CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Jan 26

- Continue: overlap graph assembly
- Begin: de Bruijn graph assembly

Notes:

- Lab 1 due Wednesday
- Assembly reading posted: spend 1.5 hours max

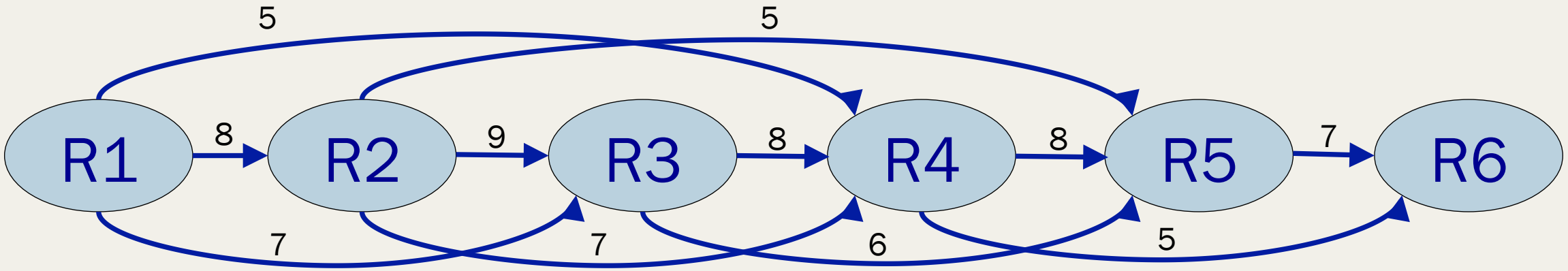
Steps of Overlap Graph Assembly (also called "overlap-layout-consensus")

1) Compute overlaps between all pairs of reads. With n = number of reads and L = length of reads, this is naively $O(n^2L^2)$. We will learn better ways of "aligning" sequences next week.

2) Construct a graph with reads as the nodes and directed, weighted edges between reads with >= T overlap.

Activity example: $L = 10$, $T = 5$

```
ATATATACTGGCGTATCGCAGTAAACGCGCCG
  R1: ACTGGCGTAT
  R2:   TGGCGTATCG
  R3:     GGCGTATCGC
  R4:       CGTATCGCAG
  R5:         TATCGCAGTA
  R6:           CGCAGTAAAC
```



Li et al, *Briefings in Functional Genomics* (2012)

Issues with overlap graphs

Bubbles

Velvet Assembler, Wikipedia

Repeats

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```

```
TAACTGTTCGCATCATCATCAT
                      CATCATCATCATCATCATCATGCATGCT
```

1 CAT

Repeats

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```

```
TAACTGTTCGCATCATCATCAT
                CATCATCATCATCATCATCATCATGCATGCT
```

2 CATs

Repeats

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```

```
TAACTGTTCGCATCATCATCAT
             CATCATCATCATCATCATCATCATGCATGCT
```

3 CATs

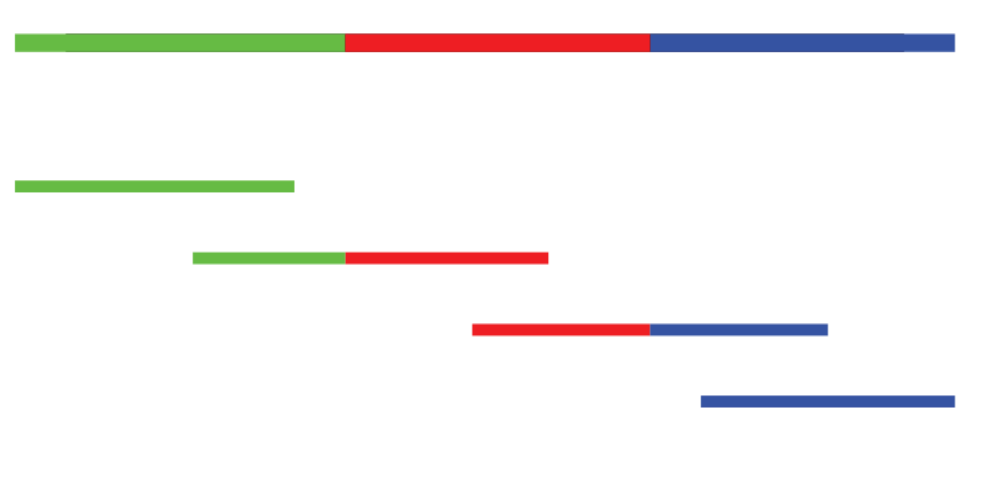
Repeats

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```

```
TAACTGTTCGCATCATCATCAT
          CATCATCATCATCATCATCATGCATGCT
```

