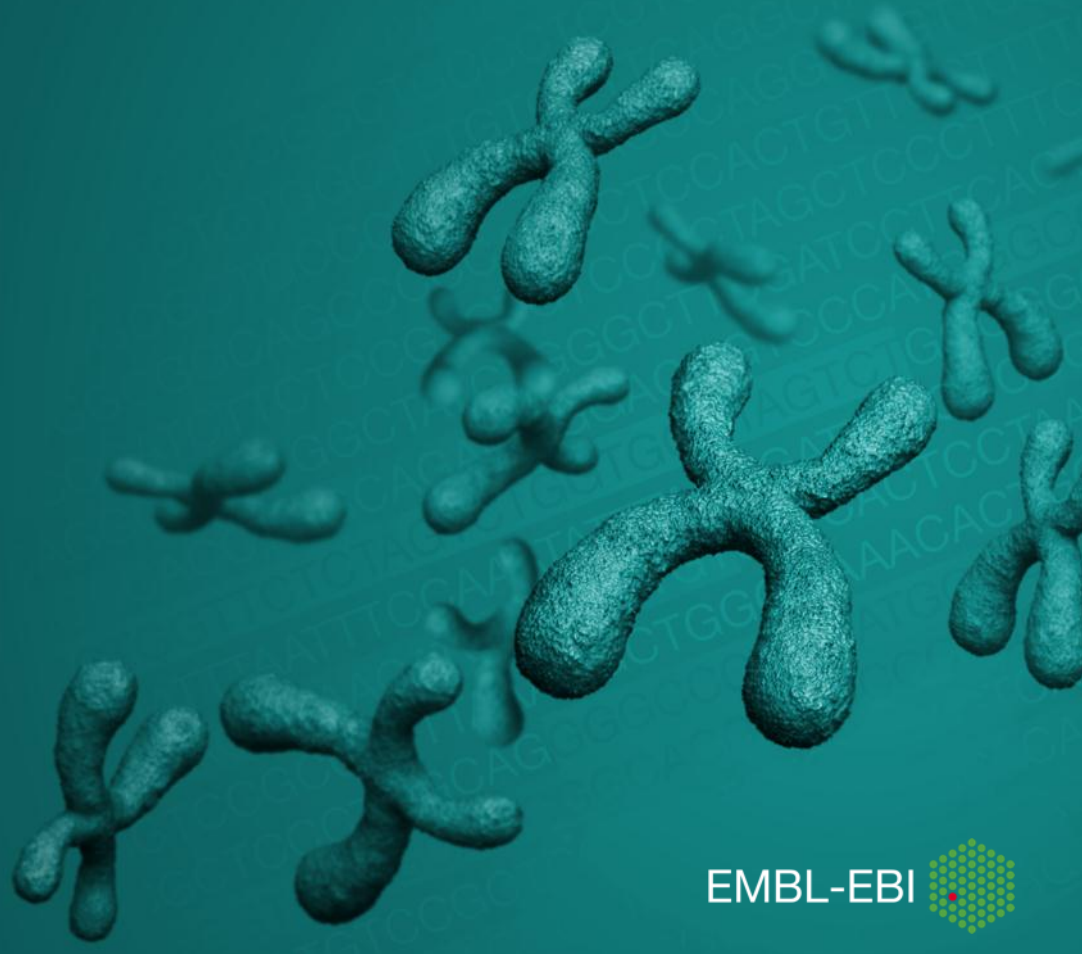
Elixir: European Bioinformatics Research Infrastructure

