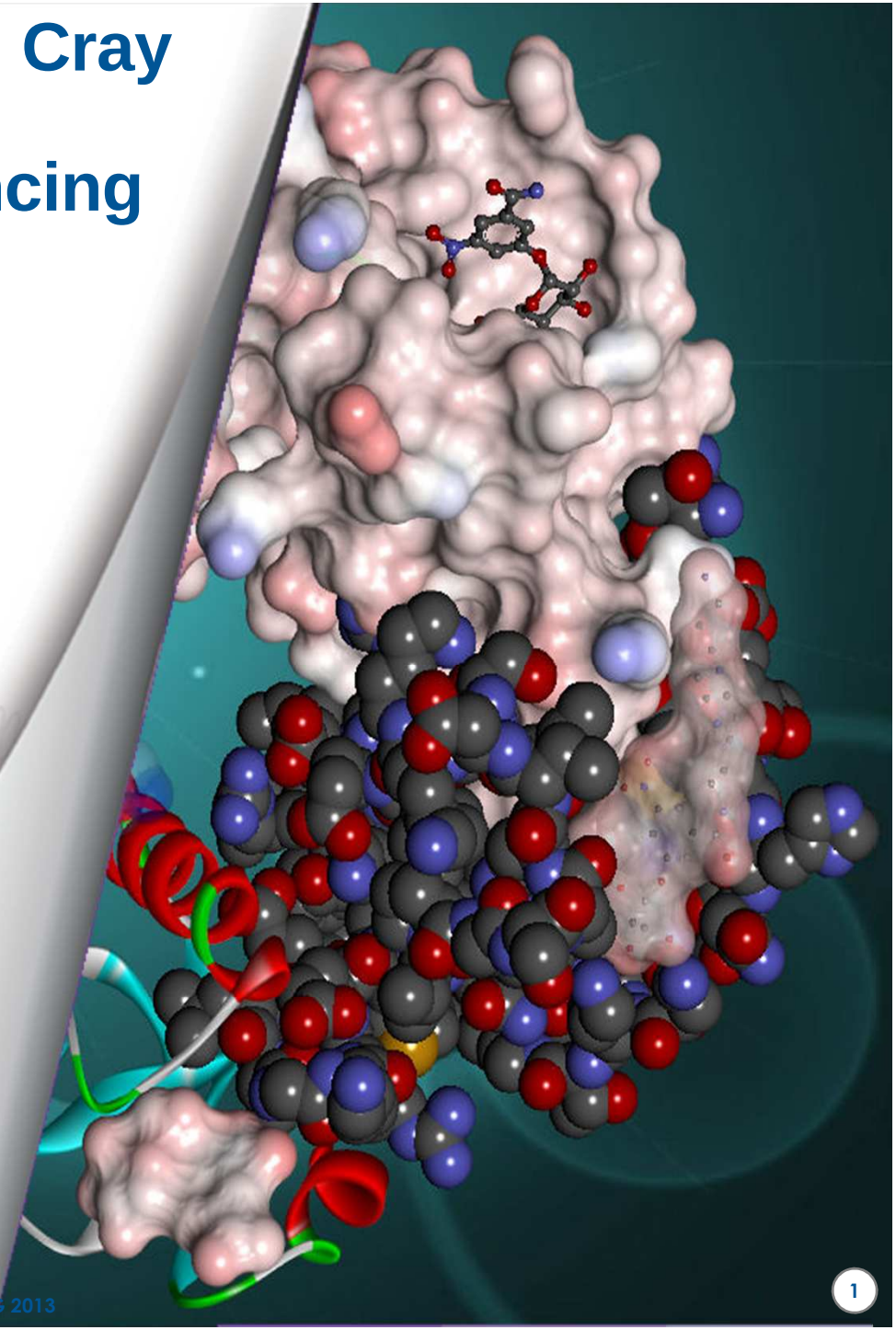
Genomic Applications on Cray supercomputers: Next Generation Sequencing Workflow

Barry Bolding

Cray Inc
Seattle, WA



THE SUPERCOMPUTER COMPANY

CUG 2013 Paper

Genomic Applications on Cray supercomputers: Next Generation Sequencing Workflow

Mikhail Kandel
Department of Computer Science
University of Illinois
Urbana-Champagne, IL
USA
kandel3@illinois.edu
Steve Behling, Nathan Schumann and Bill Long
Cray Inc
Saint Paul, MN
USA
sbehling@cray.com, nds@cray.com, longb@cray.com
Carlos P. Sosa
Cray Inc and University of Minnesota Rochester
Saint Paul, MN
USA
cpsosa@cray.com
Sébastien Boisvert and Jacques Corbeil
Département de Médecine Moléculaire, Université Laval, Québec, Canada
sebastien.boisvert.3@ulaval.ca, jacques.corbeil@crchul.ulaval.ca
Lorenzo Pesce
Computation Institute, University of Chicago, Chicago, IL USA
lpesce@uchicago.edu

Agenda

- Big Data Problem in Genomics
- Next-Generation Sequencing Work Flow
- Genomics Applications: Ray
- Summary

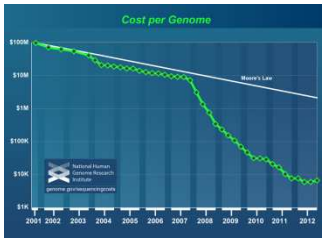
A New Dimension for HPC in Life Sciences

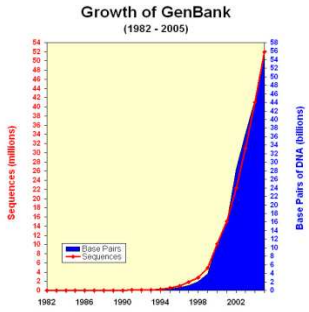
Personalized Therapeutics

Data Analytics

Simulations

In 2012 U of Illinois and NCSA completed a 100 Million atoms Chromatophore NAMD Simulation on a Cray named Blue Waters