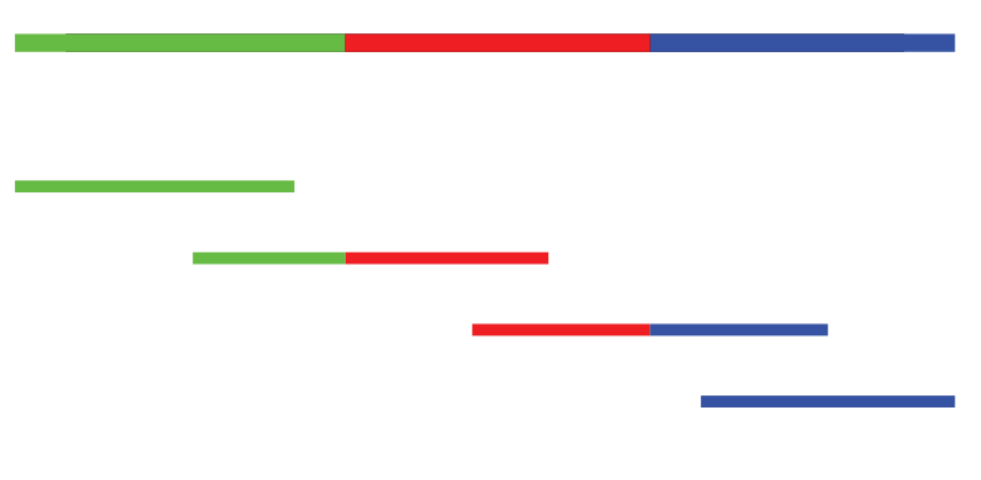
4 CATs

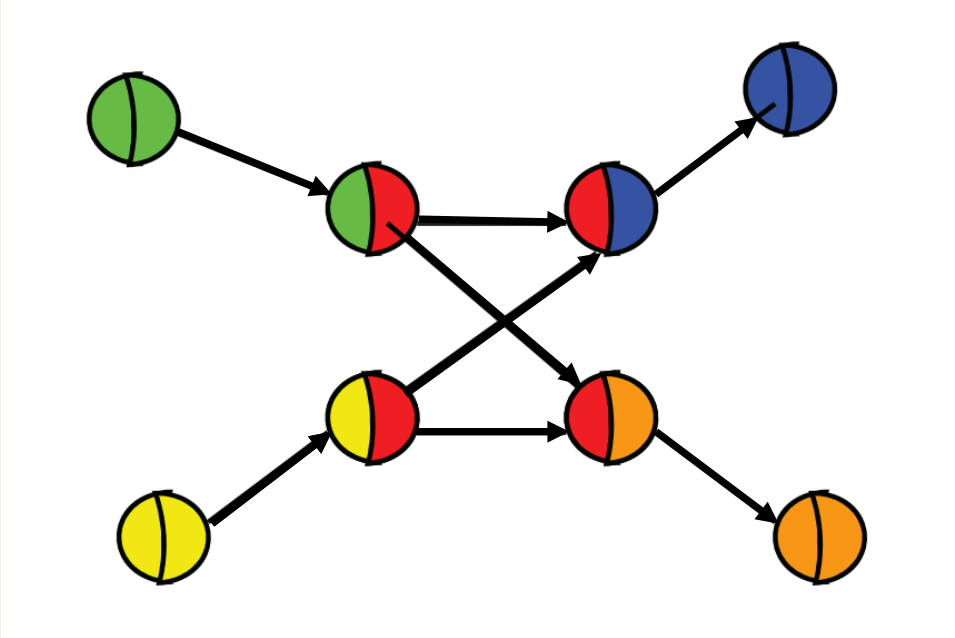
What would the graph look like for these reads?



Li et al, *Briefings in Functional Genomics* (2012)

What would the graph look like for these reads?





Li et al, *Briefings in Functional Genomics* (2012)

Back to overlap graph algorithm

Steps of Overlap Graph Assembly (also called "overlap-layout-consensus")

1) Compute overlaps between all pairs of reads. With n = number of reads and L = length of reads, this is naively $O(n^2L^2)$. We will learn better ways of "aligning" sequences next week.

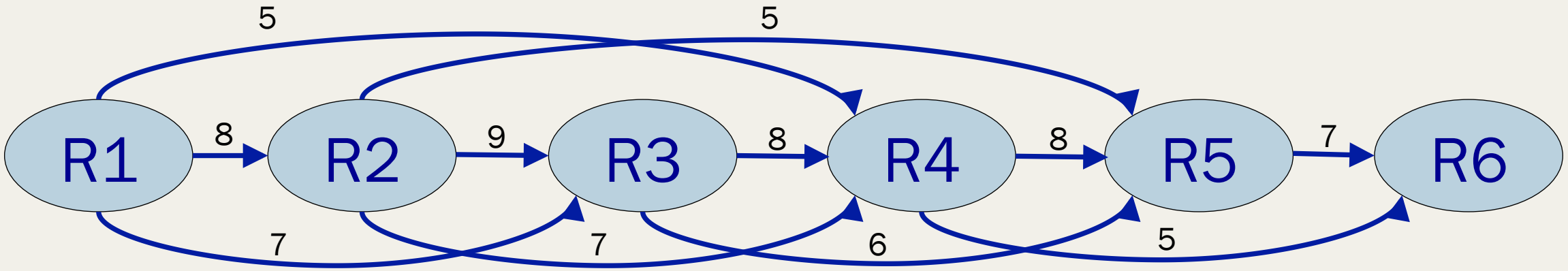
2) Construct a graph with reads as the nodes and directed, weighted edges between reads with >= T overlap.

3) "Layout" the graph and try to "group" stretches of the graph into "contigs" (short for contiguous), these are (hopefully) long portions of the original genome

4) Find a "consensus" *sequence* for each contig

Activity example: $L = 10$, $T = 5$

```
ATATATACTGGCGTATCGCAGTAAACGCGCCG
  R1: ACTGGCGTAT
  R2:   TGGCGTATCG
  R3:     GGCGTATCGC
  R4:       CGTATCGCAG
  R5:         TATCGCAGTA
  R6:           CGCAGTAAAC
```

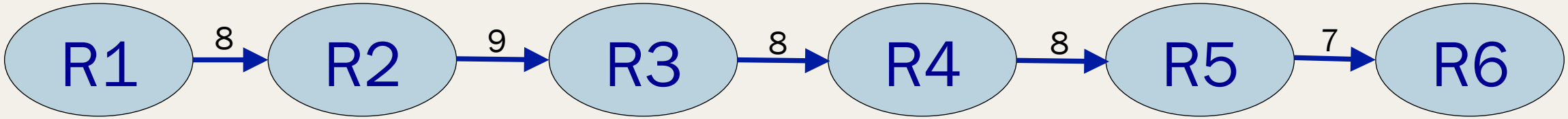


Li et al, *Briefings in Functional Genomics* (2012)

