The Impact of HPC Best Practice Applied to Next-Generation Sequencing Workflows

- Pierre Carrier, Bill Long, Richard Walsh, Jef Dawson CRAY THE SUPERCOMPUTER COMPANY
- Carlos P. Sosa UNIVERSITY OF MINNESOTA ROCHESTER Driven to Discover CRAY THE SUPERCOMPUTER COMPANY
- Brian Haas, Timothy Tickle BROAD INSTITUTE
- Thomas William TECHNISCHE UNIVERSITÄT DRESDEN

RNA-Seq software:

Trinity
Inchworm
Chrysalis
Butterfly

Agenda

- What have we **learned in computational chemistry**?
- A tale of **two communities**
- Why **HPC Best Practice**?

bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

http://biorxiv.org/content/early/2015/04/07/017665

Agenda

- What have we **learned in computational chemistry**?
- A tale of **two communities**
- Why **HPC Best Practice**?
- What are **NGS workflows**?
- What is **Trinity RNA-Seq**?

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

http://biorxiv.org/content/early/2015/04/07/017665

Agenda

- What have we **learned in computational chemistry**?
- A tale of **two communities**
- Why **HPC Best Practice**?
- What are **NGS workflows**?
- What is **Trinity RNA-Seq**?
- **HPC Best Practice in a nutshell**
- **MPI-Inchworm** algorithm

bioRχiv beta

THE PREPRINT SERVER FOR BIOLOGY

http://biorxiv.org/content/early/2015/04/07/017665

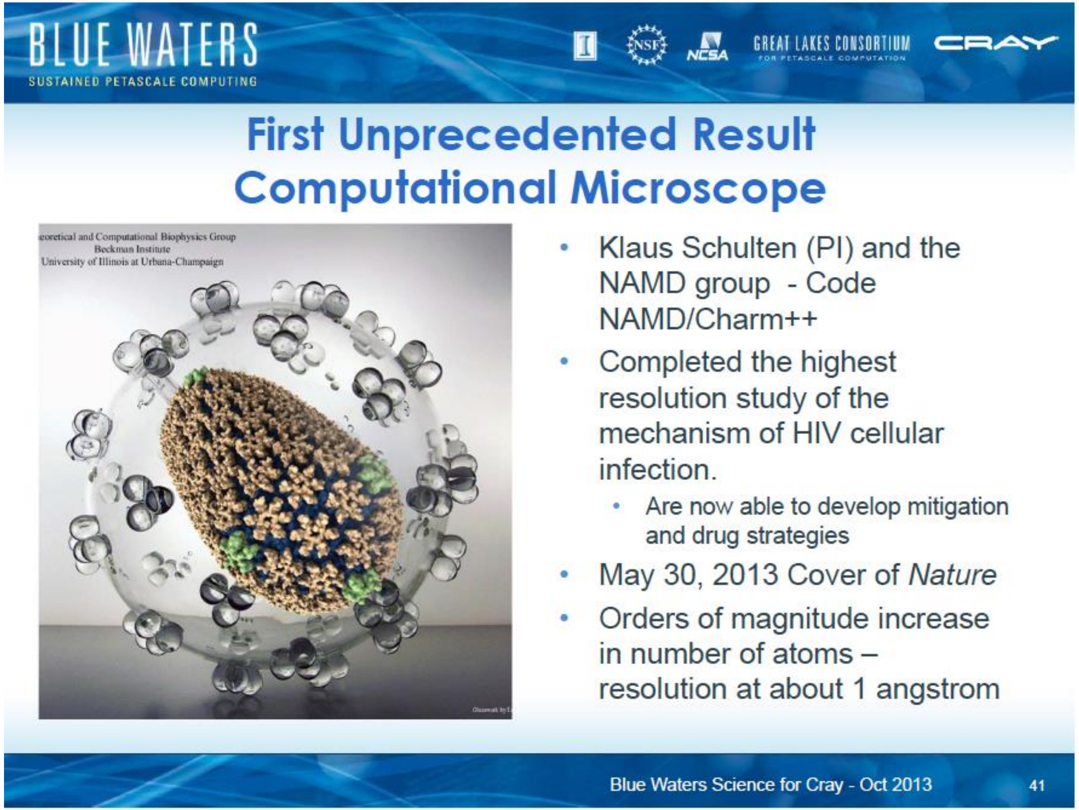
Agenda

- What have we **learned in computational chemistry**?
- A tale of **two communities**
- Why **HPC Best Practice**?
- What are **NGS workflows**?
- What is **Trinity RNA-Seq**?
- **HPC Best Practice in a nutshell**
- **MPI-Inchworm** algorithm
- Hardware resources
- **Scaling** memory and CPU
- Conclusions

bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY

http://biorxiv.org/content/early/2015/04/07/017665

Computational Chemistry Embraces HPC



BLUE WATERS
SUSTAINED PETASCALE COMPUTING

NSF
NCSA
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION
CRAY
eoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

First Unprecedented Result
Computational Microscope

- Klaus Schulten (PI) and the NAMD group - Code NAMD/Charm++
- Completed the highest resolution study of the mechanism of HIV cellular infection.
  - Are now able to develop mitigation and drug strategies
- May 30, 2013 Cover of *Nature*
- Orders of magnitude increase in number of atoms – resolution at about 1 angstrom

Blue Waters Science for Cray - Oct 2013 41

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Computational Genomics

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Example: Hartree-Fock (HF) and Density Functional Theory (DFT).
The initial problem, HF, is essentially very large to solve, often too large.
DFT, was created (in part) in order to more easily adapt the problem size and accuracy (xc functionals) to the available supercomputer

Computational Genomics

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

A Tale of Two Communities

Computational Chemistry

Computational Genomics

Method implementation has evolved together with HPC

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Computational Genomics

Method implementation has evolved together with <u>sequencing technologies</u> (NGS), not with HPC

