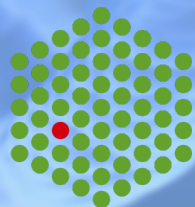


# Big data and HPC on-demand: Large-scale genome analysis on Helix Nebula – the Science Cloud

Rupert Lueck  
Head of IT Services, EMBL Heidelberg  
Helix Nebula General Assembly  
ESA/ESRIN Frascati, 16 January 2013

EMBL



**HELI  
NEBULA**  
THE SCIENCE CLOUD

# EMBL: European Molecular Biology Laboratory



- Intergovernmental Research Organization
- Supported by 20 Member States (+1 associated: )
- One of the world's foremost life science institutions
- EIROforum member
- 1500 staff  
>70 nationalities

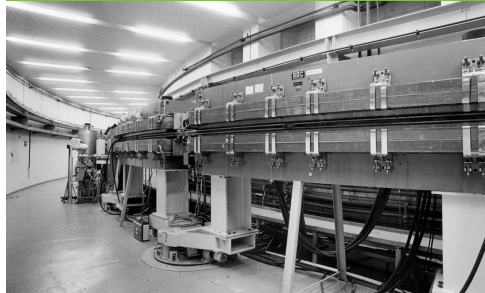
# The Five Branches of EMBL

## Heidelberg



Basic Molecular Biology  
Research  
Main Lab / Headquarters

## Hamburg



Structural Biology  
DESY

## Hinxton



European Bioinformatics  
Institute (EBI)  
Sanger Centre

## Grenoble



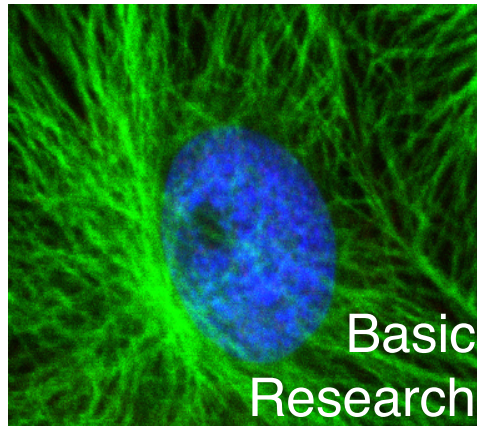
Structural Biology  
ILL, ESRF, IBS, UVHCI

