EMBL-EBI and Bioinformatics

Steven Newhouse,

Head of Technical Services, EMBL-EBI

The European Molecular Biology Laboratory

Heidelberg

Basic research

Administration

EMBO

Hamburg

Structural biology

Hinxton, Cambridge

Bioinformatics

Grenoble

Structural biology

Monterotondo, Rome

Mouse biology

EMBL staff:

1700 people

>60 nationalities

EMBL member states

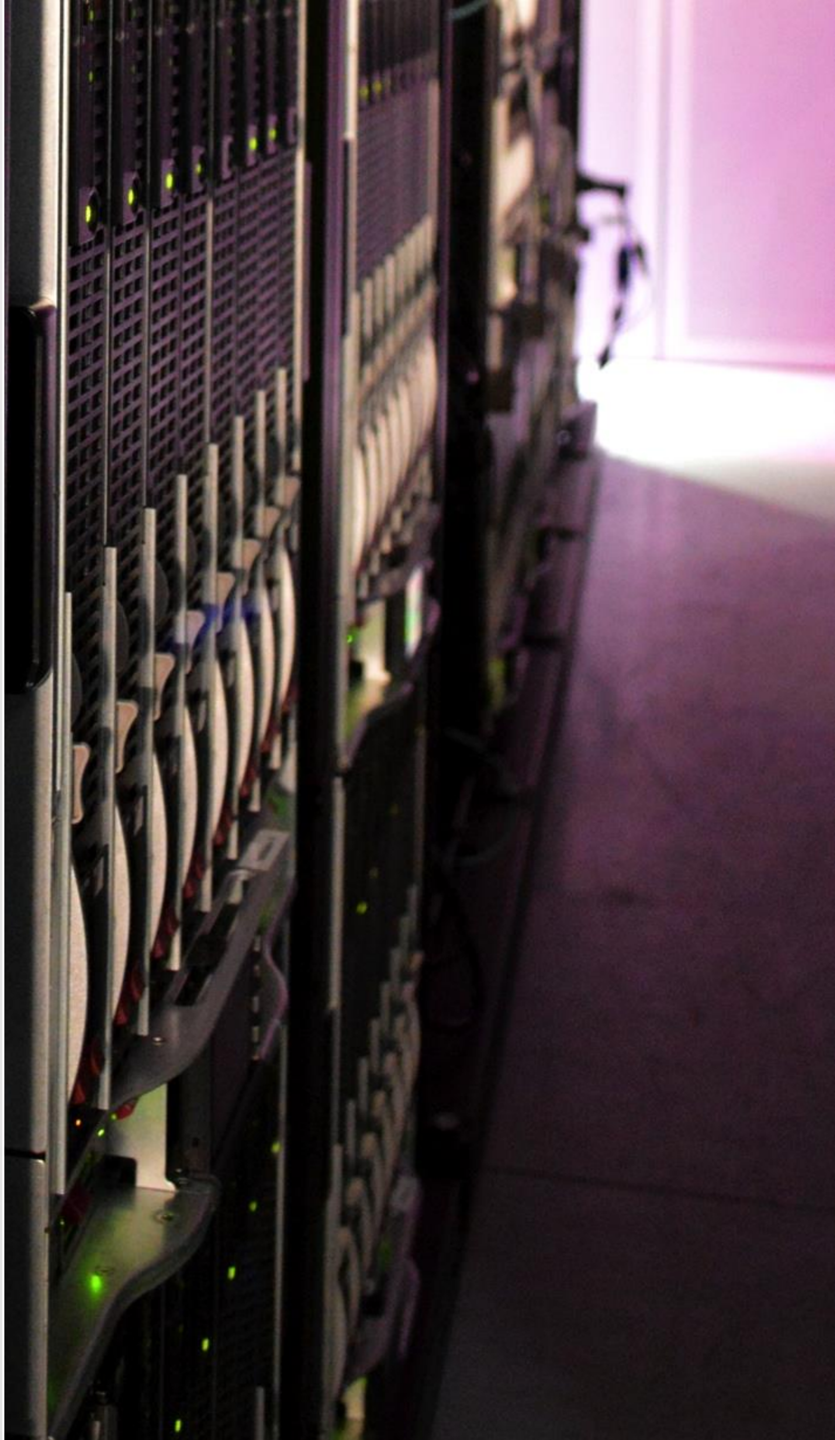
Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom

