High Performance Computing Facility

Center for Health Informatics and Bioinformatics

Accelerating Scientific Discovery and Innovation in Biomedical Research at NYULMC through Advanced Computing

Efstratios Efstathiadis, Ph.D.

Technical Director

NYU Langone Medical Center

Center for Health Informatics and Bioinformatics

NYU Langone MEDICAL CENTER

NYU Langone Medical Center

- One of nation's premier centers for excellence in clinical care, biomedical research and medical education.
- The Medical Center's tri-fold mission is to serve, discover and educate.
- NYULMC Consists of:
  - NYU School of Medicine (a division of New York University)
  - NYU Hospitals Center
- The NYU Hospitals Center is composed of 4 hospitals:
  - Tisch Hospital
  - Rusk Institute of Rehabilitation Medicine
  - NYU Hospital of Joined Diseases
  - Clinical Cancer Center
- Located in the center of Manhattan, NYC
- 17,000 employees
- www.nyulmc.org

NYU Langone
MEDICAL CENTER

Center for Health Informatics and Bioinformatics

Mission: To catalyze transformative changes in biomedicine through breakthrough computational methodological research, best practices services, state of the art infrastructure and cutting edge education.

- Informatics Research and Education
  - Computational Causal Discovery methods in biomedicine
  - Next Generation Sequencing informatics
  - Information Retrieval
  - Computational Proteomics
  - Microbiomics
  - Graduate Training Program in Biomedical Informatics
  - …..
- Service and Infrastructure
  - HPC: data storage and computing
  - Research Enterprise Data Warehouse
  - ….
- Operational and applied Collaborative Projects

NYU Langone MEDICAL CENTER

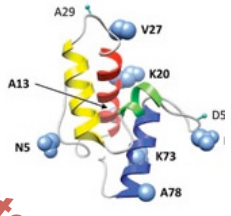
MiSeq

HiSeq 2000

GA IIx

Roche/454 FLX

RNAi



Proteomics

Leica SCN400F

Histopathology

Genome Technology Center

Data Storage

Microscopy

External Data Sources

S.O.L.V.E.