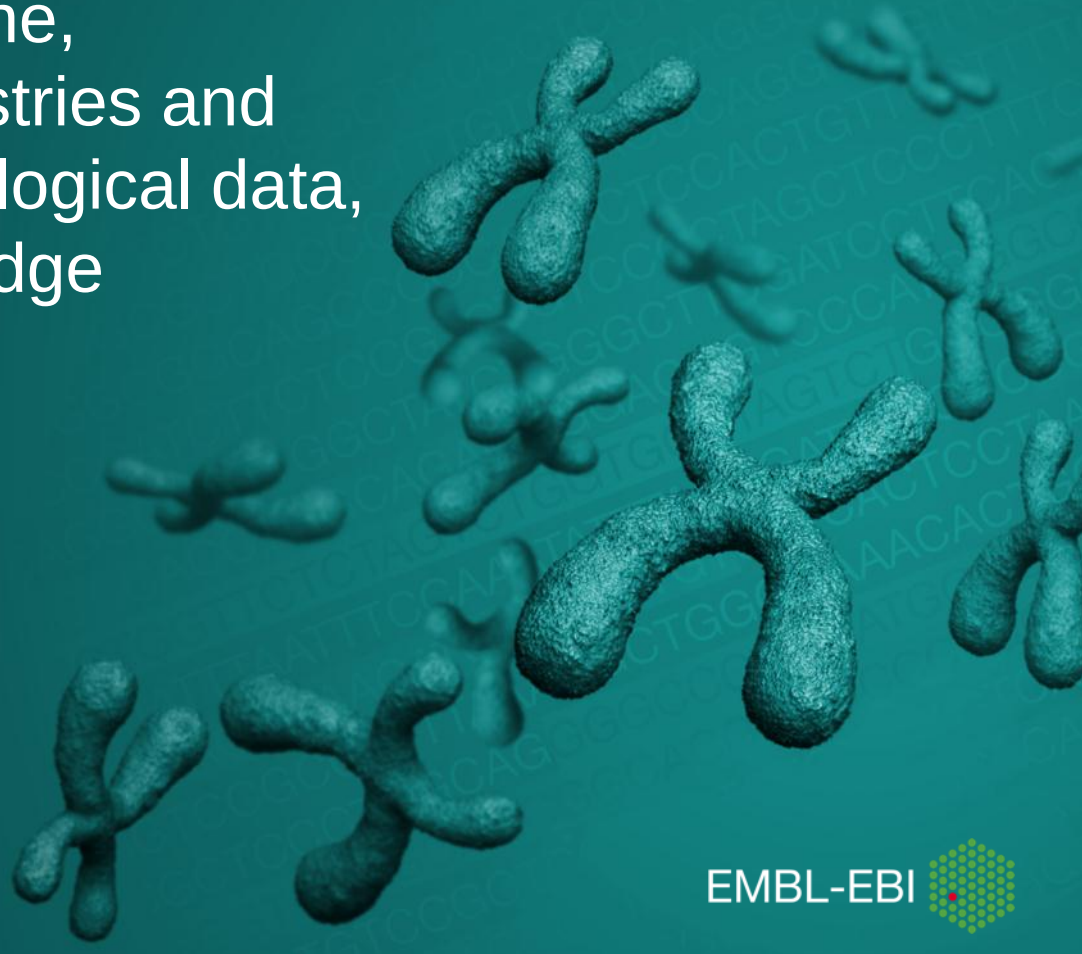
Rolf Apweiler



EMBL-EBI

EMBL-EBI Service Mission

To enable life science research and its translation to medicine, agriculture, the bioindustries and society by providing biological data, information and knowledge



EMBL-EBI

What services do we provide?

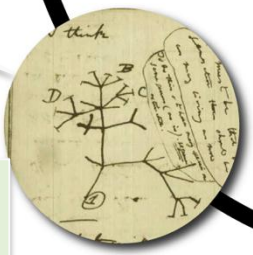
Labs around the world send us their data and we…

Archive it

Classify it

Share it with other data providers

Analyse it



…provide tools to help researchers use it

A virtuous circle

Data resources at EMBL-EBI

Genes, genomes & variation

Reactions, interactions & pathways

RNA, protein & metabolite expression

Chemical biology

Protein sequences, families & motifs

Ontologies & biological samples

Molecular & cellular structures

Scientific literature

Data resources at EMBL-EBI

Genomes & variation
- Ensembl
- Ensembl Genomes
- Genome-phenome archive
- Metagenomics

Proteins
- The Universal Protein Resource (UniProt)
- InterPro

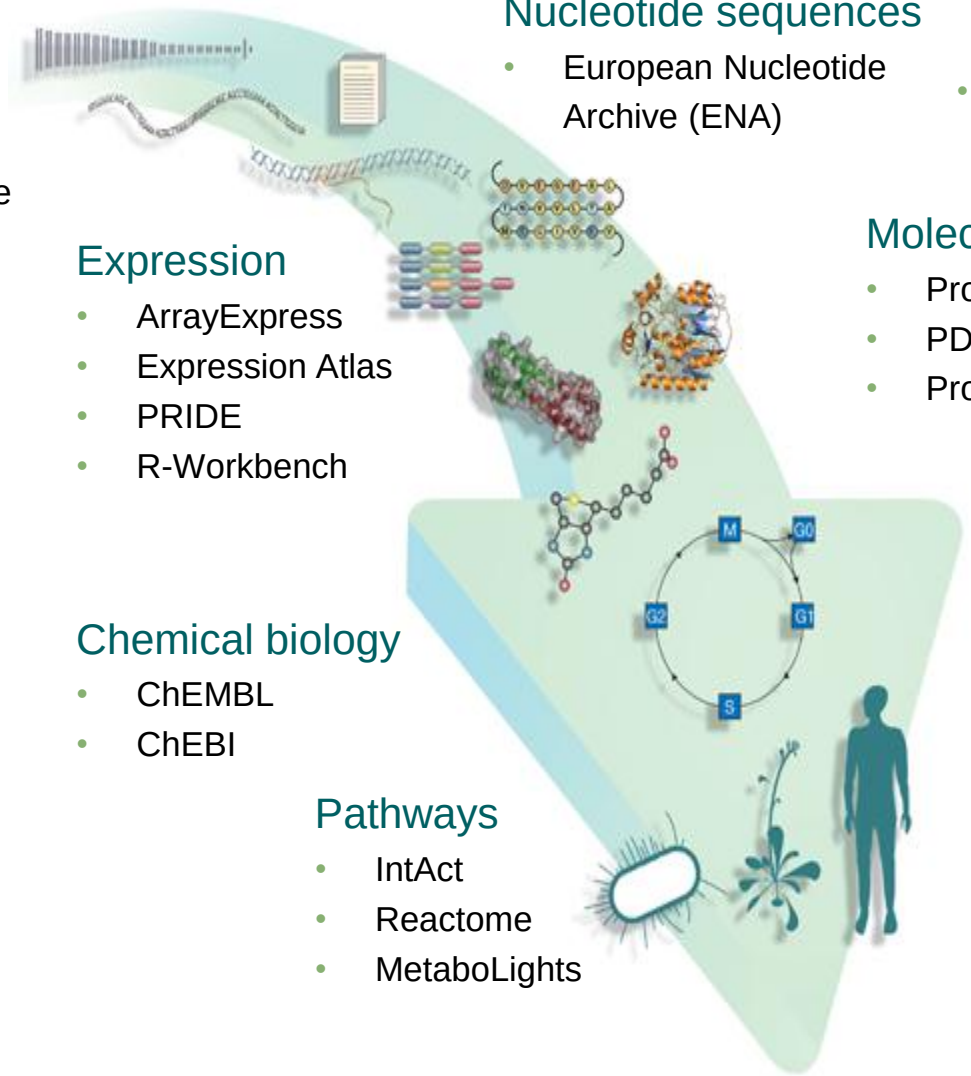
Patent sequences
- Non-redundant patent sequence dbs
- Patent compounds

Nucleotide sequences
- European Nucleotide Archive (ENA)