Activity example: $L = 10$, $T = 5$

```
ATATATACTGGCGTATCGCAGTAAACGCGCCG
  R1: ACTGGCGTAT
  R2:   TGGCGTATCG
  R3:     GGCGTATCGC
  R4:       CGTATCGCAG
  R5:         TATCGCAGTA
  R6:           CGCAGTAAAC
```

First simplification: remove edges that can be (transitively) inferred from other edges



Li et al, *Briefings in Functional Genomics* (2012)

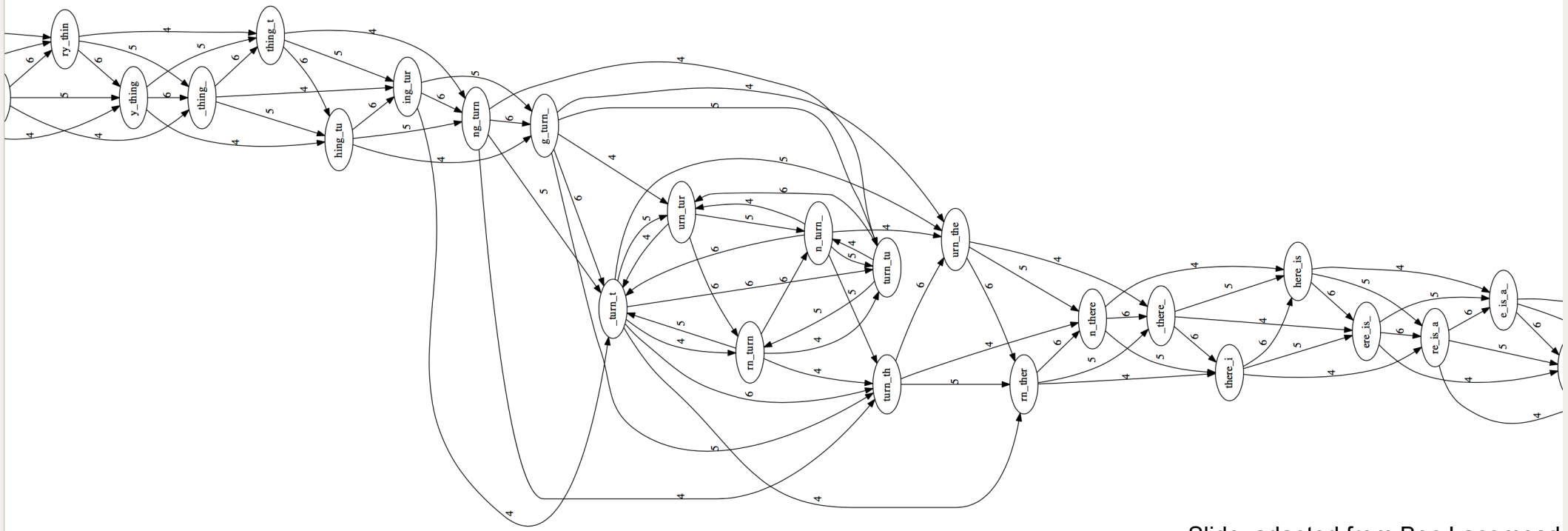
Layout

Overlap graph is big and messy. Contigs don't "pop out" at us.

