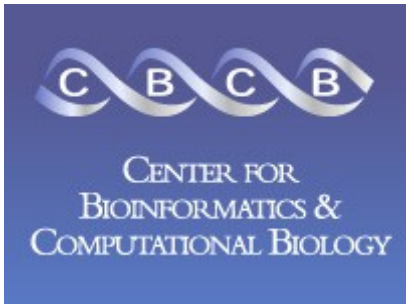
First steps towards automated metagenomic assembly

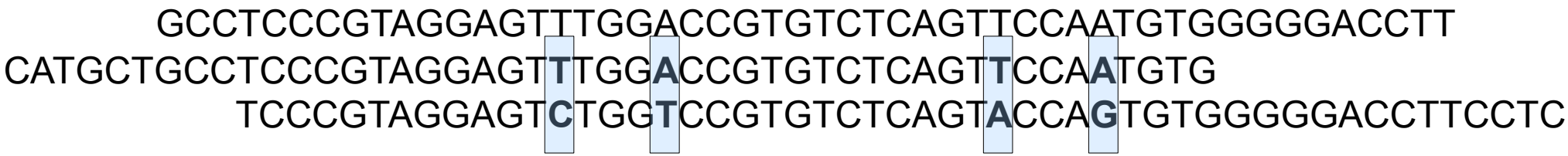
Mihai Pop
Sergey Koren
Dan Sommer

University of Maryland, College Park

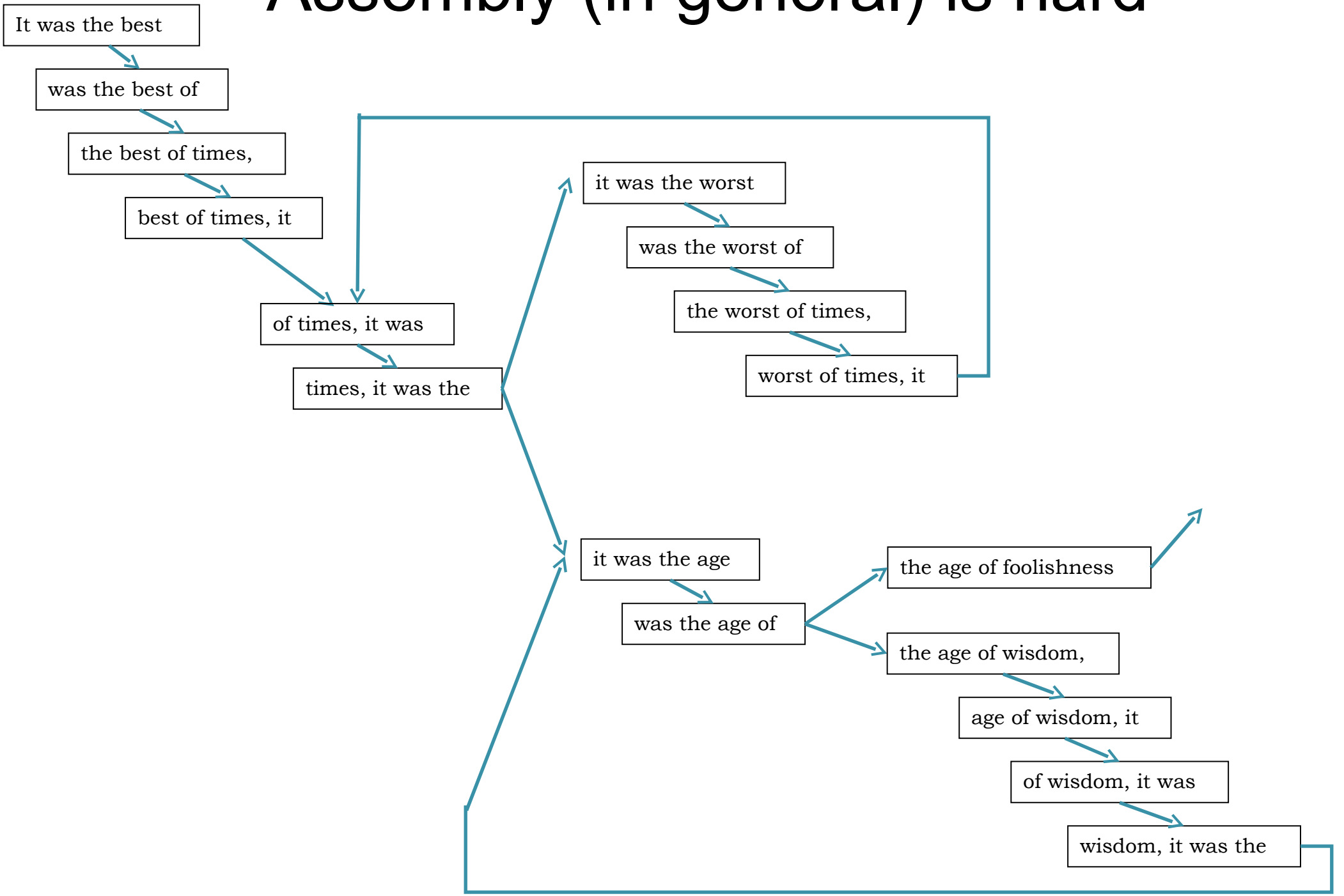


