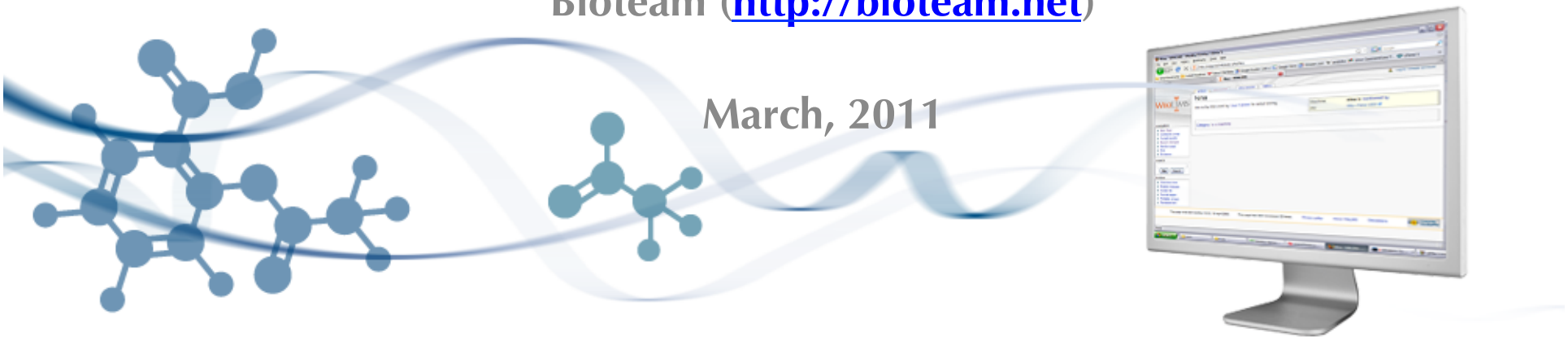
DNA Sequencing, Early 2011

Chris Dwan cdwan@bioteam.net
Bioteam (http://bioteam.net)

March, 2011



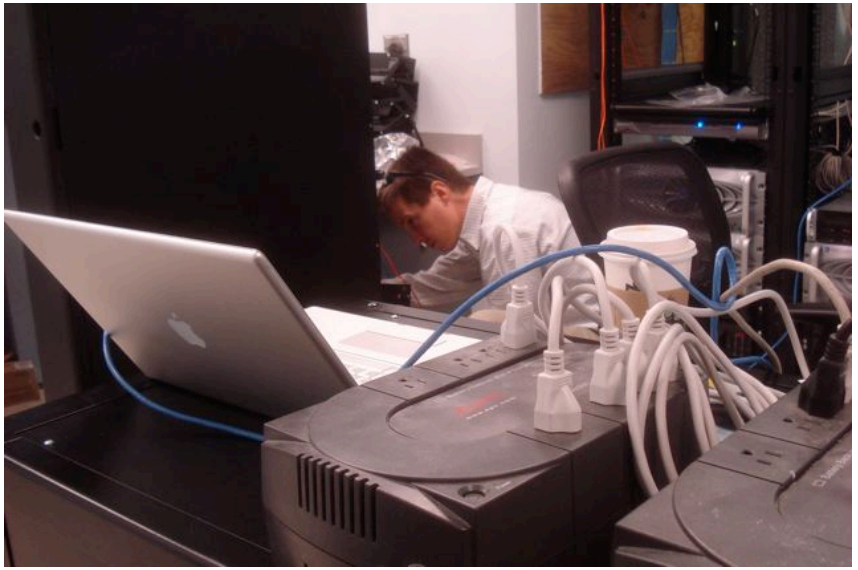
Bioteam Inc.

Independent Consulting Shop

- Incorporated: 2002
- Vendor / Technology Agnostic: No financial incentives from any vendors

Staffed by

- Scientists forced to learn high performance IT
- Many years of industry / academic experience



A fun business niche

- Bridging the gap between science, IT, and high performance computin
- *Custom LIMS solutions (WikiLIMS, MiniLIMS)*

cdwan@bioteam.net

Key Messages

For IT:

- Bioinformatics is biology
- Chemistry is changing far faster than the underlying IT infrastructure
- Successful solutions have strong stories around flexibility, scaling, capacity

For Biology:

- Petabytes and teraflops are commodities.
- Treat IT as engineering rather than science

Data management and accessibility is the current major challenge.

cdwan@bioteam.net

DNA Sequencing Technology

1869: DNA discovered

1937: Regular crystal structure ("A stupid tetranucleotide")

1952: Identified as genetic material

1953: Double helix structure discovered.

