Schatzlab Research Projects

Michael Schatz

CSH

Sept 10, 2014
Research Topics in Biology, WSBS
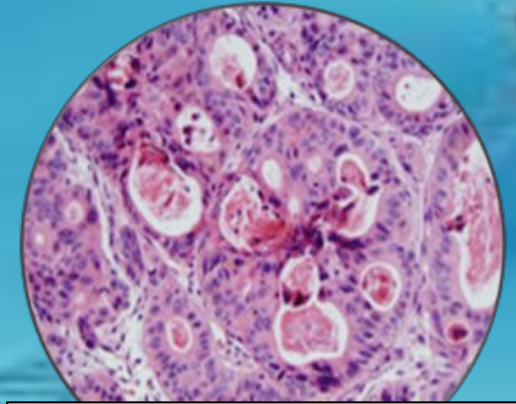
A Little About Me

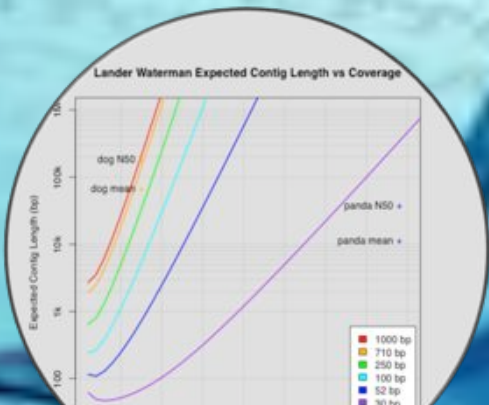# Schatz Lab Overview
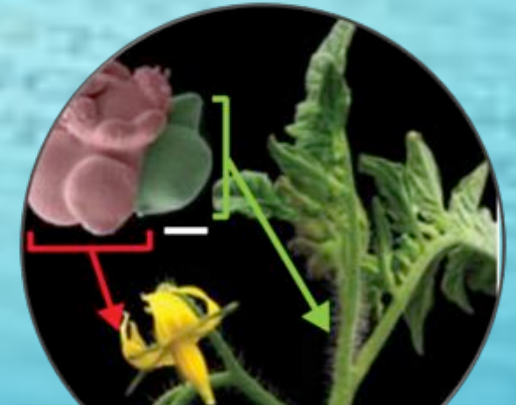
Unsolved Questions in Biology

- W
- H

- W
- H
- H

- H
- H
- H
- W

- W
- H
- W...hat drugs and treatments should we give you?

- *Plus thousands and thousands more*

The instruments provide the data, but none of the answers to any of these questions.

What software and systems will?

And who will create them?

