# Genome Assembly

4/19/17

# Let’s try assembling a set of words

- Get into groups of 2-3
- Each group will get an envelope
- Those envelopes contain short sequences of words
  - Analog to Illumina Read
- Using NO outside knowledge – assemble these fragments into a well-known passage

# Want help?

- Your group can now have
  - A mate-pair set
  - A low coverage, long read set

# What is the phrase?

In each group, take 2 minutes to answer these questions

- What challenges did you face?
- How did you solve these challenges?

# Ground Truth

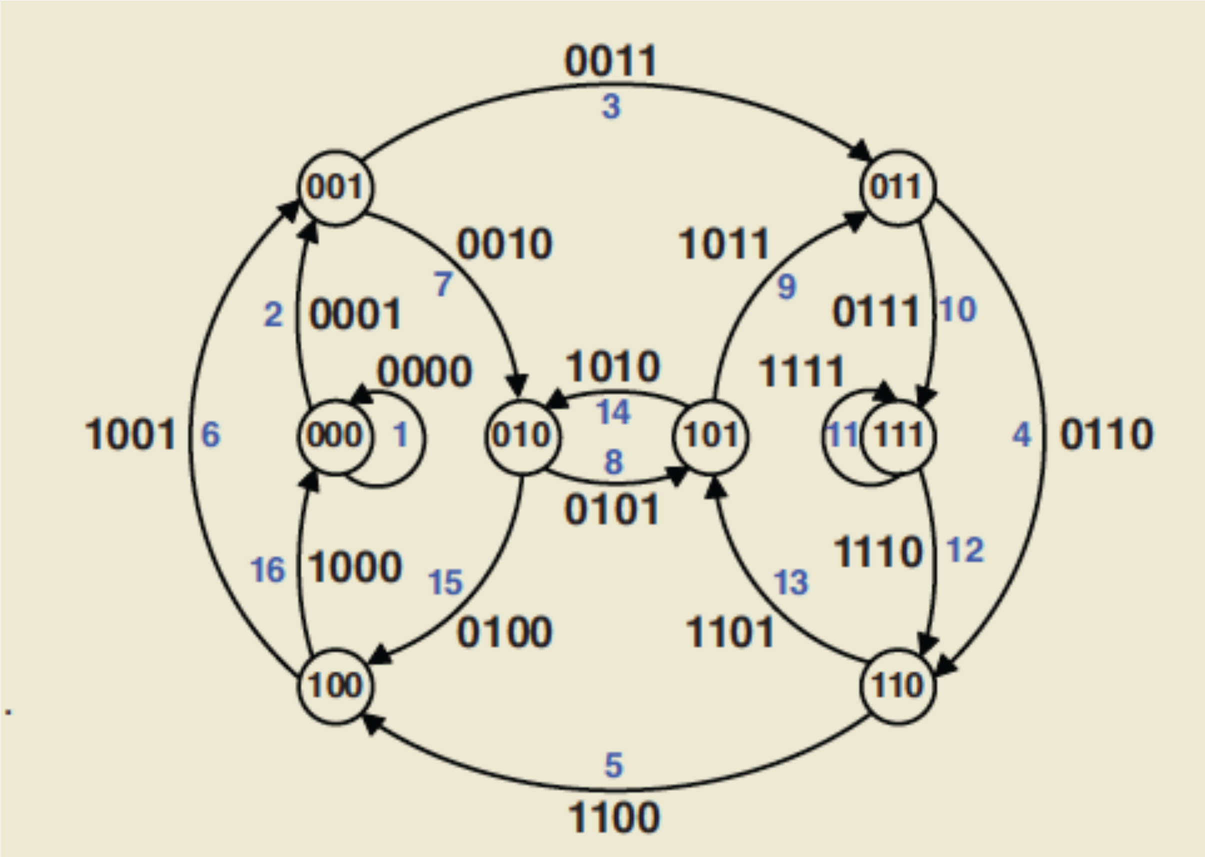
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair,…

