Computational Challenges in Life Sciences Research Infrastructures

Alvis Brazma

European Bioinformatics Institute

European Molecular Biology Laboratory

EMBL-EBI

European Bioinformatics Institute (EBI)

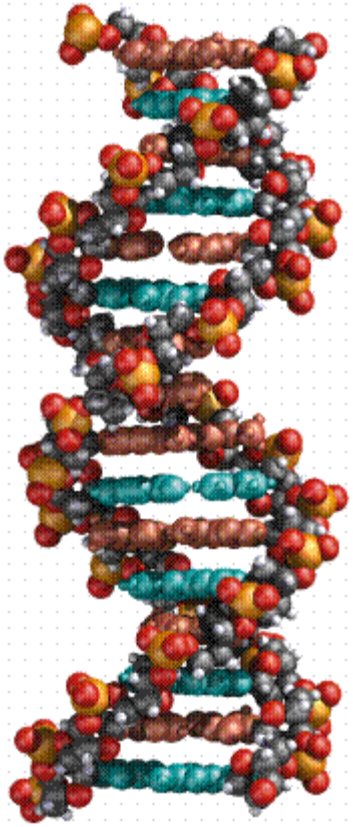
- EBI is in Hinxton, ~10 miles South of Cambridge, UK
  Wellcome Trust Genome Campus
- EBI is part of EMBL, ~like CERN for molecular biology
- ~500 scientific and IT staff at EBI
- Hosting the ELIXIR node (details to follow)

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 how to predict what they are
- Instead we determine attributes of molecules and store them in *globally accessible, open* databases for mining, exploration and model building

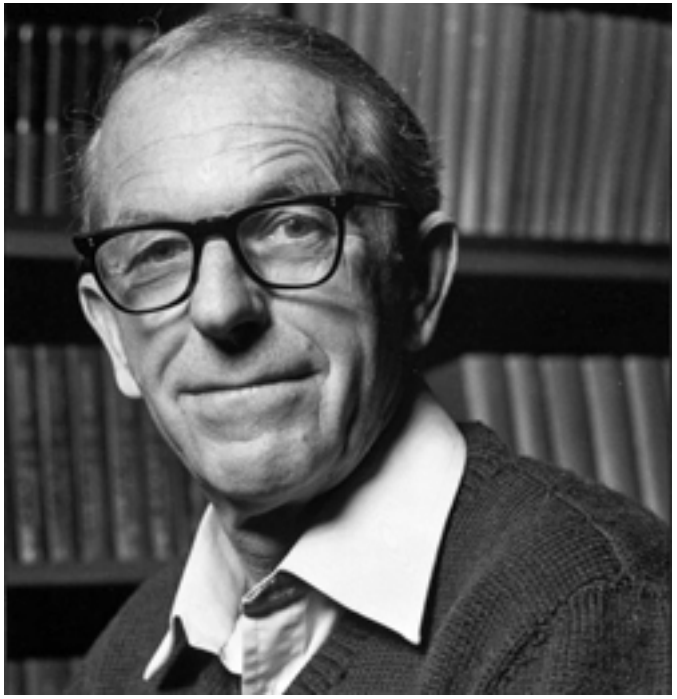
DNA molecule is **a perfect medium for recording information**

- The particular sequence of A, T, G, C in DNA has little effect on its structure – it will usually be the same the double-helix
- **Each** of the 4 different **letter**s can encode **2 bits** of **information**
- The physical distance between nucleotide pairs is about 0.34 nm
- Thus the information storage density in DNA is roughly 6×10^{8} bits/cm - approximately **12.5 CD-Roms per cm**



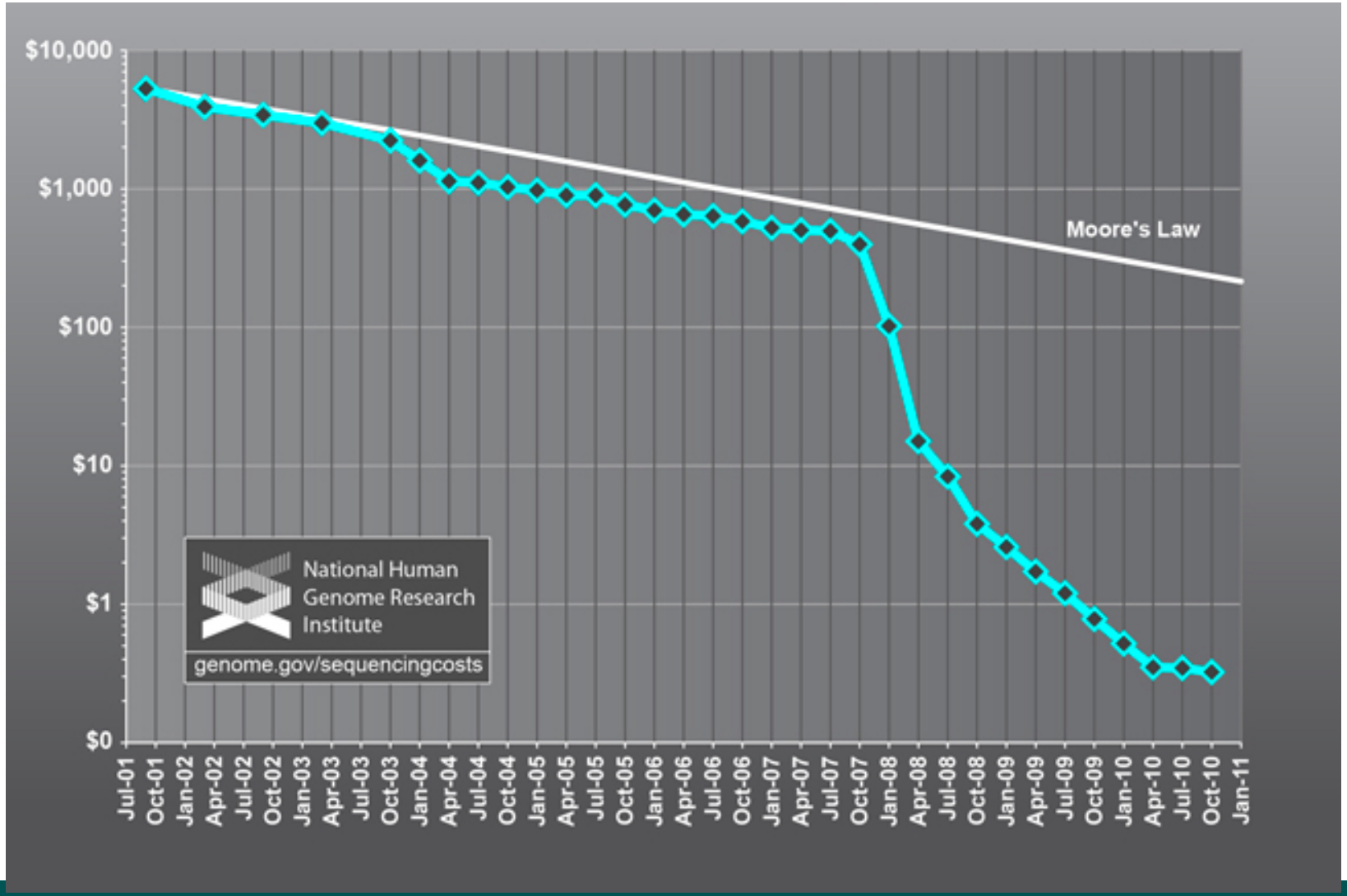
We can routinely read small fragments of DNA

