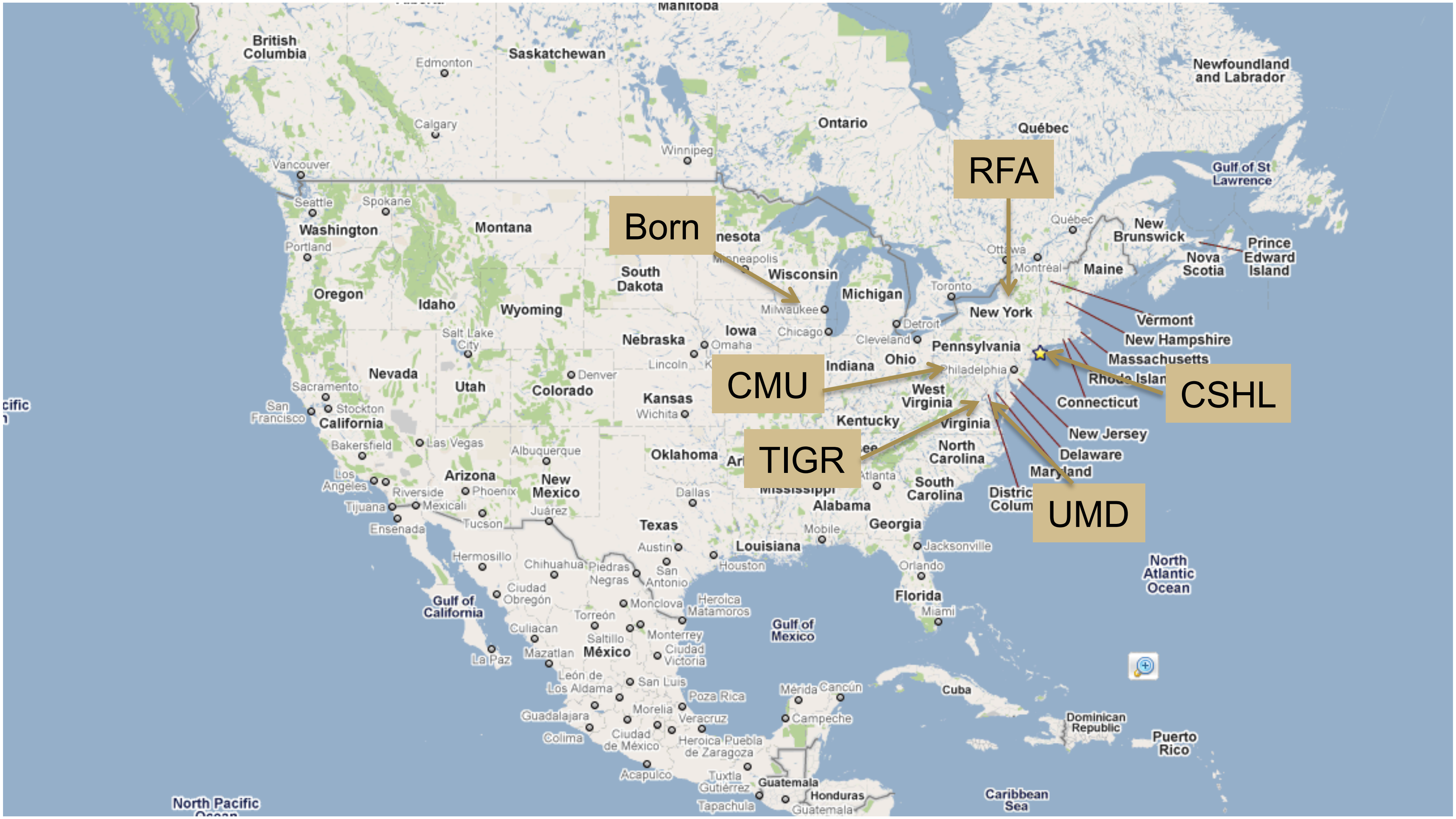
Schatzlab Research Projects

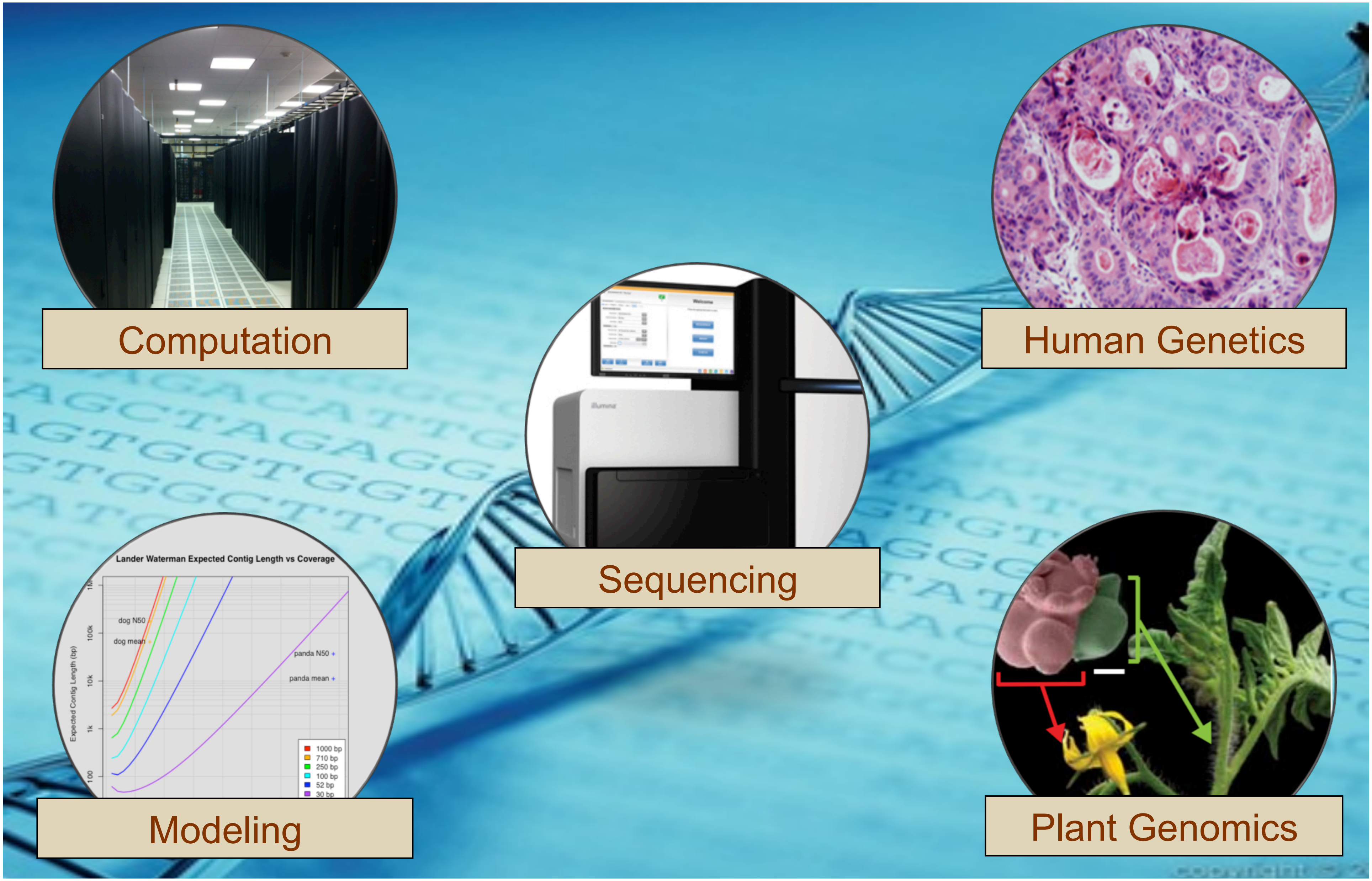
Michael Schatz

Oct 16, 2013
Research Topics in Biology, WSBS

A Little About Me



Schatz Lab Overview



Computation

Sequencing

Human Genetics

Modeling

Plant Genomics

Milestones in Molecular Biology

There is tremendous interest to sequence:

- What is your genome sequence?
- How does your genome compare to my genome?

