Large-Scale Genome Analysis on Helix Nebula – the Science Cloud

Federated Big Data Processing On-Demand

Rupert Lueck | IT Services EMBL

Helix Nebula - The Science Cloud Public Event
CERN, Geneva
14 May 2014

EMBL 40 YEARS | 1974–2014

HELIX NEBULA THE SCIENCE CLOUD

EMBL: European Molecular Biology Laboratory

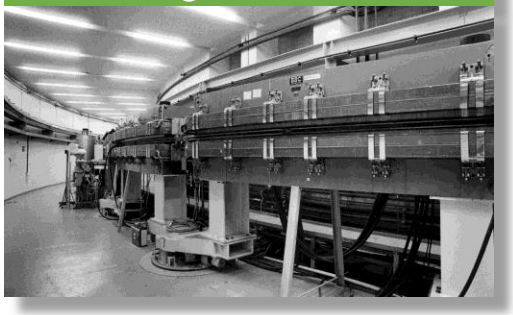
- Intergovernmental Research Organization
- Supported by 20 Member States (+1 associated:)
- One of the world‘s foremost life science institutions
- EIROforum member
- 1500 staff
  >70 nationalities

The Five Branches of EMBL

Heidelberg

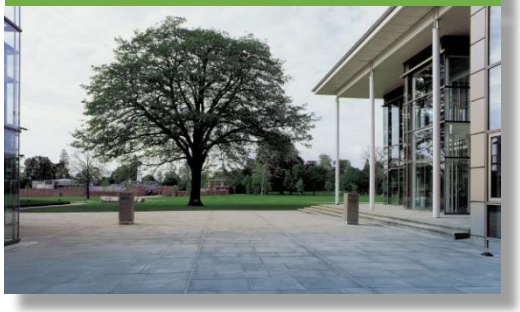
Basic Molecular Biology Research
Main Lab / Headquarters

Hamburg



Structural Biology
DESY

Hinxton



European Bioinformatics Institute (EBI)
Sanger Centre

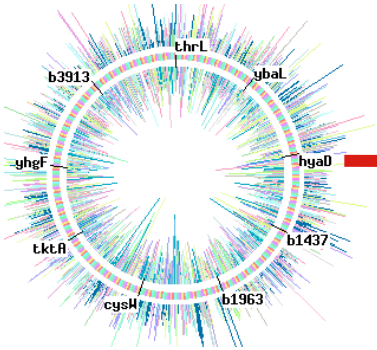
Grenoble

Structural Biology
ILL, ESRF, IBS, UVHCI

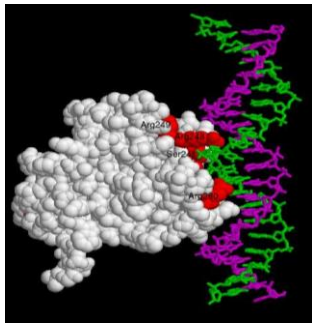
Monterotondo

Mousebiology
CNR, EMMA

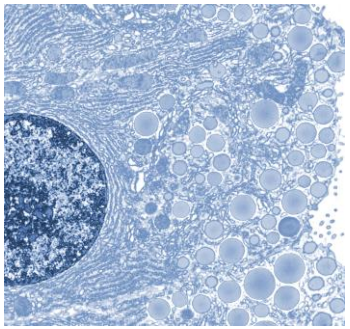
Information Biology: From Molecules to Organisms



