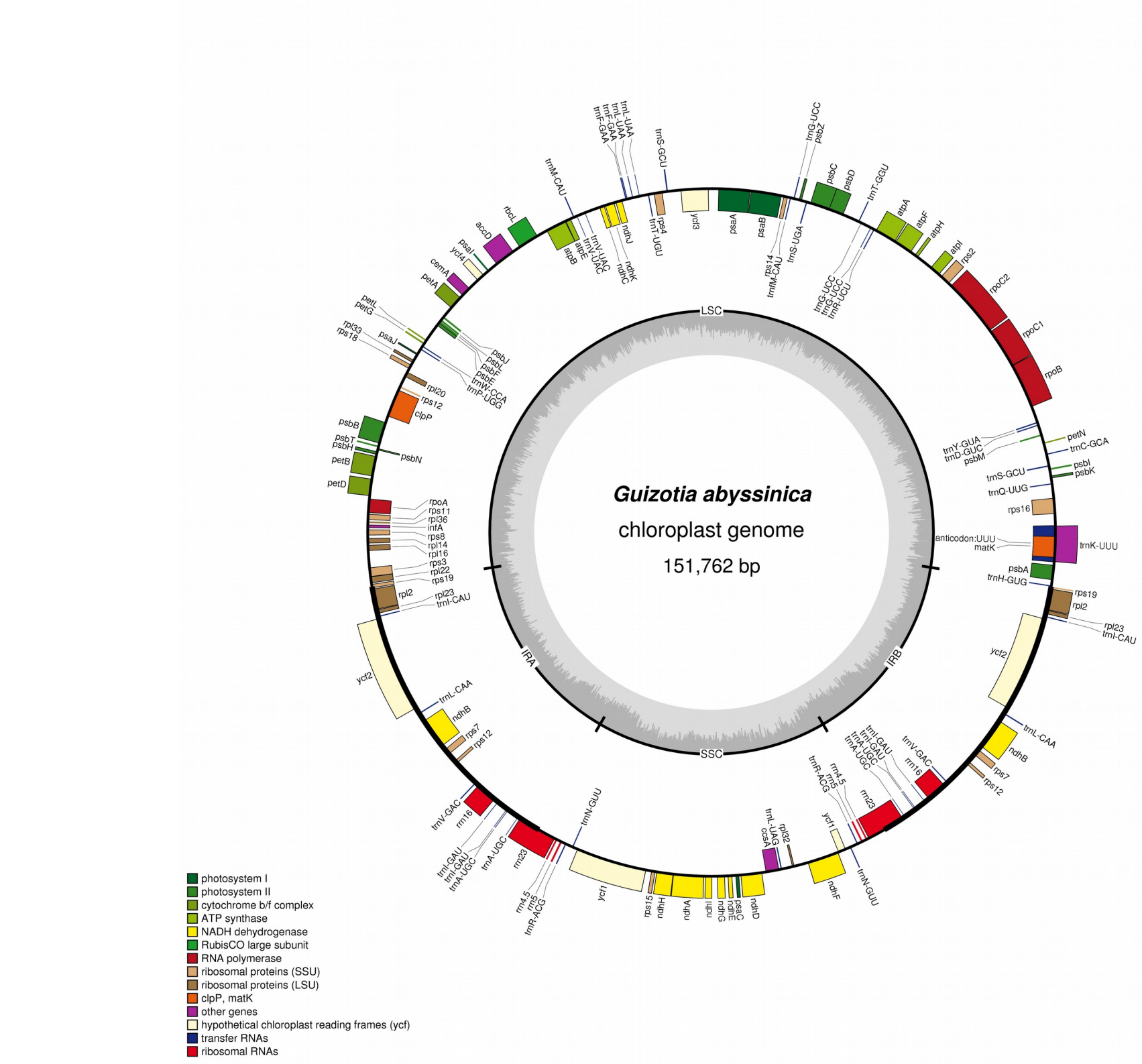
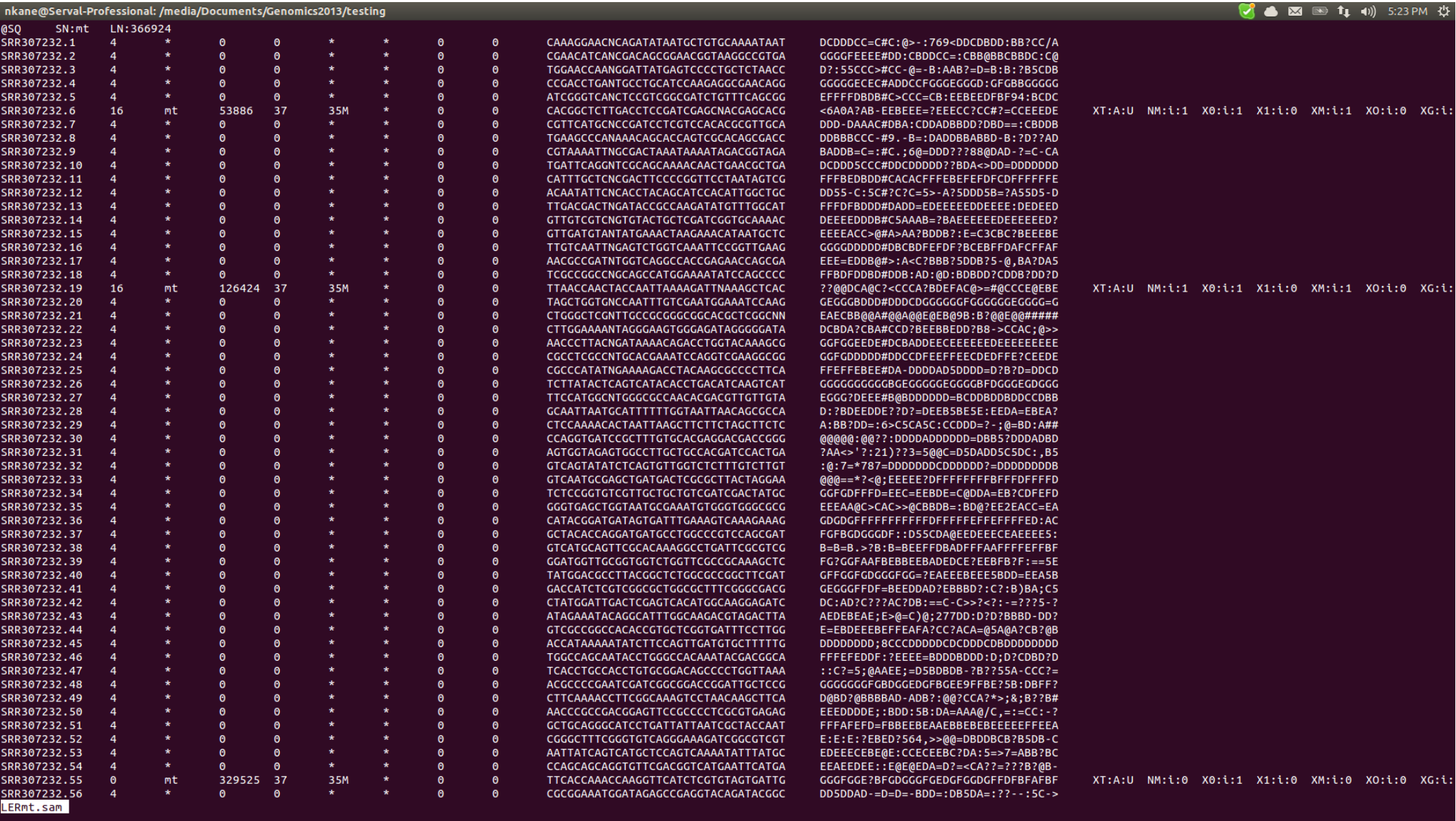



Guizotia abyssinica
chloroplast genome
151,762 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)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