- 1977-1990 – 500 bp, manual tracking
- 1990-2000 – 500 bp, computational tracking, 1D, “capillary”
- 2007-now – 20-100bp, 2D systems very cheaply, (“2nd Generation” or NGS)
- Soon >5kb, Real time “3rd Generation”

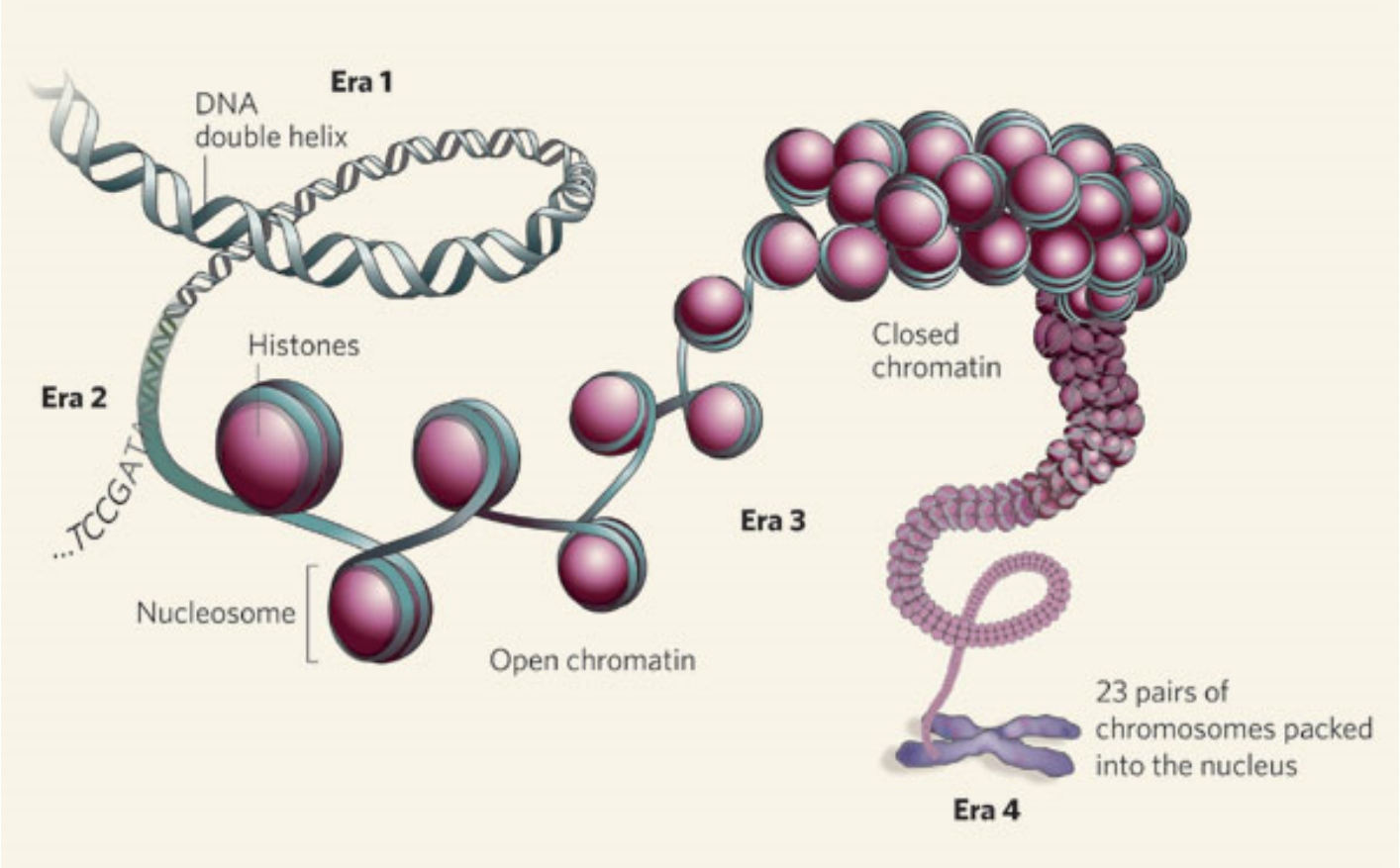


Fred Sanger, inventor of terminator DNA sequencing

Costs have come exponentially down

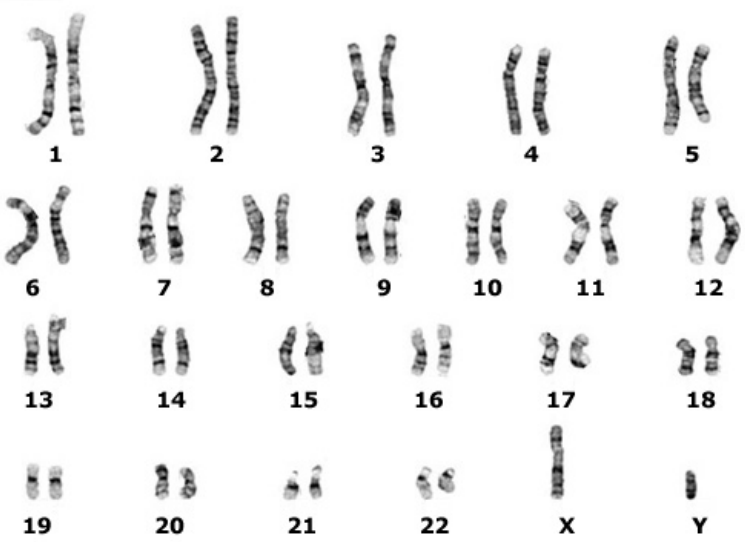


DNA is where the **genome** information is encoded



Nature 448, 548-549

The **human genome** contains DNA of the total length of 2 x **3 000 000 000 letters** packed in 23 chromosome pairs