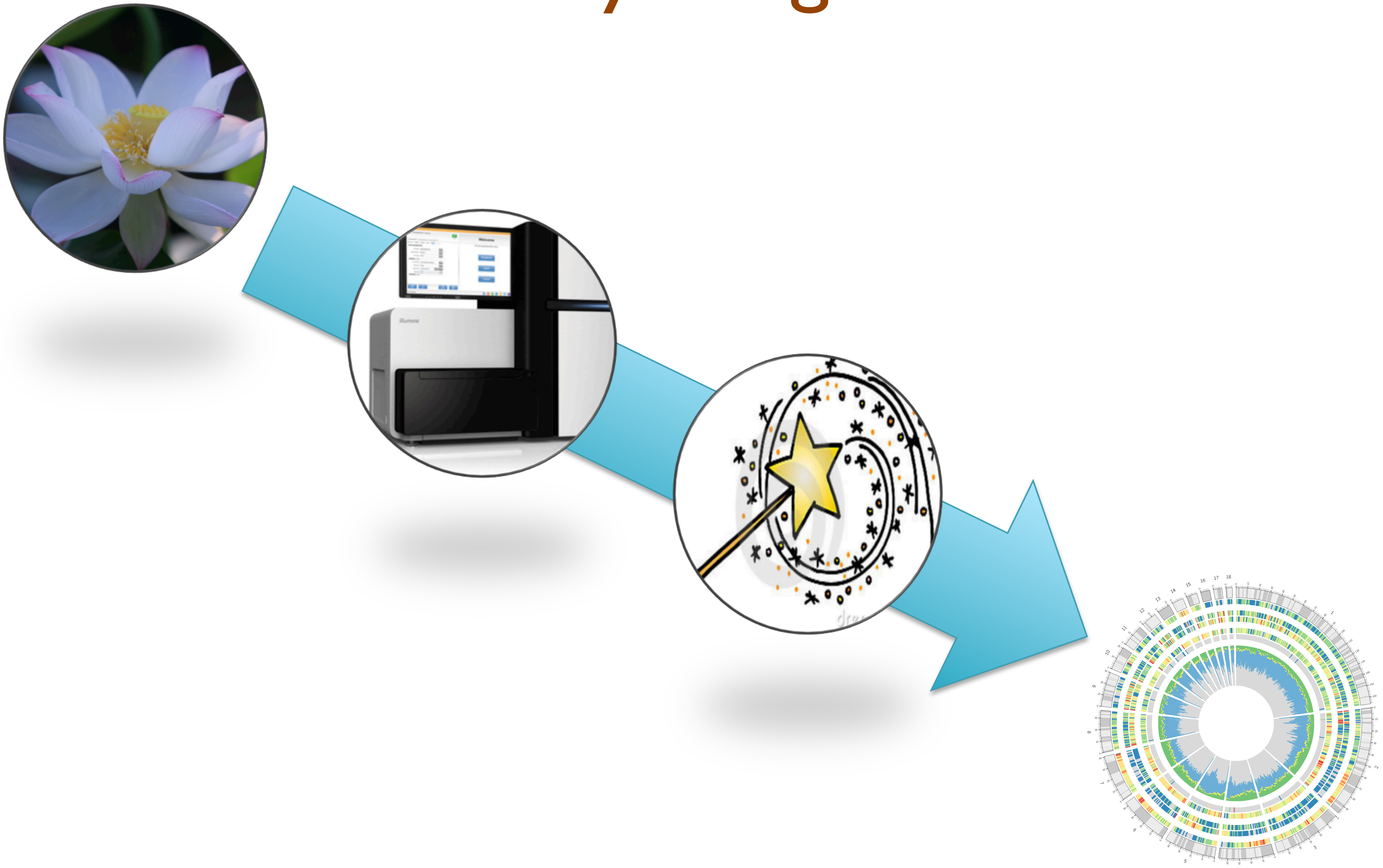
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?

- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?

- What virus and microbes are living inside you?
- How has the disease mutated your genome?
- What drugs should we give you?

