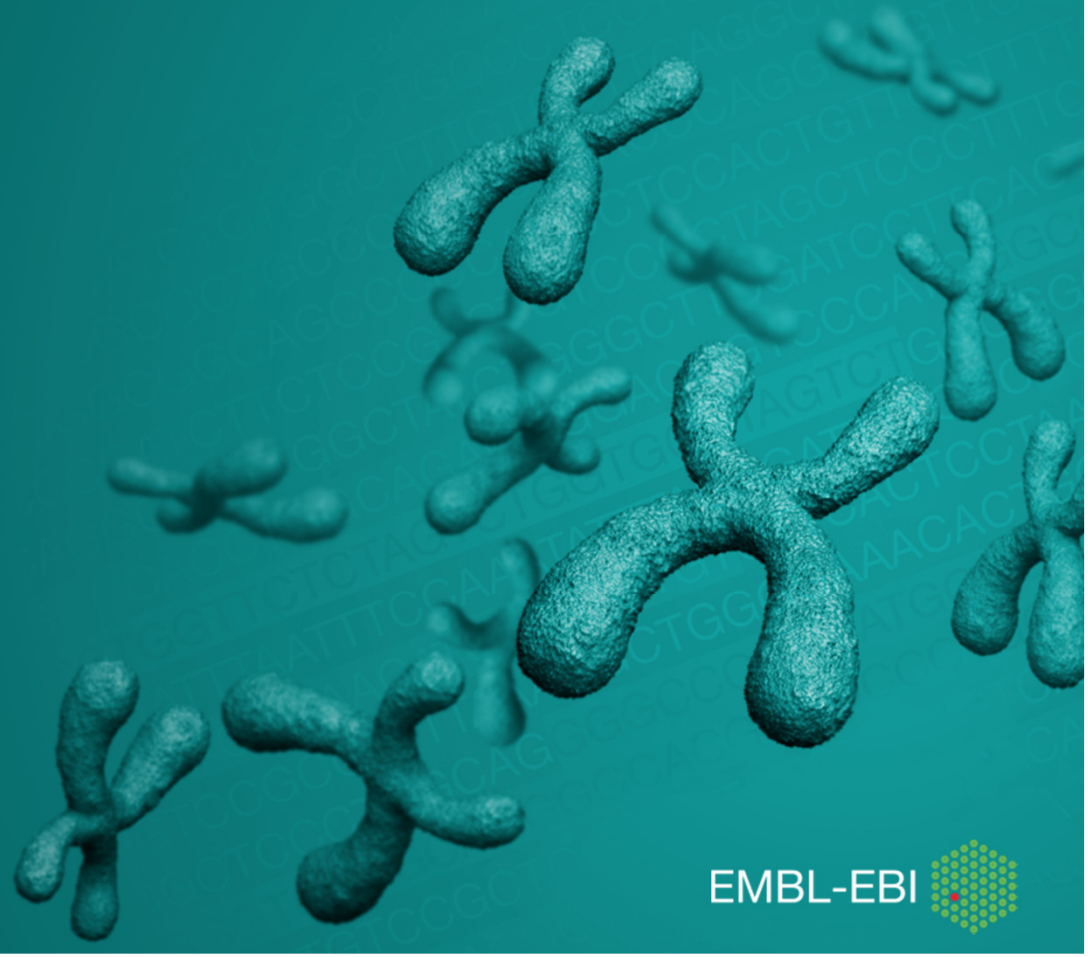
Genome Annotation

Ewan Birney (tweetable)



EMBL-EBI

Outline of the talk

- Who am I?
- A quick crash course in genomics for geeks
- Genomics 2000-2012
  - HapMap, 1000 Genomes, GWAS, ENCODE
- Route into medicine

- (Some more whimsical uses of DNA…)

Who am I?

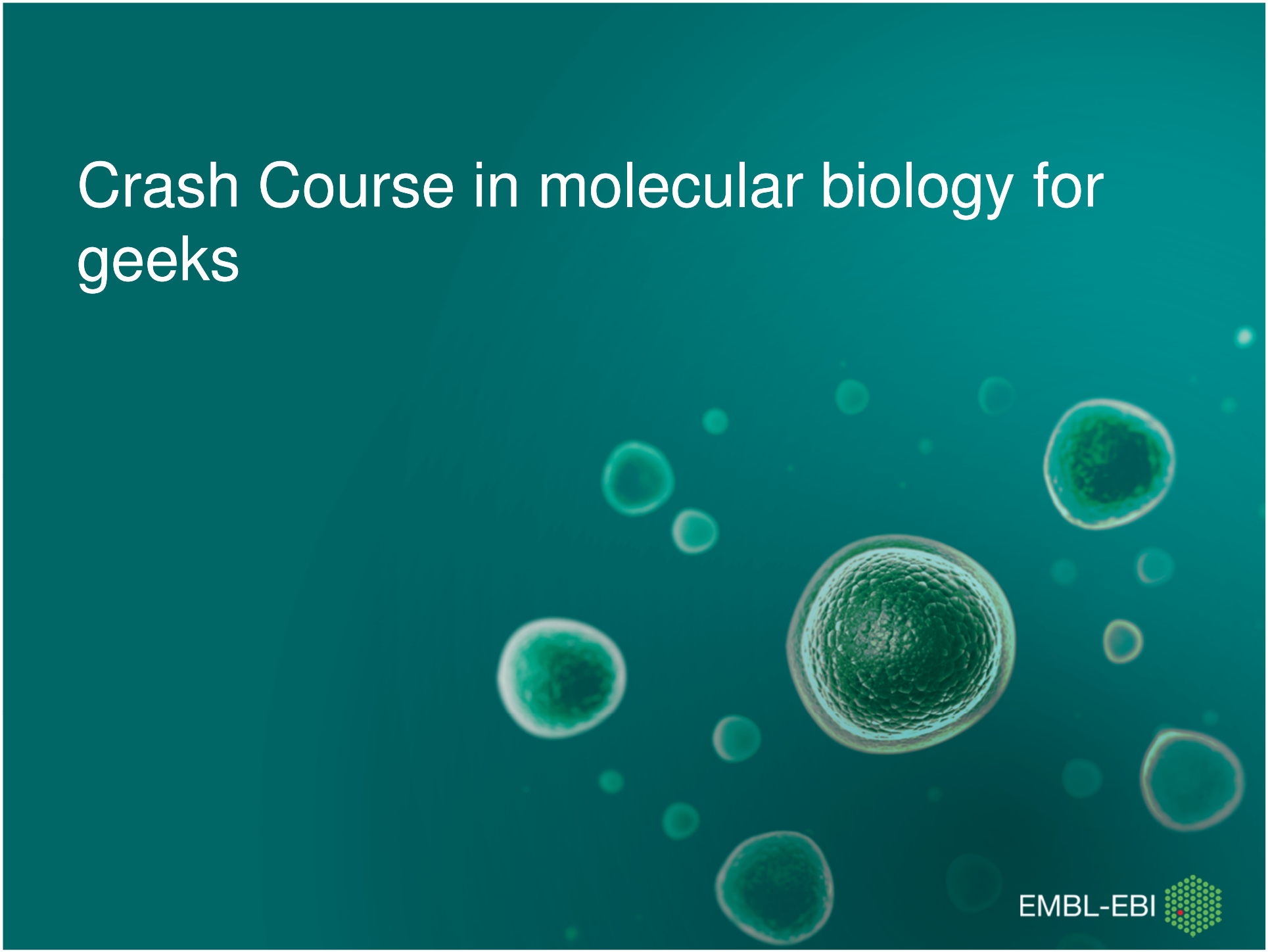
- Associate Director at European Bioinformatics Institute (EBI)
- Involved in genomics since I was 19 (almost 20 years!)
- Trained as a biochemist – most people think I am CS
- Analysed – sometimes lead –

  human/mouse/rat/platypus etc genomes

EBI is in Hinxton, South Cambridgeshire

EBI is part of EMBL, like CERN for molecular biology

Crash Course in molecular biology for geeks



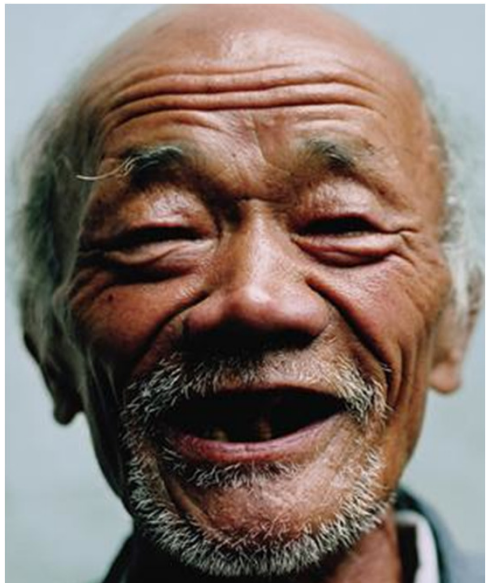
EMBL-EBI

Molecules of life

- DNA ⇔ Hard Disk
- RNA ⇔ Computer RAM
- Proteins ⇔ Computer Chips, Robotic Arms etc
- Metabolites ⇔ Electricity, Optics

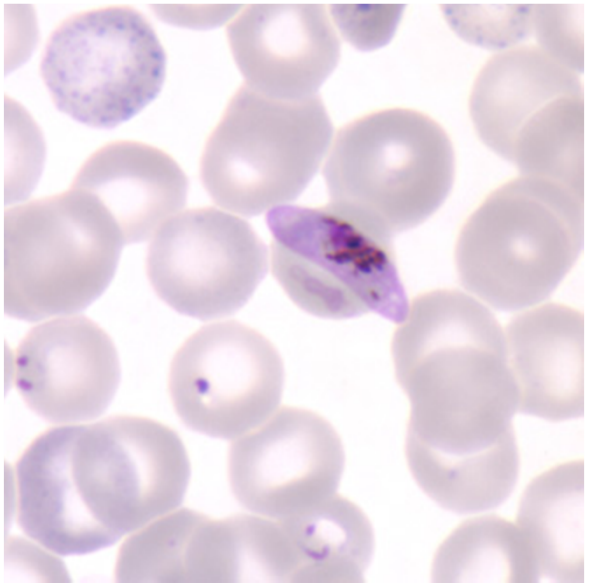
- Good theories of how these molecules fit together
  - How is the information in DNA moved to RNA
- No good theories of the precise details of these molecules
  - You simply have to *know* the Human Genome, Human RNA etc