SAM files

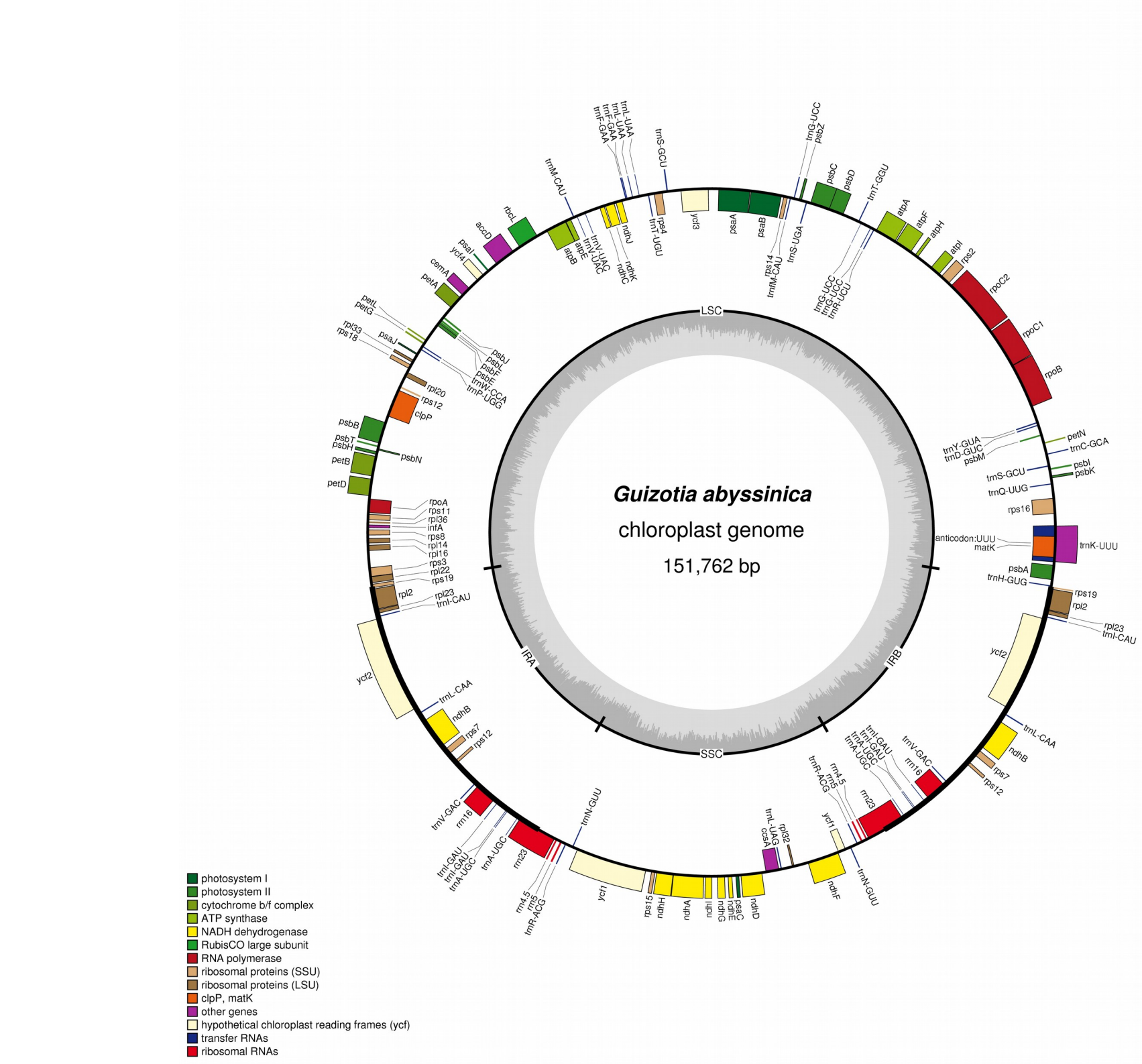

nkane@Serval-Professional: /media/Documents/Genomics2013/testing
5:23 PM
@SQ SN:mt LN:366924
SRR307232.1 4 * 0 0 * * 0 0 CAAAGGAACNCAGATATAATGCTGTGCAAAATAAT DCDDDCC=C#C:@>-:769<DDCDBDD:BB?CC/A
SRR307232.2 4 * 0 0 * * 0 0 CGAACATCANCGACAGCGGAACGGTAAGGCCGTGA GGGGFEEEE#DD:CBDDCC=:CBB@BBCBBDC:C@
SRR307232.3 4 * 0 0 * * 0 0 TGGAACCAANGGATTATGAGTCCCCTGCTCTAACC D?:55CCC>#CC-@=-B:AAB?=D=B:B:?B5CDB
SRR307232.4 4 * 0 0 * * 0 0 CCGACCTGANTGCCTGCATCCAAGAGGCGAACAGG GGGGGECEC#ADDCCFGGGEGGGD:GFGBBGGGGG
SRR307232.5 4 * 0 0 * * 0 0 ATCGGGTCANCTCCGTCGGCGATCTGTTTCAGCGG EFFFFDBDB#C>CCC=CB:EEBEEDFBF94:BCDC
SRR307232.6 16 mt 53886 37 35M * 0 0 CACGGCTCTTGACCTCCGATCGAGCNACGAGCACG <6A0A?AB-EEBEEE=?EEECC?CC#?=CCEEEDE XT:A:U NM:i:1 X0:i:1 X1:i:0 XM:i:1 XO:i:0 XG:i:
SRR307232.7 4 * 0 0 * * 0 0 CGTTCATGCNCCGATCCTCGTCCACACGCGTTGCA DDD-DAAAC#DBA:CDDADBBDD?DBD==:CBDDB
SRR307232.8 4 * 0 0 * * 0 0 TGAAGCCCANAAACAGCACCAGTCGCACAGCGACC DDBBBCCC-#9.-B=:DADDBBABBD-B:?D??AD
SRR307232.9 4 * 0 0 * * 0 0 CGTAAAATTNGCGACTAAATAAAATAGACGGTAGA BADDB=C=:#C.;6@=DDD???88@DAD-?=C-CA
SRR307232.10 4 * 0 0 * * 0 0 TGATTCAGGNTCGCAGCAAAACAACTGAACGCTGA DCDDD5CCC#DDCDDDDD??BDA<>DD=DDDDDDD
SRR307232.11 4 * 0 0 * * 0 0 CATTTGCTCNCGACTTCCCCGGTTCCTAATAGTCG FFFBEDBDD#CACACFFFEBEFEFDFCDFFFFFFE
SRR307232.12 4 * 0 0 * * 0 0 ACAATATTCNCACCTACAGCATCCACATTGGCTGC DD55-C:5C#?C?C=5>-A?5DDD5B=?A55D5-D
SRR307232.13 4 * 0 0 * * 0 0 TTGACGACTNGATACCGCCAAGATATGTTTGGCAT FFFDFBDDD#DADD=EDEEEEEDDEEEE:DEDEED
SRR307232.14 4 * 0 0 * * 0 0 GTTGTCGTCNGTGTACTGCTCGATCGGTGCAAAAC DEEEEDDDB#C5AAAB=?BAEEEEEEDEEEEEED?
SRR307232.15 4 * 0 0 * * 0 0 GTTGATGTANTATGAAACTAAGAAACATAATGCTC EEEEACC>@#A>AA?BDDB?:E=C3CBC?BEEEBE
SRR307232.16 4 * 0 0 * * 0 0 TTGTCAATTNGAGTCTGGTCAAATTCCGGTTGAAG GGGGDDDDD#DBCBDFEFDF?BCEBFFDAFCFFAF
SRR307232.17 4 * 0 0 * * 0 0 AACGCCGATNTGGTCAGGCCACCGAGAACCAGCGA EEE=EDDB@#>:A<C?BBB?5DDB?5-@,BA?DA5
SRR307232.18 4 * 0 0 * * 0 0 TCGCCGGCCNGCAGCCATGGAAAATATCCAGCCCC FFBDFDDBD#DDB:AD:@D:BDBDD?CDDB?DD?D
SRR307232.19 16 mt 126424 37 35M * 0 0 TTAACCAACTACCAATTAAAAGATTNAAAGCTCAC ??@@DCA@C?<CCCA?BDEFAC@>=#@CCCE@EBE XT:A:U NM:i:1 X0:i:1 X1:i:0 XM:i:1 XO:i:0 XG:i:
SRR307232.20 4 * 0 0 * * 0 0 TAGCTGGTGNCCAATTTGTCGAATGGAAATCCAAG GEGGGBDDD#DDDCDGGGGGGFGGGGGGEGGGG=G
SRR307232.21 4 * 0 0 * * 0 0 CTGGGCTCGNTTGCCGCGGGCGGCACGCTCGGCNN EAECBB@@A#@@A@@E@EB@9B:B?@@E@@#####
SRR307232.22 4 * 0 0 * * 0 0 CTTGGAAAANTAGGGAAGTGGGAGATAGGGGGATA DCBDA?CBA#CCD?BEEBBEDD?B8->CCAC;@>>
SRR307232.23 4 * 0 0 * * 0 0 AACCCTTACNGATAAAACAGACCTGGTACAAAGCG GGFGGEEDE#DCBADDEECEEEEEEDEEEEEEEEE
SRR307232.24 4 * 0 0 * * 0 0 CGCCTCGCCNTGCACGAAATCCAGGTCGAAGGCGG GGFGDDDDD#DDCCDFEEFFEECDEDFFE?CEEDE
SRR307232.25 4 * 0 0 * * 0 0 CGCCCATATNGAAAAGACCTACAAGCGCCCCTTCA FFEFFEBEE#DA-DDDDAD5DDDD=D?B?D=DDCD
SRR307232.26 4 * 0 0 * * 0 0 TCTTATACTCAGTCATACACCTGACATCAAGTCAT GGGGGGGGGGBGEGGGGGEGGGGBFDGGGEGDGGG
SRR307232.27 4 * 0 0 * * 0 0 TTCCATGGCNTGGGCGCCAACACGACGTTGTTGTA EGGG?DEEE#B@BDDDDDD=BCDDBDDBDDCCDBB
SRR307232.28 4 * 0 0 * * 0 0 GCAATTAATGCATTTTTTGGTAATTAACAGCGCCA D:?BDEEDDE??D?=DEEB5BE5E:EEDA=EBEA?
SRR307232.29 4 * 0 0 * * 0 0 CTCCAAAACACTAATTAAGCTTCTTCTAGCTTCTC A:BB?DD=:6>C5CA5C:CCDDD=?-;@=BD:A##
SRR307232.30 4 * 0 0 * * 0 0 CCAGGTGATCCGCTTTGTGCACGAGGACGACCGGG @@@@@:@@??:DDDDADDDDDD=DBB5?DDDADBD
SRR307232.31 4 * 0 0 * * 0 0 AGTGGTAGAGTGGCCTTGCTGCCACGATCCACTGA ?AA<>'?:21)??3=5@@C=D5DADD5C5DC:,B5
SRR307232.32 4 * 0 0 * * 0 0 GTCAGTATATCTCAGTGTTGGTCTCTTTGTCTTGT :@:7=*787=DDDDDDDCDDDDDD?=DDDDDDDDB
SRR307232.33 4 * 0 0 * * 0 0 GTCAATGCGAGCTGATGACTCGCGCTTACTAGGAA @@@==*?<@;EEEEE?DFFFFFFFFBFFFDFFFFD
SRR307232.34 4 * 0 0 * * 0 0 TCTCCGGTGTCGTTGCTGCTGTCGATCGACTATGC GGFGDFFFD=EEC=EEBDE=C@DDA=EB?CDFEFD
SRR307232.35 4 * 0 0 * * 0 0 GGGTGAGCTGGTAATGCGAAATGTGGGTGGGCGCG EEEAA@C>CAC>>@CBBDB=:BD@?EE2EACC=EA
SRR307232.36 4 * 0 0 * * 0 0 CATACGGATGATAGTGATTTGAAAGTCAAAGAAAG GDGDGFFFFFFFFFFFDFFFFFEFFEFFFED:AC
SRR307232.37 4 * 0 0 * * 0 0 GCTACACCAGGATGATGCCTGGCCCGTCCAGCGAT FGFBGDGGGDF::D55CDA@EEDEEECEAEEEE5:
SRR307232.38 4 * 0 0 * * 0 0 GTCATGCAGTTCGCACAAAGGCCTGATTCGCGTCG B=B=B.>?B:B=BEEFFDBADFFFAAFFFFEFFBF
SRR307232.39 4 * 0 0 * * 0 0 GGATGGTTGCGGTGGTCTGGTTCGCCGCAAAGCTC FG?GGFAAFBEBBEEBADEDCE?EEBFB?F:==5E
SRR307232.40 4 * 0 0 * * 0 0 TATGGACGCCTTACGGCTCTGGCGCCGGCTTCGAT GFFGGFGDGGGGFGG=?EAEEEBEEE5BDD=EEA5B
SRR307232.41 4 * 0 0 * * 0 0 GACCATCTCGTCGGCGCTGGCGCTTTCGGGCGACG GEGGGGFFDF=BEEDDAD?EBBBD?:C?:B)BA;C5
SRR307232.42 4 * 0 0 * * 0 0 CTATGGATTGACTCGAGTCACATGGCAAGGAGATC DC:AD?C???AC?DB:==C-C>>?<?:-=???5-?
SRR307232.43 4 * 0 0 * * 0 0 ATAGAAATACAGGCATTTGGCAAGACGTAGACTTA AEDEBEAE;E>@=C)@;277DD:D?D?BBBD-DD?
SRR307232.44 4 * 0 0 * * 0 0 GTCGCCGGCCACACCGTGCTCGGTGATTTCCTTGG E=EBDEEEBEFFEAFA?CC?ACA=@5A@A?CB?@B
SRR307232.45 4 * 0 0 * * 0 0 ACCATAAAAATATCTTCCAGTTGATGTGCTTTTTG DDDDDDDD;8CCCDDDDDCDCDDDCDBDDDDDDDD
SRR307232.46 4 * 0 0 * * 0 0 TGGCCAGCAATACCTGGGCCACAAATACGACGGCA FFFEFEDDF:?EEEE=BDDDBDDD:D;D?CDBD?D
SRR307232.47 4 * 0 0 * * 0 0 TCACCTGCCACCTGTGCGGACAGCCCCTGGTTAAA ::C?=5;@AAEE;=D5BDBDB-?B??55A-CCC?=
SRR307232.48 4 * 0 0 * * 0 0 ACGCCCCGAATCGATCGGCGGACCGGATTGCTCCG GGGGGGGFGBDGGEDGFBGEE9FFBE?5B:DBFF?
SRR307232.49 4 * 0 0 * * 0 0 CTTCAAAACCTTCGGCAAAGTCCTAACAAGCTTCA D@BD?@BBBBAD-ADB?:@@?CCA?*>;&;B??B#
SRR307232.50 4 * 0 0 * * 0 0 AACCCGCCGACGGAGTTCCGCCCCTCGCGTGAGAG EEEDDDDE;:BDD:5B:DA=AAA@/C,=:=CC:-?
SRR307232.51 4 * 0 0 * * 0 0 GCTGCAGGGCATCCTGATTATTAATCGCTACCAAT FFFAFEFD=FBBEEBEAAEBBEBEBEEEEEFFEEA
SRR307232.52 4 * 0 0 * * 0 0 CGGGCTTTCGGGTGTCAGGGAAAGATCGGCGTCGT E:E:E:?EBED?564,>>@@=DBDDBCB?B5DB-C
SRR307232.53 4 * 0 0 * * 0 0 AATTATCAGTCATGCTCCAGTCAAAATATTTATGC EDEEECEBE@E:CCECEEBC?DA:5=>7=ABB?BC
SRR307232.54 4 * 0 0 * * 0 0 CCAGCAGCAGGTGTTCGACGGTCATGAATTCATGA EEAEEDEE::E@E@EDA=D?=<CA??=???B?@B-
SRR307232.55 0 mt 329525 37 35M * 0 0 TTCACCAAACCAAGGTTCATCTCGTGTAGTGATTG GGGFGGE?BFGDGGGFGEDGFGGDGFFDFBFAFBF XT:A:U NM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:
SRR307232.56 4 * 0 0 * * 0 0 CGCGGAAATGGATAGAGCCGAGGTACAGATACGGC DD5DDAD-=D=D=-BDD=:DB5DA=:??--:5C->
LERmt.sam