- Charles Dickens, *A Tale of Two Cities*

# How are these challenges similar to genome assembly?

# How do we assemble genomes?

- de Bruijn graphs



# Build a graph of the sequences

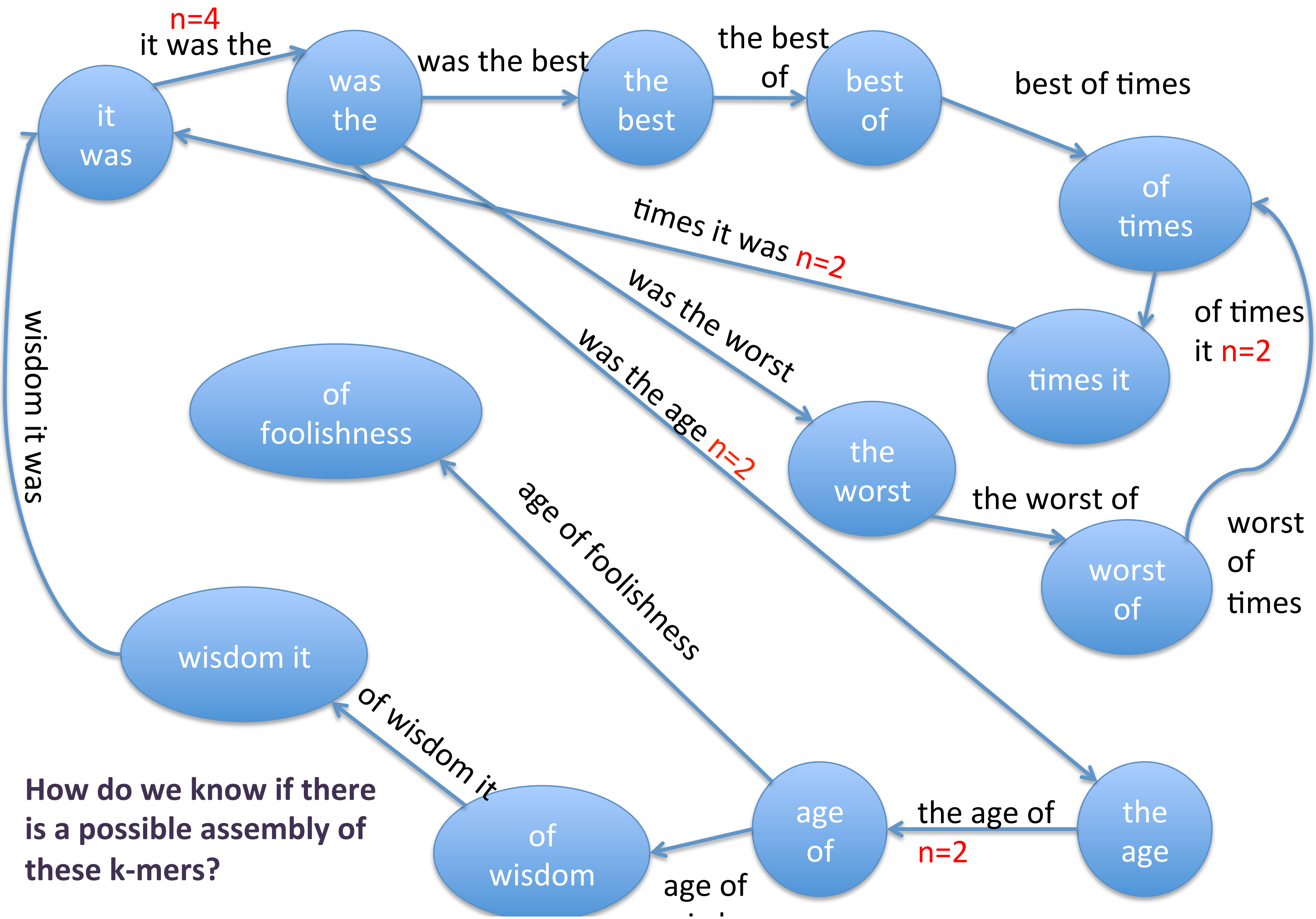
- What k-mer should we work with?
- What are the nodes?
- What are the edges?