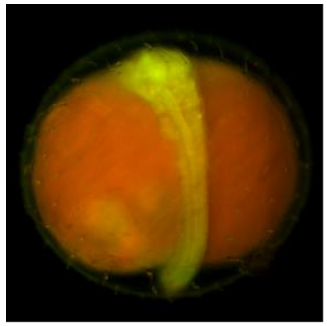
Genome



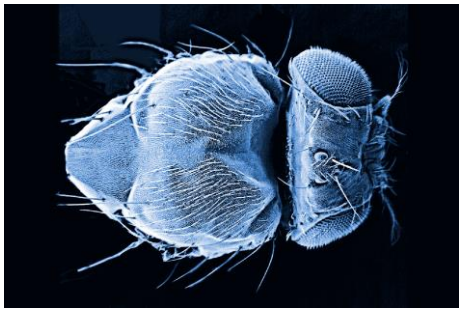
Protein/DNA



Cell



Embryo



Development

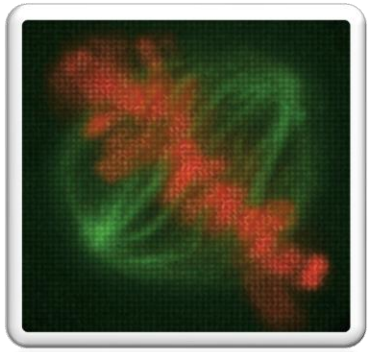
Organisms

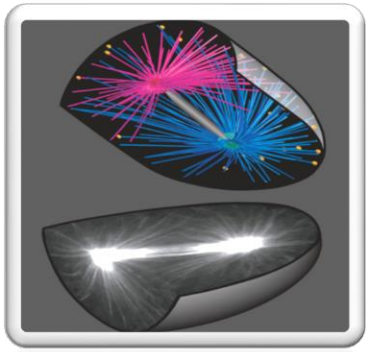
Complexity

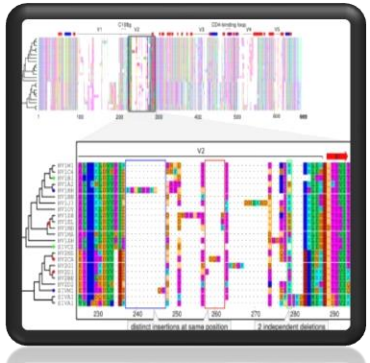
Aging

Disease

Why EMBL involves in Cloud Computing





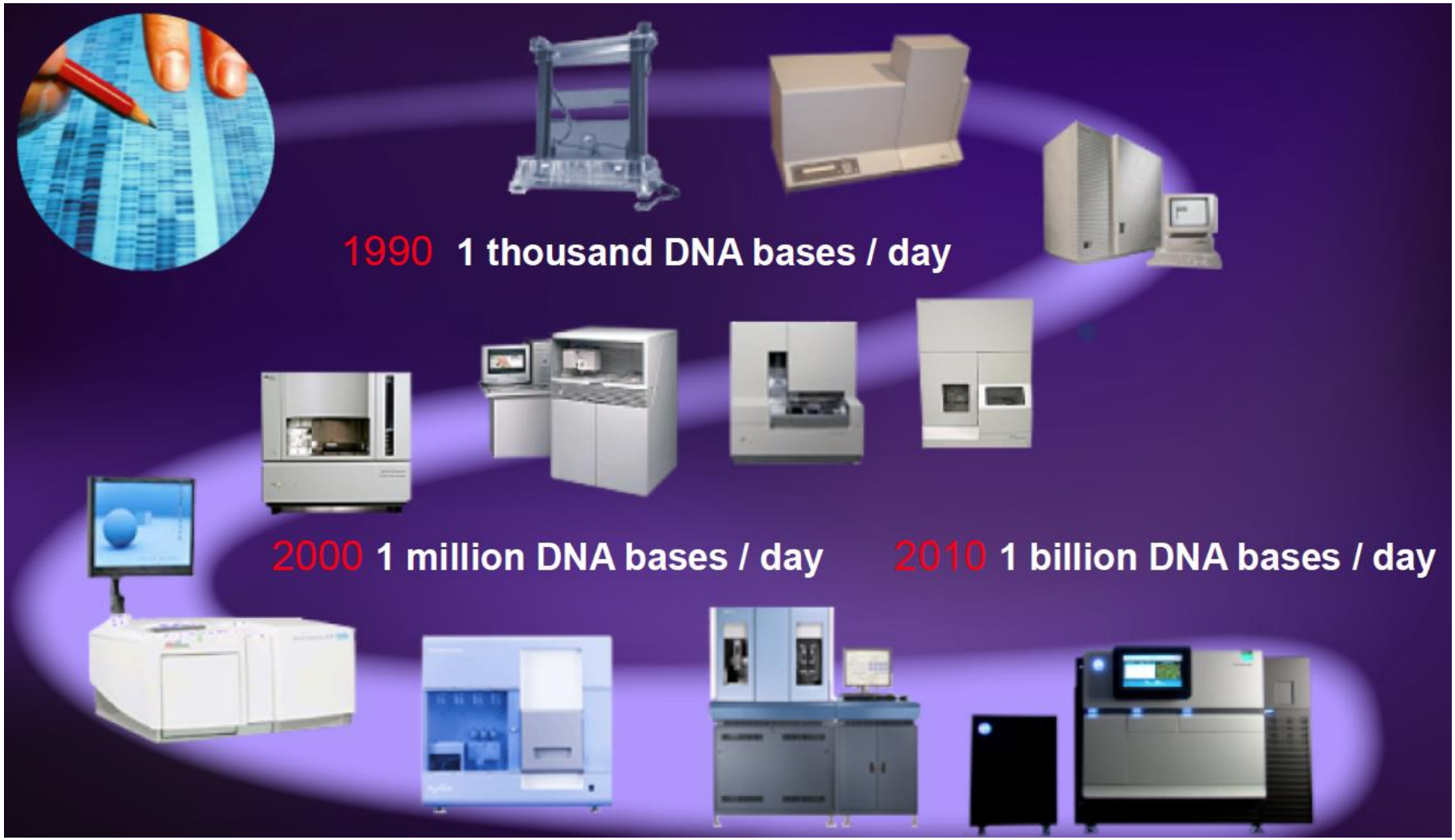


- Information Biology research at EMBL is data-driven
  - Key technologies: Imaging, Computational Biology, Next Gen Sequencing, Modeling
  - Generate tens of TeraBytes of data every week
  - Require HPC IT infrastructures to move, store, analyse and share data
  - Project activities vary over time

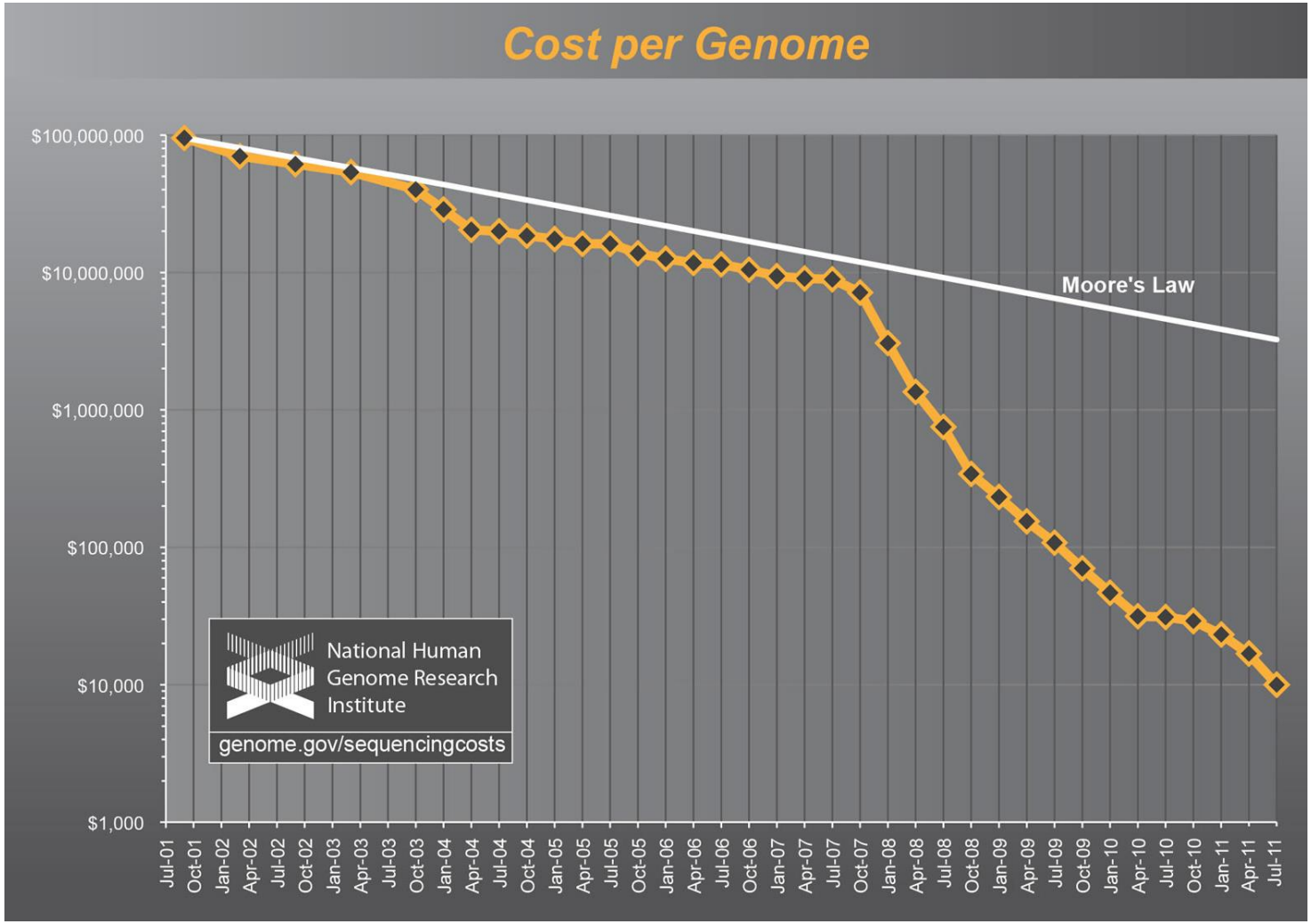
- EMBL studies Cloud computing to provision on-demand and elastic IT resources to European scientists within and outside of EMBL