- Every cell in human has two copies of full genome
- Roughly 2500 CD-roms

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*
  - As the first step of analysis we need to assemble the ‘short fragments’ into full genomes
  - But that’s only the beginning of the analysis, e.g., we need to identify *genes*, their variants, function, etc

JUNE 25/26 2011

FT Weekend Magazine

Science Special Edition

FT Global 500

Free inside 24-page pullout magazine

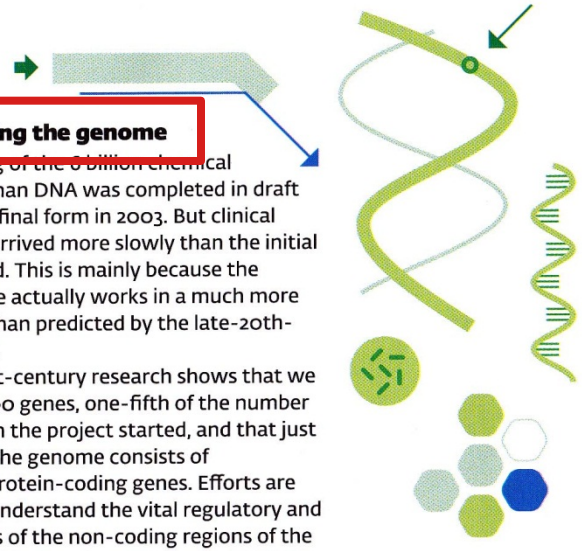
SCIENCE SPECIAL

SCIENCE'S 10 HOTTEST FIELDS

By *Clive Cookson*
Illustrations by *Leandro Castelao*

The composition of the cosmos

Cosmologists are still coming to terms with the surprising 1998 discovery that mysterious "dark energy" is accelerating the expansion of the universe. Astronomers calculate that dark energy makes up 74 per cent of the universe but they have no idea what it is. The marginally less mysterious "dark matter" makes up 22 per cent – leaving just 4 per cent for all the ordinary matter in the objects we can see directly. Dark matter is probably made of massive particles, which interact so little with ordinary matter that instruments have been unable to detect them. But the subatomic debris from smashing atoms together at Cern's Large Hadron Collider may offer a chance of identification in the future.



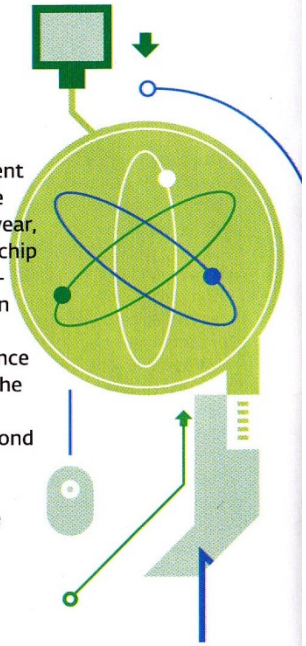
Understanding the genome

The sequencing of the 3 billion chemical "letters" of human DNA was completed in draft in 2000 and in final form in 2003. But clinical benefits have arrived more slowly than the initial hype suggested. This is mainly because the human genome actually works in a much more complex way than predicted by the late-20th-century model.

Twenty-first-century research shows that we have only 21,000 genes, one-fifth of the number predicted when the project started, and that just 1.5 per cent of the genome consists of conventional protein-coding genes. Efforts are under way to understand the vital regulatory and other functions of the non-coding regions of the genome, once dismissed wrongly as "junk DNA".

Leap for quantum computing

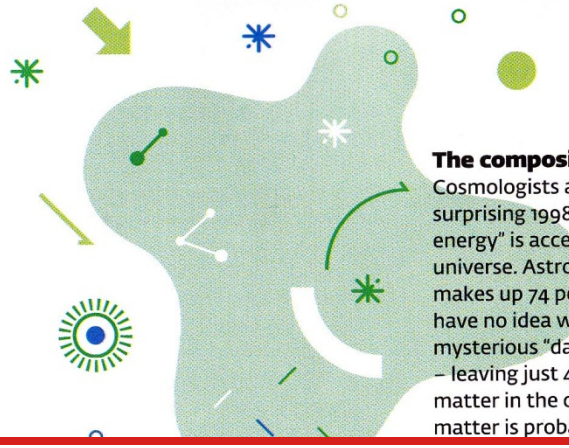
Quantum computing offers the possibility of a radical transition: a fundamentally different way of processing data. Prototype devices are beginning to emerge around the world. Last year, a team at Bristol University made a photonic chip that processes data according to the counter-intuitive rules of quantum physics, rather than conventional electronics. Because quantum particles can influence one another at a distance ("entanglement") and be in several places at the same time ("superposition"), they could in principle perform parallel calculations far beyond the capability of today's supercomputers. Governments and companies are investing hundreds of millions of dollars but formidable technical barriers must be overcome.



SCIENCE SPECIAL

SCIENCE'S 10 HO FIE

By *Clive C*
Illustrati



The composition of the cosmos
Cosmologists are still coming to terms with the surprising 1998 discovery that mysterious "dark energy" is accelerating the expansion of the universe. Astronomers calculate that dark energy makes up 74 per cent of the universe but they have no idea what it is. The marginally less mysterious "dark matter" makes up 22 per cent – leaving just 4 per cent for all the ordinary matter in the objects we can see directly. Dark matter is probably made of massive particles,

Understanding the genome

The sequencing of the 6 billion chemical "letters" of human DNA was completed in draft in 2000 and in final form in 2003. But clinical benefits have arrived more slowly than the initial hype suggested. This is mainly because the human genome actually works in a much more complex way than predicted by the late-20th-century model.