- …

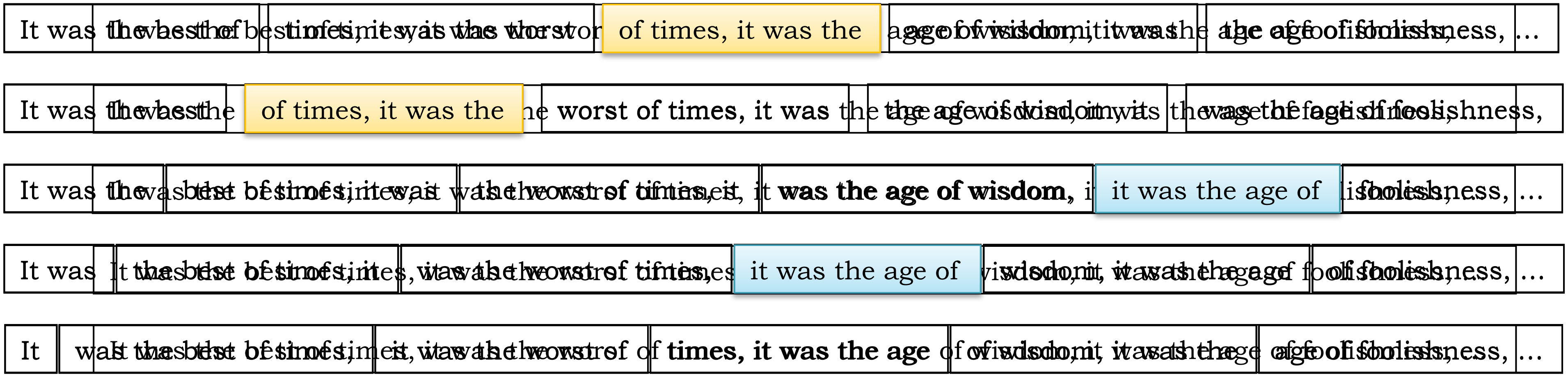
What is your genome?



Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.)
Ming, R et al. (2013) Genome Biology 14:R41

Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of <u>A Tale of Two Cities</u>
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of

age of wisdom, it was

best of times, it was

it was the age of

it was the age of

it was the worst of

of times, it was the

of times, it was the

of wisdom, it was the

the age of wisdom, it

the best of times, it

the worst of times, it

times, it was the age

times, it was the worst

was the age of wisdom,

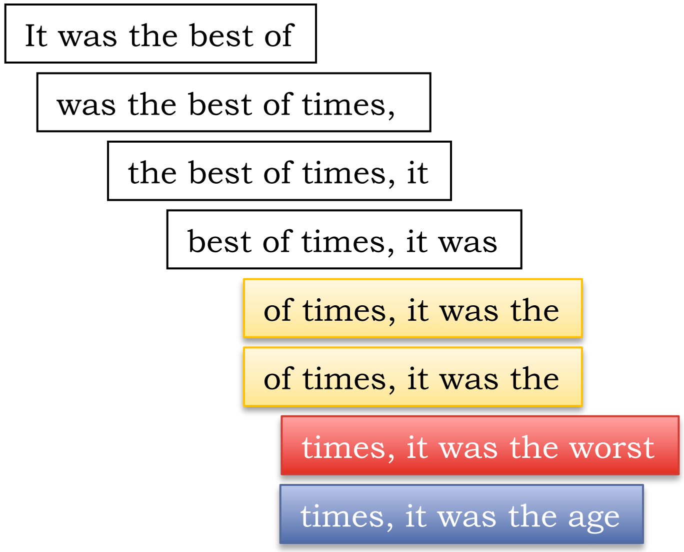
was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was



It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

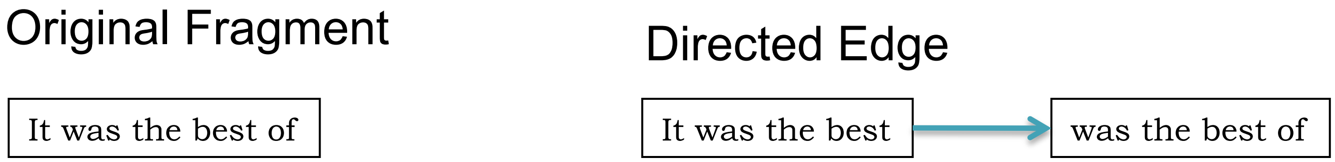
The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model sequence reconstruction as a graph problem.

de Bruijn Graph Construction

- $D_k = (V,E)$
  - V = All length-k subfragments ($k < l$)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