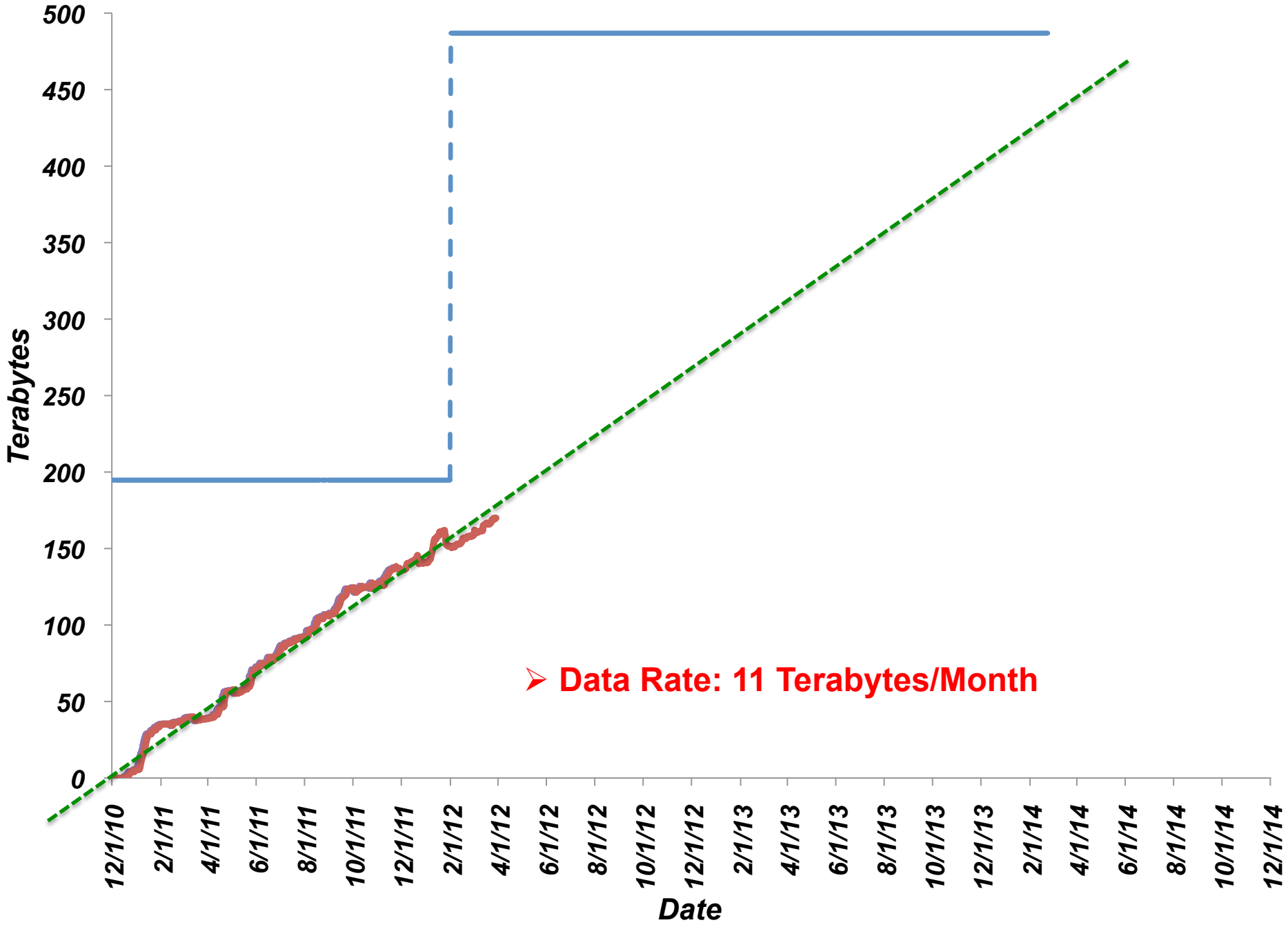
THE CANCER GENOME ATLAS

HMP

NYU Langone MEDICAL CENTER

Simulation & Analysis

Asclepius HPC Linux Cluster


500
450
400
350
300
250
200
150
100
50
0
Terabytes
12/1/10
2/1/11
4/1/11
6/1/11
8/1/11
10/1/11
12/1/11
2/1/12
4/1/12
6/1/12
8/1/12
10/1/12
12/1/12
2/1/13
4/1/13
6/1/13
8/1/13
10/1/13
12/1/13
2/1/14
4/1/14
6/1/14
8/1/14
10/1/14
12/1/14
Date
➢ Data Rate: 11 Terabytes/Month

HPC Challenges in Next Generation Sequencing (NGS)

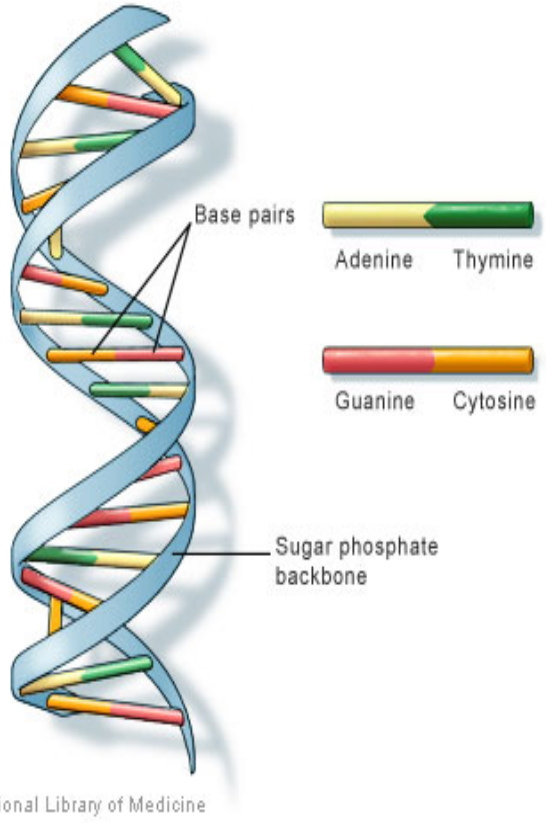
DNA Sequencing: The process of determining the exact order of the 3 billion nucleotides or bases, **A** (Adenine), **G** (Guanine), **C** (Cytosine), and **T** (Thymine) that make up a DNA molecule.

