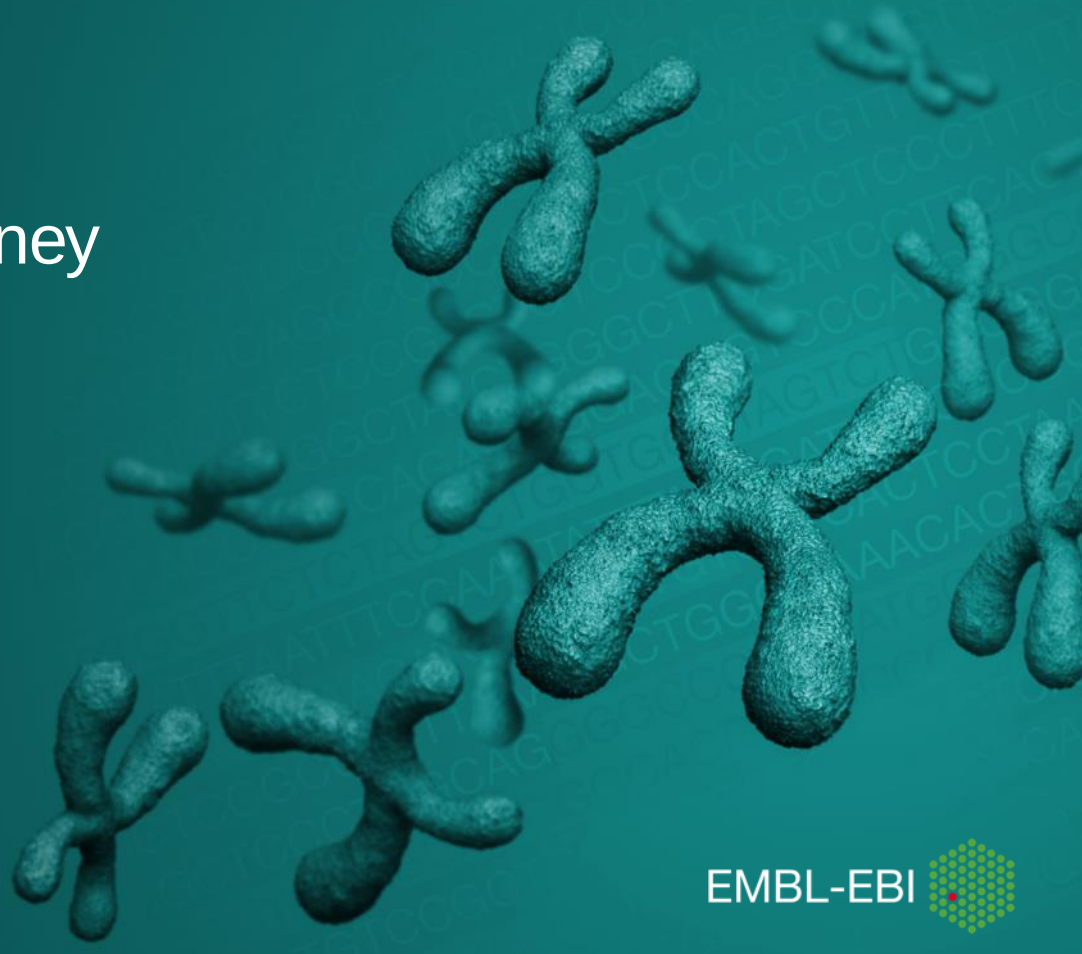
Information in the Lifesciences… The Science of the 21st Century …and why it needs infrastructure

Ewan Birney @ewanbirney

(Tweetable)



EMBL-EBI

Outline of the talk

- What is EMBL-EBI?
- Genomics 2000-2012 as an example driver
  - HapMap, 1000 Genomes, GWAS, ENCODE
- A taster of some computational problems
  - Conceptually
  - Scaling

- (Some more whimsical uses of DNA…)

What is EMBL-EBI?

- European Bioinformatics Informatics (EBI) is part of the European Molecular Biology Laboratory (EMBL)
- EMBL was founded in 1974, headquartered in Heidelberg
- Centre of excellence for molecular biology research
- Also provides services for molecular biology

Molecular Biology

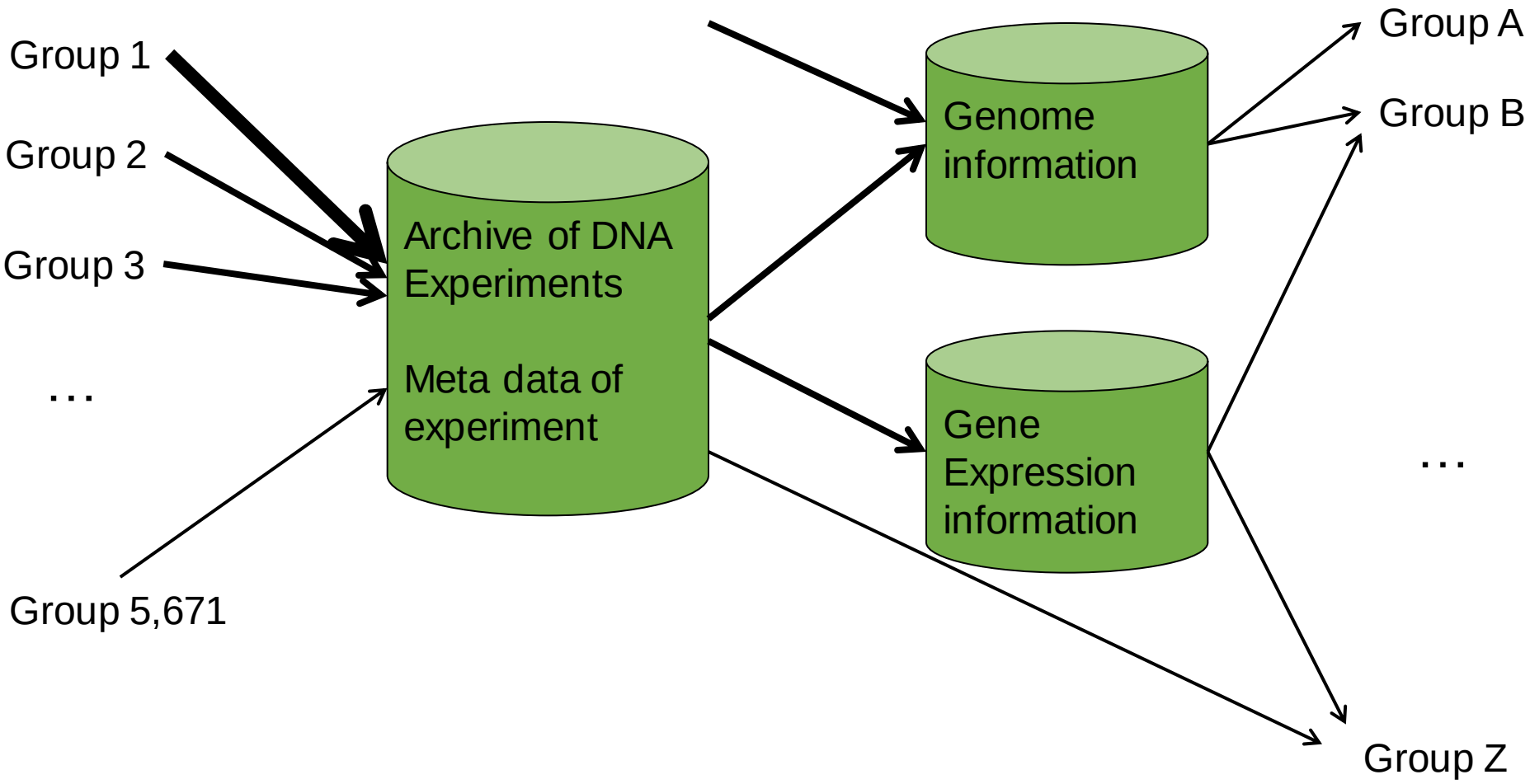
- The study of how life works – at a molecular level
- Key molecules:
  - DNA – Information store (Disk)
  - RNA – Key information transformer, also does stuff (RAM)
  - Proteins – The business end of life (Chip, robotic arms)
  - Metabolites – Fuel and signalling molecules (electricity)
- Theories of how these interact – no theories of to predict what they are
- Instead we determine attributes of molecules and store them in *globally accessible, open*, databases

EBI's mission

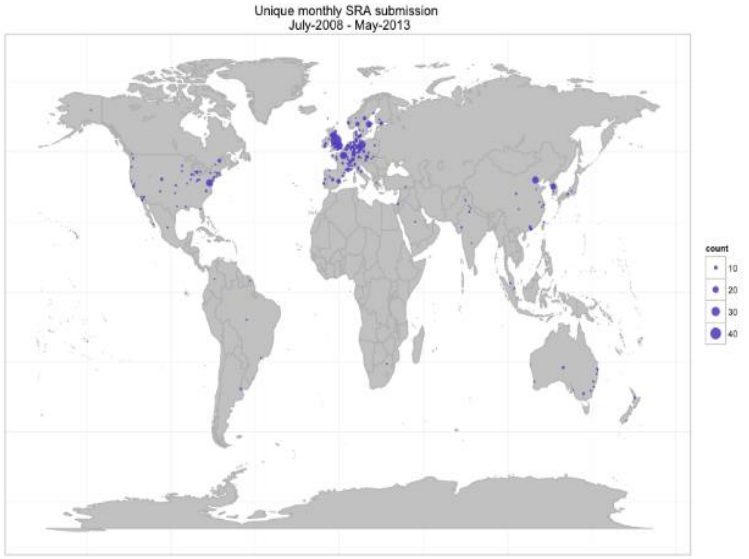
- There are five parts of EBI's mission, of which the biggest part are services.

- To provide global accessible, open data services for the molecular biology community
- To perform world leading research
- To provide appropriate user training for our services
- To support industry
- To help coordinate European efforts in bioinformatics