# What are all the 3-mers?

it was the
was the best
the best of
best of times
of times it
times it was
was the worst
the worst of
worst of times
was the age
the age of
age of wisdom
of wisdom it
wisdom it was
age of foolishness
of foolishness it
foolishness it was
was the epoch
the epoch of
epoch of belief
of belief it
belief it was
epoch of incredulity
of incredulity it
incredulity it was
was the season
the season of
season of Light
of Light it
Light it was
season of Darkness
of Darkness it
Darkness it was
was the spring
the spring of
spring of hope
of hope it
hope it was
was the winter
the winter of
winter of despair

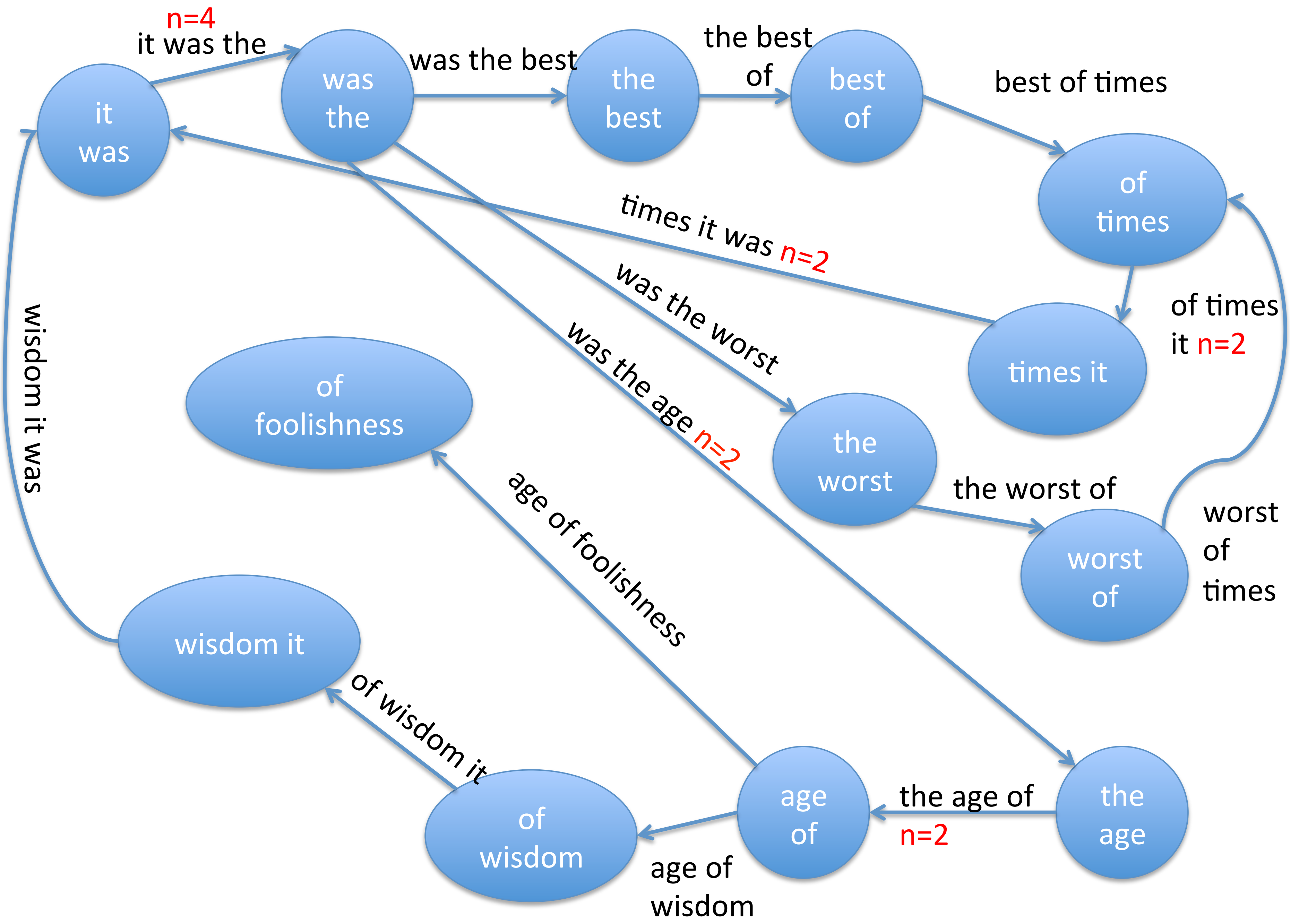
# What are all the 2-mers?

