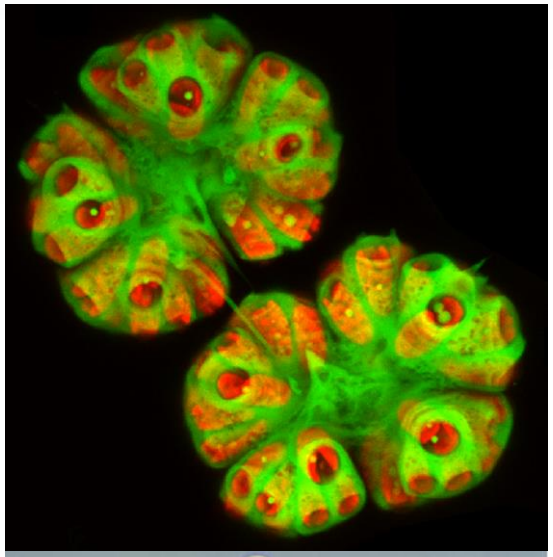
Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

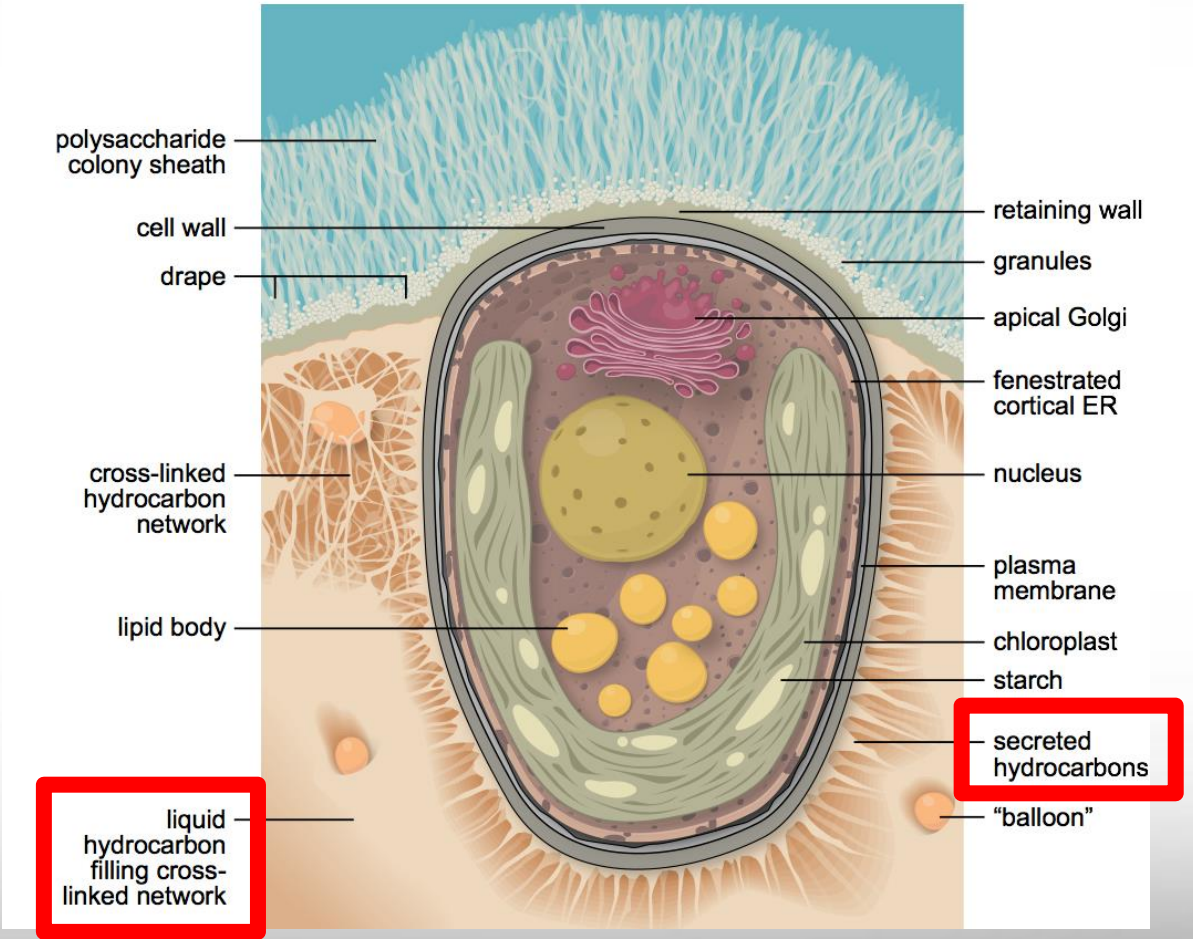
Project by Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics
Presented by Michael Dickens, High Performance Research Computing
Texas A&M University

Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

Basic model of *Botryococcus braunii* cell biology





Weiss *et al* (2012) *Eukaryotic Cell* **11**:1424-1440

Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

Why sequence the *B. braunii* genome?

JGI
JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

- ***B. braunii* is a potential source of renewable fuels and chemicals**
- *B. braunii* is found worldwide, most notably in oil and coal shale deposits
- ***B. braunii* has a very high oil content, ~40% of dry weight**
- *B. braunii* oils can be processed with conventional petroleum technology

Main project organizers:

Andy Koppisch
Northern Arizona University

Joe Chappell
University of Kentucky

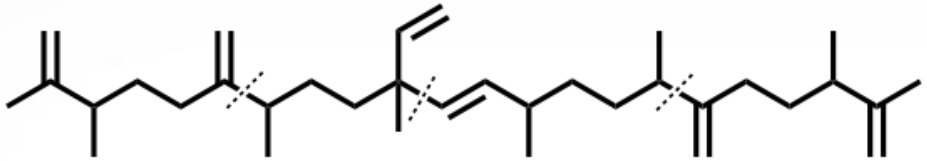
Tim Devarenne
Texas A&M University

Shigeru Okada
Tokyo University

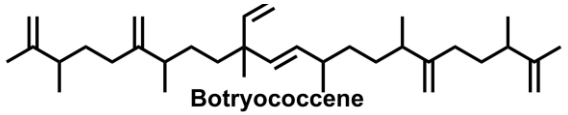
Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

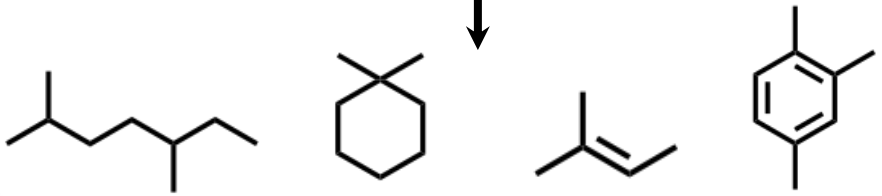
High rates of drop-in biofuel recovery



30-40% of *B. braunii* dry weight



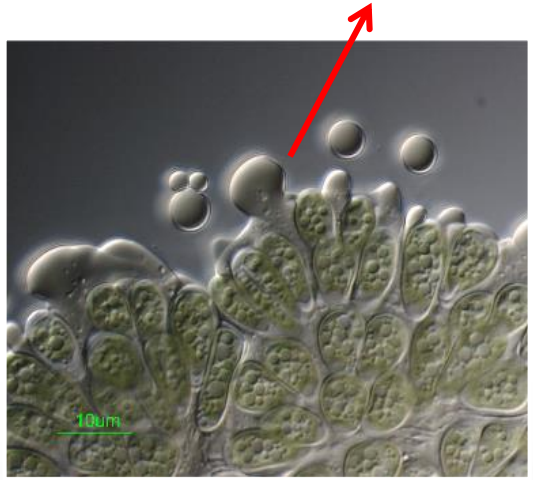
Hydrocracking
(497˚C/high pressure/catalyst)



Paraffins	Naphthenes	Olefins	Aromatics
68.5%	30.0%	<0.2%	1.4%

Distillation

Gasoline	Kerosene	Diesel	Residuals
C_5-C_{12}	C_{10}-C_{16}	C_{14}-C_{20}	>C_{70}
40-205˚C	175-325˚C	250-350˚C	>600˚C
67%	15%	15%	3%