## Monterotondo

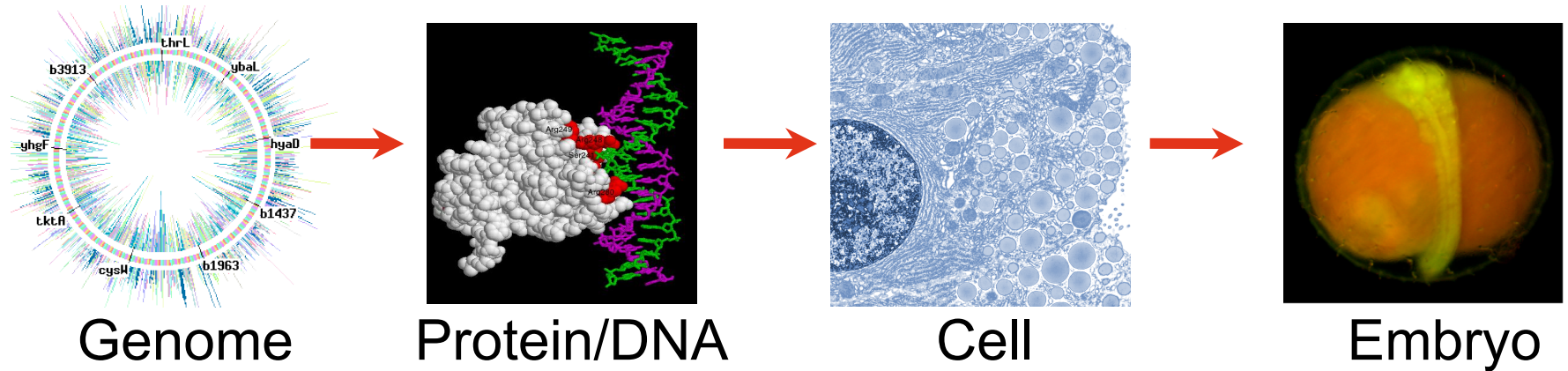


Mousebiology  
CNR, EMMA

# EMBL's Missions



# Systems Biology: From Molecules to Organisms



Development



Organisms



Complexity

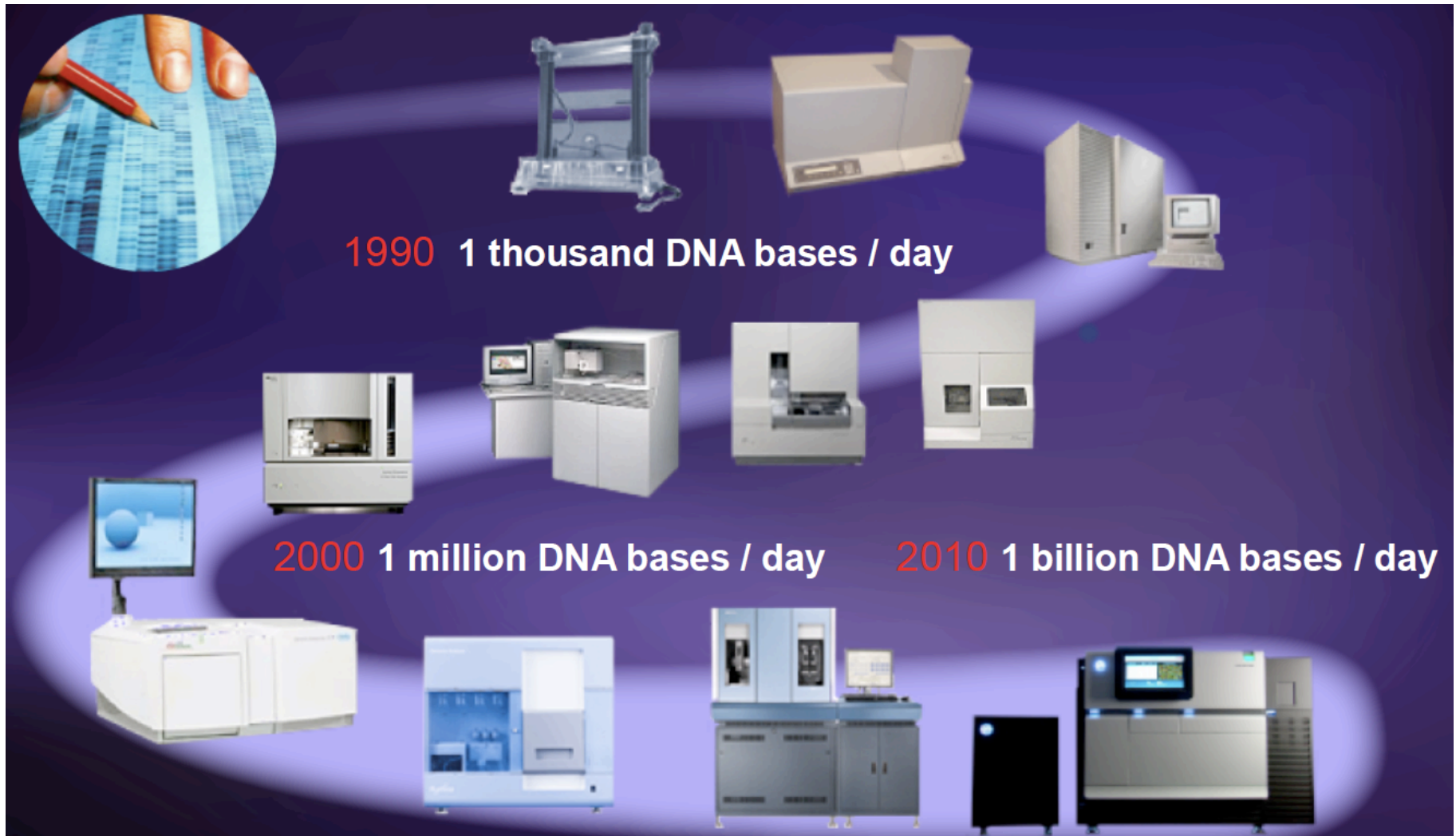


Aging

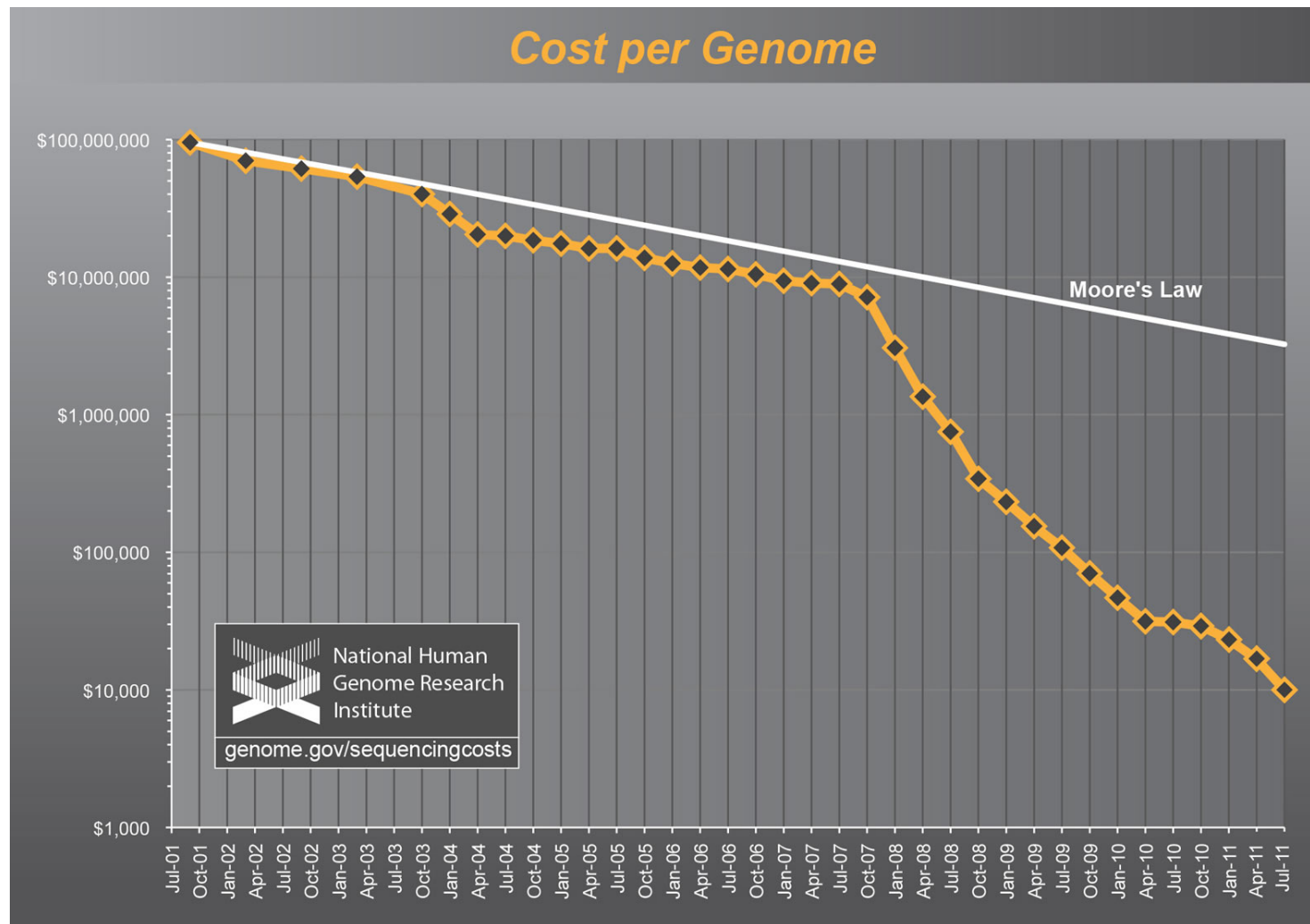


Disease

# Next Generation Sequencing (NGS) Revolution



# Cost of Sequencing Decreasing Rapidly



# Genomic Sequencing is Now an Affordable Solution

Academic  
Research  
Groups

Medical



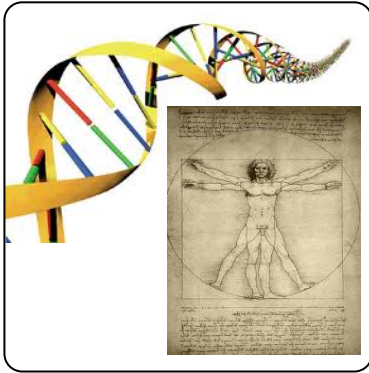
A composite image showing two web pages. The top-left page is the '1000 Genomes' website, featuring a dark header with the title '1000 Genomes' and a subtitle 'A Deep Catalog of Human Genetic Variation'. It includes a navigation bar with 'Home', 'About', and 'Data'. The main content area has sections for 'ABOUT THE 1000 GENOMES PROJECT' and 'PROJECT OVERVIEW'. The bottom-left page is the 'GENOME 10K' website, with a light blue header and a large blue DNA double helix graphic. It features a navigation bar with 'Home', 'Database &amp; Species Lists', 'News', 'Events', 'Publications', 'Participants', and 'For G10K Organizers (restricted)'. The main content area has a search bar and a section titled 'Genome 10K Project' with a description of the project's goals. The right side of the GENOME 10K page has a 'Join us' button and a link to 'Become a G10K affiliate'.

