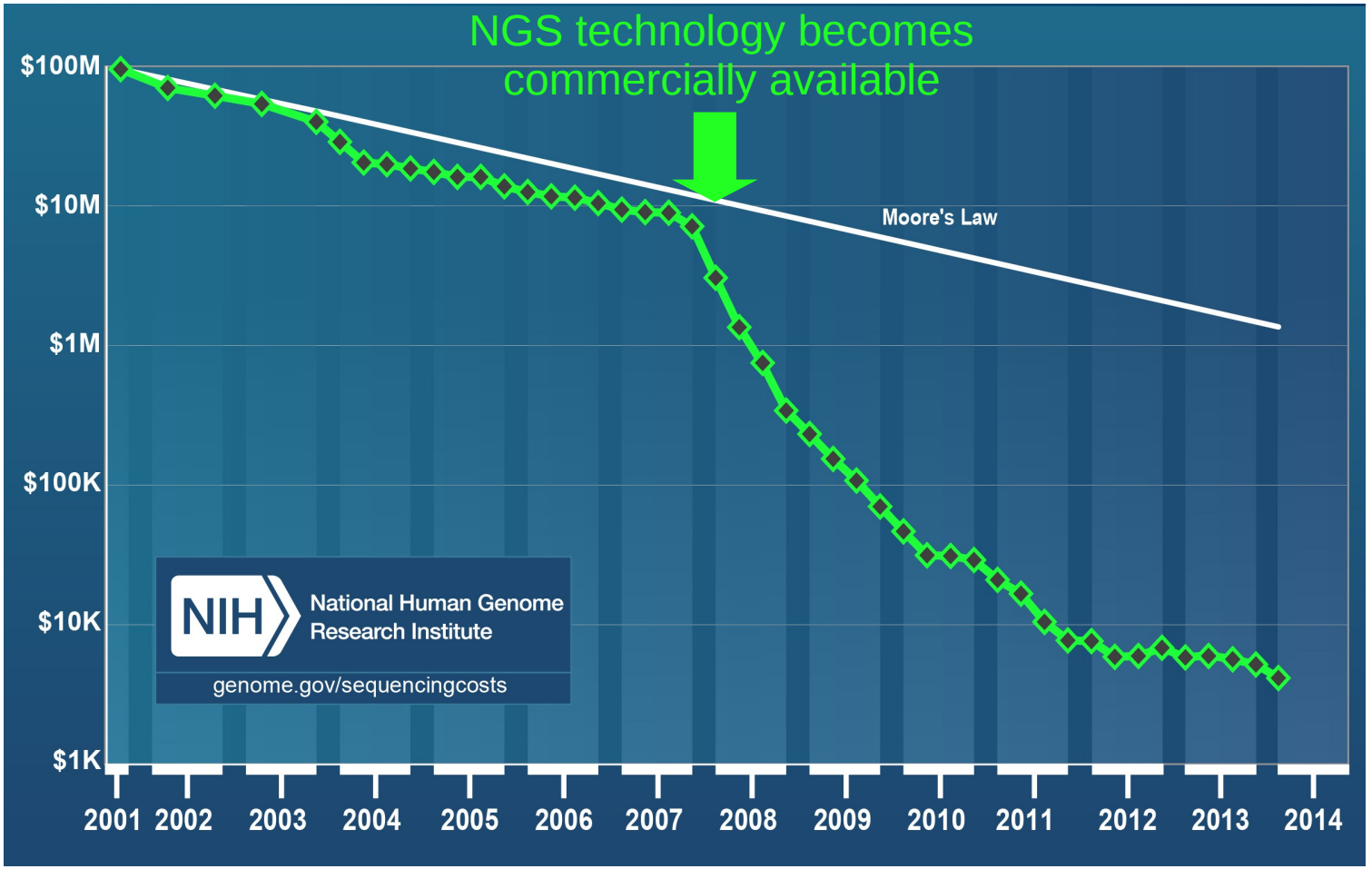
Automated and Scalable Data Management System for Genome Sequencing Data