Understar
"letters" of h
in 2000 and
benefits hav
hype sugges
human gen
complex wa
century mod

Twenty-first-century research shows that we have only 21,000 genes, one-fifth of the number predicted when the project started, and that just 1.5 per cent of the genome consists of conventional protein-coding genes. Efforts are under way to understand the vital regulatory and other functions of the non-coding regions of the genome, once dismissed wrongly as "junk DNA".

technical barriers must be overcome.

SCIENCE SPECIAL

SCIENCE'S

HO
FI

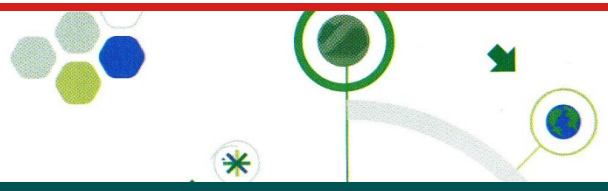
By *Cliv*
Illustra

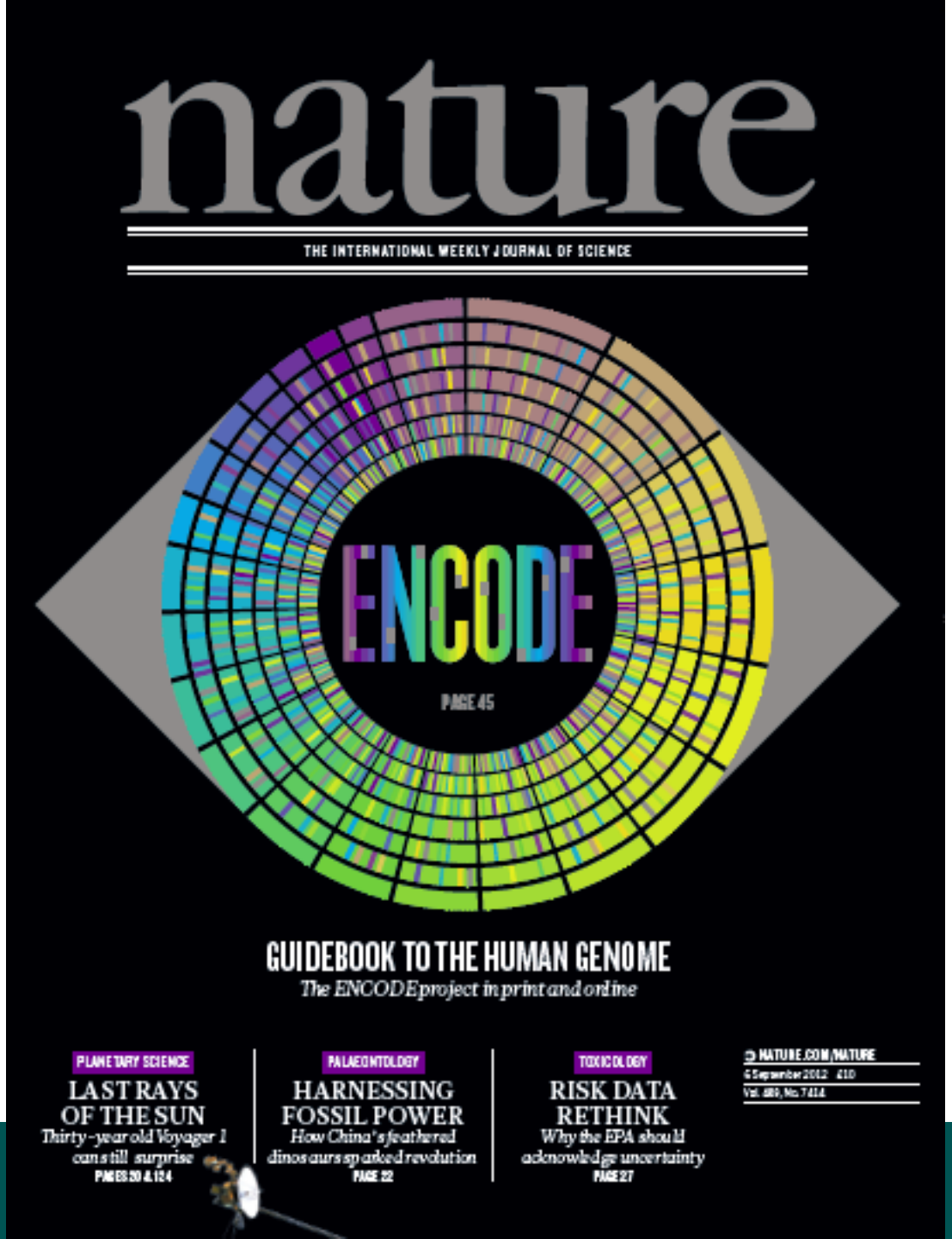
The composition of the cosmos
Cosmologists are still coming to terms with the surprising 1998 discovery that mysterious "dark energy" is accelerating the expansion of the universe. Astronomers calculate that dark energy makes up 74 per cent of the universe but they have no idea what it is. The marginally less mysterious "dark matter" makes up 22 per cent – leaving just 4 per cent for all the ordinary matter in the objects we can see directly. Dark matter is probably made of massive particles,

Understanding the genome

Twenty-first-century research shows that we have only 21,000 genes, one-fifth of the number predicted when the project started, and that just 1.5 per cent of the genome consists of conventional protein-coding genes. Efforts are under way to understand the vital regulatory and other functions of the non-coding regions of the genome, once dismissed wrongly as "junk DNA".