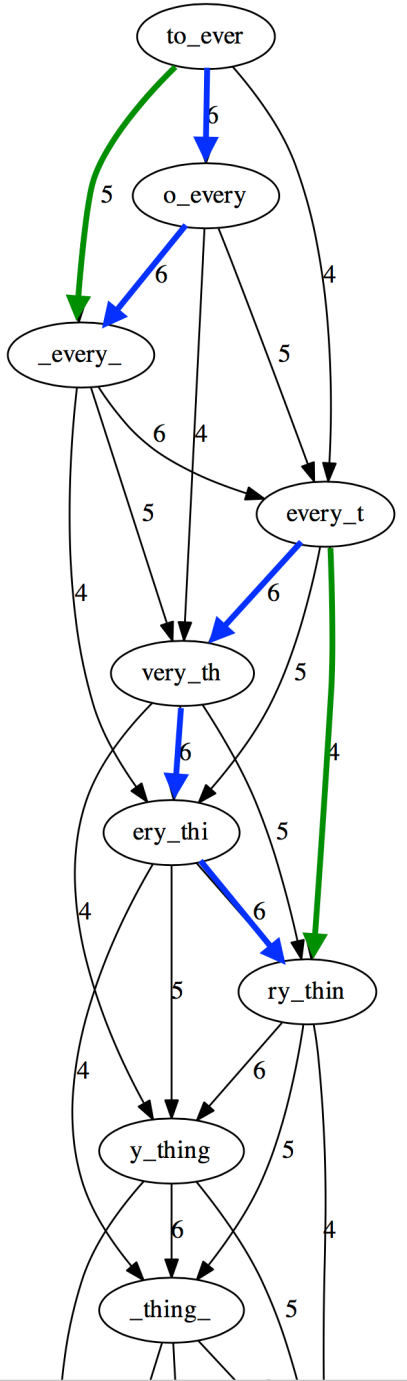
Below: part of the overlap graph for

`to_every_thing_turn_turn_turn_there_is_a_season`

$L = 7, T = 4$



Slide: adapted from Ben Langmead, John Hopkins

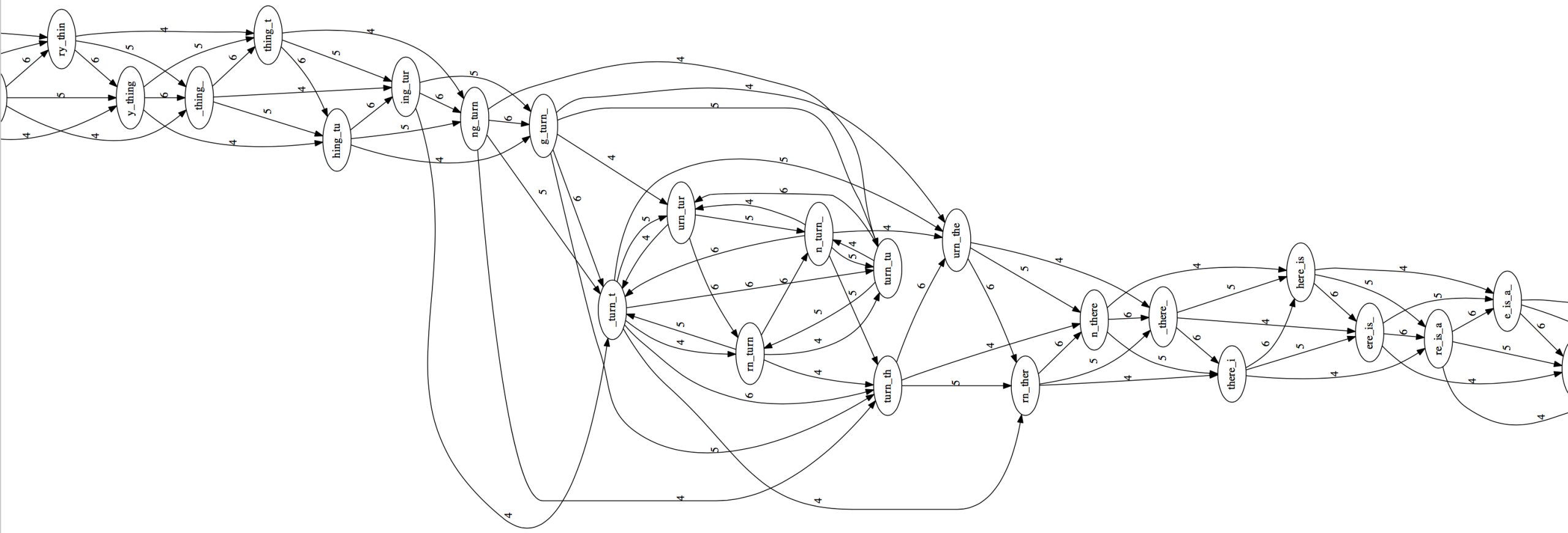


In this example: green edges can be inferred from blue

Slide: adapted from Ben Langmead, John Hopkins

Layout: remove transitively-inferrable edges

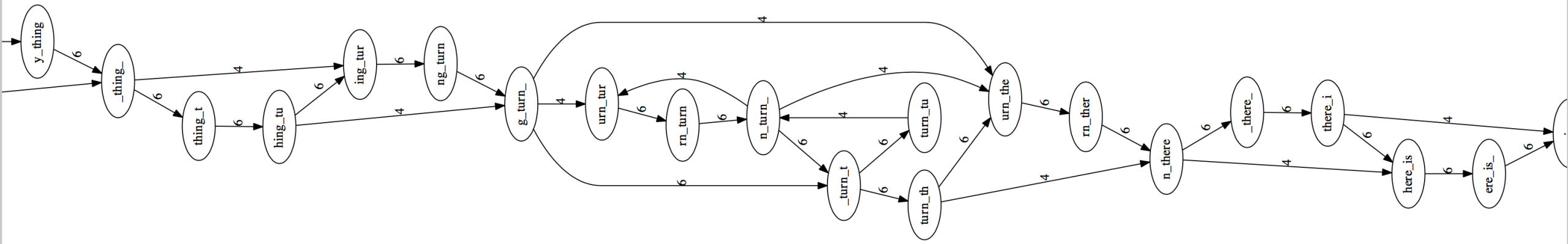
Before:



Slide: adapted from Ben Langmead, John Hopkins

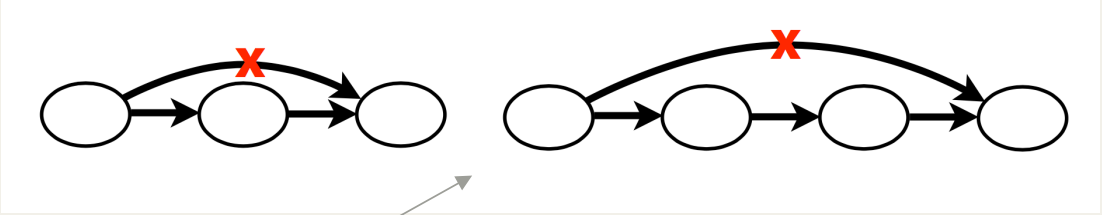
Layout: remove transitively-inferrable edges