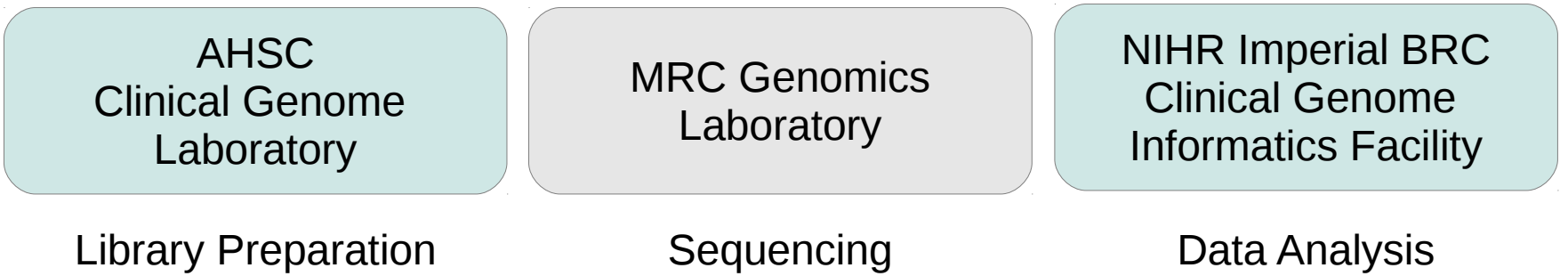
Michael Mueller
NIHR Imperial BRC
Clinical Genome Informatics Facility
Faculty of Medicine
Hammersmith Hospital Campus

Continuously falling costs have resulted in widespread adoption of next-generation sequencing (NGS) technologies

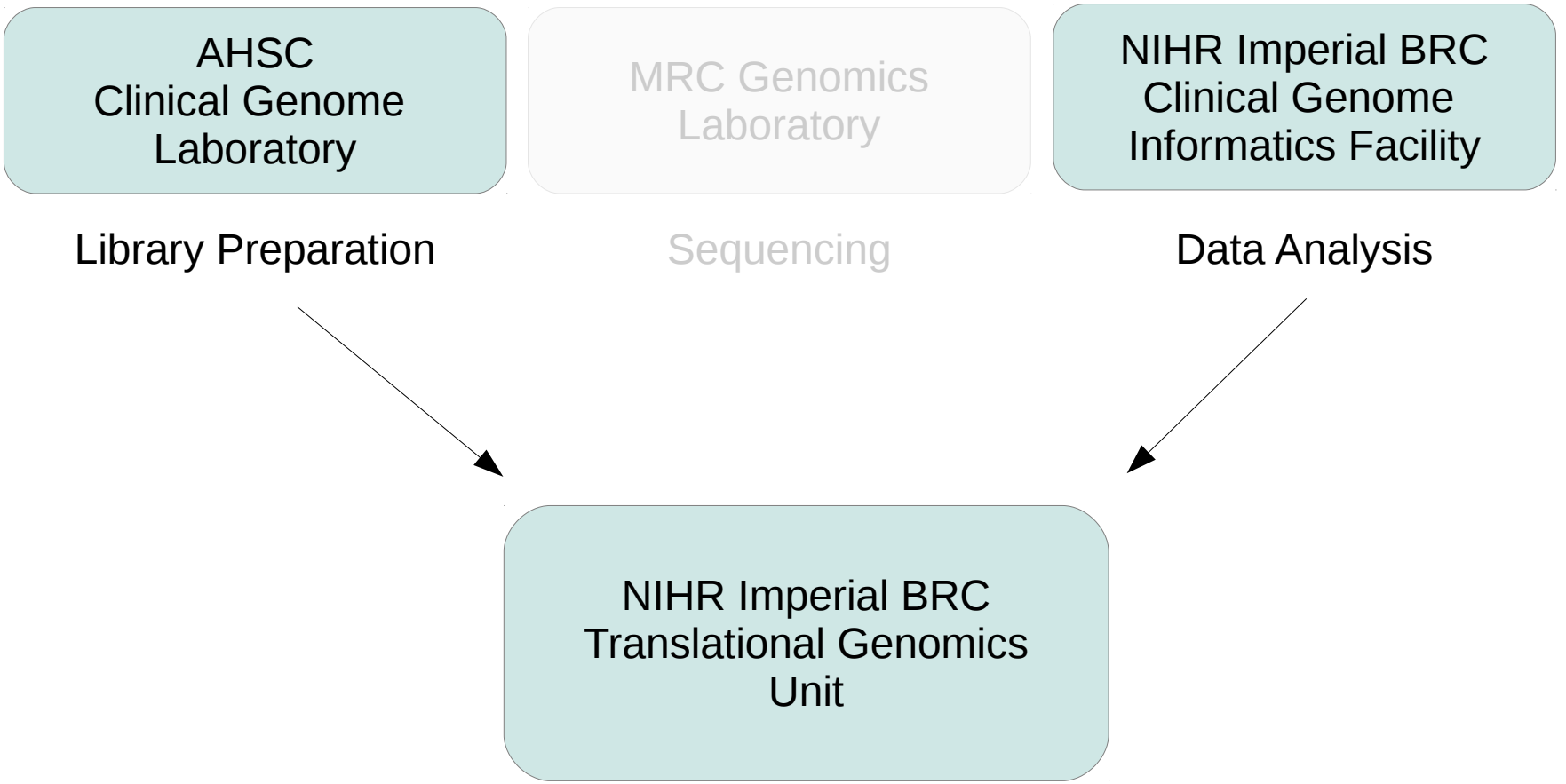
Sequencing costs per genome



NIHR Imperial BRC Translational Genomics Unit
Support for translational NGS projects