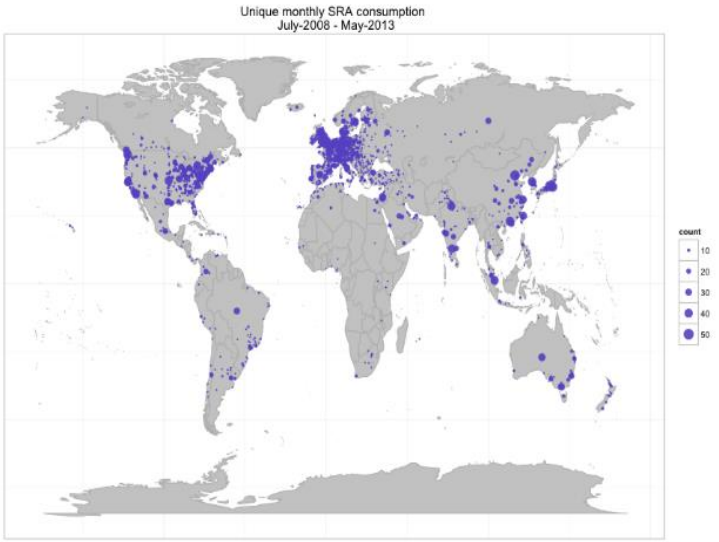
Dataflow in molecular biology



Data Flow

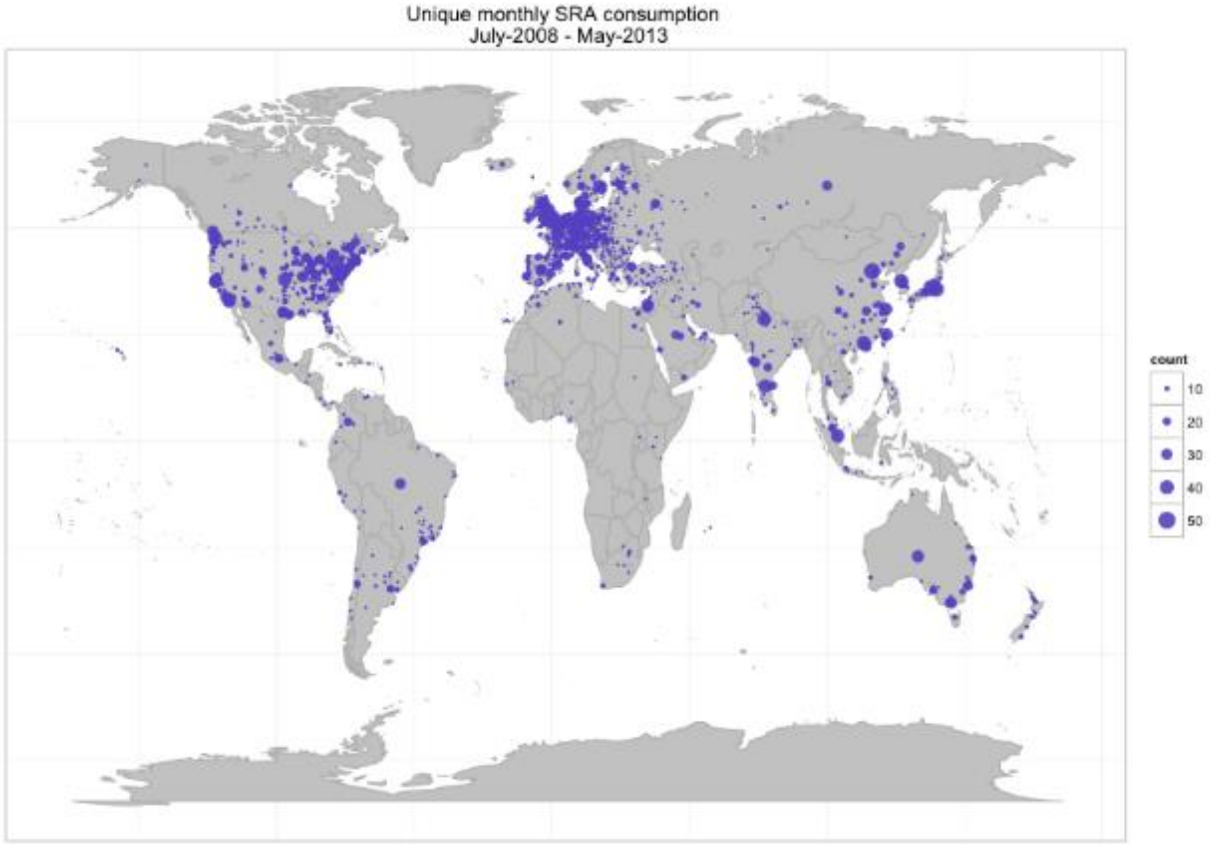


Where we get the data



Where we send the data

Distributed consumption: e.g. sequence data access

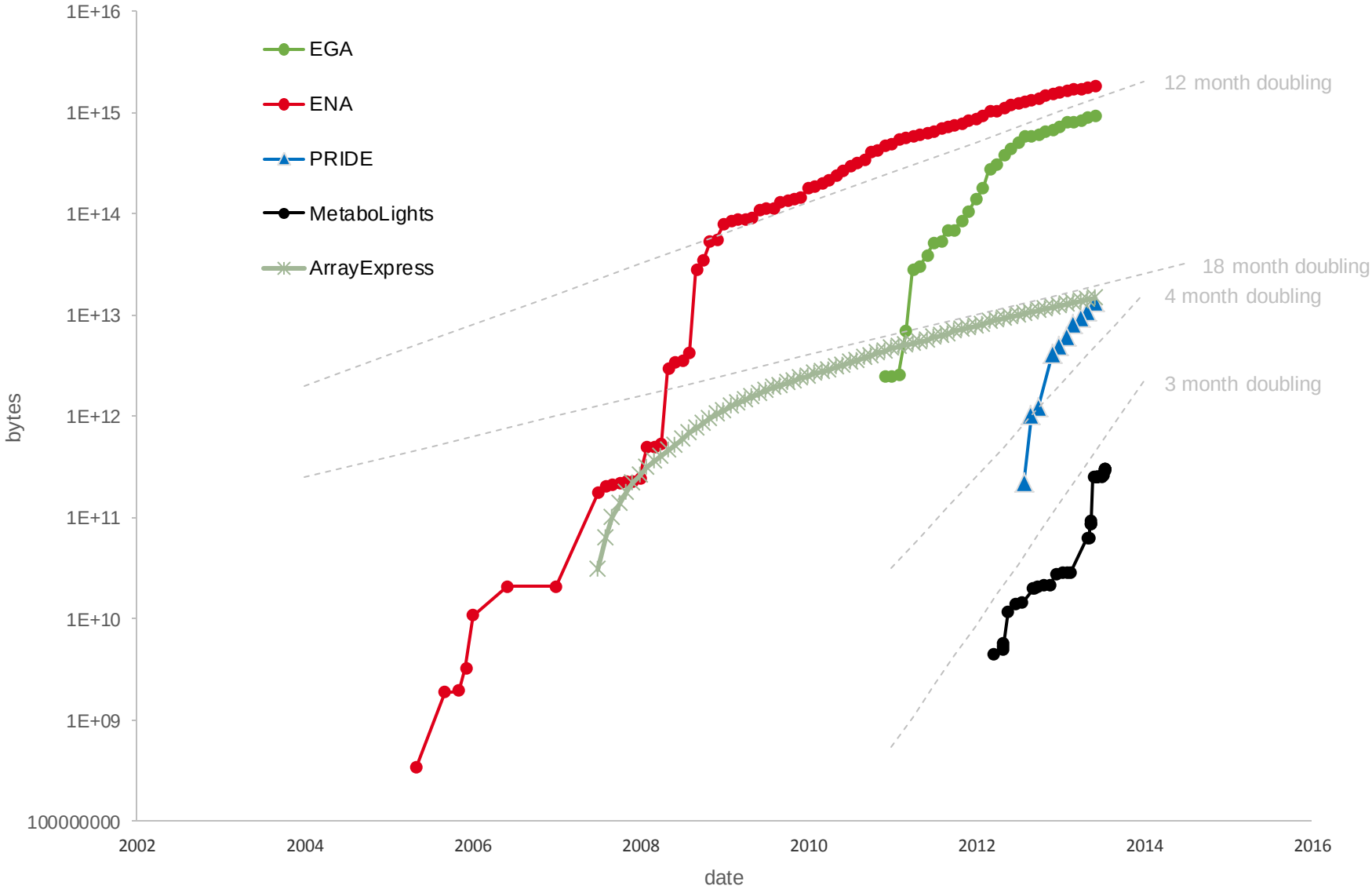


Audience

- There are over 20,000 actively submitting groups over a year across the different “archival” databases

- There are over 100,000 unique IPs that access EBI services each month

Data growth

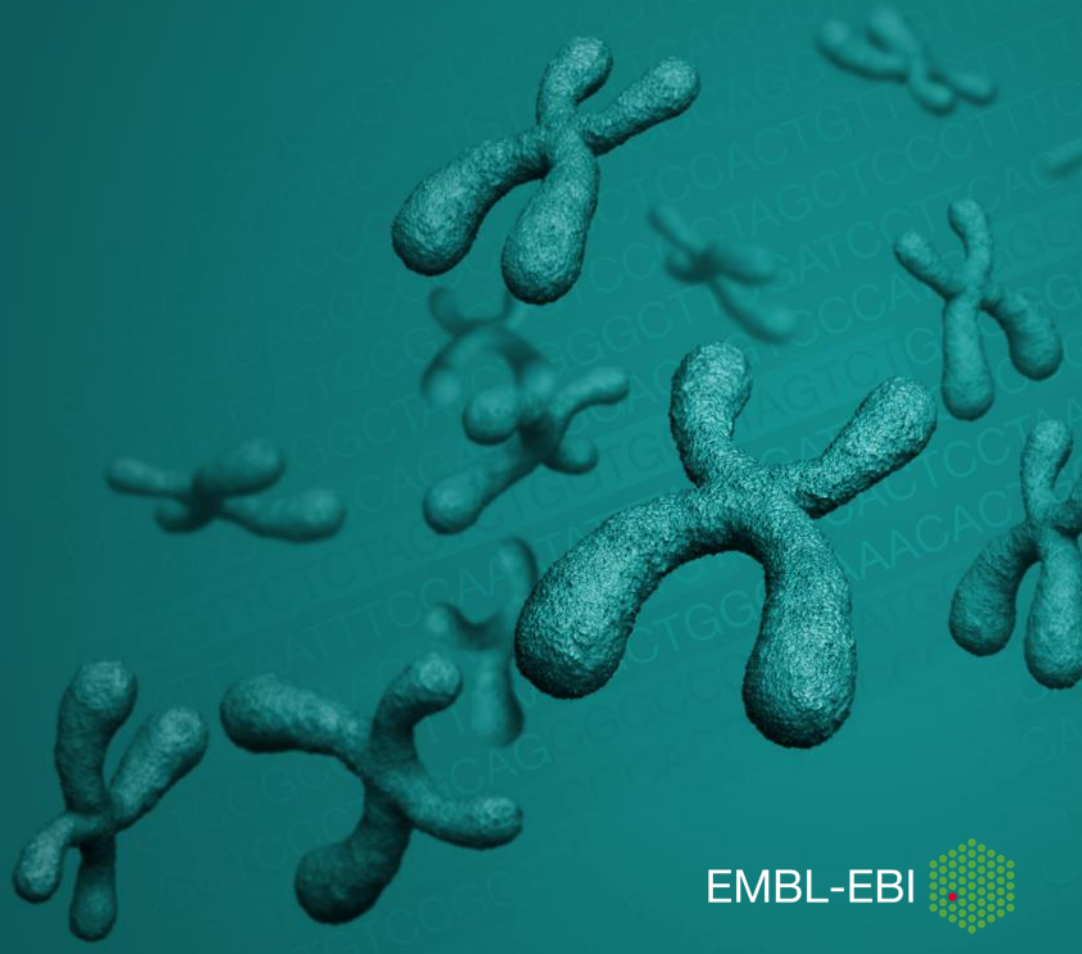

1E+16
1E+15
1E+14
1E+13
1E+12
1E+11
1E+10
1E+09
100000000
bytes
2002
2004
2006
2008
2010
2012
2014
2016
date
EGA
ENA
PRIDE
MetaboLights
ArrayExpress
12 month doubling
18 month doubling
4 month doubling
3 month doubling

This is going to get worse

- As genomics and other molecular biology components get used in medicine…
- …every medical system will need (aspects) of this data
- There an estimated 20,000 hospitals in Europe, each with potentially the need for a reliable source of genomic information

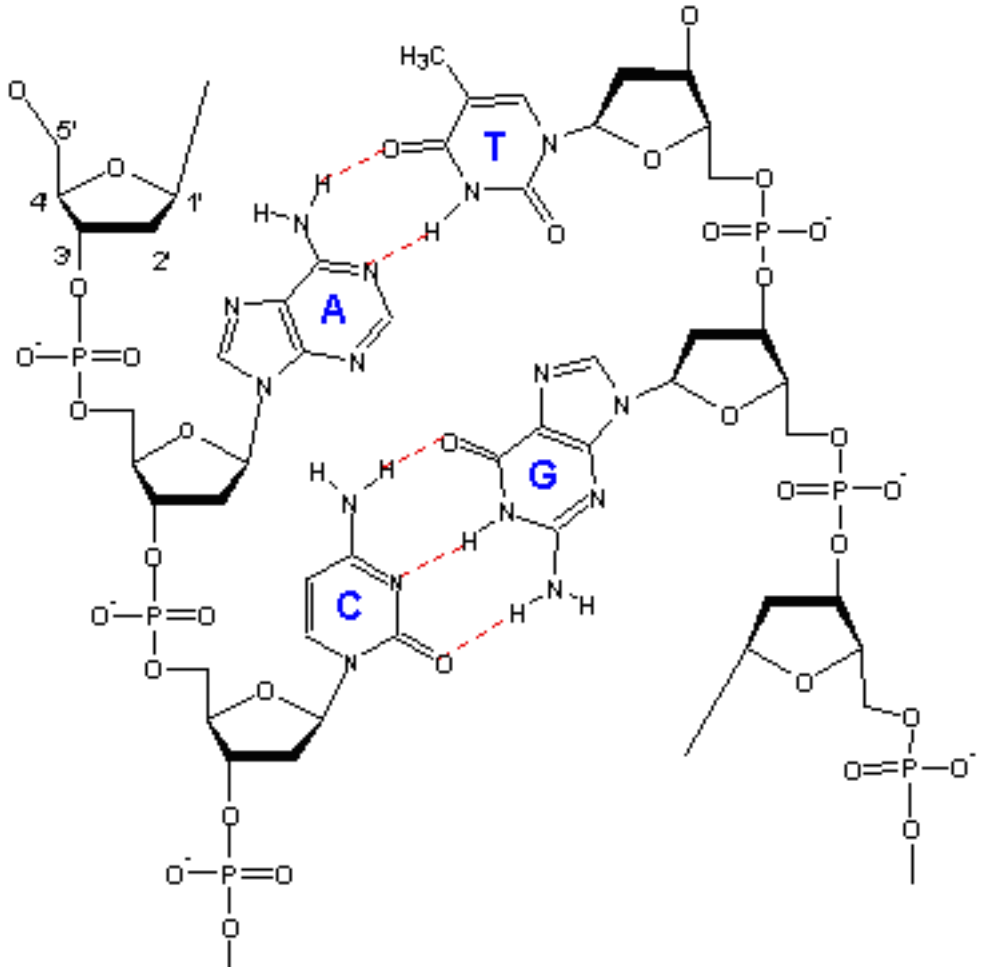
Crash Course in DNA

(one example driver)



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

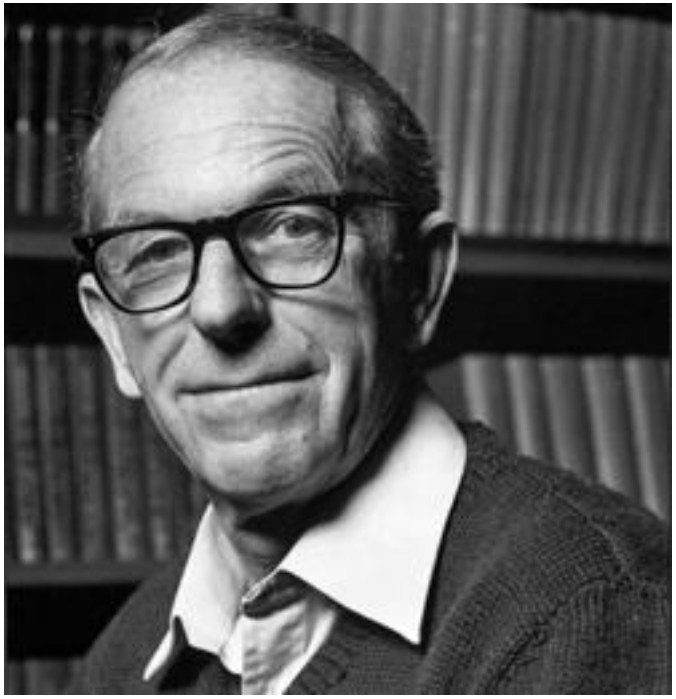
We can routinely determine small parts of DNA