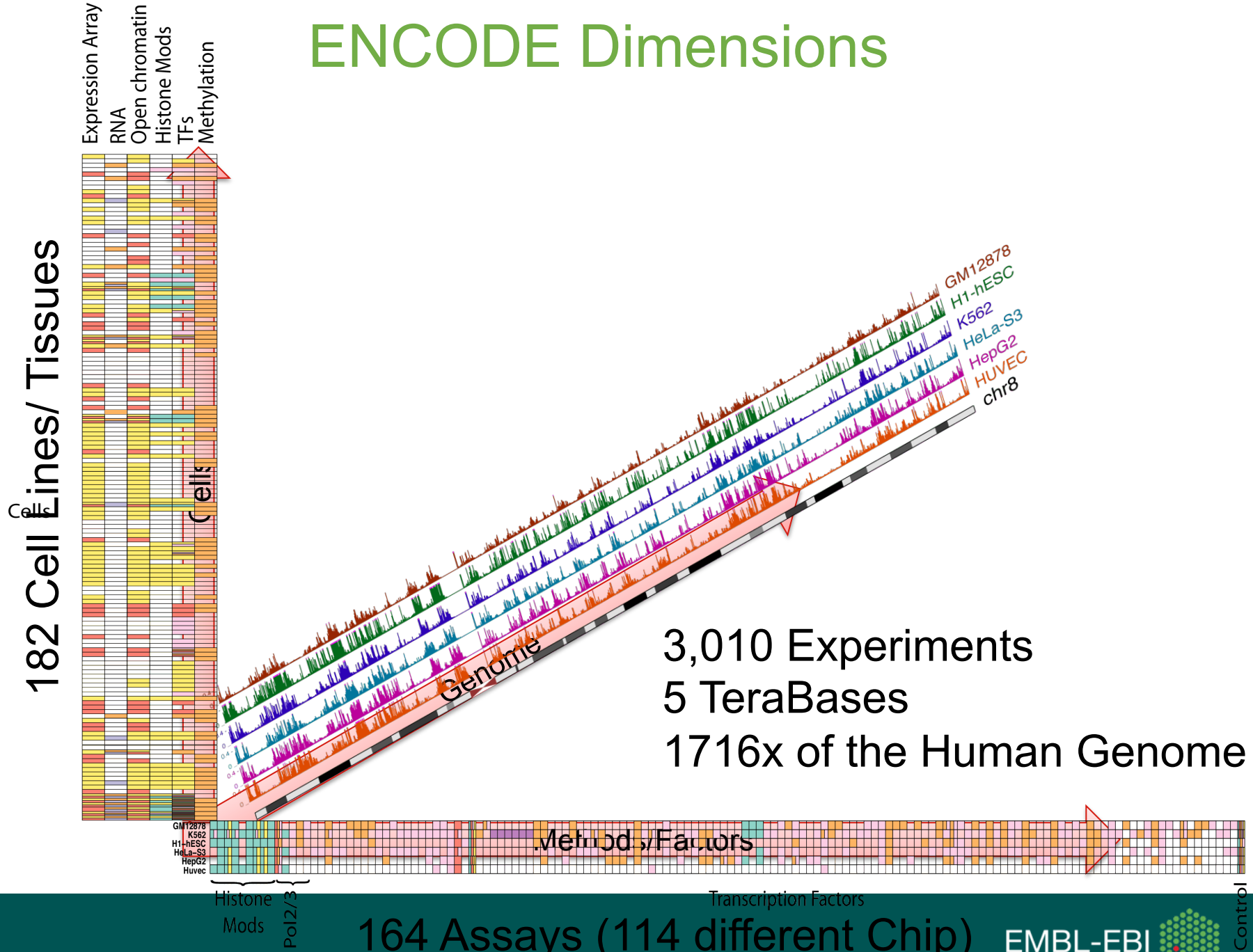
Unders
"letters"
in 2000
benefits
hype sug
human g
complex
century
Twent
have onl
predicte
1.5 per cent of the genome consists of conventional protein-coding genes. Efforts are under way to understand the vital regulatory and other functions of the non-coding regions of the genome, once dismissed wrongly as "junk DNA".




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

EMBL-EBI

ENCODE Dimensions



3,010 Experiments
5 TeraBases
1716x of the Human Genome

164 Assays (114 different Chip)

EMBL-EBI

PERSPECTIVES

International network of cancer genome projects

The International Cancer Genome Consortium*

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Systematic studies of more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies.

The genomes of all cancers accumulate somatic mutations[1]. These include nucleotide substitutions, small insertions and deletions, chromosomal rearrangements and copy number changes that can affect protein-coding or regulatory components of genes. In addition, cancer genomes usually acquire somatic epigenetic 'marks' compared to non-neoplastic tissues from the same organ, notably changes in the methylation status of cytosines at CpG dinucleotides.

A subset of the somatic mutations in cancer cells confers oncogenic properties such as growth advantage, tissue invasion and metastasis, angiogenesis, and evasion of apoptosis[2]. These are termed 'driver' mutations. The identification of driver mutations will provide insights into cancer biology and highlight new drug targets and diagnostic tests. Knowledge of cancer mutations has already led to the development of specific therapies, such as trastuzumab for *HER2* (also known as *NEU* or *ERBB2*) positive breast cancers[3] and imatinib, which

incomplete studies; (3) lack of standardization across studies could diminish the opportunities to merge and compare data sets; (4) the spectrum of many cancers is known to vary across the world; and (5) an international consortium will accelerate the dissemination of data sets and analytical methods into the user community.

Working groups were created to develop strategies and policies that would form the basis for participation in the ICGC. The goals of the consortium (Box 1) were released in April 2008 (http://www.icgc.org/files/ICGC_April_29_2008.pdf). Since then, working groups and initial member projects have further refined the policies and plans for international collaboration.

Bioethical framework

ICGC members agreed to a core set of bioethical elements for consent as a precondition of membership (Box 2). The Ethics and Policy

Vol 464|15 April 2010|doi:10.1038/nature08987

nature

PERSPECTIVES

International network of cancer genome projects

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Systematic studies of more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies

ponents of genes. In addition, cancer genomes usually acquire somatic epigenetic 'marks' compared to non-neoplastic tissues from the same organ, notably changes in the methylation status of cytosines at CpG dinucleotides.

A subset of the somatic mutations in cancer cells confers oncogenic properties such as growth advantage, tissue invasion and metastasis, angiogenesis, and evasion of apoptosis[2]. These are termed 'driver' mutations. The identification of driver mutations will provide insights into cancer biology and highlight new drug targets and diagnostic tests. Knowledge of cancer mutations has already led to the development of specific therapies, such as trastuzumab for *HER2* (also known

sets and analytical methods into the user community.

Working groups were created to develop strategies and policies that would form the basis for participation in the ICGC. The goals of the consortium (Box 1) were released in April 2008 (http://www.icgc.org/files/ICGC_April_29_2008.pdf). Since then, working groups and initial member projects have further refined the policies and plans for international collaboration.

Bioethical framework

ICGC members agreed to a core set of bioethical elements for consent as a precondition of membership (Box 2). The Ethics and Policy

Data size

- Cancer genomics projects ~30 PB (= 10^{15} B) of sequencing data
- However, the problem is not the size of the data, but how to analyse and interpret it!