> "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material"

1957: Central dogma (DNA – RNA – Protein)

1975: "Sanger" sequencing (chain termination)

2003: 2nd generation: High throughput technologies

2010: 3rd Generation: Single molecule technologies

2011: 4th Generation: "Personal"

cdwan@bioteam.net

Modern DNA Sequencing At A Glance

2005-2007

- Roche 454
- Illumina GA

2008-2009

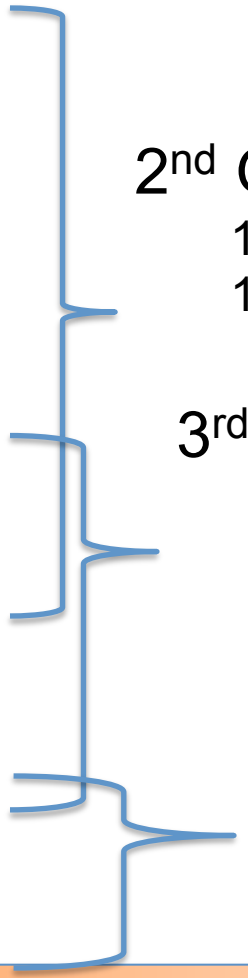
- SOLiD 3, 4
- Roche Titanium
- Illumina GA IIx

2010

- Illumina HiSeq
- Roche PI
- Pac Bio
- Ion Torrent PGM
- SOLiD Junior

2011

- Illumina MiSeq
- GridION



2nd Generation

10^3 throughput improvement
10^3 cost per base improvement

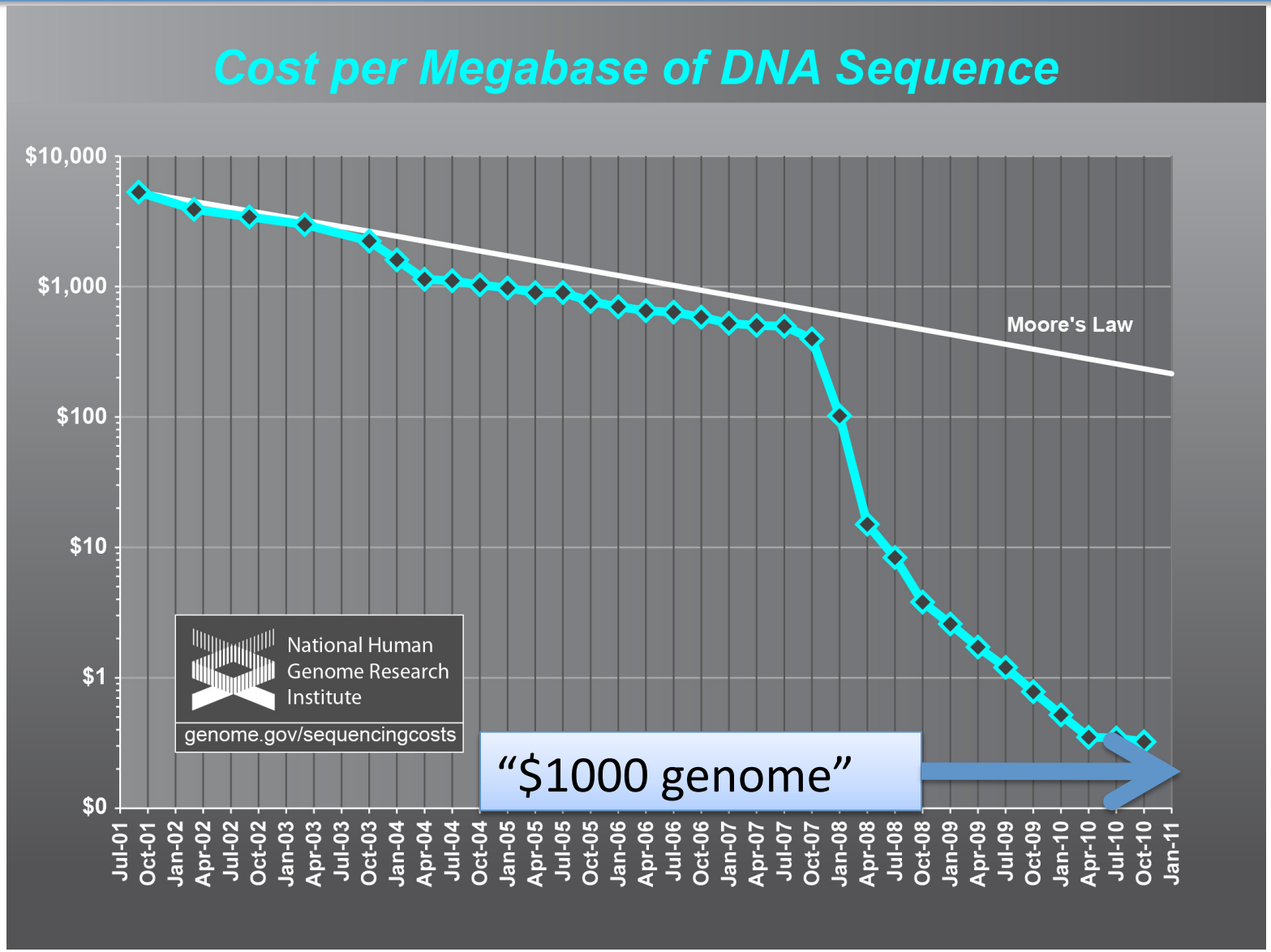
3rd Generation