Original Fragment

It was the best of

Directed Edge

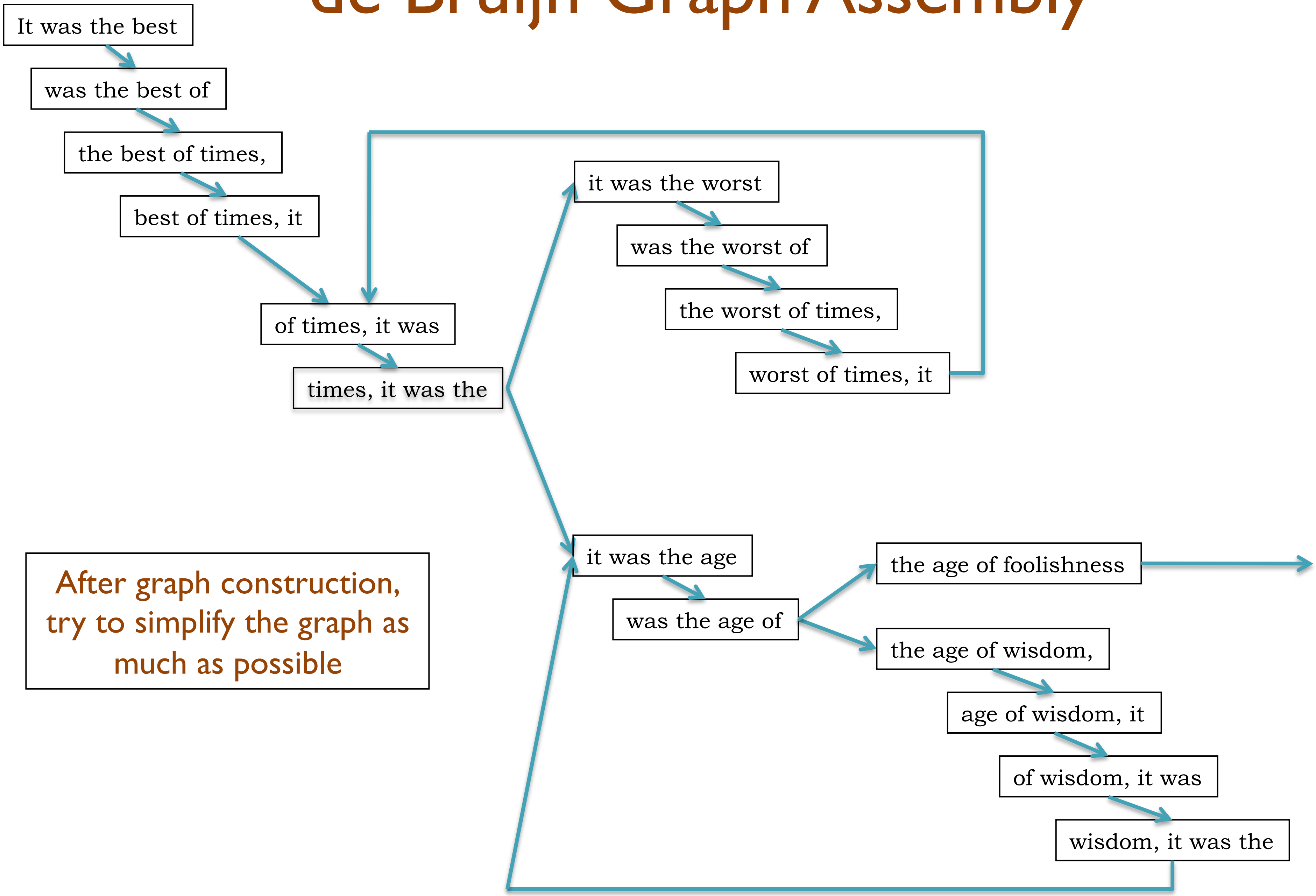
It was the best

was the best of

- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



It was the best

was the best of

the best of times,

best of times, it

of times, it was

times, it was the

it was the worst

was the worst of

the worst of times,

worst of times, it

it was the age

was the age of

the age of foolishness