Looking for *fusion genes* in breast cancer using 1000 genome RNAdata for control

- Mapping 391 tumor samples samples to the genome
  - 2172 hours = 90.5 days CPU time
  - 11730 GB memory
- Mapping 668 RNA samples from the 1000 genome project
  - 3711 hours = 155 days CPU time and
  - 20040 GB memory
- This is just one small specific one postdoc project!

Infrastructures are critical…

But we only notice them when they go wrong

Biology already relies on 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 – ArrayExpress, Expression Atlas

International Cancer Genome Consortium

Enter keywords Search

Home | Cancer Genome Projects | Committees and Working Groups | Policies and Guidelines | Media

ICGC Cancer Genome Projects

Committed projects to date: 53

Sort by: Project

Bladder Cancer United States	**Blood Cancer** United States	**Blood cancer** South Korea
Bone Cancer United Kingdom	**Brain Cancer** Canada	**Brain Cancer** United States
Breast Cancer European Union / United Kingdom	**Breast Cancer** France	**Breast Cancer** Mexico
Breast Cancer United Kingdom	**Breast Cancer** United States	**Breast cancer** South Korea

ICGC Goal: To obtain a **comprehensive** description of **genomic, transcriptomic and epigenomic changes** in **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe.

Read more »

Launch Data Portal »

Apply for Access to Controlled Data »

Announcements

4/December/2012 - The ICGC Data Coordination Center (DCC) is pleased to announce the ICGC data portal data release 11 (http://dcc.icgc.org).

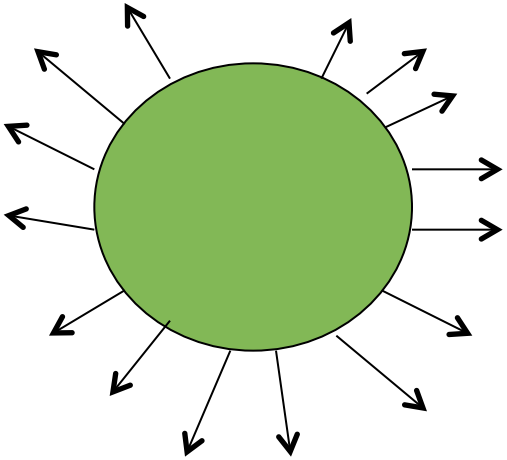
This release includes the first data release from four

..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 improve our chemical understanding
  - Of biological chemicals
  - Of chemicals which interfere with Biology
- 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

How?