1977-1990 – 500 bp, manual tracking

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

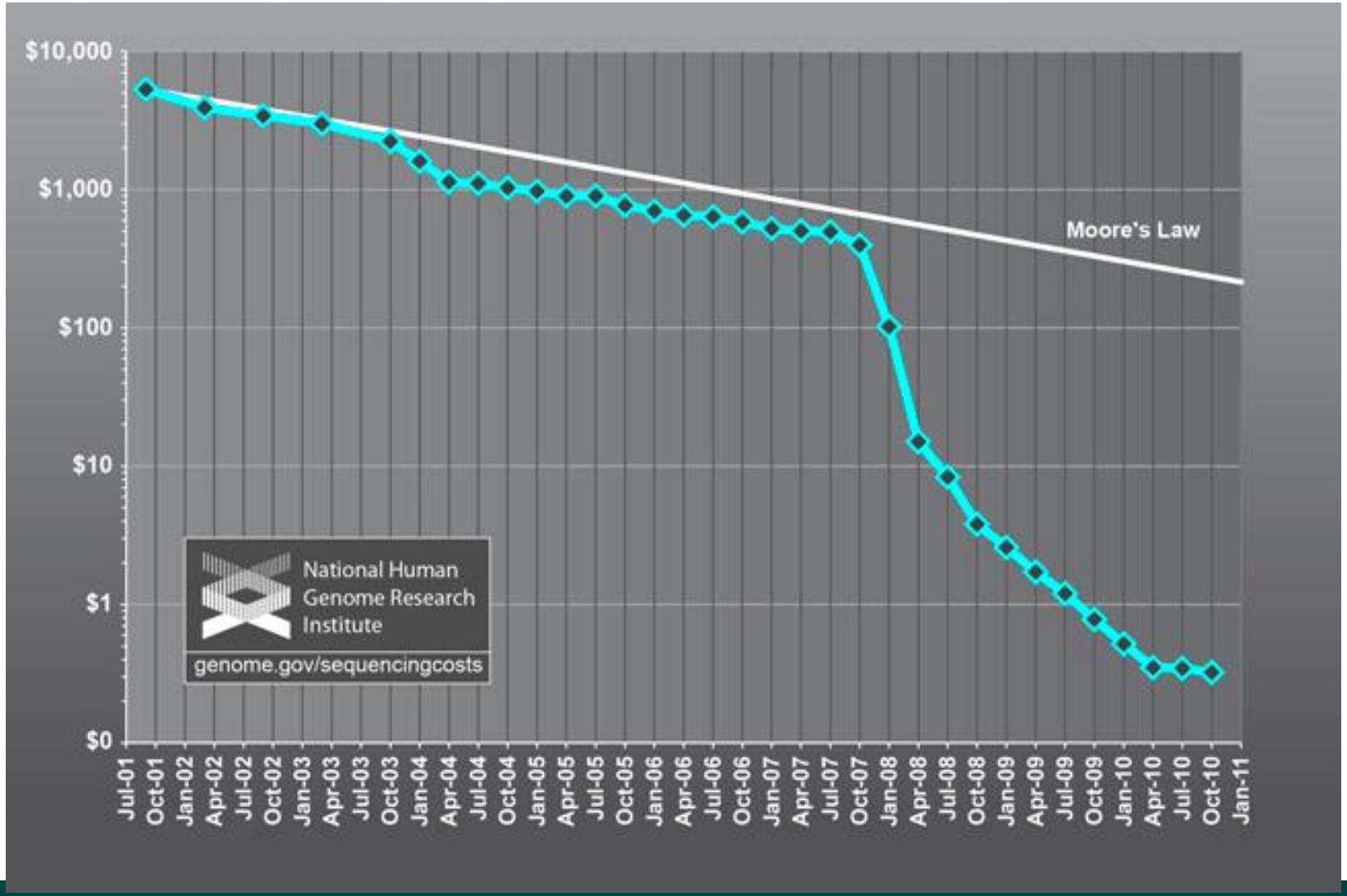
2005-2012 – 20-100bp, 2D systems, (“2nd Generation” or NGS)

2013 - ?? >5kb, Real time “3rd Generation”



Fred Sanger, inventor of terminator DNA sequencing

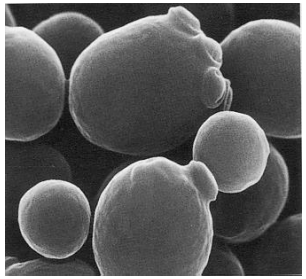
Costs have come exponentially down



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,

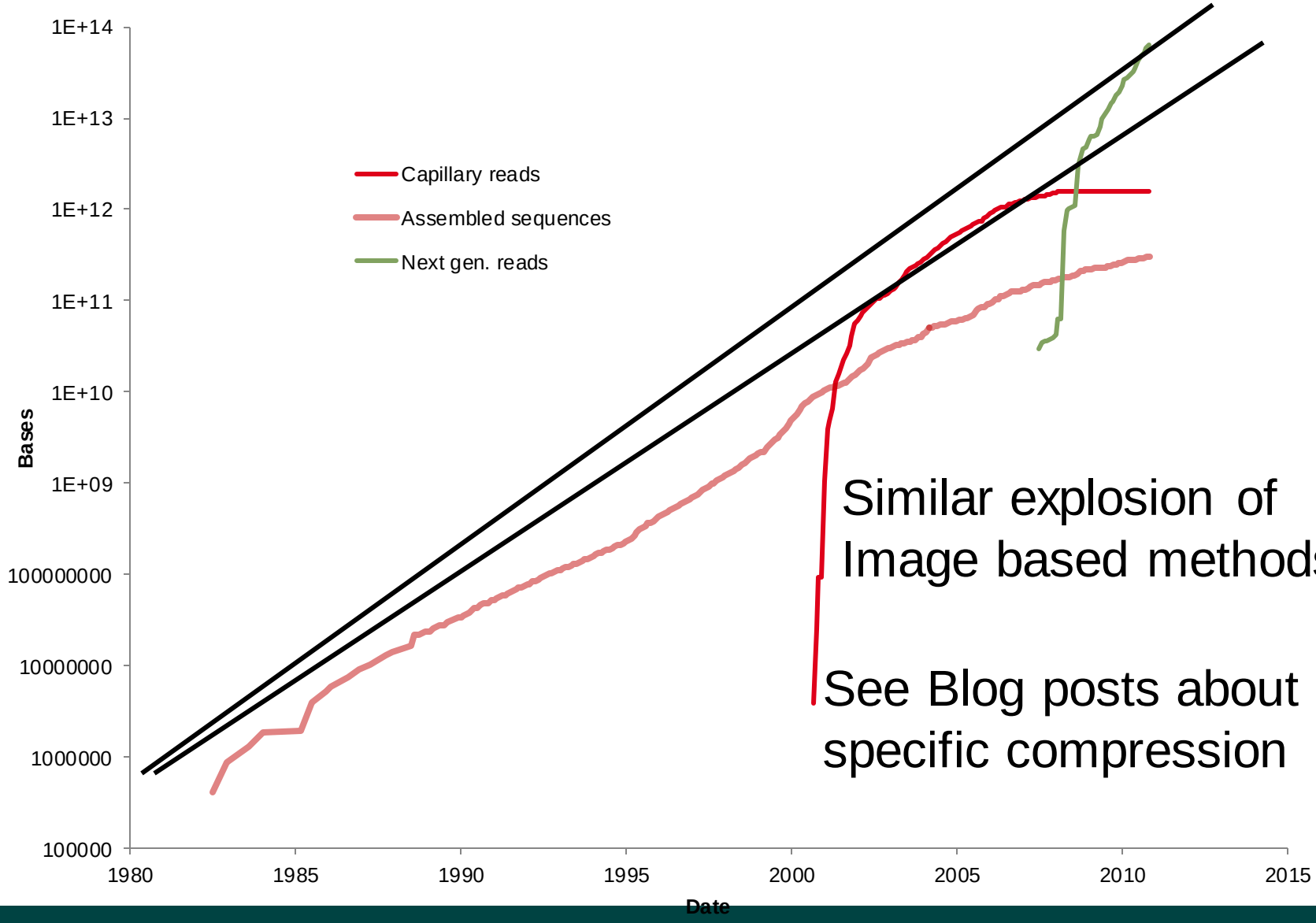


Yeast, 12e6

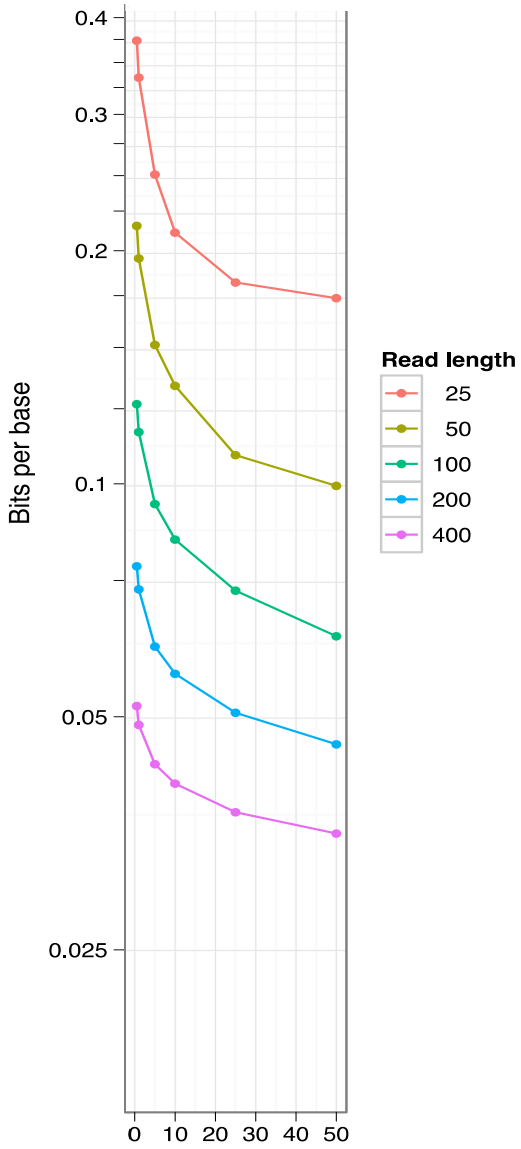
Medaka, 0.9e9

White Pine 20e9

Volumes have gone up!



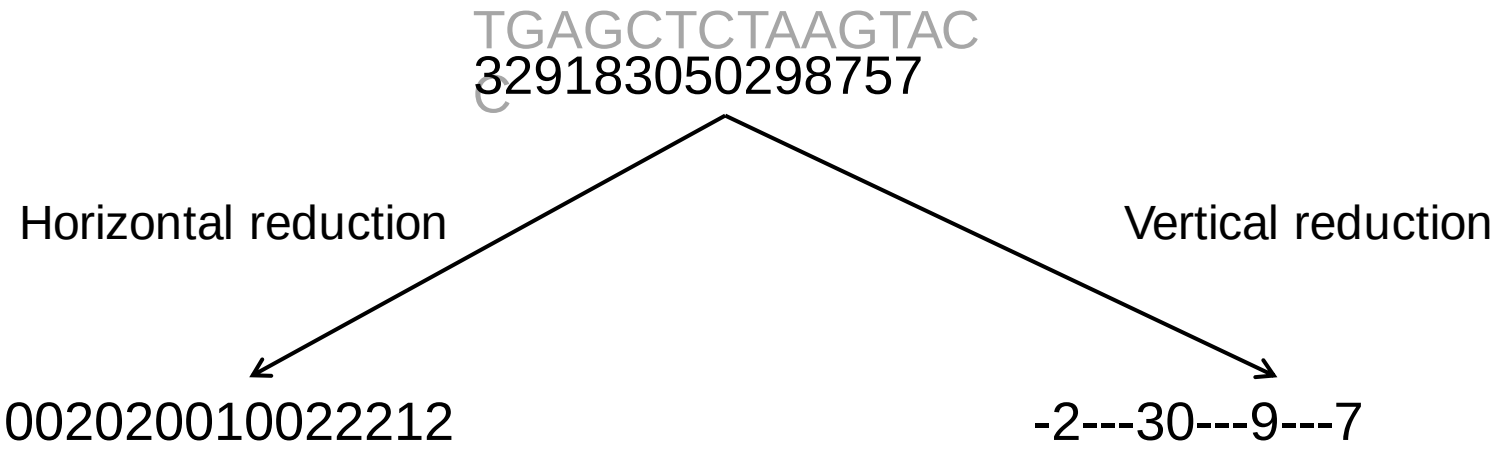
Sequence compression



- Encoding of read starts and differences
- 3.5x–100x compression over existing formats
- Scales favourably with increasing read length and density

Fritz, M.H. Leinonen, R., et al. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40.