Metagenomic assembly is impossible

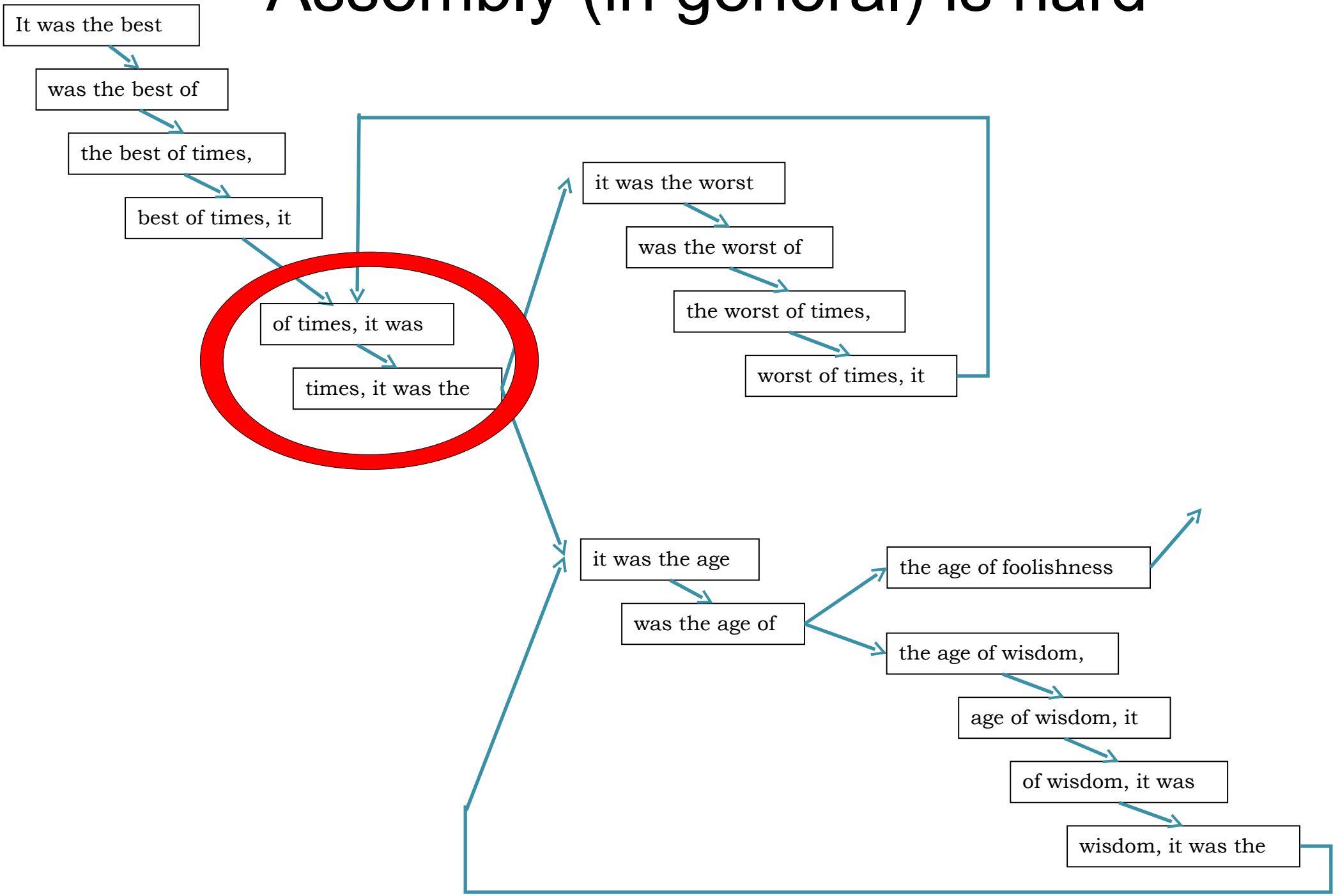
- Two competing goals:
  - assemble similar sequences from related genomes together
  - do not assemble similar sequences from unrelated genomes