EMBL Flagship Project
Genomic Assembly and
Annotation in the Cloud
EMBL
HELIX NEBULA
THE SCIENCE CLOUD

Next Generation Sequencing (NGS) Revolution



Cost of Sequencing Decreasing Rapidly



Genomic Sequencing is Now an Affordable Solution

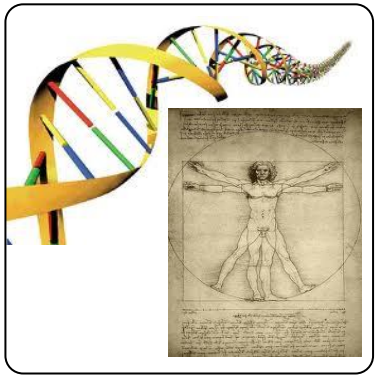
Academic
Rese
Grou
utical

Genomic sequencing is
now an affordable solution

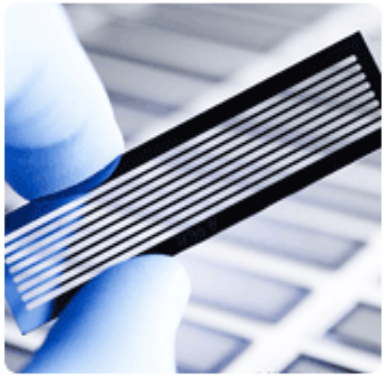
however ...