Quality compression



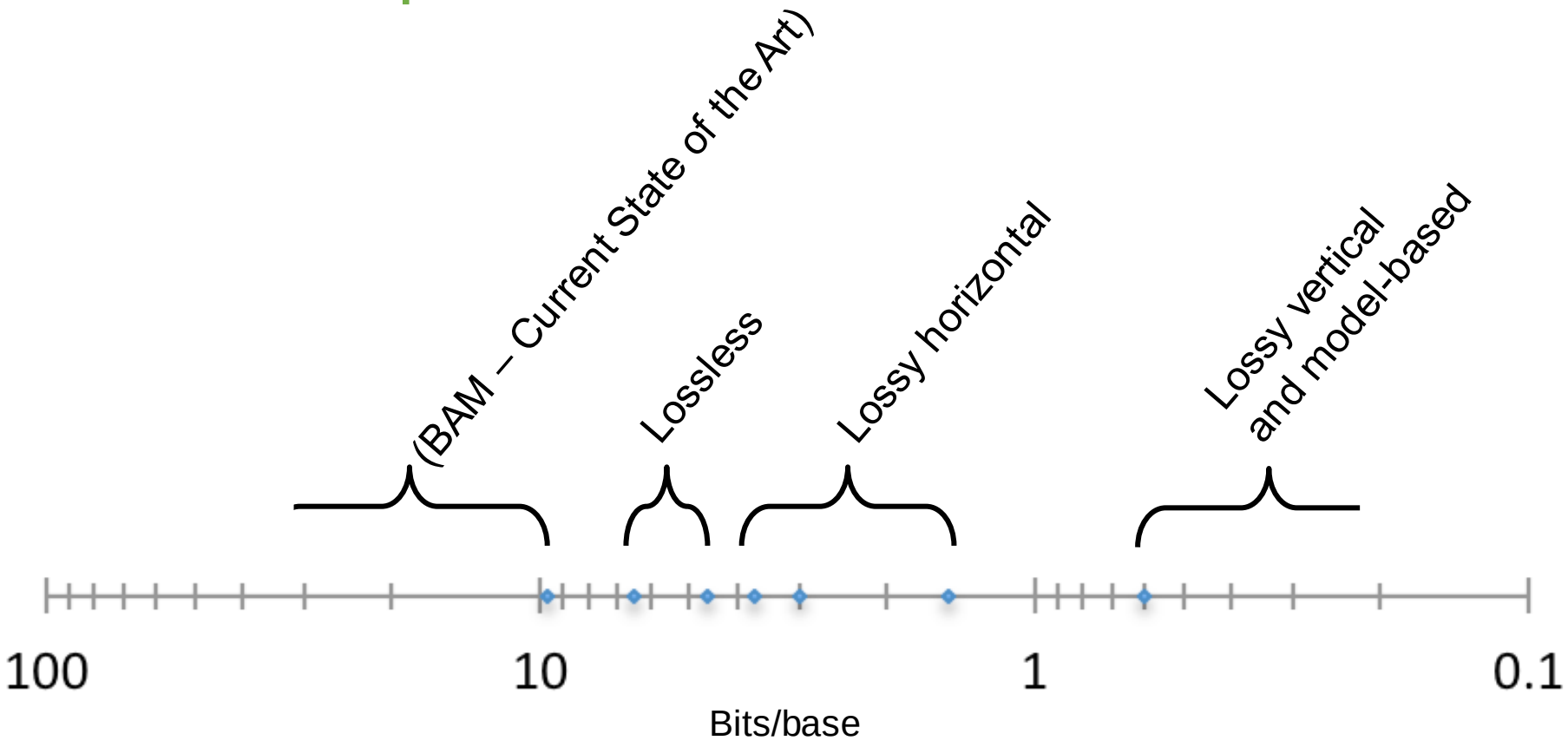
Quality compression now: simple, horizontal reduction

Photograph from MichaelMaggs, http://en.wikipedia.org/wiki/File:Amanita_muscaria_(fly_agaric).JPG

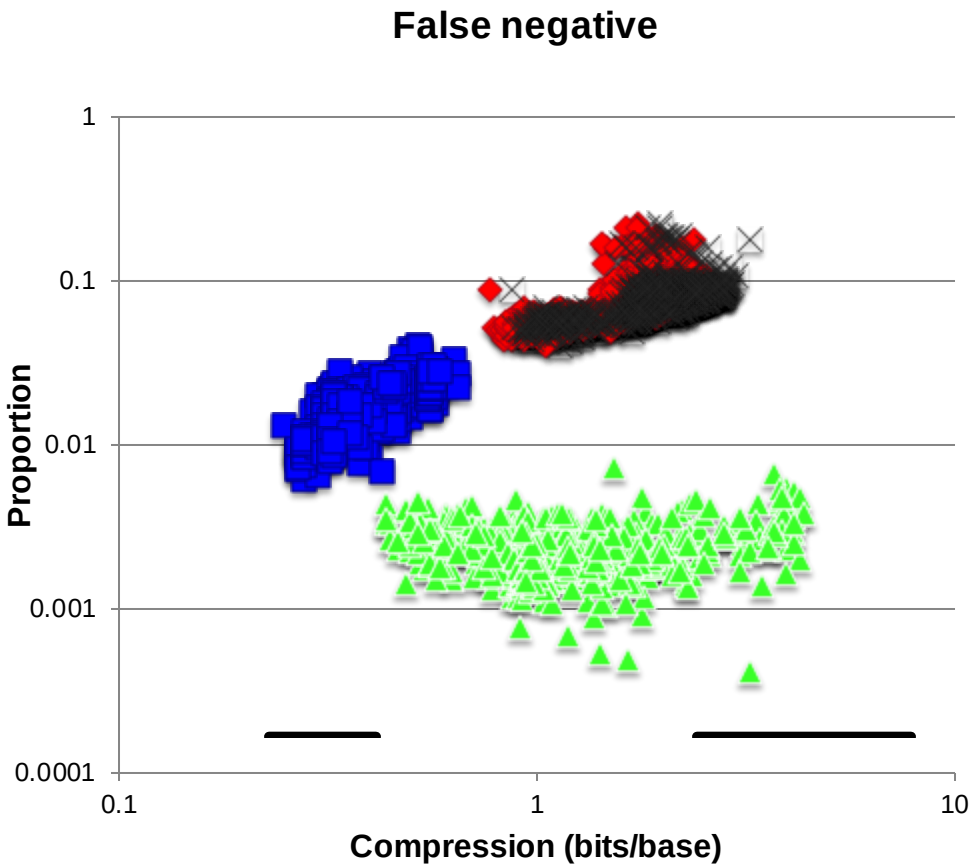
Quality compression now: vertical reduction

Jong-Seok Lee et al. (2009), http://mmspg.epfl.ch/files/content/sites/mmspl/files/shared/lee_icme.pdf

CRAM: in practice



CRAM: impact on analysis

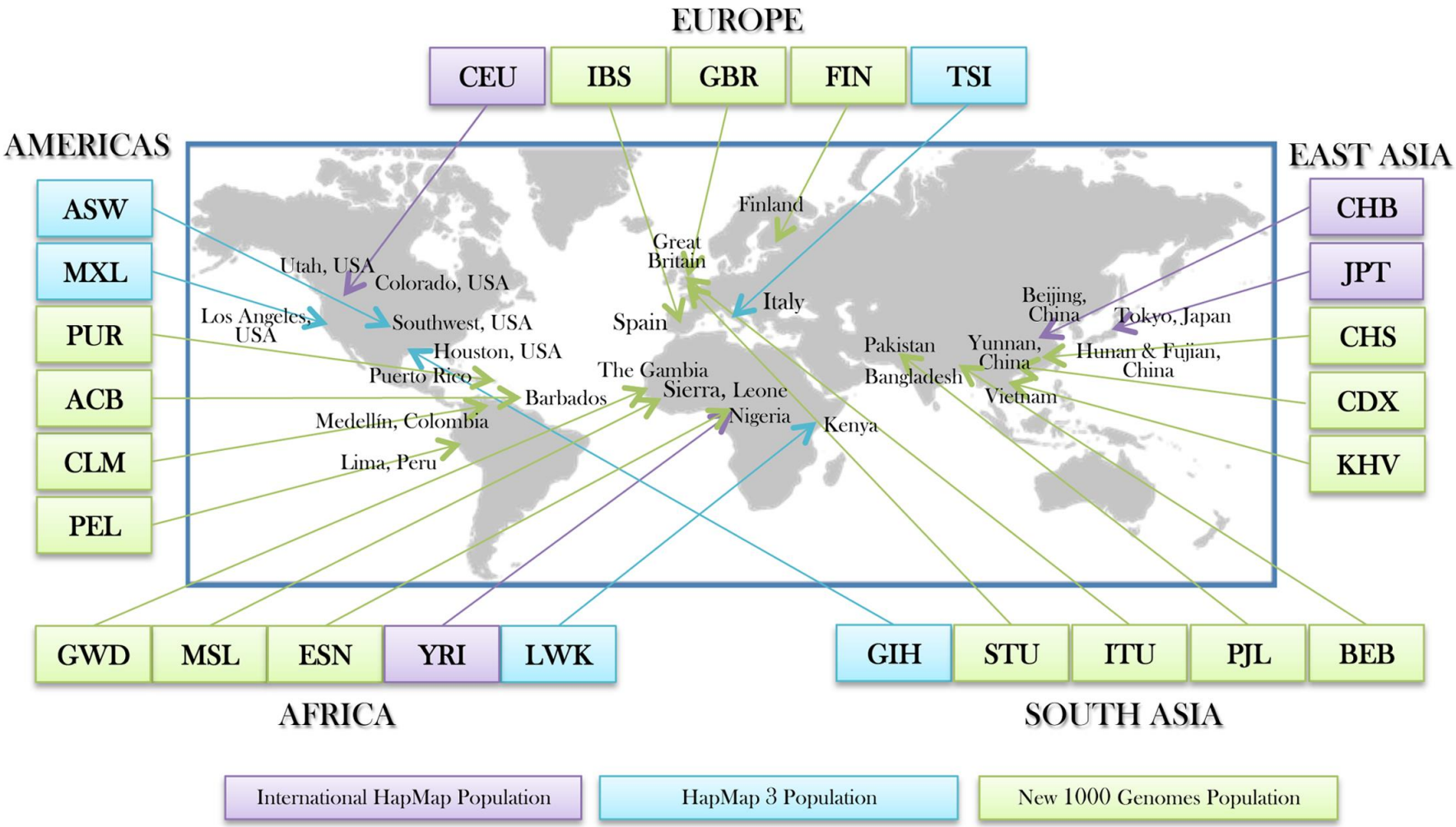


- Ongoing trials with community data: 1000 genomes, CGCI cancer genomes

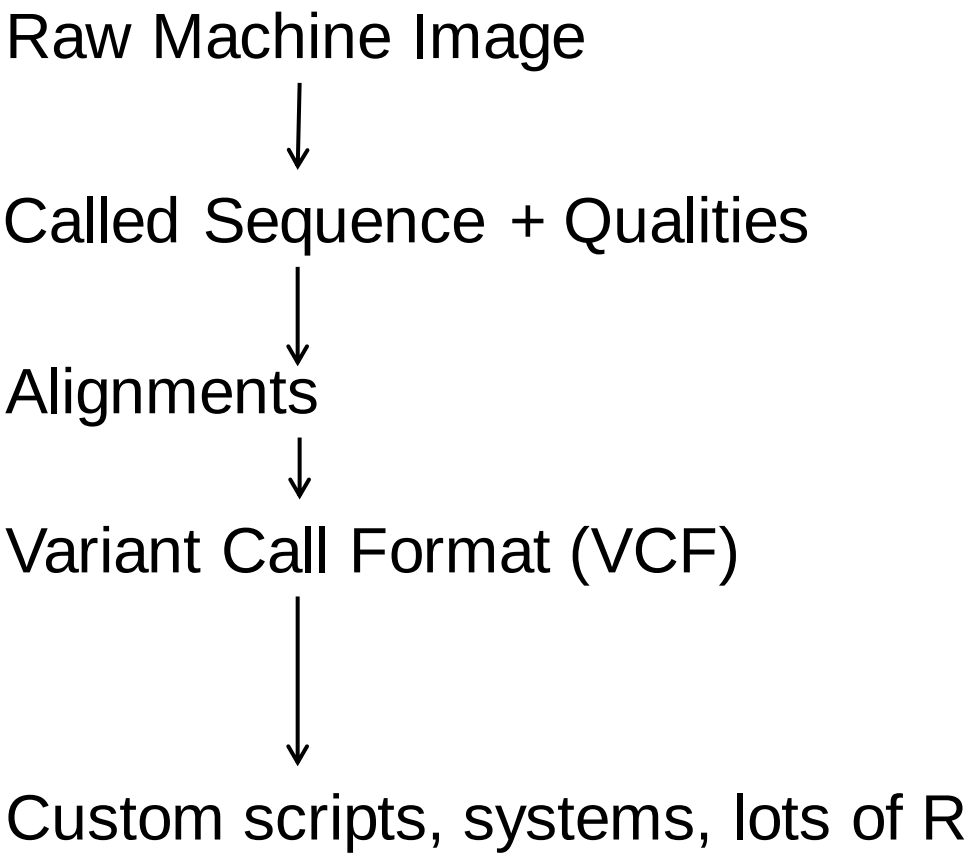
Human Genome project

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

2535 Genomes



Data flow


Raw Machine Image
Called Sequence + Qualities
Alignments
Variant Call Format (VCF)
Custom scripts, systems, lots of R

Size

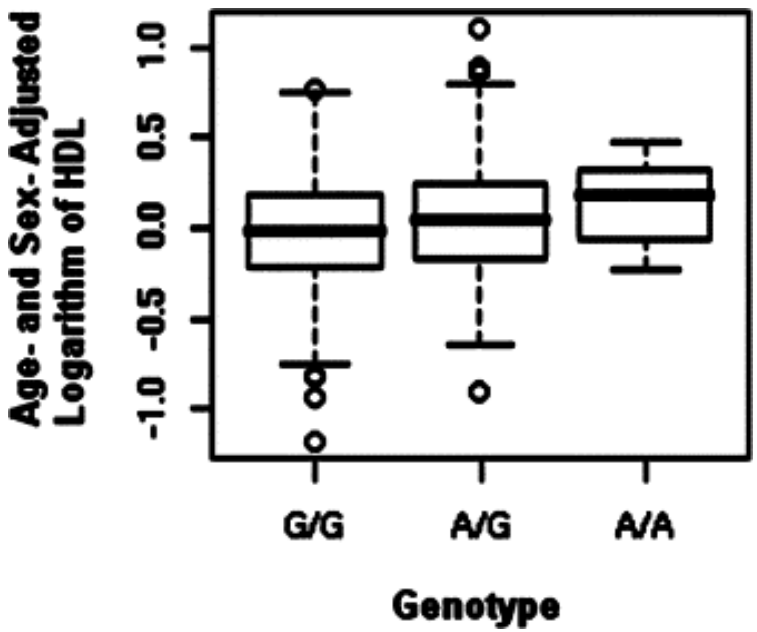
There are 449458 files on the ftp site
There are 481T of data on the ftp site
There are 26 populations
There are 2854 samples
There are 79072 gigabases of low coverage sequence
28753 x coverage in low coverage
There are 35607 gigabases of exome sequence