the age of wisdom,

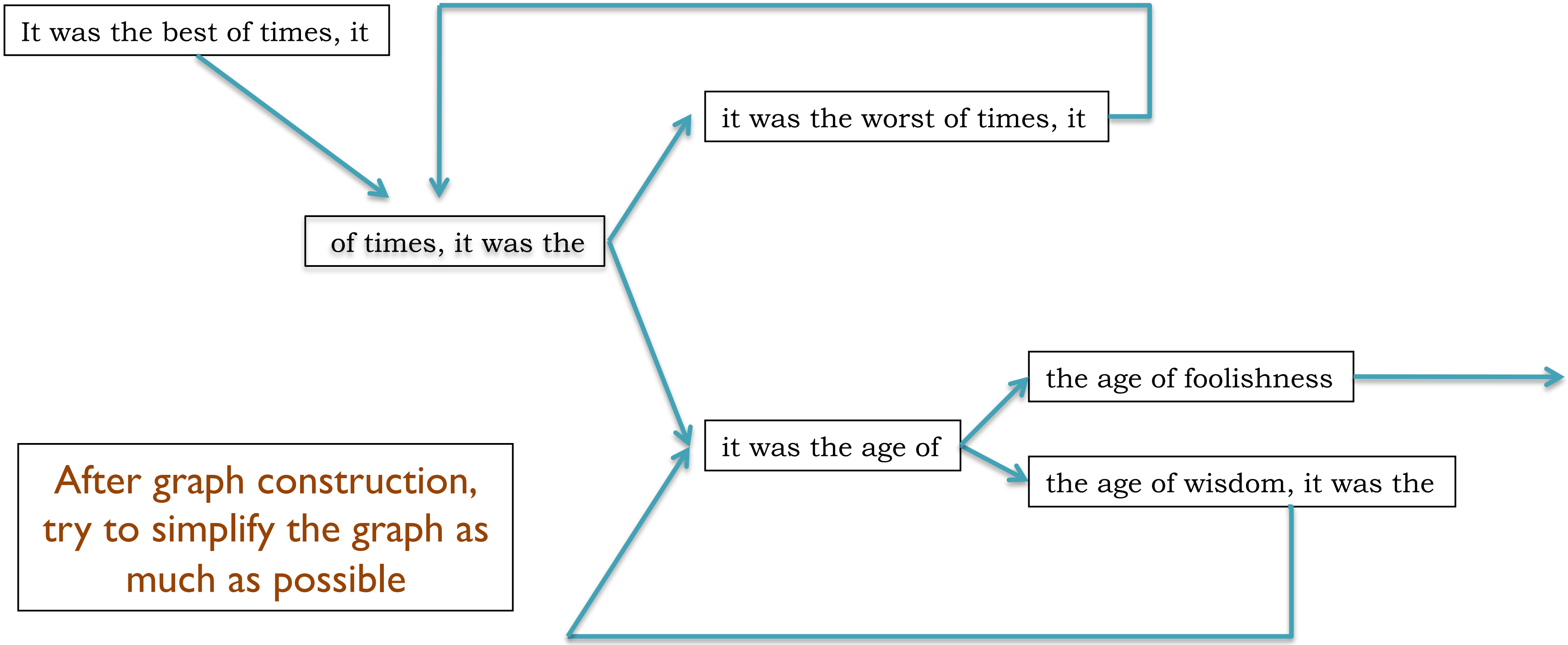
age of wisdom, it

of wisdom, it was

wisdom, it was the

After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly


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

After graph construction, try to simplify the graph as much as possible

The full tale

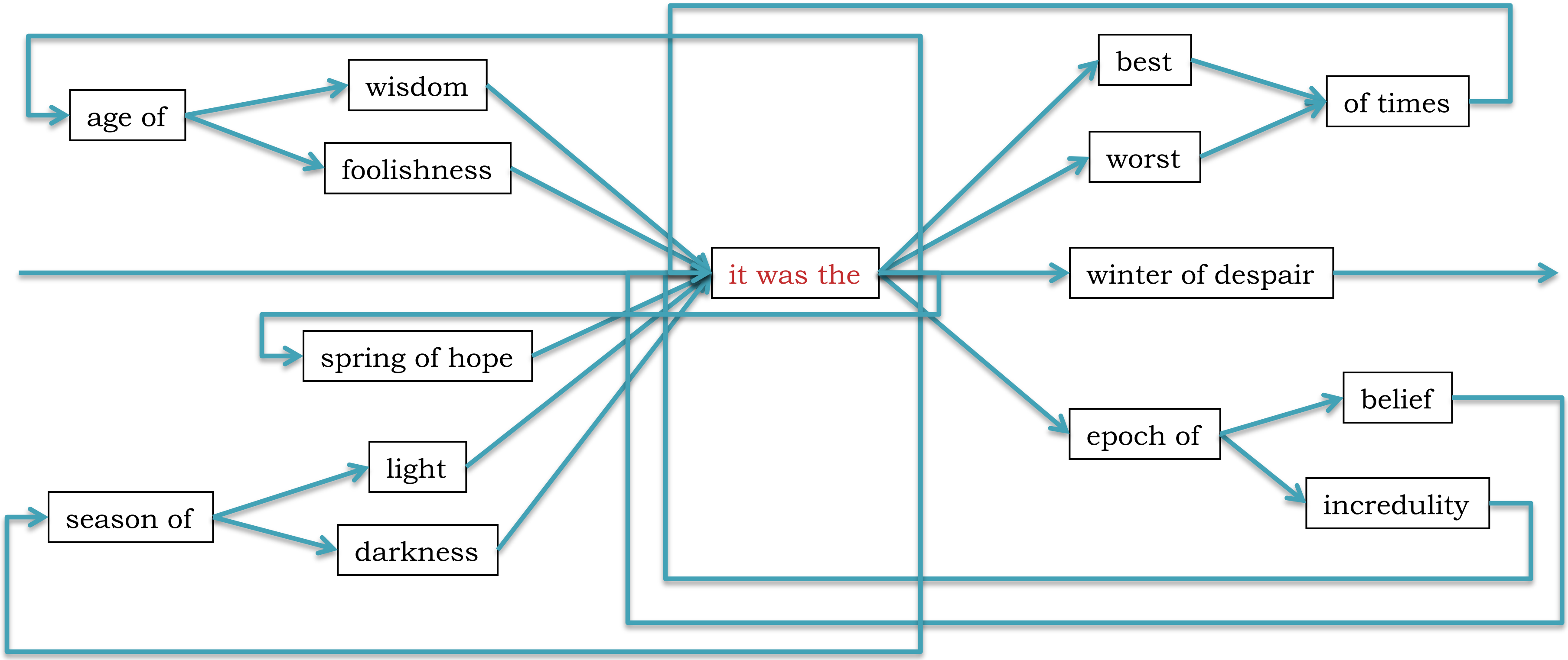
… it was the best of times it was the worst of times …