Read the Sequence to Study the Organism

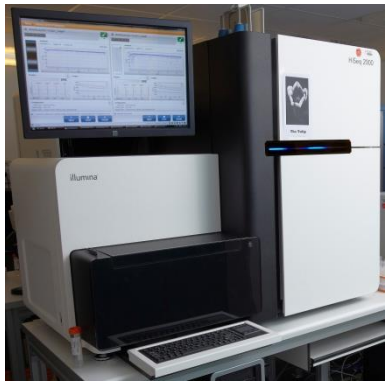
Extract DNA



Prepare



Sequence

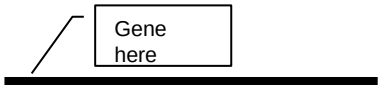


Lab

In Silico

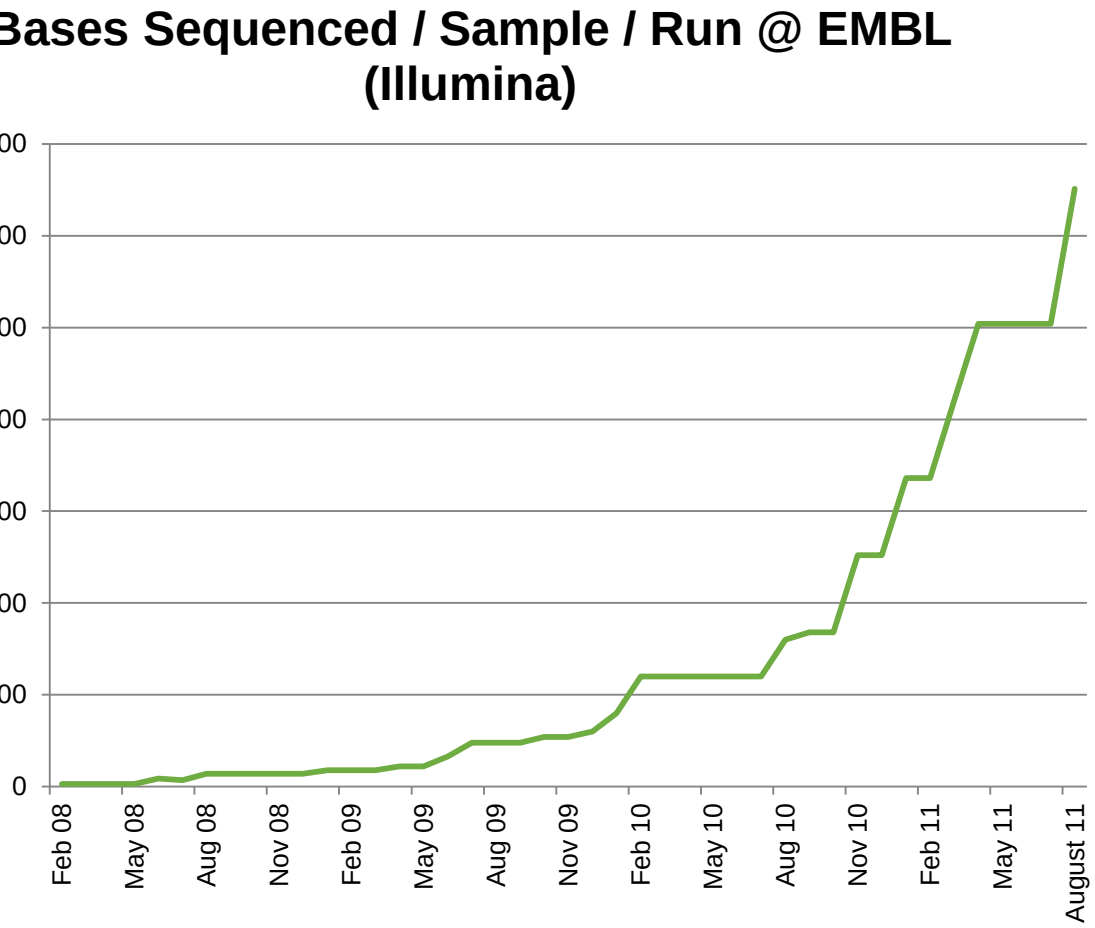
Assemble

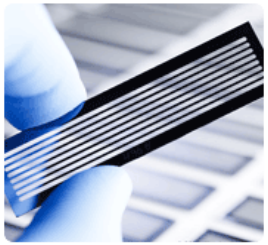
Annotate



Requires Computing Infrastructure & Expertise

Problem - Technology Explosion with NGS





Sequence Production & IT Infrastructure at EMBL

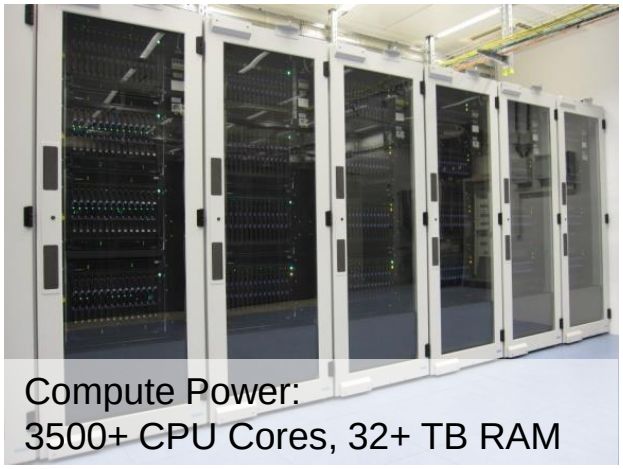
5 x Ilumina HiSeq2000 / 2500

1x NextSeq 500

1 x MiSeq

1 x Ion Torrent

30+ TB data each week



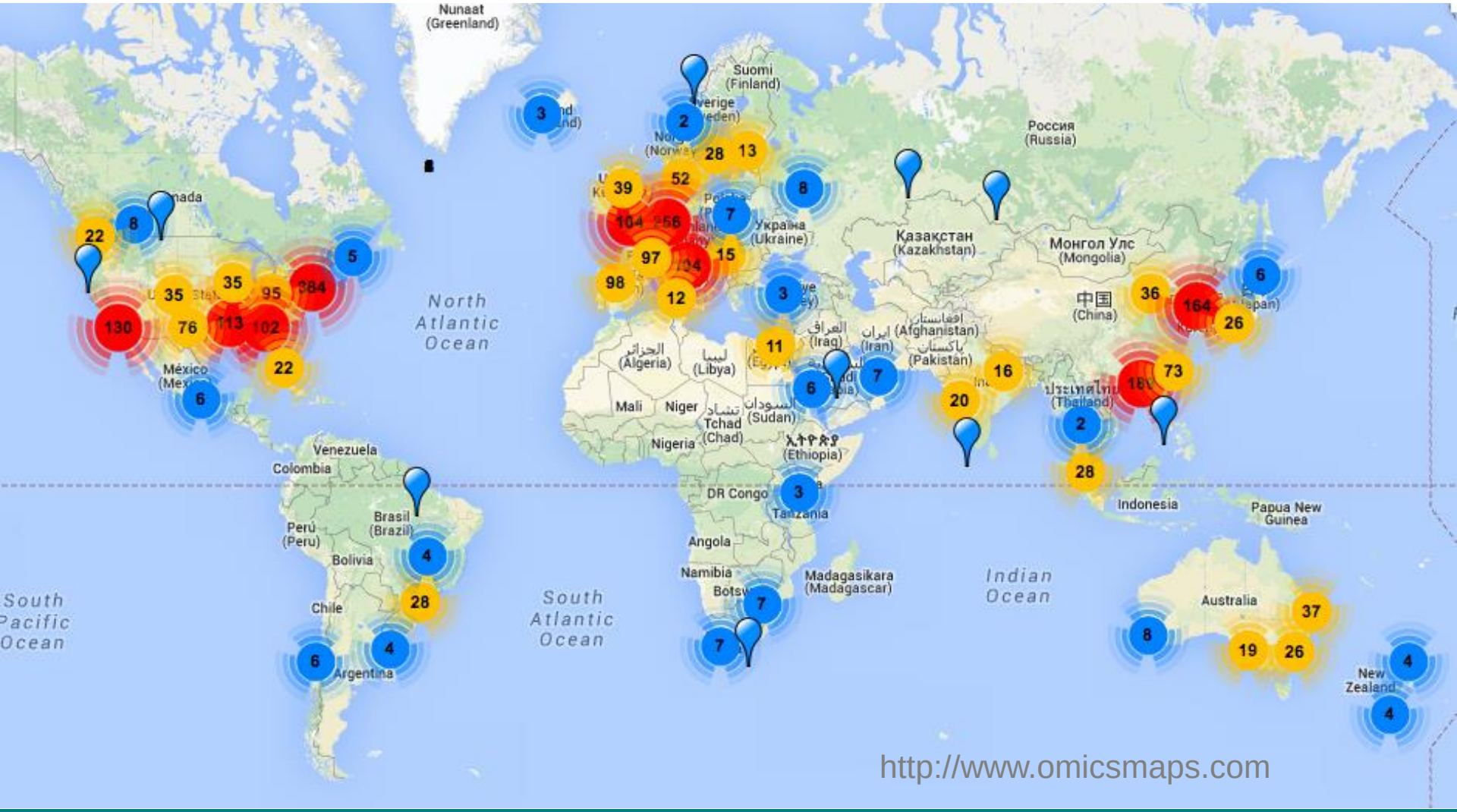
Compute Power:
3500+ CPU Cores, 32+ TB RAM

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+ Ilumina 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