Convert to fastq

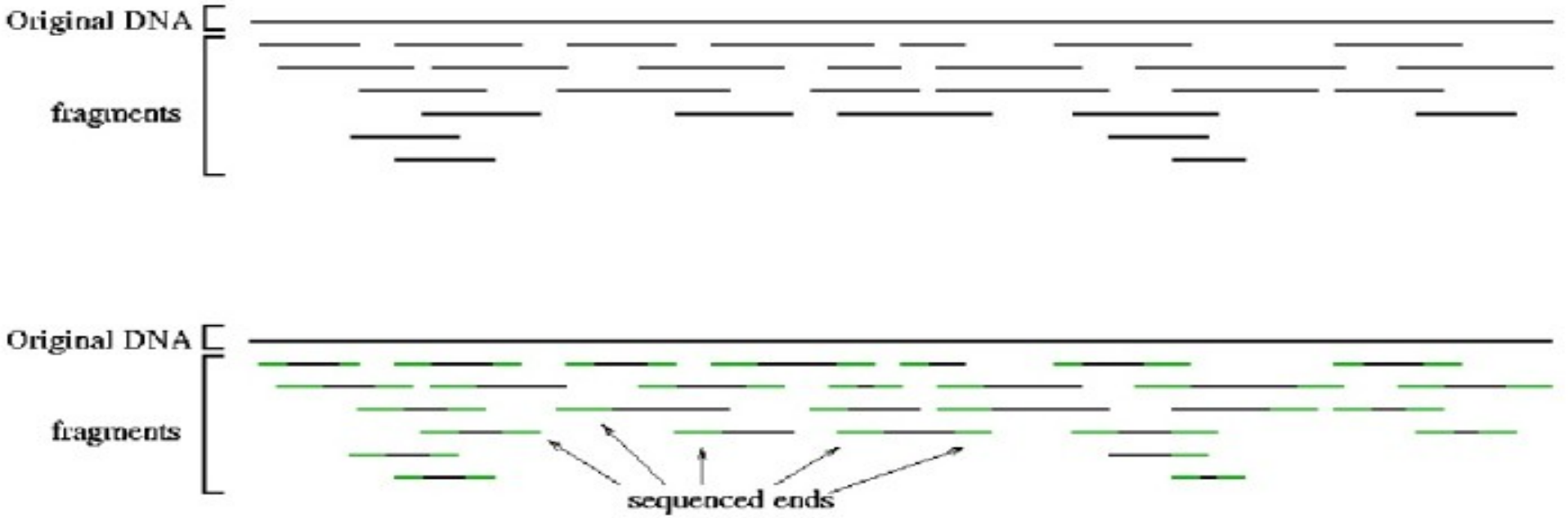
Convert to fastq

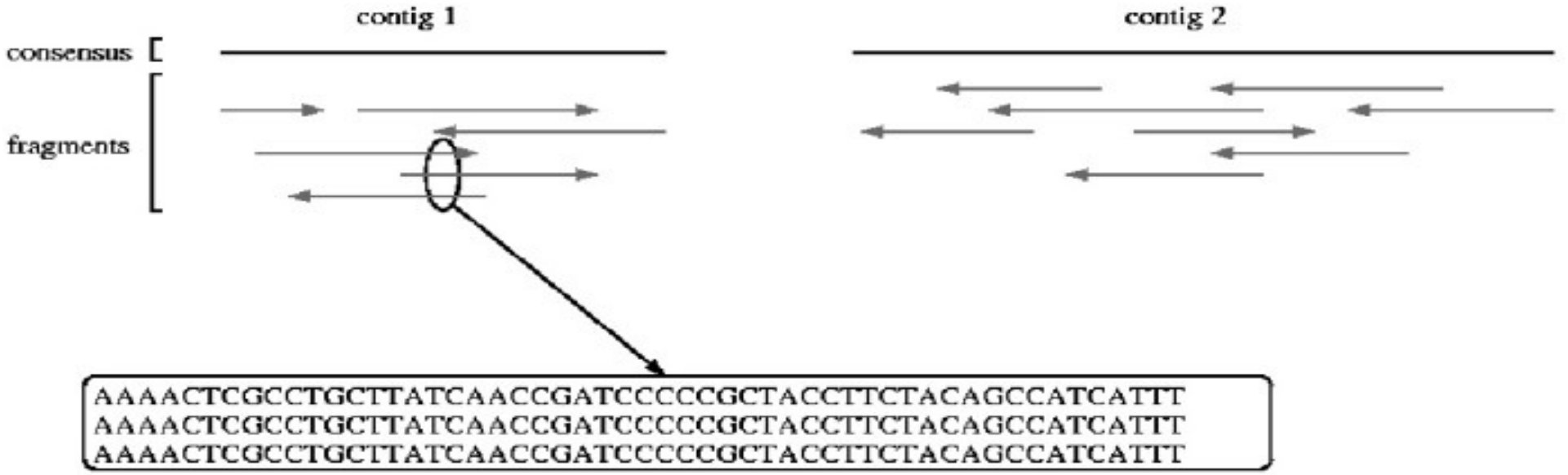
De novo genome assembly

```
@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
```