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



**How do we know if there is a possible assembly of these k-mers?**

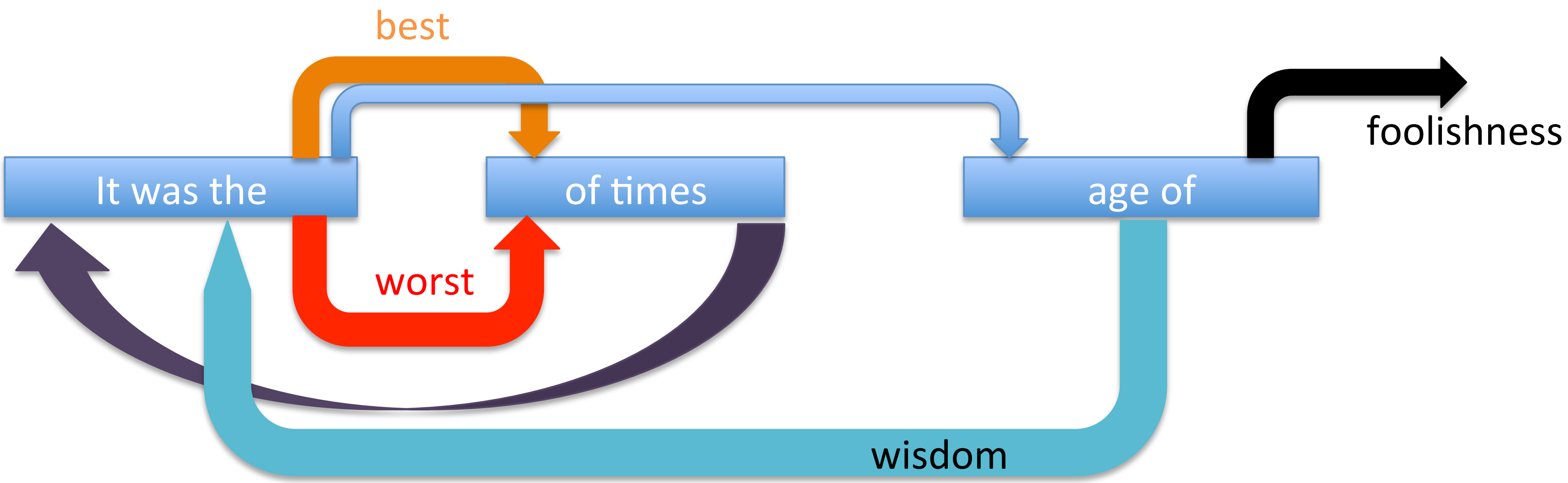
All nodes except beginning and end have equal indegrees and outdegrees


n=4
it was the
it was
was the
was the best
the best
the best of
best of
best of times
of times
of times it n=2
times it
times it was n=2
was the worst
the worst
the worst of
worst of
worst of times
was the age n=2
the age
the age of n=2
age of
age of wisdom
of wisdom
of wisdom it
wisdom it
wisdom it was
age of foolishness
of foolishness

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

It was the age of wisdom it was the best of times it was the worst of time it was the age of foolishness



Other permutations that begin with it was and end with of foolishness

How determine what order these phrases go in?

# Mate-Pairs

16 word long fragments, we have the beginning and end of each, with the end backwards.

It was the … age the was      of wisdom it … was it belief
the epoch of … Darkness of season   it was the … age the was
of foolishness it … it incredulity of
was the season … of spring the

How do you align these?