Expression
- ArrayExpress
- Expression Atlas
- PRIDE
- R-Workbench

Chemical biology
- ChEMBL
- ChEBI

Pathways
- IntAct
- Reactome
- MetaboLights

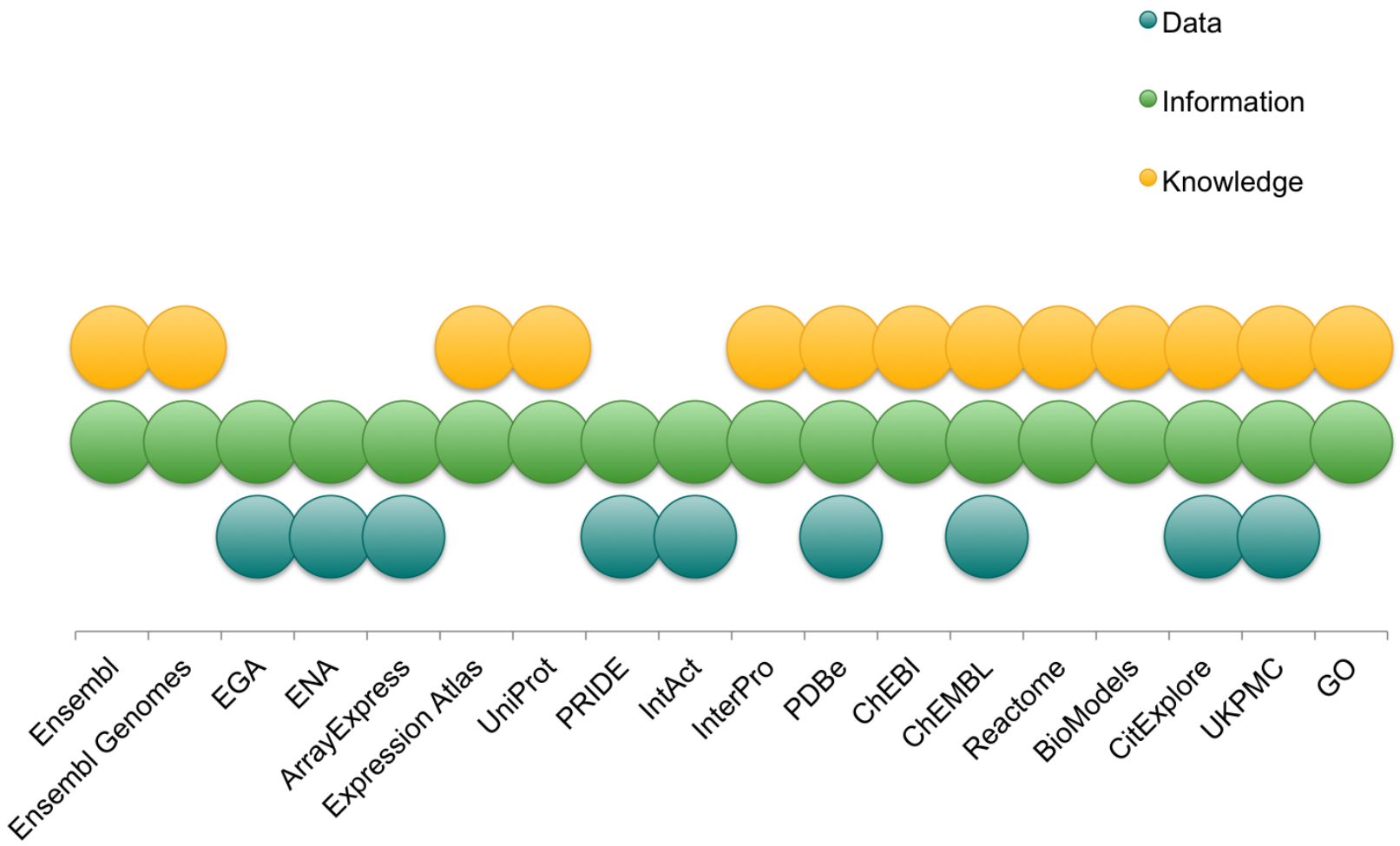


Literature & ontology
- Europe PubMed Central
- Gene Ontology

Molecular structures
- Protein Data Bank in Europe
- PDBsum
- ProFunc

Systems
- BioModels
- Enzyme Portal
- BioSamples

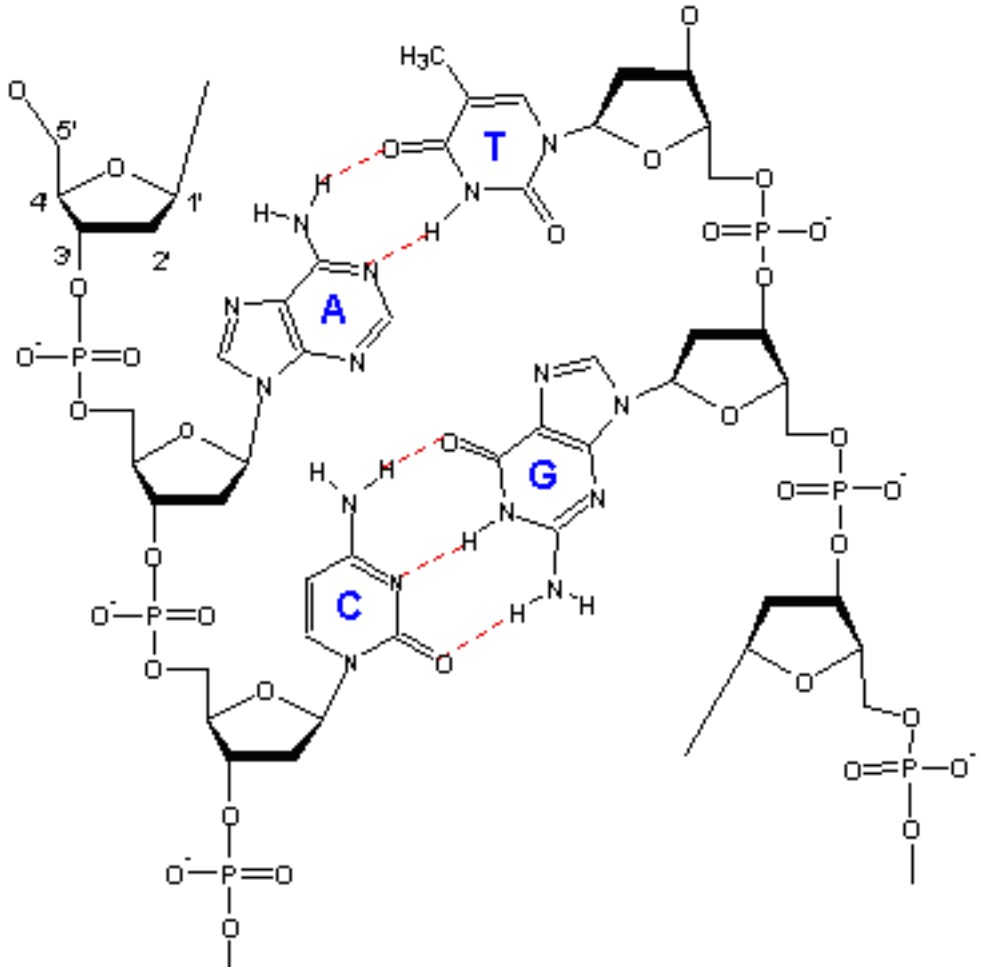
From Data To Knowledge



Crash Course in genomics for geeks

EMBL-EBI

DNA is a covalently linked polymer nearly always found in anti-parallel, non covalent pairs