Diversity of Life… unity of molecules



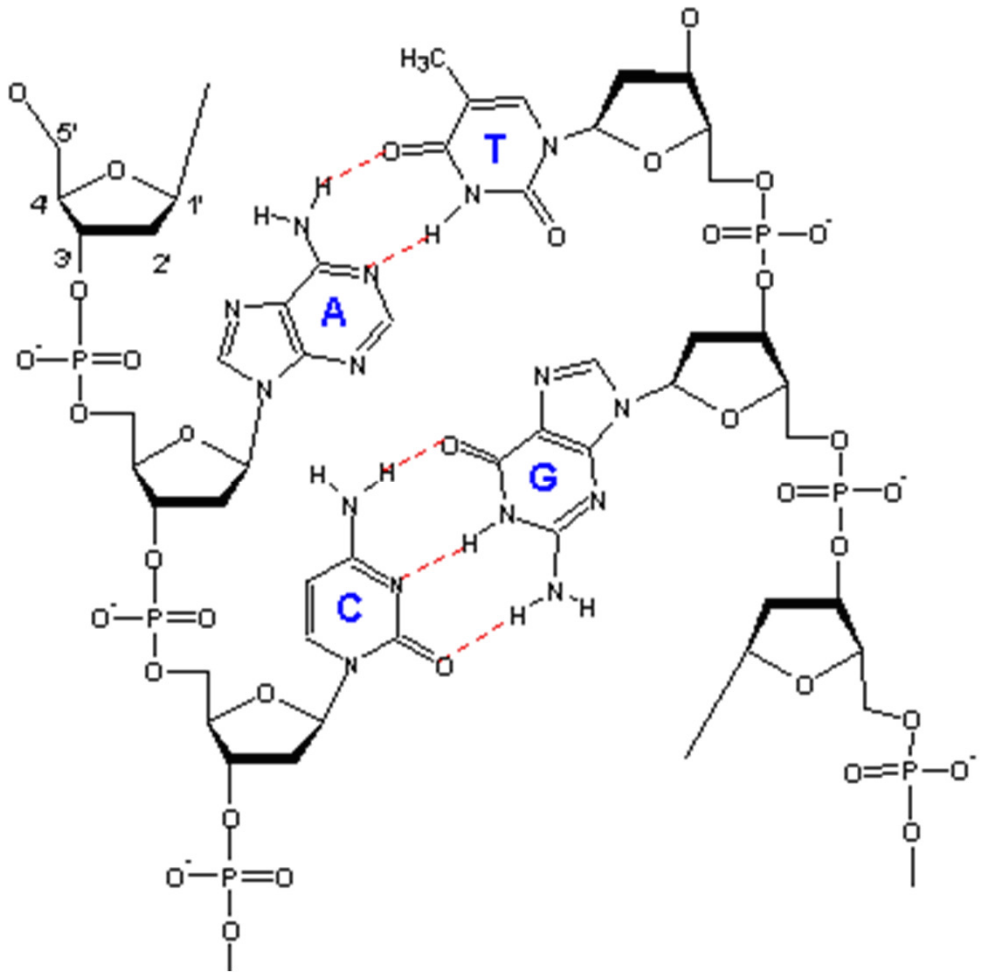
DNA
RNA
Proteins
Metabolites

DNA
RNA
Proteins
Metabolites



DNA
RNA
Proteins
Metabolites

DNA is a simple (ish) chemical



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

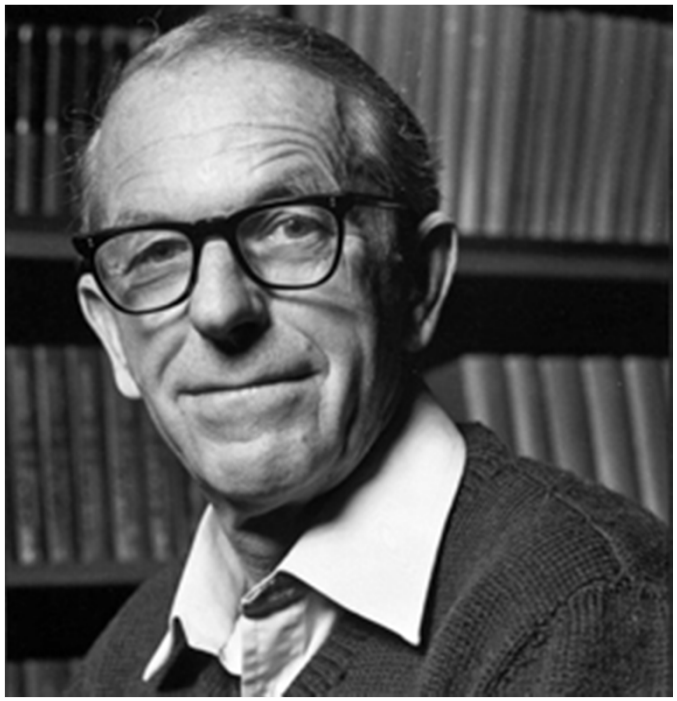
We can routinely determine small parts of DNA

1977-1990 – 500 bp, manual tracking

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

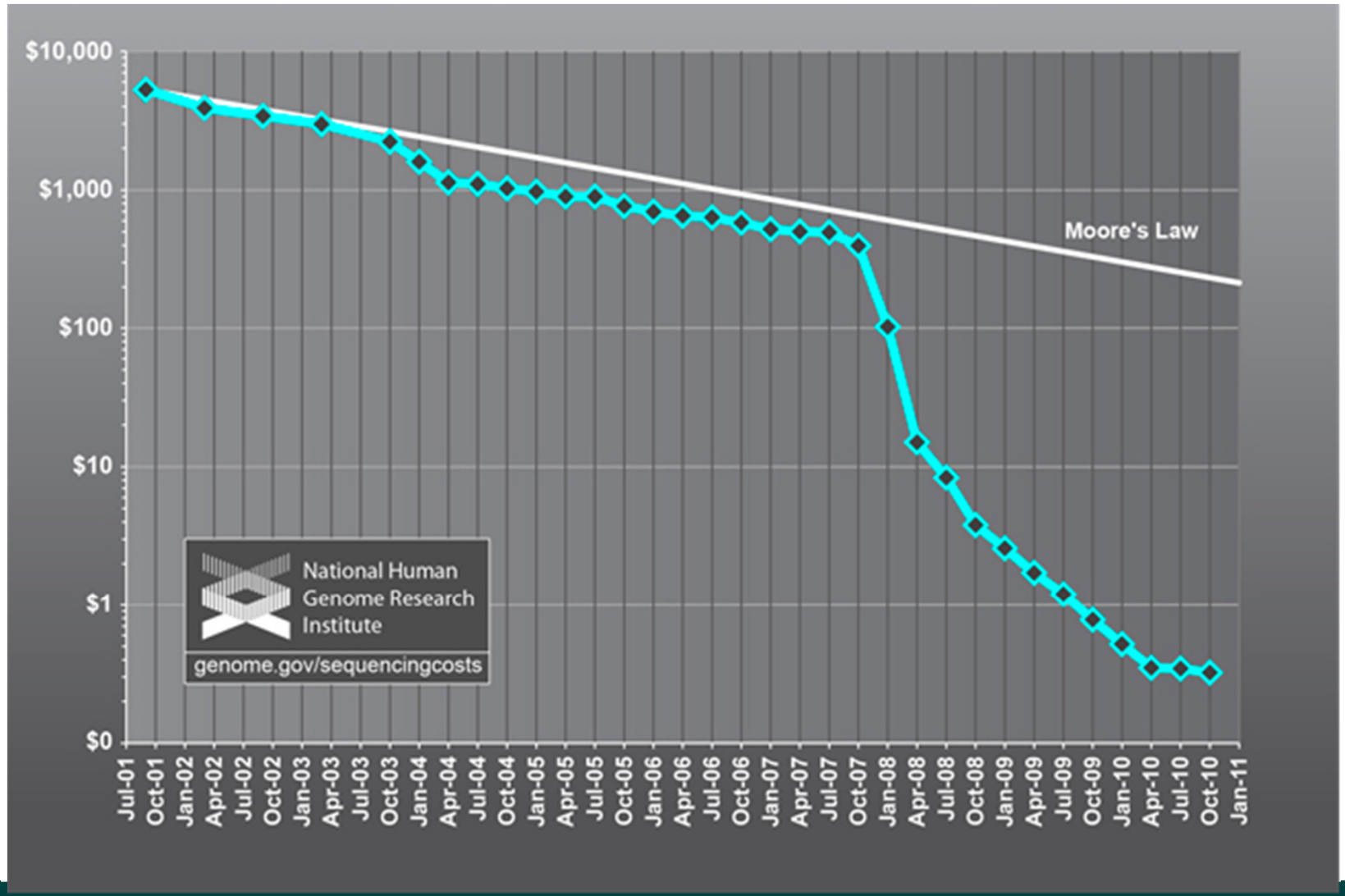
2005-2012 – 20-100bp, 2D systems, (“2nd Generation” or NGS)