After removing edges that skip one node

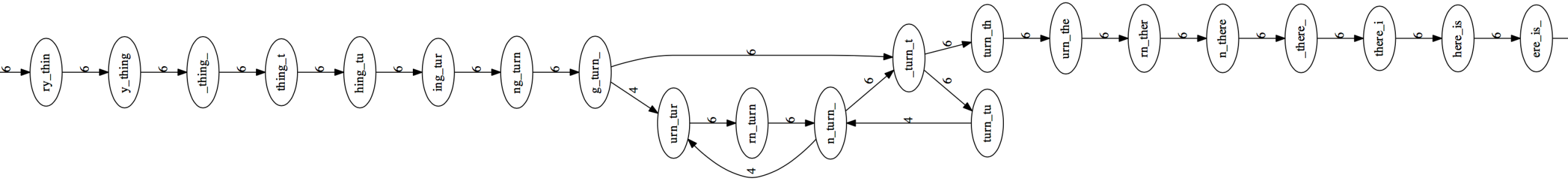


Slide: adapted from Ben Langmead, John Hopkins

Layout: remove transitively-inferrable edges



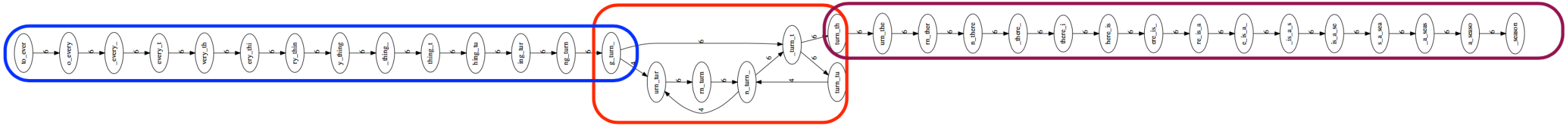
After removing edges that skip one or two nodes



Slide: adapted from Ben Langmead, John Hopkins

Layout

Emit *contigs* corresponding to the non-branching stretches



Contig 1

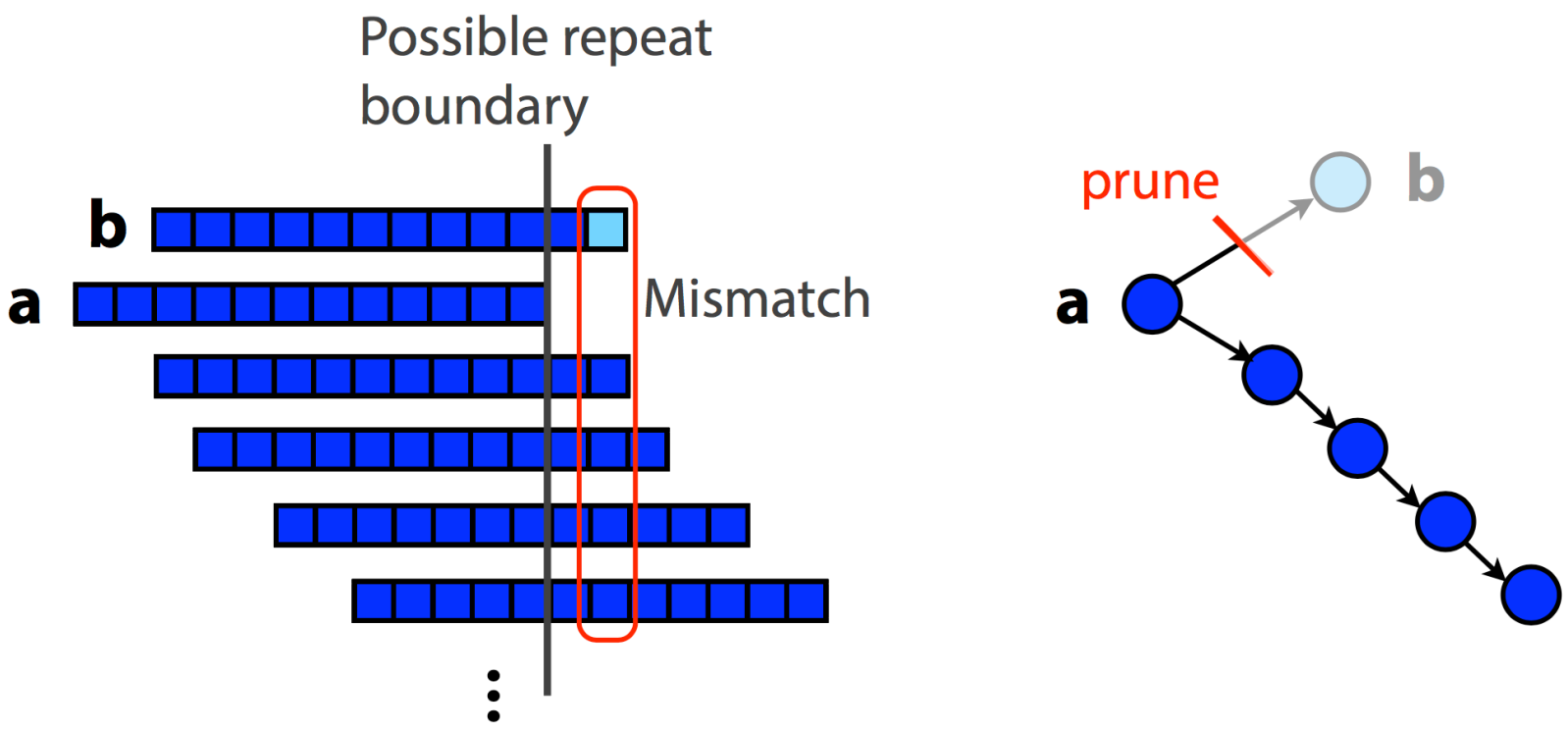
`to_every_thing_turn_`

Contig 2

`turn_there_is_a_season`

Unresolvable repeat

In practice, layout step also has to deal with spurious subgraphs, e.g. because of sequencing error



Mismatch could be due to sequencing error or repeat. Since the path through **b** ends abruptly we might conclude it's an error and prune **b**.