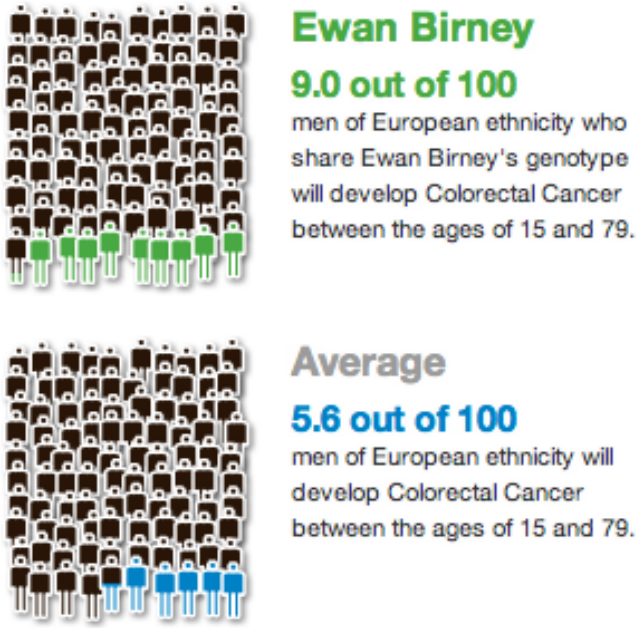
There was 235GB of sequence in the ENA before the 1000

Data Availability

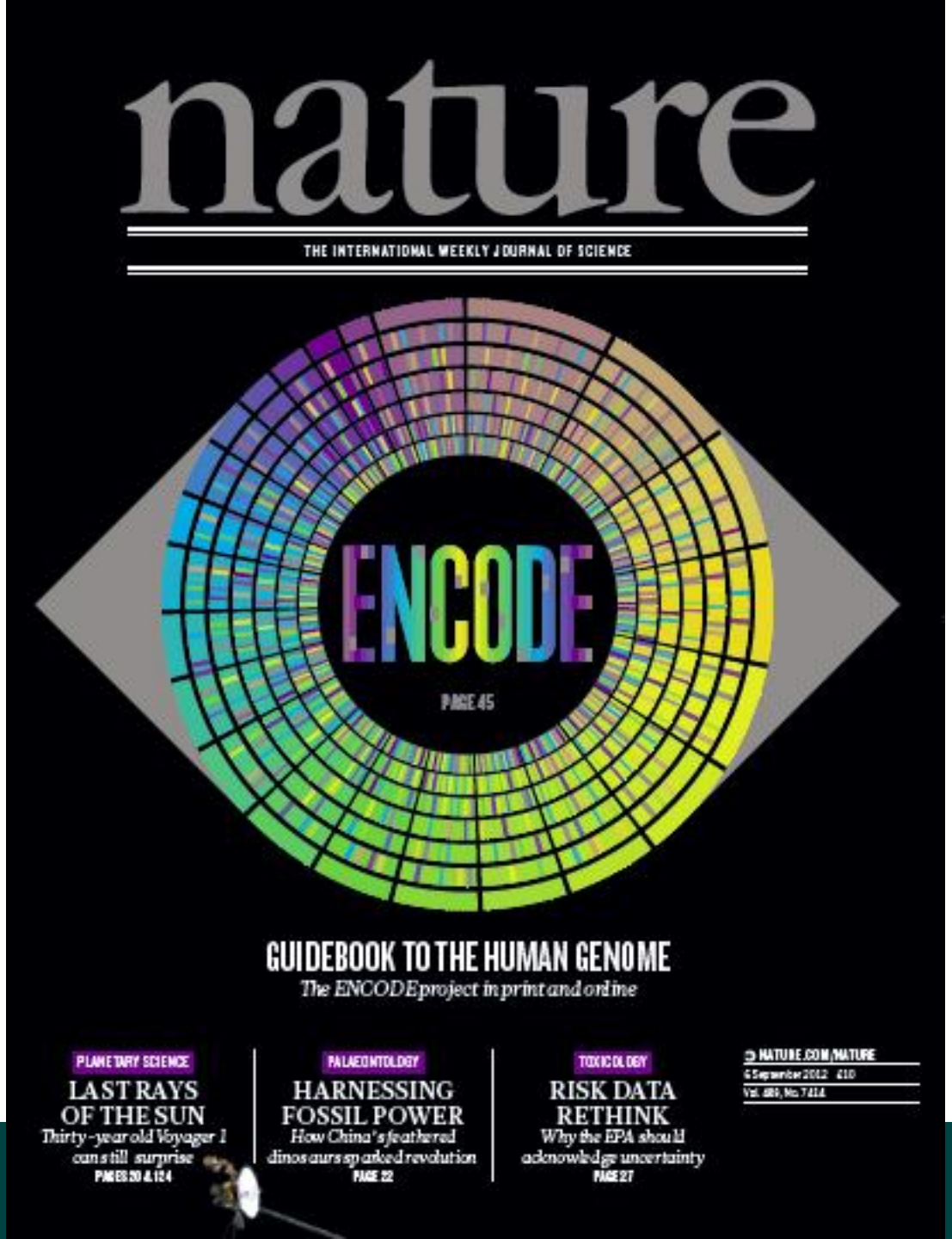
- Mirrored FTP Site (one in UK, one in US)
- HTTP mounted
- Accessible via Aspera
- FTP site is Mirrored to AWS (This is a selective mirror as disk is donated by Amazon)
- Ensembl Browser

… and associate this with traits or disease

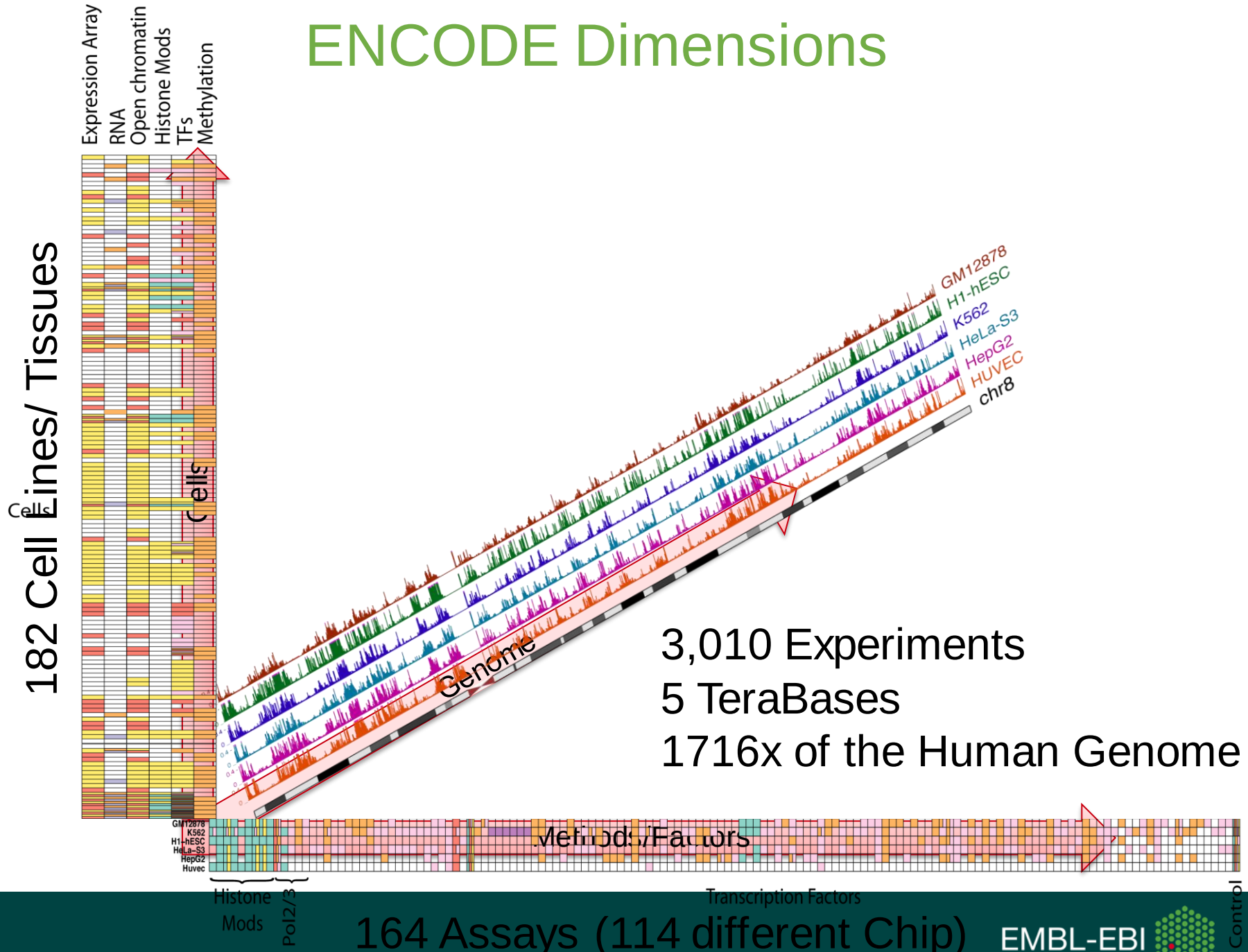




(you can infer the majority of the genome by knowing a base
About 1 every 5,000 to 10,000 bases – the experiments to
Look at this density is far cheaper than sequencing)


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
ENCODE
PAGE 45
GUIDEBOOK TO THE HUMAN GENOME
The ENCODE project in print and online
PLANETARY SCIENCE
LAST RAYS OF THE SUN
Thirty-year old Voyager 1 can still surprise
PAGES 30 & 124
PALAEONTOLOGY
HARNESSING FOSSIL POWER
How China's feathered dinosaurs sparked revolution
PAGE 22
TOXICOLOGY
RISK DATA RETHINK
Why the EPA should acknowledge uncertainty
PAGE 27
NATURE.COM/NATURE
6 September 2012 £10
Vol. 489, No. 7414

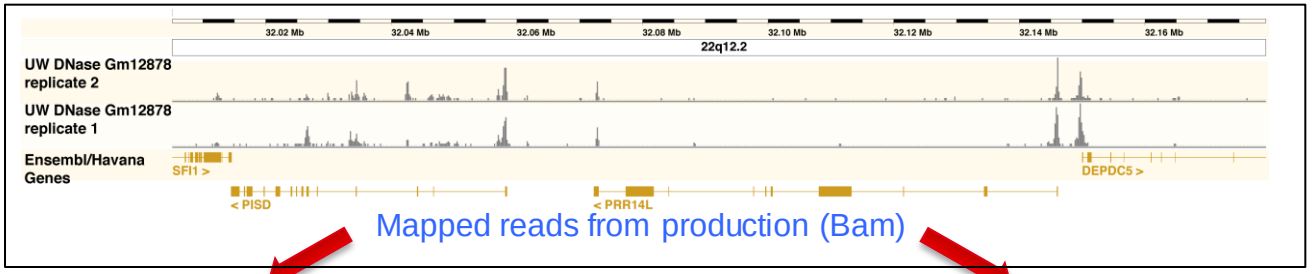
ENCODE Dimensions



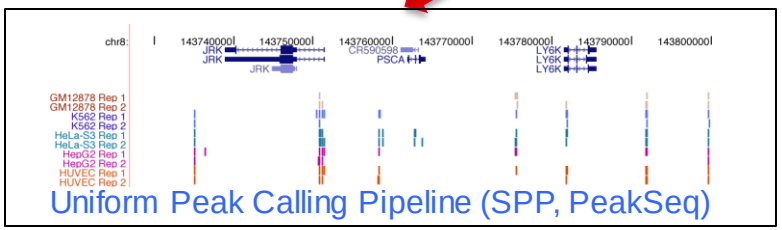
EMBL-EBI

ENCODE Uniform Analysis Pipeline

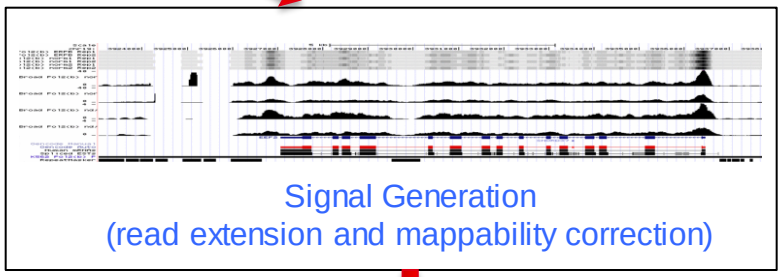
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



Mapped reads from production (Bam)

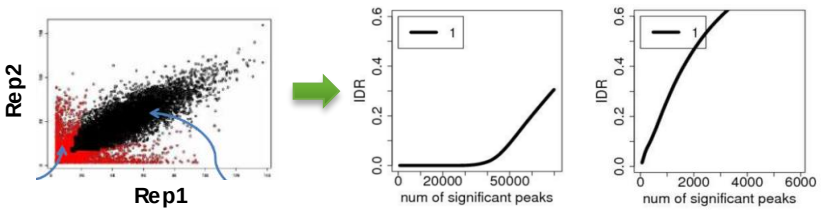


Uniform Peak Calling Pipeline (SPP, PeakSeq)