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

it was the ________________________________ was the age
it was the __________________________________was the
age
of foolishness it _________________________________of incredulity it
of wisdom _________
_______________________________belief it was
the epoch of __________________
_____________________season of Darkness
was the
season ____________________________________the spring of

# How did these help?

What about the long reads?

How did you use these?

Which was more useful, mate-pairs or long reads?

# Genome Assemblers

- Most use a De Bruijn graph assembly

Challenges

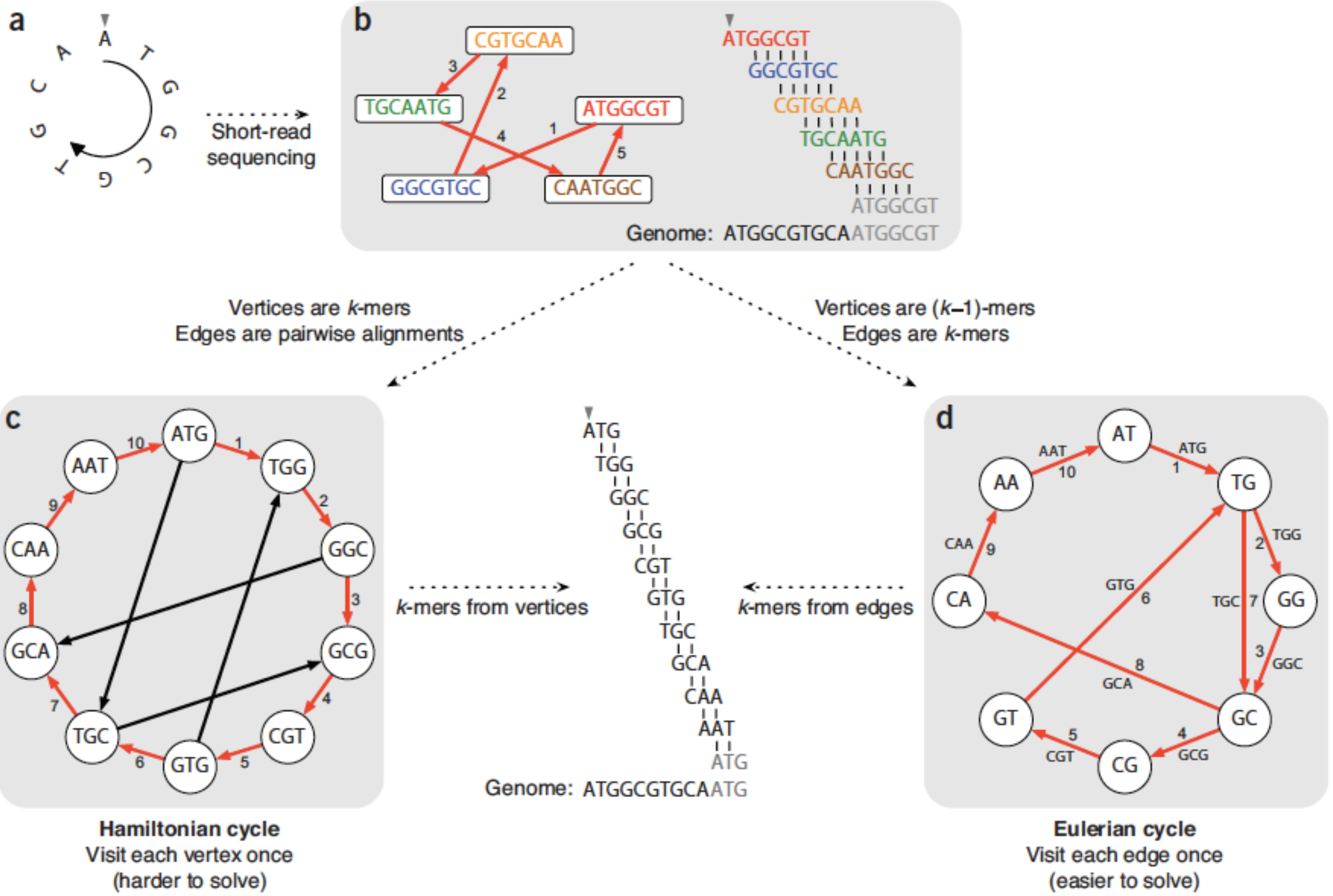
- Need a lot of data
  - 5-6 fold coverage in short reads
  - Mate-pairs or long reads
- Need a lot of memory
  - Building large graphs then traversing them to build up contigs
- Many designed for short reads so don't fully take advantage of longer reads