Slide: adapted from Ben Langmead, John Hopkins

Consensus

```
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
```

Take reads that make up a contig and line them up

```
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
```

Take *consensus*, i.e. majority vote

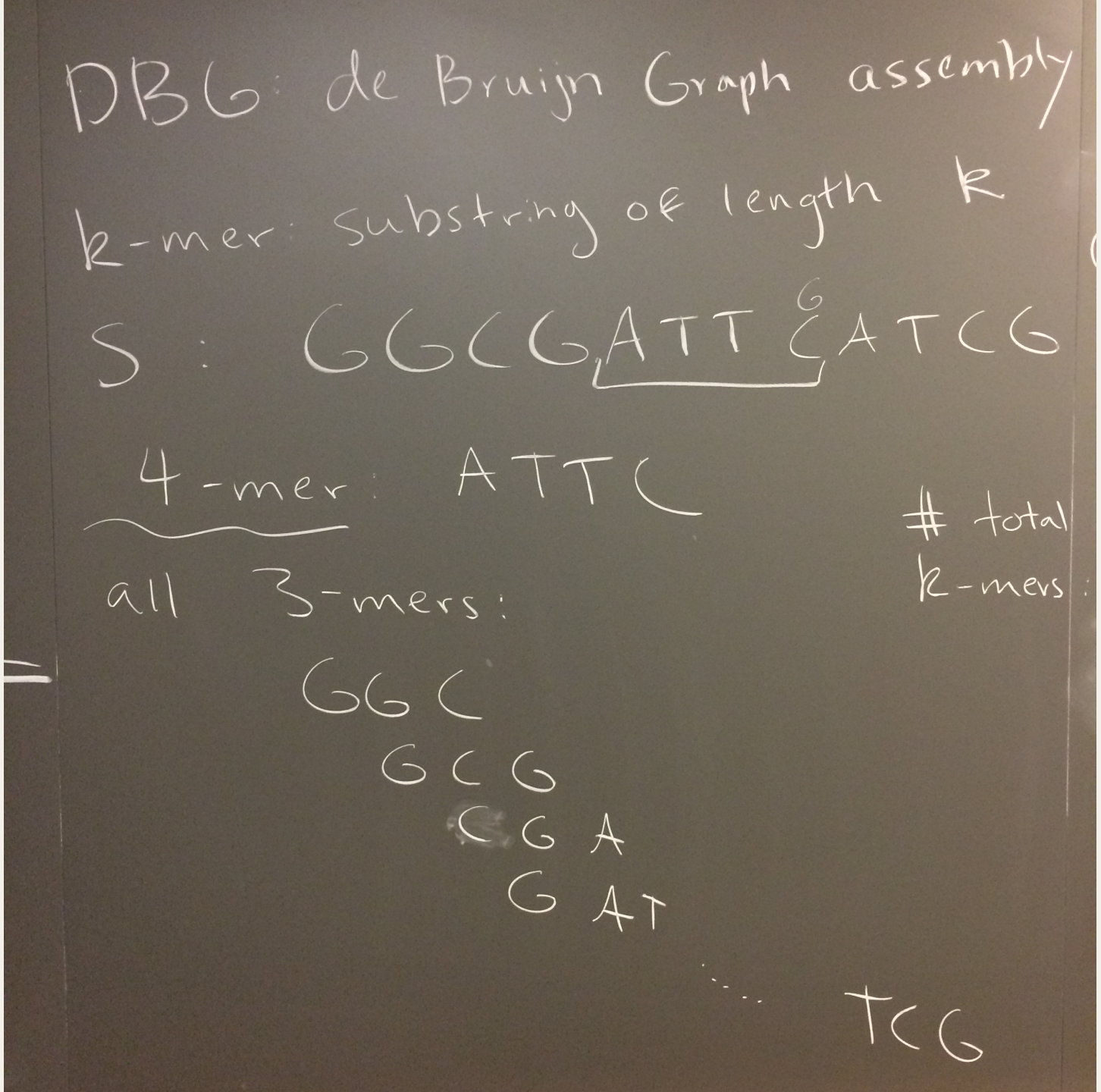
Slide: adapted from Ben Langmead, John Hopkins