We represent it as strings, not worrying about one pair of the two polymers

>6 dna:chromosome chromosome:GRCh37:6:133017695:133161157:1
GCAGCAAGACAGAAGTGACTCATACATACAAGGGATCCCCAATAAGATTATCGGCAGATT
TCTCATCAATAACTTTGGAGACCACAAAGCATTGAGCTGATATATTTAAAGTACTGAAAG
AAAAAAAAATCTGACAACCAAGAATTCTATATCCATCAGAACTGCCCTTCAAAAGGGAGG
GAGAAATGAAGACATTCTCAGATTTGAGAAGAAAGGAAAGAGAGAAGGGAGGGGAGGGGA
GAGGAGGGGAGGGGAGGAGAGGAGAGGAGAGGGCACAGTGGCTCACGCCTGTAATCCTAG
CACTTTGCAAGACTGAGGCCAGTGGAACACCTGAGGTCAGGAGATCGAGACCATCCTGGC
TAACACGGTGAAACCCCGTCTCCACTAAAAATACAAAAAATTAGCCAGGCGTGGTGGCAG
GTGCCTGTAGTTCCAGCTACTCAGGAGGCTGAGGCAGCAGAATGGCGTGAACTCGGGAGG
TGGAGCTTGCAGTGAGCTGAGATTGCGCCCCTGCACTCCAGCCTGGGTGACAGAGTGAGA
CTCTGTCTCAAAAAAATAAAAAGTTTAAAAATATTTTAAAAAAAGAAAGAAAGAAGGGAG

1 monomer is called a “base pair” – bp

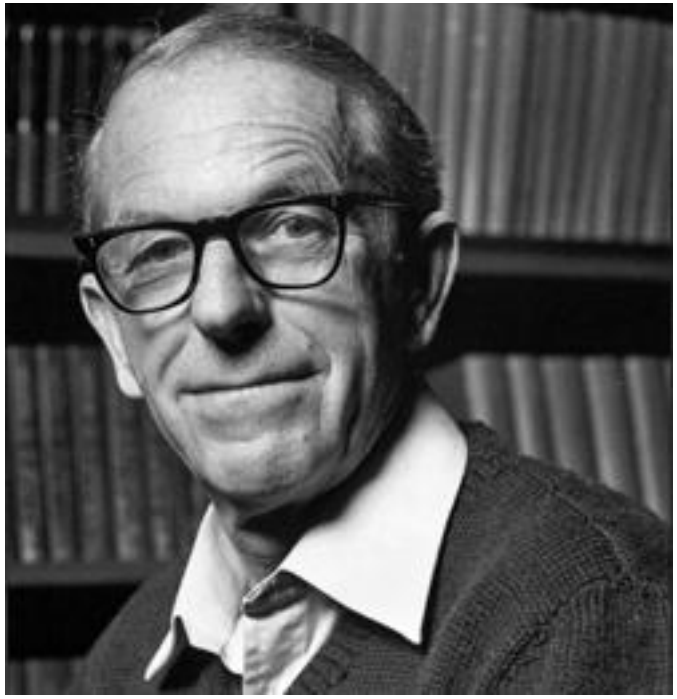
We can routinely determine small parts of DNA

1977-1990 – 500 bp, manual tracking

1990-2000 – 500 bp, computational tracking, 1D, “capillary”

2005-2012 – 20-100bp, 2D systems, (“2nd Generation” or NGS)

2012 - ?? >5kb, Real time “3rd Generation”

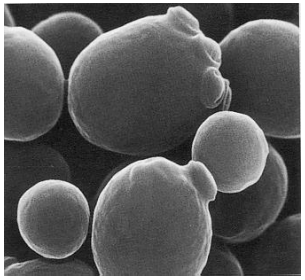


Fred Sanger, inventor of terminator DNA sequencing

A genome is all our DNA

Every cell has two copies of 3e9bp (one from mum, one from dad) in 24 polymers (“chromosomes”)