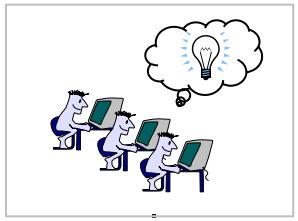
World Map of High-throughput Sequencers



EMBL 40 YEARS | 1974–2014

EMBL Helix Nebula Flagship Project

Access



Scientists

Data acquisition

NGS Labs

On-Demand
Large-Scale Genome Analysis

Whole-genome assembly &
genome annotation

HELIX NEBULA
THESCIENCECLOUD

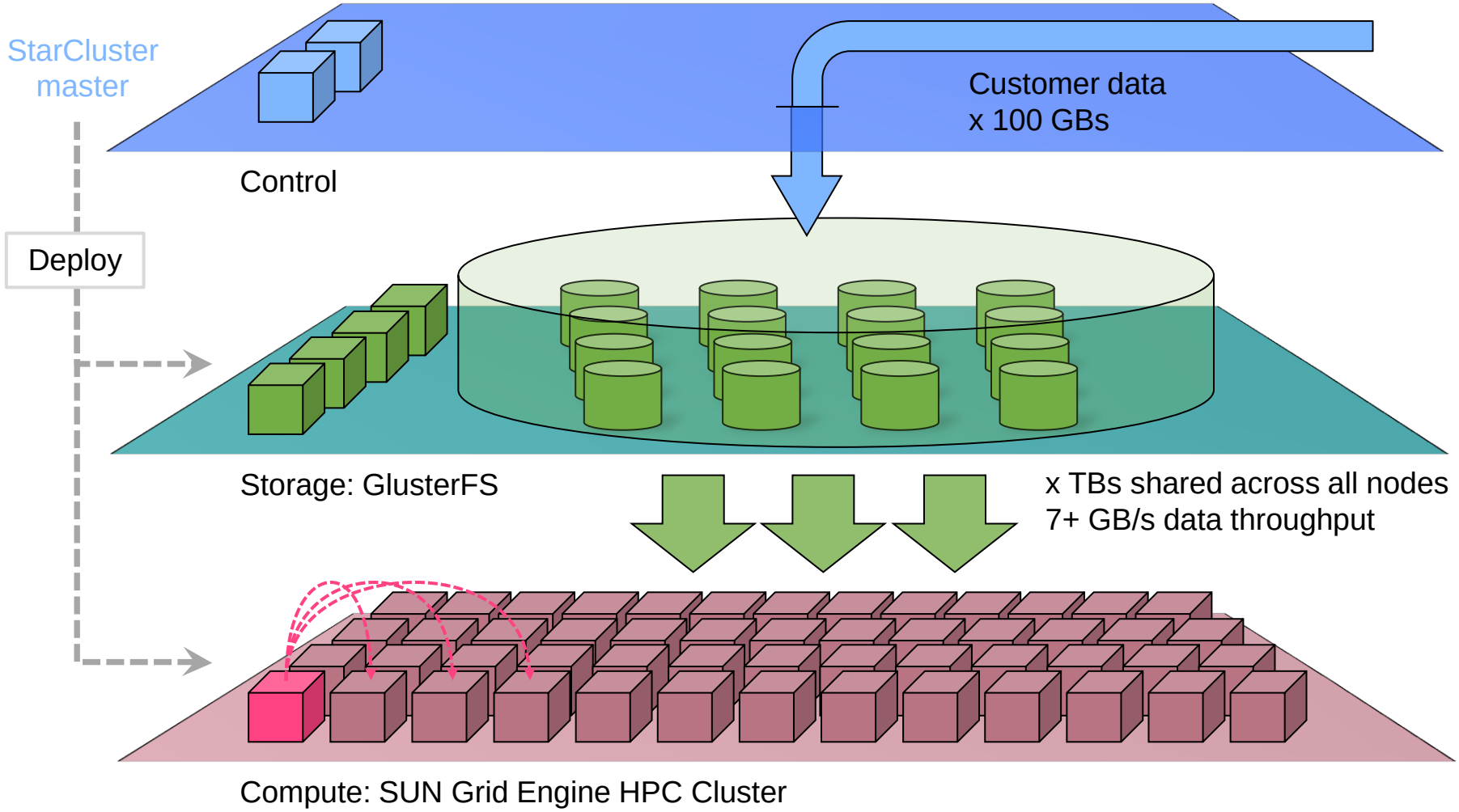
Cloud Service

Integration
with other
cloud services
/ Archiving

Key software components

Assembly pipeline	SGA by Simpson, JT & Durbin, R http://genome.cshlp.org/content/22/3/549.long
Annotation pipeline	Ensembl
Shared file system	glusterFS, Fraunhofer FS
On-demand dynamic HPC cluster provisioning	StarCluster http://star.mit.edu/cluster/