Genomic sequencing is  
now an affordable solution

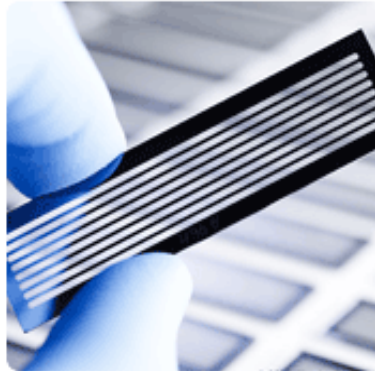
**however ...**

# Read the Sequence to Study the Organism

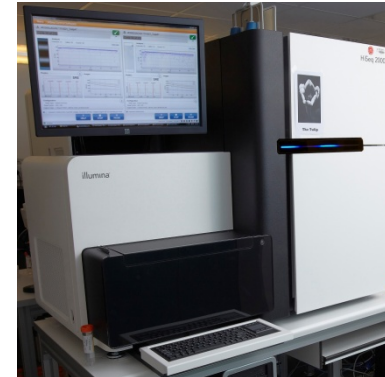
Extract DNA



Prepare



Sequence



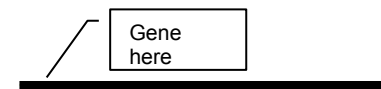
Lab

Assemble



Annotate

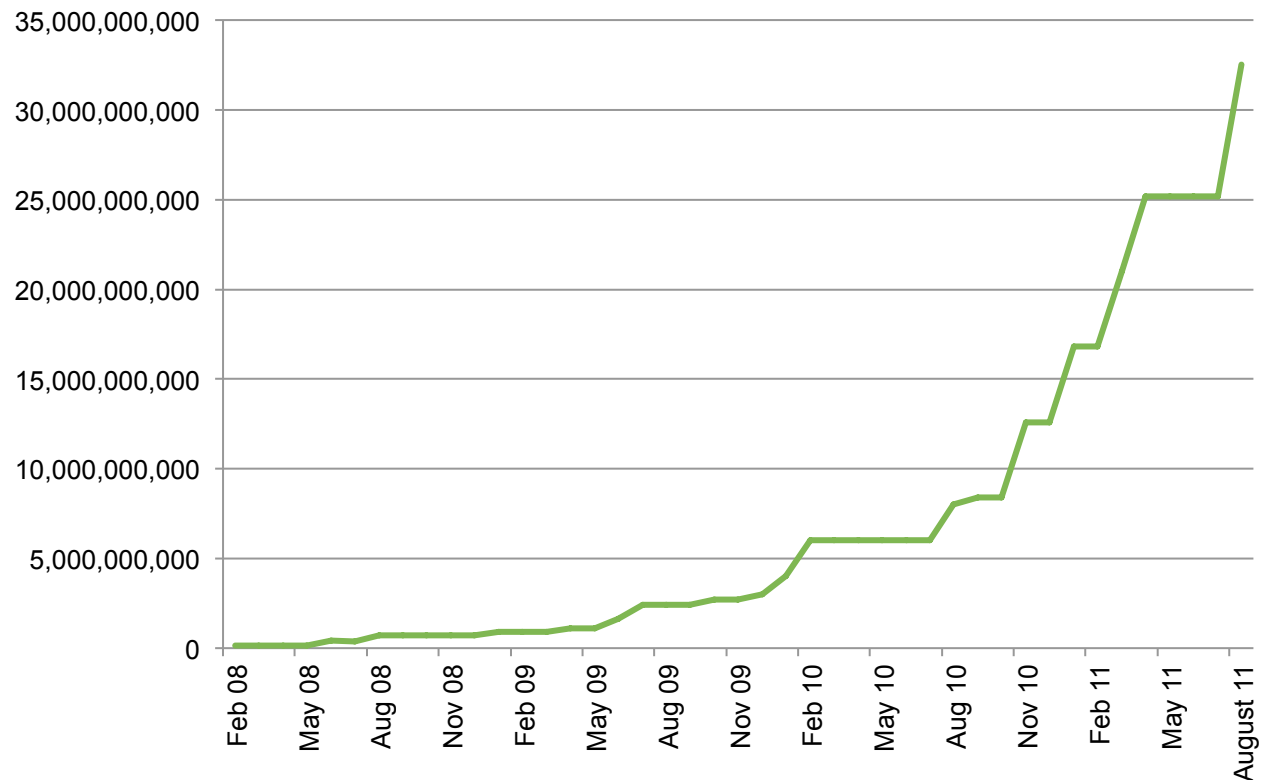
In Silico



Requires Computing Infrastructure & Expertise

# Problem - Technology Explosion with NGS

**Bases Sequenced / Sample / Run @ EMBL  
(Illumina)**



# Sequence Production & IT Infrastructure at EMBL

4 x Illumina HiSeq2000



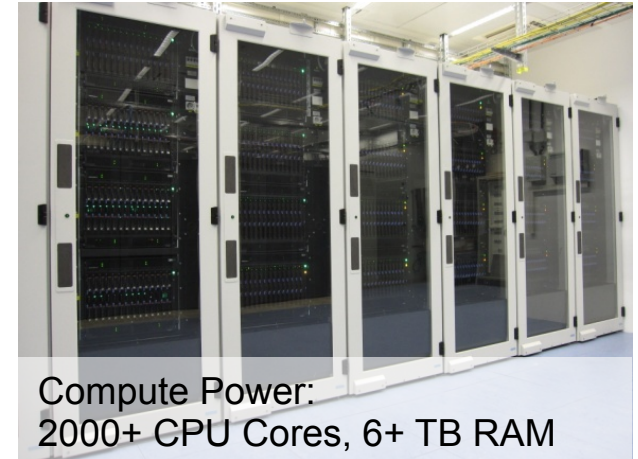
1 x MySeq



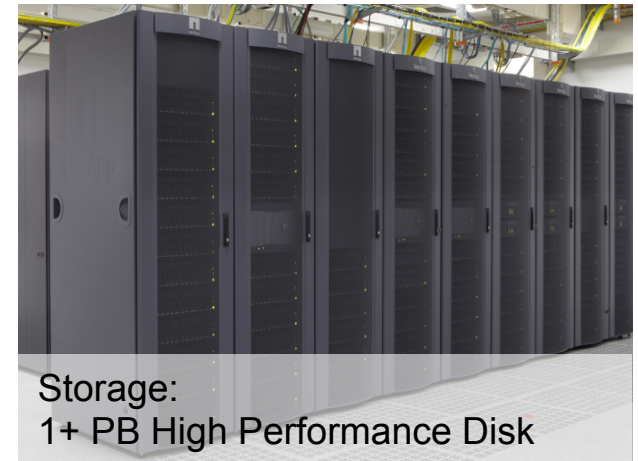
1 x Ion Torrent



25+ TB data  
each week



Compute Power:  
2000+ CPU Cores, 6+ TB RAM

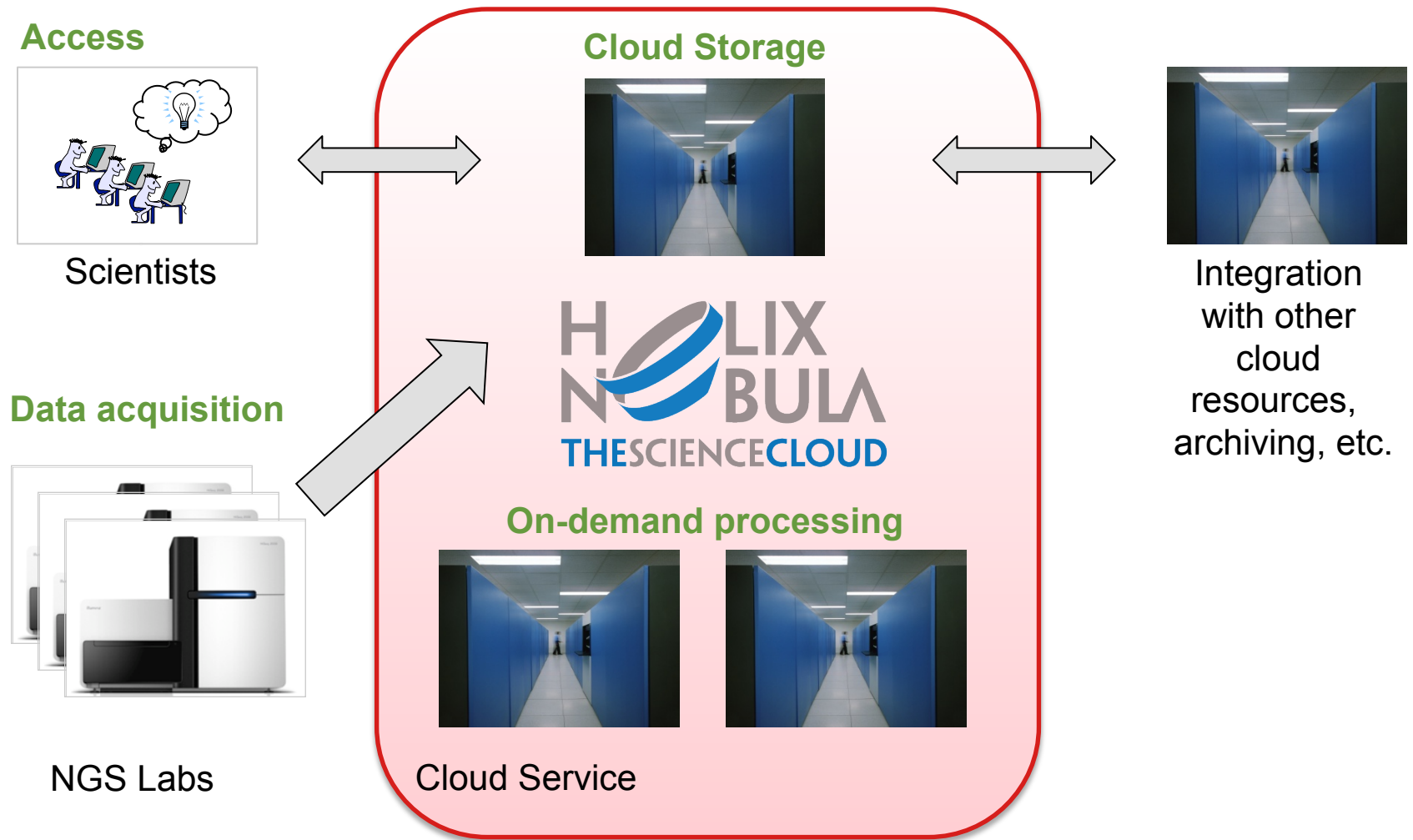


Storage:  
1+ PB High Performance Disk

# NGS - The Big Picture

- ~ 8.7 million species in the world (estimate)