Ecoli: 4e6,



Yeast, 12e6

Medaka, 0.9e9

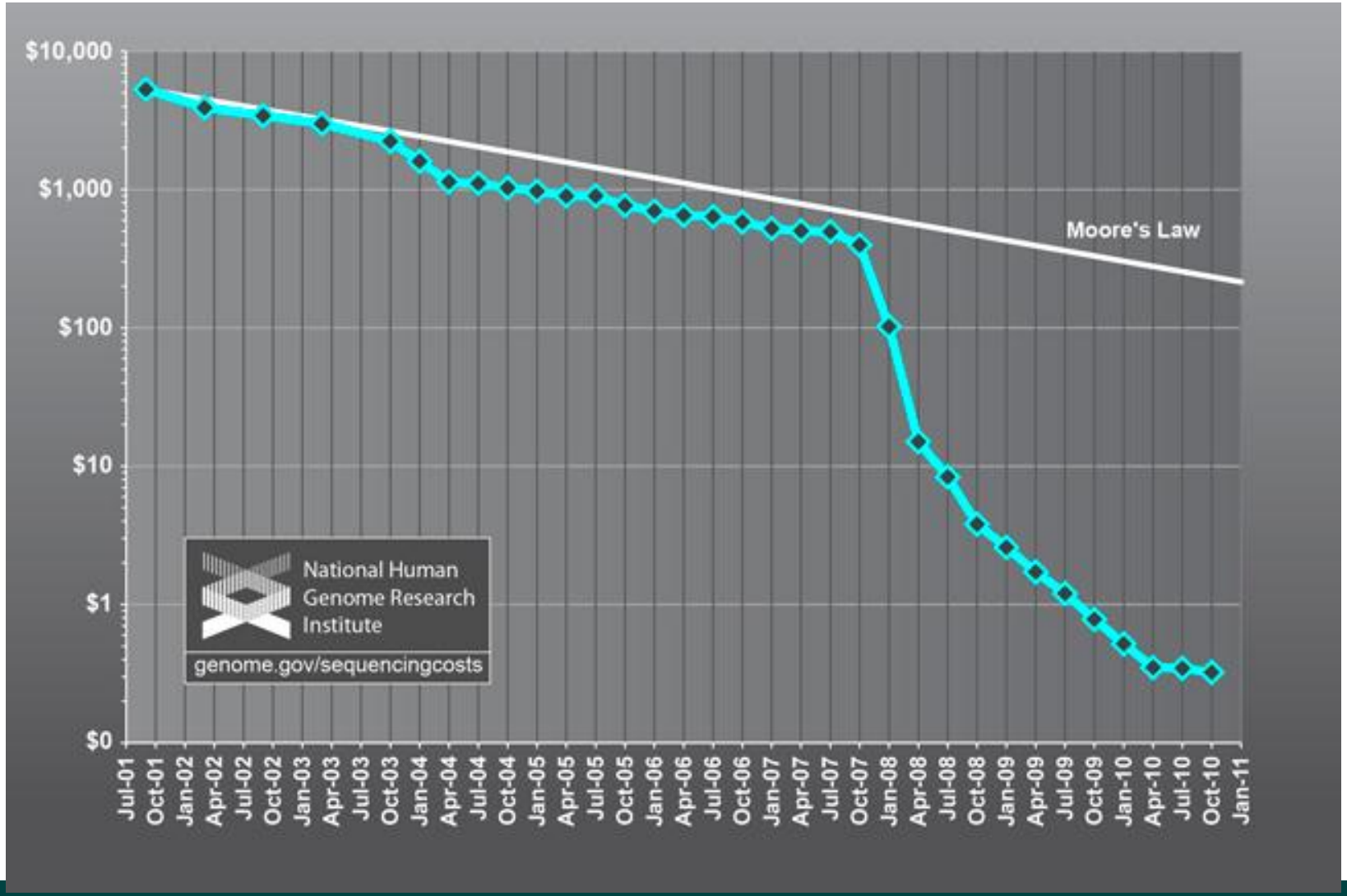
White Pine 20e9

Human Genome project

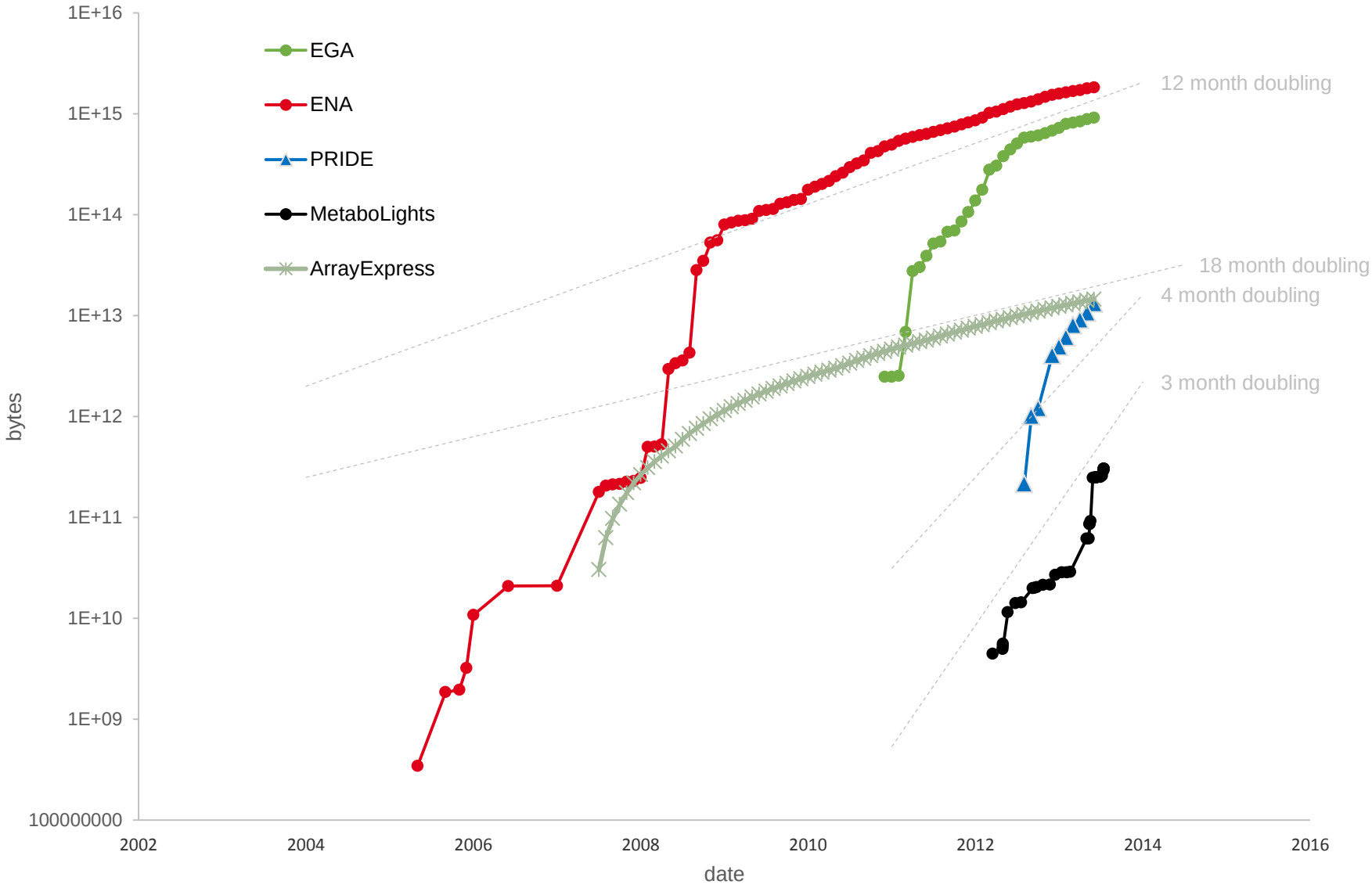
- 1989 – 2000 – sequencing the human genome
  - Just 1 “individual” – actually a mosaic of about 24 individuals but as if it was one
  - Old school technologies