… 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 …

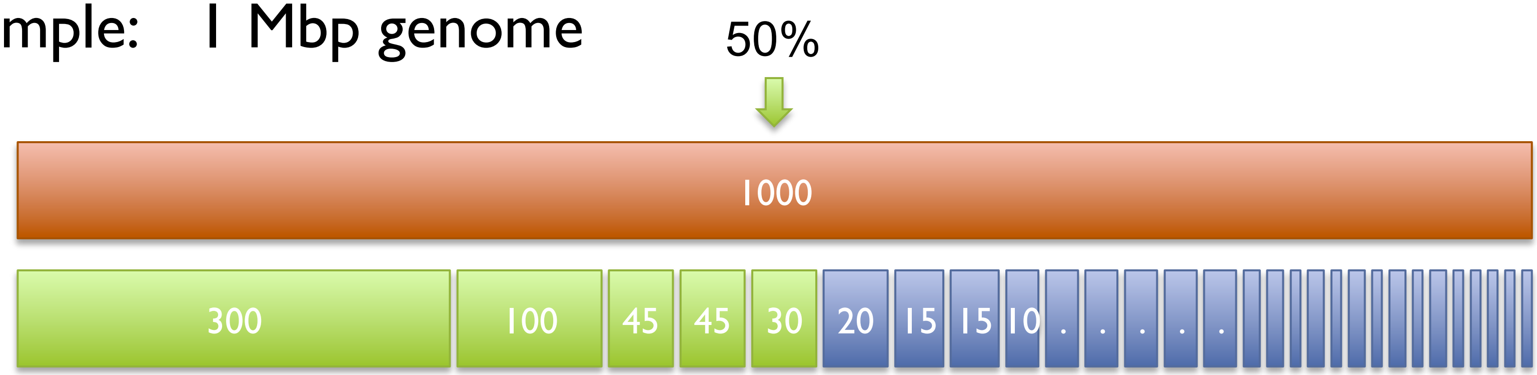
… it was the spring of hope it was the winder of despair …