In 1998 Prof Kollman study folding of Villin Headpiece Sub-Domain Observed for a Full Microsecond on a Cray T3E

Personalized Medicine Era

Big Data Era

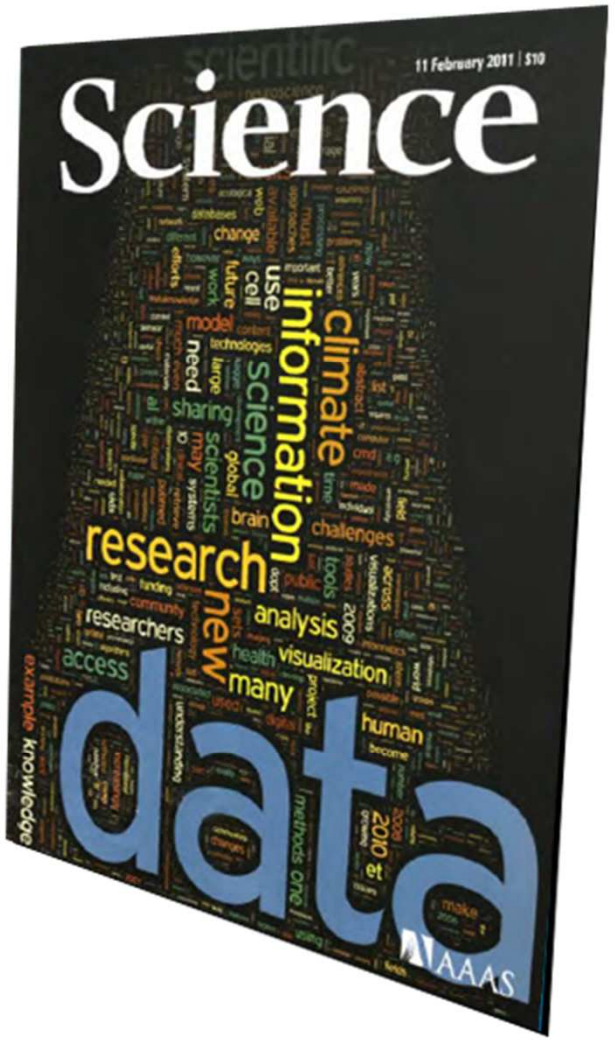
Omics Era

Structural Biology Era

Selected Milestones in HPC