Associate member states: Argentina, Australia

ICELAND
FINLAND
NORWAY
SWEDEN
DENMARK
IRELAND
UNITED KINGDOM
NETHERLANDS
BELGIUM
GERMANY
LUXEMBOURG
FRANCE
SWITZERLAND
AUSTRIA
CROATIA
ITALY
PORTUGAL
SPAIN
GREECE
Associate Member State
AUSTRALIA
ISRAEL

EMBL-EBI MISSION

To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress



European Bioinformatics Institute (EBI)

- International, non-profit research institute
- Europe’s hub for biological data services and research
- 570 members of staff from 53 nations
- Funded primarily by member states and research bodies (EC, USA, UK, Wellcome Trust)

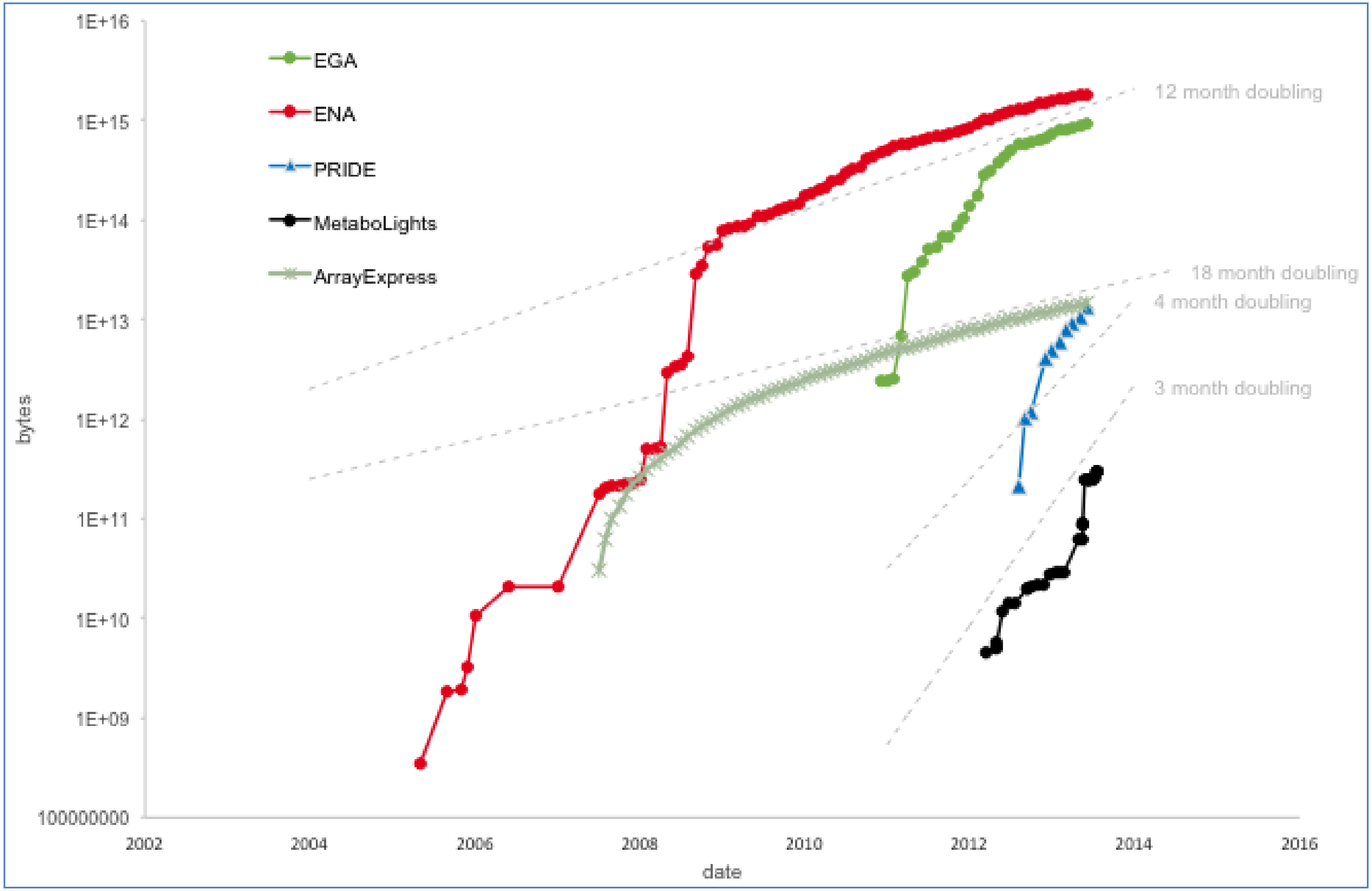
What is bioinformatics?