<u>2 main "next-generation" sequencing technologies:</u>

Short reads

illumina

A <u>"Reads"</u> = a "puzzle piece" of nucleotides (the "ATGC"s)

Long reads

PACIFIC BIOSCIENCES

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Computational Genomics

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

2 main "next-generation" sequencing technologies:

A "Reads" = a "puzzle piece" of nucleotides (the "ATGC"s)

Short reads

illumina

Short reads: ~**100s** nucleotides

Long reads: ~**1000s** nucleotides

Long reads

PACIFIC BIOSCIENCES

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Computational Genomics

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

Emerging technologies:

Short reads

illumina

Short reads: ~**100s** nucleotides

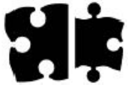
A **“Reads” = a “puzzle piece” of nucleotides (the “ATGC”s)**

Long reads: ~**1000s** nucleotides

Long reads

PACIFIC BIOSCIENCES

…Even longer reads



NANOPORE

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Computational Genomics

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

sequencing

Short reads

illumina

- ✓ Smallest genomes
- ✓ Bacteria
- ✓ Drosophila
- ✓ Mouse
- ✓ **Human**
- ✓ …
- ✓ Spruce
- X Axolotl (not done)
- X Largest genome

SIZE

Software

- ✓ Develop/compute on laptop
- ✓ On laptop
- ✓ On laptop
- ✓ On workstation
- ✓ **On workstation**
- ✓ …
- ✓ On supercomputer?

A Tale of Two Communities

Computational Chemistry

Method implementation has evolved together with HPC

Computational Genomics

Method implementation has evolved together with <u>sequencing technologies</u> (NGS), not with HPC

<u>sequencing</u>

Short reads

illumina

- ✓ <u>Smallest genomes</u>
- ✓ Bacteria
- ✓ Drosophila
- ✓ Mouse
- ✓ **Human**
- ✓ …
- ✓ Spruce
- X Axolotl (not done)
- X <u>Largest genome</u>

software

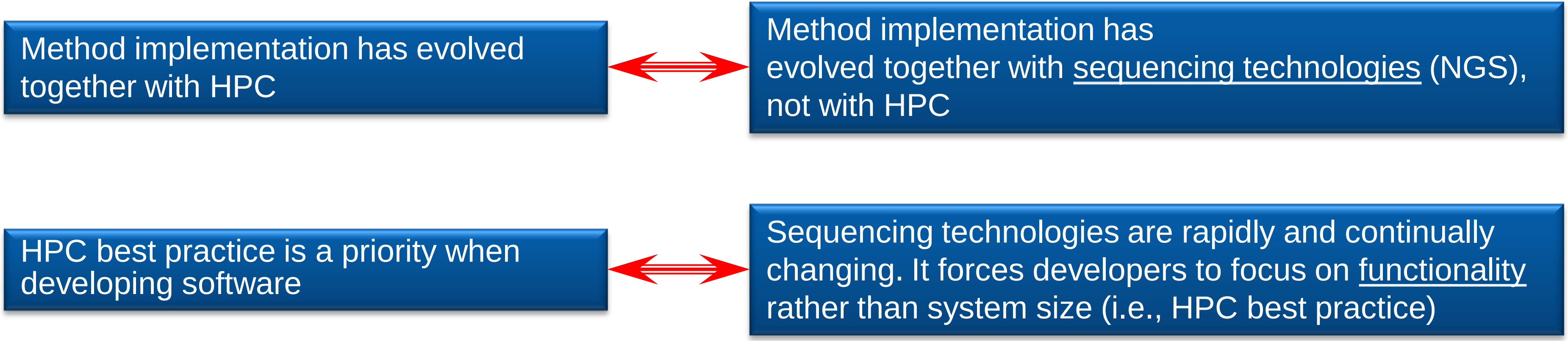
Long reads

PACIFIC BIOSCIENCES

- ✓ **<u>Small genome</u>**
- ✓ Bacteria
- ✓ Drosophila
- X Fungus
- X Mouse
- X **Human**
- X …
- X Axolotl (not done)
- X Largest genome

SIZE

A Tale of Two Communities



Computational Chemistry

Method implementation has evolved together with HPC

HPC best practice is a priority when developing software

Computational Genomics

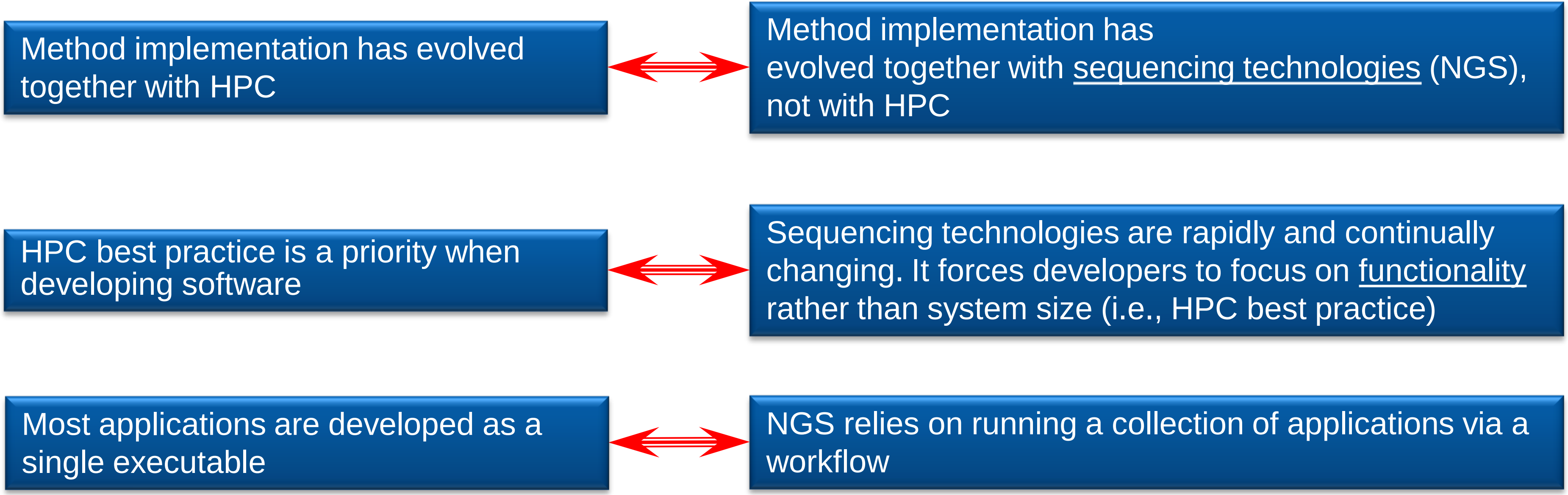
Method implementation has evolved together with sequencing technologies (NGS), not with HPC