Selected Milestones in Life Sciences

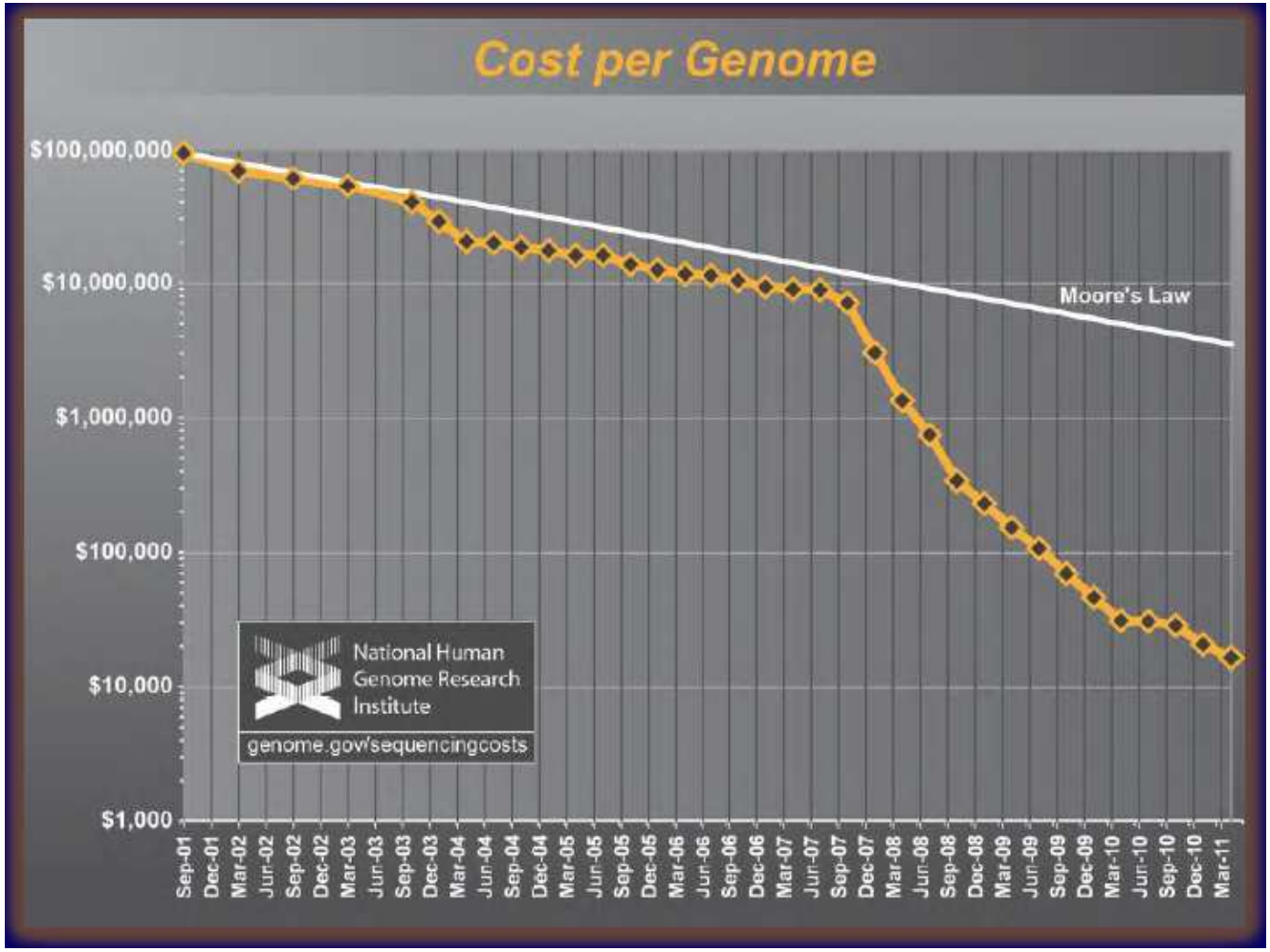
The Era of Big Data



Source: Eric Green, Director, National Institute of Health: NextGen 101 Workshop

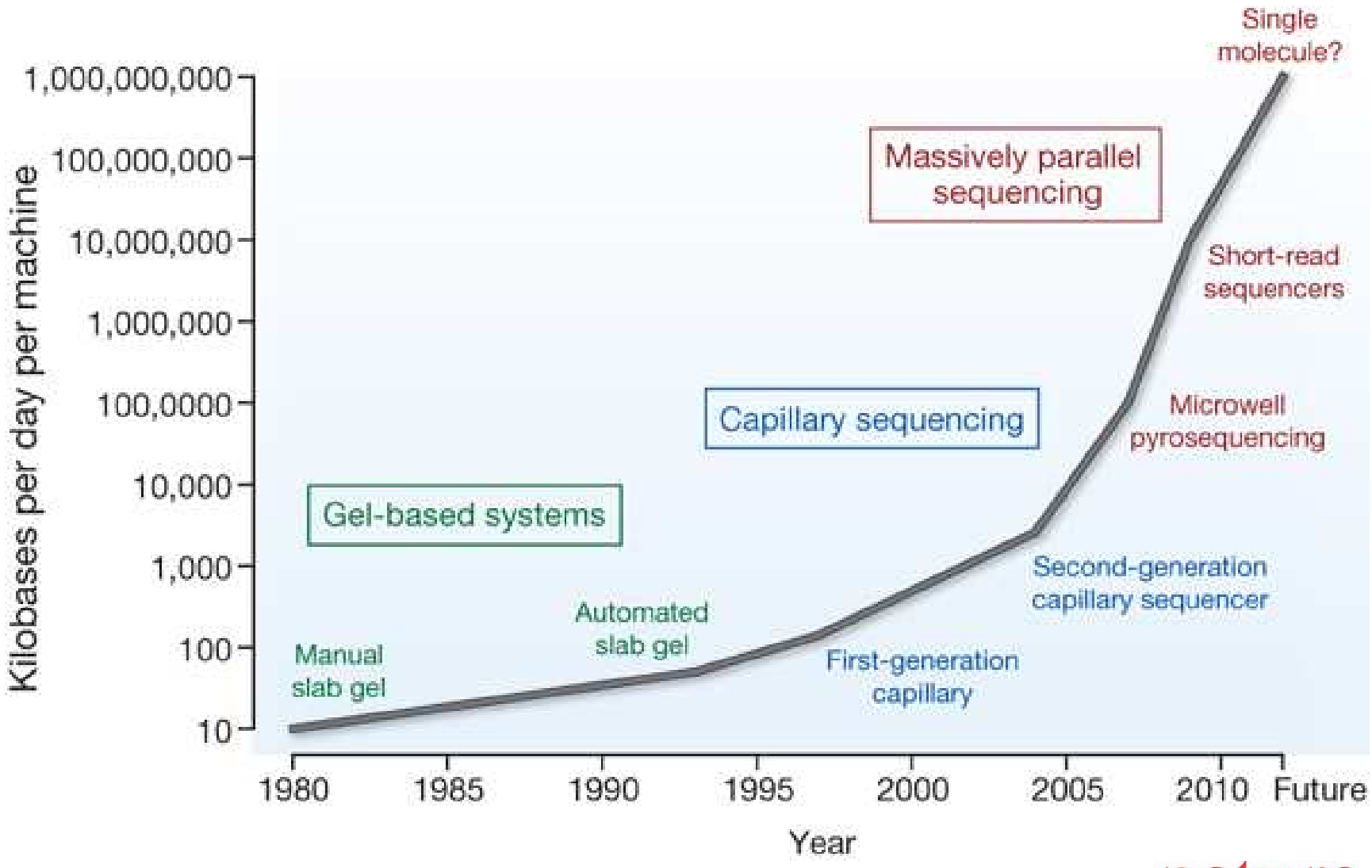
Next Generation Sequencing (NGS)