- The science of storing, retrieving and analysing large amounts of biological information
- An interdisciplinary science involving:
  - biologists
  - biochemists
  - computer scientists
  - mathematicians.

EBI Provides Services and Data Resources

- Data Resources
  - Public and Managed Access
  - Individual sequence or bulk download
- Services
  - Web & programmatic access for common tools
  - Run ‘jobs’ on EMBL-EBI hardware
- Volume and variety of genomic data expanding
  - EMBL-EBI data doubling every year - replication is challenging
  - Infrastructure currently 50,000 CPUs & 46PB

Data Growth: A Community Challenge



The 1000 Genomes Project

- Has sequenced 2504 individuals from 26 populations around the globe
- Established a reference human haplotype structure
  - ~80M variants with phased genotypes for the 2504 individuals
- Defined standards for variation data sequencing, storage and analysis
- The data is completely open and is available from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

The 1000 Genomes Project

- FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
  - Raw Data Files
  - Accessible by Aspera
  - Accessible by Globus Grid FTP
- AWS Amazon Cloud: http://aws.amazon.com/1000genomes/
  - FTP mirror
- Web site: http://www.1000genomes.org
  - Release Announcements
  - Documentation
- Ensembl Style Browser: http://browser.1000genomes.org
  - Browse 1000 Genomes variants in Genomic Context
  - Variant Effect Predictor
  - Data Slicer
  - Other Tools

The International Genome Sample Resource

The International Genome Sample Resource (IGSR), a Wellcome Trust funded project that will be built on the foundation of the 1000 Genomes Project starts in 2015.

IGSR plans to:

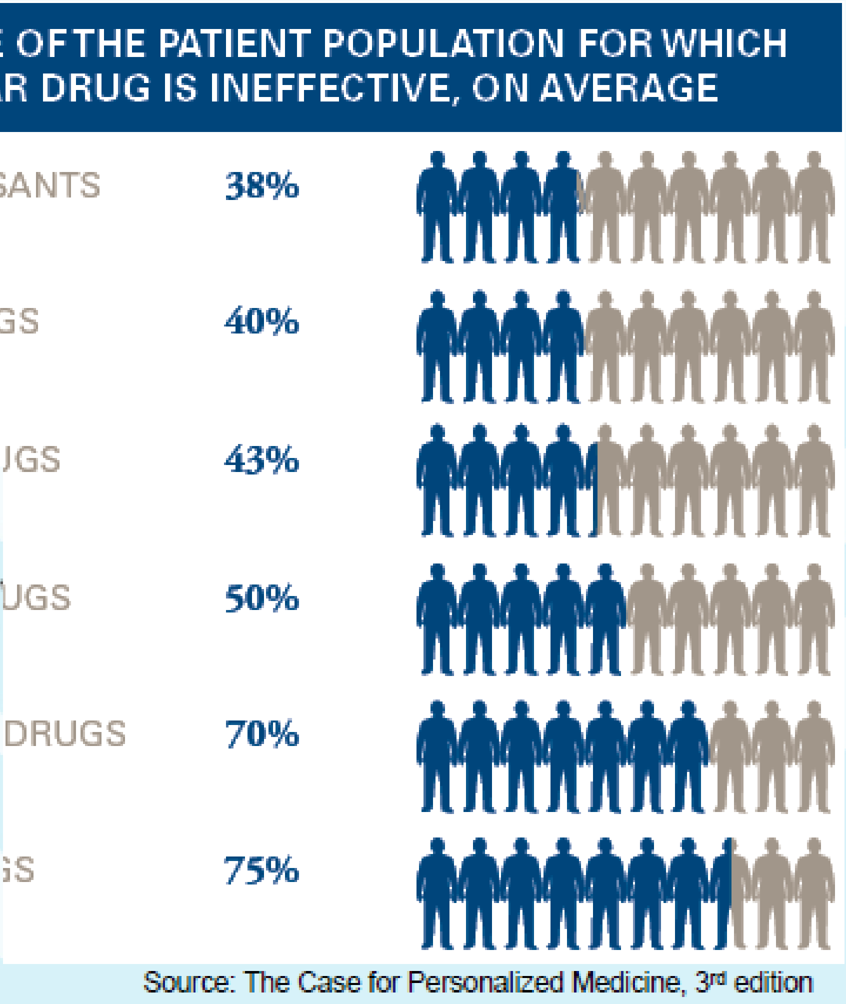
- Maintain the existing 1000 genomes data and move to GRCh38
- Collect other data sets generated on the Coriell Cell Lines including Geuvadis
- Add new populations to expand the global diversity of the variant catalog