Sequencing technologies are rapidly and continually changing. It forces developers to focus on functionality rather than system size (i.e., HPC best practice)

A Tale of Two Communities

Computational Chemistry

Computational Genomics



Method implementation has evolved together with HPC

Method implementation has evolved together with sequencing technologies (NGS), not with HPC

HPC best practice is a priority when developing software

Sequencing technologies are rapidly and continually changing. It forces developers to focus on functionality rather than system size (i.e., HPC best practice)

Most applications are developed as a single executable

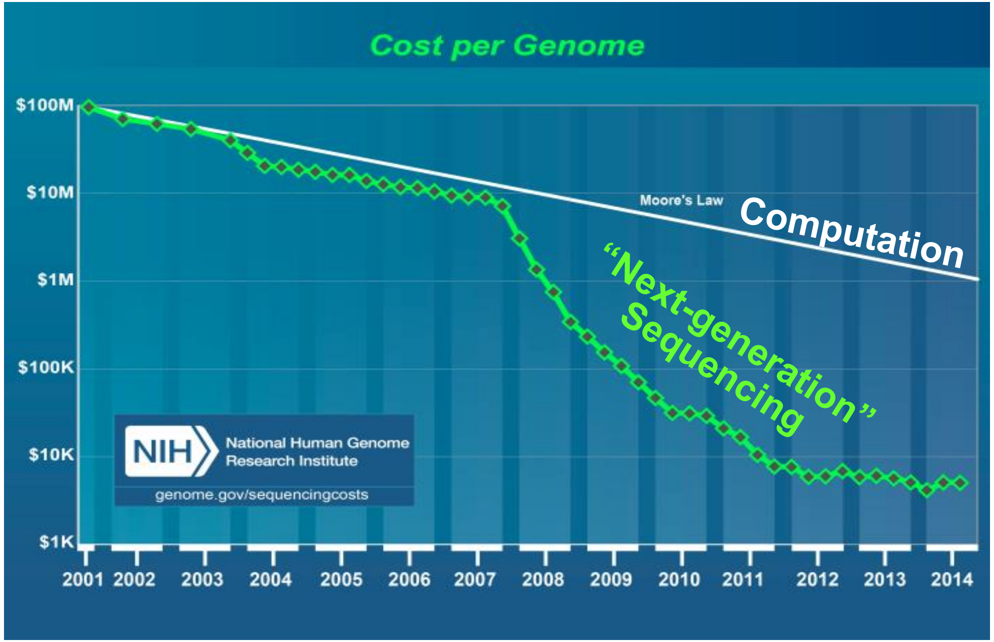
NGS relies on running a collection of applications via a workflow

NGS Bottlenecks are Now Computational

Experiment
(cost)

Bottleneck Shift

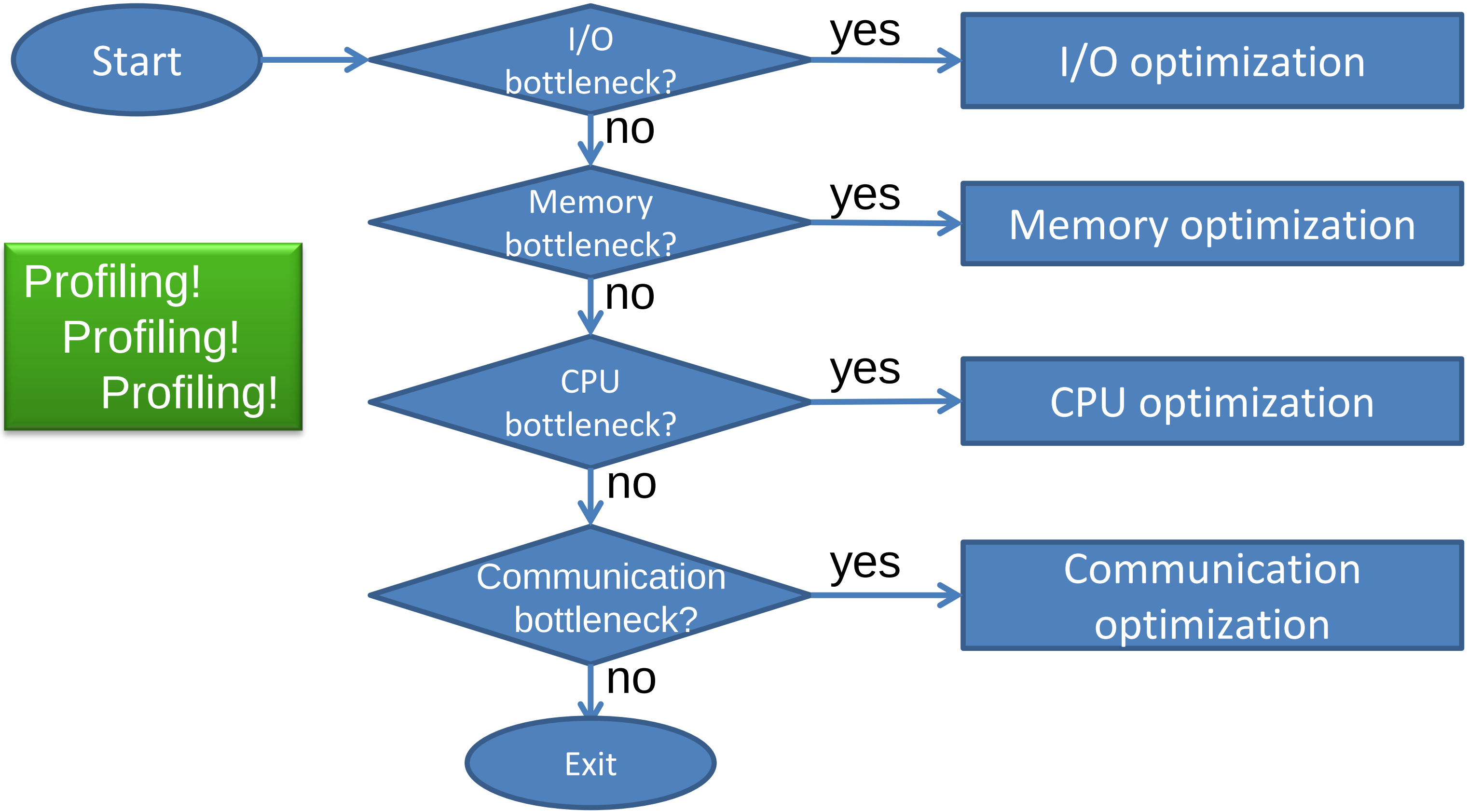
Computation
(time-to-solution)