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

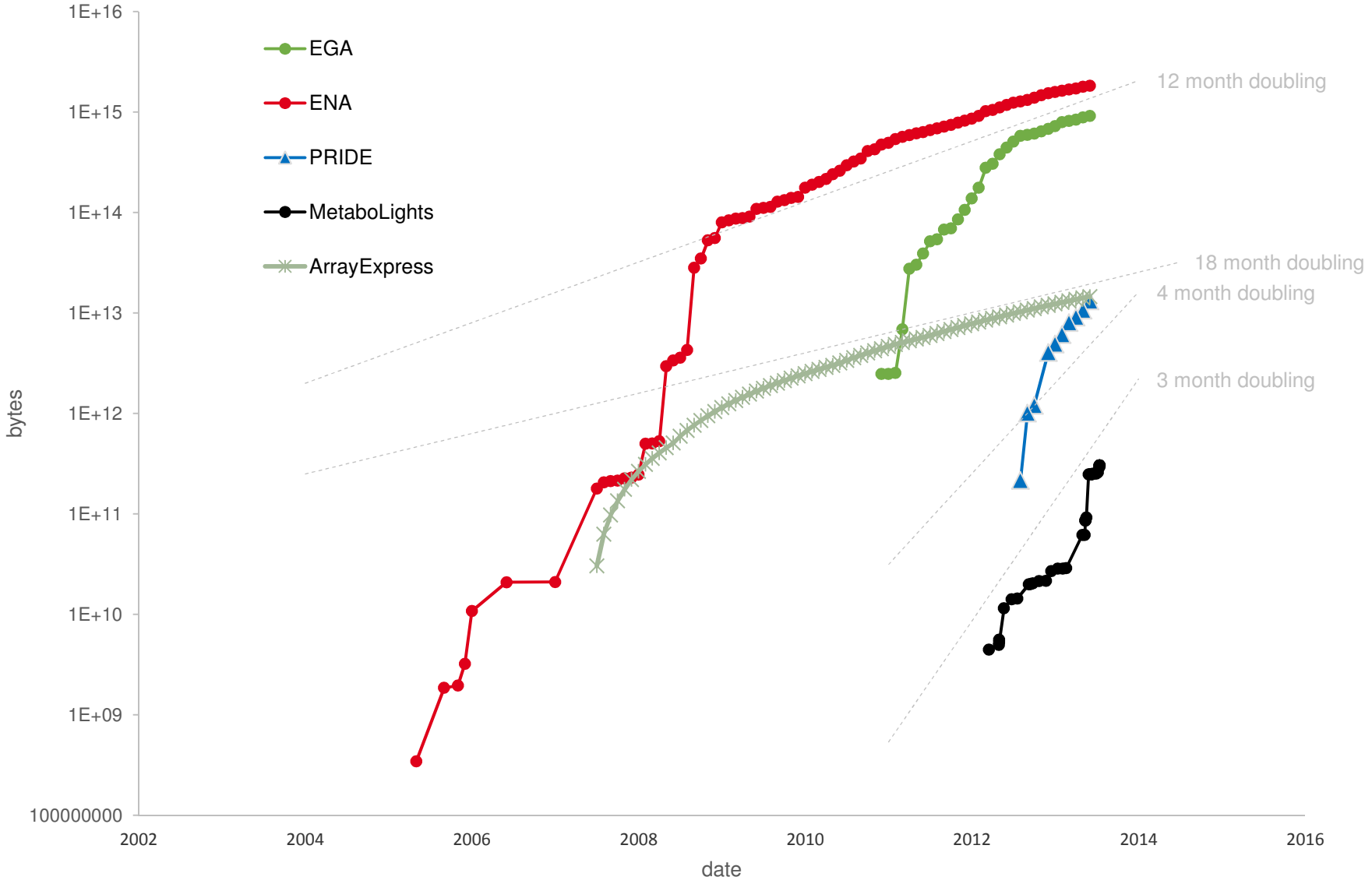


Fred Sanger, inventor of terminator DNA sequencing

Costs have come exponentially down



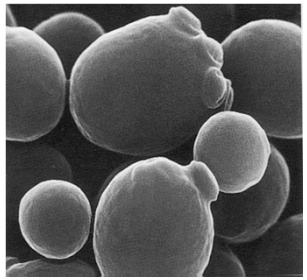
Data growth


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

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

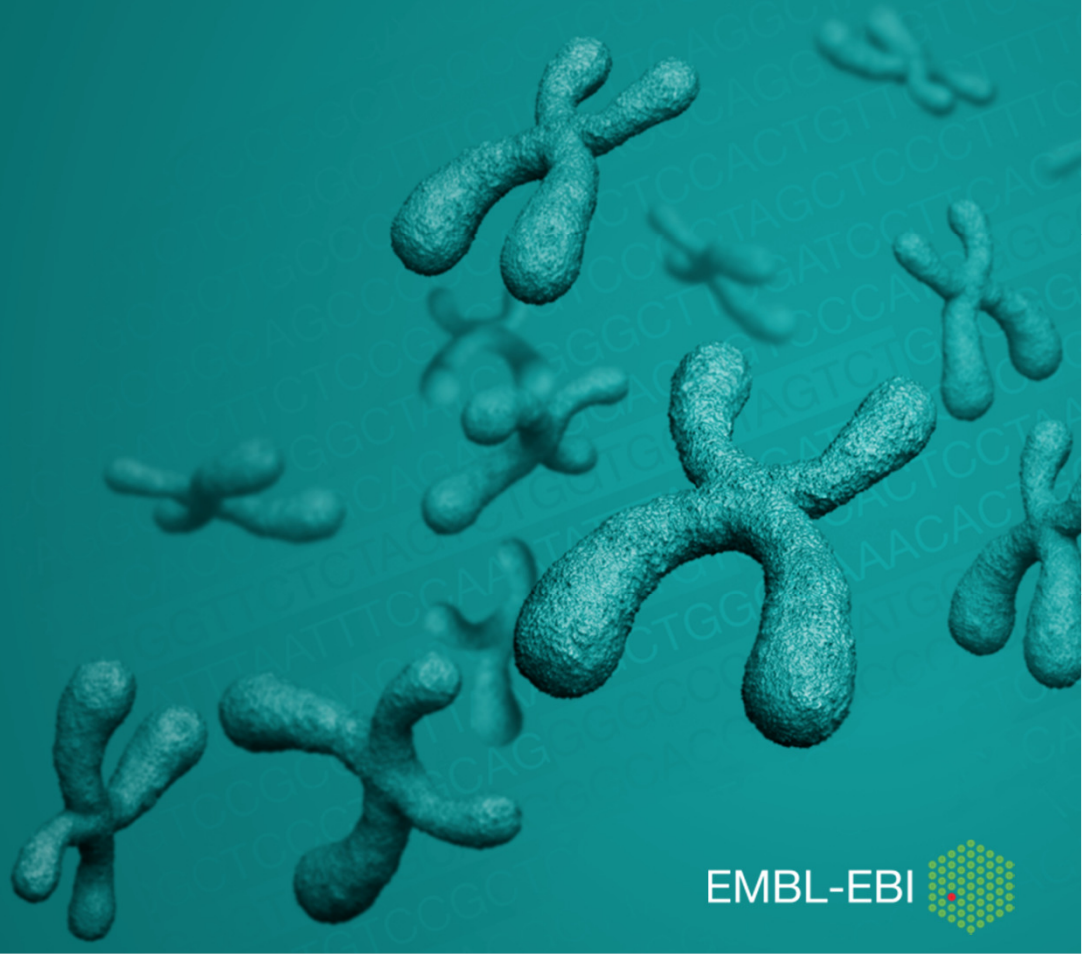
Medaka, 0.9e9

White Pine 20e9

Human Genome project

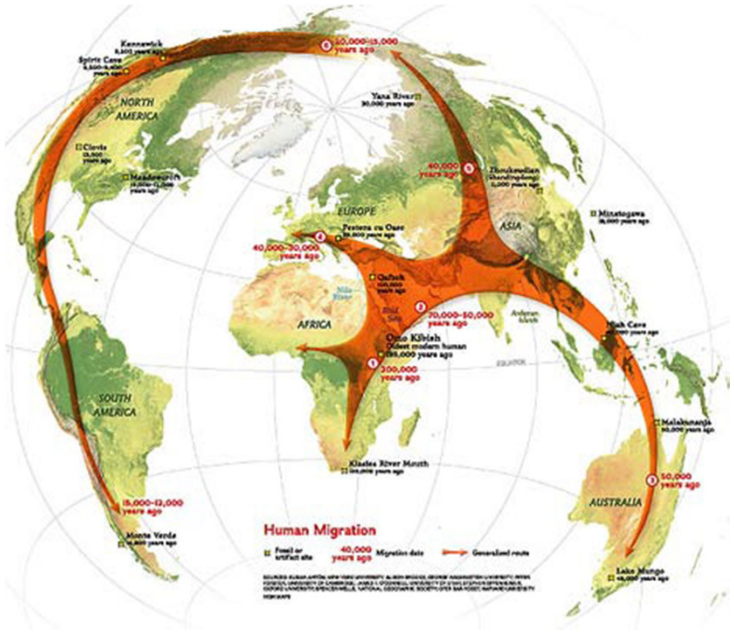
- 1989 – 2000 – sequencing the human genome
  - Just 1 “individual” – actually a mosaic of about 24 individuals but as if it was one
  - Old school technologies
  - A bit epic
- Now
  - Same data volume generated in ~3mins in a current large scale centre
  - It’s all about the *analysis*

What happened next?



EMBL-EBI

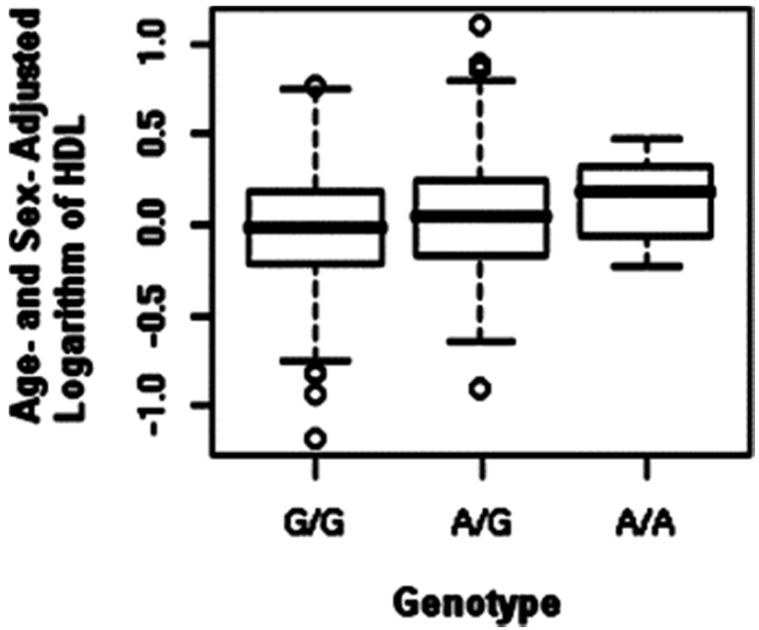
We looked into human variation

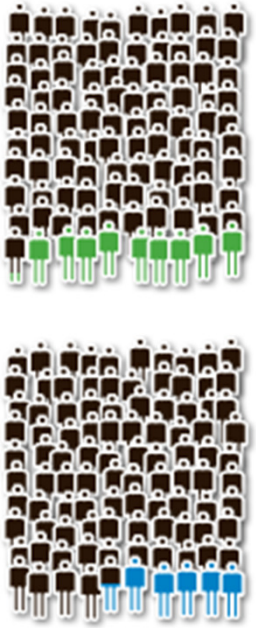


3 in 10,000 bases between any two individuals are different (a bit more between Africans)

The similarity of a European to an African (any population) is Only marginally smaller than European to European (2 or 3%).
Only a minute amount of DNA is unique to any population

… and associate this with traits or disease





Ewan Birney

9.0 out of 100

men of European ethnicity who share Ewan Birney's genotype will develop Colorectal Cancer between the ages of 15 and 79.

Average

5.6 out of 100

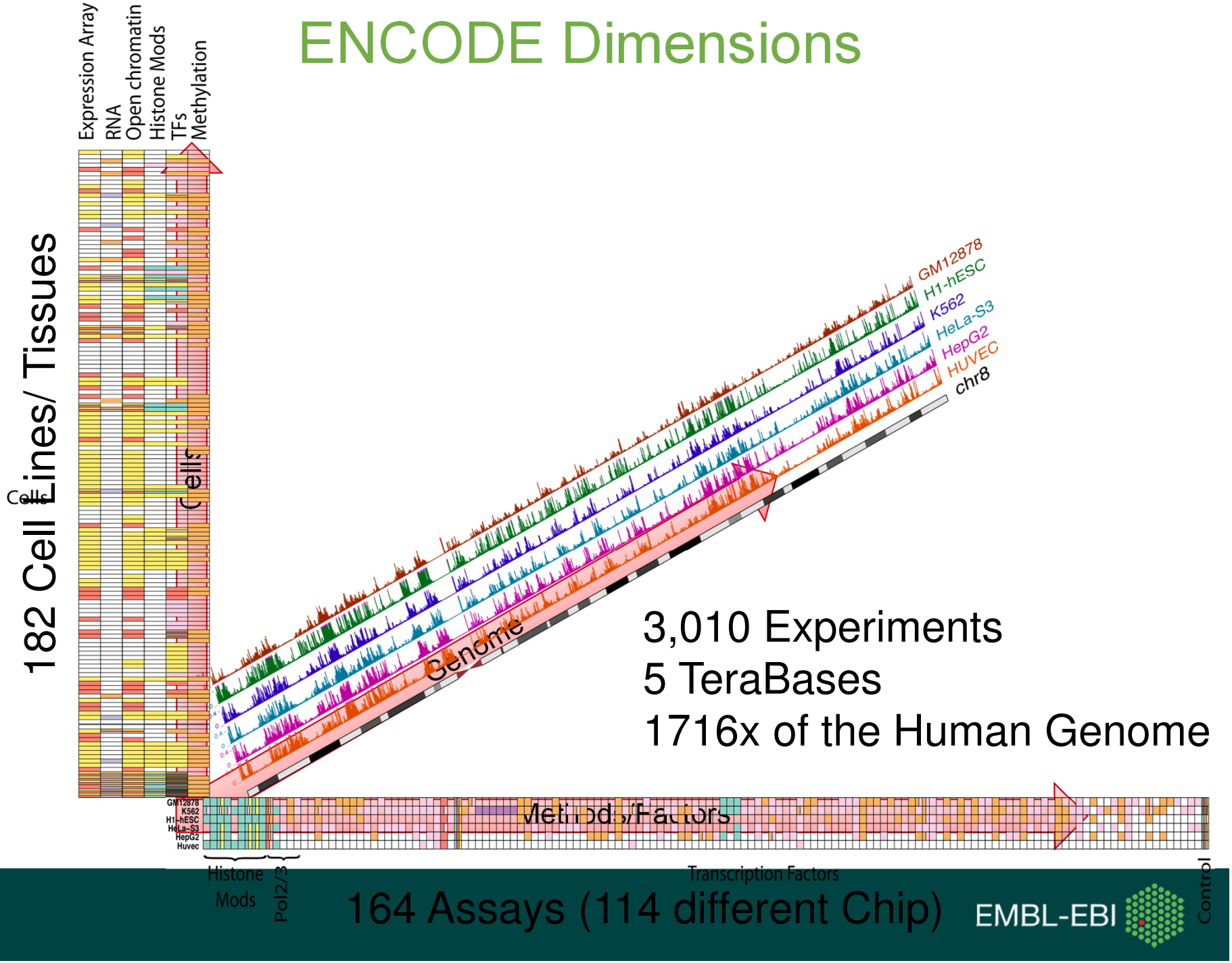
men of European ethnicity will develop Colorectal Cancer between the ages of 15 and 79.

(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 20 & 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 410

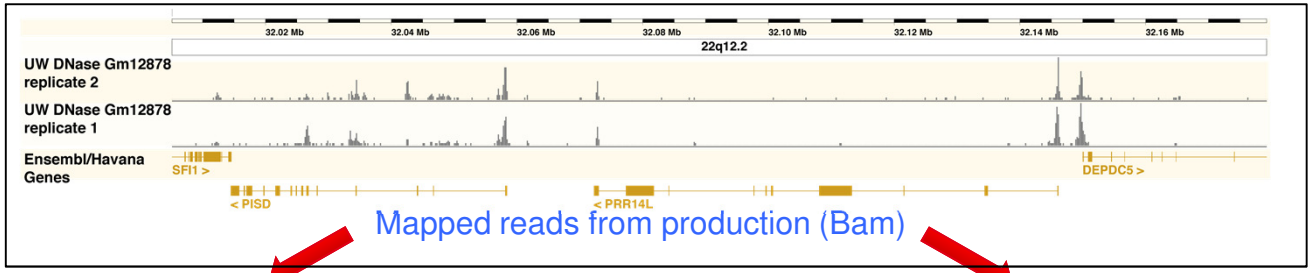
EMBL-EBI

ENCODE Dimensions

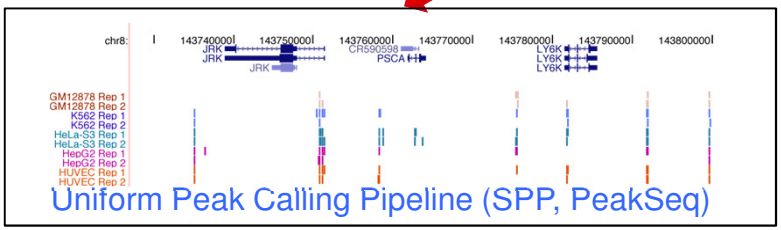


ENCODE Uniform Analysis Pipeline

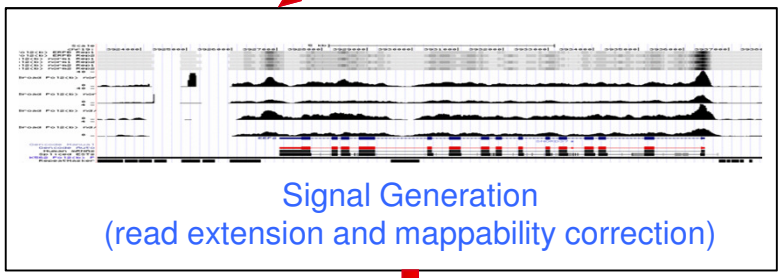
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