London Phenome Centres

- Gaining insights into clinical and epidemiological questions
  - Led by Jeremy Nicholson, Imperial College
- Clinical Phenome Centre at St Mary’s
  - How the metabolome of individual patients changes during the patient journey in the hospital
- Contribute data into MetaboLights data resource @ EBI
  - Issues with handling and analysing this data
- Goal: More individualised response to the patients
  - Lead to better more cost-effective outcomes

Personalised Medicine

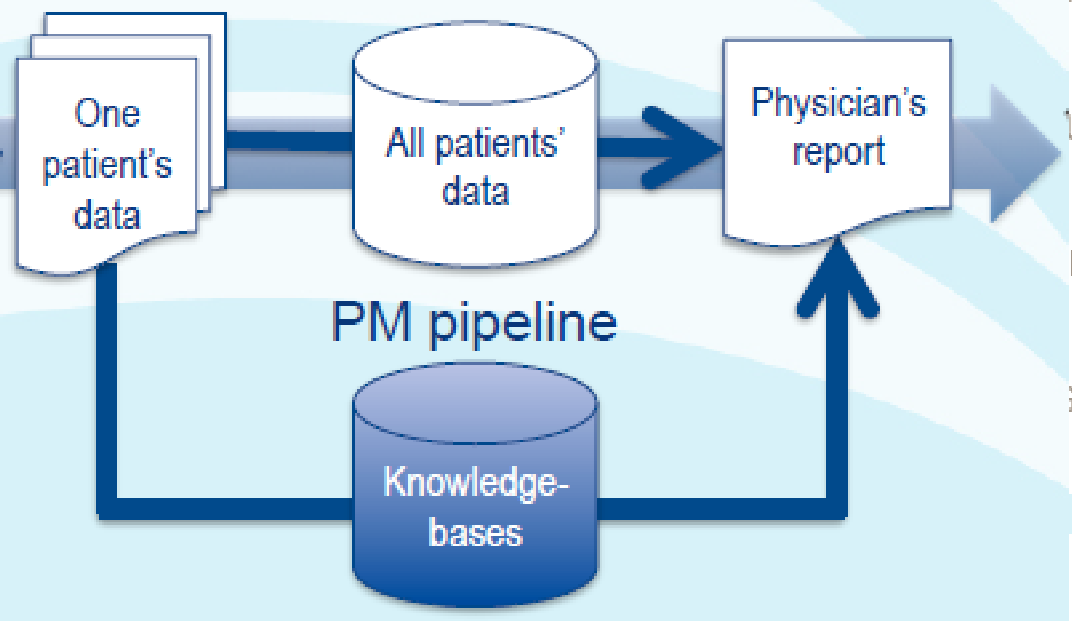
- Reduced sequencing costs enable new techniques
  - Past: 13 yrs & £2Bn
  - Now: 2 days & £1000



PERCENTAGE OF THE PATIENT POPULATION FOR WHICH A PARTICULAR DRUG IS INEFFECTIVE, ON AVERAGE

ANTI-DEPRESSANTS (SSRIs)	38%
ASTHMA DRUGS	40%
DIABETES DRUGS	43%
UGS	50%
DRUGS	70%
GS	75%

Source: The Case for Personalized Medicine, 3rd edition



The Challenge Facing Services @ EMBL-EBI

- Current ‘Software as a Service’ model needs optimisation
  - Web and programmatic access to services (3M unique users)
- Need to support complex analysis scenarios
  - Access to both public and managed access data sets
  - Bespoke workflows and tools across a variety of domains
- Hard for users to replicate data sets for local analysis
  - ‘Infrastructure as a Service’ brings local analysis to EMBL-EBI