NIHR Imperial BRC Translational Genomics Unit
Support for translational NGS projects



Integrated support for translational NGS projects
Library Preparation – Sequencing – Data Analysis

Data generation, storage and processing

Translational Genomics Unit

Illumina HiSeq 2500
up to 8TB
of raw data per run

Illumina MiSeq
up to 100GB
of raw data per run

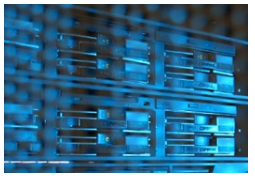
Dell PowerEdge Server
40TB storage

High Performance Computing Service

PC cluster (cx1)
dedicated access to
320 cores
20TB + 80TB storage

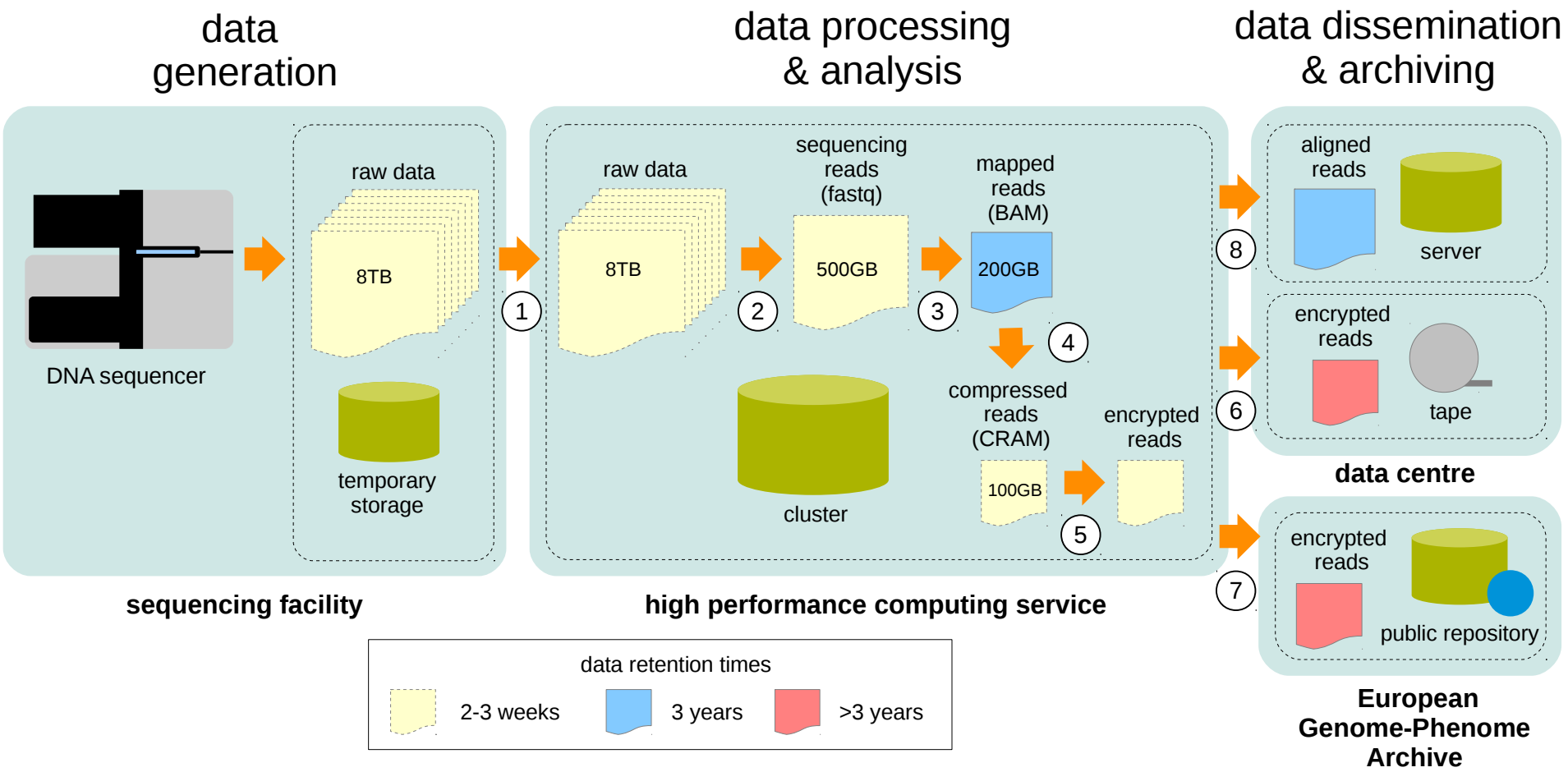
SGI Altix UV (ax3)
shared access to
384 cores
28TB storage