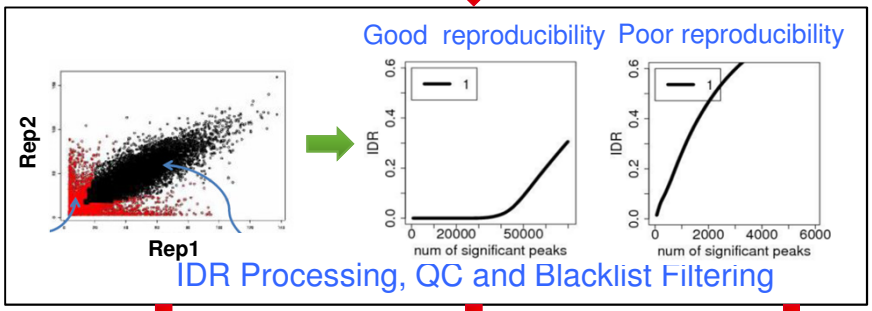
Mapped reads from production (Bam)



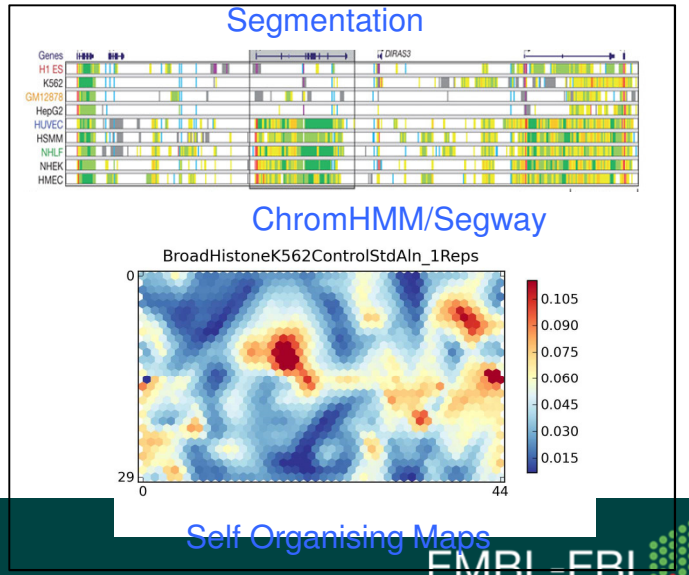
Uniform Peak Calling Pipeline (SPP, PeakSeq)

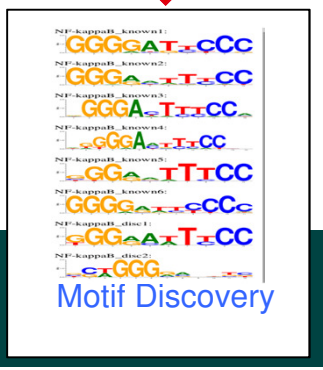


Signal Generation
(read extension and mappability correction)

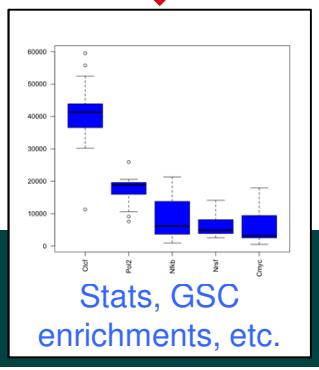


IDR Processing, QC and Blacklist Filtering

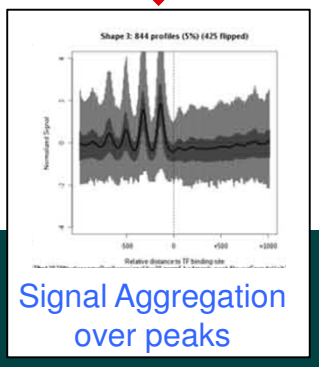




Motif Discovery



Stats, GSC enrichments, etc.

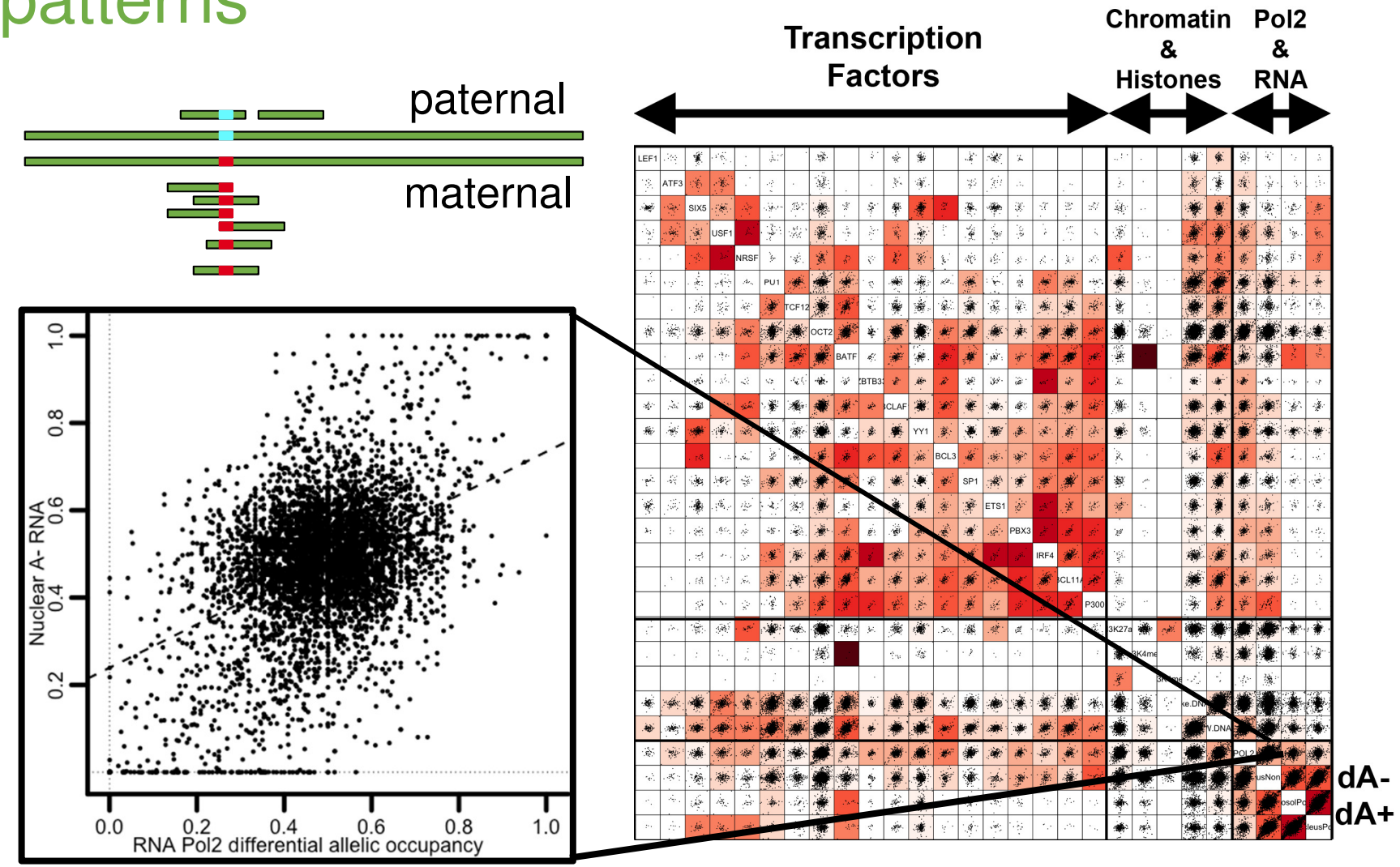


Signal Aggregation over peaks

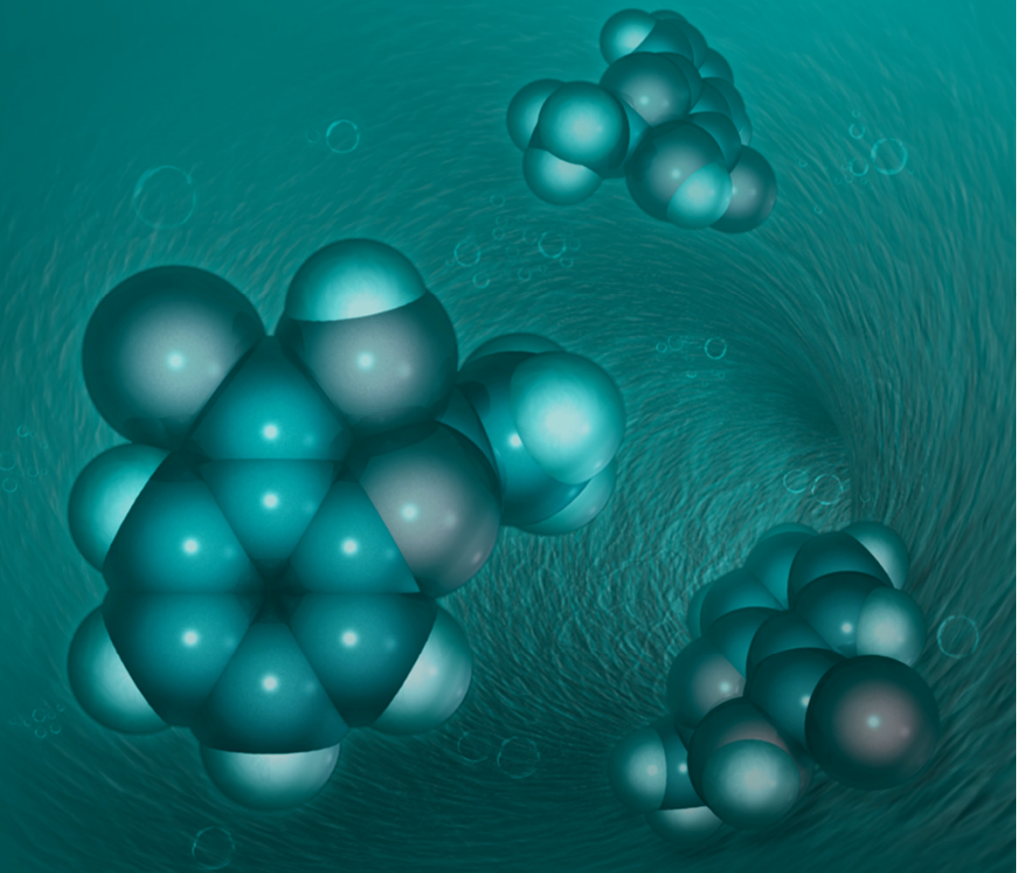
EMBL-EBI

Large-scale analysis of allelic occupancy patterns

Bob Altschuler, Tim Reddy, Joel Rozowsky, Xainjun Dong



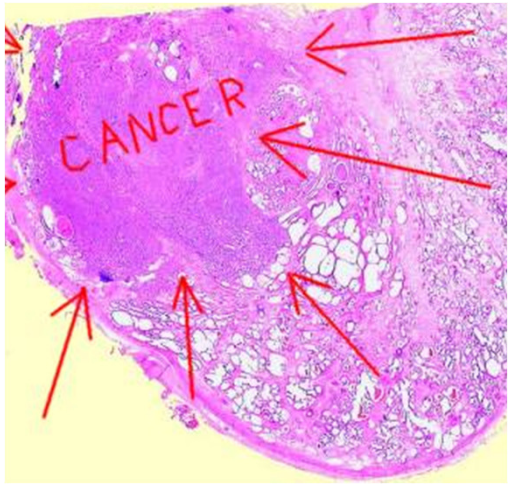
Impact on Medicine



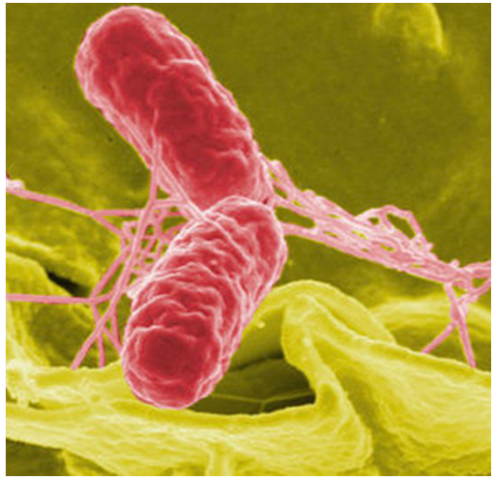
EMBL-EBI

3 big areas of impact for medicine

Germ line
Risk to disease



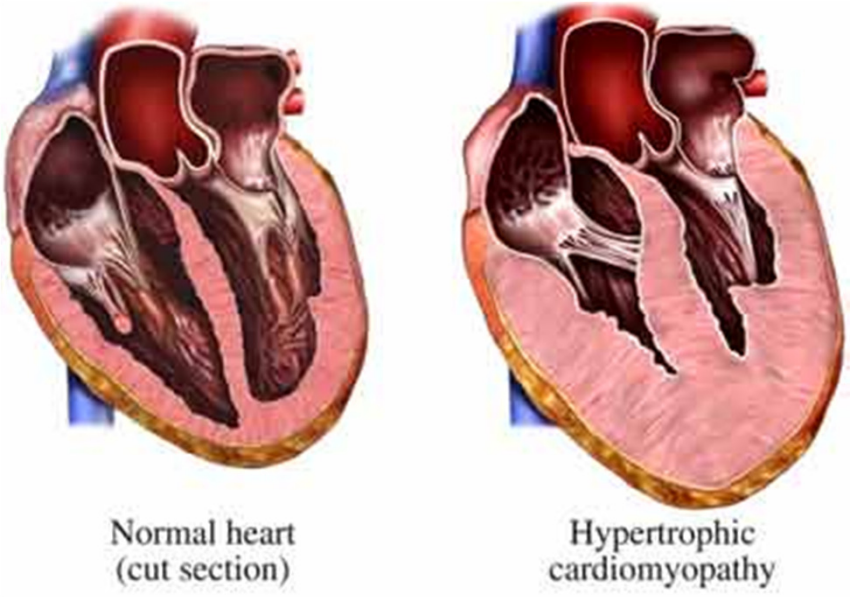
“Precision” cancer medicine



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”