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 |

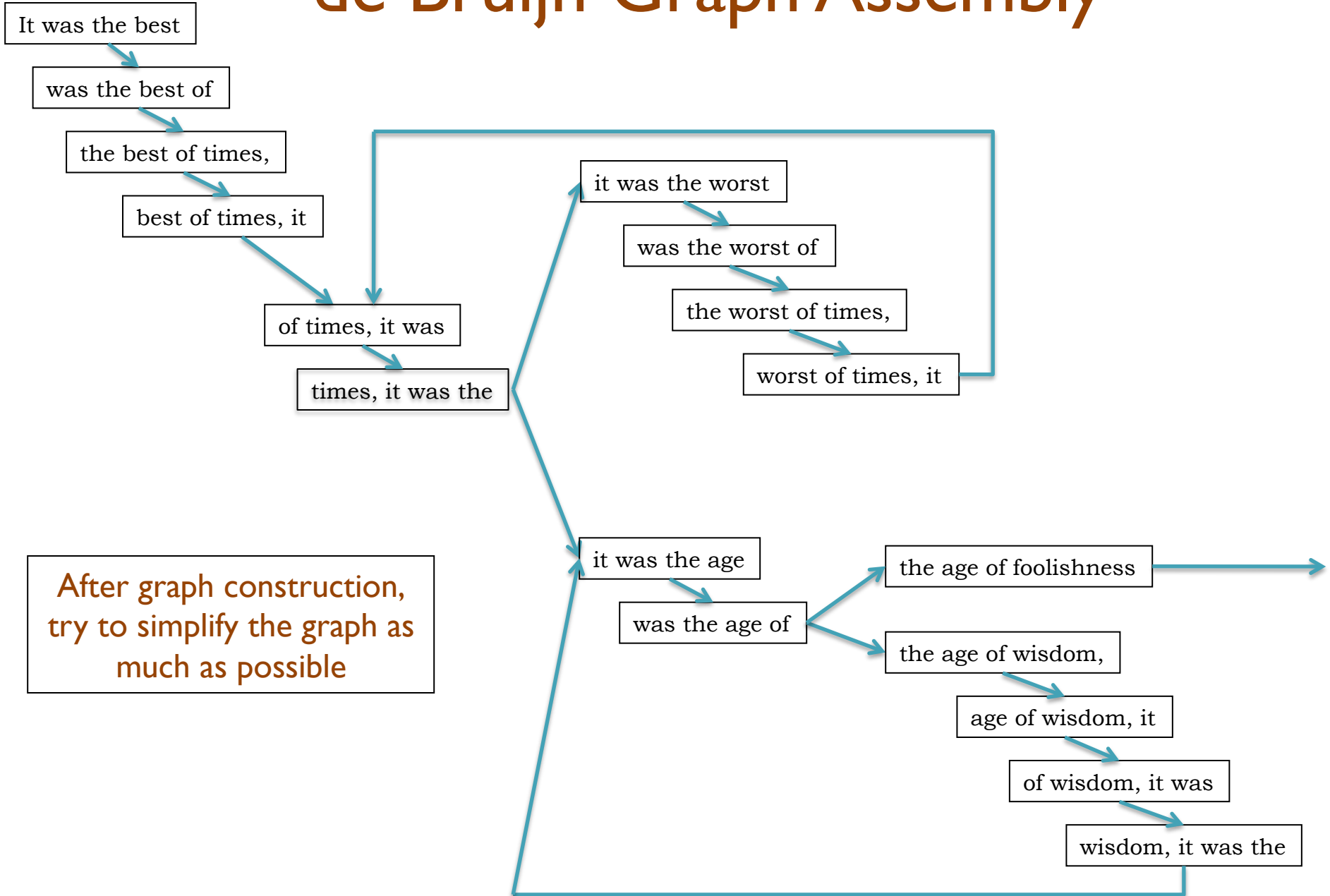Directed Edge

| It was the best | → | was the best of |

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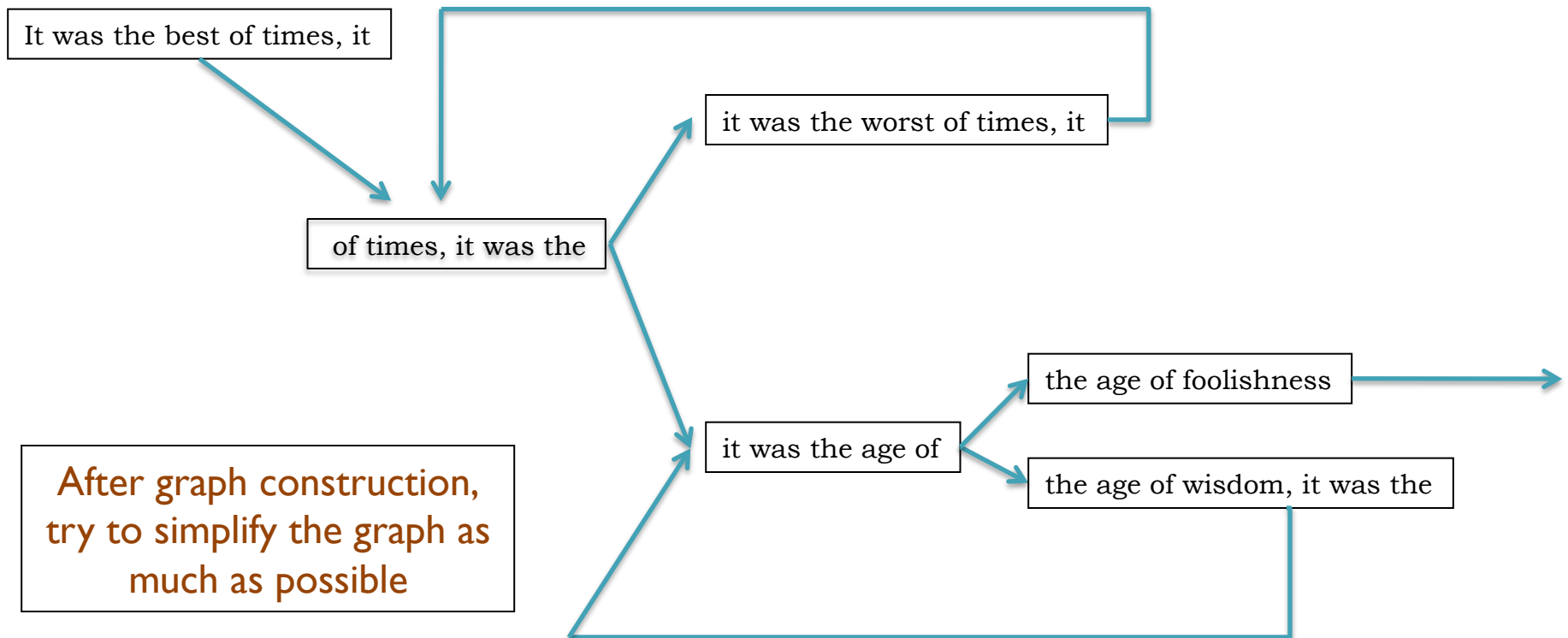 - Overlaps between sequences implicitly computed