Issues with overlap graph assembly

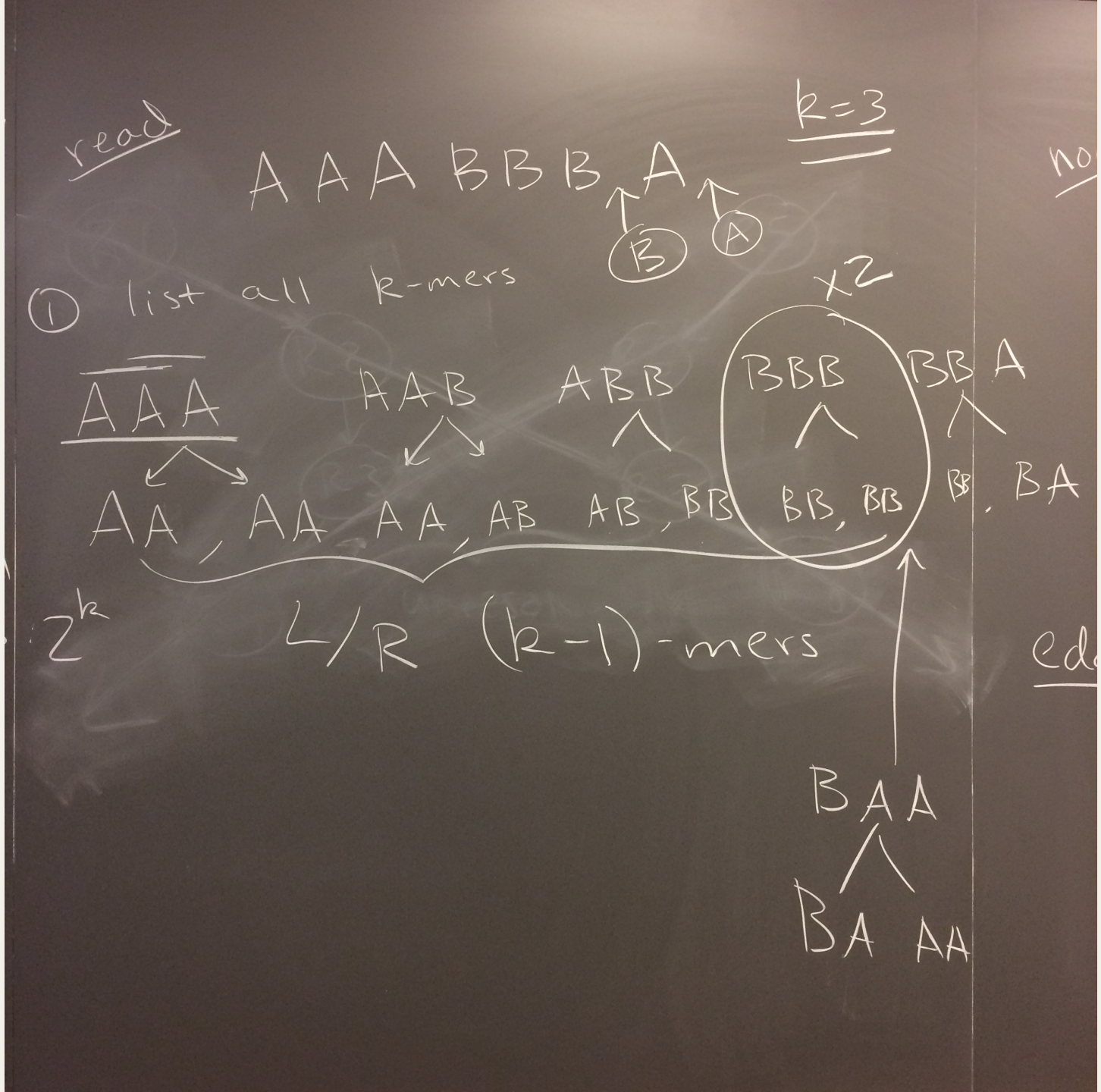
- Next-generation sequencing produces 100’s of millions (or even billions) of reads
- With one node per read this is computationally intractable for large genomes
- What if the nodes in our graph were not reads?