Data Centre



HP ProLiant Server
15TB storage

Data generation, storage and processing



(1) transfer of raw data from local storage server to HPC Service
(2) extraction of sequencing read information from raw data
(3) mapping of sequencing reads to reference genome sequence
(4) reference based compression of mapped read data
(5) encryption of compressed read data
(6) local archiving of encrypted read data on tape
(7) remote archiving at public repository
(8) local dissemination of mapped read data

Data Management System Requirements

- Ensure data integrity and security
- Avoid unnecessary data replication
- Facilitate data access for analysis and sharing
- Allow for a high degree of automation
- Scalable
- Comply with College and funder data preservation requirements

integrated Rule-Oriented Data System

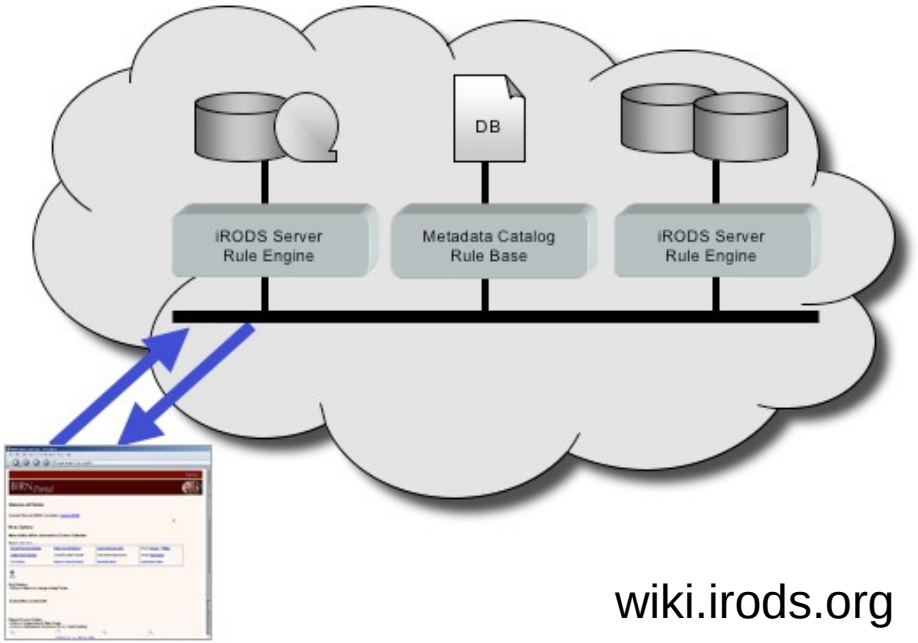
"...software middleware that manages a highly controlled collection of distributed digital objects, while enforcing user-defined Management Policies across the multiple storage locations"

Integrated Rule-Oriented Data System

- Open source under a BSD license
- Developed by the Data Intensive Cyber Environments research group (University of North Carolina at Chapel Hill (UNC) and the University of California, San Diego (UCSD)) and collaborators
- Uniform interface to distributed and heterogeneous data storage resources
- Implements a **logical namespace** and maintains **metadata catalogue** (iCAT) on data-objects
- Features highly-configurable **rule engine** to manage the processing, sharing, replication, transfer, and preservation of distributed data collections
- various end-user client applications: command line (i-Commands), web client, iRODS Explorer for Windows, Java and PHP client APIs
- Used for the management of genome sequencing data at the Wellcome Trust Sanger Institute

integrated Rule-Oriented Data System

Peer-to-peer server architecture



wiki.irods.org

Rule Engine

- Management policies expressed as computer actionable rules
- Management procedures expressed as sets of remotely executable Micro-services
- Rules control execution of Micro-services
- State information generated by Micro-services is stored in metadata catalog (iCAT)

Dalia Kasperaviciute
Bioinformatician
Clinical Genome Informatics Facility

Alona Sosinsky
Bioinformatician
Clinical Genome Informatics Facility

Anna Zekavati
Head
Clinical Genome Laboratory

Emmanuel Savage
Laboratory Technician
Clinical Genome Laboratory

Jorge Ferrer
Theme Leader
Genetics & Genomics NIHR Imperial BRC

Simon Burbidge
Manager
Imperial HPC Service

Steve Lawlor
Manager
ICT Computer Room