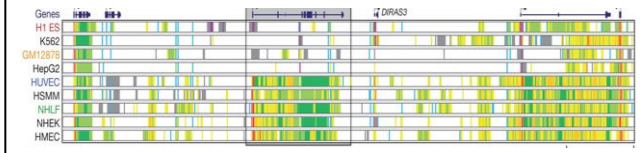
Signal Generation
(read extension and mappability correction)

Good reproducibility Poor reproducibility

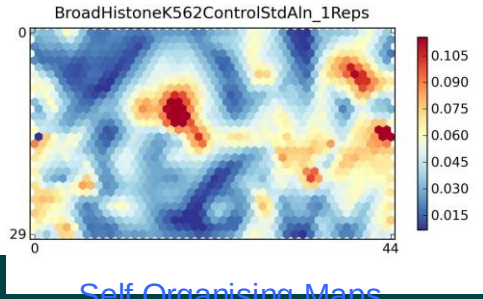


IDR Processing, QC and Blacklist Filtering

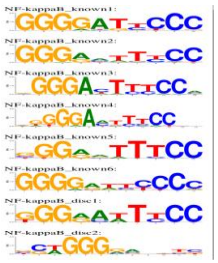
Segmentation



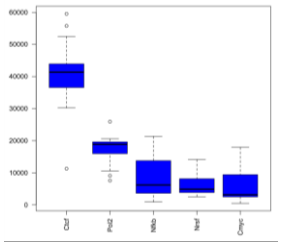
ChromHMM/Segway



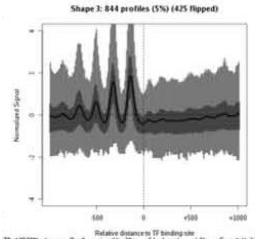
Self Organising Maps



Motif Discovery



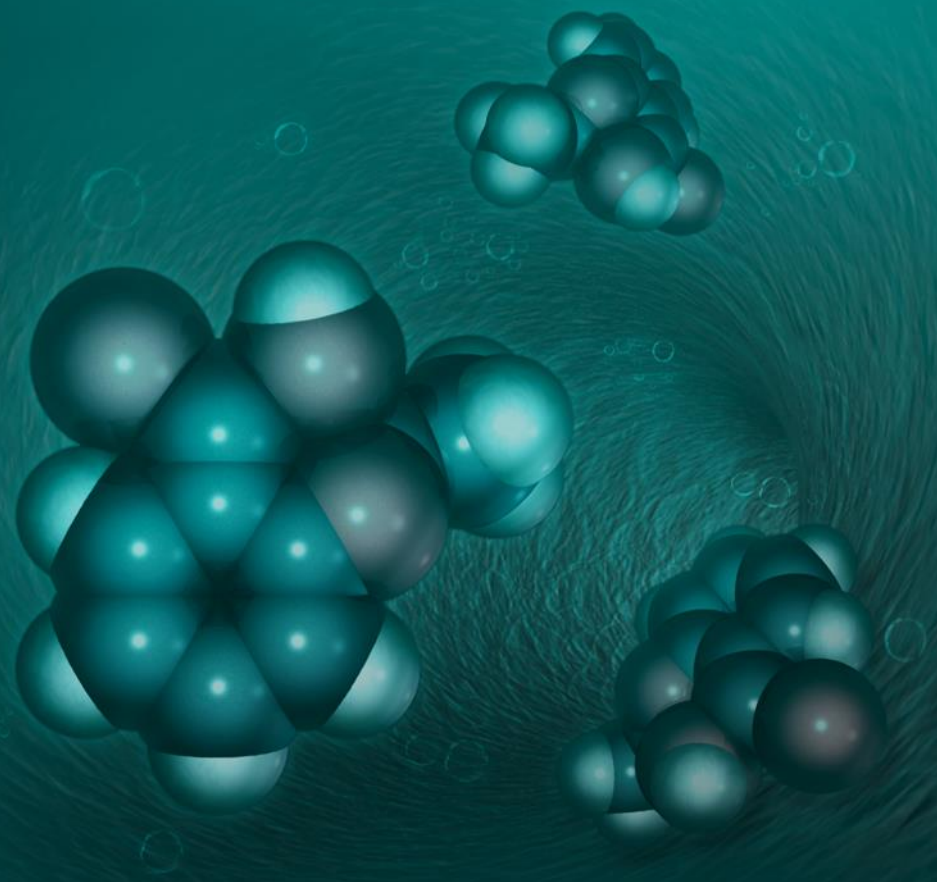
Stats, GSC enrichments, etc.



Signal Aggregation over peaks

EMBL-EBI

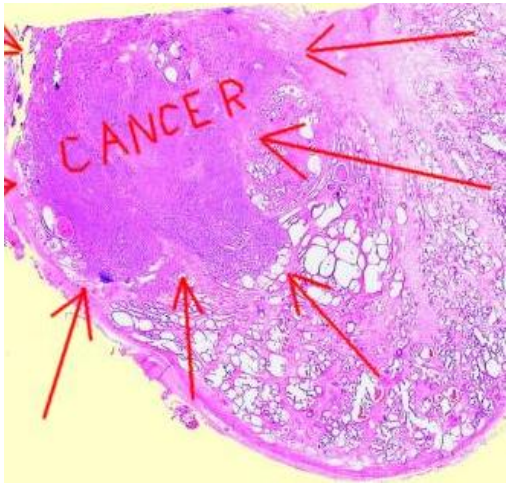
Impact on Medicine



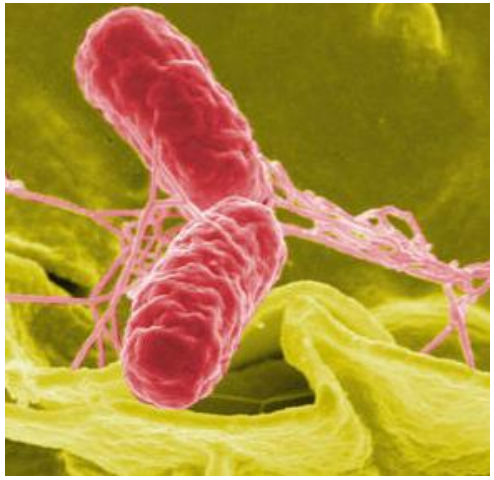
EMBL-EBI

3 big areas of impact for medicine

Germ line
Risk to disease



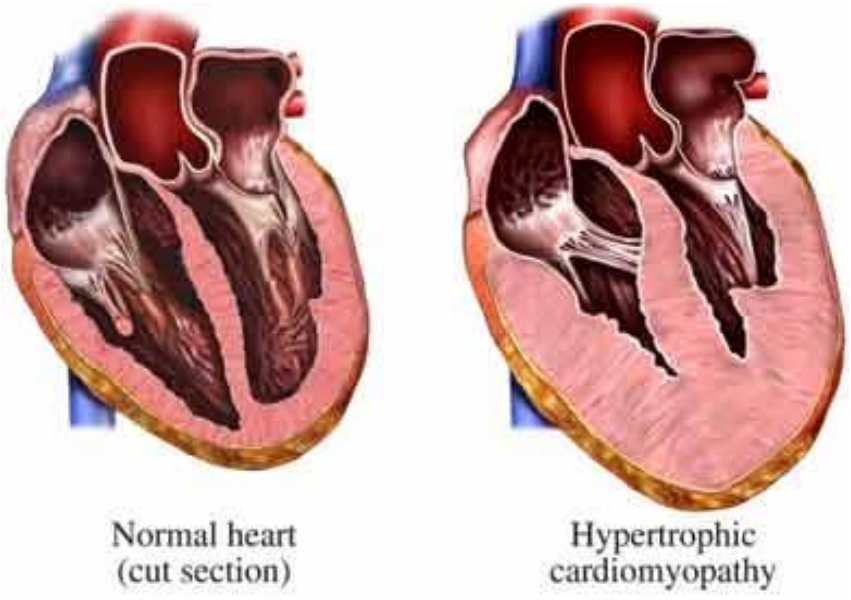
“Precision” cancer medicine



Pathogens + Hospital acquired infections

Germ Line impact

- Everyone has differential risk of disease
- But the shift in risk is small
- Perhaps 1 to 2% have a striking change in risk to a serious disease (>10 fold) which is “actionable”