- Now
  - Same data volume generated in ~3mins in a current large scale centre
  - It’s all about the *analysis*

Costs have come exponentially down



Data growth


1E+16
1E+15
1E+14
1E+13
1E+12
1E+11
1E+10
1E+09
100000000
bytes
2002
2004
2006
2008
2010
2012
2014
2016
date
EGA
ENA
PRIDE
MetaboLights
ArrayExpress
12 month doubling
18 month doubling
4 month doubling
3 month doubling

EBI's technical infrastructure

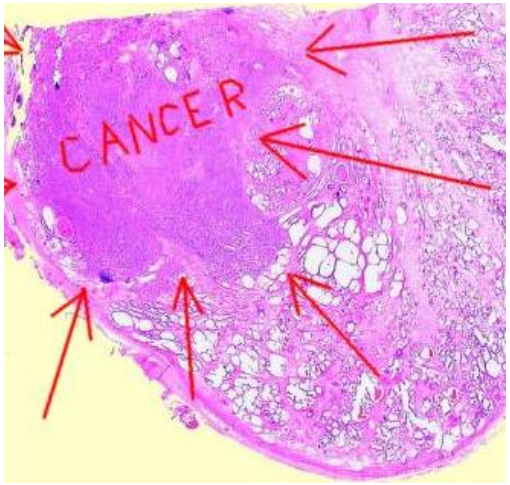
- 30 PB of "raw" disk
  - Big archives on two systems, no tape backup (analysis is recovery would be very hard; disaster recovery by institutional replication in US)
- ~20,000 cores in 2 major farms
- A VMware Cloud ("Embassy Cloud") allowing remote users to directly mount large datasets (in pilot mode)
- 4 machine rooms; 2 in London, 2 in Cambridge
- Janet uplink at 10 Gbit/sec