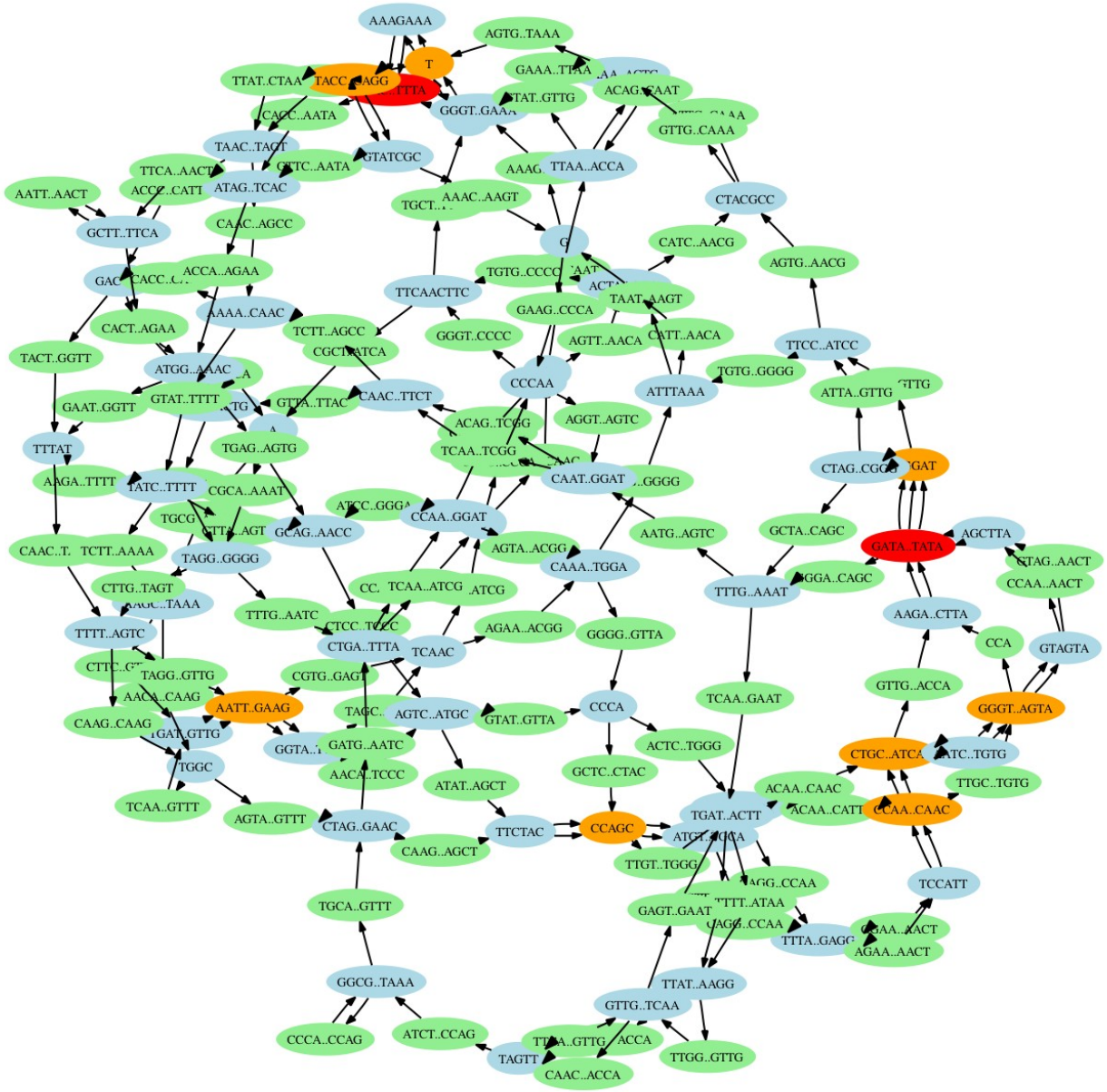


Assembly (in general) is hard



Assembly (in general) is hard

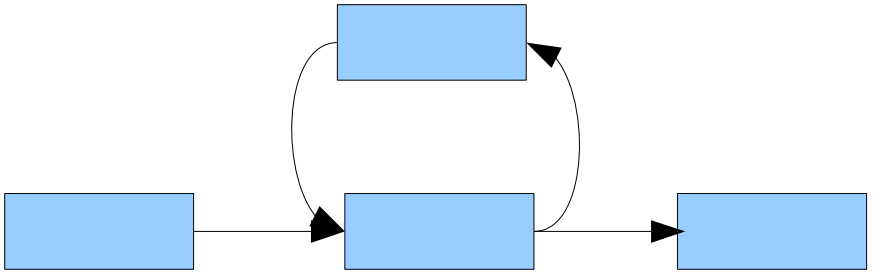



AAAGAAA
AGTG..TAAA
T
TTAT..CTAA
TACC..AAGG
GAAA..TTAA
ACAG..GAAT
GGGT..GAAA
GTAT..GTTG
GTTG..CAAA
TAAC..TAGT
GTATCGC
TTCA..AACT
GTTC..AATA
AAAG..TTAA..ACCA
AATT..AACT
ATAG..TCAC
AAAC..AAGT
CTACGCC
GCTT..TTCA
CAAC..AGCC
CATC..AACG
AGTG..AACG
ACCA..AGAA
TTCAACTTC
TGTG..CCCC
TAAT..AAGT
GAAG..CCCA
CACT..AGAA
AAAA..CAAC
TCTT..AGCC
GGGT..CCCC
AGTT..AACA
CATT..AACA
TACT..GGTT
CGCT..ATCA
TTCC..ATCC
ATGG..AAAC
CCCAA
TGTG..GGGG
ATTTAAA
ATTA..GTTG
GAAT..GGTT
GTAT..TTTT
GTTA..TTAC
CAAC..TTCT
AGGT..AGTC
ACAG..TCGG
TTTAT
TGAG..AGTG
TCAA..TCGG
AAGA..TTTT
TATC..TTTT
CGCA..AAAT
CAAT..GGAT
CTAG..CGGA
GAT
ATCC..GGGA
CCAA..GGAT
GCTA..CAGC
AATG..AGTC
GATA..TATA
AGCTTA
TGCG..
GTTA..AGT
GCAG..AACC
CAAC..T..
TCTT..AAAA
AGTA..ACGG
CAAA..TGGA
GTAG..AACT
TAGG..GGGG
GGGA..CAGC
CCAA..AACT
CTTG..TAGT
CC..TCAA..ATCG
TTTG..AAAT
AAGA..CTTA
AAGC..TAAA
TTTG..AATC
CTCC..TCCC
CCA
TTTT..AGTC
AGAA..ACGG
GGGG..GTTA
GTAGTA
CTGA..TTTA
TCAAC
CTTC..GT..
TAGG..GTTG
CGTG..GAGT
GTTG..ACCA
AACA..CAAG
AATT..GAAG
TAGC..
AGTC..ATGC
GTAT..GTTA
CCCA
TCAA..GAAT
GGGT..AGTA
CAAG..CAAG
GAT..GTTG
GGTA..
GATG..AATC
ACTC..TGGG
CTGC..ATCA
ATC..TGTG
TGGC
AACA..TCCC
GCTC..CTAC
TTGC..TGTG
TCAA..GTTT
ATAT..AGCT
ACAA..CAAC
AGTA..GTTT
CTAG..GAAC
TTCTAC
CCAGC
TGAT..ACTT
ACAA..CATT
CCAA..CAAC
CAAG..AGCT
ATCT..ACCA
TTGT..TGGG
TCCATT
TGCA..GTTT
GAGT..GAAT
TTTT..ATAA
AAGG..CCAA
TTTA..GAGG
AGAA..AACT
GGCG..TAAA
TTAT..AAGG
GTTG..TCAA
CCCA..CCAG
ATCT..CCAG
TTAA..GTTG
ACCA
TAGTT
TTGG..GTTG
CAAC..ACCA

Assembly (in general) is hard

- Repeats lead to ambiguity in reconstruction of genome
  (complexity exponential in # of repeats)

- Insufficient coverage:
  - gaps
  - obscures "true" assembly

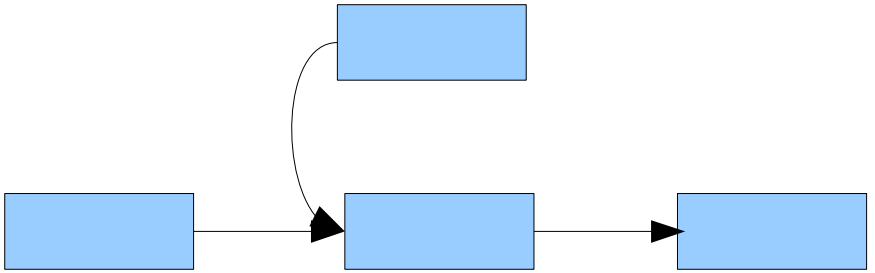


- Sequencing errors compound all other challenges

Assembly (in general) is hard

- Repeats lead to ambiguity in reconstruction of genome
  (complexity exponential in # of repeats)