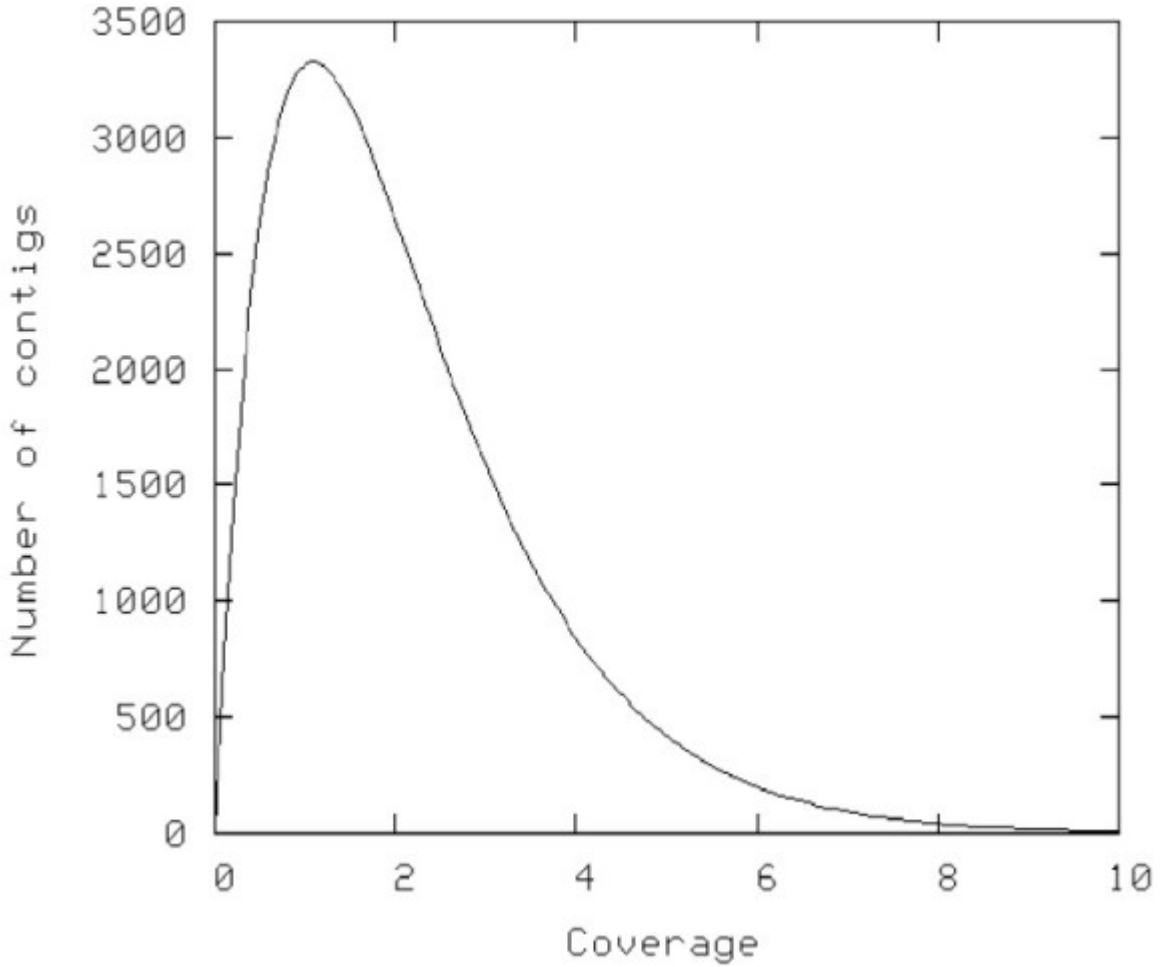

Guizotia abyssinica
chloroplast genome
151,762 bp
LSC
IRA
IRB
SSC
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)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