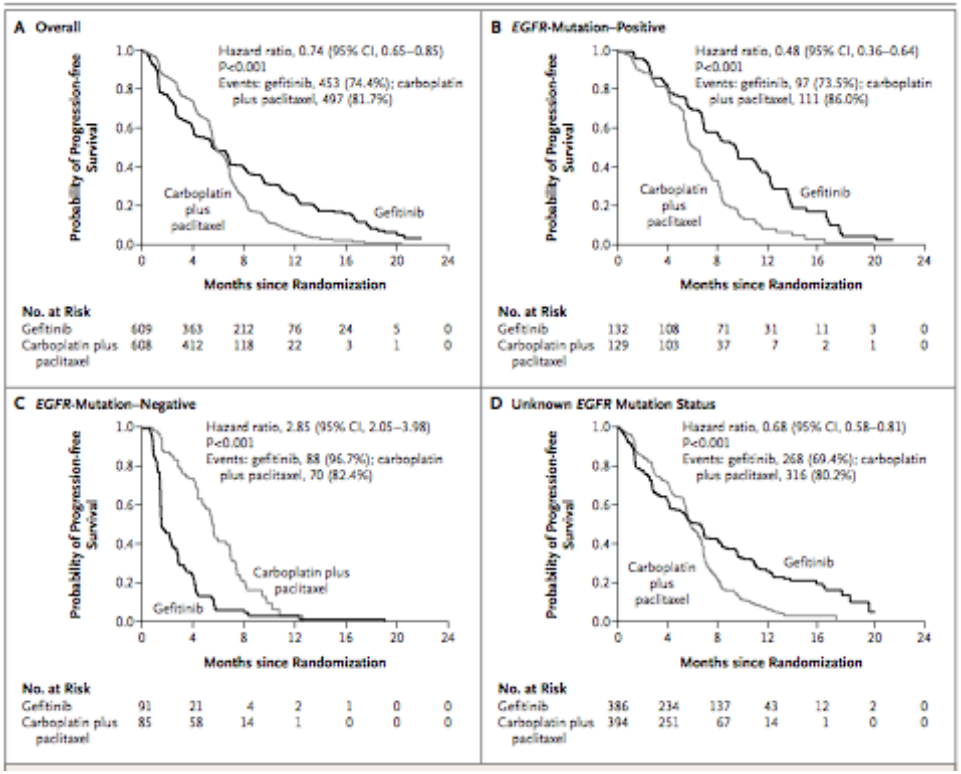


- This goes up to 3-4% if you count some less clinically worrying diseases

1:500 people have HCM
1:500 people have FH

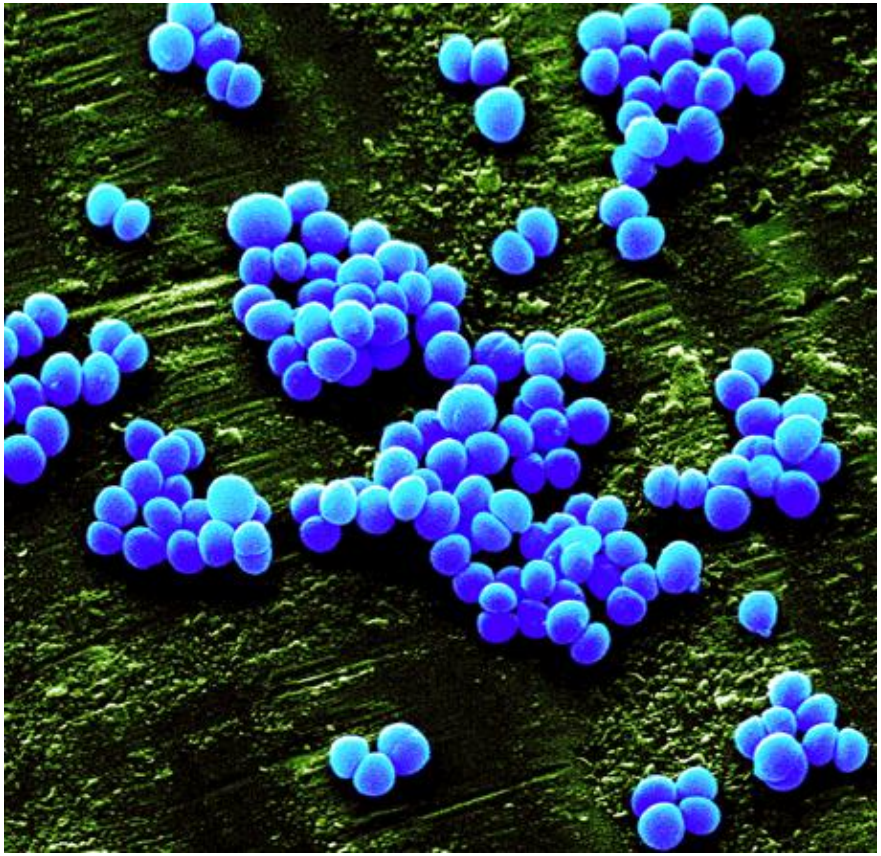
Precision cancer diagnosis

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

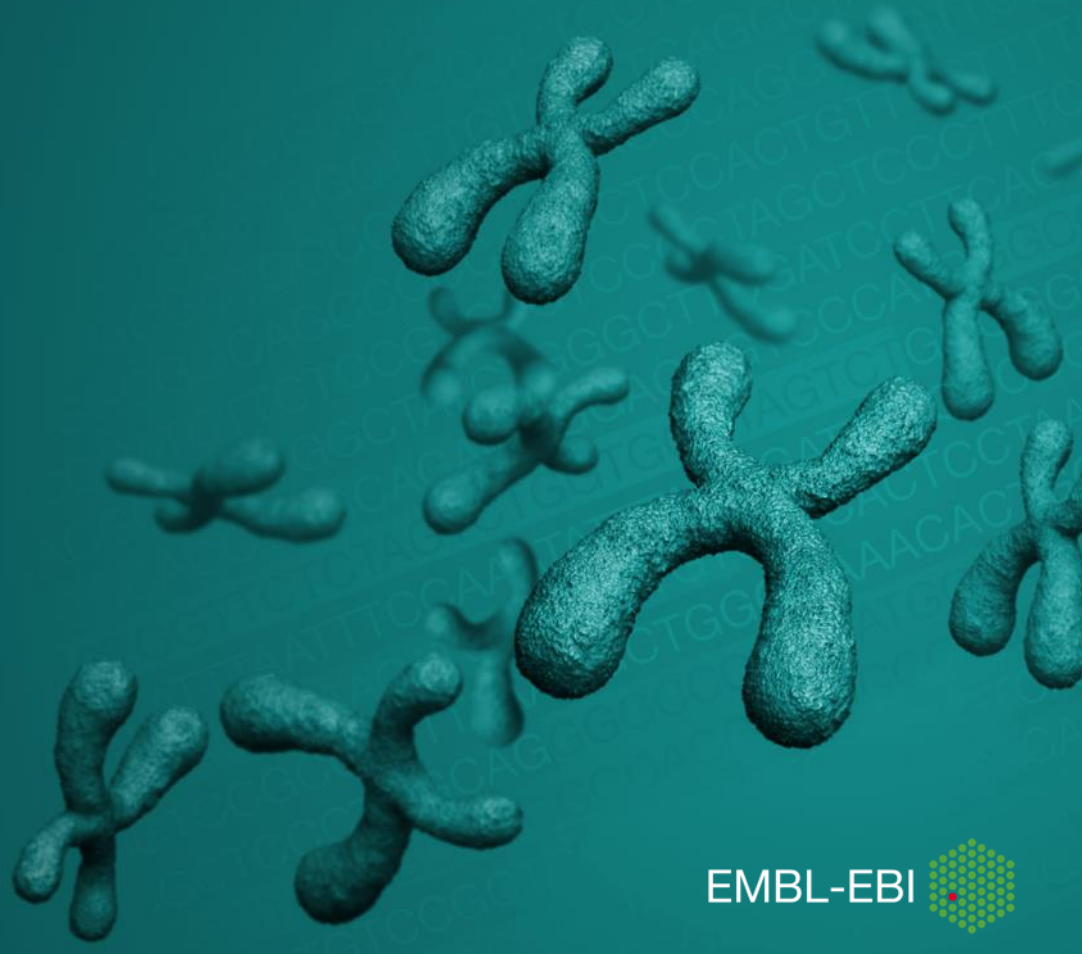


Pathogens

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

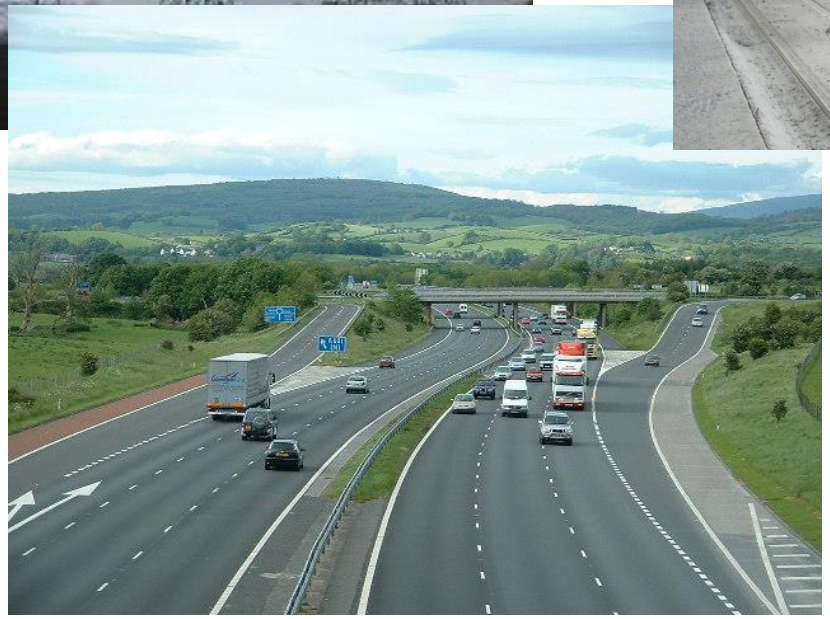


Why we need an infrastructure…



EMBL-EBI

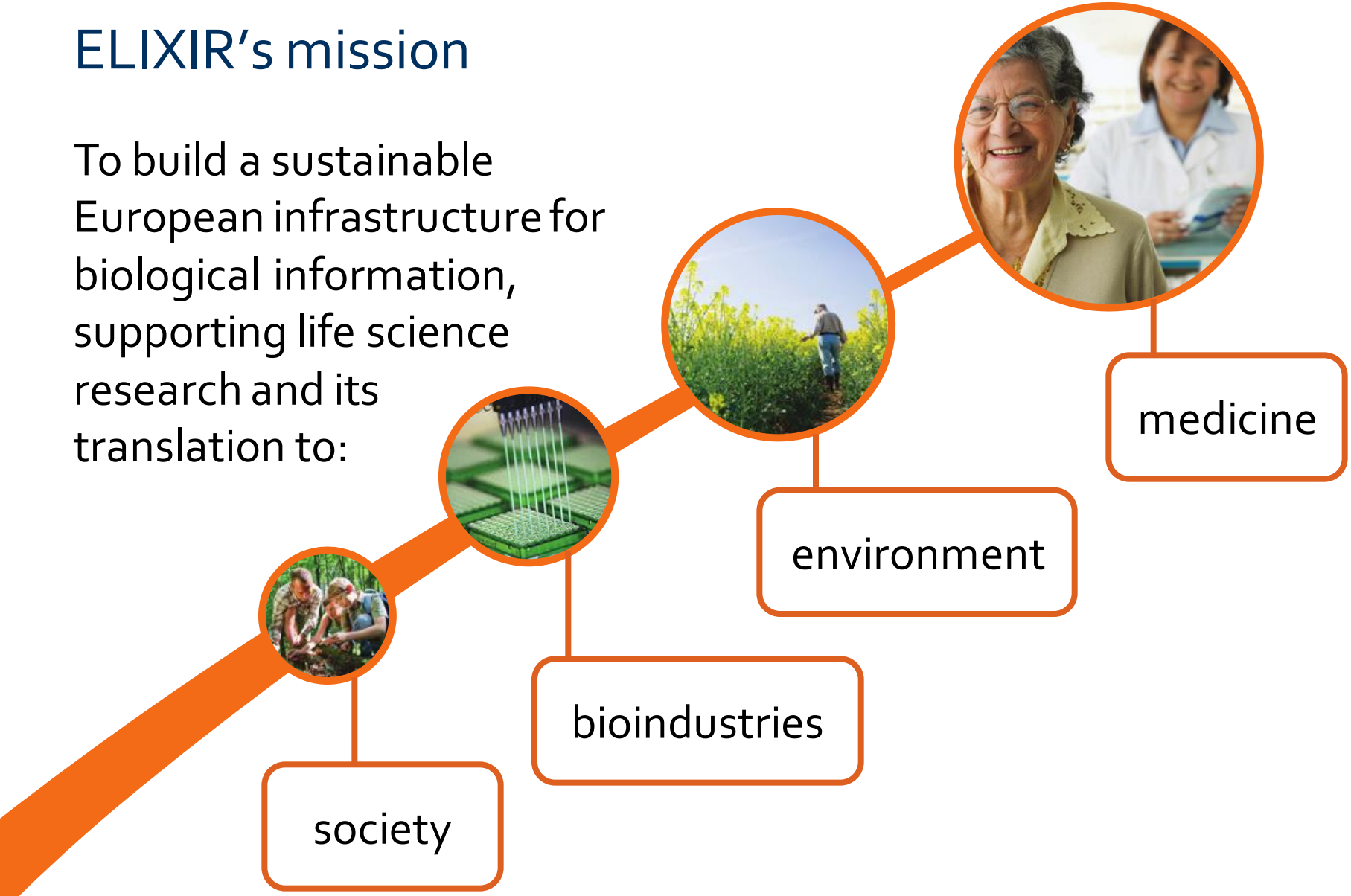
Infrastructures are critical…



But we only notice them when they go wrong

ELIXIR's mission

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

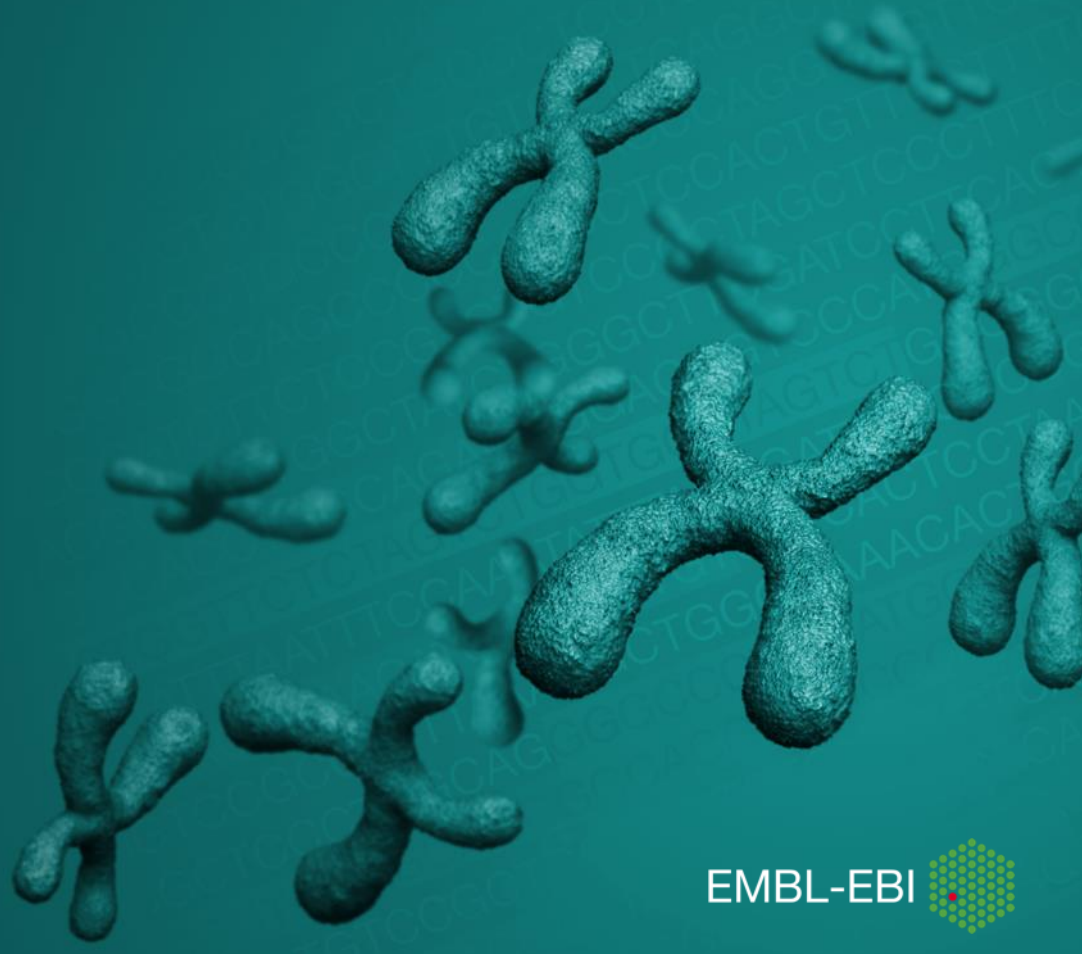


EMBL-EBI

Things to discuss with CERN

- Look ahead about data storage
  - The cost of different solutions, future technologies
  - I/O management, how one handles hot vs cold data
- Data specific compression
  - How best to handle making bespoke data compression routines
- Network services
  - Both to “core” groups and “every day” groups
- Cloud and virtualisation
  - Both for internal use and as a data distribution approach