Impact

- Algorithm Innovation
- Enhanced Clinical Applications
- Enhanced NGS Research

HPC Best Practice



Start

I/O bottleneck?

yes

I/O optimization

no

Memory bottleneck?

yes

Memory optimization

no

CPU bottleneck?

yes

CPU optimization

no

Communication bottleneck?

yes

Communication optimization

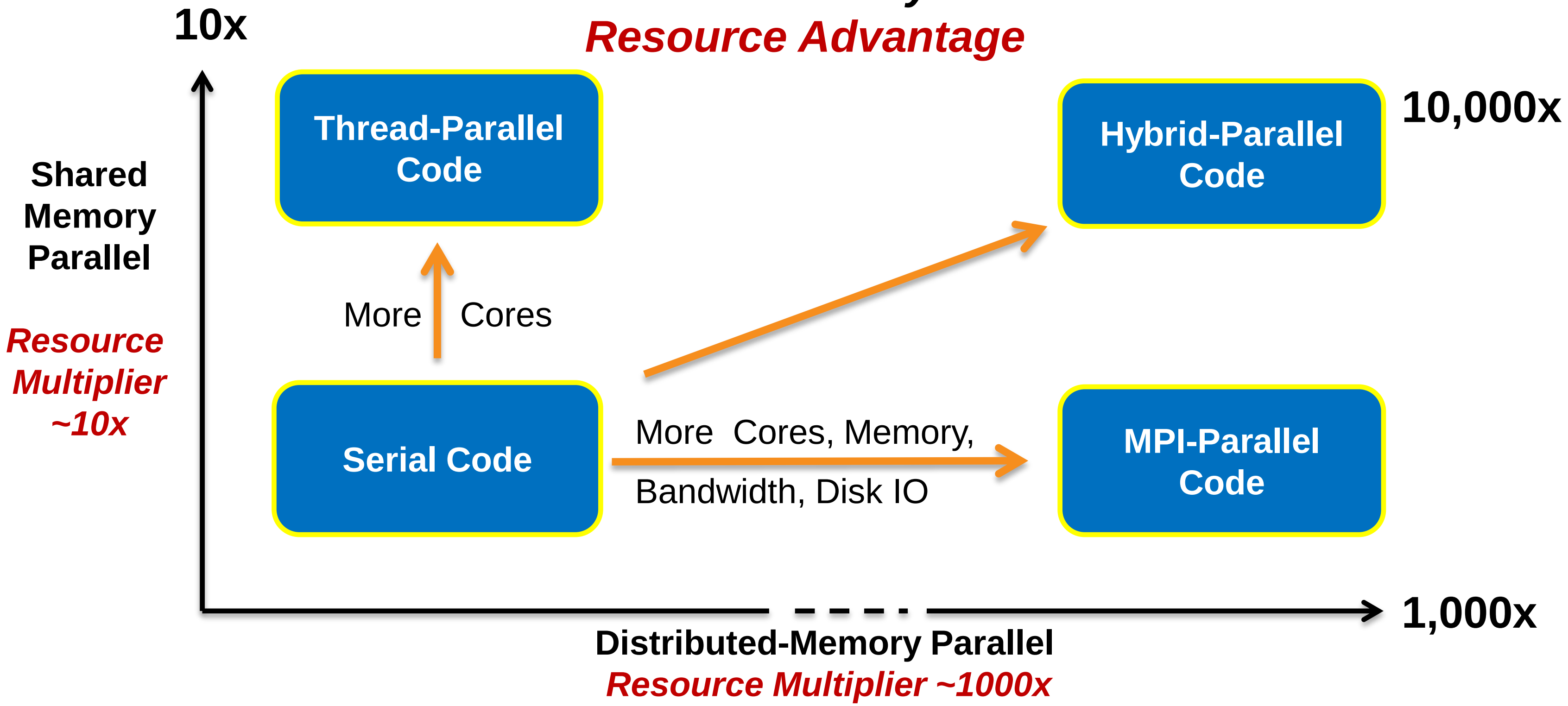
no

Exit

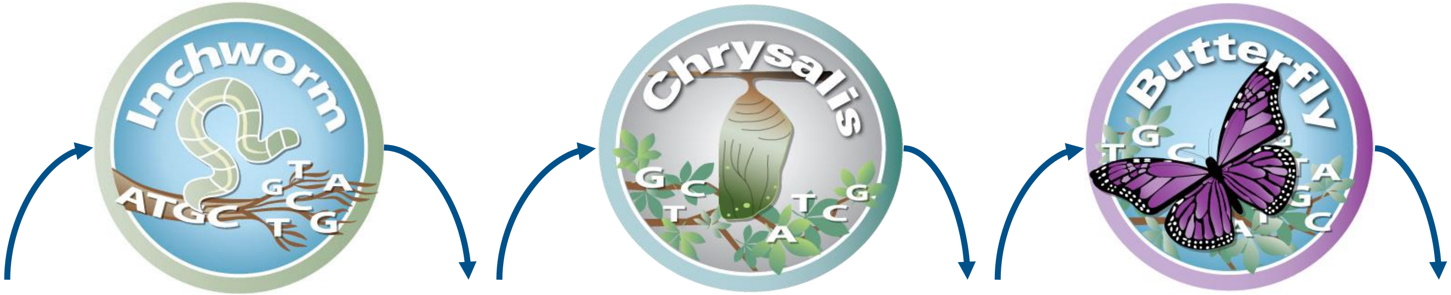
Profiling!
Profiling!
Profiling!

Why HPC Best Practice?

The Distributed-Memory Parallelism
Resource Advantage