# Creating DeBruijn graphs for genomes

- We don't use full length reads to assemble the graph.
  - Why?
  - Did you have all possible 4-mers from phrase in your library?

# Creating DeBruijn graphs for genomes

- We don’t use full length reads to assemble the graph.
  - Need all possible k-mers to be represented in graph, so breaks reads into small chunks to represent all possible k-mers
  - Usually 21-100 bp k-mers used



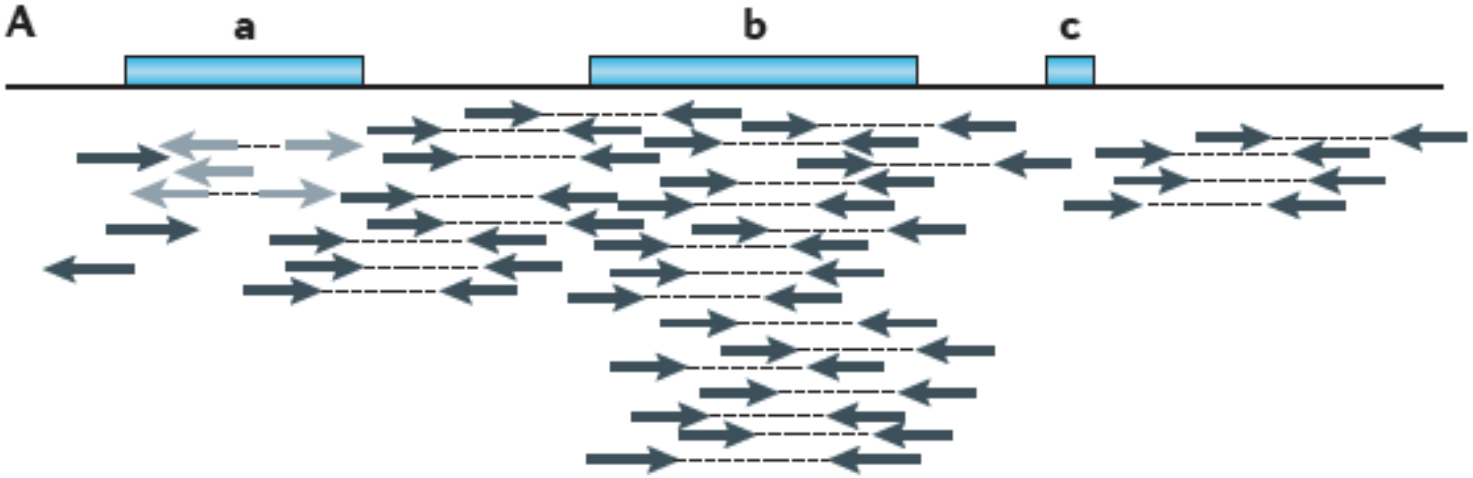
Hamiltonian cycle is NP-complete

Compeau, PE, et al.;*Nature Biotechnology* **29**, 987–991 (2011) doi:10.1038/nbt.2023