Next Generation Sequencing

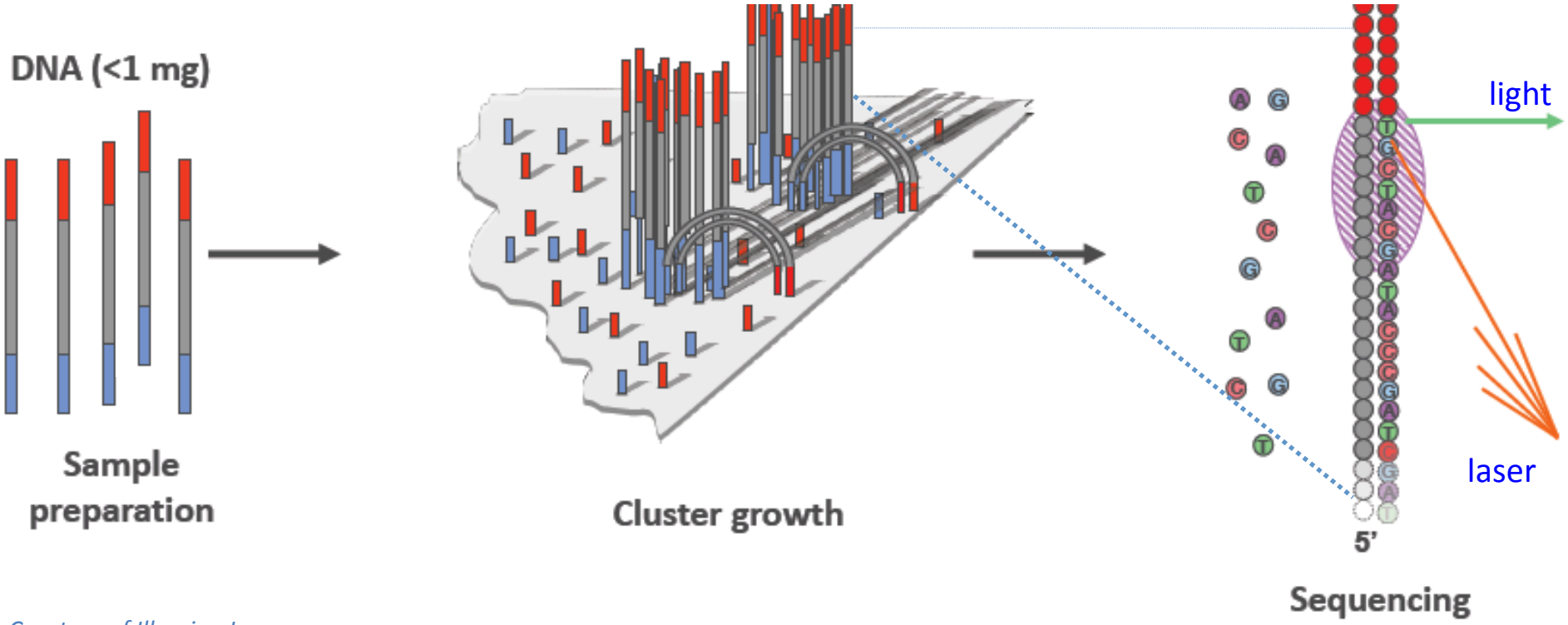
Massively Parallel
High Throughput
Low cost per base

February 2001:
Ten-year international effort produced 22.5 x 10^9 bases of DNA sequence, costing $3B

February 2011:
One NGS instrument produces 60 x 10^9 bases per day, costing several thousand dollars



