HIGH-PERFORMANCE BIOLOGICAL COMPUTING
University of Illinois at Urbana Champaign

Using High Performance Computing in Computational Genomics

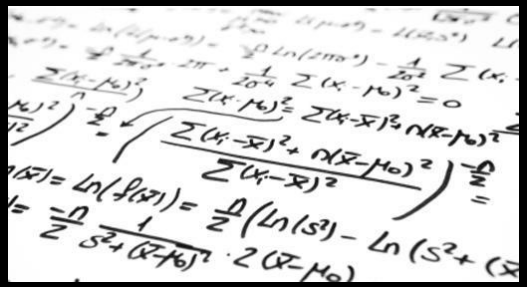
Liudmila Sergeevna Mainzer
SPIN
June 27, 2016

High Performance Biological Computing

A Core Facility Anchored in Research and Technology

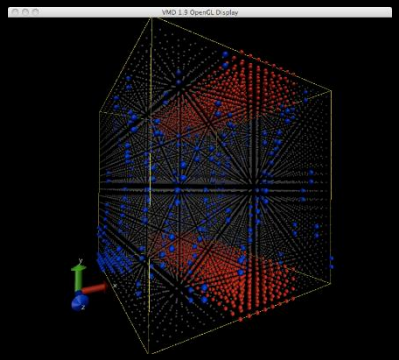
IGB, Carver Biotechnology Center, NCSA

Infrastructure: hardware, software, data



User support: experimental design, analysis, statistics

Training: short courses, workshops



Applied R&D: scalability, optimal HPC architectures

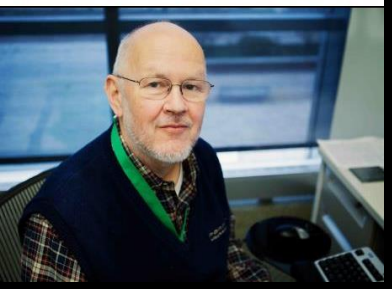
International engagement

Applied R&D at HPCBio

- Testing and benchmarking new methods
  - Survey of literature and social media, technology trends
  - Benchmarking of new methods using our own datasets, comparison to current best practice
  - Establishing consistent benchmarks for accuracy and performance (synthetic data, consistency checks)
- Scaling of existing methods
  - Sizes and numbers of datasets are constantly increasing → scaling issues
  - Explore how to best use large scale computational resources (Blue Waters, Clouds) for the analysis of large and complex datasets
- Computational systems research
  - Explore the behavior of best practice workflows when deployed on different systems architectures

Computational Genomics at NCSA and UIUC

INSTITUTE FOR GENOMIC BIOLOGY



Victor Jongeneel,
Director of HPCBio

CSL: COORDINATED SCIENCE LAB

Ravi Iyer,
Professor of ECE

- Architecture:

  What kind of computer architecture is best suited for bioinformatics work?

- Performance bottlenecks:

  What are the performance bottlenecks for bioinformatics work, on different architectures?

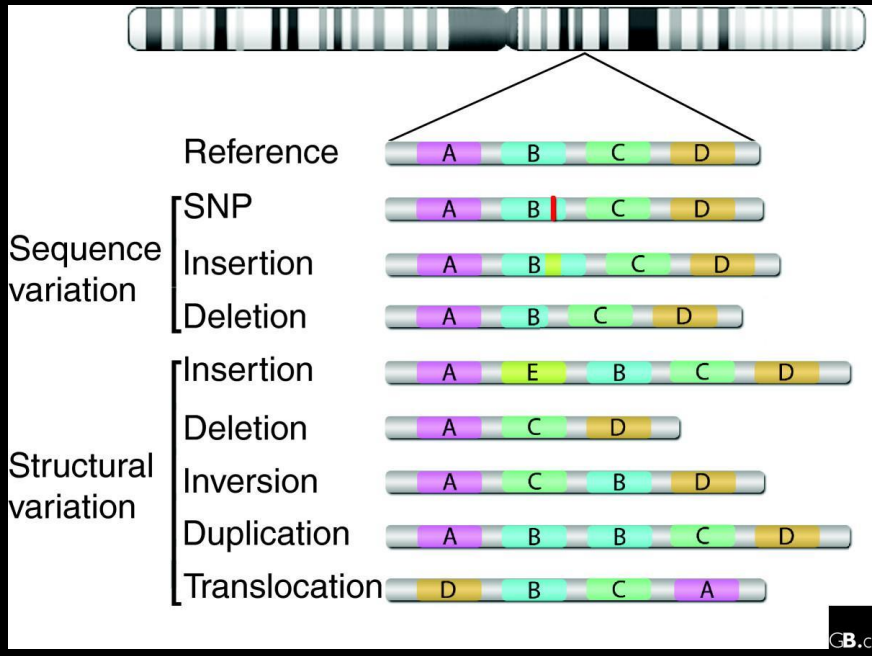
- Future:

  How to structure the bioinformatics workflows for best performance on the architectures upcoming in the next 5, 10, 20 years?

CompGen INITIATIVE NCSA BLUE WATERS SUSTAINED PETASCALE COMPUTING CRAY

What is Genomic Variant Calling?
Why do we think it is important?
Why does it need high performance computing?

Genomic Variant = a difference in the genetic code



Rahim *et al. Genome Biology* 2008 **9**:215

Even a single variant in a single gene can lead to a drastic difference in phenotype

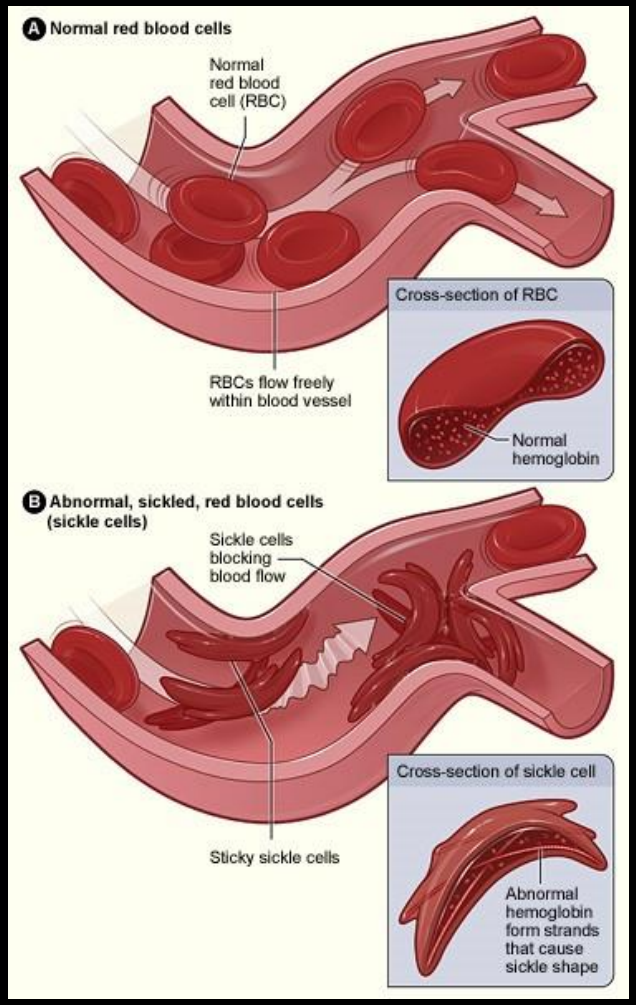


Image from http://www.nhlbi.nih.gov/

Sickle-cell anemia is a Mendelian disease.

NHGRI:
Since 2011, Centers for Mendelian Genomics sequenced >20,000 human exomes.

Human exome ~ 2% human genome

1 sample ~ 10 GB sequencing data
20,000 samples ~ 200 TB sequencing data

Data footprint, scaling up

1 B	"Hello World" = 12 B	
x1000 = 1 KB	1 page of code ~ 6 KB	floppies, memory sticks
x1000 = 1 MB	this presentation ~ 2.5 MB	
x1000 = 1 GB	Soybean sequencing data ~ 4 - 69 GB	
x1000 = 1 TB	human tumor/normal sample pair, WGS	laptops, servers
x1000 = 1 PB	daily data production	clusters, supercomputers