- Insufficient coverage:
  - gaps
  - obscures "true" assembly

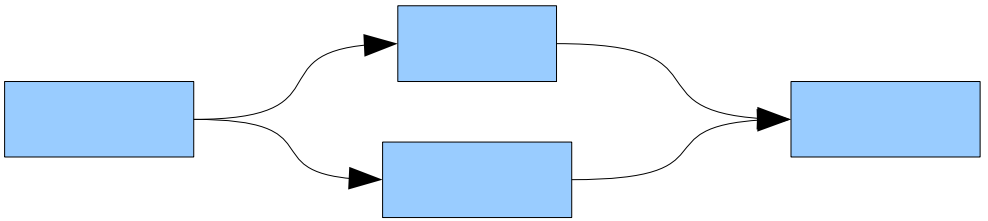


- Sequencing errors compound all other challenges

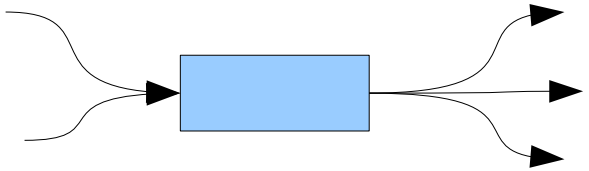
Goals for a metagenomic assembler

- Must work well for clonal data
  - handle repeats
  - handle errors
  - deal with low coverage regions

- Must deal with polymorphisms
  - distinguish between errors and polymorphisms



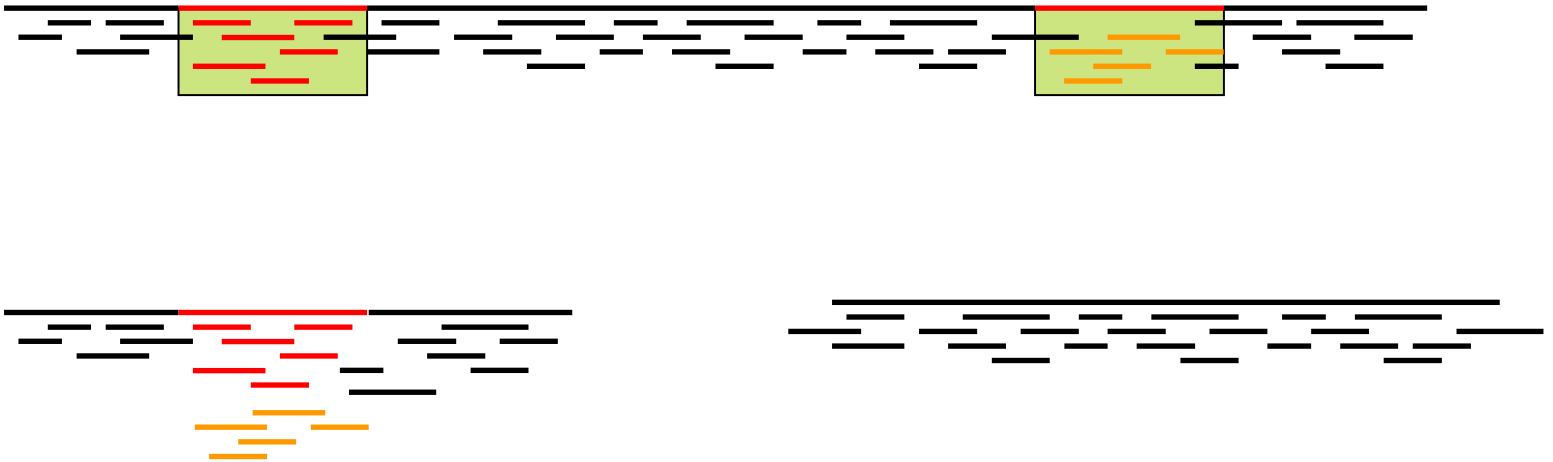
  - distinguish between repeats and polymorphisms