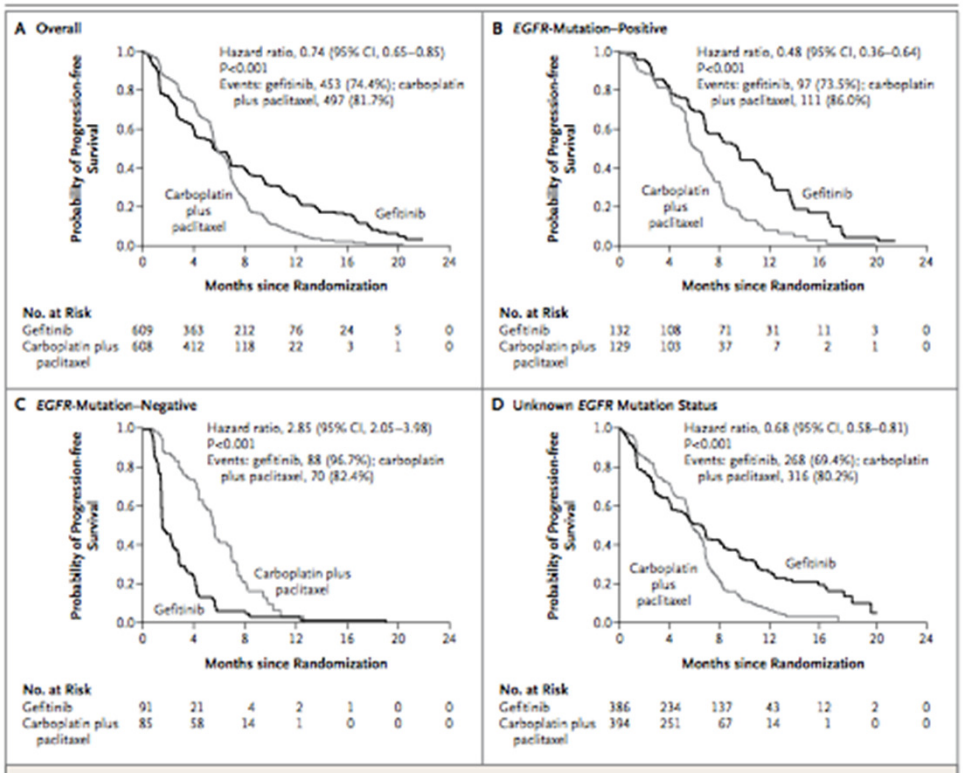


- This goes up to 3-4% if you count some less clinically worrying diseases

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

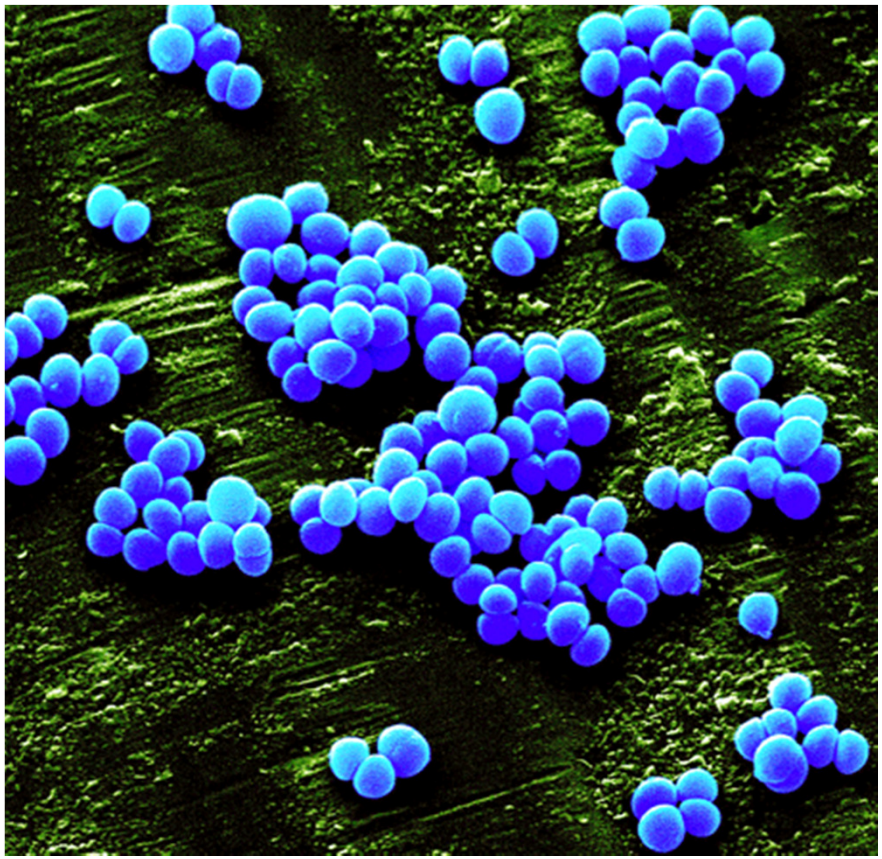
Precision cancer diagnosis

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

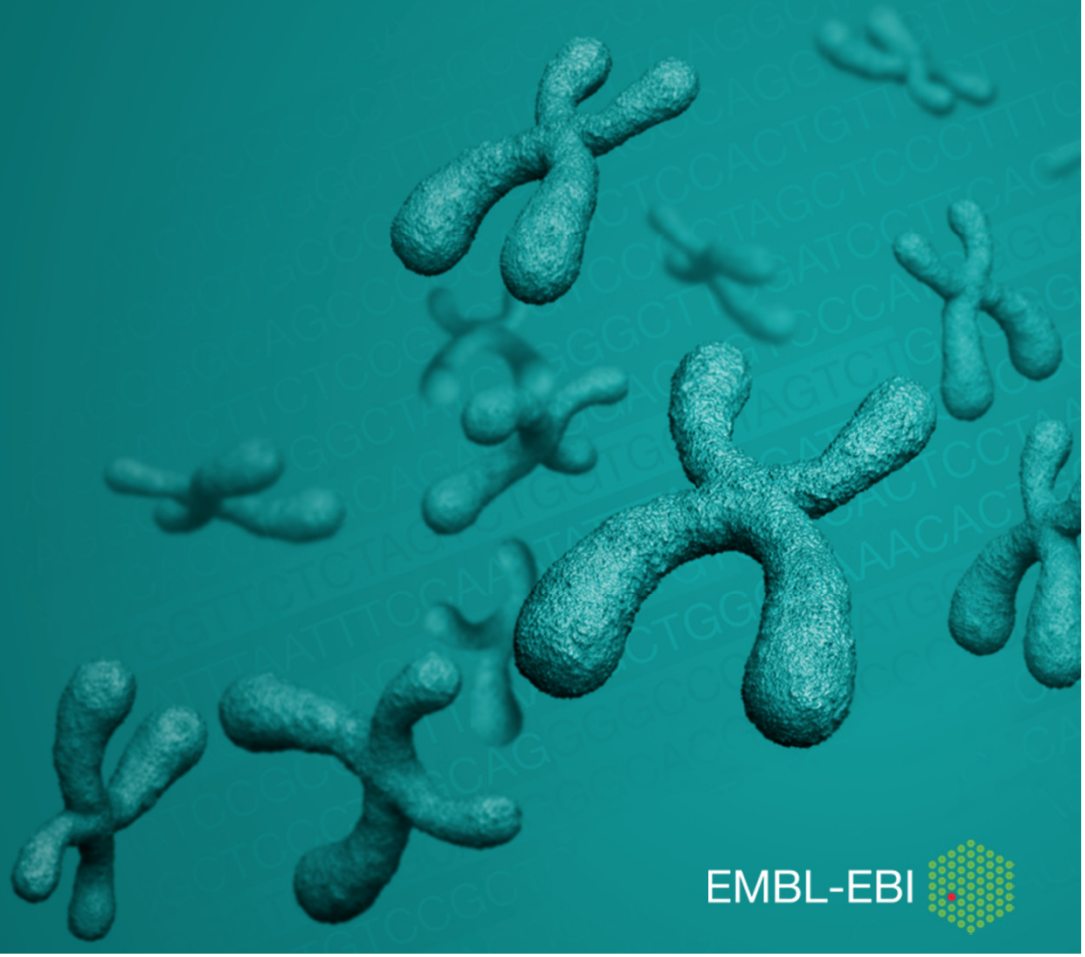


Pathogens

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

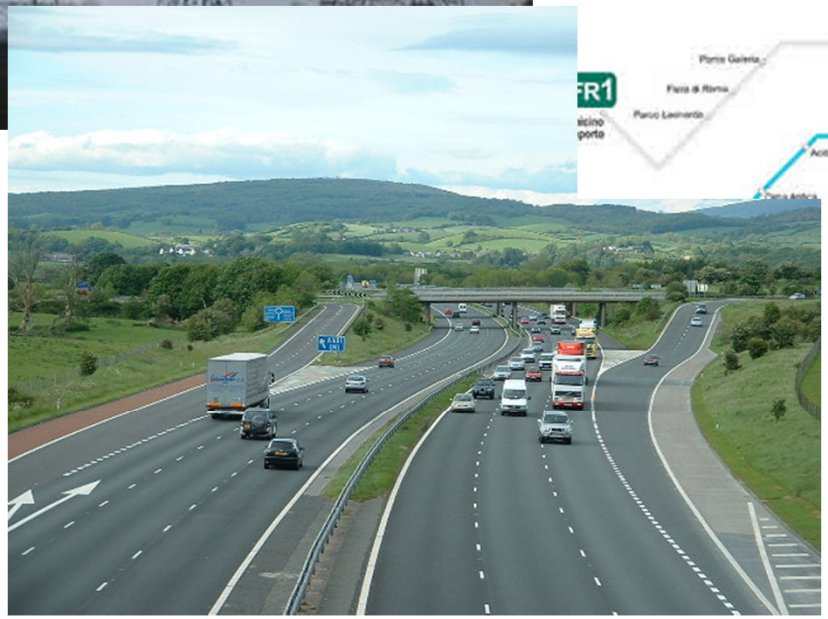


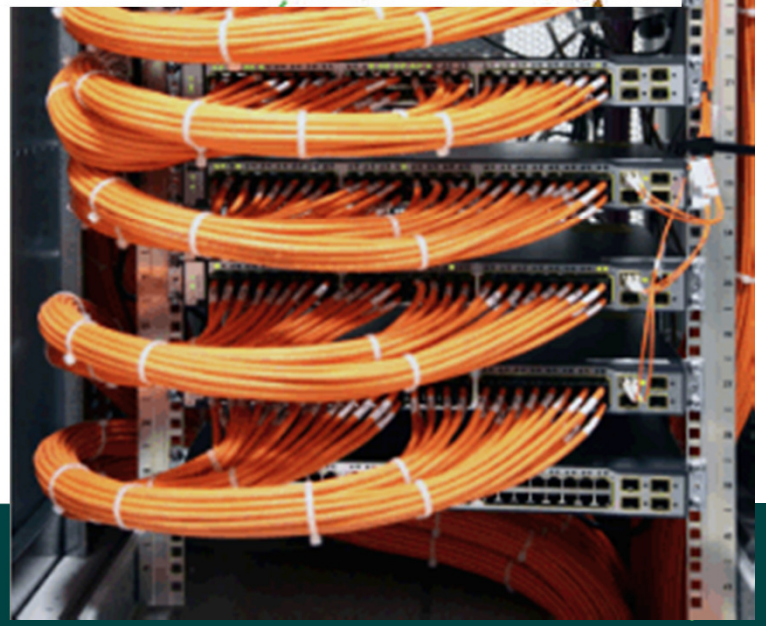
Why we need an infrastructure…



EMBL-EBI

Infrastructures are critical…





But we only notice them when they go wrong

Biology already needs an information infrastructure

- For the human genome
  - (…and the mouse, and the rat, and… x 150 now, 1000 in the future!) - Ensembl
- For the function of genes and proteins
  - For all genes, in text and computational – UniProt and GO
- For all 3D structures
  - To understand how proteins work – PDBe
- For where things are expressed
  - The differences and functionality of cells - Atlas