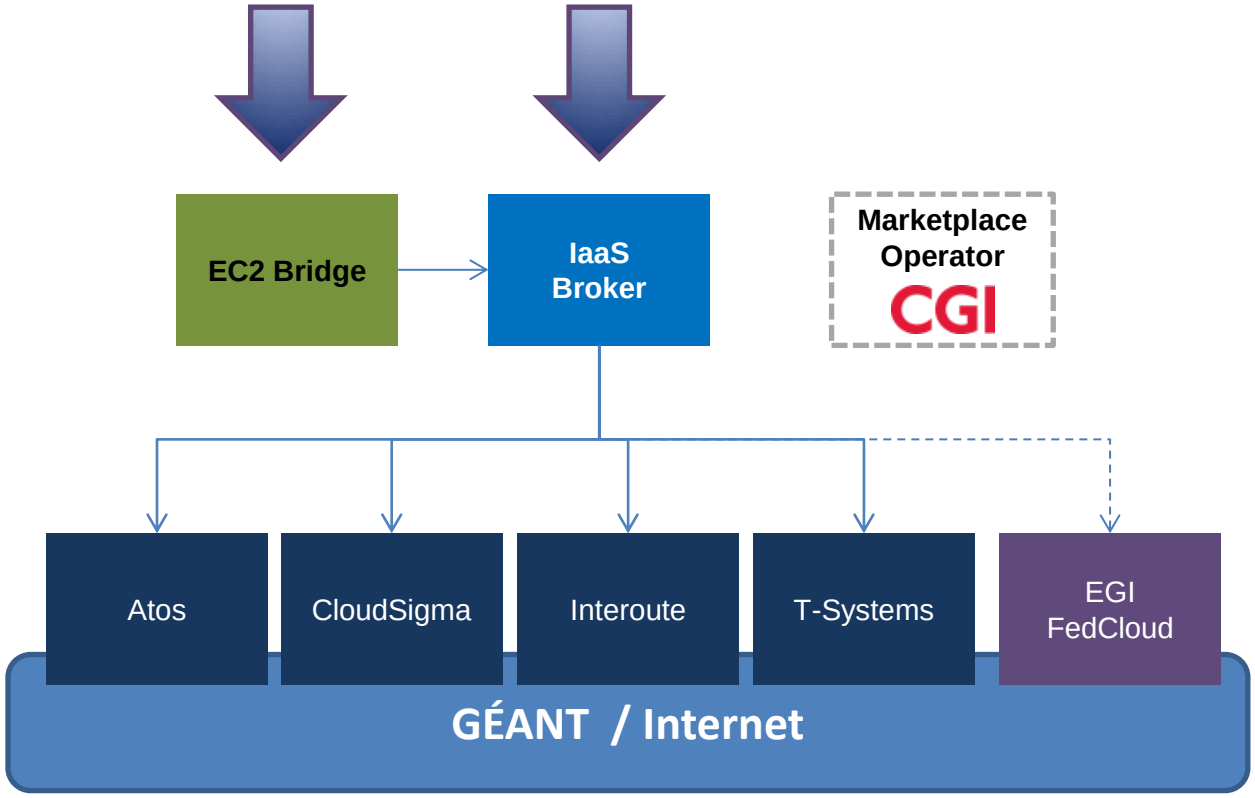
EMBL Dynamic Cloud Architecture



HNX CLOUD MARKETPLACE

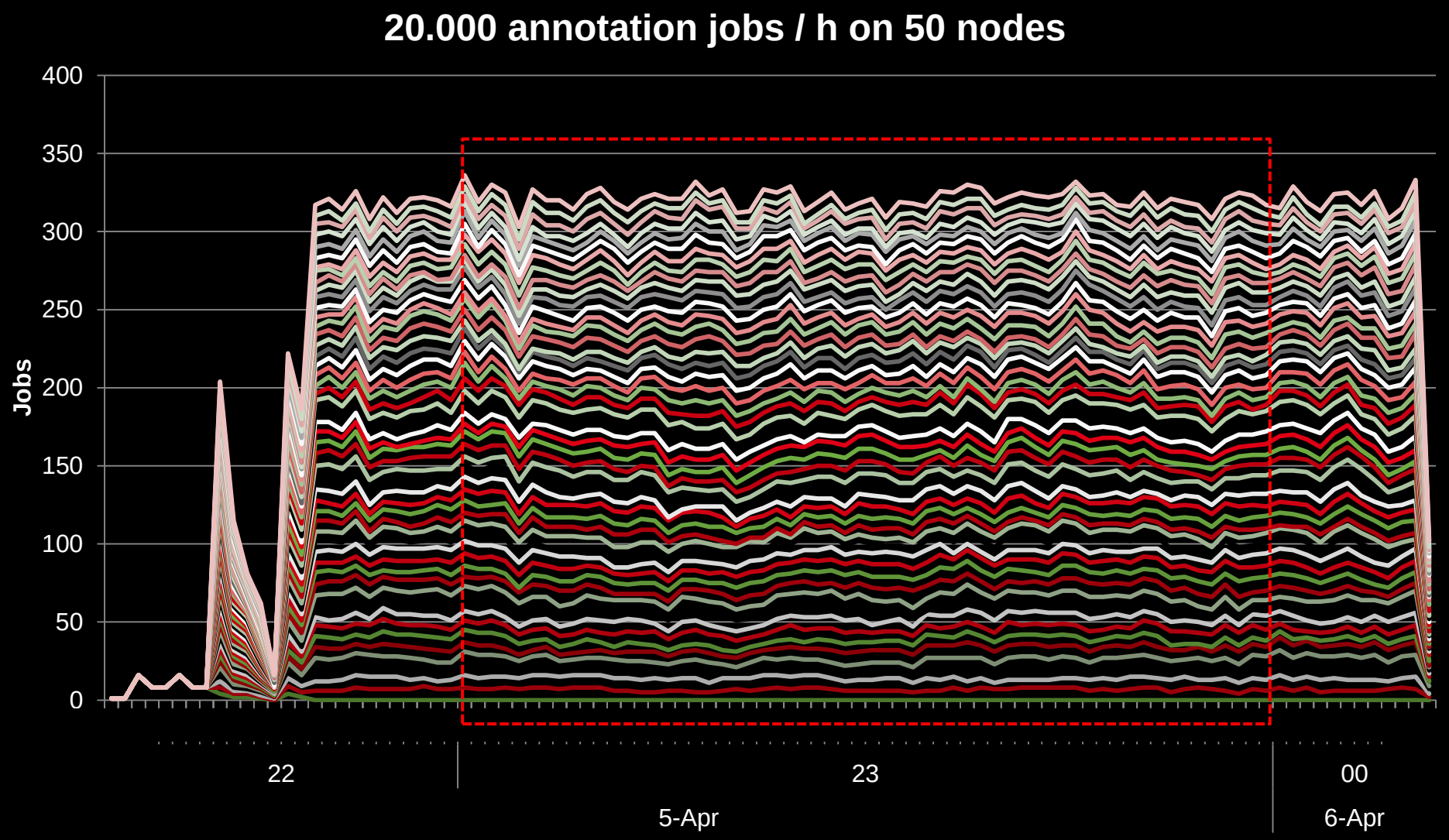
HELIX NEBULA THESCIENCECLOUD

HNX v1.0

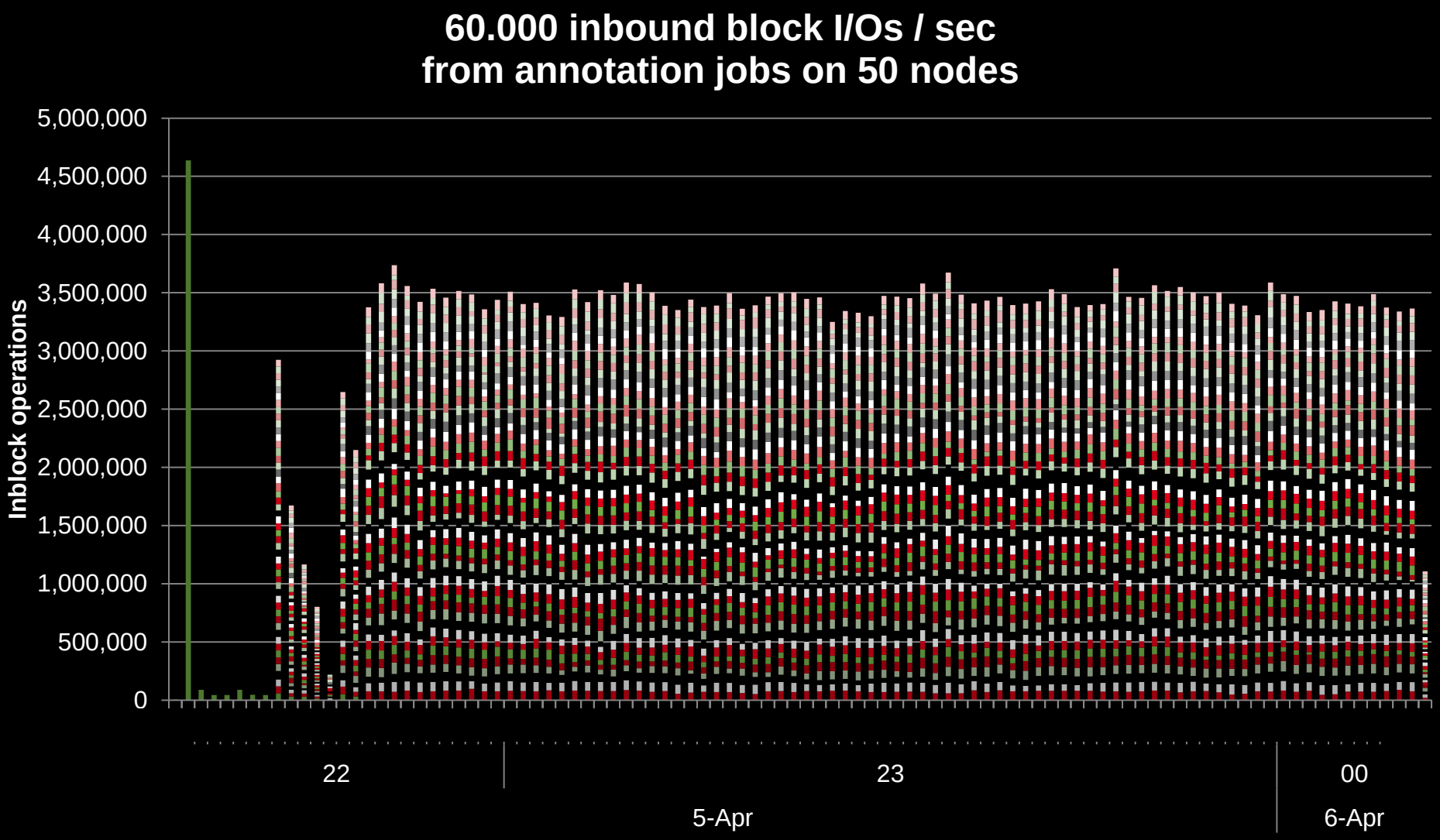


Courtesy of HNX

Sun Grid Engine cluster troughput



GlusterFS throughput



Pilot test results

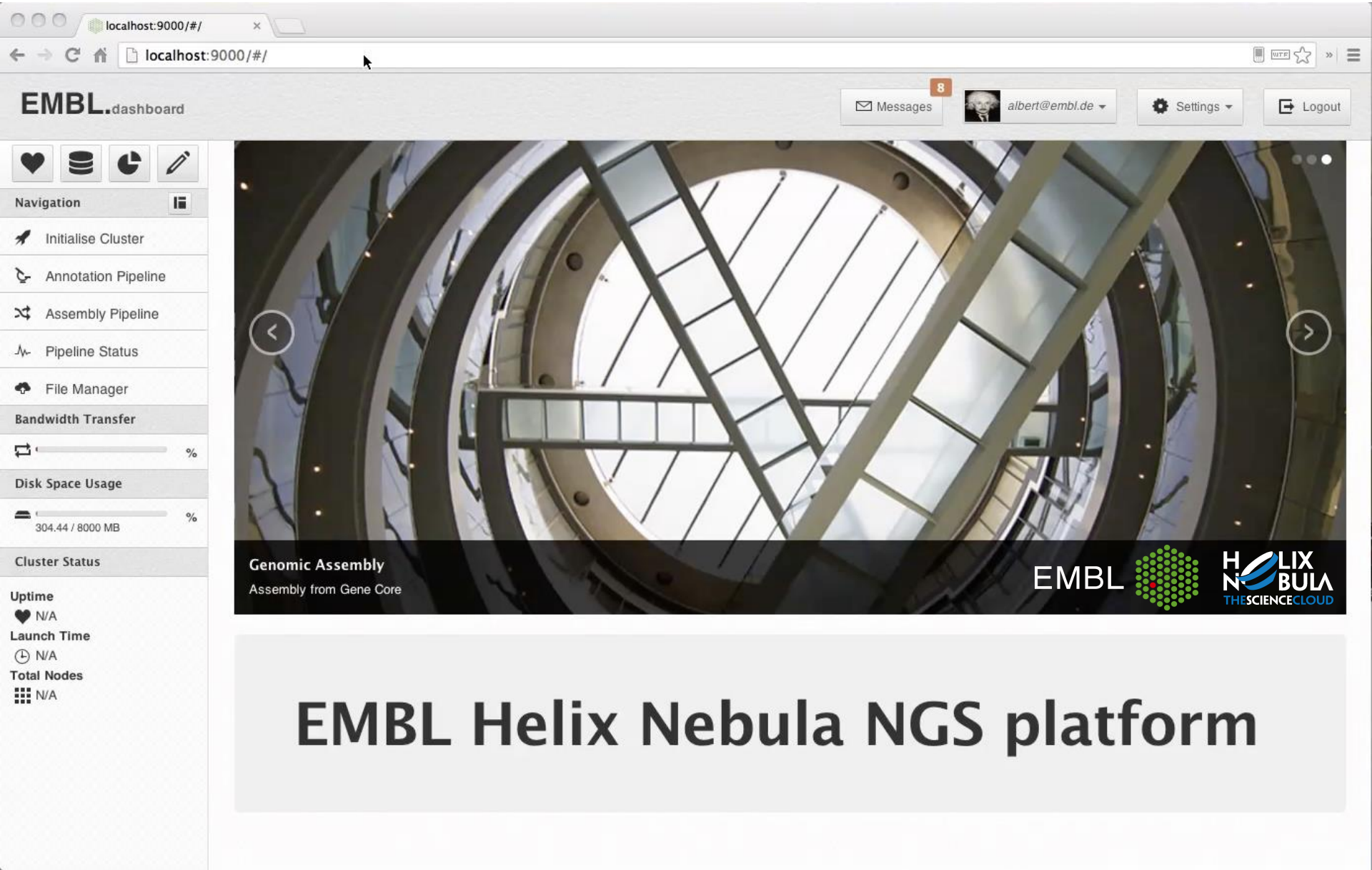
Successful tests of all vendors deployed so far

- StarCluster API integration