  - enable discovery of polymorphisms/variation

Repeat detection

- Basic idea: deeply covered unitigs are repeats

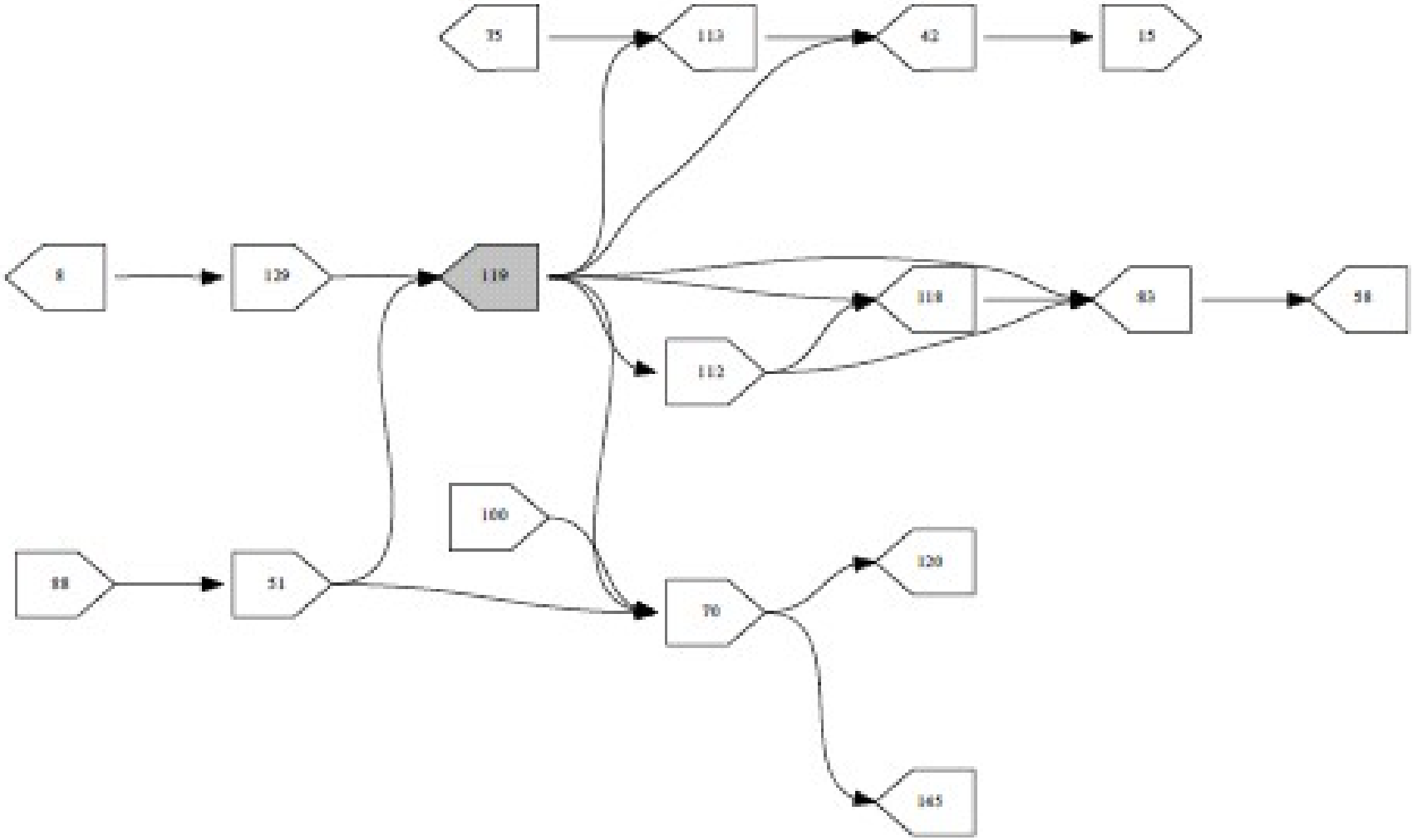


- But.... in metagenomics the deeply covered contigs are often abundant organisms
(which should be assembled)

Rusch, D. B., A. L. Halpern, et al. (2007). "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific." PLoS Biol **5**(3): e77.

Better repeat detection

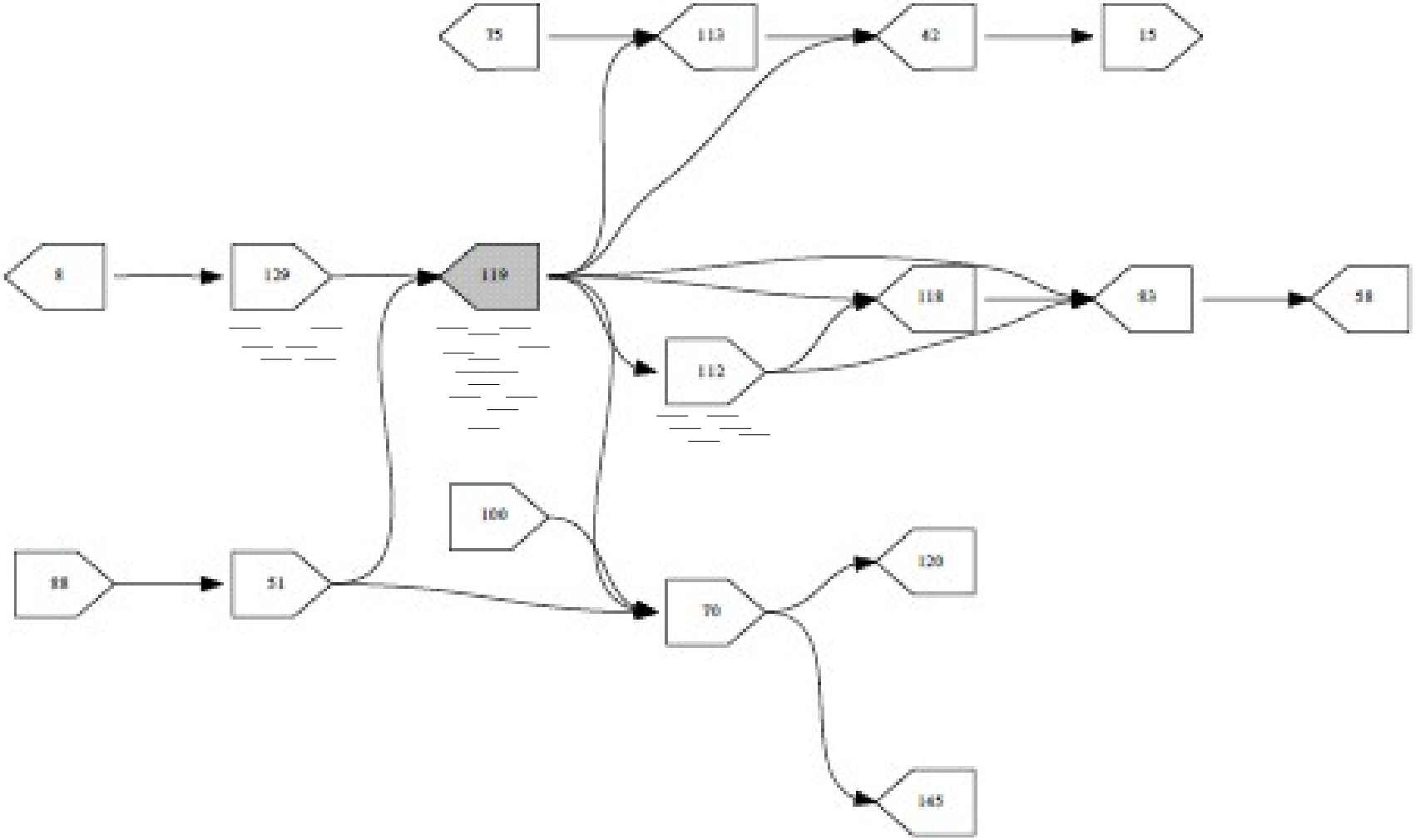
- Observation: repeats "tangle" the assembly graph



- Solution: find nodes "central" to the graph

Better repeat detection...2

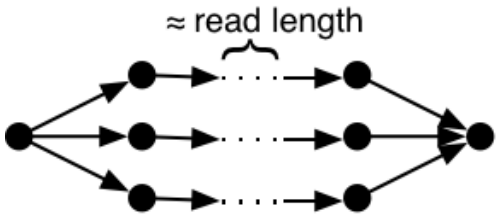
- Observation: depth of coverage is good <u>local</u> marker of repeats