TrinityRNA-Seq – How it works:

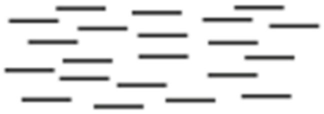


RNA-Seq “reads”

Linear “contigs”

de-Bruijn graphs

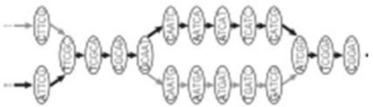
Transcripts + Isoforms



Short reads

illumina

>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66



Thousands of disjoint graphs

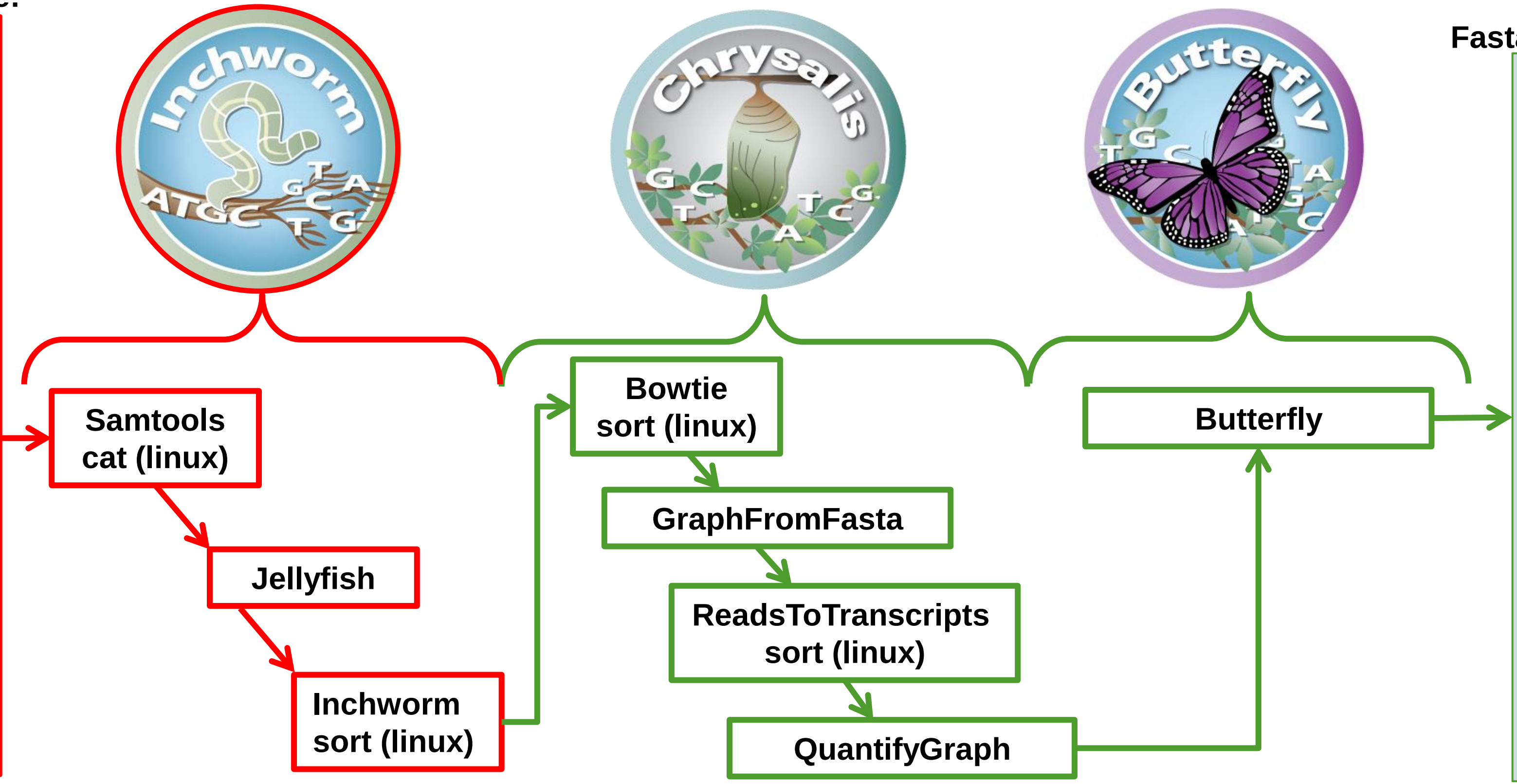
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

https://github.com/trinityrnaseq/trinityrnaseq/releases/tag/v2.0.6

TrinityRNA-Seq – How it works:

Fastq file:

```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACATTTTCTCAGTATTCCATCTAGCTGC
+
C<CCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCCCCCCCCCCCCCCCCACCCCCACCCC=
@61DFRAAXX100204:1:100:10497:13422/2
GAGTTACTGGTAAGACGCTTACACCTATAACTCAAGGTCGGAATAGTCCCTCCAGTCCCTTTAGTAACCCAGTGGC
…
```



Samtools
cat (linux)

Jellyfish

Inchworm
sort (linux)

Bowtie
sort (linux)

GraphFromFasta

ReadsToTranscripts
sort (linux)

QuantifyGraph

Butterfly

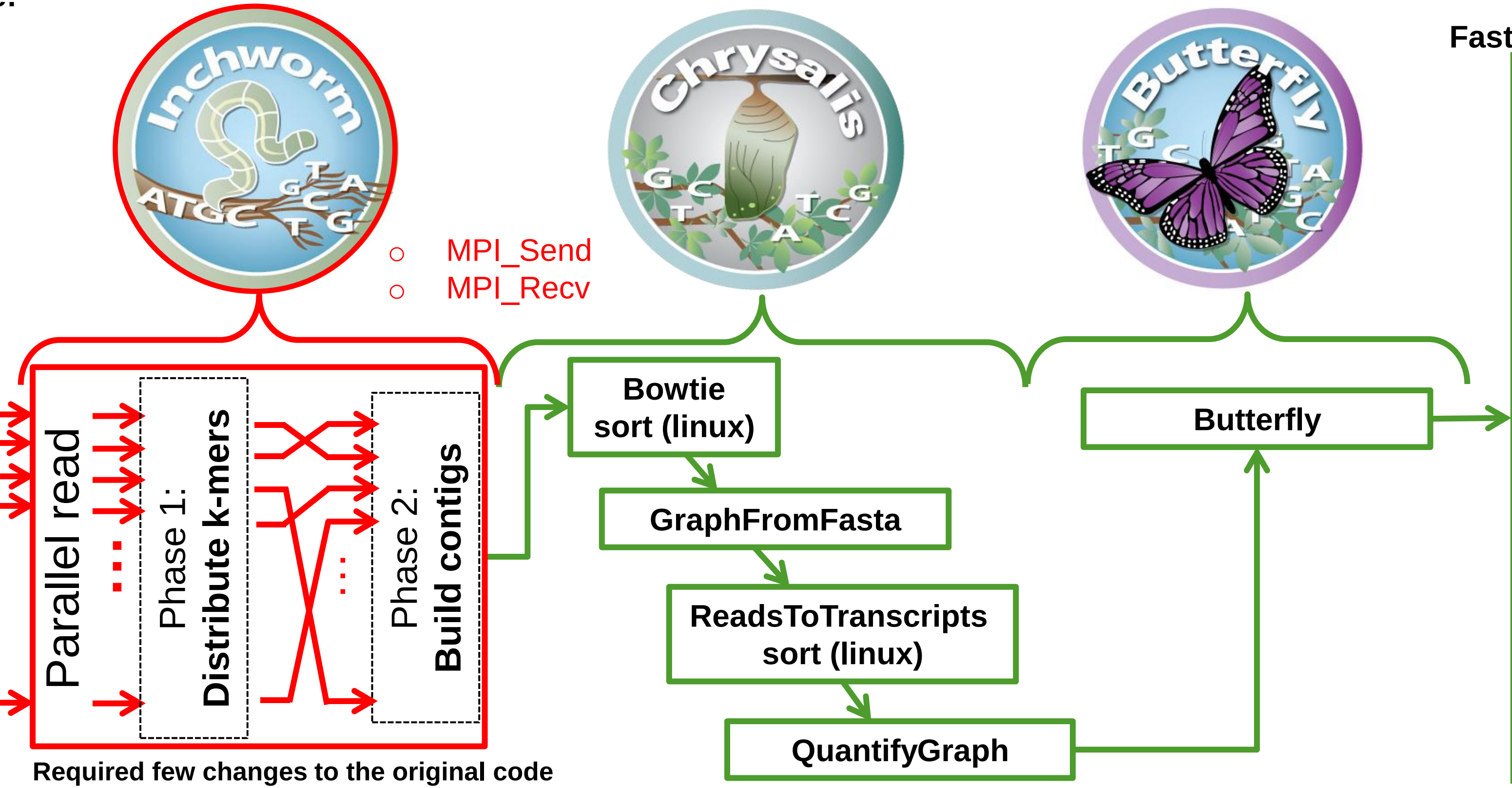
Fasta file:

```
>comp0_c0_seq1 len=240 path=[1:0-239]
GGACTTGTCTTCAAGTCCATTTTCTTTCCTCGGGGAATCTCCCCTTCCCGACACCCACAC
CGAACTCCGGGAGCCAGGAAGTACACTGCTTGCAAGACGCCTTTGCAGCCTGCTCCCTCC
TCCTGGCACCCCACCTGCATCTGCAGACTATCCCCAGGCTTGGCCCTGAGCCCCAGCTAT
CCCAAACGTGGTCTCCCTTCTCTCTCTCTCAGAGACCGTGCGGGCCATGACCCA
>comp7_c0_seq1 len=215 path=[1394:0-214]
…
```

MPI-Inchworm Algorithm

Fastq file:

```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACATTTTCTCAGTATTCCATCTAGCTGC
+
C<CCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCCCCCCCCCCCCCCCACCCCCACCC=
@61DFRAAXX100204:1:100:10497:13422/2
GAGTTACTGGTAAGACGCTTACACCTATAACTCAAGGTCGGAATAGTCCCTCCAGTCCCTTTAGTAACCCAGTGGC
...
```