- ~ 7 billion people
- Sequencers exist in both large centres & small research groups
- 200+ Illumina HiSeq sequencers in Europe alone
  - capacity to sequence 1600 human genomes / month
- Largest centre: Beijing Genomics Institute (BGI)
  - ~140 HiSeq
- ~1500 Hiseq devices worldwide today
  - 3-6 PB / day
  - 1.1 – 2.2 ExaBbytes / year

# EMBL Flagship: Large-scale Genome Analysis



# Proof of Concept

- Multiple Cloud providers
  - ATOS / Sixsq
  - CloudSigma
  - T-Systems
- Each tested steps with increasing complexity
- Major software components to test

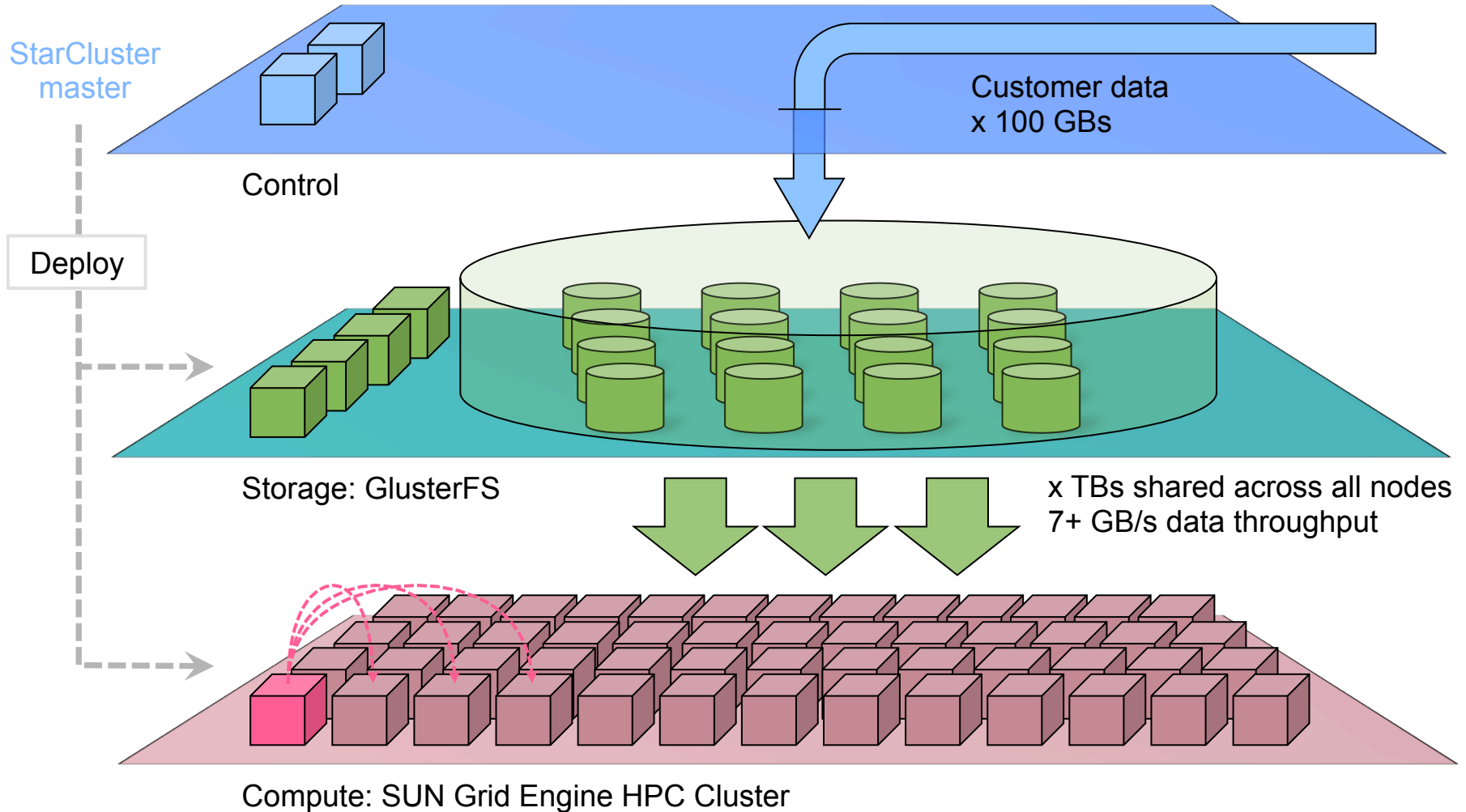
Assembly pipeline	SGA by Simpson, JT & Durbin, R <a href="http://genome.cshlp.org/content/22/3/549.long">http://genome.cshlp.org/content/22/3/549.long</a>
Annotation pipeline	Ensemble
Shared File system	e.g. glusterFS
StarCluster	<a href="http://star.mit.edu/cluster/">http://star.mit.edu/cluster/</a>