Sequencing By Synthesis



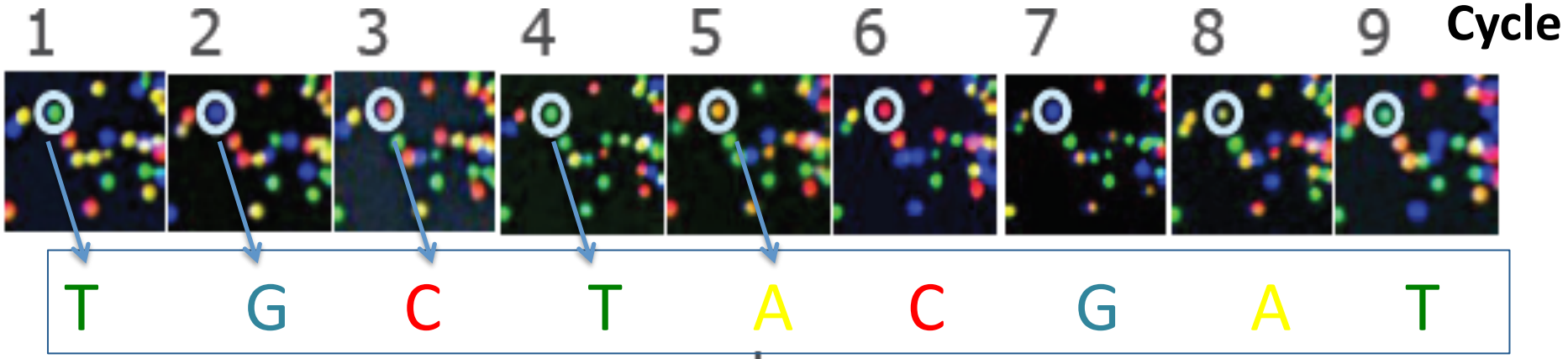
Courtesy of Illumina Inc.

High Throughput Sequencing ➔ High Data Throughput ➔ Dramatic increase in data storage and computing requirements

Dramatic Increase in Data Storage and Computing Requirements

Image Acquisition

Very Precise, High Resolution imaging: millions of images (.tiff) ➔ Terabytes of data



Courtesy of Illumina Inc.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EAOD@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Gigabytes of Text Output

Millions of DNA reads (Billions of Nucleotides and associated quality scores)

NGS Data Output

Whole Human Genome Sequencing requires a ~30x coverage

- **Uneven Coverage** - Poisson distribution of small DNA reads
- **Sequencing Errors** - machine/chemistry (~ 1% => 30Mbp)
- **Systematic Biases** - some regions are harder to sequence
- **Alignment Problems** - gaps, repeats, etc.
- **Quality Factor** - additional data/metadata

Making Sense of the Data

Raw Data ➔ Genetic Variation ➔ Biological Function

- Where in the Genome did each of the reads originate from?
- Identify Genetic Variations: How is the individual different from the "idealized" individual represented by the reference or other genomes?

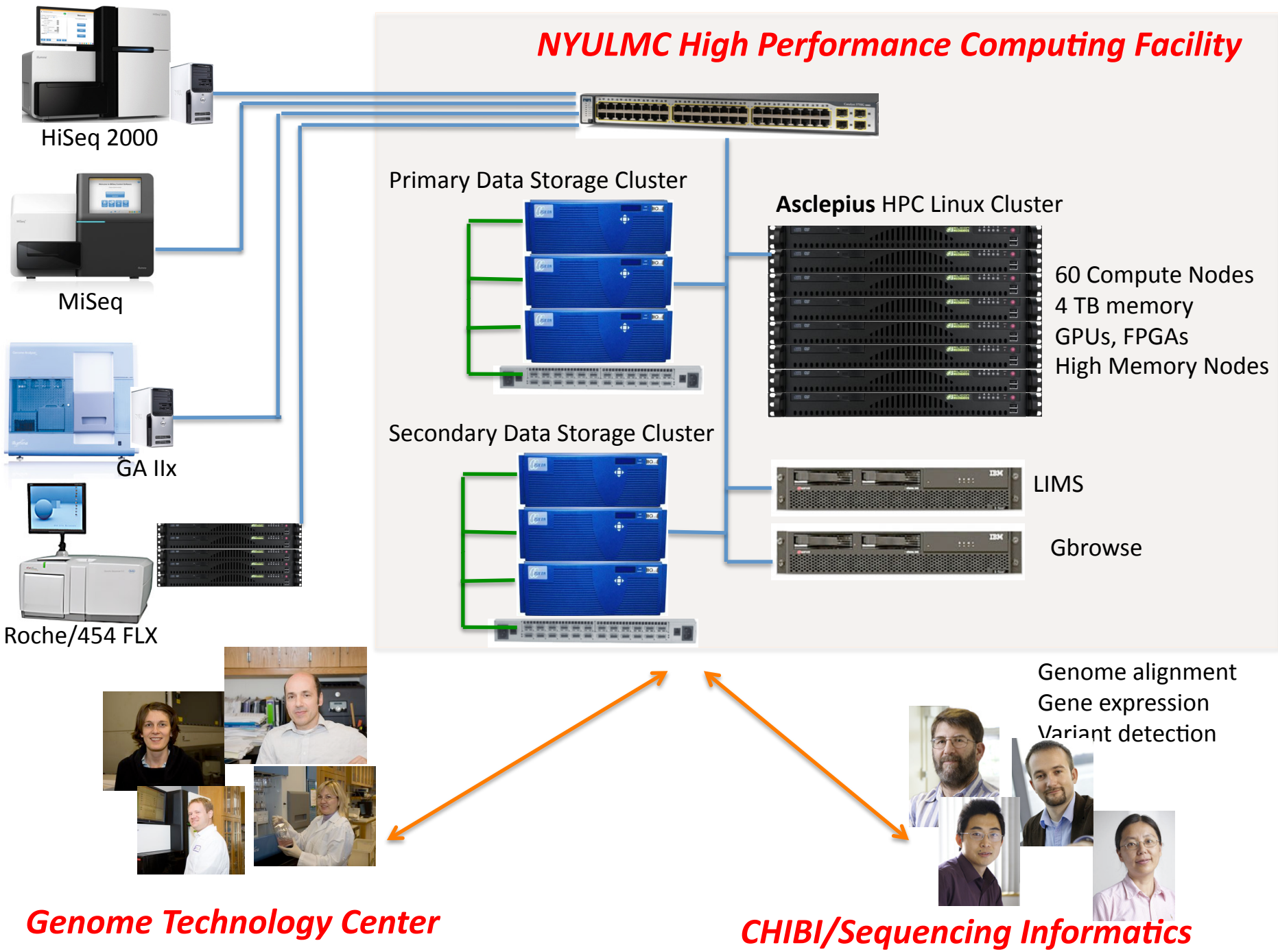
Sequence Alignment: Computationally and Data Intensive task