- auto-provision 50-node cluster setups
- Successful end-to-end tests of bioinformatics pipelines
- Using 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

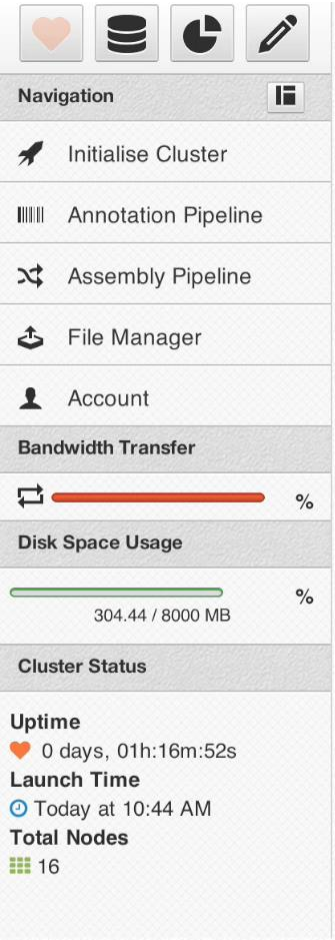
HNX Deployment so far

- Dynamically launched cloud based HPC clusters using a set of EC2 calls initiated by StarCluster
- EMBL NGS web portal-based cluster launch is fully automated
- Initial performance test matching pre-HNX findings