Impact on Medicine

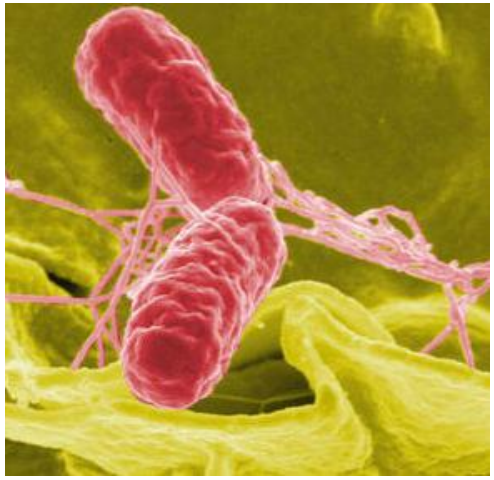
EMBL-EBI

3 big areas of impact for medicine

Germ line
Risk to disease



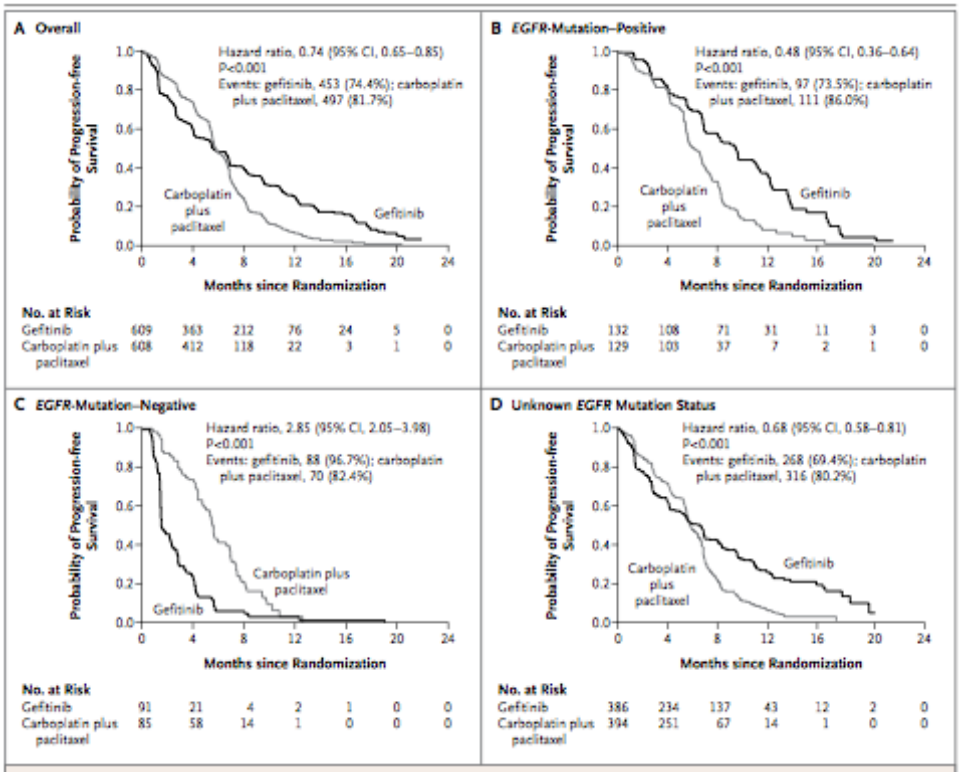
“Precision” cancer medicine



Pathogens +
Hospital acquired infections

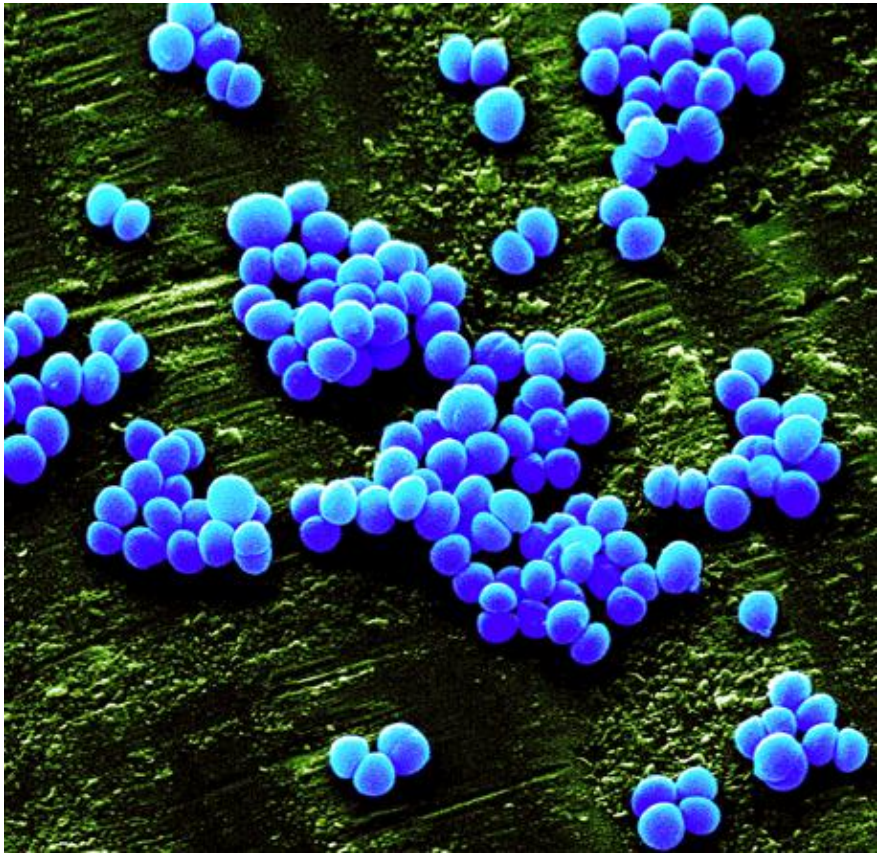
Precision cancer diagnosis

- Cancer is a genomic disease
- By sequencing a cancer you can understand its molecular form better
- Particular molecular forms respond to particular drugs



Pathogens

- Sequencing provides a clear cut diagnosis of pathogens
- Can also be used to sequence environments (eg, hospitals)



Why do we need ELIXIR?

- Creating a robust infrastructure for biological and clinical information is a bigger task than any individual organisation or nation can take on alone
- Life Sciences has huge data needs and by far the largest research community:
  - Data deluge - 30 Petabytes storage at EMBL-EBI
  - ~3 million life science researchers in Europe
  - >9 million web hits a day at EMBL-EBI alone
  - 1 million unique users per year

ELIXIR

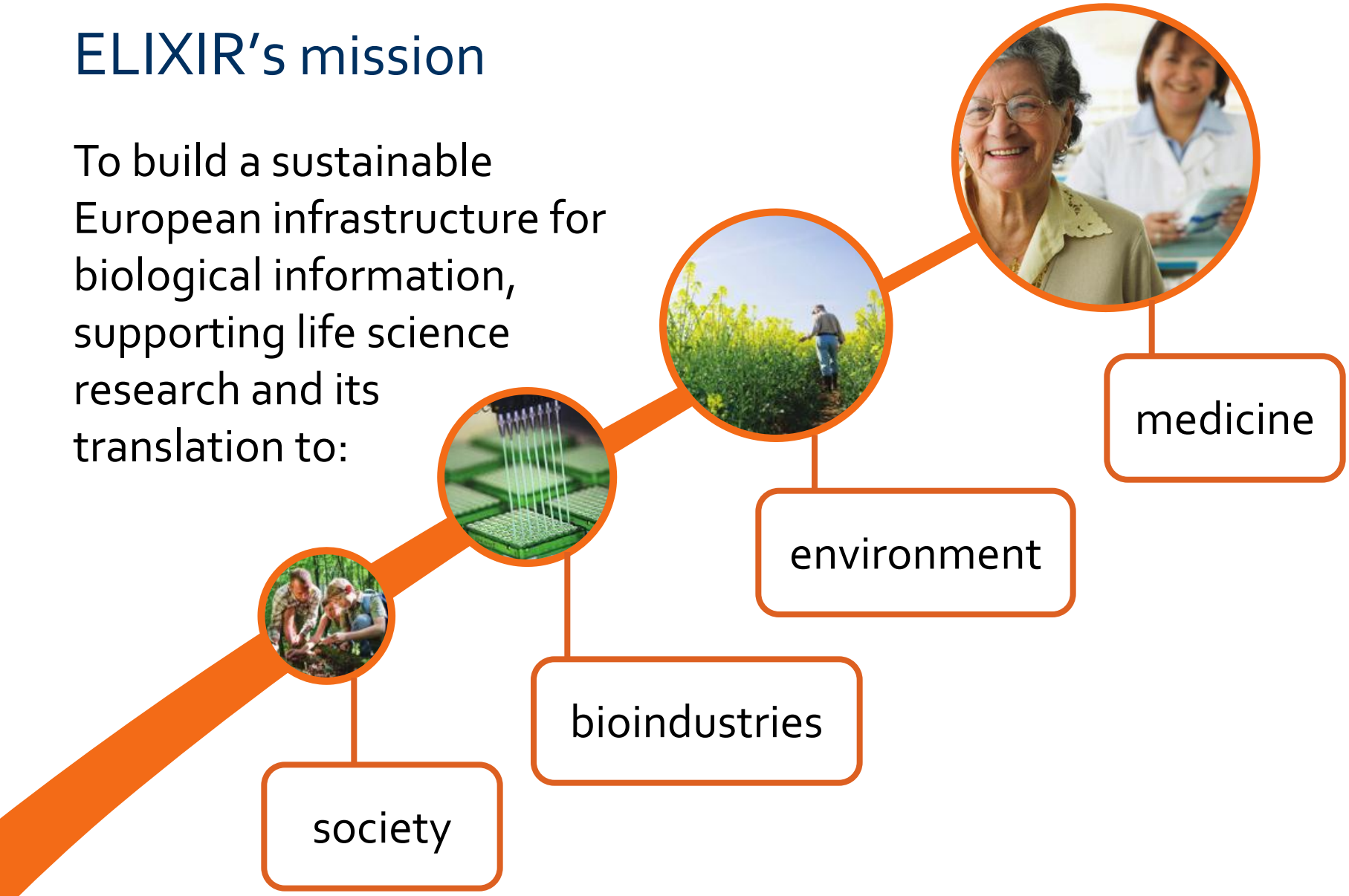
Safeguarding the results of life science research in Europe