Single molecule **and / or**
Higher throughput **and / or**
Newest version

4th Generation?

Personal **and / or**
25% more GridCloud

cdwan@bioteam.net



cdwan@bioteam.net

Major Differentiators

Scientific:

- Chemistry: DNA, cDNA, RNA, …
- Read length: "Long Reads" (500+ bp) vs. "Short Reads" (< 100bp)
- Run duration: 15 minutes to 7 days

IT:

- Data volume per run: 1 – 900 GB
- Onboard computing resource: Nothing, single server, small cluster

Operational:

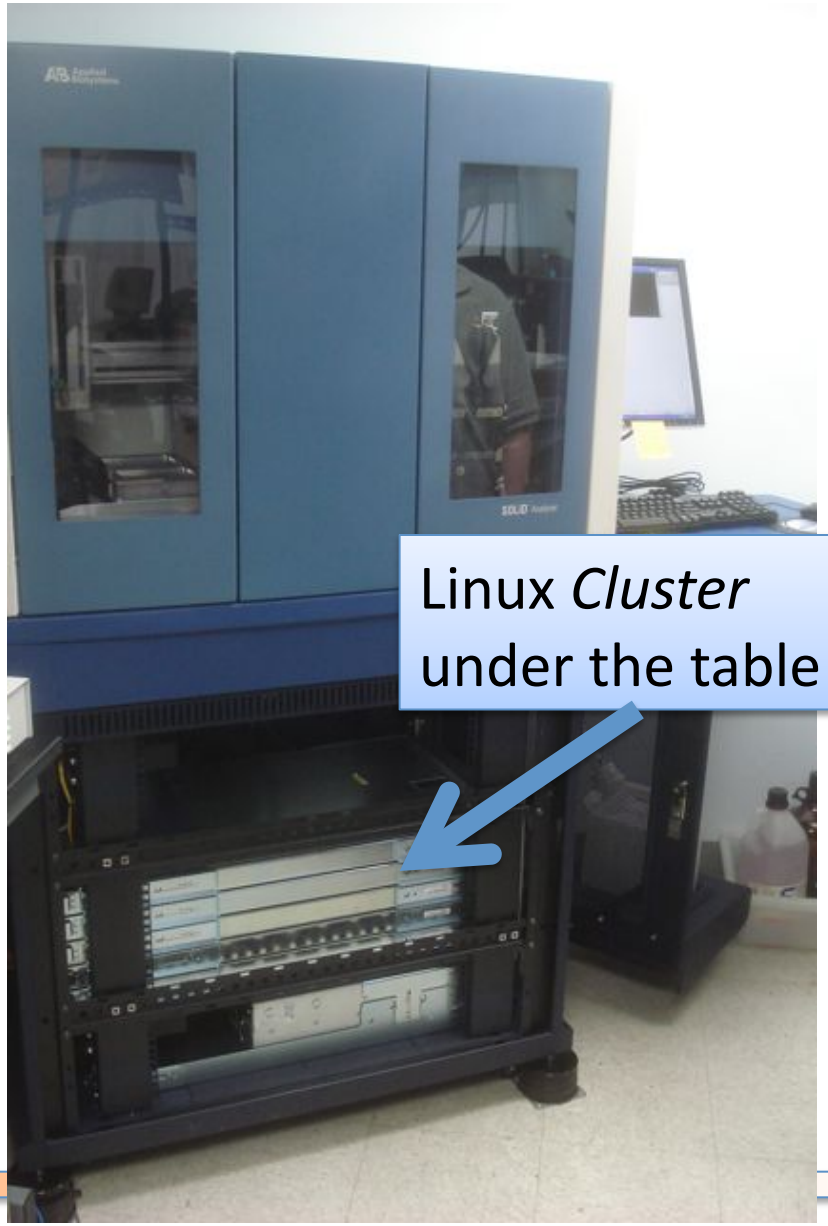
- "High Throughput" vs 9 to 5 labs
- Data retention policy: "none" through "life of the company plus 3 years"
- Operational availability
- Reagent cost per run: $100 to $10,000
- Instrument cost: $20,000 to $750,000