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 …
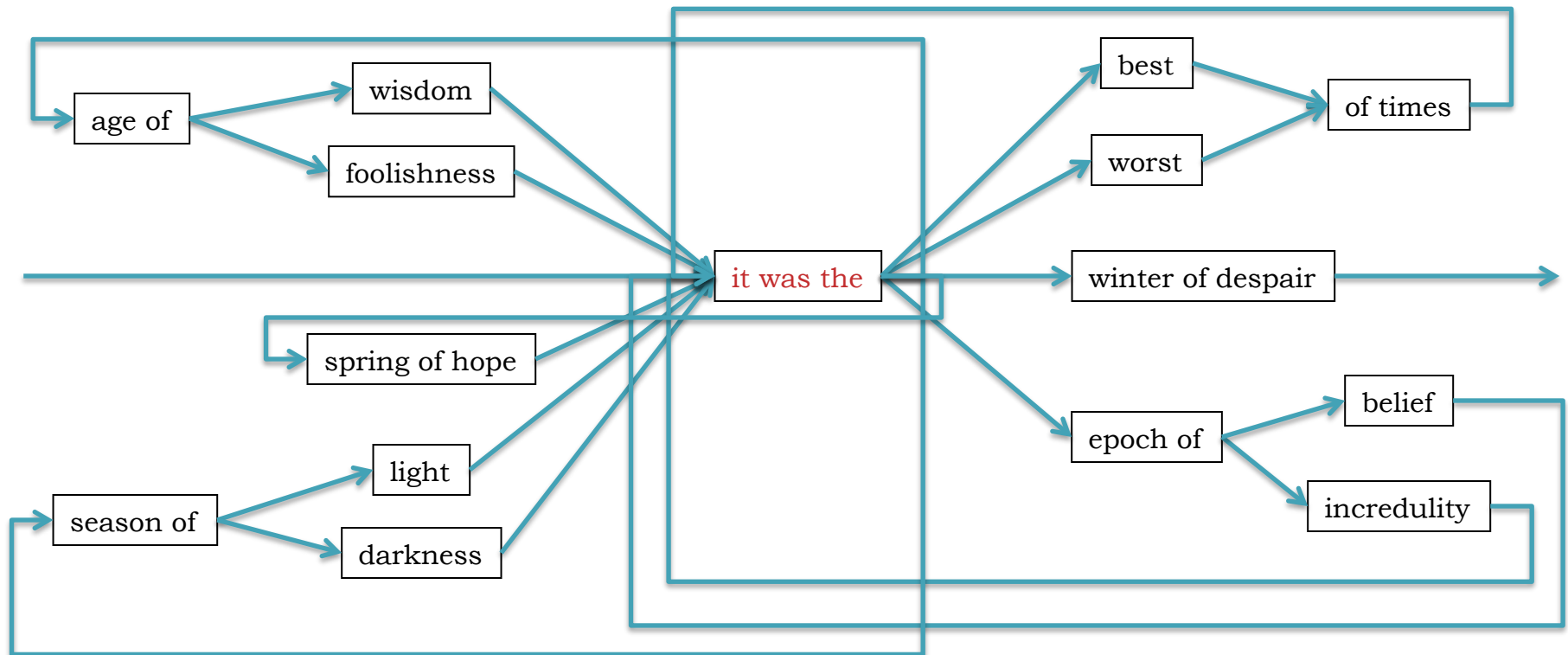
Introductions