European Life Sciences Infrastructure for Biological Information

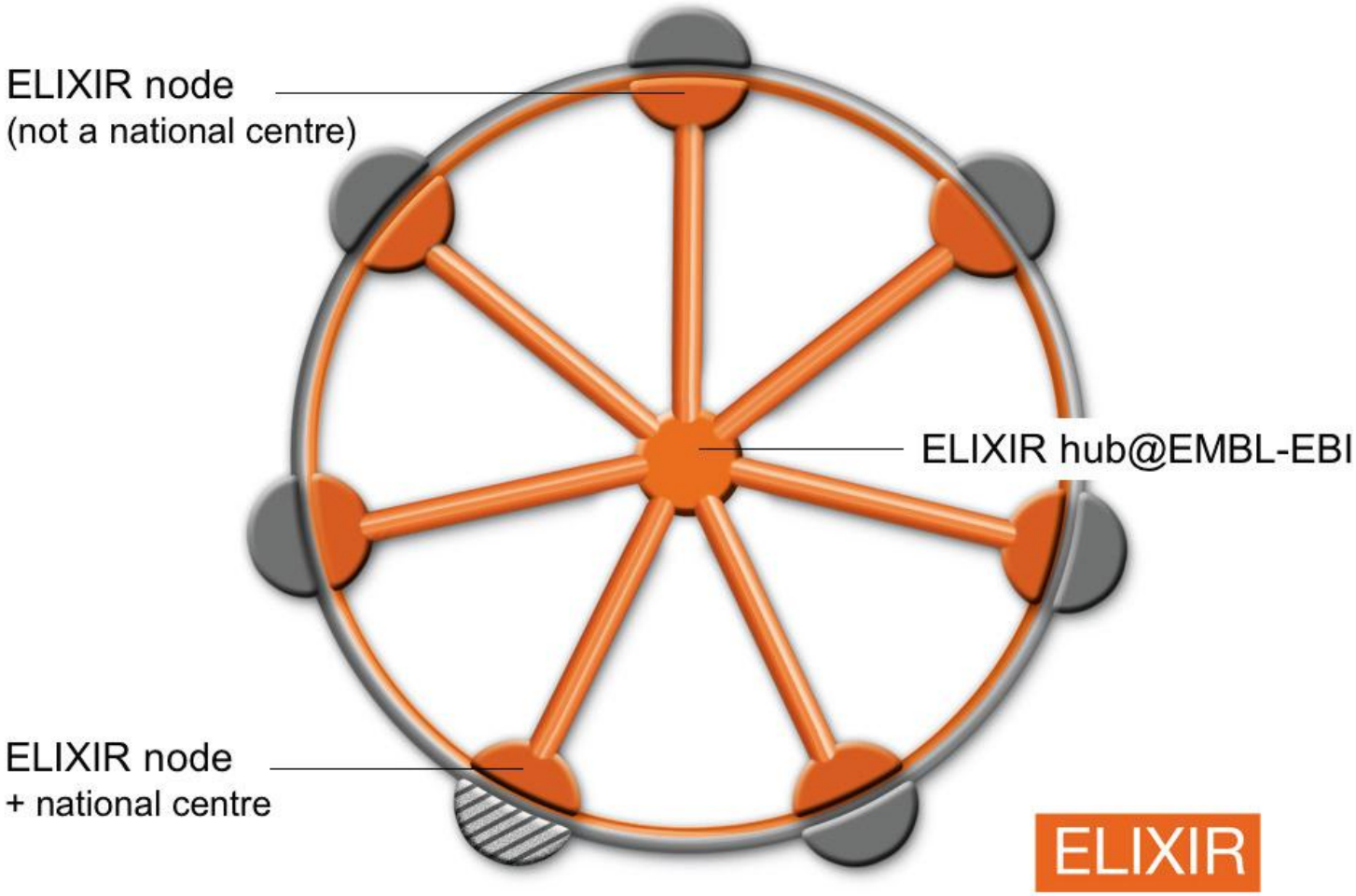
www.elixir-europe.org

ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:



A distributed pan-European infrastructure



Services offered by ELIXIR

- ELIXIR services are open access and free of charge:
- **Data** – Global access to biological data including human, animal, crop and marine: Very large user community (e.g. EMBL-EBI >9 million hits per day)
- **Tools -** Integration of existing tools to enable data access and mining by developing an interoperable tools infrastructure
- **Training -** Deployment of specialist and general training courses and workshops, including eLearning. 'Train the trainer' activities for new Member States
- **Standardisation -** Coordinate development of standards for biological and medical nomenclature and controlled vocabularies and ontologies
- **Industry** - Support to industry through localised, bespoke projects and SME training

Summary - ELIXIR members co-ordinate their national bioinformatics efforts, reduce fragmentation and providing users with simple interface for data

Sixteen countries have signed up

- 16 countries plus EMBL have signed the Memorandum of Understanding (MoU) to participate
- Countries now work towards signing ELIXIR Consortium Agreement (ECA)
- UK, Sweden and Switzerland have signed ECA
- More are expected to follow in the coming months…

Czech republic
Denmark
Estonia
Finland
Greece
Israel
Italy
Netherlands
Norway
Portugal
Slovenia
Spain
Sweden
Switzerland
United Kingdom

Future challenges for life-science data services