De Bruijn Graph (DBG) Assembly

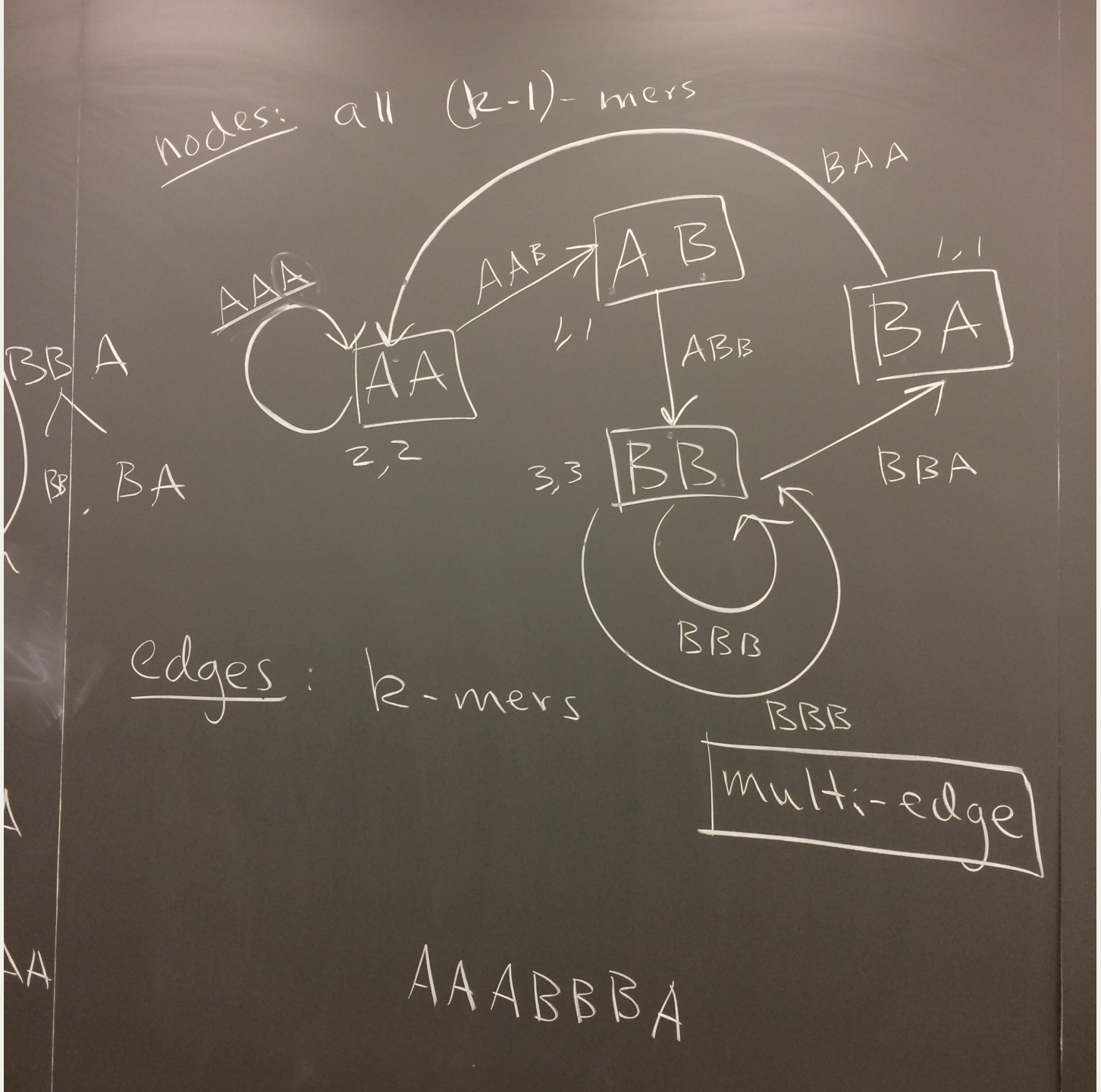
Examples: Ben Langmead, John Hopkins



Examples: Ben Langmead, John Hopkins

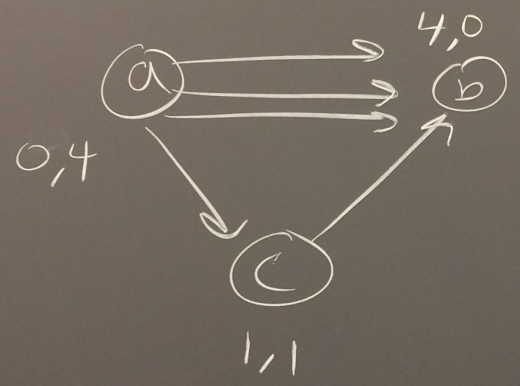


Examples: Ben Langmead, John Hopkins



Examples: Ben Langmead, John Hopkins

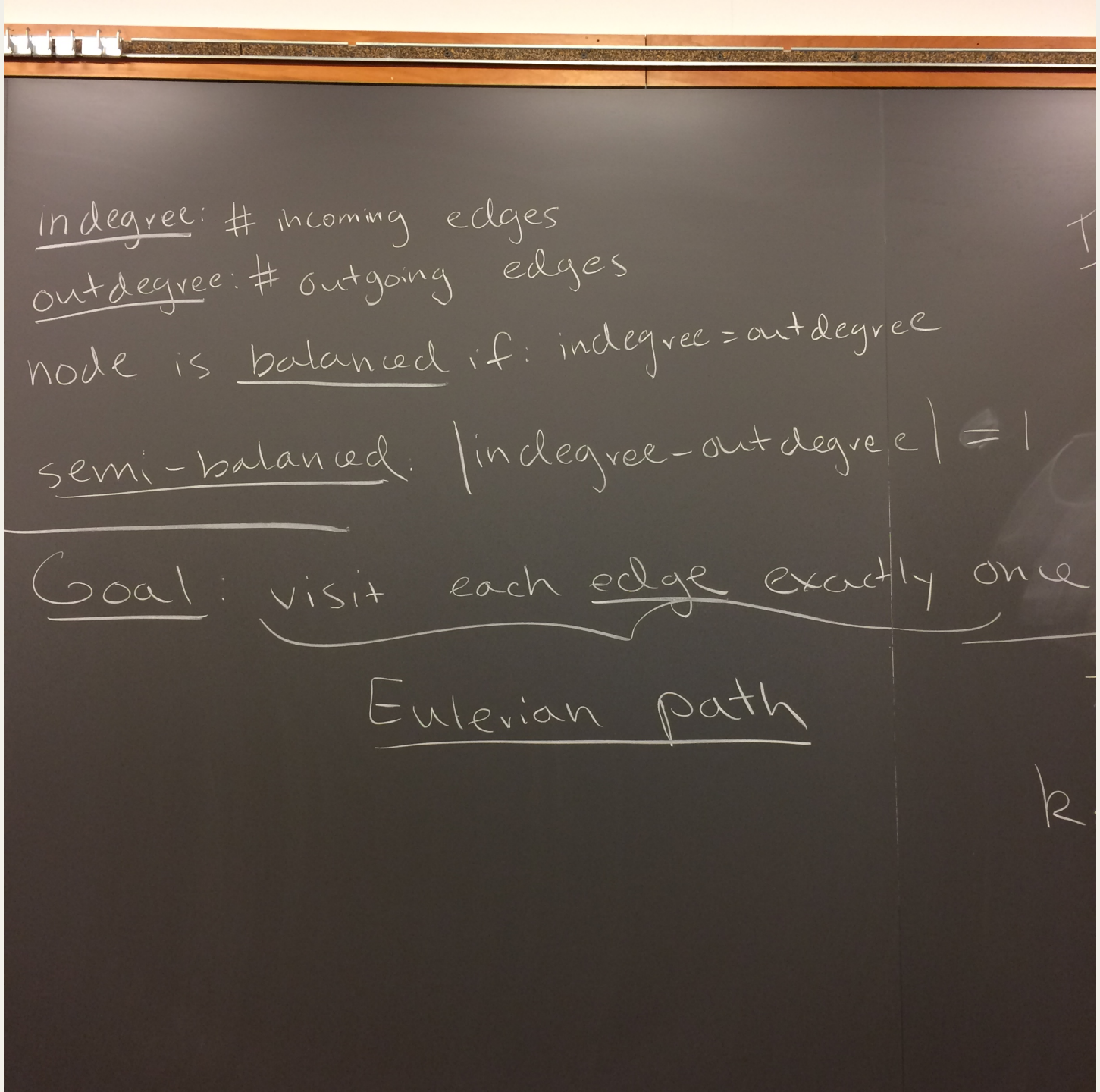
de Bruijn graph: directed multigraph.



$V = \{a, b, c\}$

$E = \{(a,b), (a,b), (a,b), (a,c), (c,b)\}$

Examples: Ben Langmead, John Hopkins



Examples: Ben Langmead, John Hopkins