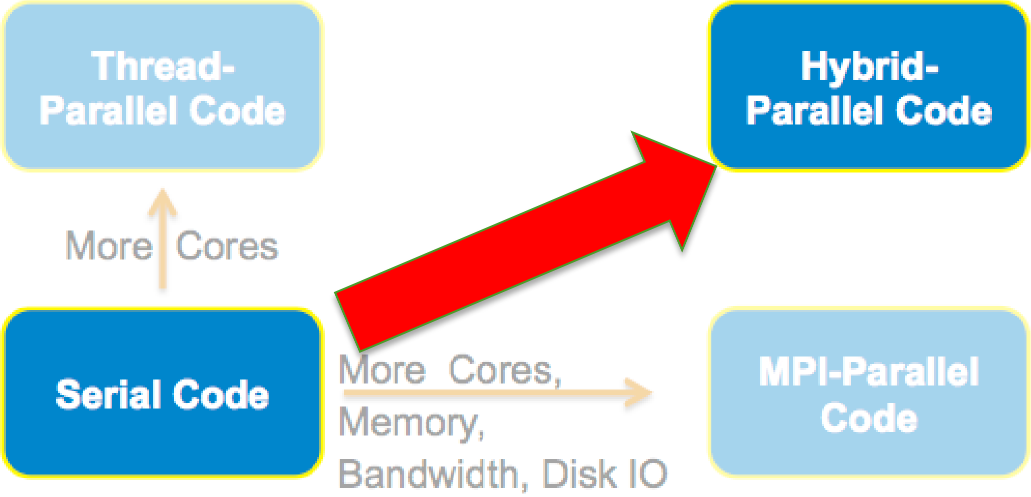
- ○ MPI_Send
- ○ MPI_Recv

Required few changes to the original code

Fasta file:

```
>comp0_c0_seq1 len=240 path=[1:0-239]
GGACTTGTCTTCAAGTCCATTTTCTTTCCTCGGGGAATCTCCCCTTCCCGACACCCACAC
CGAACTCCGGGAGCCAGGAAGTACACTGCTTGCAAGACGCCTTTGCAGCCTGCTCCCTCC
TCCTGGCACCCCACCTGCATCTGCAGACTATCCCCAGGCTTGGCCCTGAGCCCCAGCTAT
CCCAAACGTGGTCTCCCTTCTCTCTTTCTCTCTCTCAGAGACCGTGCGGGCCATGACCCA
>comp7_c0_seq1 len=215 path=[1394:0-214]
...
```

MPI-Inchworm Algorithm



- Phase 1: Distribute k-mers
- Phase 2: Build contigs

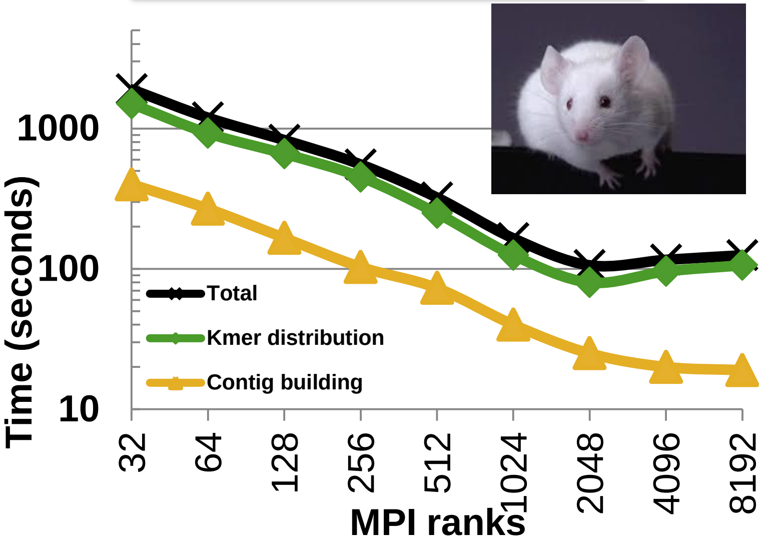
Hardware Resources

- Cray XC40, 64-bit Intel® Xeon® E5-2698 V3 "**Haswell**"
- **16 core** 2.3 GHz processor
- **Two processors** per compute node and **384 processors per cabinet**
- The processor **peak performance** per core is 36.8 GF
- The memory consists of **128 GB** DDR4-**2133 MHz** per compute node
- **Memory bandwidth is 120 GB/s per node**
- The system interconnect is **Cray Aries multilevel dragonfly topology**

Essentially, 2 cabinets XC40 (2x192 nodes) is ideal for running MPI-inchworm (one run after another).