Tyler Garvin

CNV and transcriptome analysis of single cells

Giuseppe Narzisi

Indels related to Autism and other human diseases

Maria Nattestad

Hi-C Chromatin Interactions within plants

Hayan Lee

Modeling of genome assembly performance
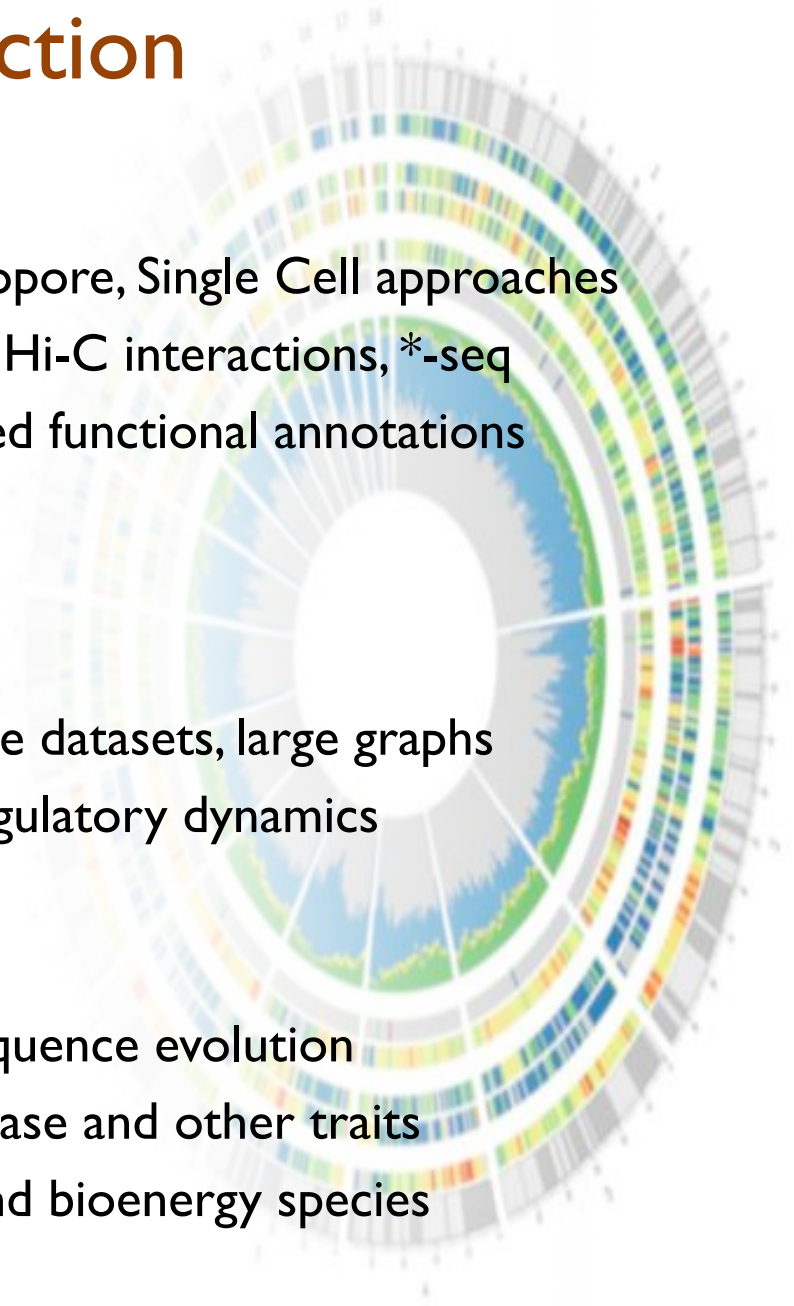
Understanding Genome Structure & Function

Biotechnology

- Sequencing: Illumina, PacBio, Oxford Nanopore, Single Cell approaches
- Biochemical assays: RNA-seq, Methyl-seq, Hi-C interactions, *-seq
- More accurate sequencing & more detailed functional annotations

Algorithmics

- Highly scalable algorithms and systems
- Indexing and analyzing very large sequence datasets, large graphs
- Constructing Pan-genomes & inferring regulatory dynamics

Comparative Genomics

- Cross species comparisons, models of sequence evolution
- Identifying mutations associated with disease and other traits
- Genotype-to-phenotype of agricultural and bioenergy species

Acknowledgements

Biological Data Sciences
Anne Carpenter, Michael Schatz, Matt Wood
Nov 5 - 8, 2014

Thank you

http://schatzlab.cshl.edu
@mike_schatz