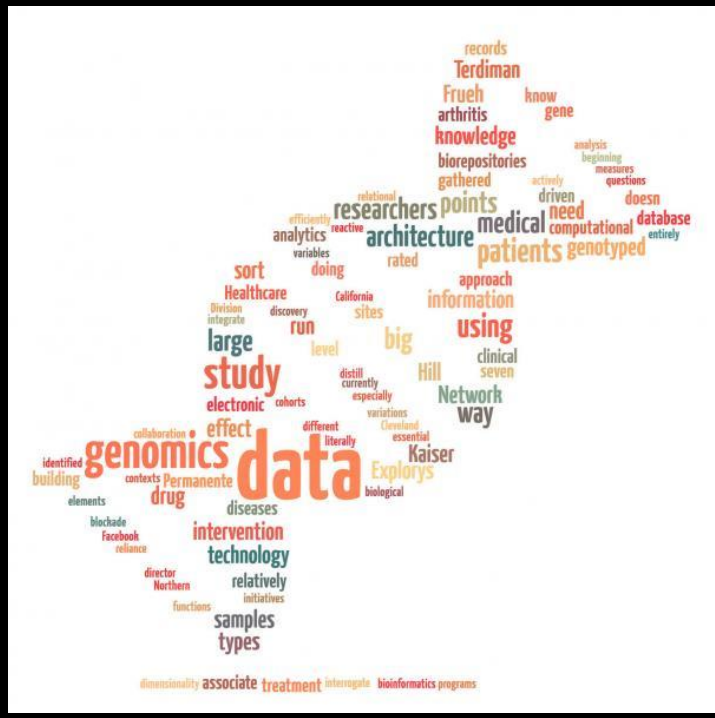
Biocluster at IGB:	700 TB of project space
iForge at NCSA:	600 TB of project space
Blue Waters:	26 PB of disk storage

20,000 of the NHGRI WES samples ~ 200 TB sequencing data

Petascale storage requirements

Complex traits are influenced by many variants, frequently outside coding regions:

- BMI
- Human height
- Alzheimer's disease
- Diabetes
- Stroke
- Autism
- Heart disease
- Intelligence
- Fertility



NHGRI:
Centers for Common Disease Genomics plan to sequence ~200,000 whole human genomes.

1 sample ~ 200 GB sequencing data (depth-dependent)

200,000 samples ~ 40 PB sequencing data → input data to the variant calling process

2015: Obama announced Precision Medicine Initiative

" to bring us closer to curing diseases like cancer and diabetes – and to give all of us access to the personalized information we need to keep ourselves and our families healthier."

"I want the country that eliminated polio and mapped the human genome to lead a new era of medicine – one that delivers the right treatment at the right time,"

U.S. President Barack Obama delivers his State of the Union address to a joint session of the U.S. Congress on Capitol Hill in Washington, January 20, 2015. Reuters/Jonathan Ernst

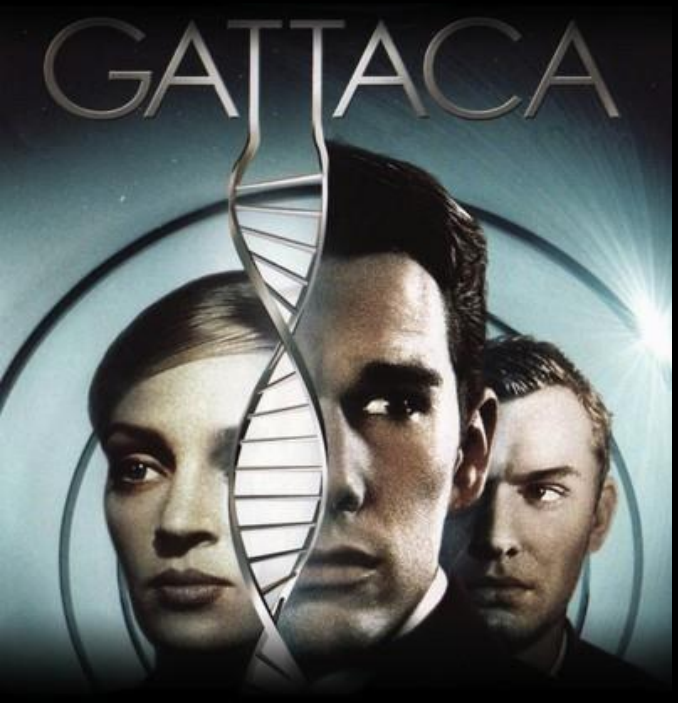
NIH http://www.nih.gov/precisionmedicine/

Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.