DNA extraction, sequencing, assembly

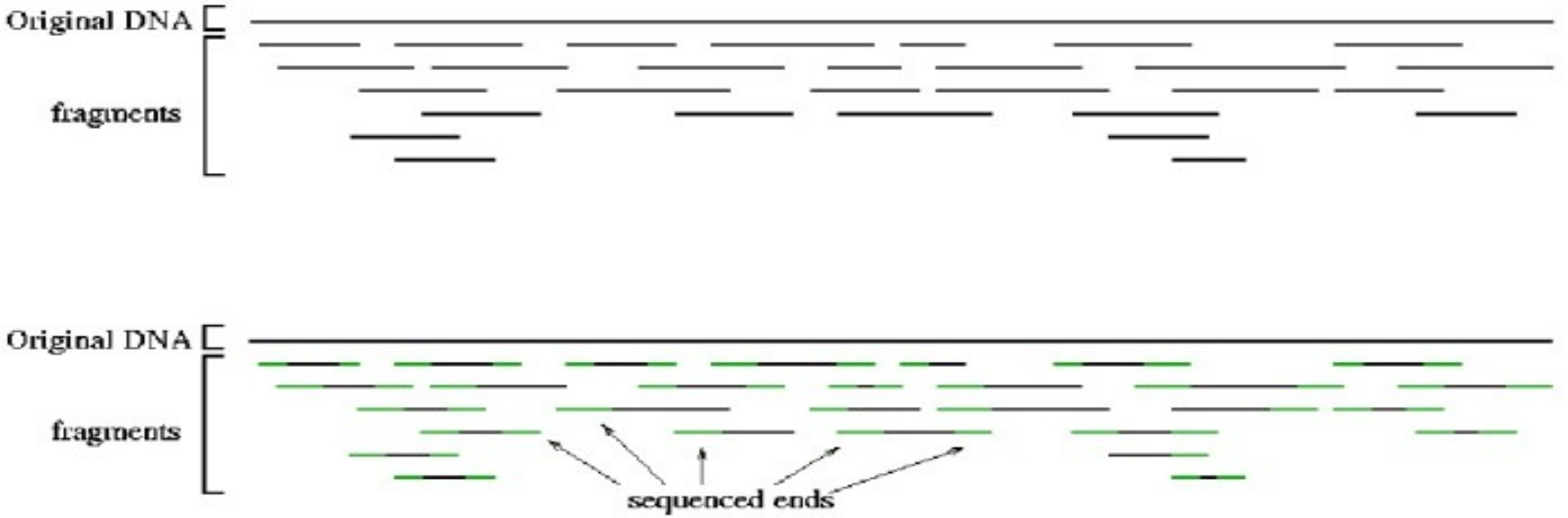


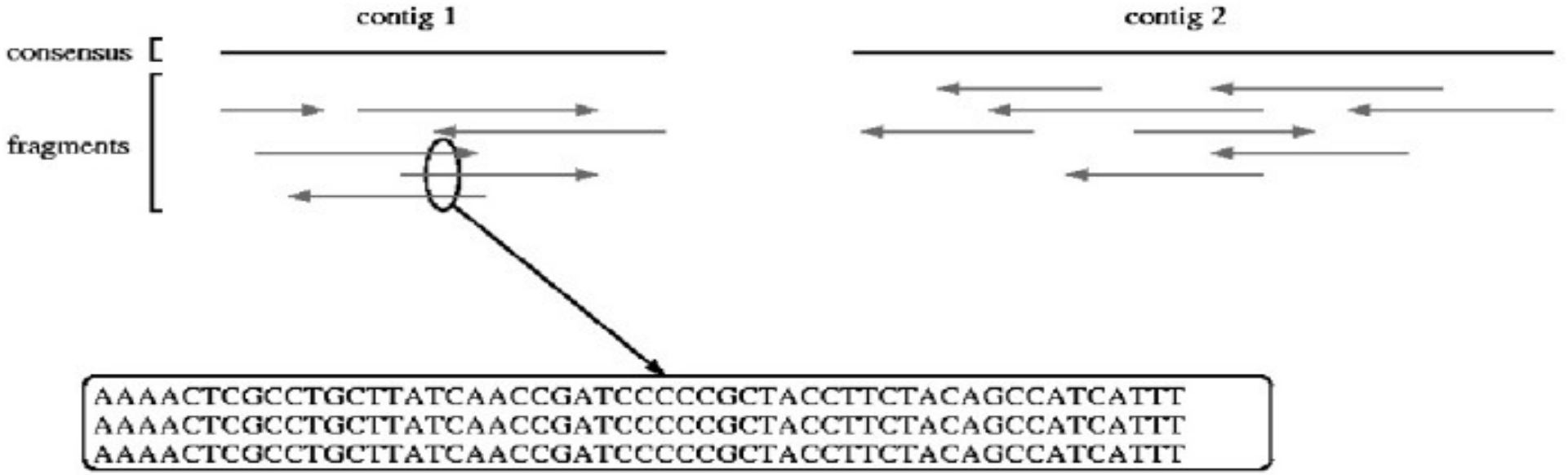


Number of contigs vs. genome coverage

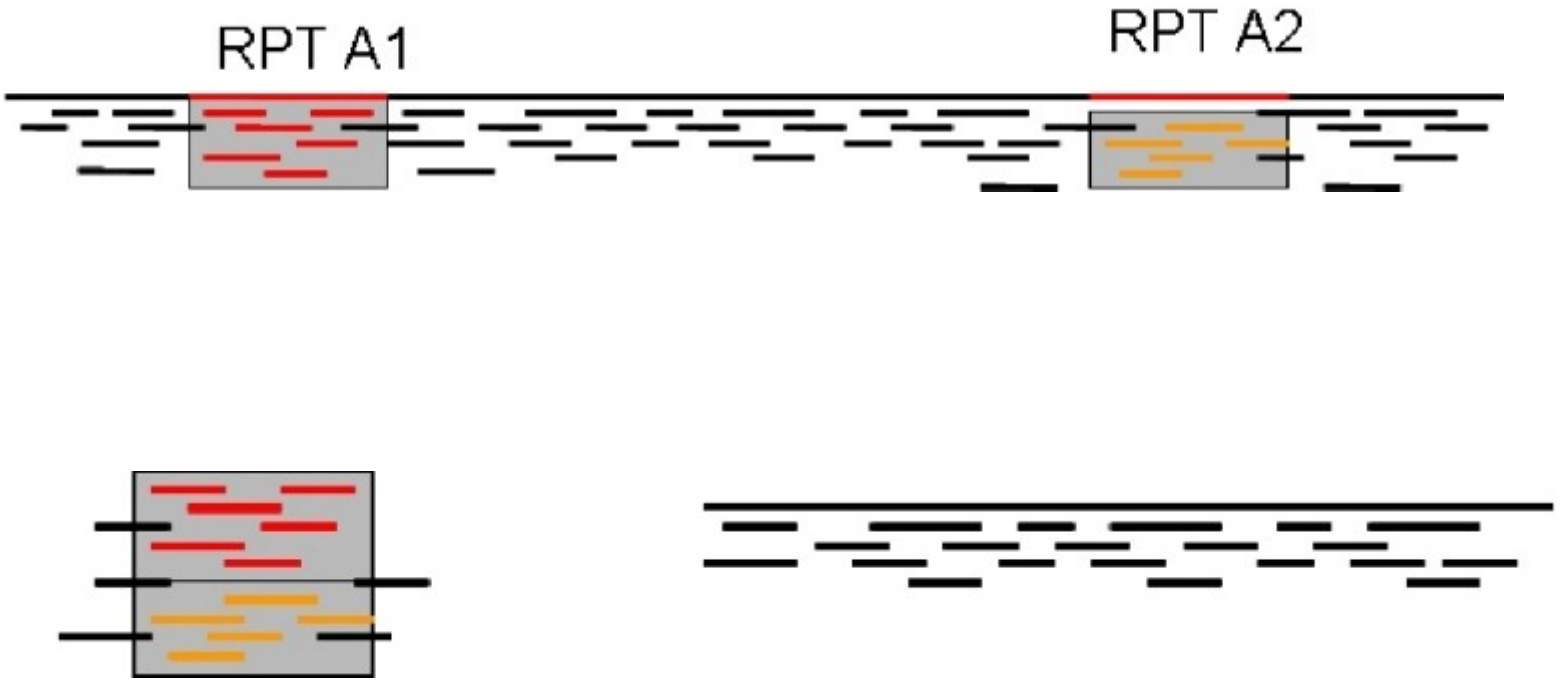


DNA extraction, sequencing, assembly





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

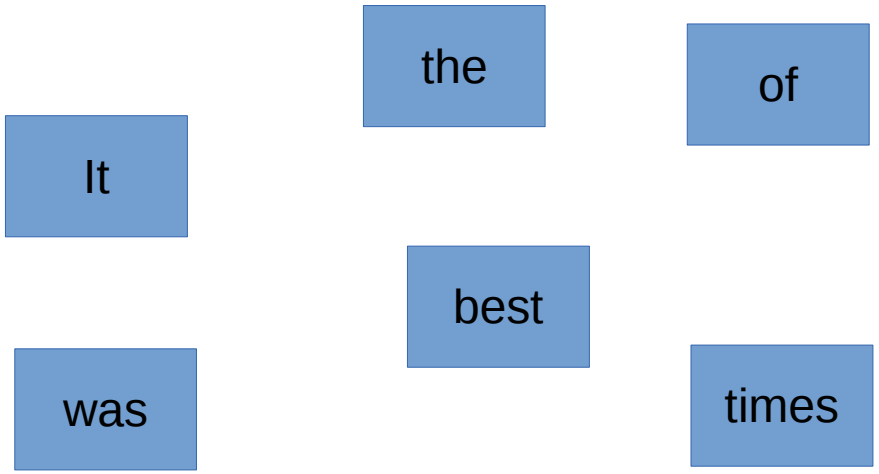
Assembly algorithms

- 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

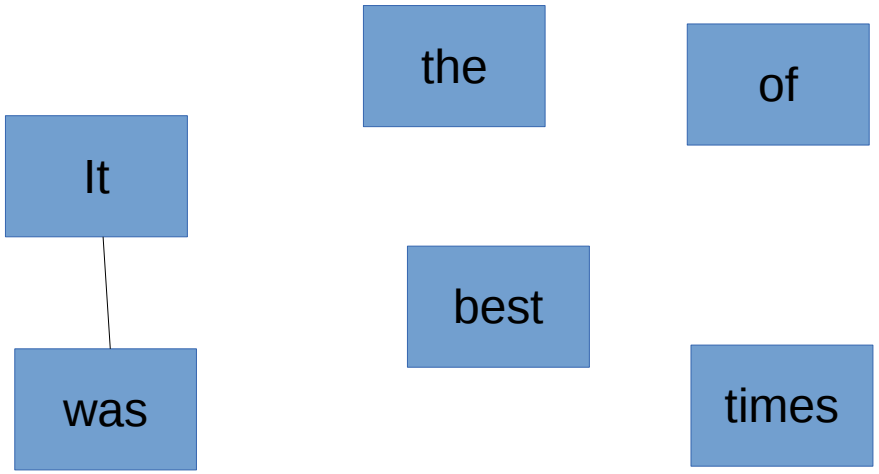
Joining reads with k-mers

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



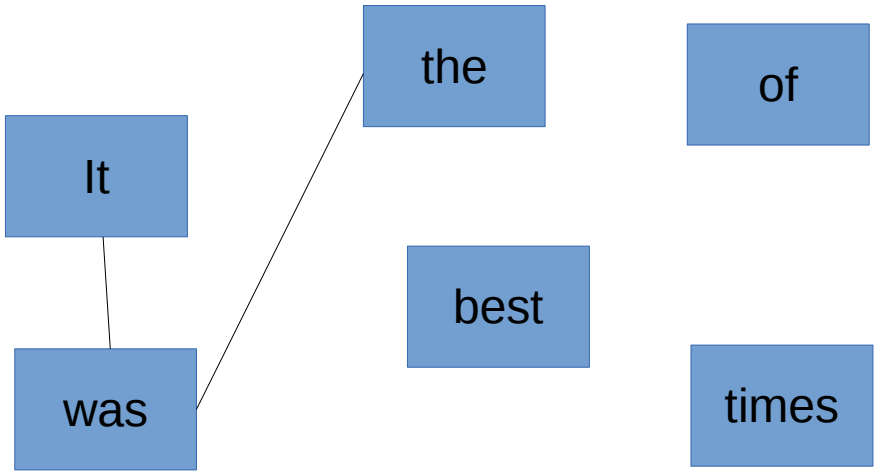
Joining reads with k-mers

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



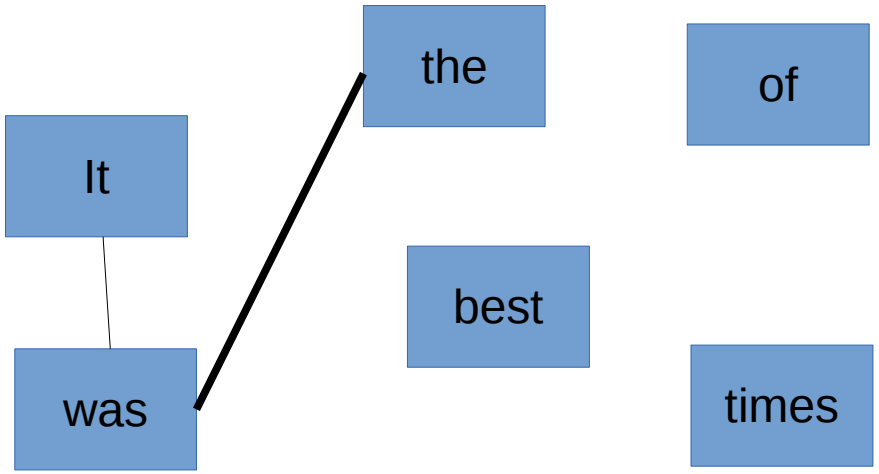