N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



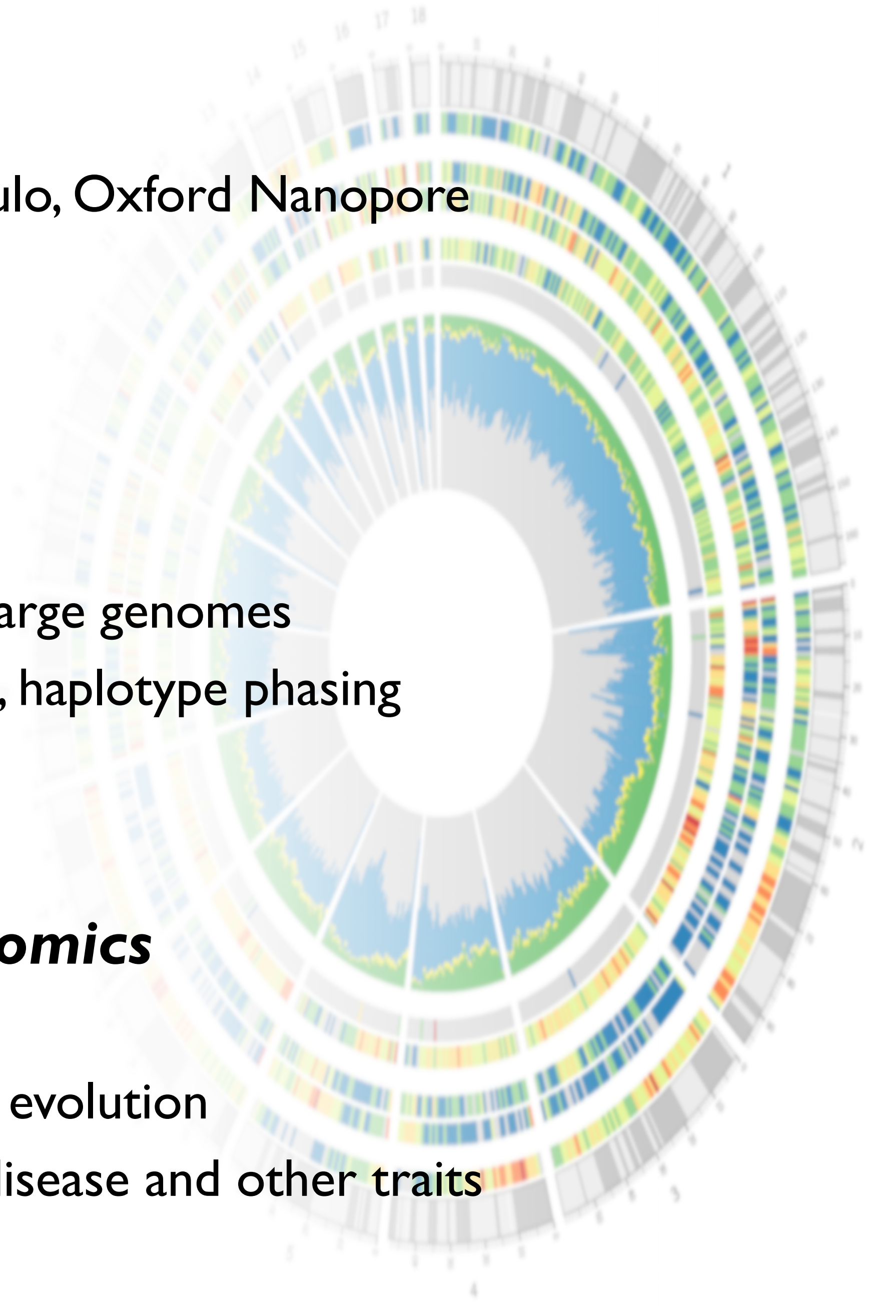
N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

N50 values are only meaningful to compare when base genome size is the same in all cases

Research Dimensions



1. ***New Biotechnology***
   - Sequencing: Pacific Biosciences, Moleculo, Oxford Nanopore
   - Mapping: BioNanoGenomics, OpGen
   - Faster/Cheaper/Better assemblies

2. ***Algorithmics***
   - Algorithms for assembling extremely large genomes
   - Improved error correction, scaffolding, haplotype phasing
   - Analyzing populations of genomes

3. ***Annotation & Comparative Genomics***
   - Identifying functional elements
   - Cross species comparisons, models of evolution
   - Identifying mutations responsible for disease and other traits

Acknowledgements

Schatz Lab

Giuseppe Narzisi

Shoshana Marcus

James Gurtowski

Srividya Ramakrishnan

Hayan Lee

Rob Aboukhalil

Mitch Bekritsky

Charles Underwood

Tyler Gavin

Alejandro Wences

Greg Vurture

Eric Biggers

Aspyn Palatnick

CSHL

Hannon Lab

Gingeras Lab

Iossifov Lab

Levy Lab

Lippman Lab

Lyon Lab

Martienssen Lab

McCombie Lab

Ware Lab

Wigler Lab

IT Department

NBACC

Adam Phillippy

Sergey Koren

SFARI

SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

National Human
Genome Research
Institute

U.S. DEPARTMENT OF
ENERGY

NSF

Thank You!

http://schatzlab.cshl.edu
@mike_schatz