# StarCluster & Sun Grid Engine

## Dynamic cluster provisioning

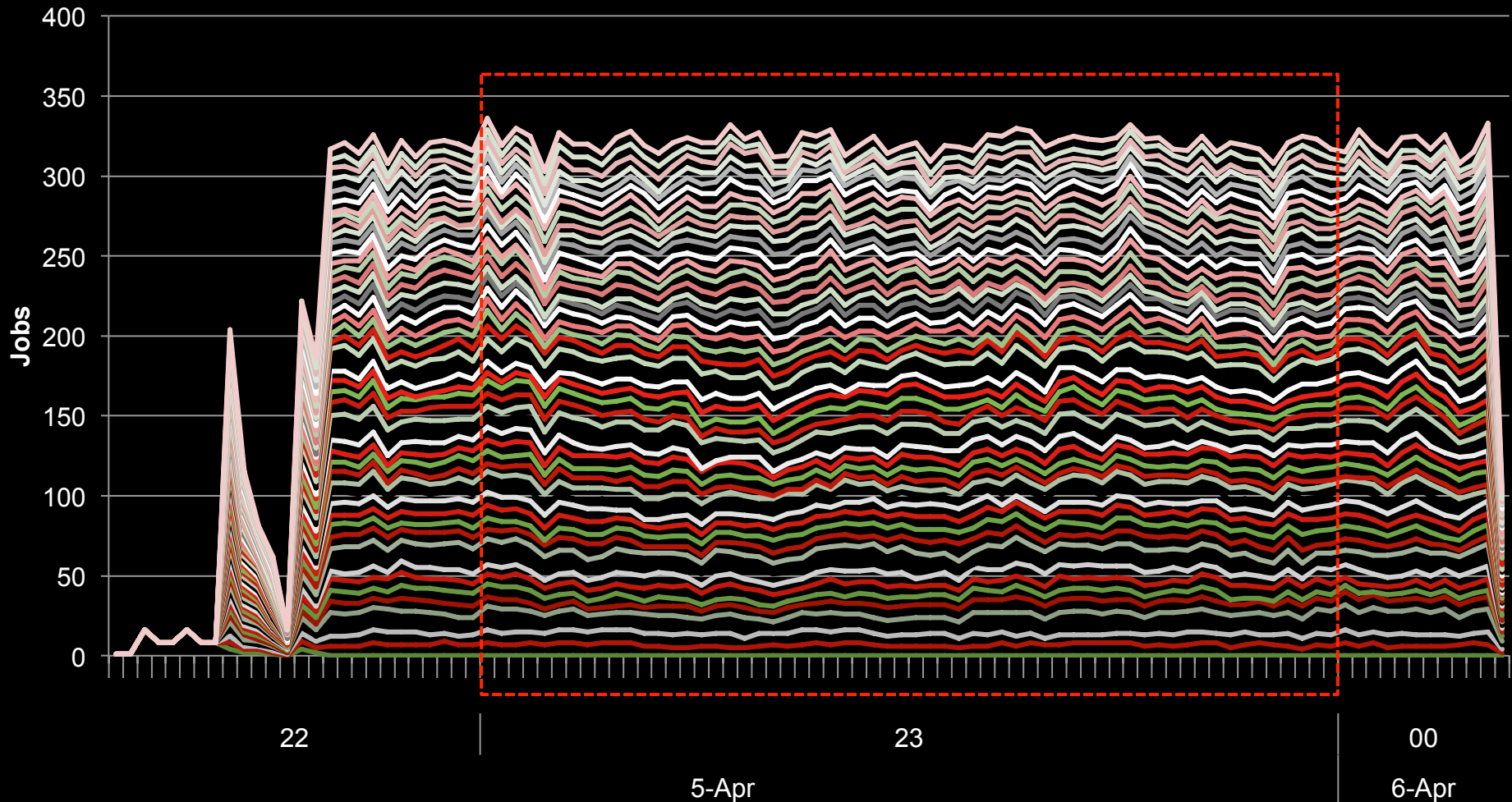
- StarCluster – Dealing with the Fluctuating Workload
  - Manages provisioning of images and setting up of cluster
  - Requires sets of EC2 APIs to work
  - It monitors the number of jobs in the queue and launches more instances
  - Terminates them when no longer required
- Sun Grid Engine
  - Single image running in two modes – master/worker
  - Post-launch configuration