Sustained Petascale and Exascale storage requirements

What if we had to genotype every baby being born? = 500 genomes/day in the state of Illinois

NIH http://www.nih.gov/precisionmedicine/



NERVE CONDITION - PROBABILITY 60%,
MANIC DEPRESSION - 42%,
OBESITY - 66%,
ATTENTION DEFICIT DISORDER - 89%
HEART DISORDER - 99%
EARLY FATAL POTENTIAL
LIFE EXPECTANCY - 33 YEARS

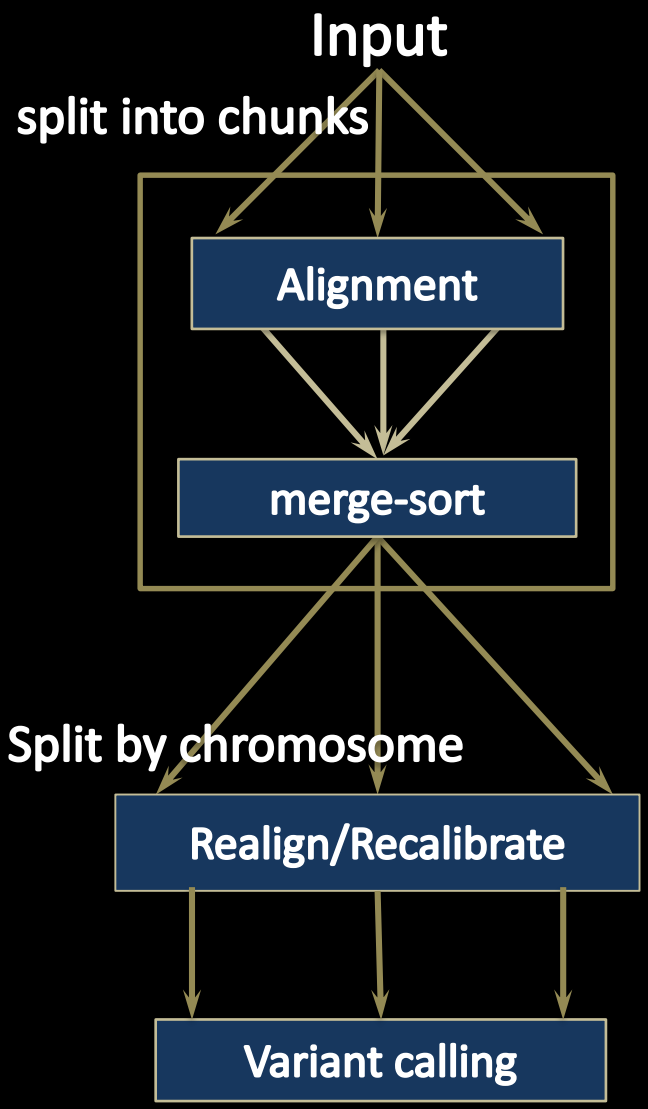
Sustained Petascale and Exascale storage requirements

What if we had to genotype every baby being born? = 500 genomes/day in the state of Illinois

NIH http://www.nih.gov/precisionmedicine/

- **Input**
  - 300-600 GB/genome
  - 150-300 TB/day when analyzing 500 genomes/day
- **Intermediary**
  - 3 TB per sample with intermediaries
  - **0.3-1.5 PB/day when analyzing 500 genomes/day**
- **Output**
  - Tiny: < 500 M per sample

Compute requirements: Node count, not flops



1. Alignment

500 jobs for BWA

If chunking input data: 5,000 jobs for Novoalign

2. Split by chromosome

25 chromosomes * 500 genomes = 12,500 jobs

3. Realign/Recalibrate

25 chromosomes * 500 genomes = 12,500 jobs

4. Variant calling

25 chromosomes * 500 genomes = 12,500 jobs

Blue Waters

Node Type	Cray XE6	Cray XK7
CPU	2 x AMD “Interlagos” Opteron 6276	1 x AMD “Interlagos” Opteron 6276
GPU	NA	1 x Nvidia “Kepler” Tesla K20x
Total Nodes	**22,640**	**4,224**
Total x86 Cores	362,240	33,792
Cores/Node	16 FP x86_64 cores, 2.45 GHz	8 FP x86 Cores, 2.45 GHz; 2688 CUDA cores
Memory/Node	64 GB	32 GB (CPU) + 6 GB (GPU)
Storage	**26.4 petabytes (disk)**, 380 petabytes (nearline)	
Interconnect	**Cray “Gemini” 3D Torus**	
OS	Cray Linux 6	