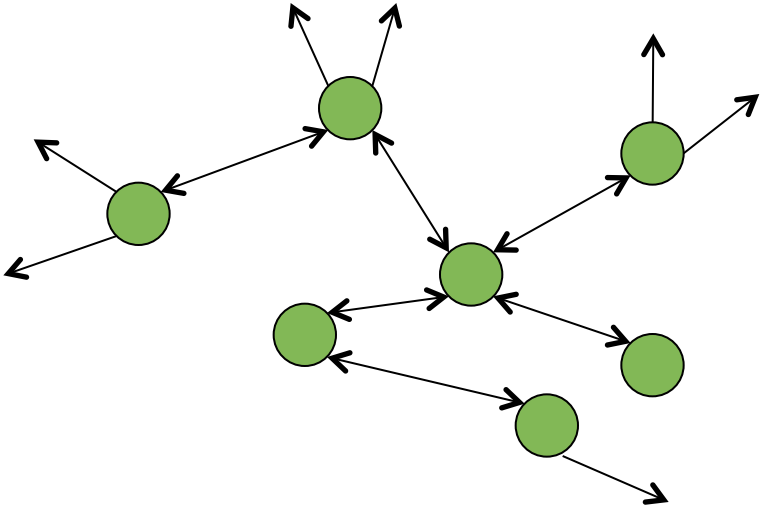
Fully Centralised



Pros: Stability, reuse,
Learning ease

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

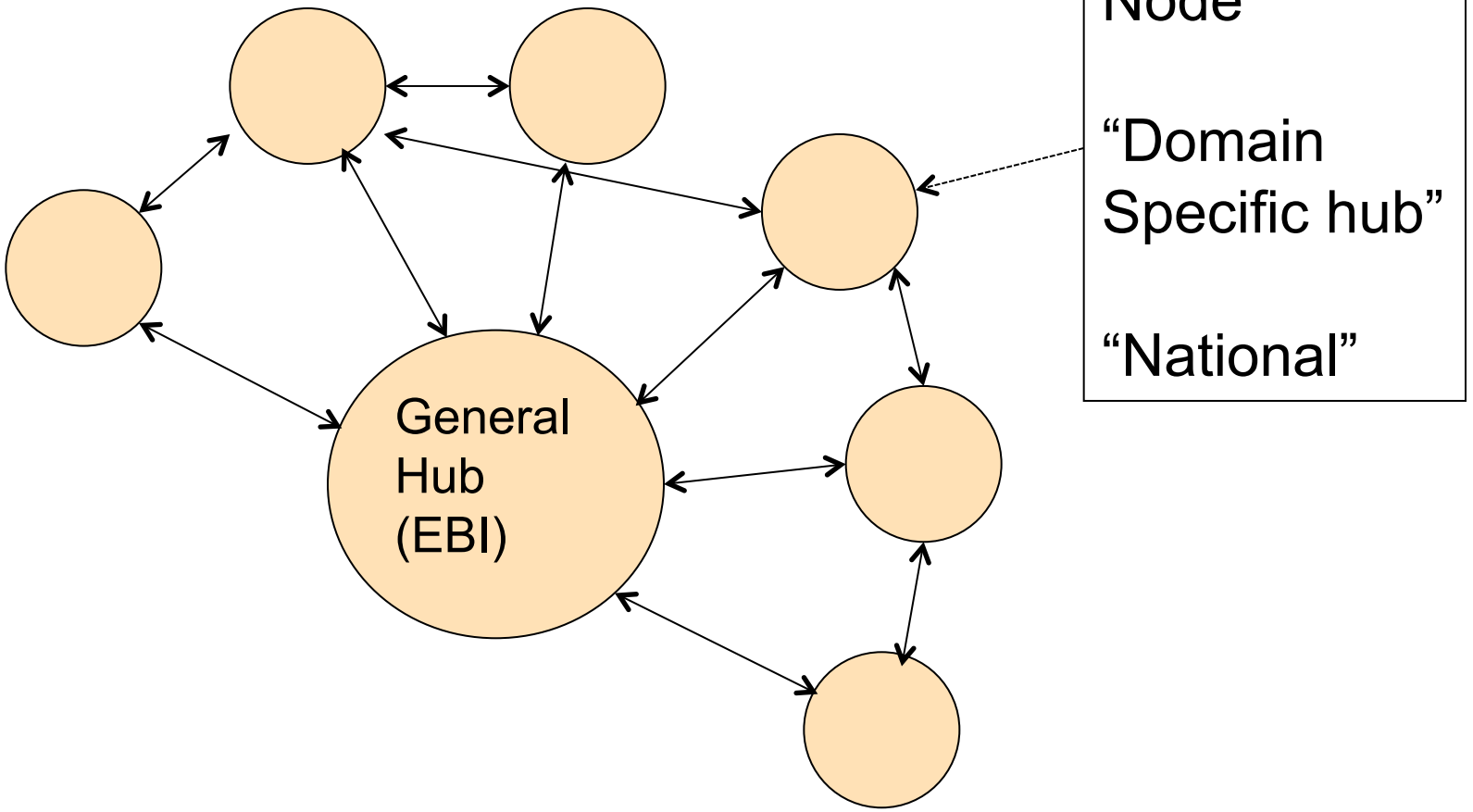
Fully Distributed



Pros: Responsive, Geographic
Language responsive

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

Robust network with a strong hub



European Life Sciences Infrastructure - ELIXIR

ELIXIR's mission

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

EMBL-EBI

Overlapping Networks

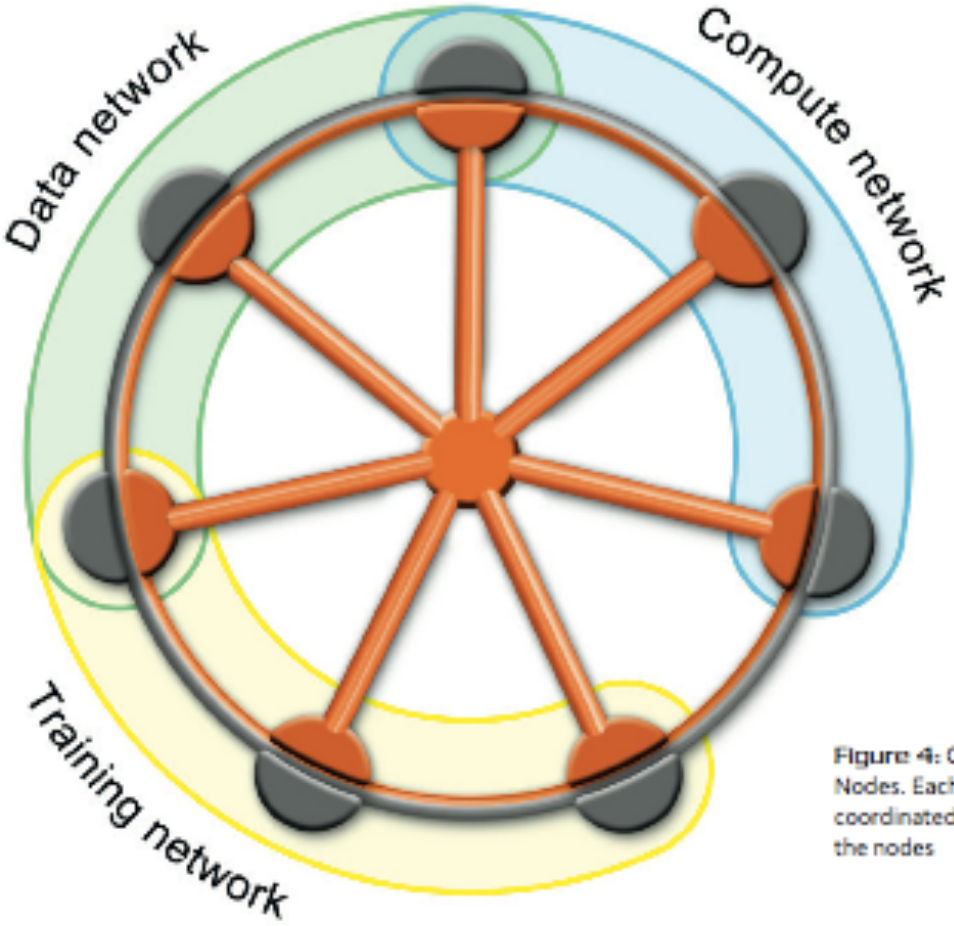


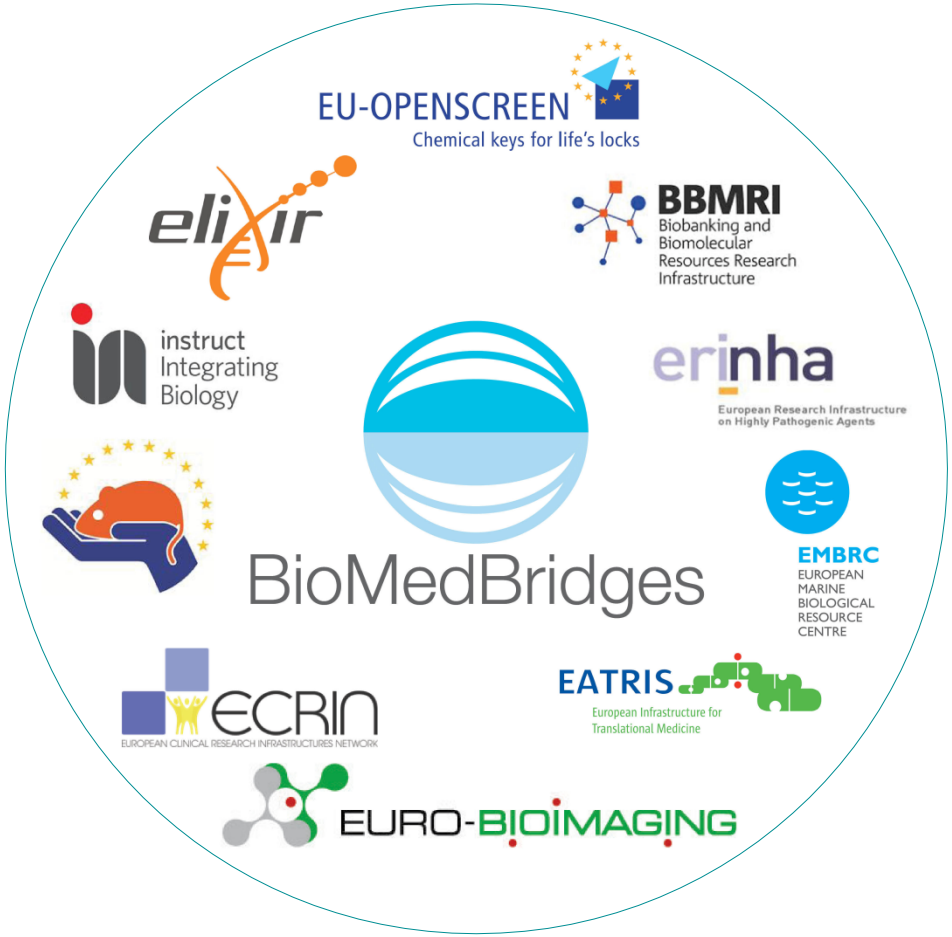
Figure 4: Coordinati
Nodes. Each network
coordinated by the hu
the nodes