Liquid hydrocarbons are easily recovered from colony

60 - 70% of crude *B. braunii* hydrocarbons converted to gasoline

Comparable to petroleum

Hillen et al. (1982) *Biotechnol Bioeng* **24**:193

Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

B. braunii whole-genome sequencing with Illumina

Library Name	Library Type	Insert Size	Total Sequence Reads	Read Length	Genome Size	Coverage
SXPX	Paired End	800 bp	499,073,402	250 bp	166 Mb	~750x

Genome sequence can be used to identify genes involved in hydrocarbon production

Genomic DNA

AATAATGTCAATTTGGTAGATATCAGAGAGTTTTATGTTGACAAAGATGG

AATAATGTCA	GATATCAGAGA	ATGTTGACAAA
AATAATGTCAA	GATATCAGAGAG	ATGTTGACAAA
ATAATGTCAAT	TATCAGAGAGT	GTTGACAAAG
ATAATGTCAAT	TATCAGAGAGT	GTTGACAAAG
TAATGTCAATT	TCAGAGAGT	TTGACAAAGAT
TGTCAATTTGG	CAGAGAGT	TTGACAAAGAT
TGTCAATTTGGT	CAGAGAGT	TGACAAAGATG
AATTTGGTAGAT	GAGAGT	GACAAAGATGG
TTGGTAGATAT		CAAAGATGG
TGGTAGATATC		AAAGATGG

Computational Assembly

AATAATGTCAATTTGGTAGATATCAGAGAGTNNNATGTTGACAAAGATGG

Reconstructed DNA Sequence

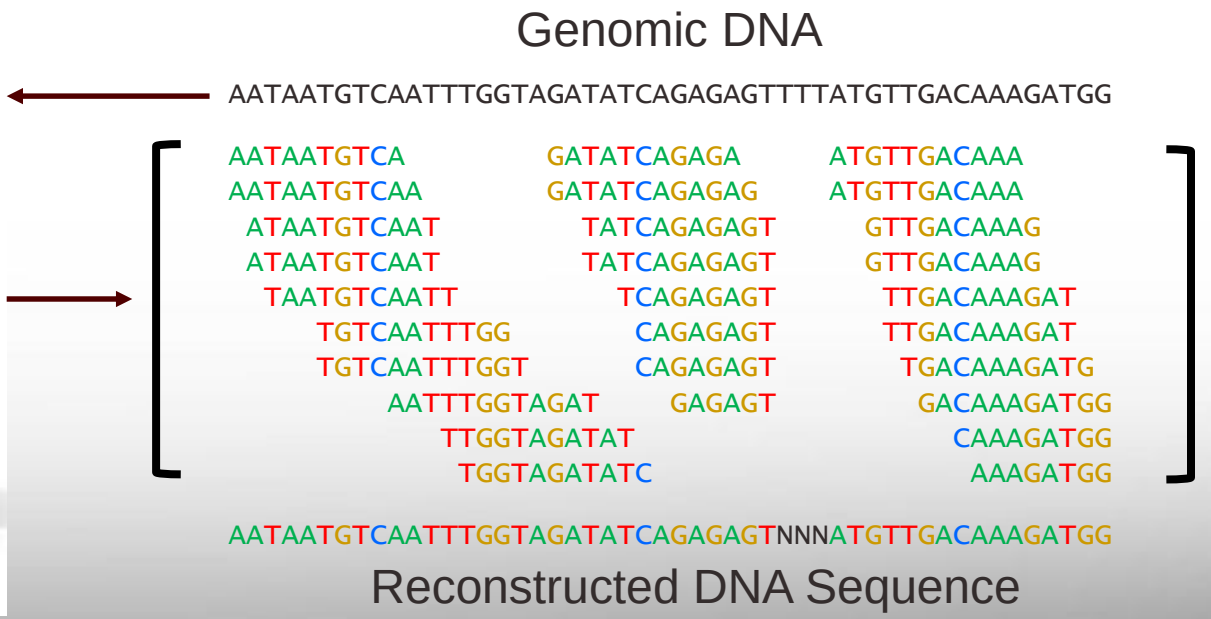
Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

B. braunii whole-genome sequencing with Illumina

Library Name	Library Type	Insert Size	Total Sequence Reads	Read Length	Genome Size	Coverage
SXPX	Paired End	800 bp	499,073,402	250 bp	166 Mb	~750x