# EMBL Dynamic Cloud Architecture



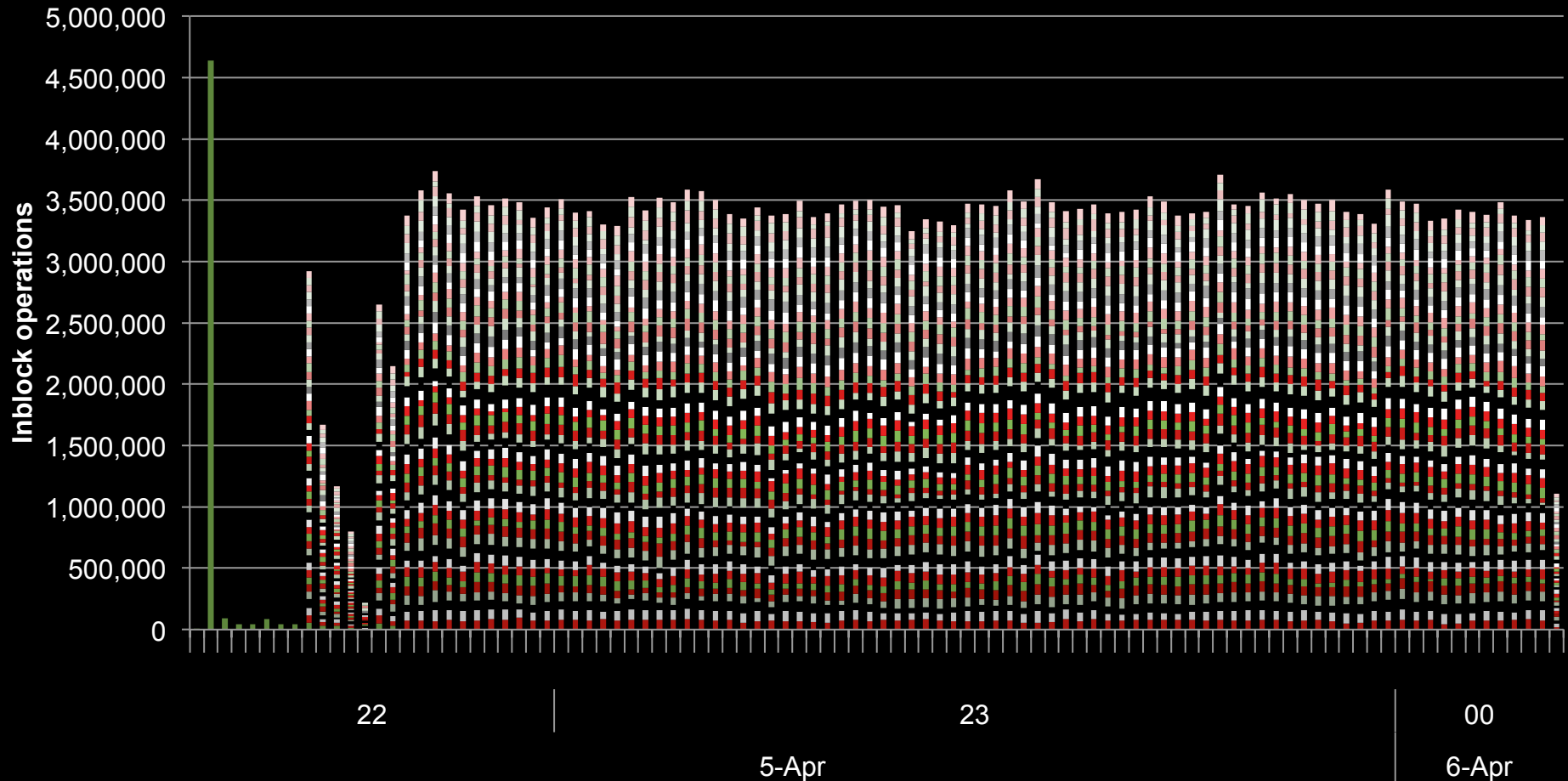
# Sun Grid Engine cluster throughput

20.000 annotation jobs / h on 50 nodes



# GlusterFS throughput

**60.000 inbound block I/Os / sec  
from annotation jobs on 50 nodes**




Successful tests of all vendors deployed so far

- StarCluster API integration
- auto-provision 50-node cluster setups
- real world large genome sequencing data
- 100,000s of jobs
- mix of quick parallel jobs and long running serial jobs
- glusterFS stability under high I/O levels
- Initial hurdles (e.g. image deployment, StarCluster integration, network setup) solved