ELIXIR memorandum or understanding signed by

Other infrastructures needed for biology

- EuroBioImaging
  - Cellular and whole organism Imaging
- BioBanks (BBMRI)
  - European populations – in particular for rare diseases, but also for specific sub types of common disease
- Mouse models and phenotypes (Infrafrontier)
  - A baseline set of knockouts and phenotypes in our most tractable mammalian model
- Robust molecular assays in a clinical setting (EATRIS)
  - The ability to reliably use state of the art molecular techniques in a clinical research setting

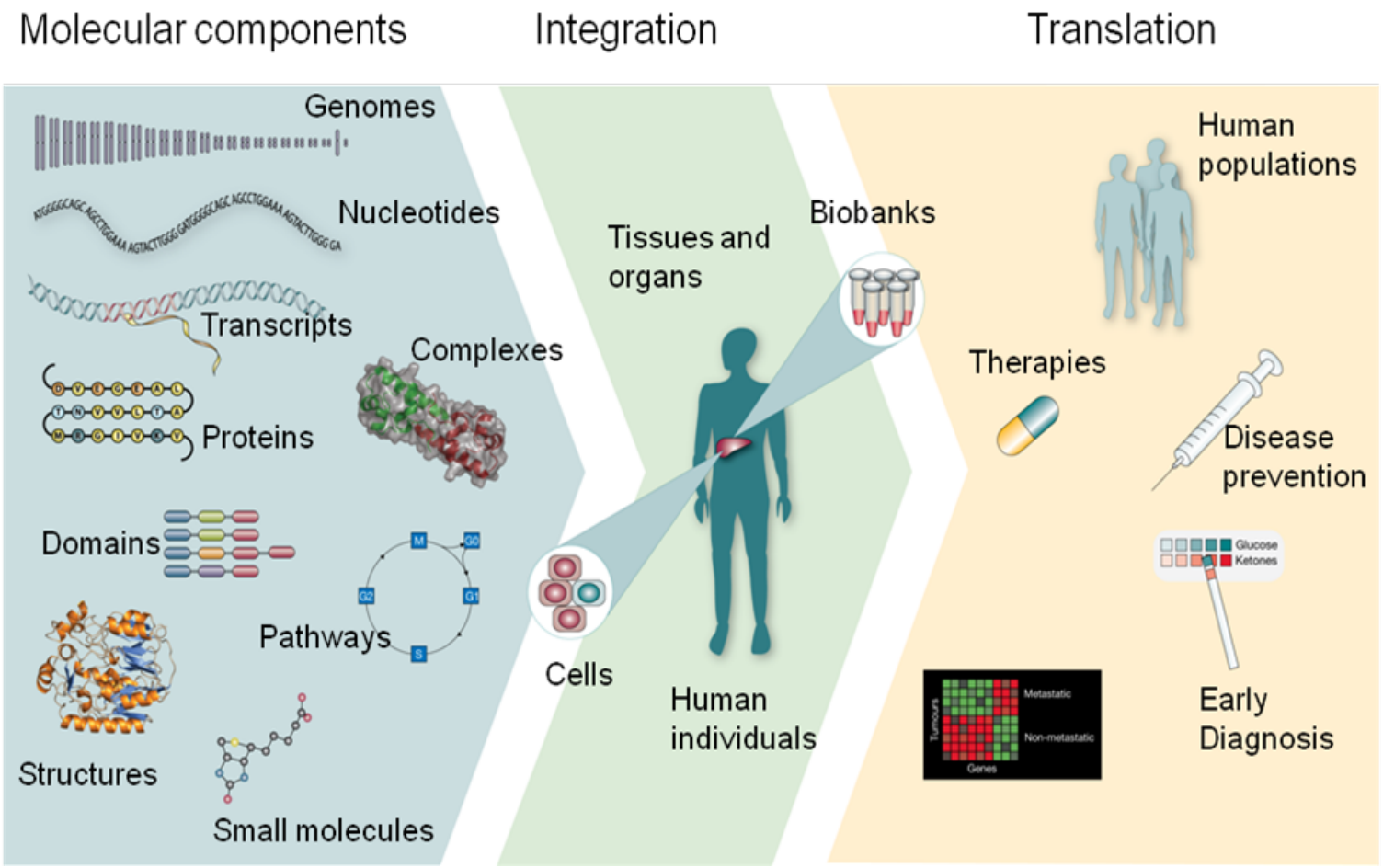
BioMedBridges - Ten new biomedical sciences research infrastructures: stronger through common links



WP11: Technology Watch

- Comprises representatives of GÉANT, DANTE, EGI.eu, PRACE & CERN as well as technical experts from the ESFRI BMS RIs
  - Brings together the technical experts
  - Facilitates adoption of e-Infrastructure technologies
  - Communicates advice from the ICT Infrastructures and the e-Infrastructures to the BioMedBridges partners

From Molecules to Medicine…



Funding and acknowledgements

- Ewan Birney, who provided many of the slides
- Functional Genomics Group at EBI
- IT Systems Group at EBI
- Funders
  - EMBL member countries
  - European Commission FP7 grants – EurocanPlatform, CAGEKID, ELIXIR, BioMedBridges, GEUVADIS, IMI EMIF