Genome sequence can be used to identify genes involved in hydrocarbon production



Workflow of Assembly By Short Sequences (ABySS): A parallel *de novo* genome assembler with MPI support

(1) ABYSS-P

Read (l = 12):
ATCATACATGAT

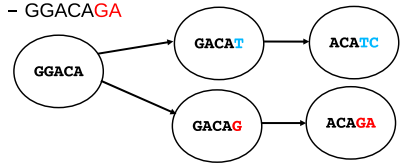
k-mers (k = 9):
ATCATACAT
TCATACATG
CATACATGA
ATACATGAT

k-mer

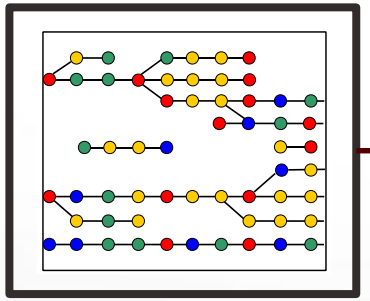
De Bruijn graph

- A simple graph for k = 5
- Two reads
  - GGACATC
  - GGACAGA

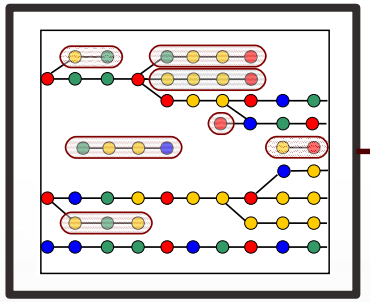
MPI



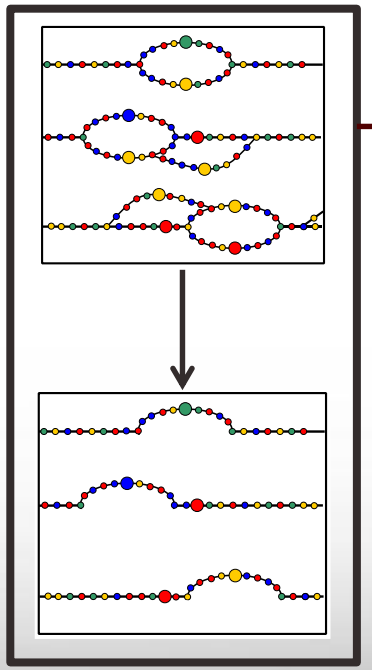
(2) AdjList



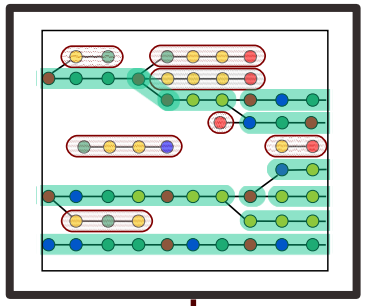
(3) Prune tips



(4) Pop bubbles



(5) Generate contigs



map 1

map 2

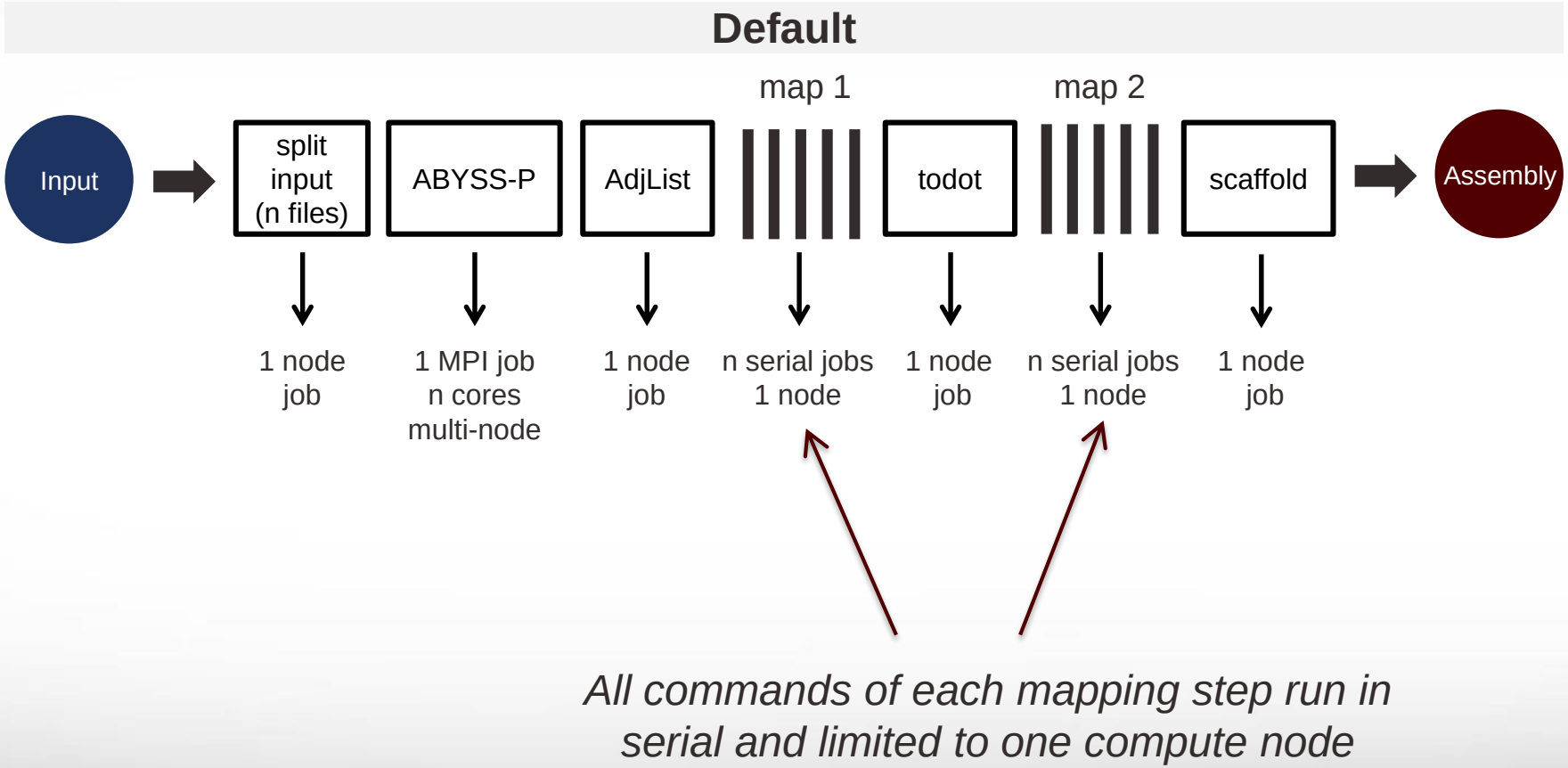
scaffold

http://www.bcgsc.ca/platform/bioinfo/software/abyss
Slide Material From: Shaun D. Jackman (http://sjackman.github.io/)

Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

Default and modified ABySS execution pipelines

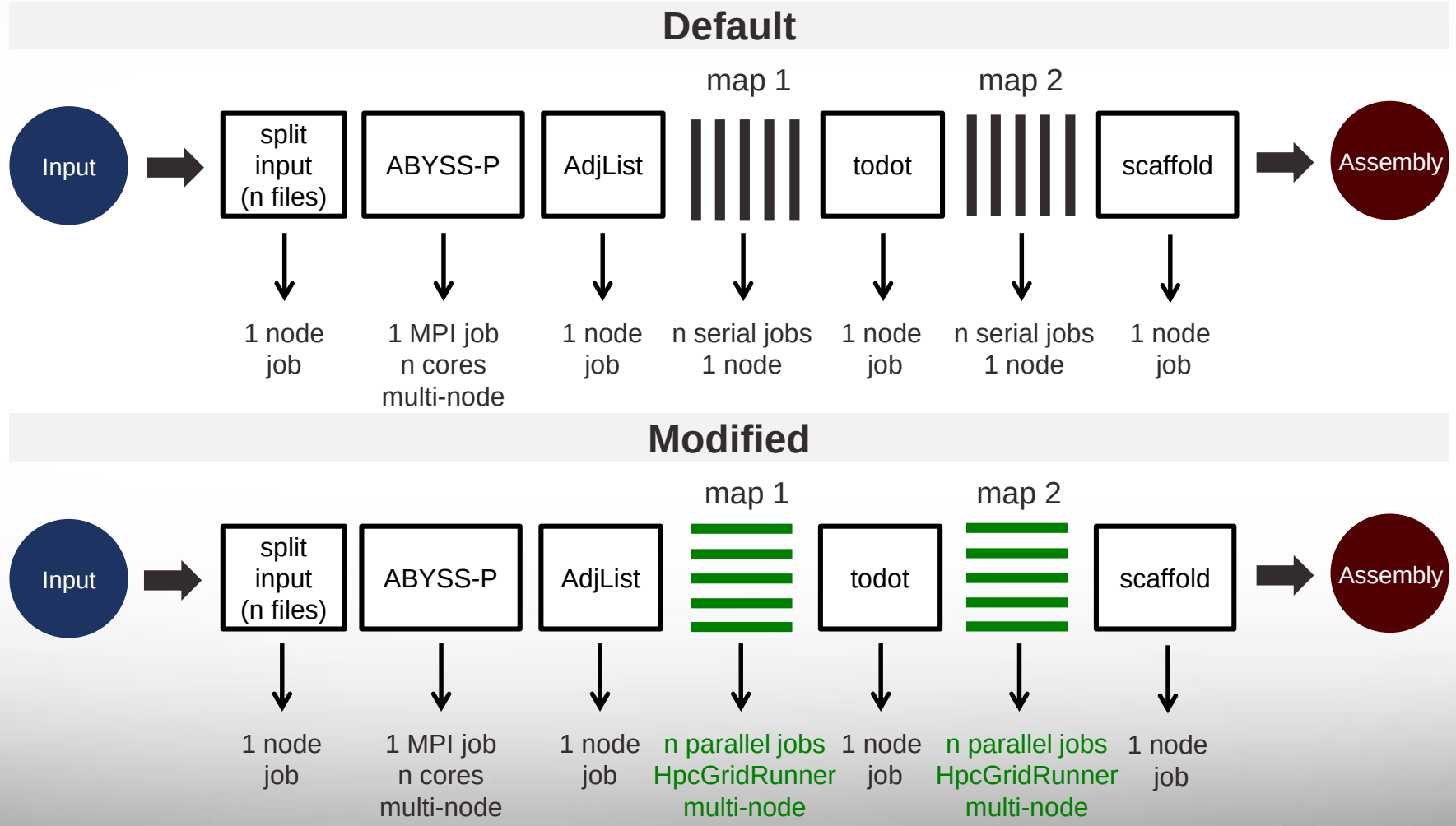


All commands of each mapping step run in serial and limited to one compute node

Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

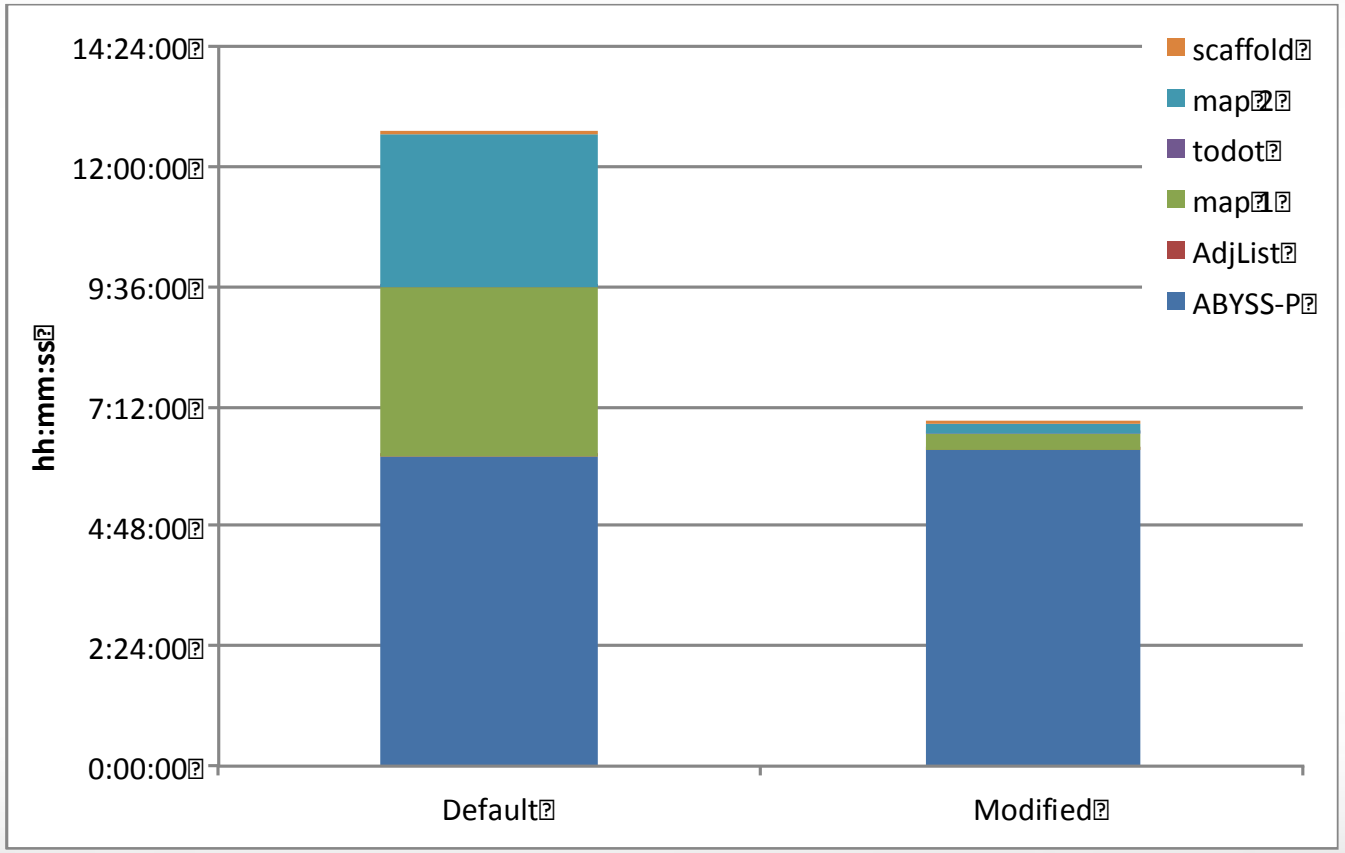
Default and modified ABySS execution pipelines



Improving HPC resource utilization in the genome assembly of a biofuel producing green algae

Dan Browne, PhD Candidate, Devarenne Lab, Biochemistry & Biophysics Department, Texas A&M University

Assembly times of default and modified pipelines