And… just for fun…

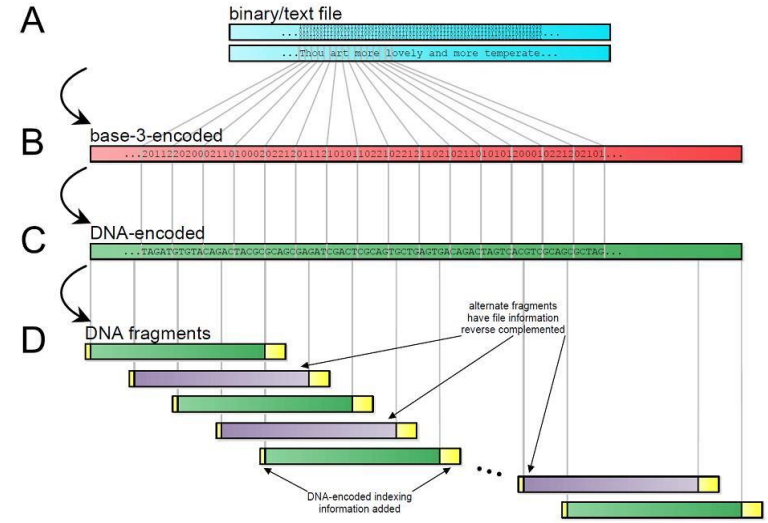


EMBL-EBI

Over a beer…

Ha! At some point all the data we Store is going to be DNA…

Of course, the cost effective way To store this would be as DNA…



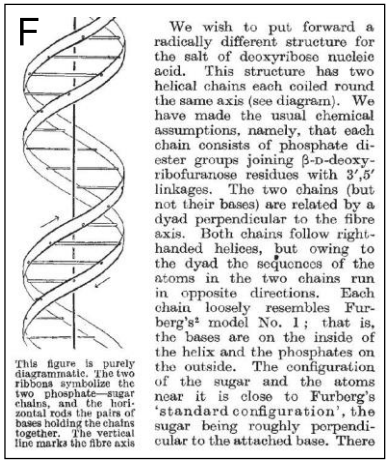
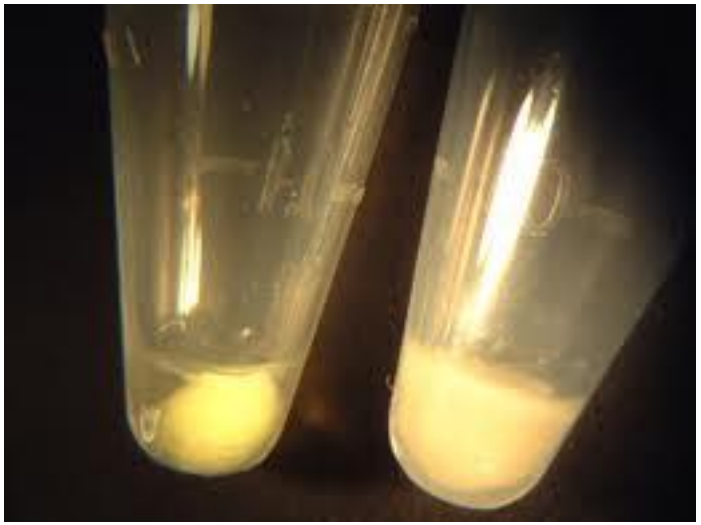


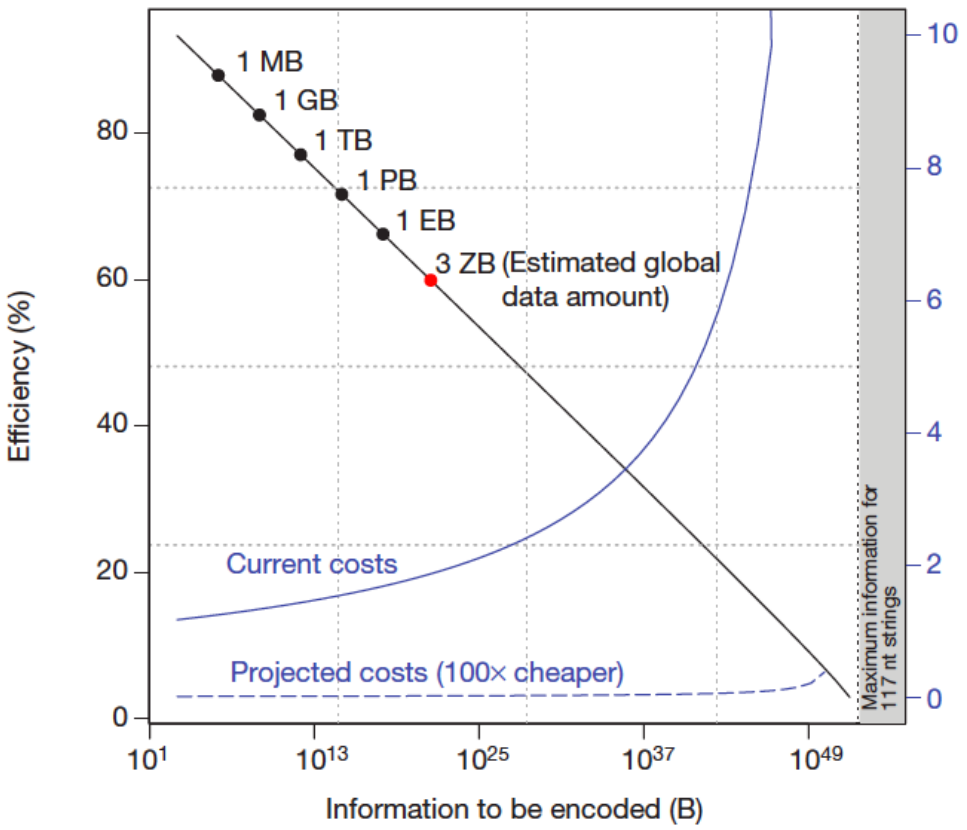
Figure 2 | Digital information encoded in DNA. Digital information (**A**, in blue), here binary digits holding the ASCII codes for part of Shakespeare's sonnet 18, was converted to base-3 (**B**, red) using a Huffman code. This in turn was converted *in silico* to our DNA code (**C**, green), with no homopolymers, which formed the basis for a large number of overlapping DNA segments each containing 100 bases of encoded information (**D**, green or, with alternate segments reverse complemented for added data security, violet) and with orientation and indexing DNA codes added (yellow, as described in the text). These strings were synthesised, sequenced and decoded. **E**, A digital photograph of the EMBL-European Bioinformatics Institute (JPEG 2000 format) and **F**, an extract of the Watson and Crick (1953) paper[10] (PDF format) that were among the files encoded in DNA and successfully recovered in this study.

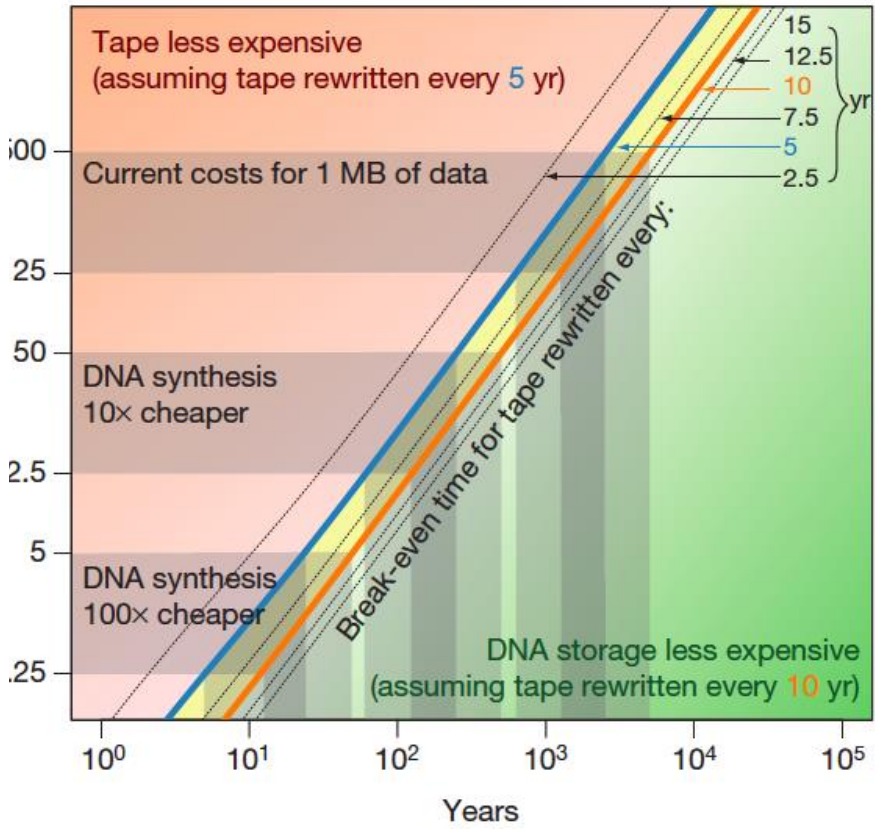
1 g == 1 PB (with redundancy)



EMBL-EBI

Scaleable? Cost effective?





Questions?

(you can follow me on twitter @ewanbirney)
I blog and update this on Google Plus publically

ATOMIC COORDINATES FOR SUBTILISIN BPN (OR NOVO)

Download PDB file | View in 3D | Similar structures | Quaternary structure

The structure was published by Alden, R.A., Birktoft, J.J., Kraut, J., Robertus, J.D., and Wright, C.S., in 1971 in a paper entitled "Atomic coordinates for subtilisin BPN' (or Novo)." (abstract).

This crystal structure was determined using X-ray diffraction at a resolution of 2.5 Å and deposited in 1972.

The experimental data on which the structure is based was not deposited.

The PDB entry contains the structure of SUBTILISIN BPN'. This molecule has the UniProt identifier P00782 (SUBT_BACAM). The sample contained 275 residues which is < 90% of the natural sequence. Out of 275 residues 275 were observed and are deposited in the PDB.

The molecule is most likely monomeric.

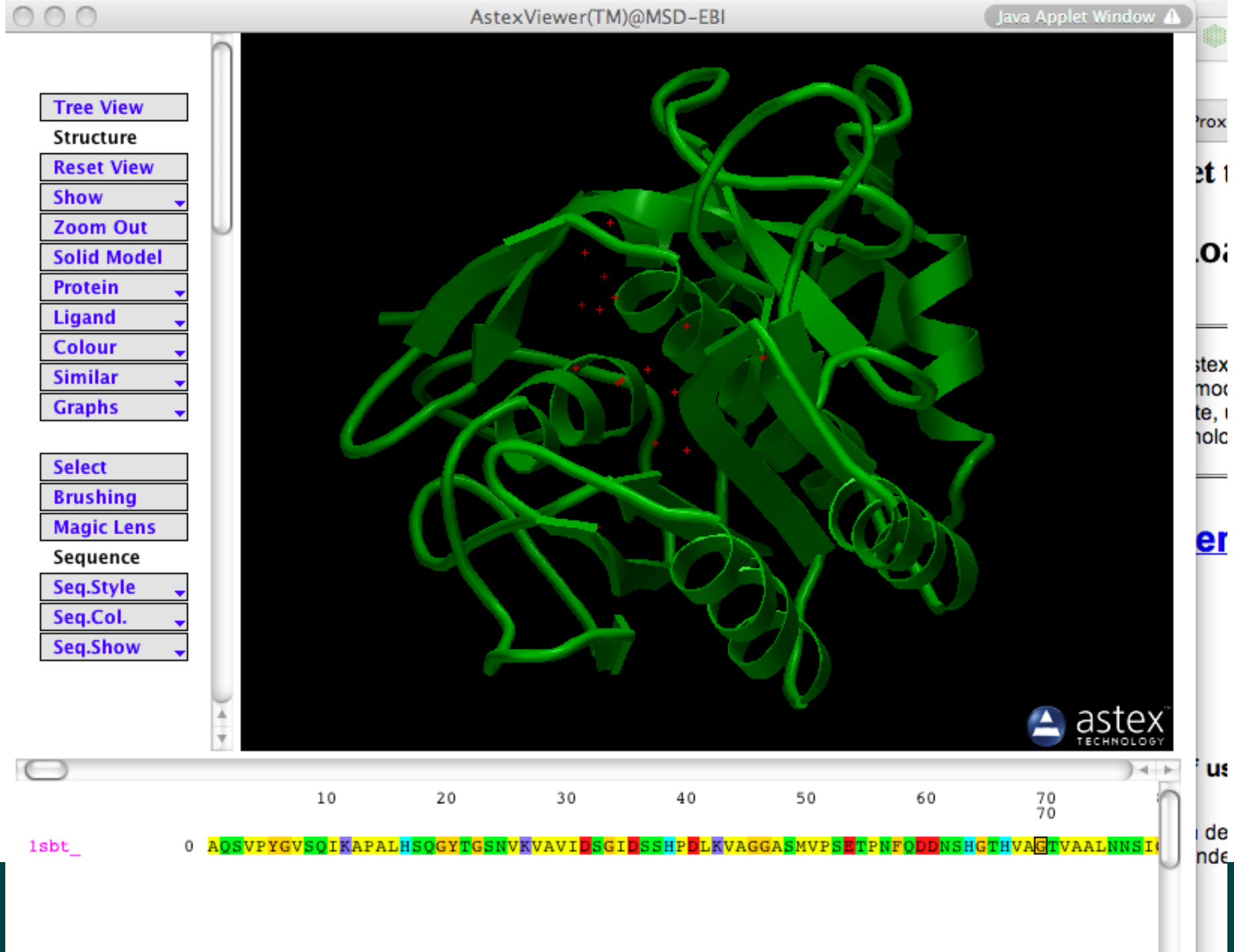
The following tables show cross-reference information to other databases (to obtain a list of all PDB entries sharing the same property or classification, click on the magnifying glass icon):

Chain	Name	UniProt	Name of source organism	% of UniProt sequence present in the sample	Residues in the sample molecules	% of residues observed
A	SUBTILISIN BPN'	P00782 (108-382) (SUBT_BACAM)	Bacillus amyloliquefaciens	< 90%	275	100%

This entry contains 1 unique UniProt protein:

UniProt accession	Name	Organism	PDB
P00782 (108 - 382)	SUBTILISIN BPN'	Bacillus amyloliquefaciens	Related PDB sequences UniProt coverage

EMBL-EBI


AstexViewer(TM)@MSD-EBI
Java Applet Window
Tree View
Structure
Reset View
Show
Zoom Out
Solid Model
Protein
Ligand
Colour
Similar
Graphs
Select
Brushing
Magic Lens
Sequence
Seq.Style
Seq.Col.
Seq.Show
astex
TECHNOLOGY
10
20
30
40
50
60
70
70
1sbt_
0
AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLKVAGGASMVPSETPNFQDDNSHGTHVAGTVAALNNSI