- Cost per sequenced human genome



- Source: genome.gov/SequencingCosts/

Big Data Generation Evolution and Explosion



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

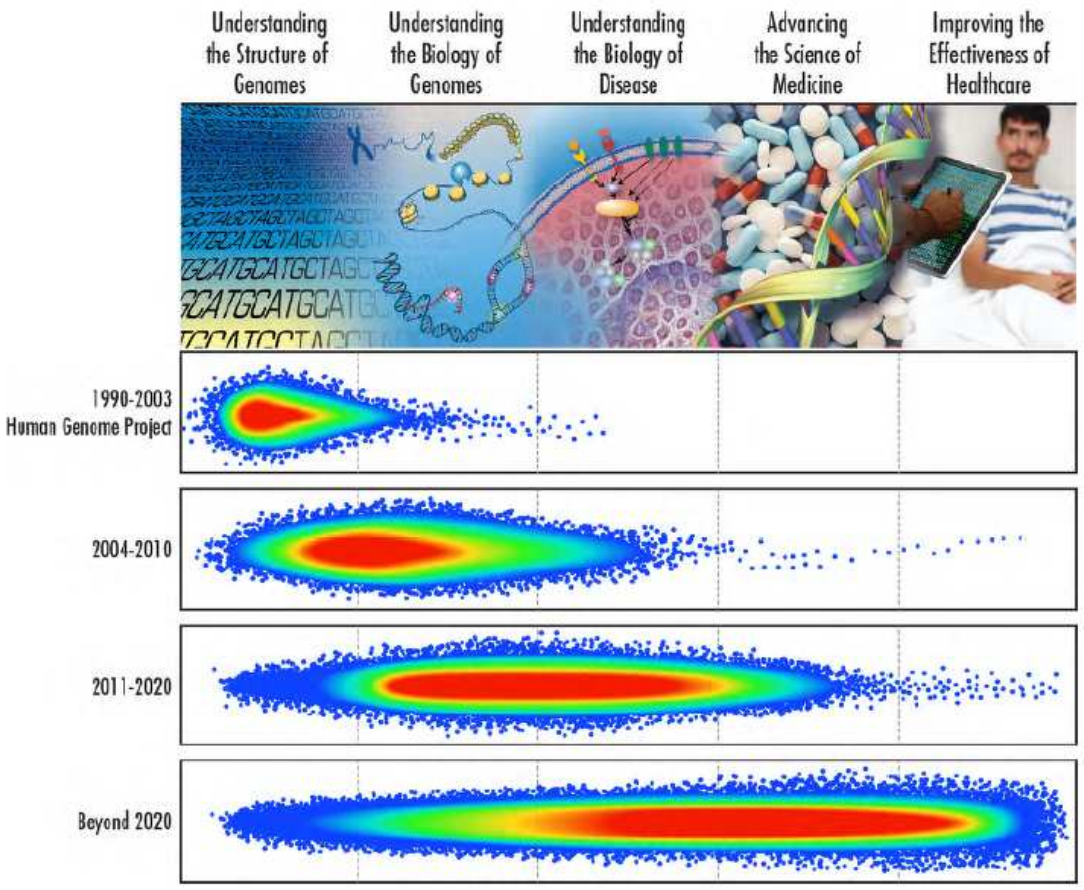
Current Processing and Analytics Tools Cannot Cope with the Amount of Data Generated in Genomics

Think about this: you need to collect the **1.5 GB** for each person and likely extract out the genetic markers. Then you need to analyze the cancer and the treatment data.

According to The American Cancer Society, **12,549,000** people in the U.S. have cancer. So at 1.5 GB per person, that comes out to about **18.8 PB** of data — **and this does not include the genetics of the cancer.**

Henry Newman - Cancer, Big Data and Storage

Data Explosion in the Quest for Personalized Medicine



Source: Eric Green, Director, National Institute of Health: NextGen 101 Workshop

Ericka C. Hayden, NATURE, VOL 482, 16 FEBRUARY 2012

FEBRUARY 2001 Human genome draft completed by competing teams.

APRIL 2008 First sequence of an individual human, James Watson.

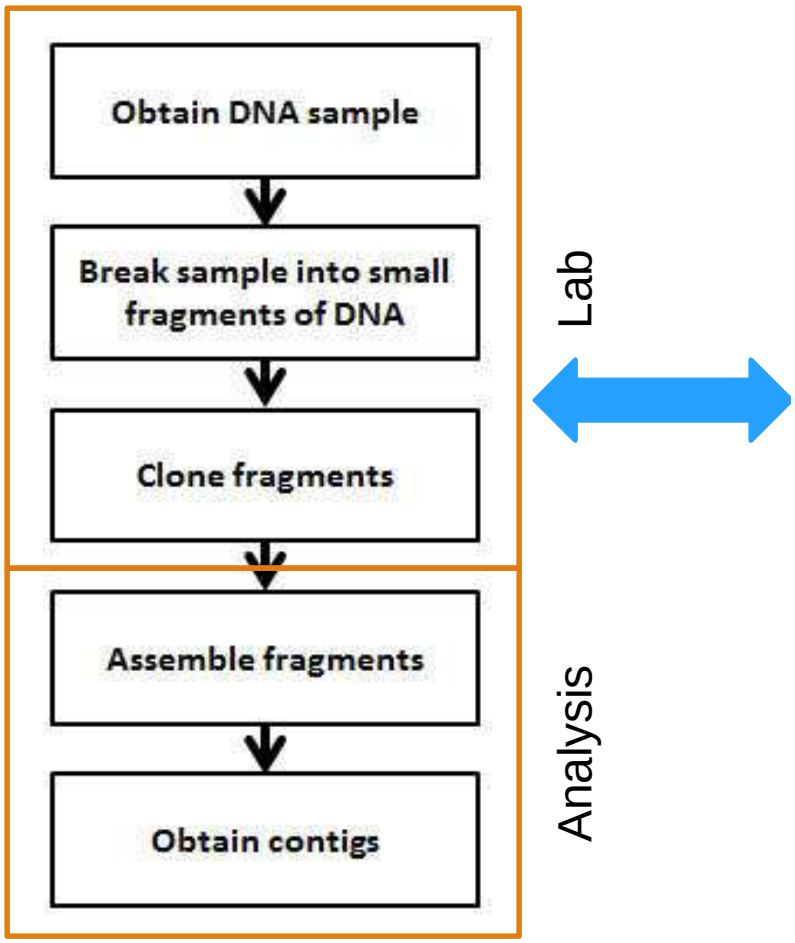
MARCH 2010 First sequenced family uncovers causative gene for Miller syndrome.