..But this keeps on going…

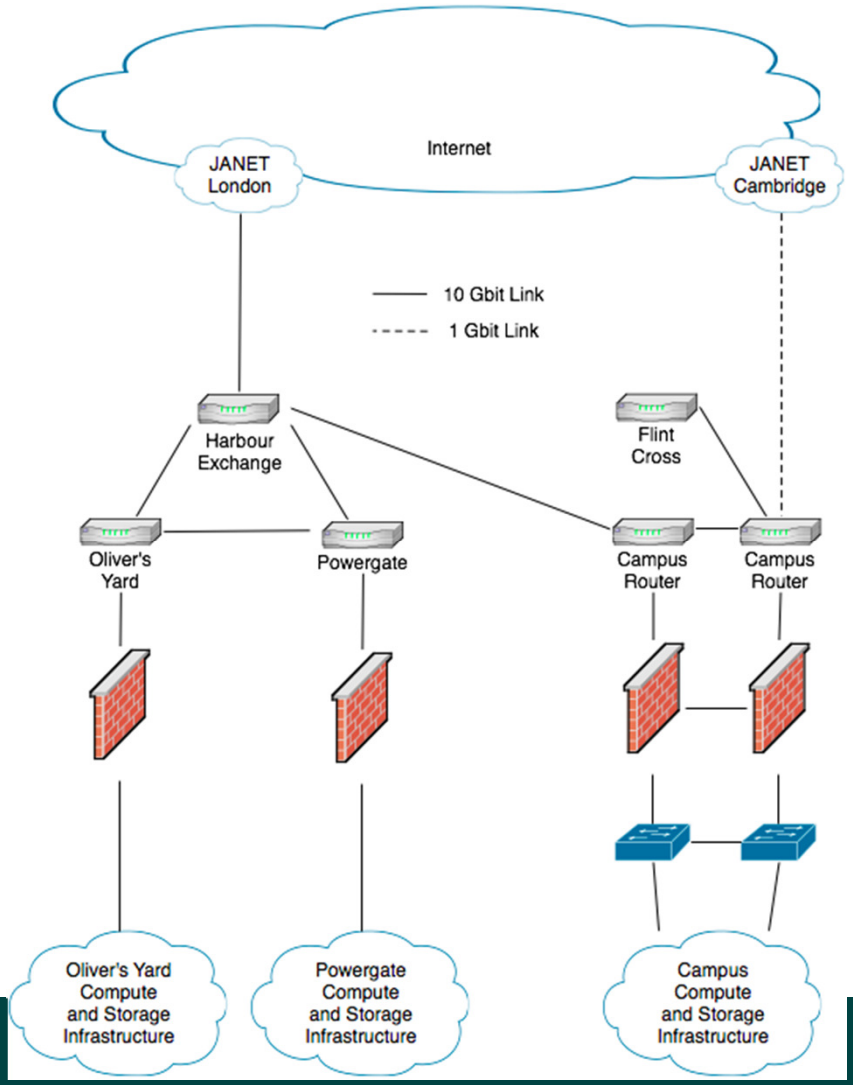
- We have to scale across all of (interesting) life
  - There are a lot of species out there!
- We have to handle new areas, in particular medicine
  - A set of European haplotypes for good imputation
  - A set of actionable variants in germline and cancers
- We have to improve our chemical understanding
  - Of biological chemicals
  - Of chemicals which interfere with Biology

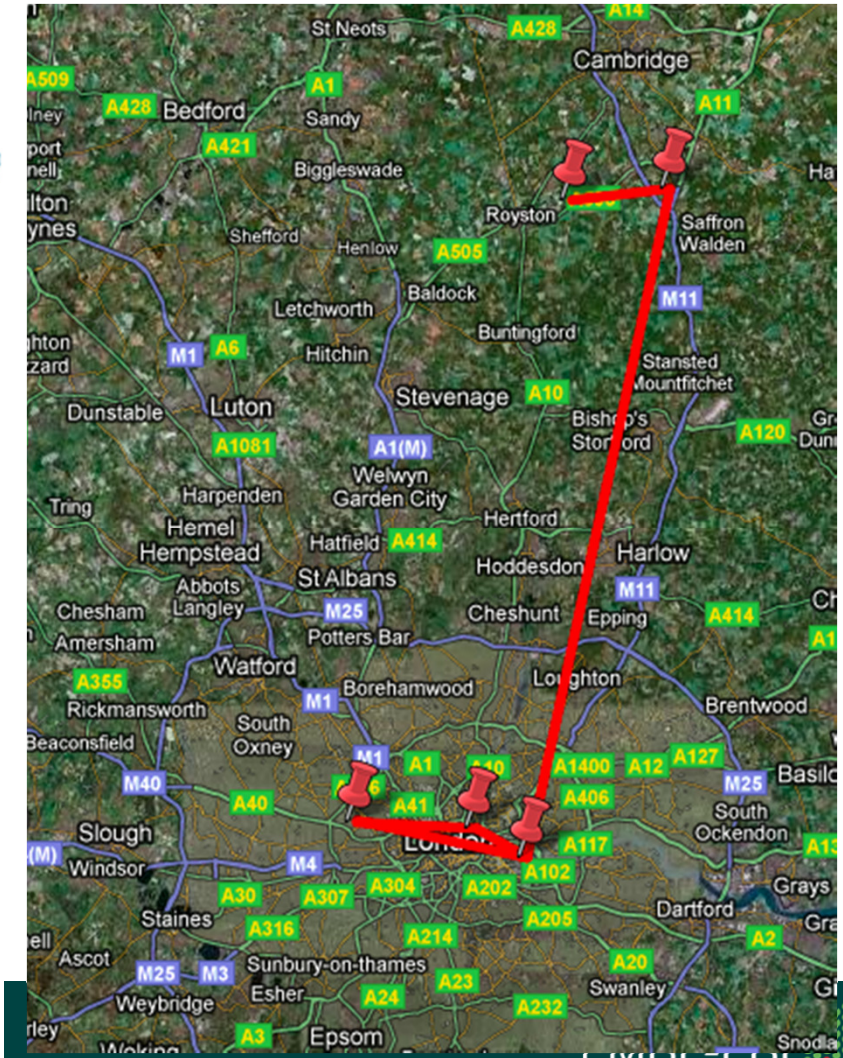
EBI's technical infrastructure

- 30 PB of 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

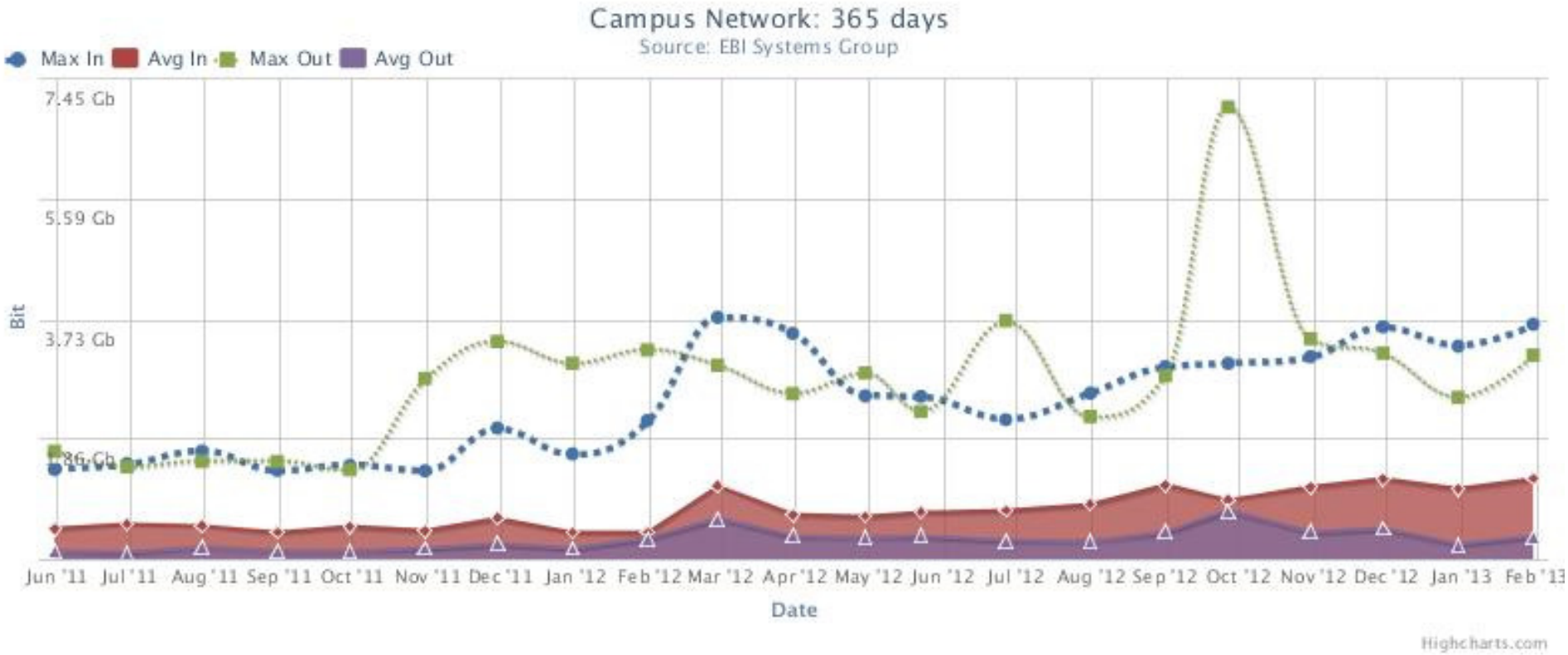
- Janet uplink at 10 Gbit/sec

Machine room architecture

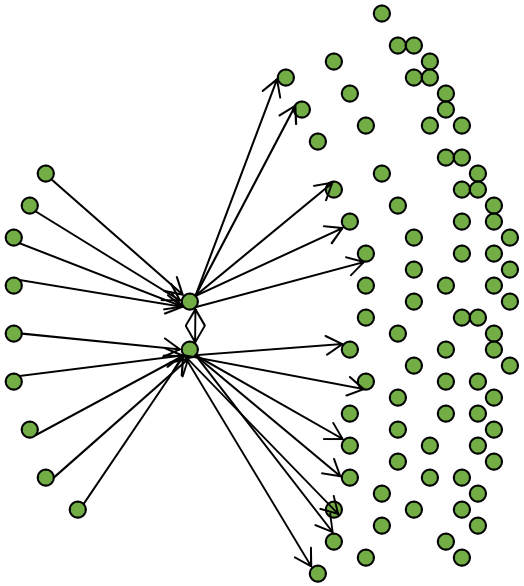




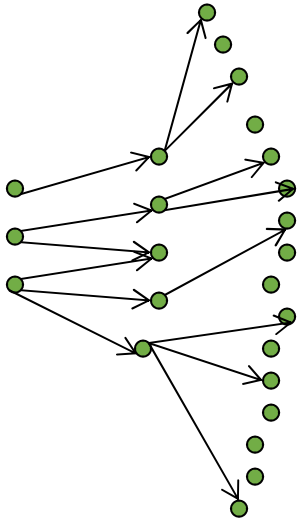
Network usage



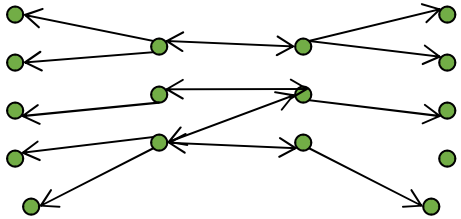
Distribution Patterns are different:



Genomics



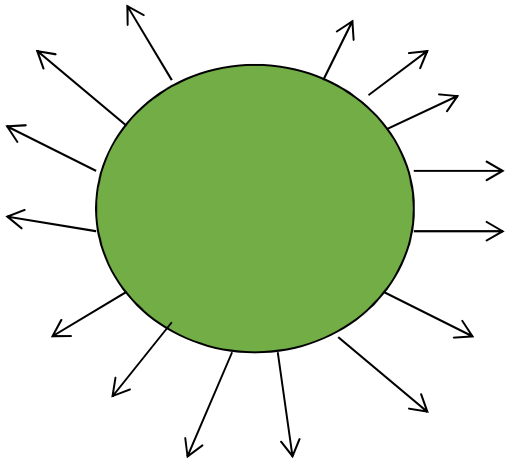
High Energy Physics



Astronomy

How?