Joining reads with k-mers

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



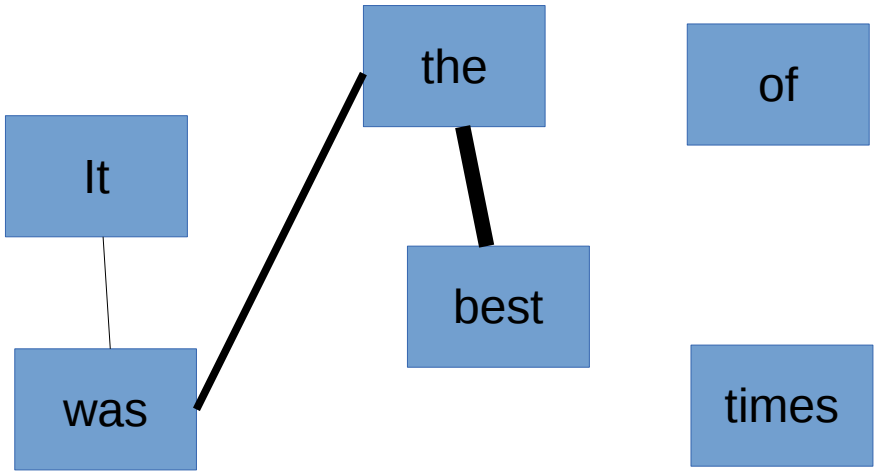
Joining reads with k-mers

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



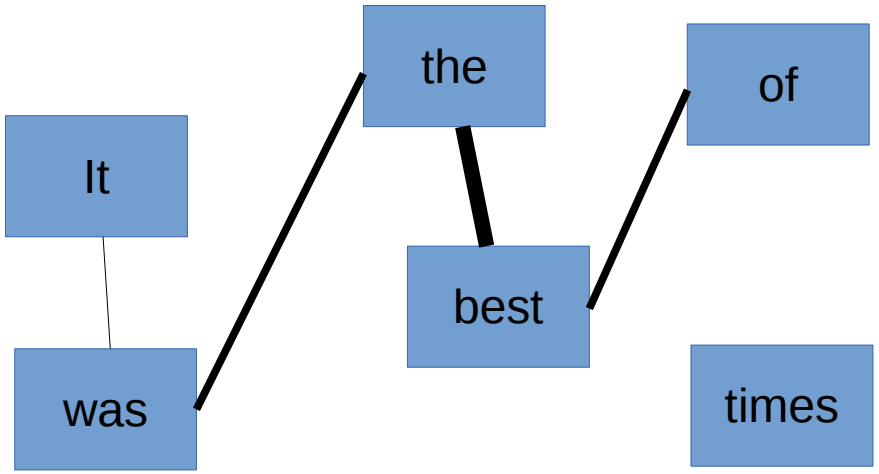
Joining reads with k-mers

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



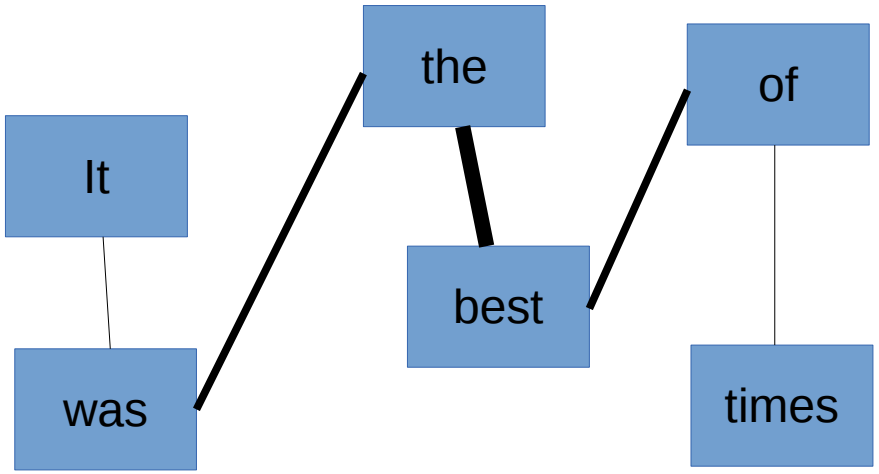
Joining reads with k-mers

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



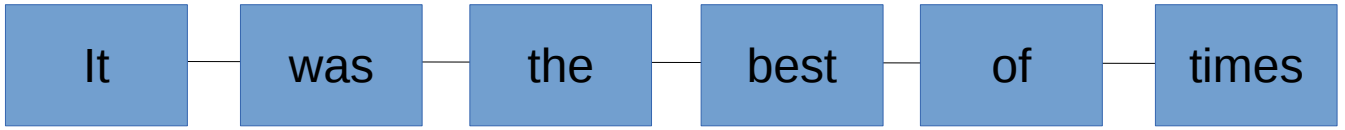
Joining reads with k-mers

- 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

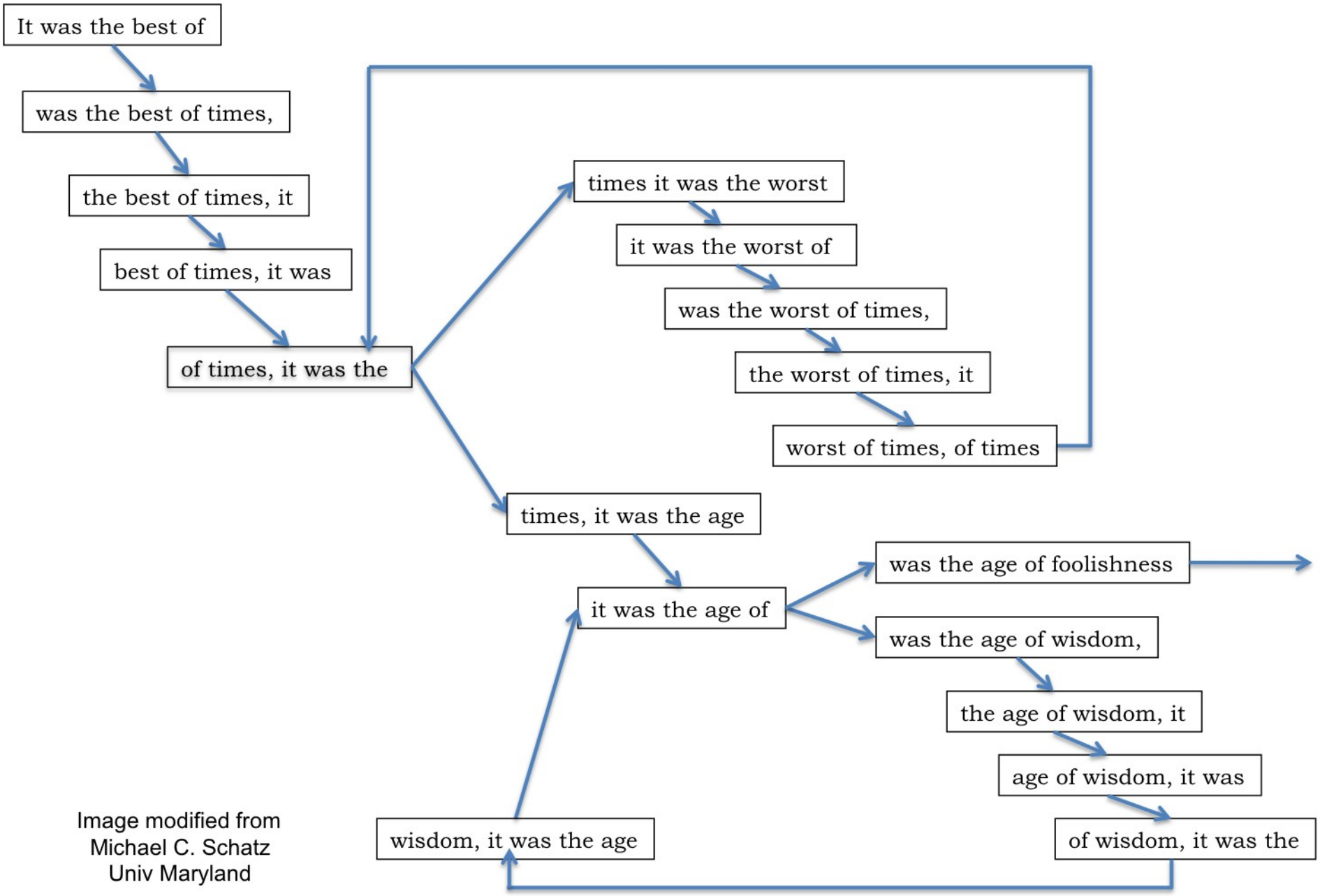
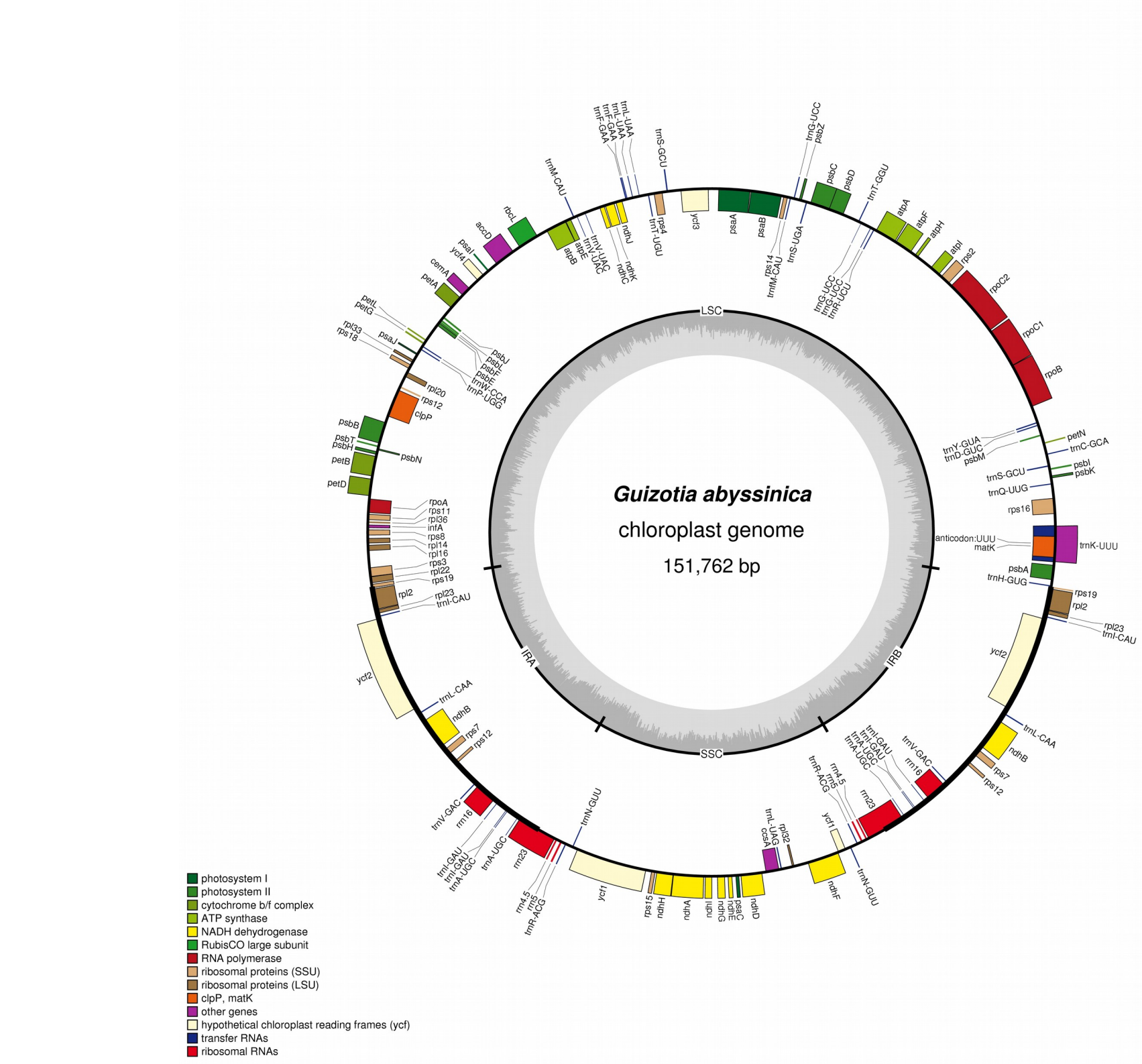
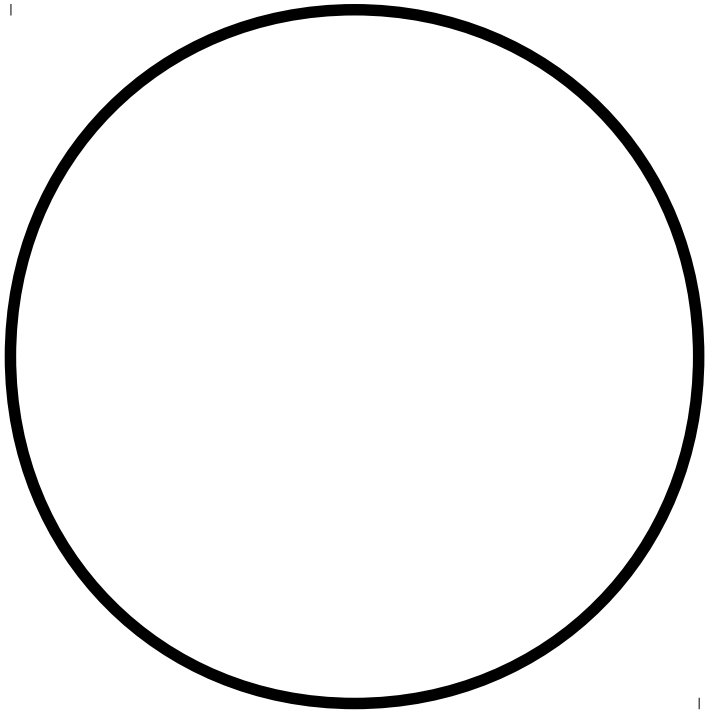