- A *Moving Target*:
  - Changes in sequencing technology
  - Shifting of biologists' interests.

- *Large volumes of sequencing data*:
  - Fast, Sophisticated algorithms (often trading accuracy for speed)
  - Scalable tools (multi-threaded/multi-process)
  - Efficient memory use (GPUs seem to be a good fit!)

- Plethora of unsupported, serial tools, developed from "scratch" by researchers.

- No commercial commitment


NYULMC High Performance Computing Facility
HiSeq 2000
MiSeq
GA IIx
Roche/454 FLX
Primary Data Storage Cluster
Secondary Data Storage Cluster
Asclepius HPC Linux Cluster
60 Compute Nodes
4 TB memory
GPUs, FPGAs
High Memory Nodes
LIMS
Gbrowse
Genome alignment
Gene expression
Variant detection
Genome Technology Center
CHIBI/Sequencing Informatics

Cloud Computing in Next-Gen Sequencing

(1) Infrastructure as a Service (IaaS) and Data Sharing

- Self-Service, easy to deploy, ability to scale, availability of resources (reference genomes, public databases, etc.)

(2) Cloud-enabled services made available over the internet

- Ready to use services deployed on the cloud with predefined (and configurable) pipelines, no need for local installation.
- Examples: Galaxy, UCSC Genome Browser, etc.