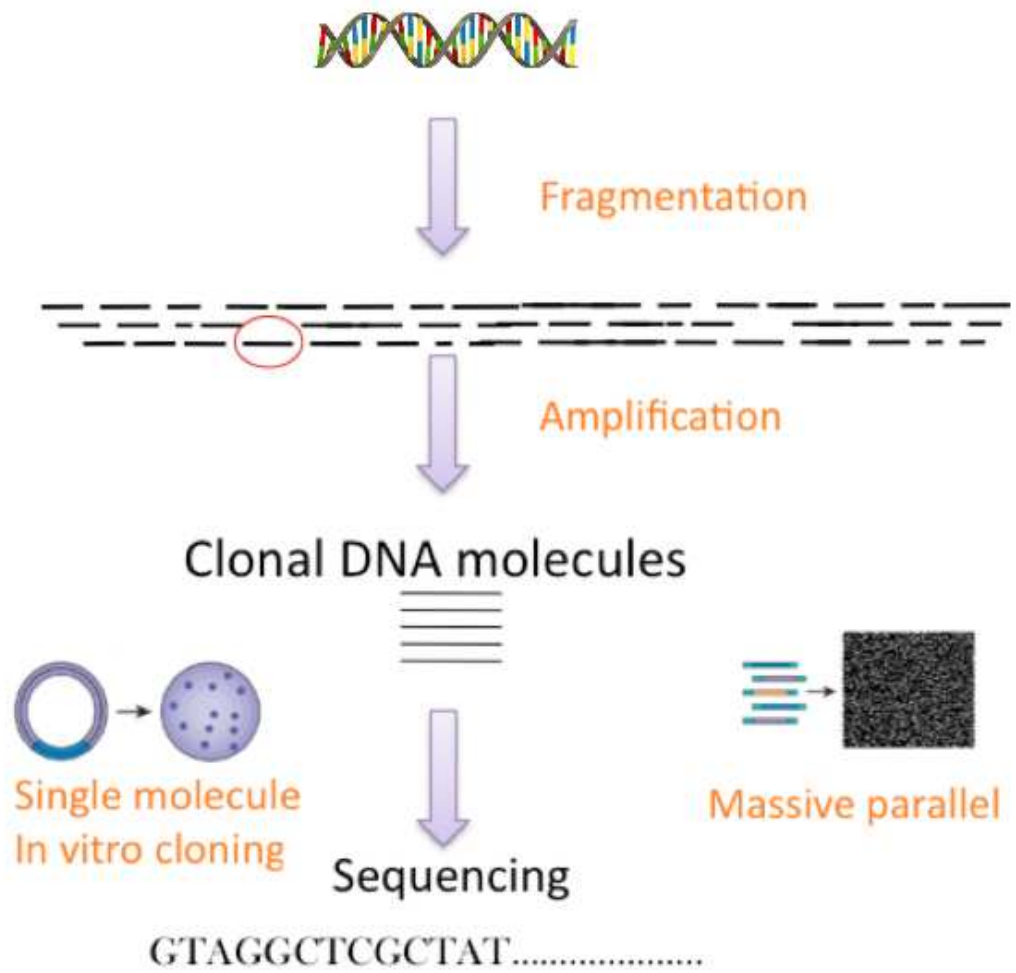
JUNE 2010 Doctors help to restore health of Nicholas Volker (**pictured**) after sequencing indicates that his inflammatory bowel disease could be alleviated by a bone-marrow transplant.

APRIL 2011 Sequencing spares a woman with leukaemia from undergoing a bone-marrow transplant.

JUNE 2011 Doctors report using whole-genome sequencing to improve treatment for a patient with the movement disorder dopa-responsive dystonia.

What is Sequencing?





Best Practices for Variant Calling with the GATK, Broad Institute, MIT, Cambridge, MA
Image source: http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-0A-Intro_to_NGS.pdf

WorkFlow Based on Illumina Sequencer

Next Generation Sequencers

TeraBytes of images (tiff)