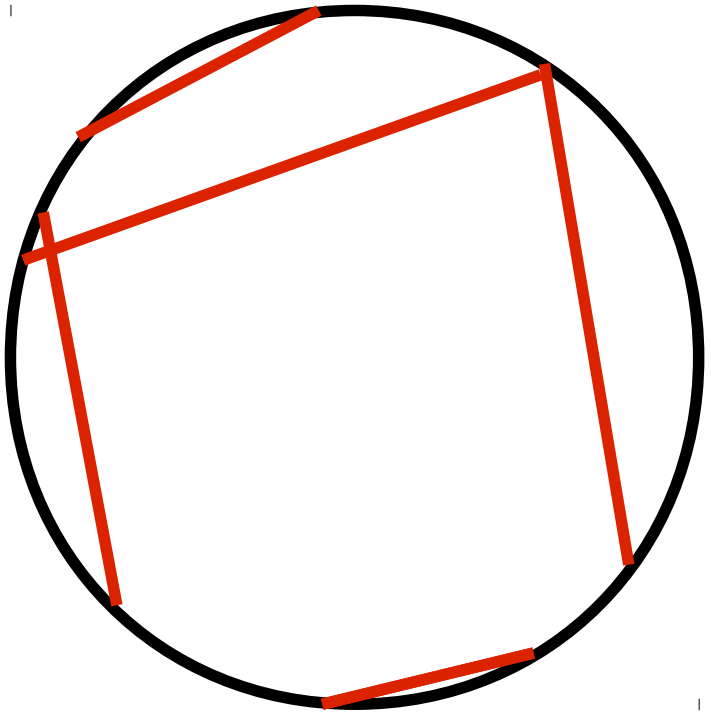
Image modified from
Michael C. Schatz
Univ Maryland

Assembly algorithms

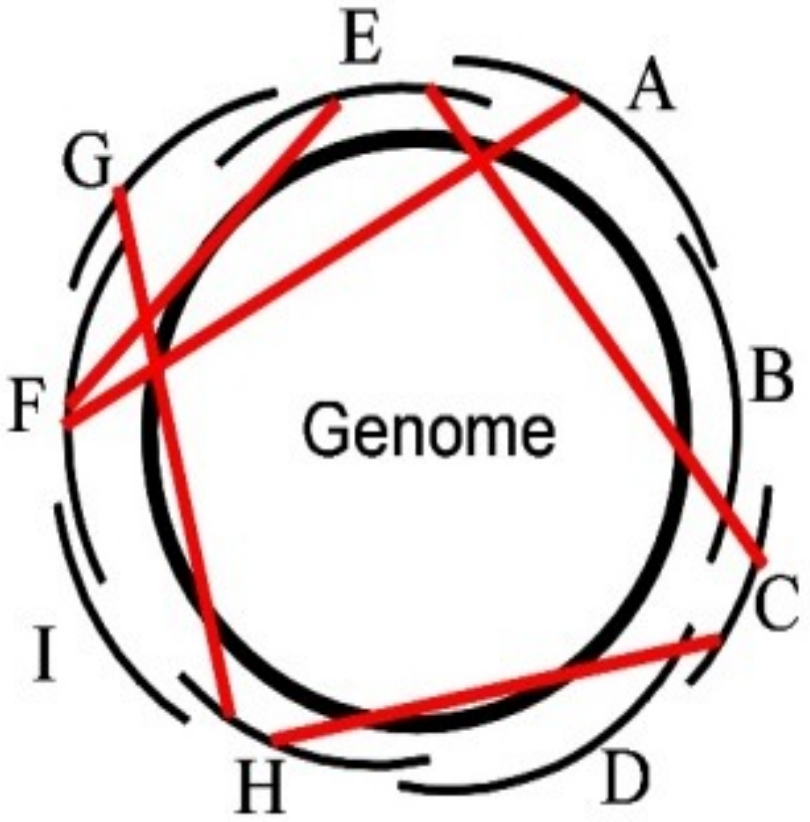
- It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything 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