# How do we know we have correctly assembled the genome?

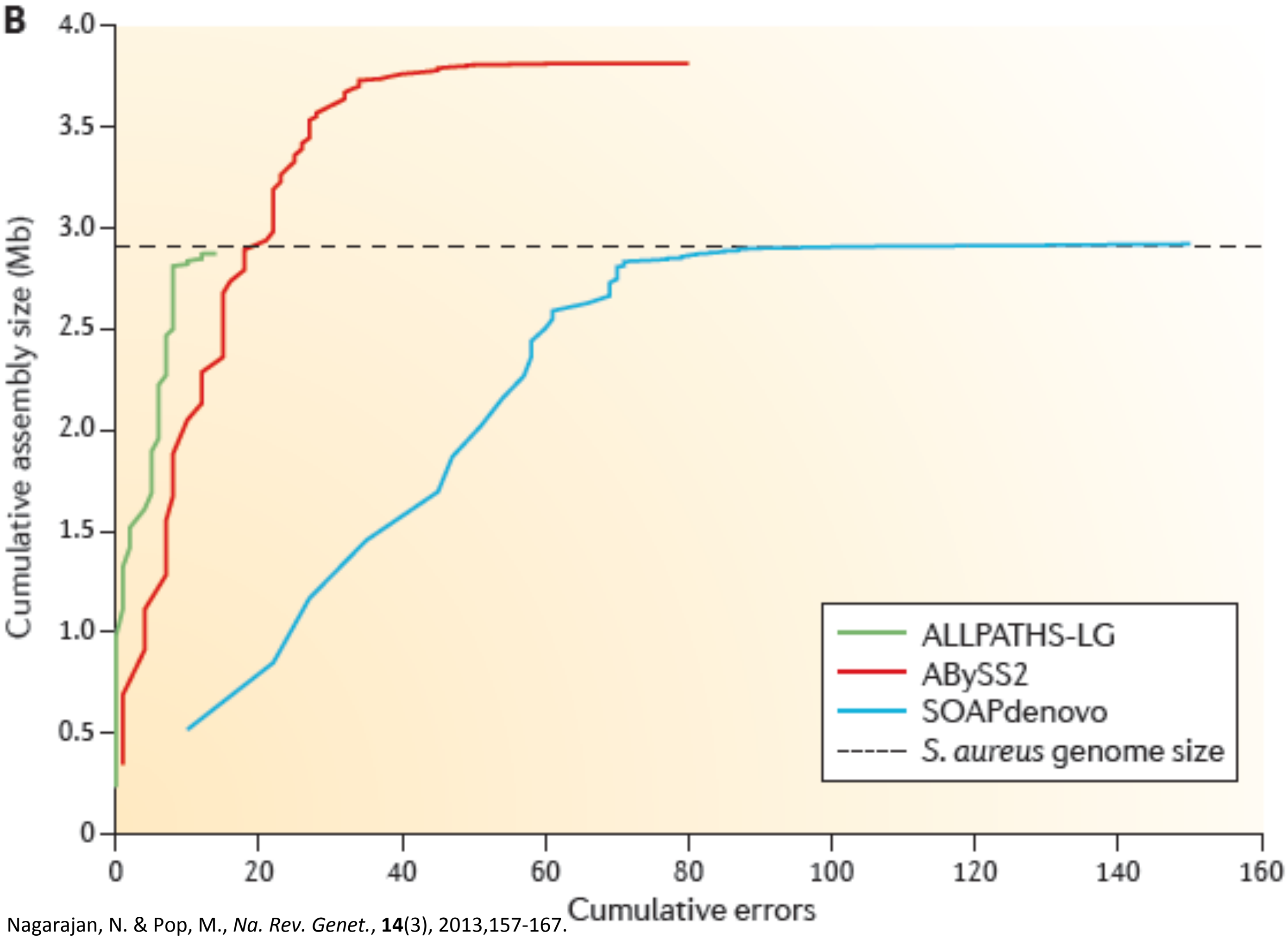
# How do we know we have correctly assembled the genome?

- Hard to know – **no ground truth**
- Use measures that correlate with correctness
  - Total size
  - Contiguity (N50)
  - Mate-pair alignment orientation
  - All regions have similar coverage
    - when does this not apply?
  - Similarity to similar genomes
  - Matches to transcriptome data



Nagarajan, N. & Pop, M., *Na. Rev. Genet.*, **14**(3), 2013,157-167.

If mate pairs misalign, what might be going on?



Nagarajan, N. & Pop, M., *Na. Rev. Genet.*, **14**(3), 2013,157-167.

Our measures of genome assembly don't always correlate with correctness