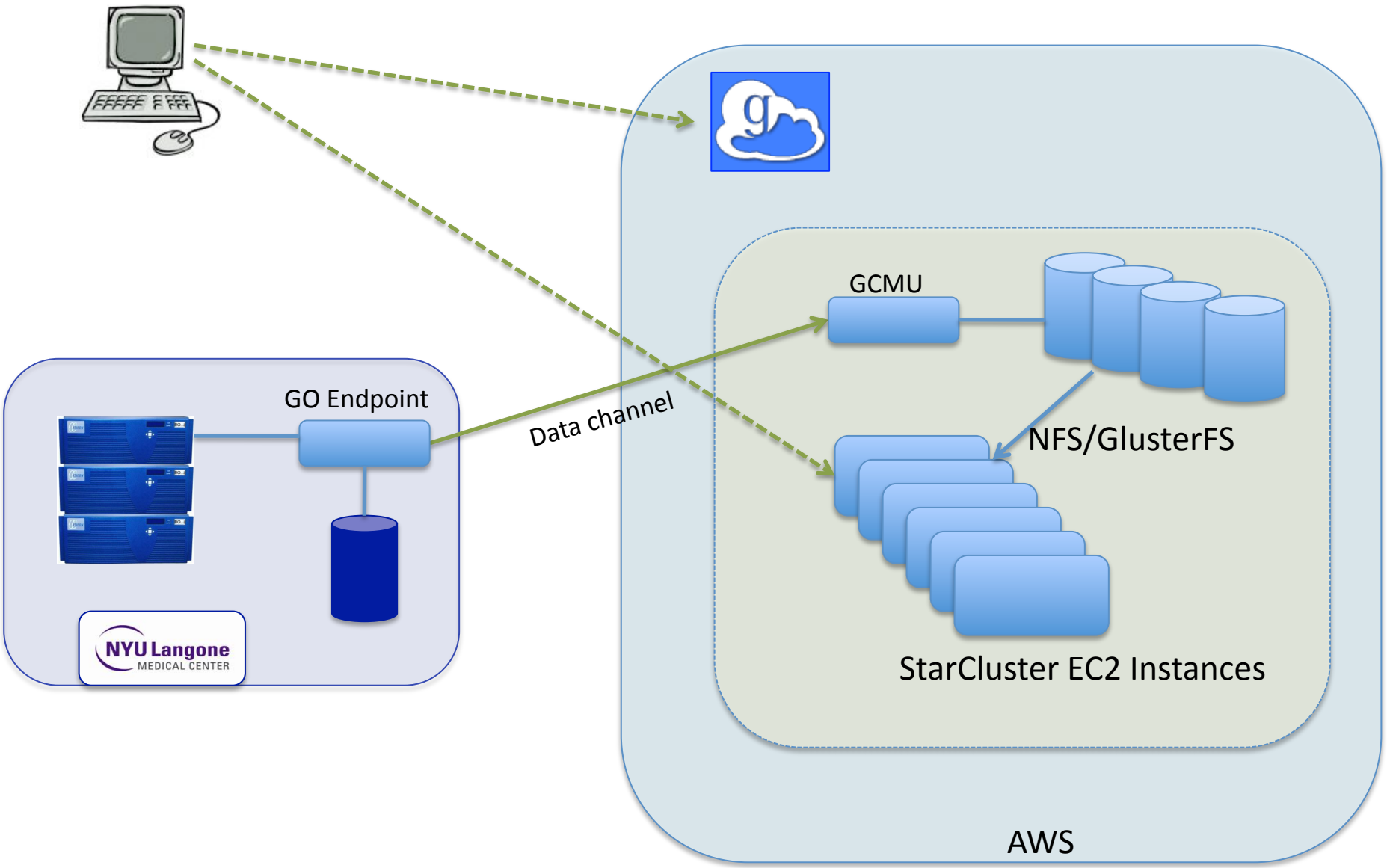
(3) NGS as a Service: Commercial pipelines use cloud resources to analyze customer NGS data. Examples: DNAnexus, Genome Quest, etc.

(4) Ready-to-use, scalable pipelines.

- Pre-configured pipelines that use virtualization and Map Reduce to expedite NGS data analysis
- Examples: CloudBurst, CloudBLAST, cloVR, crossbow, etc.

- Biggest Challenge:
  - Uploading large data sets to the cloud, in an efficient, cost-effective, fault-tolerant, easy to use way.
- Typical data set: A few hundred GigaBytes of FASTQ or BAM files. It should take ~ 1 hr over Gbps

- scp, sftp, http: known limitations in WAN
- Shipping disks: huge latencies, low automation potential, error prone
- GridFTP: Challenging GSI infrastructure


GCMU
GO Endpoint
Data channel
NFS/GlusterFS
StarCluster EC2 Instances
NYU Langone
MEDICAL CENTER
AWS

- GO enabled data intensive analysis on the cloud.
  - Performance (a factor of 10 improvement over scp/sftp)
  - Minimum Firewall Requirements
  - Eliminates unnecessary duplication of large data sets
  - No need to VPN/on-site login to initiate transfers
  - Cost-effective

- Increasing Need for Data Sharing solutions:
  - New York Genome Center
  - NYC Workshop on HPC for Biomedical Research coming up in May