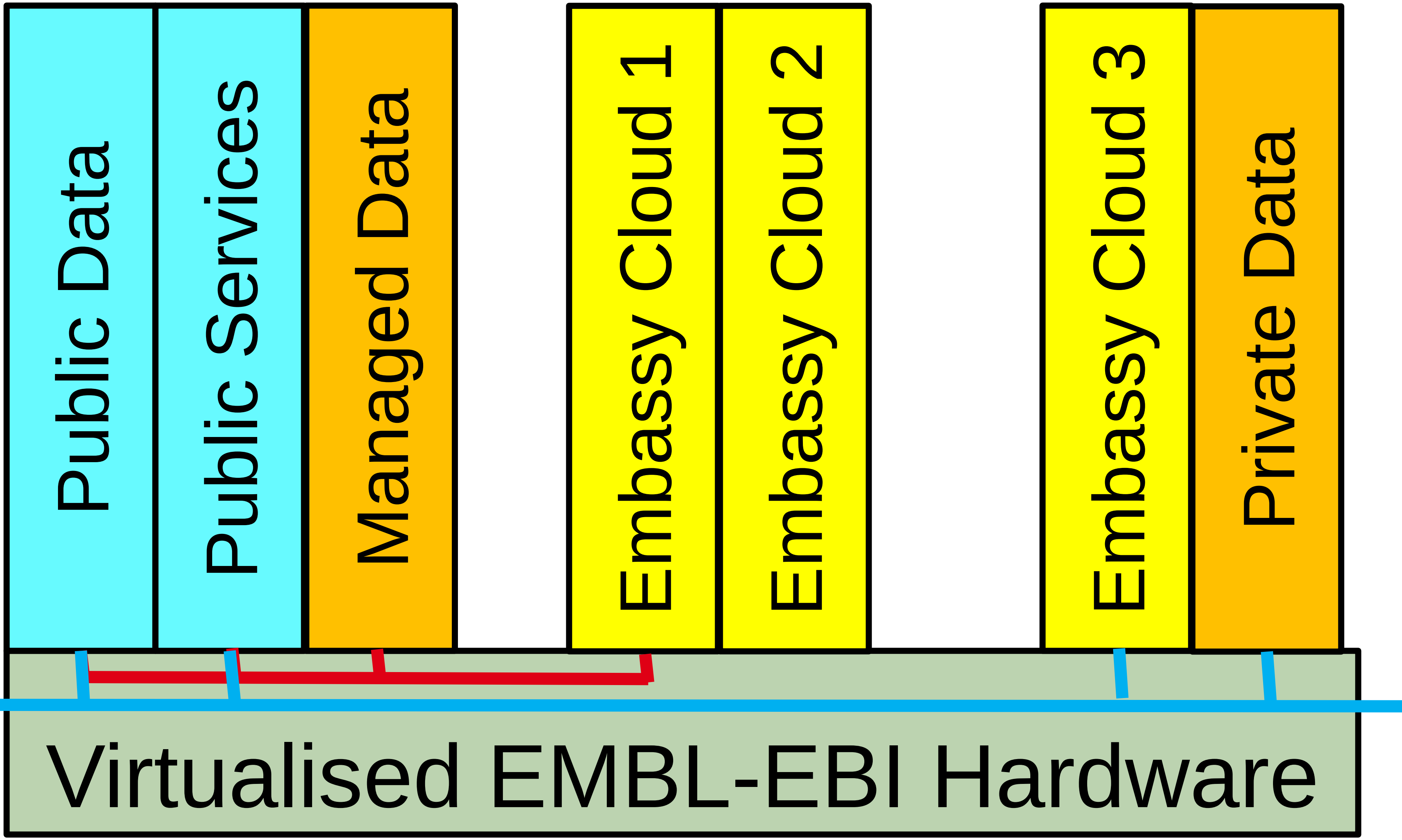
Embassy Cloud

- Work directly with EMBL-EBI data
  - High bandwidth
  - Low latency
  - Robust, secure environment
- Not in competition with commercial cloud services
- Other cloud initiatives:
  - ELIXIR-facing cloud support
  - HELIX Nebula

Genes & genomes · Expression · Prote
Structures
Pathways
EMBL-EBI
Embassy Cloud
Literature · Ontology & taxonomy · Cros