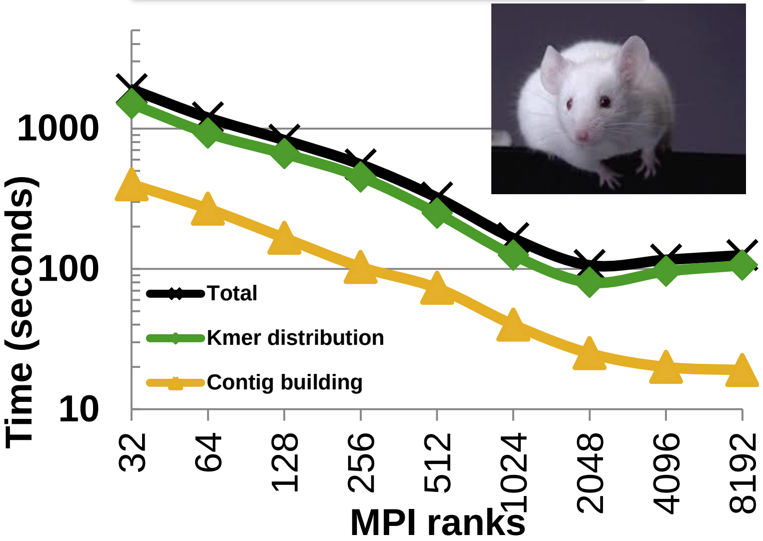
Compute and Memory Scalability

Compute Scalability

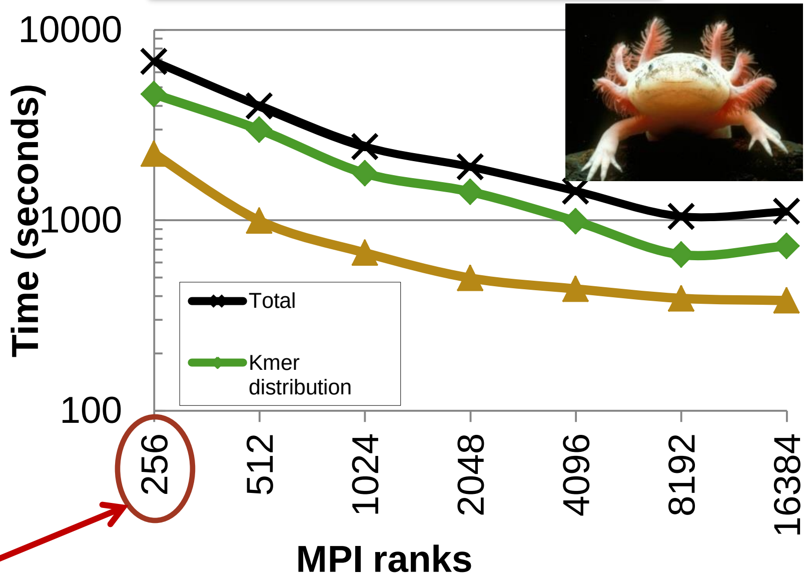


Compute and Memory Scalability

Compute Scalability



Memory Scalability

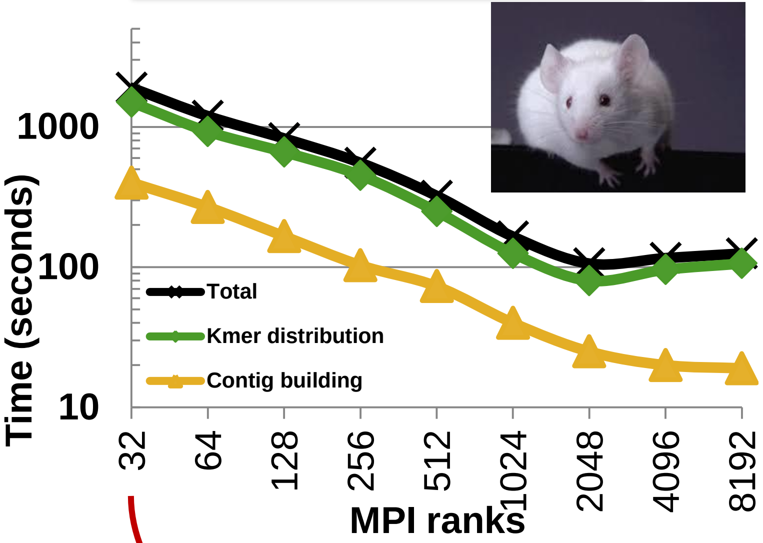


8 nodes minimum

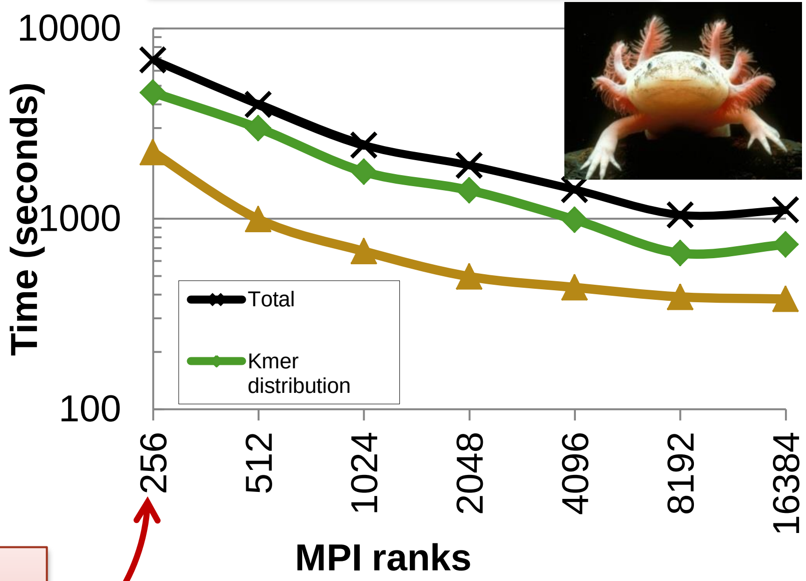
Salamander (~30GB) computation is impossible to perform using the serial code, on 64 or 128 GB nodes; requires using special memory node

Compute and Memory Scalability

Compute Scalability



Memory Scalability



Distributed memory allows research on problems that otherwise would not be feasible

Salamander (~30GB) computation is impossible to perform using the serial code, on 64 or 128 GB nodes; requires using special memory node

Conclusions

- **HPC Best Practice** can be applied to the **parallelization of NGS workflows.**
- The **distributed MPI-Inchworm** can now **utilize 4096 and more cores.**
- **Any bioinformatics** workflow can greatly **benefit from HPC.**
- Distributed-memory parallelism **eliminates** the need for **hybrid configurations** with large shared-memory nodes.

Acknowledgements

- We thank **Cray Inc**. for the computing time on the **XC30/XC40 marketing machine.**
- We thank **Broad Institute principal investigator Aviv Regev** for generously supporting Trinity development efforts.
- We thank **Jessica Whited** at the **Brigham Regenerative Medicine Center** for access to the axolotl RNA-Seq data.
- Research reported in this publication was supported by the **National Cancer Institute of the National Institutes of Health (NIH) under Award Number 1U24CA180922-01.**
- The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Q&A

CODE: https://github.com/trinityrnaseq/trinityrnaseq/releases/tag/v2.0.6

CUG paper: http://biorxiv.org/content/early/2015/04/07/017665

Pierre Carrier: pcarrier@cray.com