PC with large file systems

Storage

Transferred Storage

Approx. 1 TeraByte

Formatted into BCl file

Translate to fastq files Approx. 100 GB

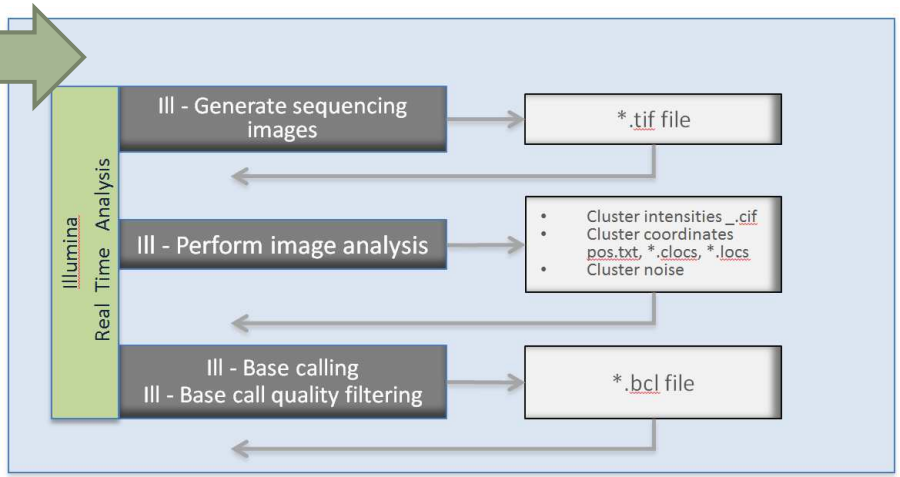
Selected Applications

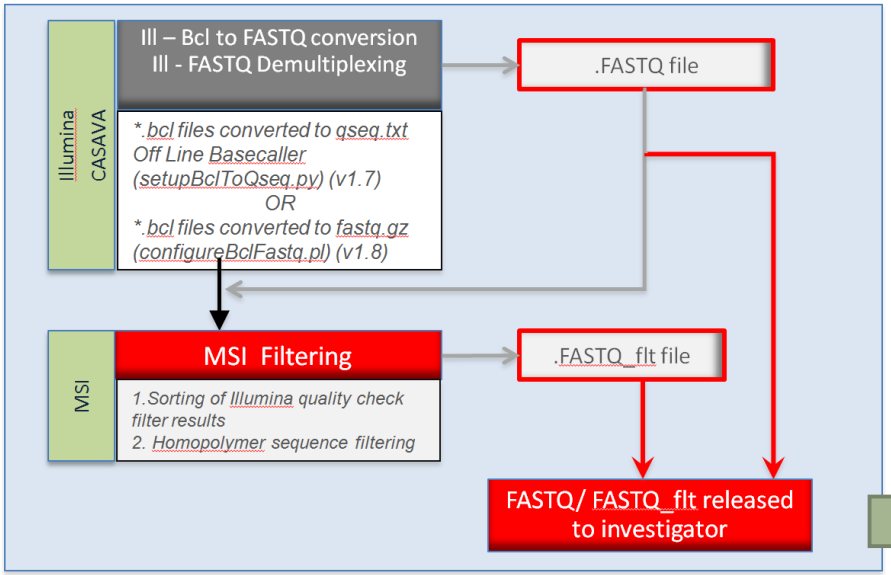
- Abyss
- AFABC
- ALLPATH
- AMOS
- Amplicon Noise
- BamTools
- BEAST
- BEDTools
- BEERS
- Bioconductor
- BLAST+
- Blat
- BlueGnome
- Bowtie
- BreakDancer
- BWA
- CAP3
- CASAVA
- ClustalW/X
- Cufflinks
- Decypher
- EMBOSS
- FastQC
- GATK
- GMAP
- Hmmer
- Illumina GenomeStudio
- MIRA