- HPC resource utilization: 50 cores (5 cores/node * 10 nodes)
- Assembly time reduced by 46% using modified ABySS pipeline.
- Modified pipeline eliminated 45 cores being idle for almost 6 hours.

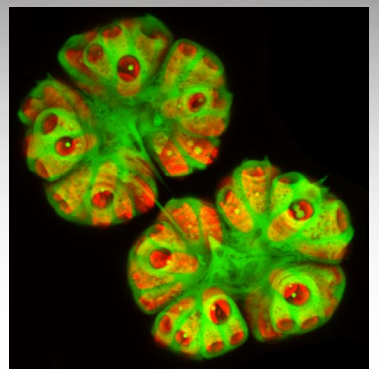
Devarenne Lab 2015

http://devarennelab.tamu.edu

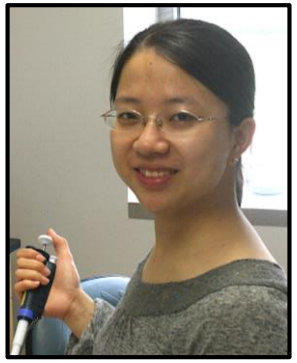
Department of Biochemistry & Biophysics

Tim Devarenne, PhD
Associate Professor

Hem Thapa
Grad Student



Botryococcus braunii



Dongyin Su
Grad Student

Mehmet Tatli
Grad Student

Dan Browne
Grad Student

Incheol Yeo
Grad Student

Victoria Yell
Undergrad Student