Large scale plant and animal genotyping

Ongoing and future projects

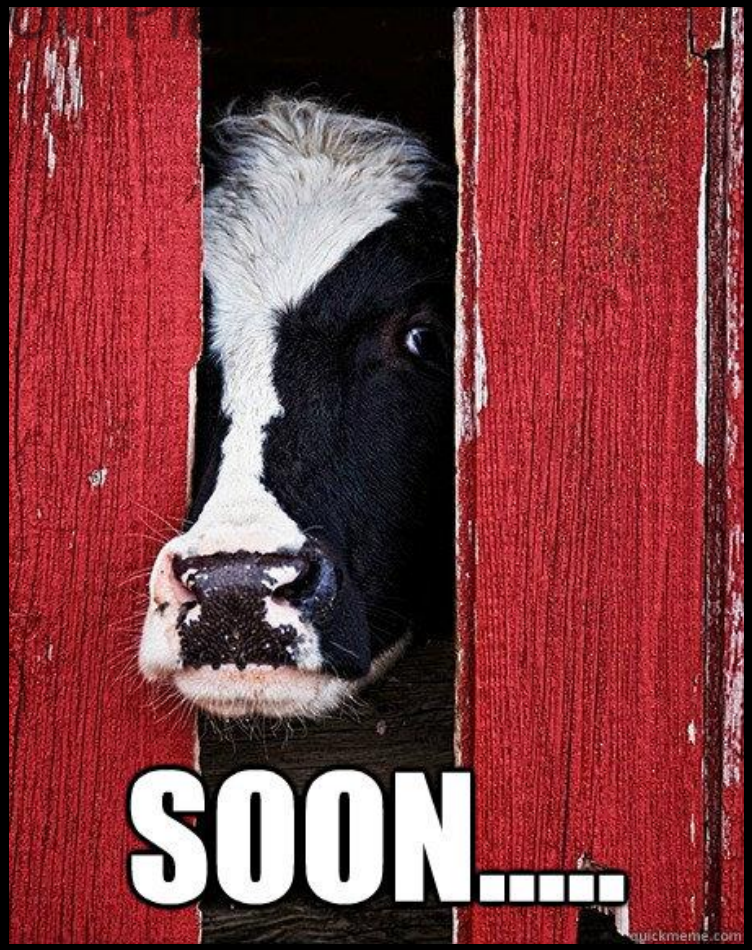
- 1000+ Arabidopsis genomes
- 3,000 Rice varieties
- 1000 Fungal genomes project
- Genome10K: 16,000 Vertebrates
- 5,000 Insect genomes

Complex traits of note

- Plant biomass
- Nutritive content of grain:
  oil, protein, vitamins, minerals
- Parasite resistance
- Milk volume
- Muscle mass

Purebreddairycattle.com

It won’t stop here



Purebreddairycattle.com

The best computer for genomics

Next generation of Blue Waters:

- HPC expertise and a solid, dedicated support team like that on BW is absolutely essential

- Must have ~>256 GB RAM per node
- Nodes must have internal storage: 1-4 TB
- We want lots of cores: 32-64

Acknowledgements

INSTITUTE FOR GENOMIC BIOLOGY

HPCBio

Victor Jongeneel

Gloria Rendon

Chris Fields

NCSA

NCSA Industry Engagement

Evan Burness

Jim Long

Wayne Hoyenga

CRAY

Cray

Bob Fiedler

Carlos Sosa

Pierre Carrier

Richard Walsh

Bill Long

Jef Dawson

Blue Waters support team

Greg Bauer

Victor Anisimov

Ryan Mokos

Kalyana Chadalavada

Alex Parga

Jeremy Enos

Andriy Kot

Jason Alt

Craig Steffen

CompGen INITIATIVE

CompGen

Ravi Iyer

Subho Banerjee

Arjun Athreya

Zachary Stephens

Innovative Systems Lab

Volodymyr Kindratenko

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

H3A bionet: UCT

Gerrit Botha

Ayton Meintjes

Nicola Mulder