https://www.msi.umn.edu/sw/cat/genetics

Detailed Sample Workflow: Illumina Sequencer

HiSeq 2500
http://www.illumina.com



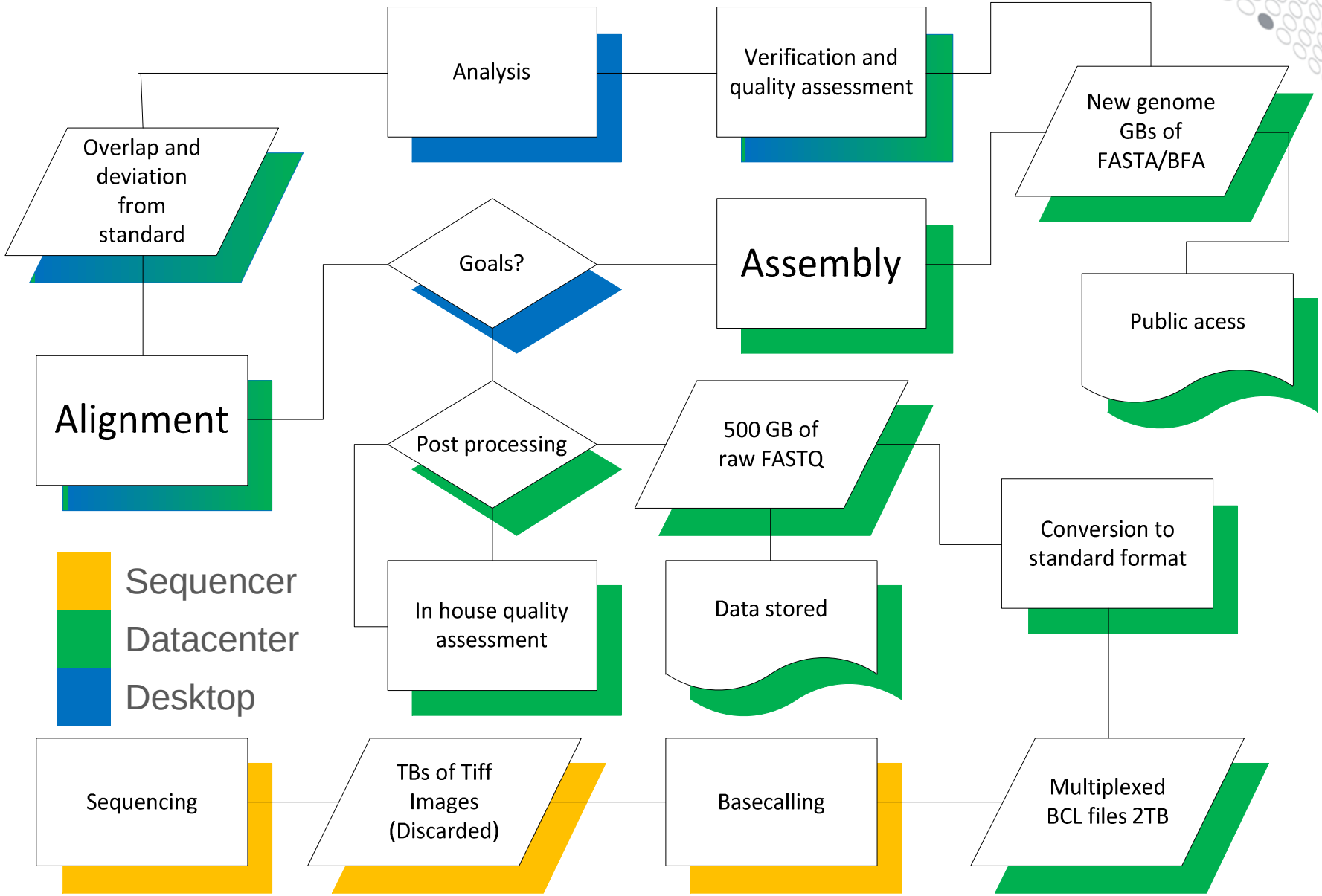


Analysis

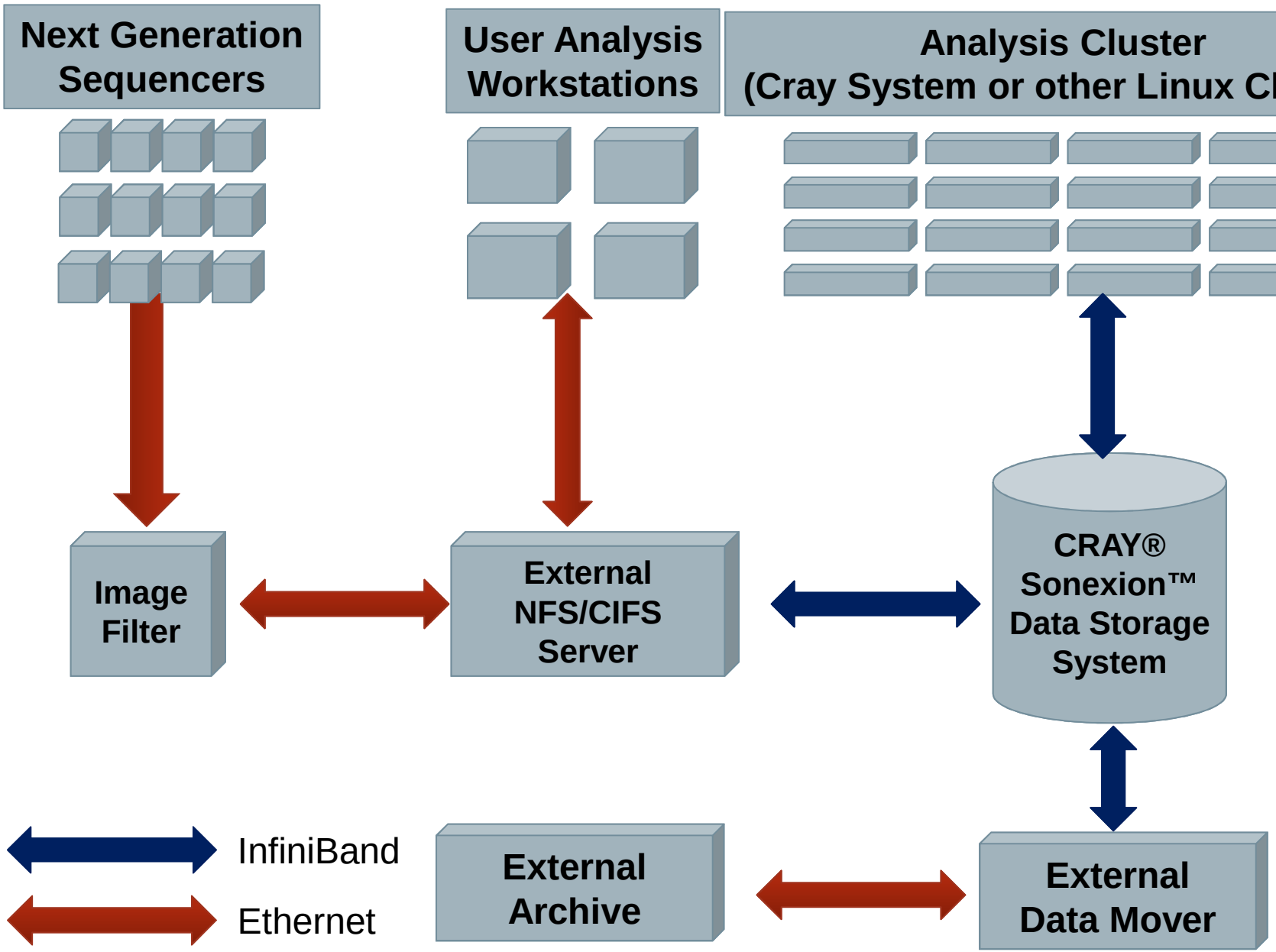
- Whole Genome Resequencing
- Targeted Resequencing
- DeNovo Assembly
- Chromatin Immuno-prescription Sequencing
- Methylation Analysis
- Whole Transcriptome Analysis
- Small RNA Analysis
- Gene Expression
- RNA Structure
- Metagenomics
- Exomics

Data Analysis

Workflow Based on IT Resources Required



Cray Genomics Work Flow Diagram



Analysis: Leveraging Cray Solutions

Analysis

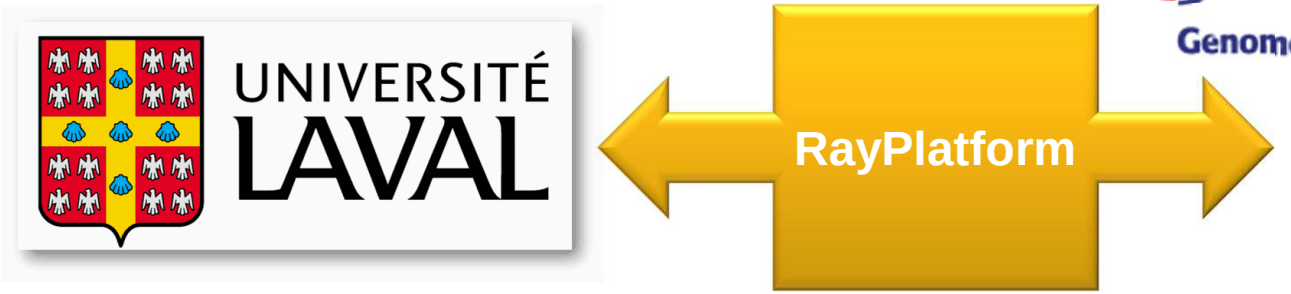
- Whole Genome Resequencing
- Targeted Resequencing
- De Novo Assembly
- Chromatin Immuno-prescription Sequencing
- Methylation Analysis
- Whole Transcriptome Analysis
- Small RNA Analysis
- Gene Expression
- RNA Structure
- Metagenomics
- Exomics