Embassy Cloud Concept

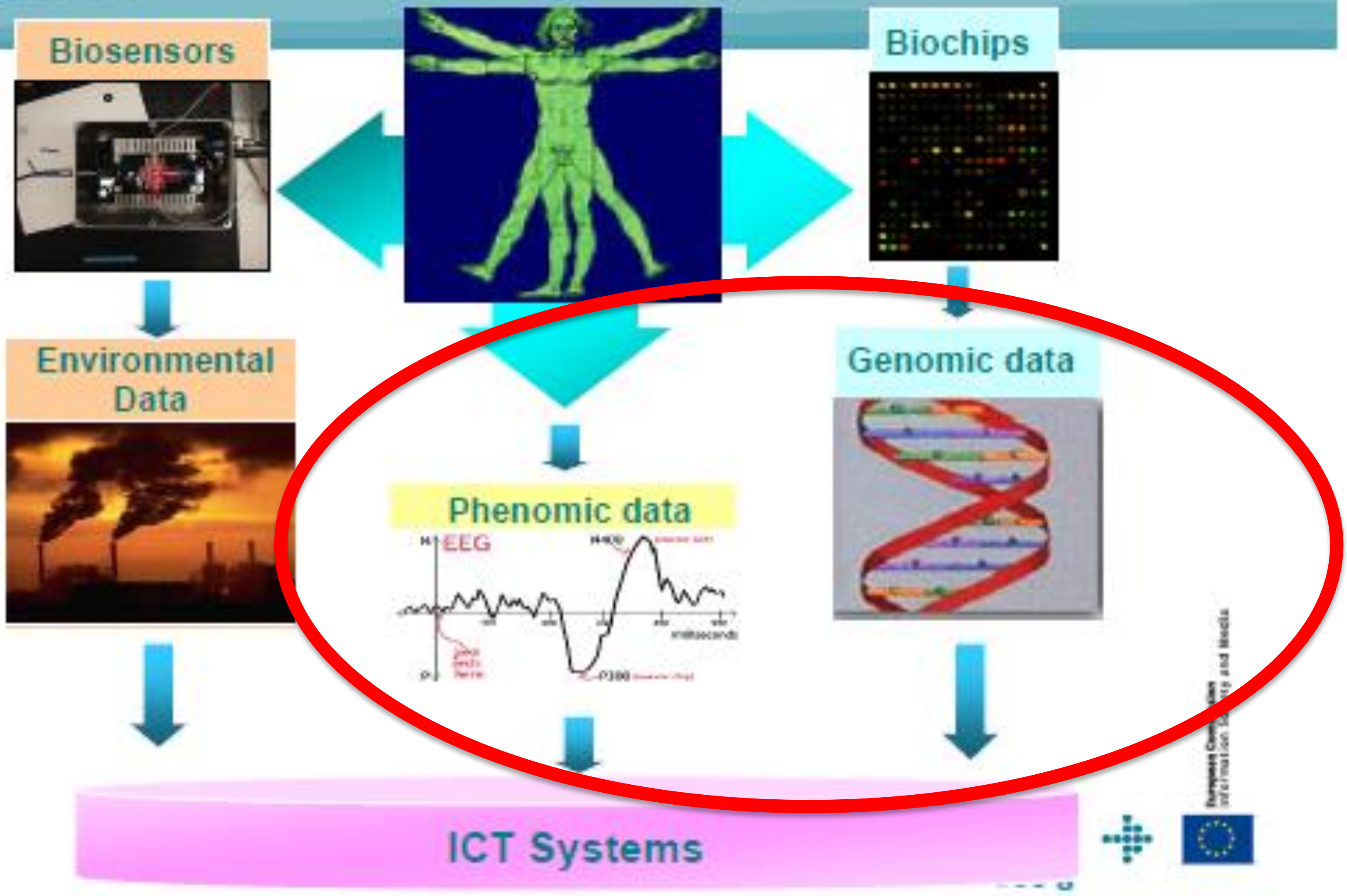
PanCancer



Typical Uses

- Web Application Hosting
  - Limited need for resources & VMs
  - CTTV: Host intranet, databases, …
- Data Staging
  - Undertake submission from local machine (following data staging) rather from remote location
  - BRAEMBL: Remote submission unreliable due to file upload
- Data Analysis
  - Large scale management and analysis of data
  - PanCancer: 1,000 cores, 2.5 TB RAM, 0.5 PB HDD

Towards full picture of individual's health status



Biosensors

Biochips

Environmental Data

Genomic data

Phenomic data

ICT Systems

Courtesy Marco Manca

Summary

- EMBL-EBI is the source of life-science data in Europe
- Variety of technical resources for collaborators
    - Provided by the Technical Services cluster
- Open for collaboration across all data resources
    - Submitting data
    - New data analysis techniques
    - New approaches to accessing and manipulating data
- Contact: steven.newhouse@ebi.ac.uk