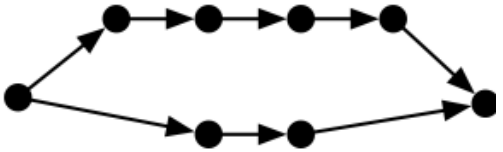
- Solution: run local depth-of-coverage statistics

Detection of variation

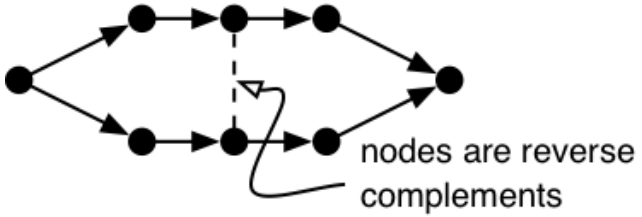
- Key idea: find assembly motifs that look like genome variation



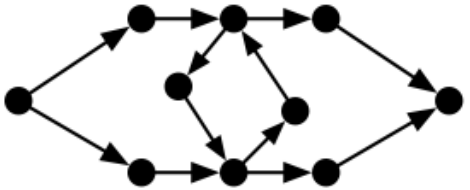
(a) SNPs



(b) Indels



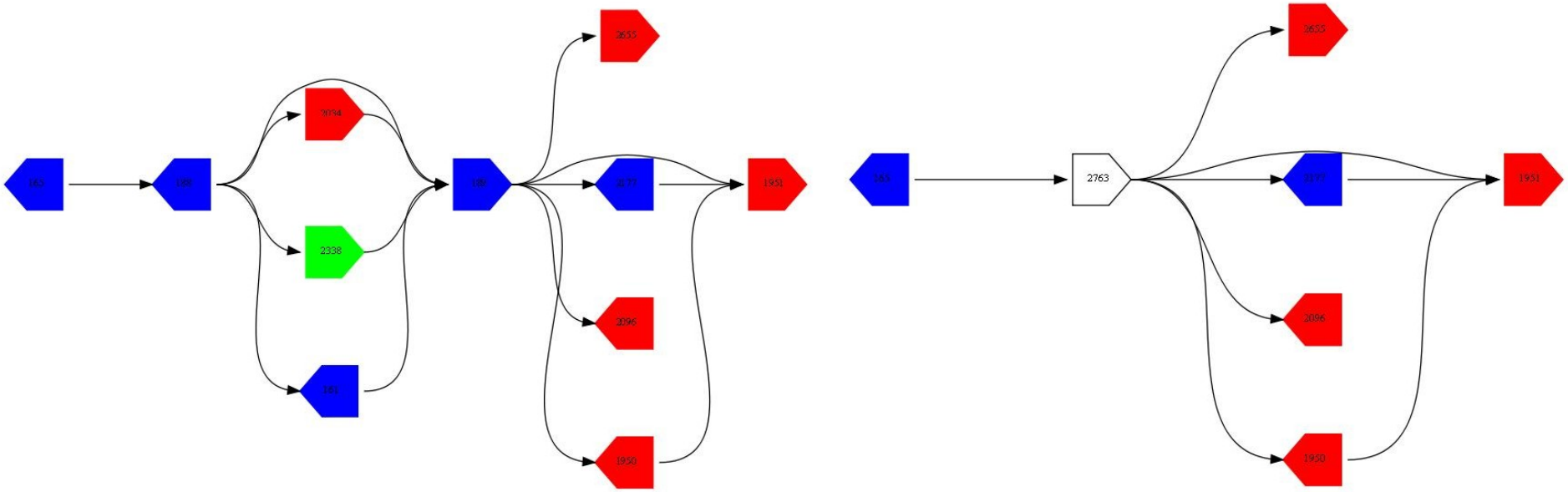
(c) Inversions



(d) Transpositions

Scaffolding through variation