# User Interface Genome Analysis





**Genome.analysis**

Messages 8

 albert@embl.de

Settings

Logout



**Configure Annotation Pipeline**

**Navigation**

File Manager


Initialise Cluster

Assembly Pipeline


Annotation Pipeline

Pipeline Status





**Bandwidth Transfer**





 %

**Disk Space Usage**


 304.44 / 8000 MB %

**Stats**

Jobs	Completed
 <b>21,501</b>	 <b>308</b>
 5%	 8%






Servers	Credit
 <b>40</b>	 <b>\$376</b>
 1%	 26%

**Right now**





**34**  Posts

**Bootstrap RNAseq Pipeline** Create pipeline Cancel

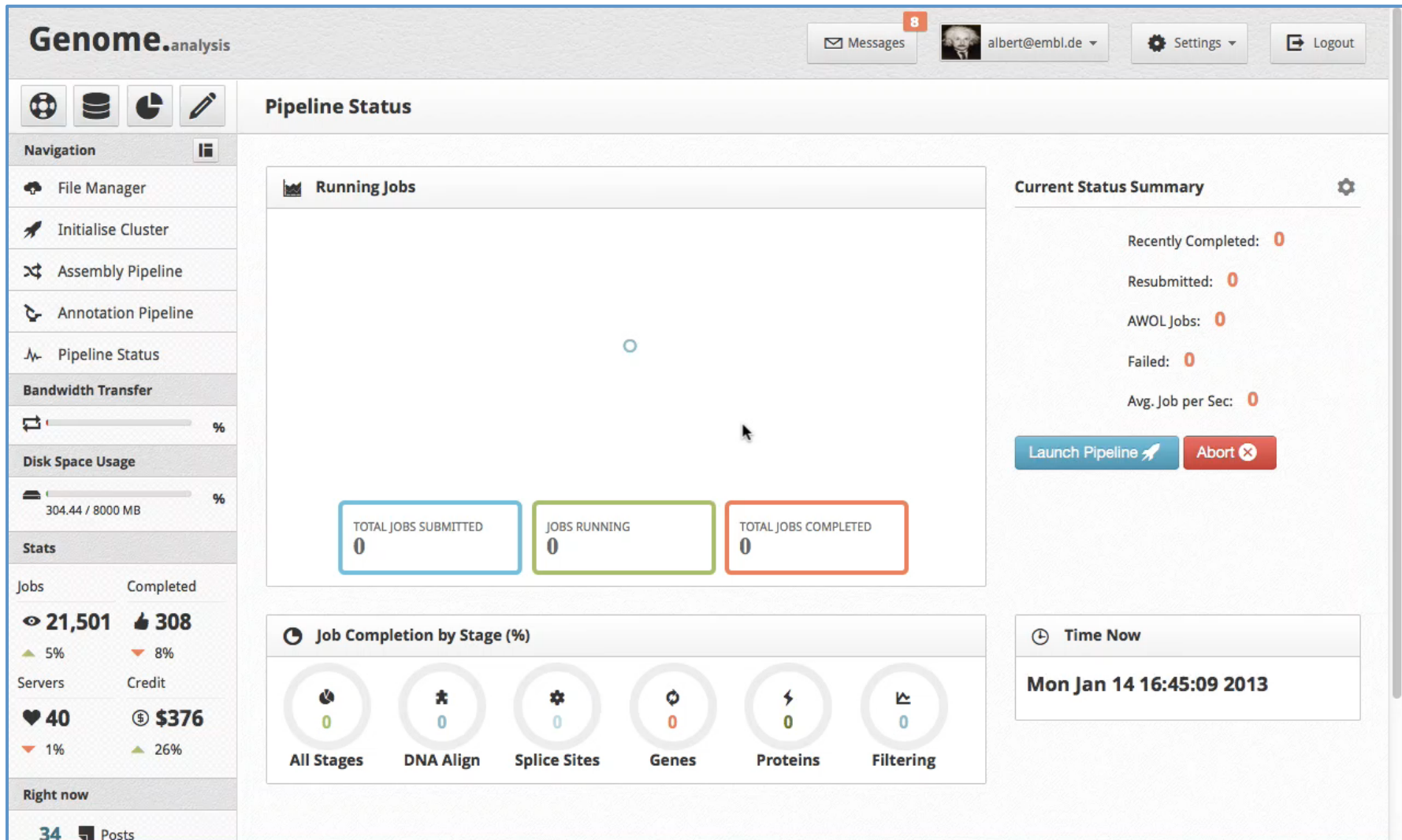
Time now: Mon Jan 14 16:54:09 2013

	0%
	0%
	0%
	0%
	0%

**Pipeline Bootstrap Status**

	<b>0/8</b>	FTP fetch Ensembl MySQL tables
	<b>0/32</b>	Creating Databases
	<b>0/6</b>	Config
	<b>0/3</b>	Installing Ensembl API

# User Interface Genome Analysis



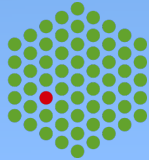
# Next steps

- Integration of EMBL flagship with first implementations of the Helix Nebula Blue Box
- Extensive testing of this federated cloud
- Feed back results to Helix Nebula requirements management
- Continue work on the UI
- Prepare for putting EMBL genome analysis pipeline into production in 2013

# Helix Nebula PoC Acknowledgements



EMBL



Michael Wahlers  
Jonathon Blake  
Tobias Rausch  
Jürgen Zimmermann  
Vladimir Benes  
Christian Boulin  
Rupert Lueck

EMBL- EBI

Stephen Keenan  
Paul Flicek

**AtoS**

**CloudSigma** 

**T** . . . **Systems** . . .