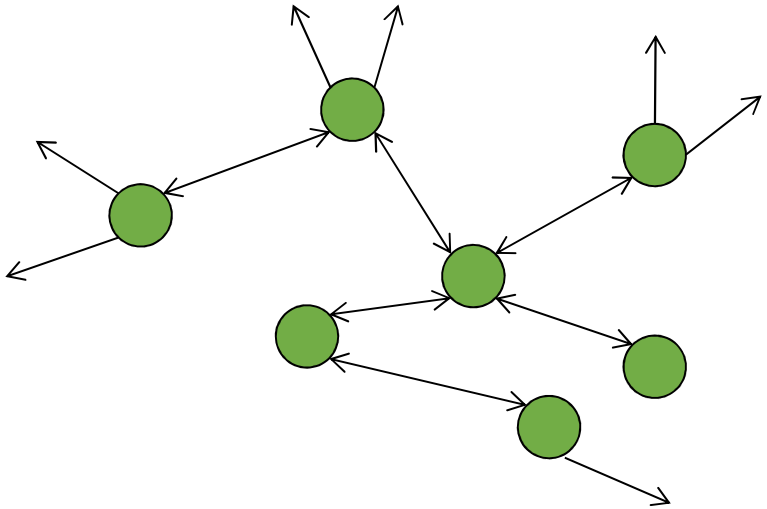
Fully Centralised



Pros: Stability, reuse, Learning ease

Cons: Hard to concentrate Expertise across of life sciences Geographic, language placement Bottlenecks and lack of diversity

Fully Distributed

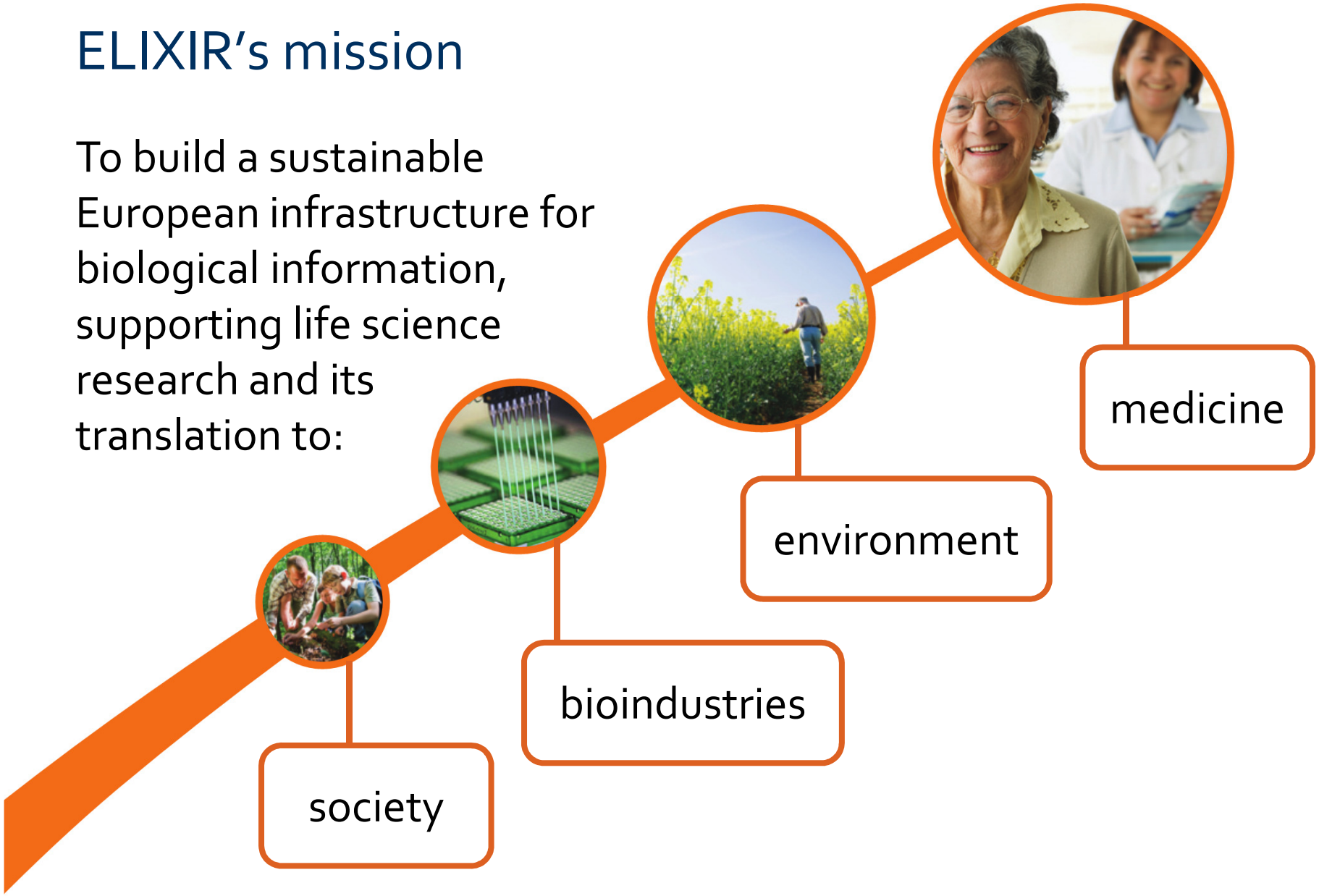


Pros: Responsive, Geographic Language responsive

Cons: Internal communication overhead Harder for end users to learn Harder to provide multi-decade stability

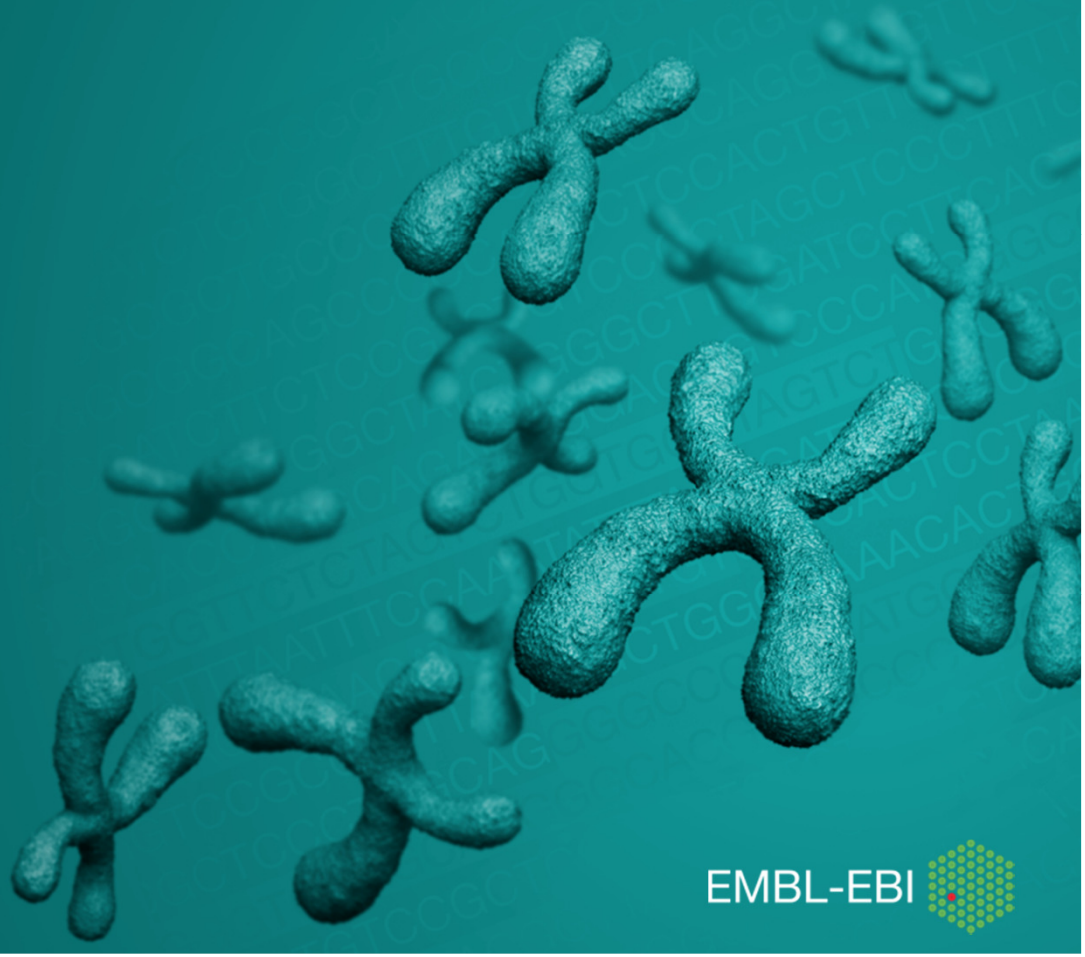
ELIXIR's mission

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



EMBL-EBI

And… just for fun…

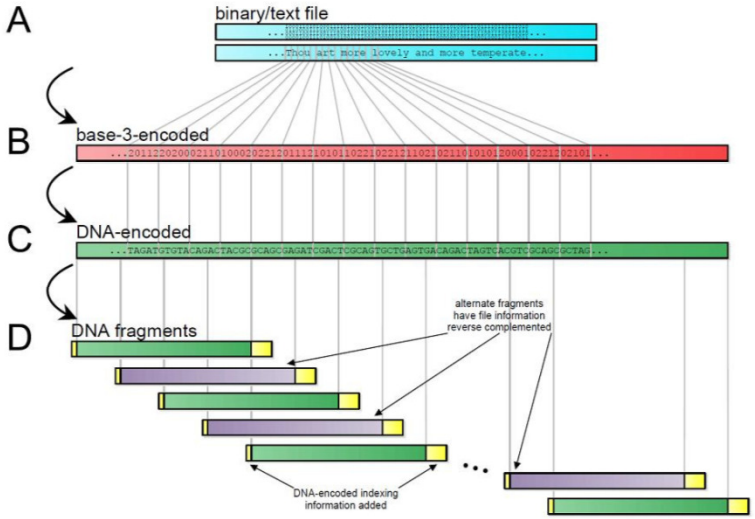


EMBL-EBI

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…



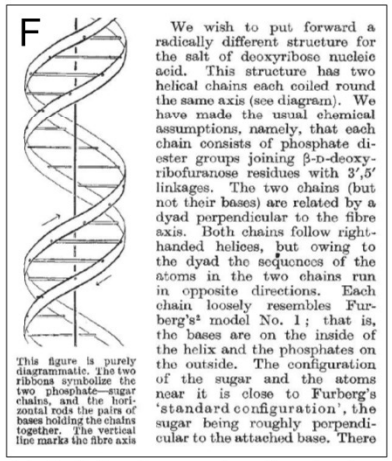
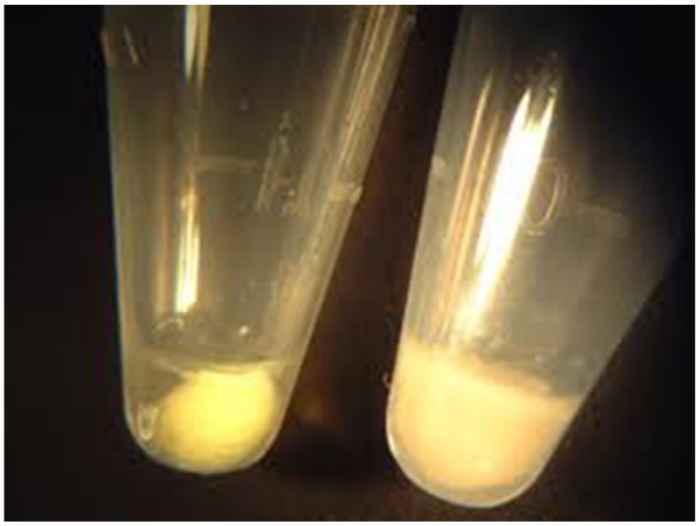
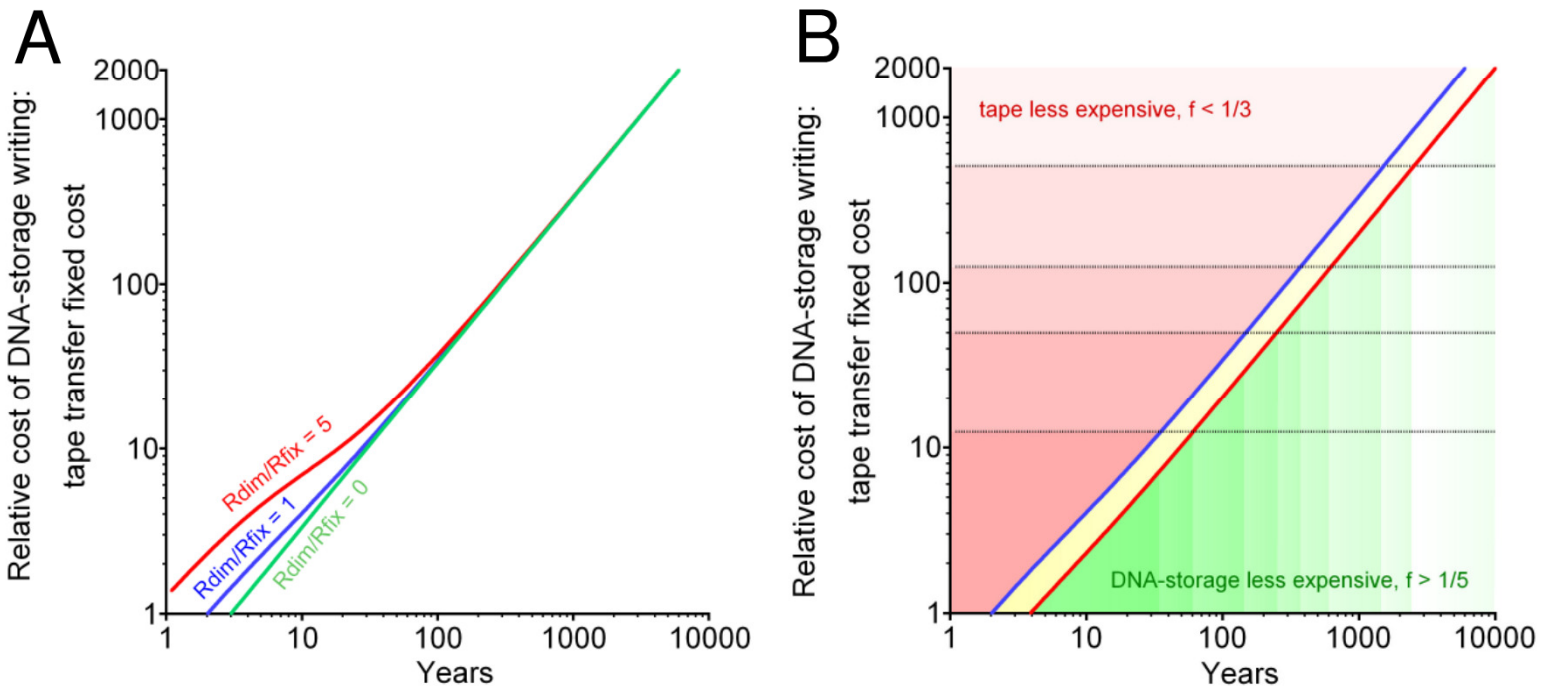