- Identify motifs and collapse them

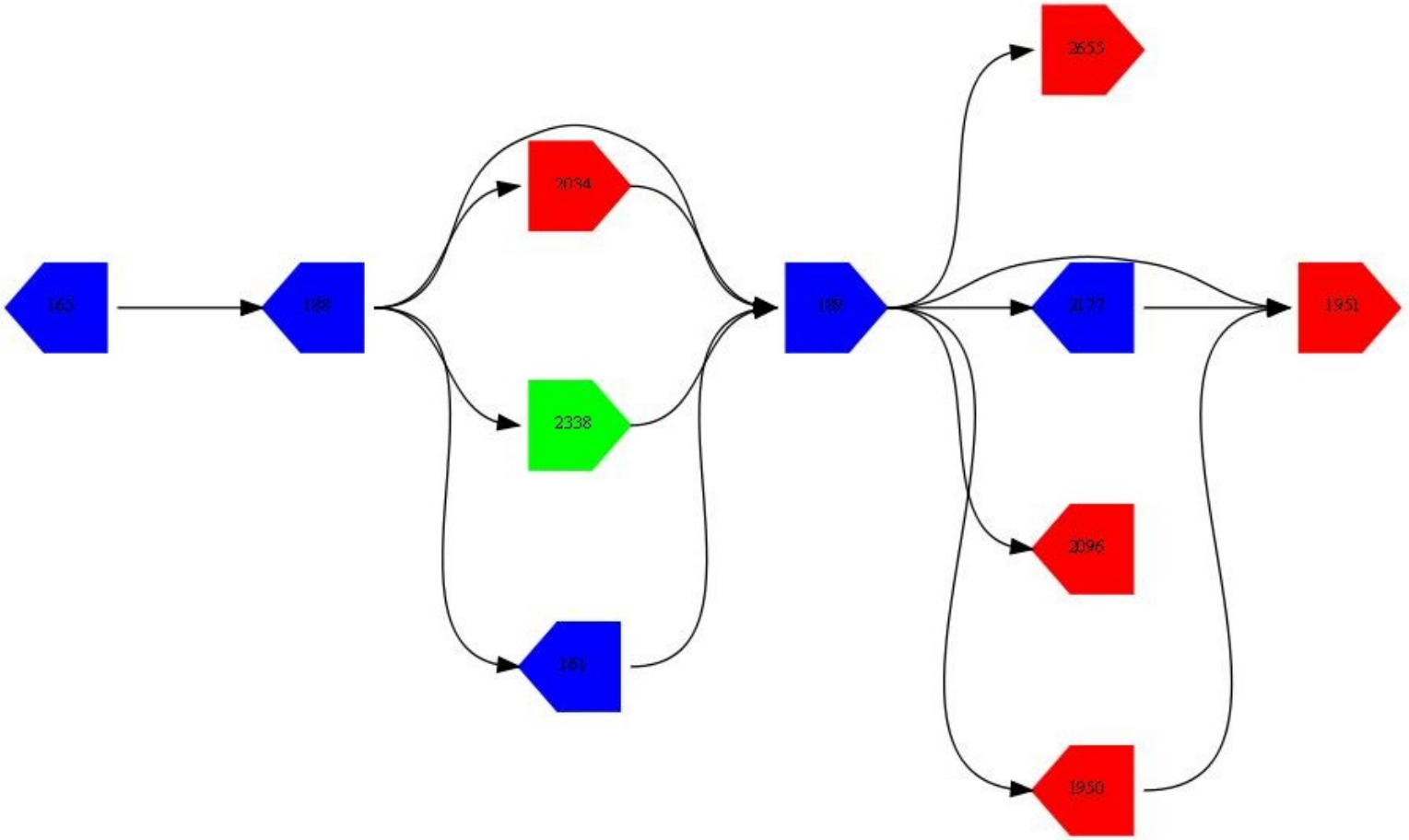


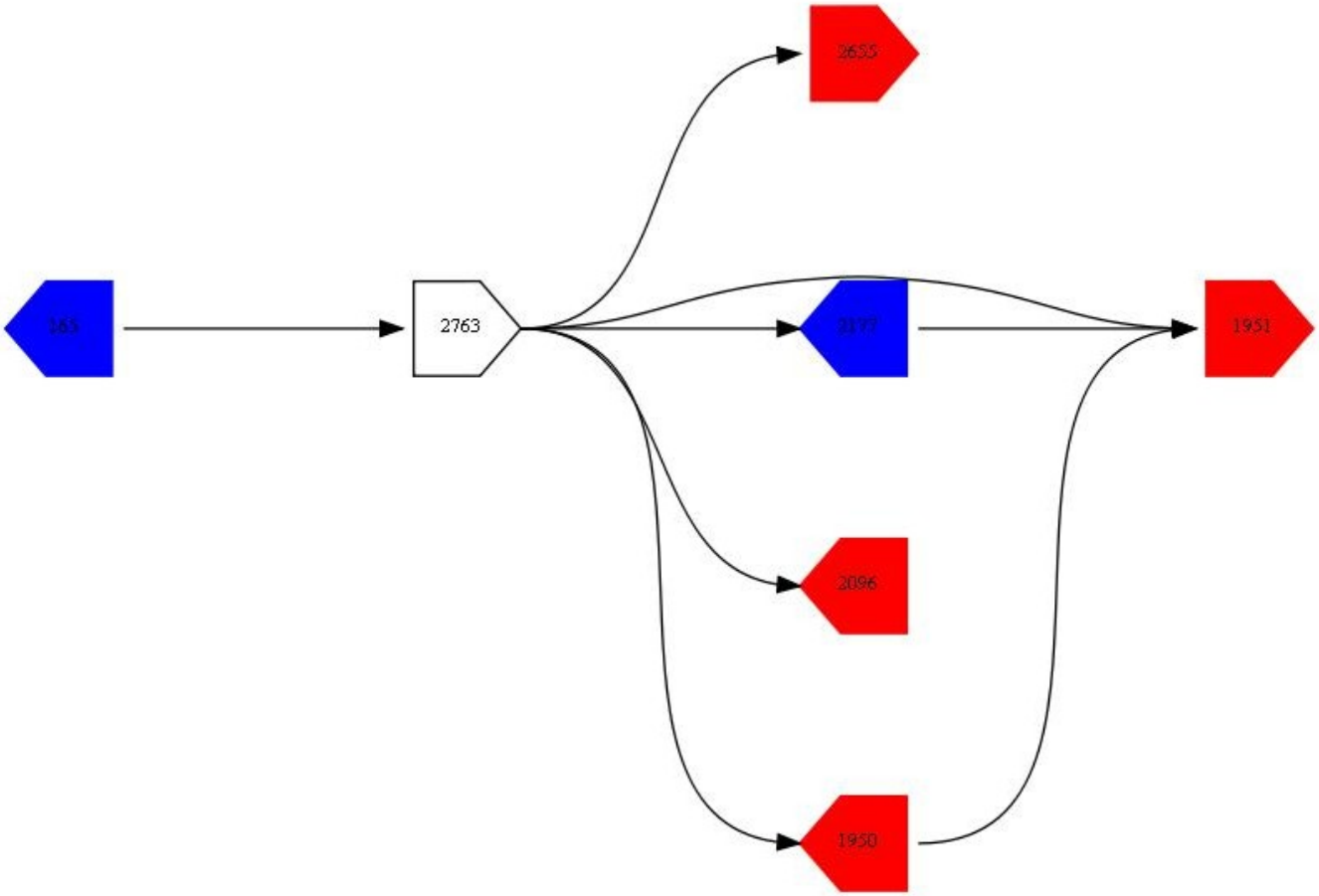
Detection of variation...cont

- Still work in progress
  - simple motifs can be found
  - looking for "enriched" motifs
  - still analyzing motifs found by our code

- Basic idea (find motif, collapse it) very computer graphic-y (that's how video games work...)