cdwan@bioteam.net

454 Instrument

cdwan@bioteam.net

Illumina and SOLiD Instruments



cdwan@bioteam.net

Ion Torrent PGM

Linux server (TorrentServer) somewhere nearby …

cdwan@bioteam.net

Generalized Next Gen Workflow

Data acquisition

- Chemistry and image capture

Primary Analysis (once per sample)

- Base calling

Secondary Analysis

- Multiple instrument runs
- Contig assembly
- Mapping to reference genomes

Downstream Analysis

- Here there be dragons

cdwan@bioteam.net

Complete Genomics

Human re-sequencing service provider

- Recently released 40+ human genomes

"Cloud Sequencing"

Solving the distribution problem

- All data stored in S3
- Primary datasets shipped to customers on SATA, direct from Amazon

cdwan@bioteam.net

Management, Not Just Storage

Data volumes are becoming manageable

- 10TB to 200TB of primary data per instrument year
- 20% to 500% additional space for derived data (seriously)
- On-instrument data reduction and migration away from optical sensors

Data management is the problem

- Post-it notes, emails, and descriptive file names are simply not going to cut it
- Underlying science changes too fast for all but the most agile LIMS

User Expectations are still (always) a problem

- "I can get a terabyte from CostCo for $80" (Feb 2010)
- End users tend to have no clue about the true cost of keeping data online and accessible

cdwan@bioteam.net

~~Best~~ Decent Practices

Retain descriptive, structured filenames and paths for primary data

- These have saved me more times than I care to admit
- When the LIMS freaks out …

Deleting data must be a scientific decision

- IT staff are (rightfully) terrified of deleting your stuff
- Management: Insist on annual (at least) data audits

Communicate, communicate, communicate

- Scientific staff and leadership are actually quite savvy
- IT staff and leadership are also quite savvy
- Resist the temptation to obscure straightforward goals and mutual benefit behind complex SLAs.

cdwan@bioteam.net

Backups / Data Lifecycle Management

Backups have been a sick joke since 2008

- Most groups simply build disk to disk replicas, preferably in different buildings
- Long term archives may go on tape
- Don't look back

Data retention policies can exceed the lifespan of the organization

All systems fail eventually

Three major categories of data

User space (code, reports)

Data archive / warehouse

Fast attached 'scratch' space

cdwan@bioteam.net

Power for 5PB and 5,000 core cluster

The Ugly Physical

- Power and cooling issues will stop your project cold
- Timelines for facilities work start at six months

cdwan@bioteam.net

Conclusions

2011: Data Tsunami seems to be manageable

Data lifecycle management, accessibility, and mining are the current challenges

Cloud / scriptable infrastructure on demand is a game changer

It's going to be an interesting decade.

cdwan@bioteam.net