Cray started a collaborative effort with Université Laval:

- To achieve the goal of assembling a human genome in less than 1 hour
- Ray can achieve it in 10 hours and other assemblers in days
- This tool will help toward the goal of personalized medicine

Bioinformatics and Computational Biology Project

Collaborating to enable rapid genomics analysis



Ray: Hybrid De Novo Assembler

Accelerating Genomic Applications is a Collaborative effort

The goal in De Novo assembly is to correctly assemble short reads into longer sequences

- Representing contiguous genomic regions

Current Next Generation Sequencing technologies offer increase in throughput and decrease in cost and time

Most software is available to assemble reads from specific NGS system

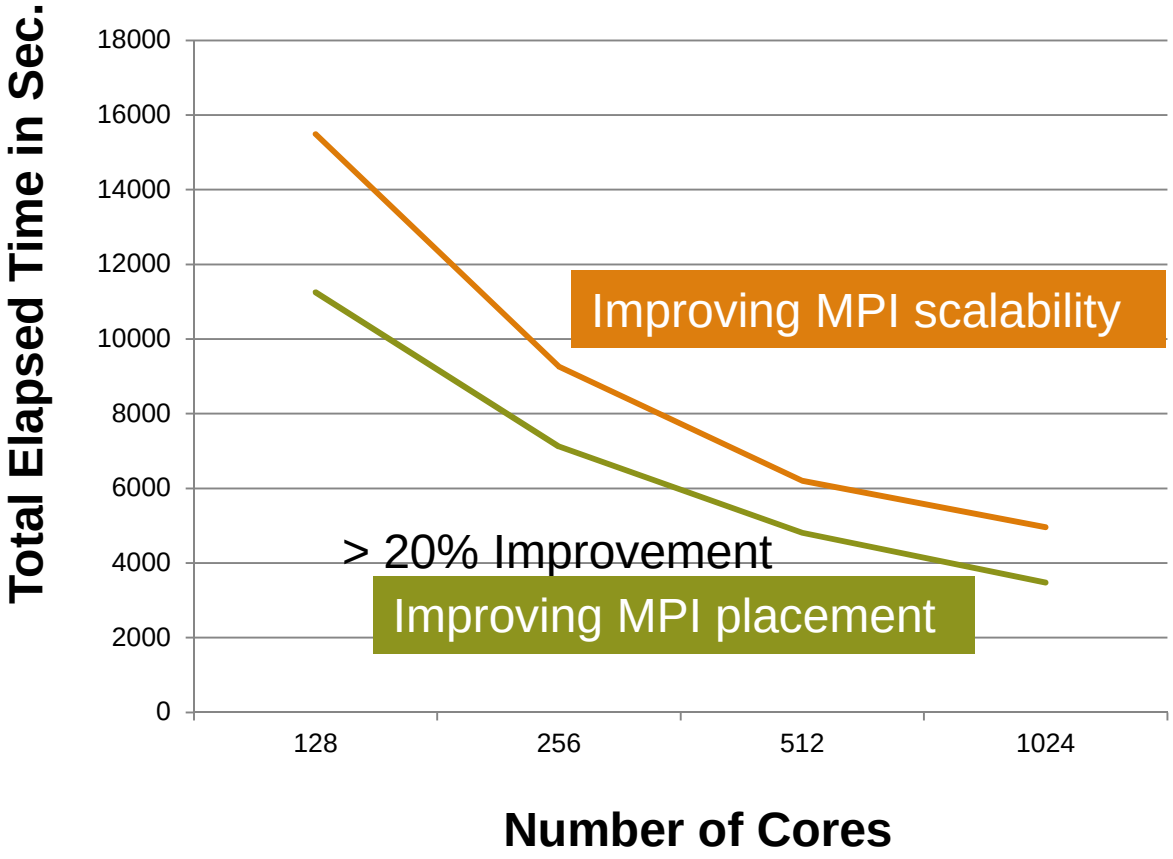
Ray has been developed to assemble reads obtained from a combination of sequencing platforms

- S. Boisvert, F. Laviolette, and J. Corbeil, J. Comp. Biol. 17, 1519-1533(2010)

Ray Parallel Performance

Human gut gene catalog
Metagenomics
124 Individuals, 577 GB generated
Beijing Genomic Institute

Application Tuning



- The Ray benchmarks shown in this study were run on a Cray XE6 system with AMD Opteron™ Interlagos IL-16
- Clock frequency of the core of 2.1 GHz

Future Work

- Ray represents a major step forward in overcoming some of the major challenges facing genome assembling today
- This is particularly true for large datasets that otherwise are intractable
- To achieve the goal of sequencing even large data sets additional optimization work will be required
  - We'll put particular attention to the bidirectional extension of seeds
- Continue to explore the role of Lustre, large scale data storage and storage optimization in detail with the parallel, high-throughput genomics workflow.
- Extend the scalability of Ray in collaboration with ORNL on Titan

Summary

- NGS machines technology have provided critical tools for deciphering DNA
- The cost of one Mb of DNA sequence has gone down from about $5,000 in 2001 to approximately $0.78 in 2009
- Assembler programs have been created to assist in the process of assembling genomic data
- *Data coming from sequencers outpaces Moore's law, it is critical to develop tools and procedures that can accurately and efficiently keep pace with the data production*
- Ray provides next generation of assemblers
- Ray represents a major step forward in overcoming some of the major challenges facing genome assembling today

Acknowledgements

Biomedical Informatics and Computational Biology (BICB)

- Prof Claudia Neuhauser, UMR
- Michael Olesen, UMR

Université Laval

- Prof Jacques Corbeil
- Sébastien Boisvert

Cray Inc

- Per Nyberg, Segment Team
- Sonexion Team
- Cray Inc resources (Cray XE6)

http://www.cray.com

Thank You!