Figure 2 | Digital information encoded in DNA. Digital information (**A**, in blue), here binary digits holding the ASCII codes for part of Shakespeare's sonnet 18, was converted to base-3 (**B**, red) using a Huffman code. This in turn was converted *in silico* to our DNA code (**C**, green), with no 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.



Cost effective?



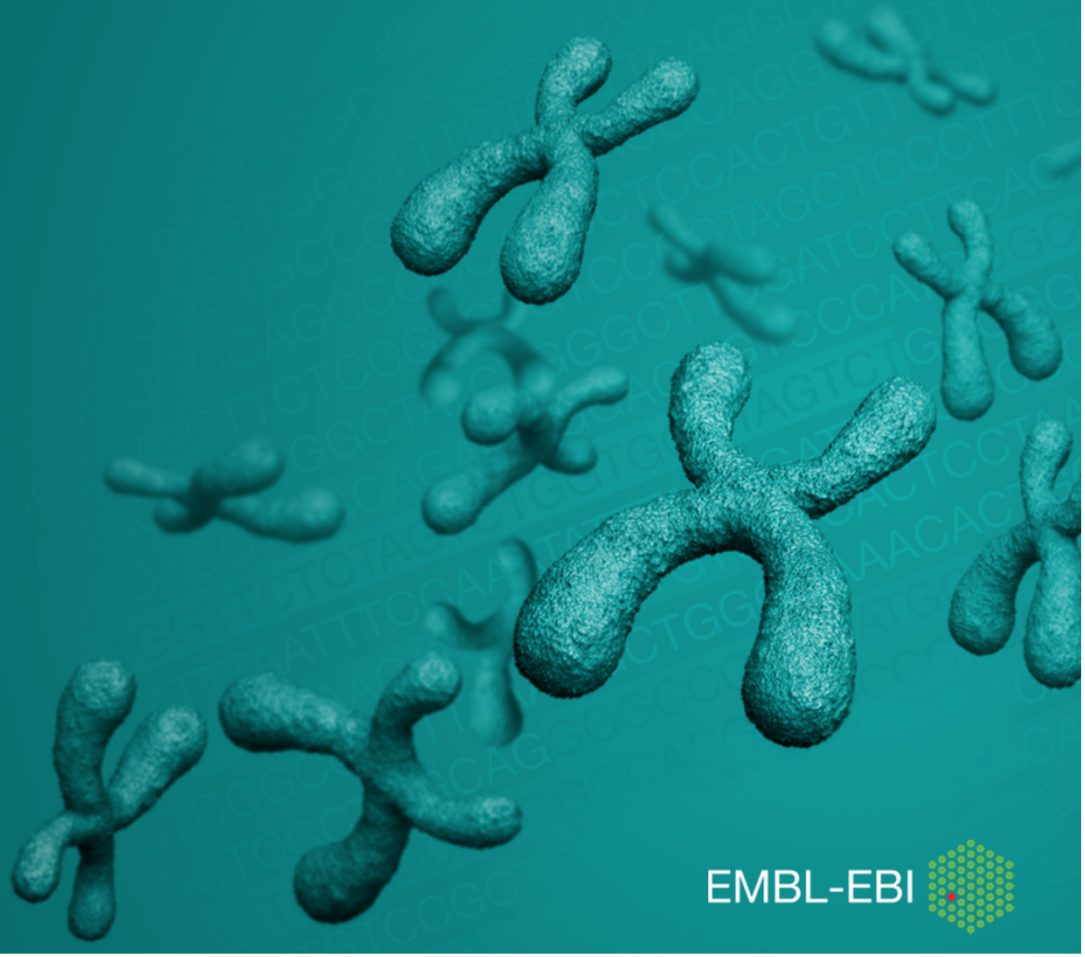

Dave Simonds

Questions?

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

EMBL-EBI is funded by the 20 member states of EMBL, Wellcome Trust, European Union FP7, NIH, BBSRC, MRC and over 20 other funding agencies

How to integrate?



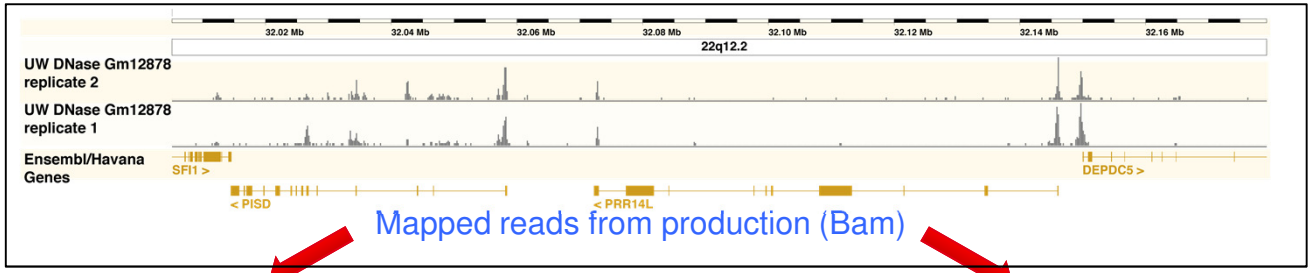
EMBL-EBI

Integration levels

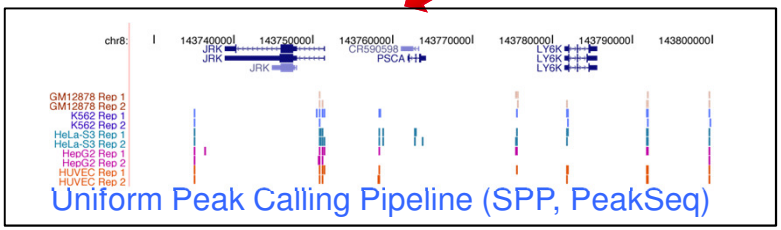
- Low
  - Access ability, formats, identifier tracking, volume
- Medium
  - Concepts, Ontologies, Samples
- High
  - Statistical, Domain ontology, discovery

ENCODE Uniform Analysis Pipeline

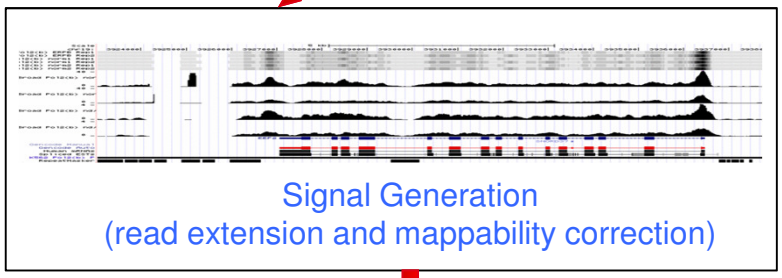
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



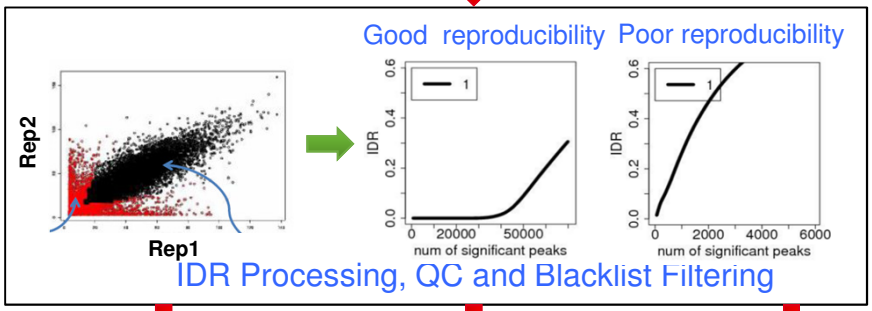
Mapped reads from production (Bam)



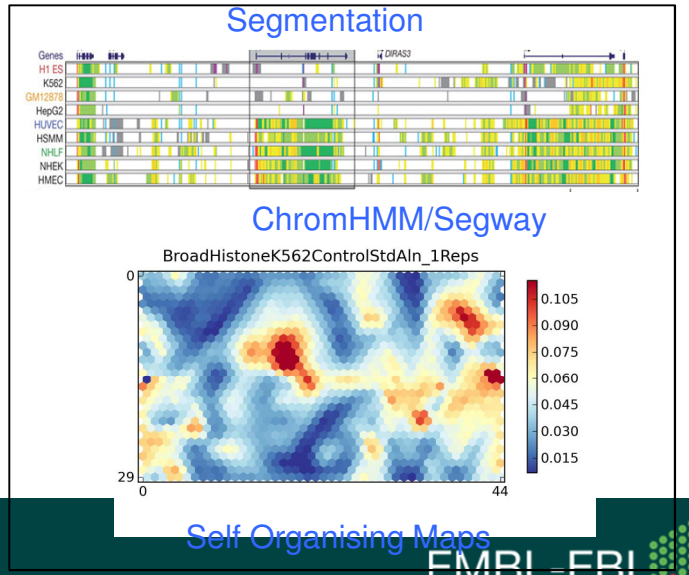
Uniform Peak Calling Pipeline (SPP, PeakSeq)

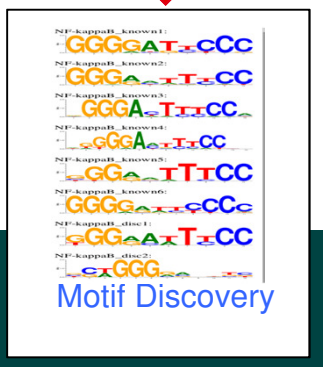


Signal Generation
(read extension and mappability correction)

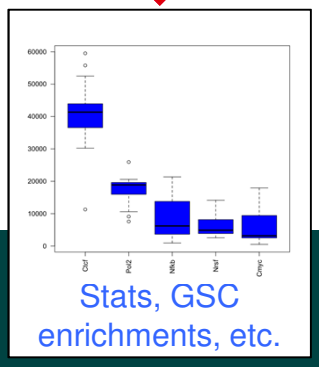


IDR Processing, QC and Blacklist Filtering

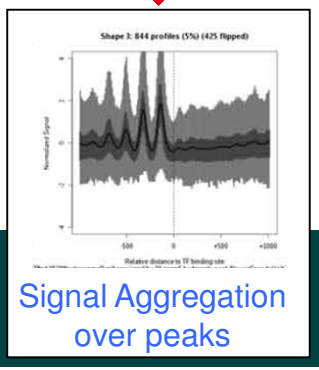




Motif Discovery



Stats, GSC enrichments, etc.



Signal Aggregation over peaks

EMBL-EBI

Engineering is not so easy