EMBL NGS Cloud Portal

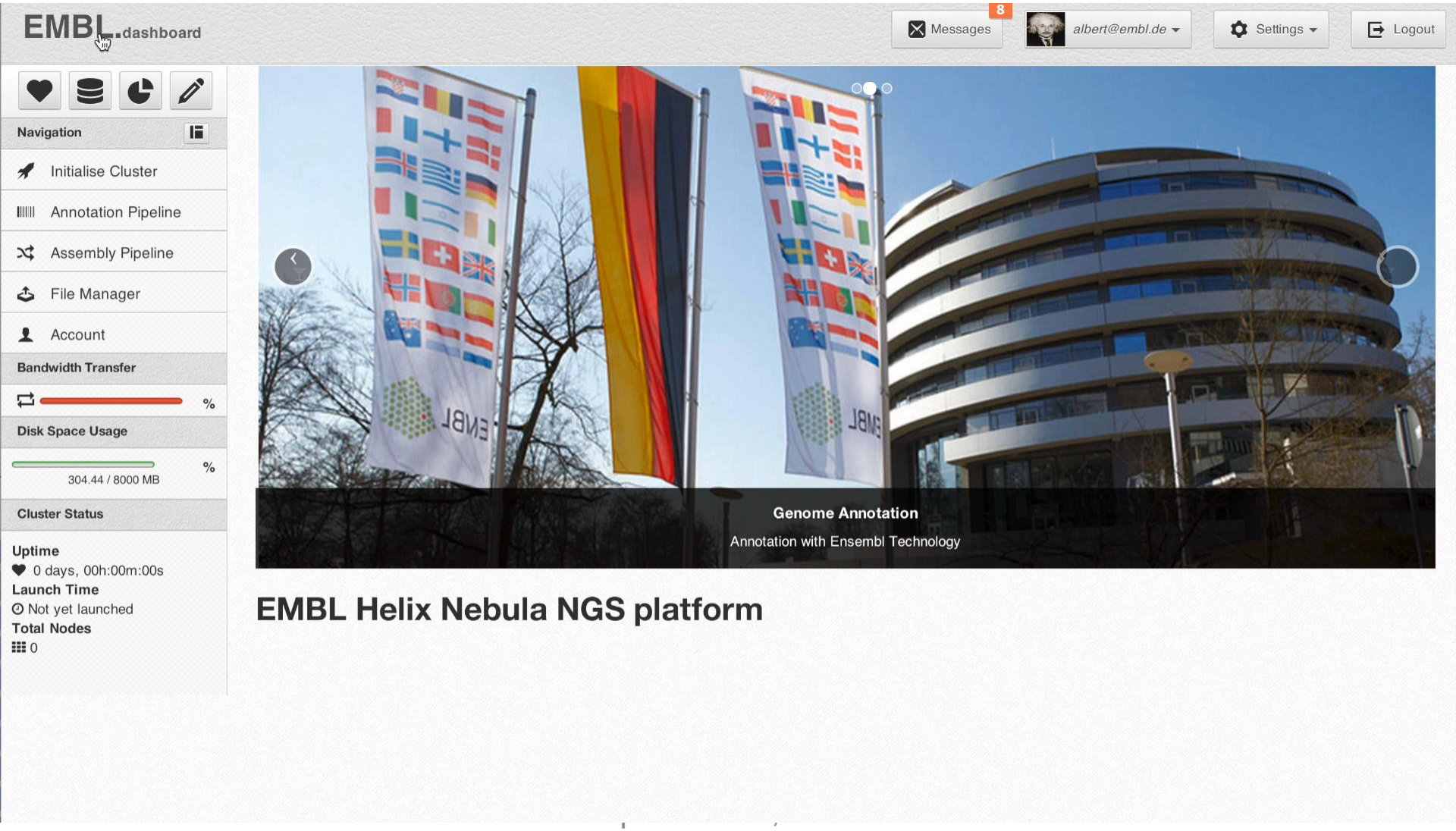


Globus-based Data Transfer into the Cloud

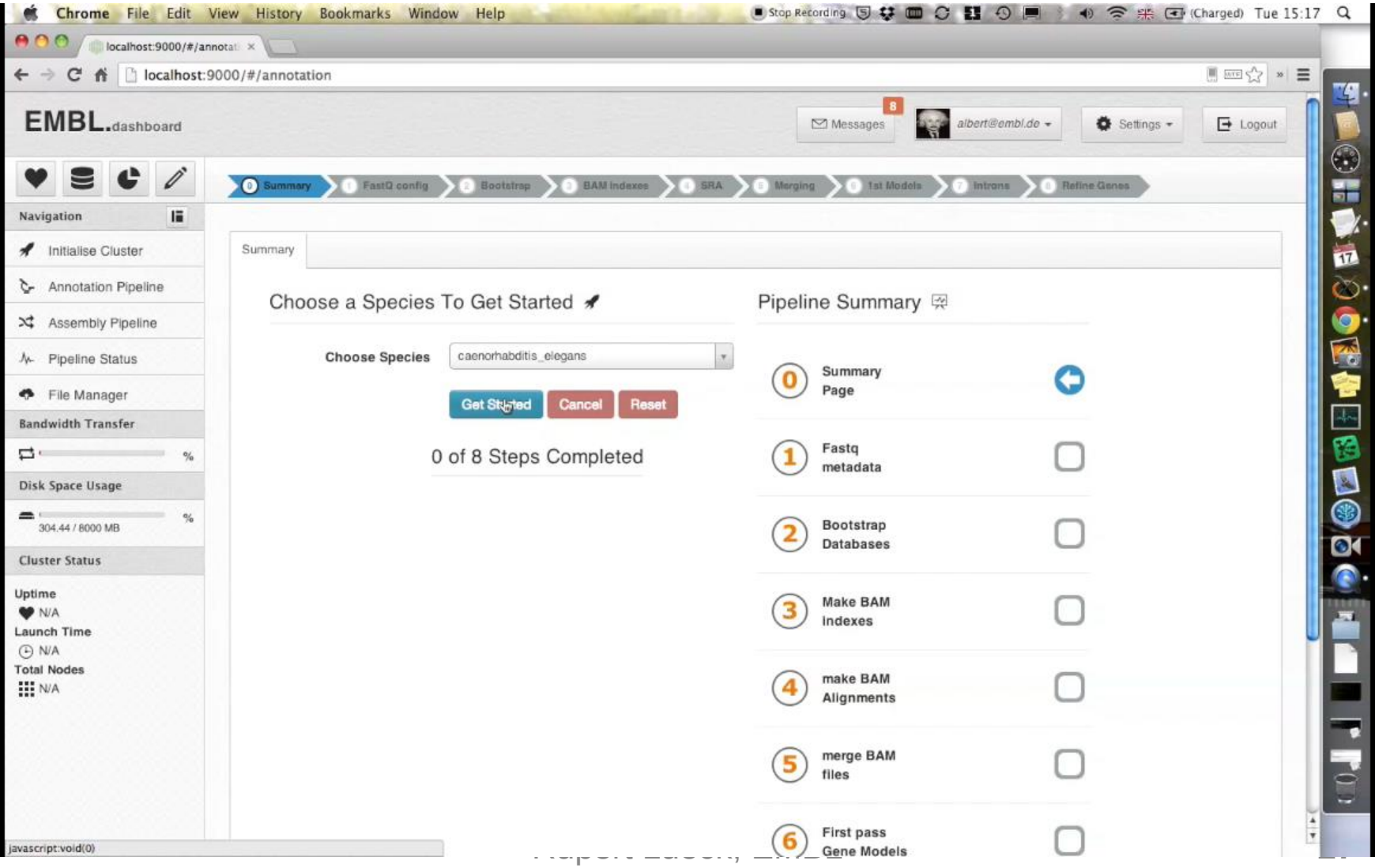


EMBL Helix Nebula NGS platform

StarCluster-driven Cluster Launch



Annotation Pipeline Run



Next Steps

- Further evaluation of EMBL software stack on HNX production platform
- Testing with more suppliers
- End-to-end tests of bioinformatics pipelines
- Scale up and stability testing
- Further develop EMBL genome analysis pipeline towards production

Acknowledgements

EMBL

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

EMBL- EBI

Stephen Keenan
Paul Flicek

In Memoriam

Christian Boulin

EMBL Director of Core Facilities and Services

† 27 April 2014