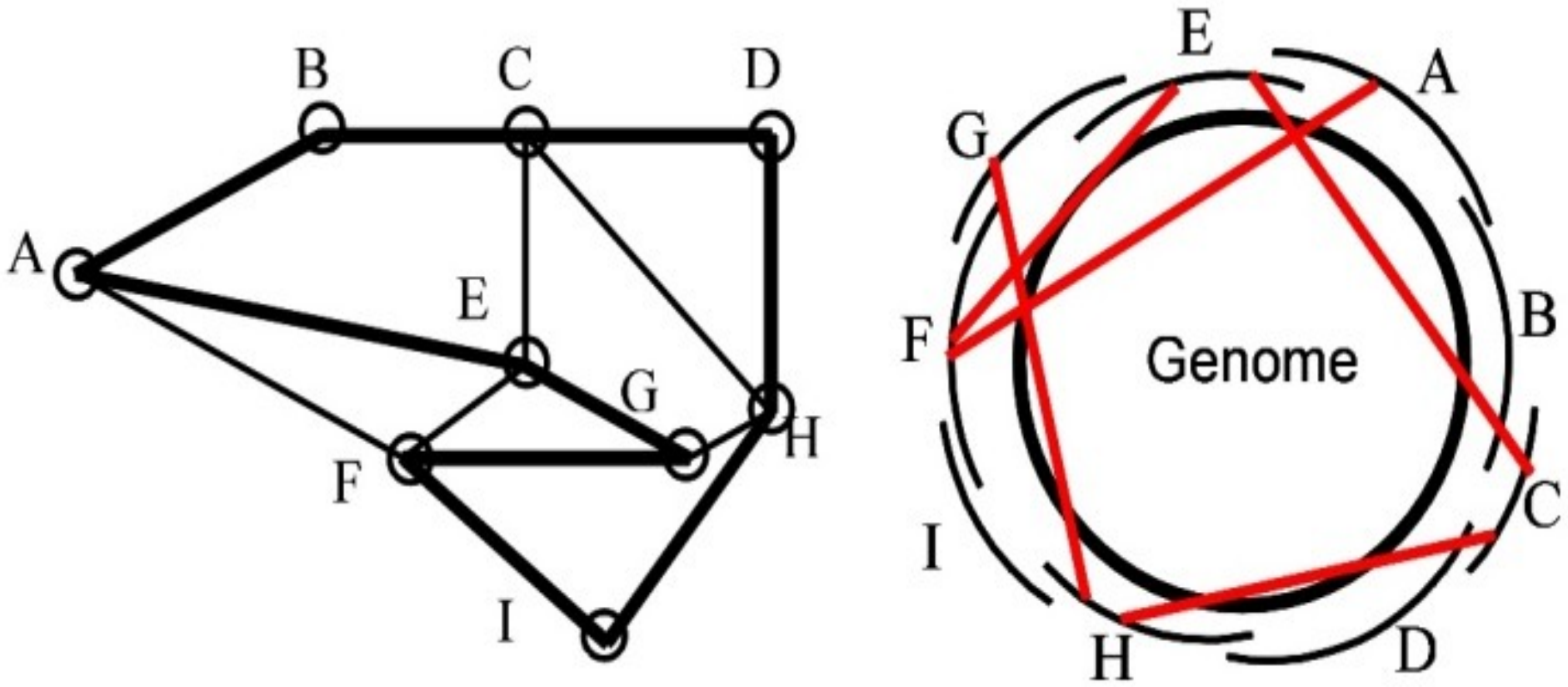




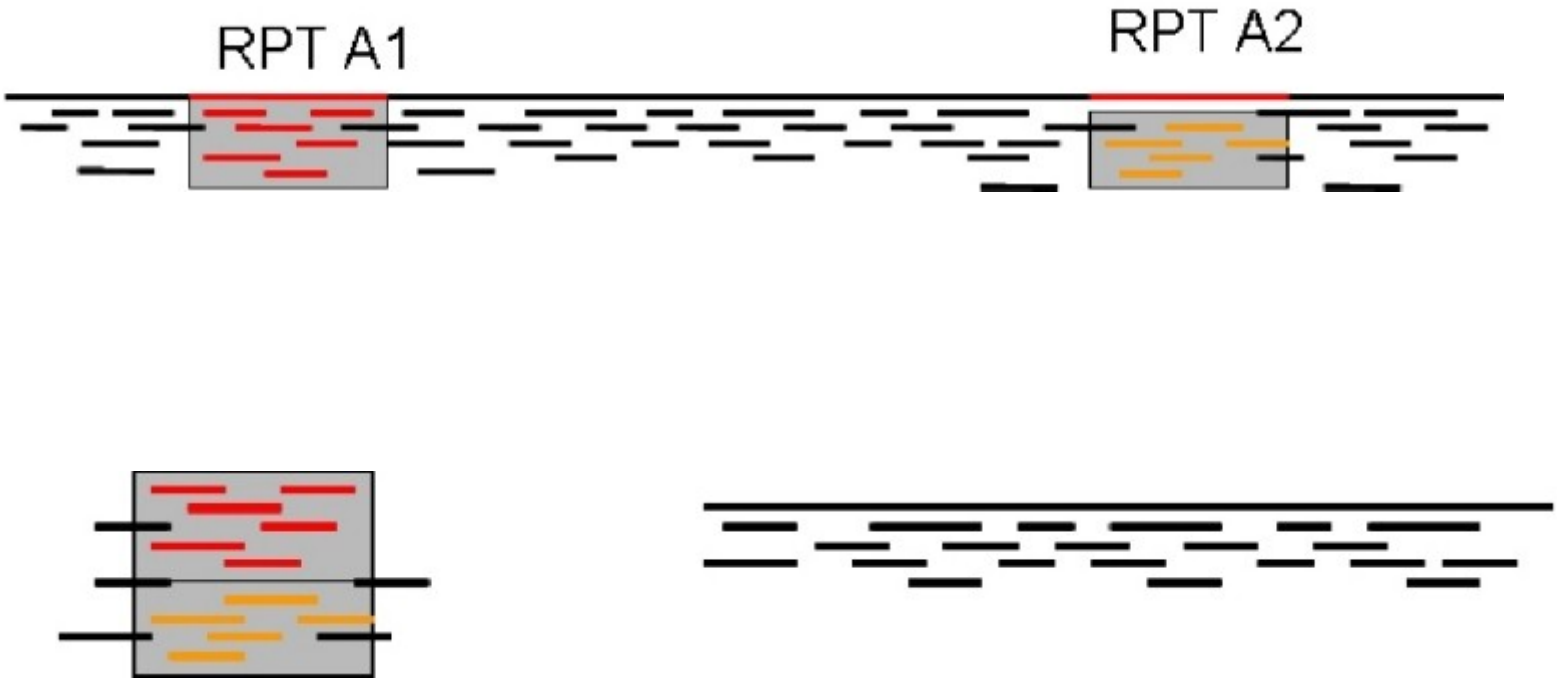
K-mers connect reads -> assembly



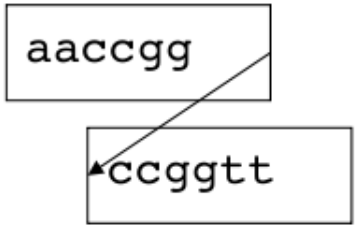
K-mers connect reads -> assembly



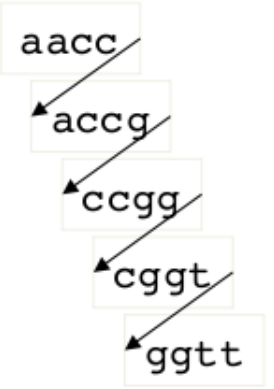
Repeats can cause errors

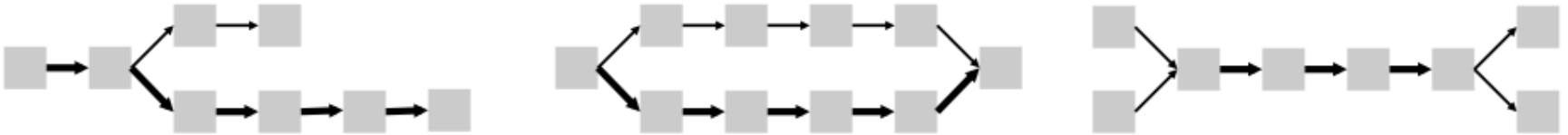


Graph Algorithms



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.

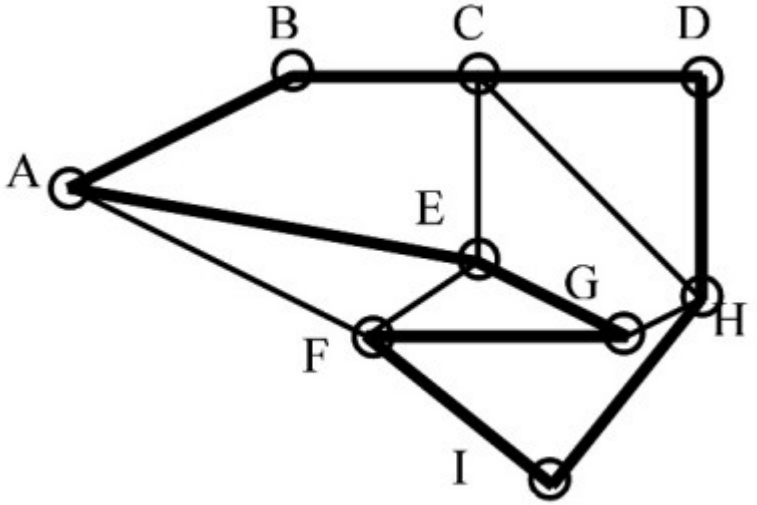


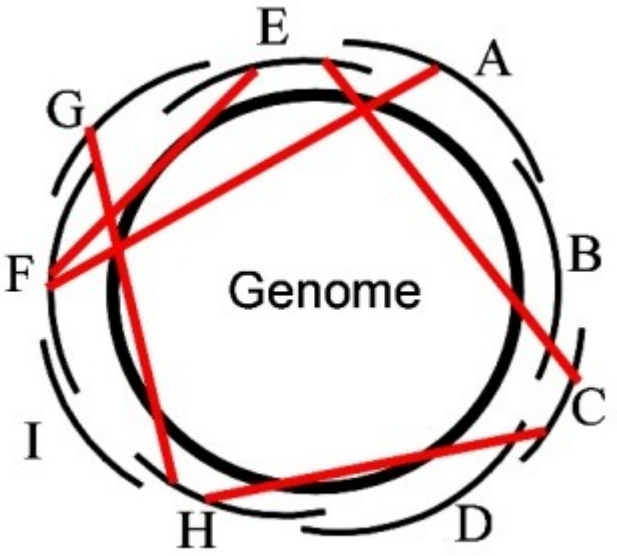


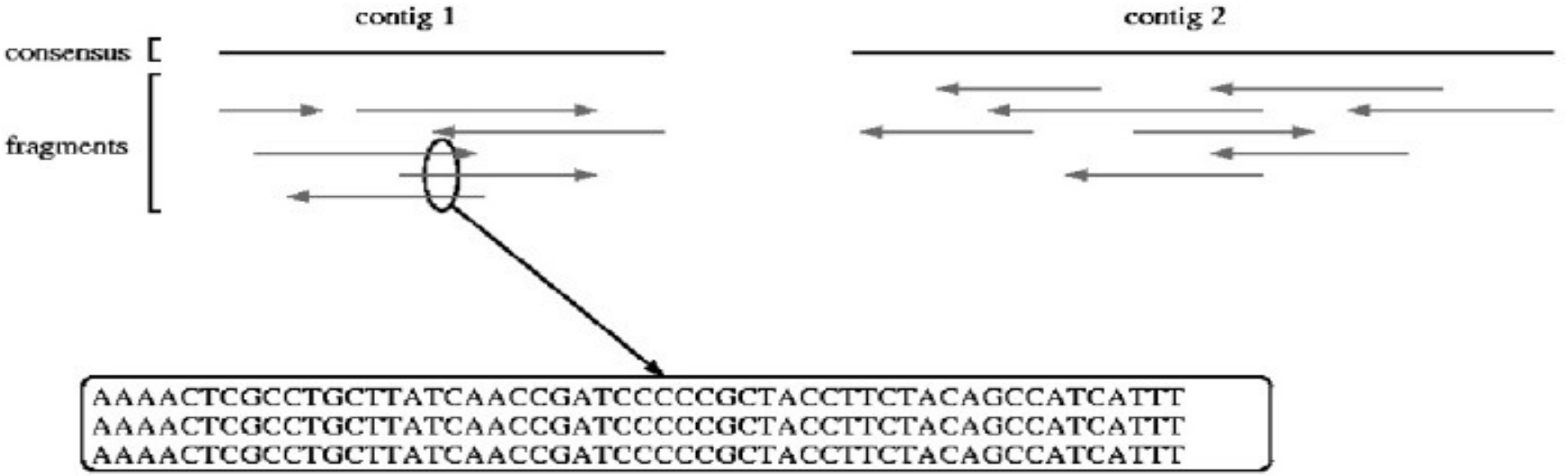
Assembly algorithms reduce graph complexity. Boxes are reads or K-mers. Edges are overlaps or matches. Edge thickness can indicate amount of support in reads.

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