- Corrolary: This can lead to interactive visualizations of assembly graphs


165
189
2034
2338
161
189
2655
2177
1951
2096
1950


2655
365
2763
2177
1951
2096
1950

Bambus 2

- http://www.cbcb.umd.edu/software/bambus
- Can be used with output from most assemblers (tested with CA, Minimus, Newbler)

- Good repeat detection

Organism	ASM		# Repeats	# TP	# FP	# FN	SN	SP
Brucella suis 1330	Bambus 2	Component-Joining	6	6	0	5	54.54	100
		Local Coverage	14	5	9	6	45.45	94.61
		Total	20	11	9	0	100	94.61
	CA		30	8	22	0	100	79.43
Acid Mine	Bambus 2	Component-Joining	93	13	80	44	22.80	99.25
		Local Coverage	2,508	25	2483	32	43.85	76.77
		Total	2,601	38	2,563	19	66.66	76.03
	CA		4,749	49	4,700	8	85.96	51.88
	CA-met		1,126	24	1,102	28	46.15	82.47

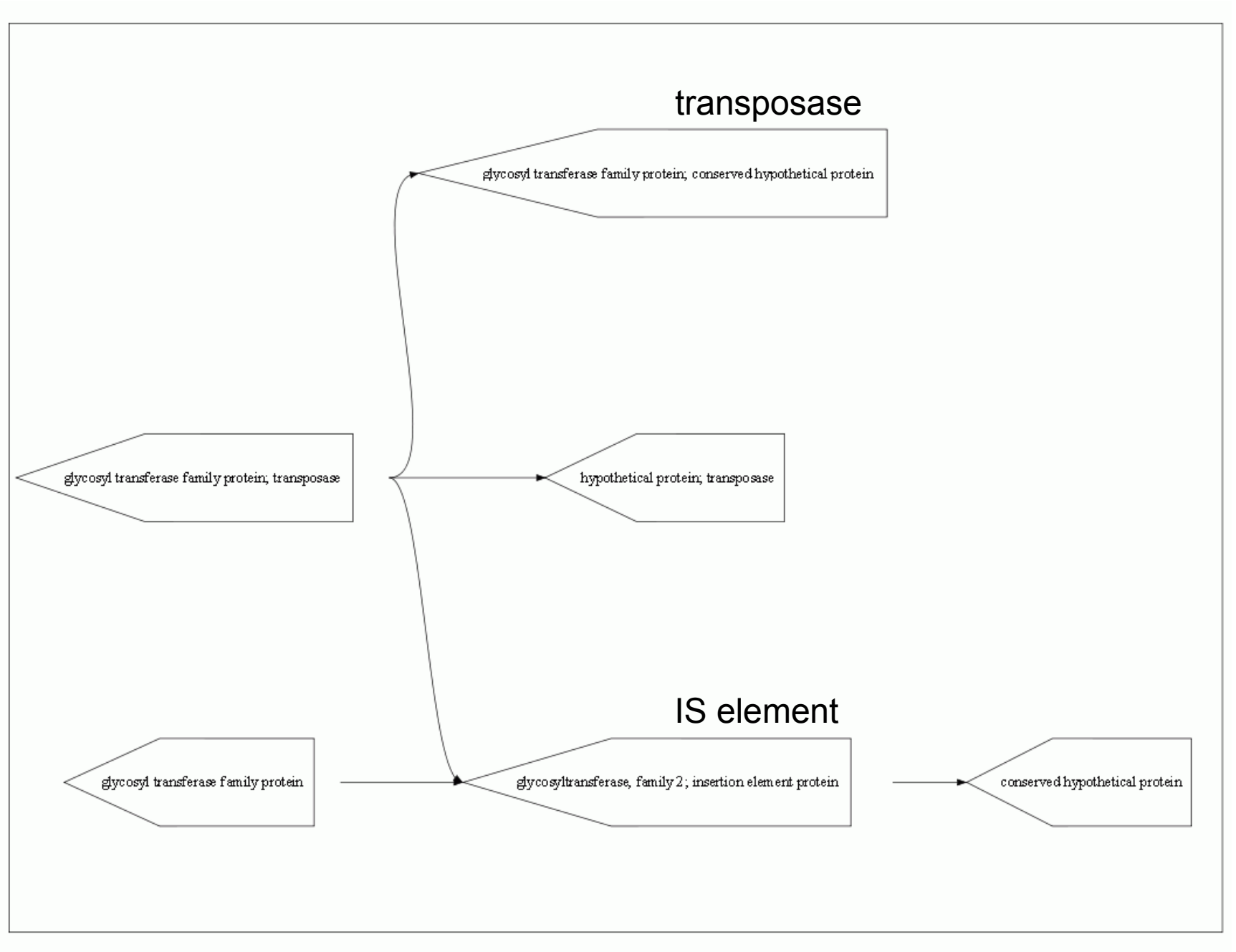
Bambus 2

- Good genome reconstruction

ASM	Organism	# Scfs	% Genome
Published	*Leptospirillum sp* Group II '5-way CG'	70	81.86 %
	Leptospirillum sp Group III	474	120.49 %
	Ferroplasma acidarmanus Type I	170	112.66 %
	Ferroplasma sp Type II	59	138.54 %
	Thermoplasmatales archaeon Gpl	410	121.44 %
CA	*Leptospirillum sp* Group II '5-way CG'	198	102.42 %
	Leptospirillum sp Group III	277	82.04 %
	Ferroplasma acidarmanus Type I	151	91.21 %
	Ferroplasma sp Type II	342	112.48 %
	Thermoplasmatales archaeon Gpl	405	87.93 %
CA-met	*Leptospirillum sp* Group II '5-way CG'	101	97.14 %
	Leptospirillum sp Group III	234	81.69 %
	Ferroplasma acidarmanus Type I	62	90.15 %
	Ferroplasma sp Type II	90	99.46 %
	Thermoplasmatales archaeon Gpl	179	83.60 %
Bambus 2	*Leptospirillum sp* Group II '5-way CG'	109	102.40 %
	Leptospirillum sp Group III	103	84.78 %
	Ferroplasma acidarmanus Type I	26	102.13 %
	Ferroplasma sp Type II	237	112.04 %
	Thermoplasmatales archaeon Gpl	167	94.90 %

Variation motif

Glycosyl transferase hypervariable locus in *Leptospirillum*



Future work

- Better documentation
- Better integration with other assemblers
- Tool for inspection of scaffolding data
- New types of variation
- Variation analysis toolkit

Acknowledgments

- Steven Salzberg
- Carl Kingsford
- Sergey Koren
- Dan Sommer
- Bo Liu
- Mohammad Ghodsi
- Ted Gibbons
- Todd Treangen

NIH R01-HG-004885
NSF IIS-0812111